# ALLDATA 2016

The Second International Conference on Big Data, Small Data, Linked Data and Open Data

ISBN: 978-1-61208-457-2

**KESA 2016**

The International Workshop on Knowledge Extraction and Semantic Annotation

February 21 - 25, 2016

Lisbon, Portugal

**ALLDATA 2016 Editors**

Venkat Gudivada, East Carolina University, USA
Dumitru Roman, SINDEF/University of Oslo, Norway
Maria Pia di Buono, University of Salerno, Italy
Mario Monteleone, University of Salerno, Italy

# ALLDATA 2016

## Forward

The Second International Conference on Big Data, Small Data, Linked Data and Open Data (ALLDATA 2016), held between February 21-25, 2016 in Lisbon, Portugal continued a series of events bridging the concepts and the communities devoted to each of data categories for a better understanding of data semantics and their use, by taking advantage from the development of Semantic Web, Deep Web, Internet, non-SQL and SQL structures, progresses in data processing, and the new tendency for acceptance of open environments.

The volume and the complexity of available information overwhelm human and computing resources. Several approaches, technologies and tools are dealing with different types of data when searching, mining, learning and managing existing and increasingly growing information. From understanding Small data, the academia and industry recently embraced Big data, Linked data, and Open data. Each of these concepts carries specific foundations, algorithms and techniques, and is suitable and successful for different kinds of application. While approaching each concept from a silo point of view allows a better understanding (and potential optimization), no application or service can be developed without considering all data types mentioned above.

The conference had the following tracks:

- Big data
- Linked data

The conference also featured the following symposium:

- **KESA 2016*, The International Workshop on Knowledge Extraction and Semantic Annotation***

We take here the opportunity to warmly thank all the members of the ALLDATA 2016 technical program committee, as well as the numerous reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors that dedicated much of their time and effort to contribute to ALLDATA 2016. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations and sponsors. We also gratefully thank the members of the ALLDATA 2016 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope ALLDATA 2016 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in the area of all data. We also hope that Lisbon, Portugal, provided a pleasant environment during the conference and everyone saved some time to enjoy the beauty of the city.

### ALLDATA 2016 Advisory Committee

Mark Balas, Embry-Riddle Aeronautical University in Daytona Beach, USA
Yeh-Ching Chung, National Tsing Hua University, Taiwan
Dumitru Roman, SINDEF/University of Oslo, Norway
Venkat Naidu Gudivada, East Carolina University, USA
Andreas Schmidt, University of Applied Sciences Karlsruhe | Karlsruhe Institute of Technology, Germany
Fabrice Mourlin, Université Paris-Est Créteil Val de Marne, France
Dan Tamir, Texas State University, USA

### IARIA Advisory

Dumitru Roman, SINDEF/University of Oslo, Norway

### KESA 2016 Chairs

Maria Pia di Buono, University of Salerno, Italy
Mario Monteleone, University of Salerno, Italy
Annibale Elia, University of Salerno, Italy

# ALLDATA 2016

## Committee

### ALLDATA 2016 Advisory Committee

Mark Balas, Embry-Riddle Aeronautical University in Daytona Beach, USA
Yeh-Ching Chung, National Tsing Hua University, Taiwan
Dumitru Roman, SINDEF/University of Oslo, Norway
Venkat Naidu Gudivada, East Carolina University, USA
Andreas Schmidt, University of Applied Sciences Karlsruhe | Karlsruhe Institute of Technology, Germany
Fabrice Mourlin, Université Paris-Est Créteil Val de Marne, France
Dan Tamir, Texas State University, USA

### ALLDATA 2016 Technical Program Committee

Babak Abbasi, RMIT University, Australia
Rajeev Agrawal, North Carolina A&T State University, USA
Mary Akinyemi, University of Lagos, Nigeria
Hossein Asghari, Loyola Marymount University,  Los Angeles, USA
Simon Reay Atkinson, University of Sydney, Australia
Valentina E. Balas, Aurel Vlaicu University of Arad, Romania
Philipp Berger, Hasso-Plattner-Institut, Germany
Sandjai Bhulai, VU University Amsterdam, Netherlands
Simone Braun, CAS Software AG, Germany
Peter Breuer, CTO of Hecusys LLC, UK
Borut Čampelj, Ministry of Education, Science and Sport / School of Business and Management Novo mesto, Slovenia
Lijun Chang, University of New South Wales, Australia
Rachid  Chelouah, EISTI, France
Haifeng Chen, NEC Laboratories America, USA
Yue Chen, Florida State University, USA
Chi-Hung Chi, CSIRO, Australia
Tsan-Ming Choi (Jason), Hong Kong Polytechnic University, Hong Kong
Esma Nur Cinicioglu, Istanbul University, Turkey
Alexandru Costan, INRIA / INSA Rennes, France
Jorge R. Cuellar, Siemens AG, Germany
Cinzia Daraio, Sapienza University of Rome, Italy
Maria Cristina De Cola, IRCCS Centro Neurolesi "Bonino-Pulejo", Italy
Noel De Palma, University Joseph Fourier, France

Dorien DeTombe, International Research Society on Methodology of Societal Complexity, Netherlands
Mohamed Y. Eltabakh, Worcester Polytechnic Institute (WPI), USA
Serpil Erol, Gazi University, Turkey
Gustav Feichtinger, Vienna University of Technology, Austria
Denise Beatriz Ferrari, Instituto Tecnológico de Aeronáutica, Brazil
Paola Festa, Universita' degli Studi di Napoli "FEDERICO II", Italy
Yangchun Fu, University of Texas at Dallas, USA
Fausto P. Garcia, Universidad Castilla-La Mancha, Spain
Clemens Grelck, University of Amsterdam, Netherlands
Jerzy Grzymala-Busse, University of Kansas, USA
Venkat Naidu Gudivada, East Carolina University, USA
Ali Fuat Guneri, Yildiz Technical University, Turkey
Gür Emre Güraksin, Afyon Kocatepe University, Turkey
Ragib Hasan, University of Alabama at Birmingham, USA
Wen-Chi Hou, Southern Illinois University, USA
Joshua Zhexue Huang, Shenzhen University, China
Nilesh Jain, Intel Labs, USA
Hai Jiang, Arkansas State University, USA
David Kaeli, Northeastern University, USA
Atsushi Kanai, Hosei University, Japan
Sokratis K. Katsikas, University of Piraeus, Greece
Dukka KC, North Carolina A&T State University Greensboro, USA
Rasib Khan, University of Alabama at Birmingham, USA
Jinho Kim, Kangwon National University, South Korea
Alexander Lazovik, University of Groningen, Netherlands
Defu Lian, Big data research center - University of Electronic Science and Technology of China, China
Kun Liu, KTH Royal Institute of Technology, Sweden
Claudio Lucchese, ISTI-CNR, Italy
Victor E. Malyshkin, Russian Academy of Science, Russia
Cezary Mazurek, Poznan Supercomputing and Networking Center (PSNC), Poland
Roger Menday, Fujitsu, UK
Armando B. Mendes, Azores University, Portugal
Pablo Moscato, University of Newcastle, Australia
Hidemoto Nakada, National Institute of Advanced Industrial Science and Technology, Japan
Mirco Nanni, KDD Lab - ISTI-CNR, Italy
Sadegh Nobari, Skolkovo Institute of Science and Technology, Russia
Cyril Onwubiko, Research Series Ltd, UK
Mario Pavone, University of Catania, Italy
Yonghong Peng, University of Bradford, UK
Jaroslav Pokorny, Charles University, Czech Republic
Meikel Poess, Oracle Corporation, USA
Loganathan Ponnambalam, Institute of High Performance Computing - A*STAR, Singapore

Filip Radulovic, Universidad Politécnica de Madrid, Spain
Stefan Rass, Alpen-Adria-Universität Klagenfurt, Austria
Valderi Reis Quietinho Leithardt, Federal University of Rio Grande do Sul (UFRGS), Brazil
Dumitru Roman, SINDEF/University of Oslo, Norway
Paolo Romano, University of Lisbon / INESC-ID, Portugal
Ismael Sanz, Universitat Jaume I, Spain
Hiroyuki Sato, University of Tokyo, Japan
Stefanie Scherzinger, Regensburg University of Applied Sciences (OTH Regensburg), Germany
Ingo Schwab, Hochschule Karlsruhe, Germany
Sharad Sharma, Bowie State University, USA
Suzanne Michelle Shontz, University of Kansas, USA
Patrick  Siarry, Université de Paris 12, France
Srivathsan Srinivasagopalan, Cognizant, USA
Bela Stantic, Griffith University, Australia
Yun Tian, California State University, Fullerton, USA
Arthur Tórgo Gómez, Universidade do Vale do Rio dos Sinos (UNISINOS), Brazil
Henry Tufo, University of Colorado at Boulder, USA
Antonino Tumeo, Pacific Northwest National Laboratory, USA
Bhekisipho Twala, University of Johannesburg, South Africa
Liqiang Wang, University of Central Florida, USA
Hironori Washizaki, Waseda University, Japan
Ouri Wolfson, University of Illinois, USA
Chase Wu, New Jersey Institute of Technology, USA
Yinglong Xia, IBM Research, USA
Feng Yan, College of William and Mary, USA
Hongzhi Yin, University of Queensland, Australia
Feng Yu, Youngstown University, USA
Stefan Zander, FZI Research Center for Information Technology, Germany
Daqiang Zhang, School of Software Engineering - Tongji University, China
Vincent Zheng, Advanced Digital Sciences Center, Singapore
Sotirios Ziavras, New Jersey Institute of Technology, Newark, USA

**KESA 2016**

**IARIA Advisory**

Dumitru Roman, SINDEF/University of Oslo, Norway

**KESA 2016 Chairs**

Maria Pia di Buono, University of Salerno, Italy
Mario Monteleone, University of Salerno, Italy
Annibale Elia, University of Salerno, Italy

**KESA 2016 Technical Program Committee**

Rodrigo Agerri, University of the Basque Country, Spain
Ahmet Aker, University of Sheffield, UK
Flora Amato, University of Naples, Italy
Sergey Balandin, Tampere University of Technology / FRUCT Group, Finland
Abel Browarnik, Tel Aviv University, Israel
Maaike de Boer, TNO and Radboud University, Netherlands
Antoine Doucet, University of La Rochelle, France
Kavallieratou Ergina, University of the Aegean, Greece
Xavier Blanco Escoda, Universitat Autònoma de Barcelona, Spain
Zhisheng Huang, VU University Amsterdam, Netherlands
Pavel Klinov, Complexible Inc, USA
Kristina Kocijan, University of Zagreb, Croatia
Gijs Koot, TNO, Netherlands
Giuseppe Laquidara, X23 Ltd., Italy
Kun Lu, University of Oklahoma, USA
Antonino Mazzeo, University of Naples, Italy
Thiago Pardo, University of São Paulo, Brazil
Jan Radimsky, University of South Bohemia, Czech Republic
Giovanni Semeraro, University of Bari, Italy
Max Silberztein, University de Franche-Comté, France

## Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

## Table of Contents

# AScale: Simple and fast ETL+Q scaling for small and big data

Pedro Martins, Maryam Abbasi, Pedro Furtado
*University of Coimbra*
*Department of Informatics*
*Coimbra, Portugal*
*email: {pmom, maryam, pnf}@dei.uc.pt*

***Abstract*—In this paper, we investigate the problem of providing scalability (out and in) to Extraction, Transformation, Load (ETL) and Querying (Q) (ETL+Q) process of data warehouses. In general, data loading, transformation, and integration are heavy tasks that are performed only periodically, instead of row by row. Parallel architectures and mechanisms are able to optimize the ETL process by speeding up each part of the pipeline process as more performance is needed. We propose parallelization solutions for each part of the ETL+Q, which we integrate into a framework, that is, an approach that enables the automatic scalability and freshness of any data warehouse and ETL+Q process. Our results show that the proposed system algorithms can handle scalability to provide the desired processing speed in big-data and small-data scenarios.**

***Keywords-Algorithms; architecture; Scalability; ETL; freshness; high-rate; performance; scale; parallel processing.***

## I. INTRODUCTION

ETL tools are special purpose software used to populate a data warehouse with up-to-date, clean records from one or more sources. The majority of current ETL tools organize such operations as a workflow. At the logical level, the E (extraction) can be considered as a capture of data flow from the sources, normally more than one with high-rate throughput. Then, we have T representing transformation and cleansing of data. This corresponds to modifying data so that it will conform to an analysis schema. The L (load) represents loading the data into the data warehouse, where the data is stored to be queried and analyzed. When implementing these types of systems, besides the necessity to create all these steps, the user is required to be aware of scalability requirements that the ETL+Q (ETL and queries) might raise for this specific scenario.

When defining the ETL+Q the user must have in mind the existence of data sources, where and how the data is extracted to be transformed (e.g., completed, cleaned, validated), the loading into the data warehouse, and finally the data warehouse schema, each of these steps requires different processing capacities, resources, and data treatment. However, in some applications scenarios, (e.g., near-real-time monitoring of telecom, energy distribution or stock market) ETL can be demanding in terms of performance. Most of the time because the data volume is too large and one single, extraction, transform, loading or querying node is not sufficient. Thus, more nodes must be added to extract the data and extraction policies from the sources must be created (e.g., round-robin OR on-demand). The other phases, transformation, and load must also be scaled.

After extraction, data must be re-directed and distributed across the available transformation nodes. Again since transformation involves heavy duty tasks (heavier than extraction), more than one node should be necessary to assure acceptable execution/transformation times.

After the data is transformed and ready to be loaded, the load period must be scheduled (e.g., every night, every hour, every minute) and load time controlled (e.g., maximum load time = 5 hours). This means that, between the transformation and load process, the data must be held somewhere.

Regarding the data warehouse, in some application scenarios the entire data will not fit into a single node, and if it fits, it will not be possible to execute queries within acceptable time ranges. Thus, more than one data warehouse node is necessary with a specific schema which allows distributing, replicate, and finally query the data within an acceptable time frame.

In this paper, we study how to provide ETL+Q scalability with ingress high-data-rate in big and small data warehouses. We propose a set of mechanisms and algorithms, to parallelize and scale each part of the entire ETL+Q process, which is included in an auto-scale (in and out) ETL+Q framework. This framework is based on time bounds for the parts of the ETL+Q and/or the global ETL process, automatically scaling, to assure the desired time bounds.

The presented results prove that the proposed monitoring mechanisms and detection algorithms are able to scale-out when necessary.

In Section II, we present relevant related work in the field. Section III, we describe the architecture of the proposed system. Section IV explains the main algorithms which allow to scale-out when necessary. Section V shows the experimental results obtained when testing the proposed system. Finally, Section VI concludes the paper and discusses future work.

## II. RELATED WORK

Works in the area of ETL scheduling include efforts towards the optimization of the entire ETL workflow [6] and of individual operators in terms of algebraic optimization

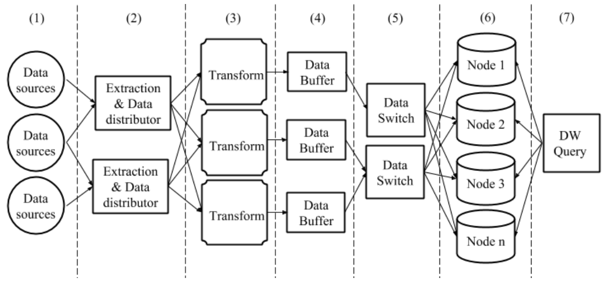


Figure 1. Total automatic ETL+Q scalability

(e.g., joins or data sort operations). The work [3] deals with the problem of scheduling ETL workflow at the data level and in particular scheduling protocols and software architecture for an ETL engine in order to minimize the execution time and the allocated memory needed for a given ETL workflow. The second aspect in ETL execution that the authors address is how to schedule flow execution at the operations level (blocking, non-parallelizable operations may exist in the flow) and how we can improve this with pipeline parallelization [2].

The work [4] focuses on finding approaches for the automatic code generation of ETL processes which is aligning the modeling of ETL processes in the data warehouse with Model Driven Architecture (MDA) by formally defining a set of QVT (Query, View, Transformation) transformations.

Related problems studied in the past include the scheduling of concurrent updates and queries in real-time warehousing and the scheduling of operators in data streams management systems. However, we argue that a fresher look is needed in the context of ETL technology. The issue is no longer the scalability cost/price, but rather the complexity it adds to the system. Previews presented recent works in the field do not address in detail how to scale each part of the ETL+Q and do not regard the automatic scalability to make ETL scalability easy and automatic. The authors focus on mechanisms to improve scheduling algorithms and optimizing workflow and memory usage. In our work, we assume that scalability in a number of machines and quantity of memory is not the issue. We focus on offering scalability for each part of the ETL pipeline process, without the nightmare of operators relocation and complex execution plans. Thus, in our work, we focus on scalability based on generic ETL process to provide the users desired performance with minimum complexity and implementations. In addition, we also support queries execution.

## III. Architecture

In this section, we describe the main components of the proposed architecture for ETL+Q scalability.

Figure 1 depicts the main processes needed to support total ETL+Q scalability with specific time bounds.

(1) Represents the data sources from where data is extracted from the system.
(2) The data distributor(s) is responsible for forwarding or replicating the raw data to the transformer nodes. The distribution algorithm to be used is configured and enforced at this stage. The data distributors (2) should also be parallelizable if needed, for scalability reasons.
(3) In the transformation nodes the data is cleaned and transformed to be loaded into the data warehouse. This might involve data look-ups to in-memory or disk tables and further computation tasks. In Figure 1 the transformation (3) is parallelized for scalability reasons.
(4) The data buffer can be in memory, disk file (batch files) or both. In periodically configured time frames/periods, data is distributed across the data warehouse nodes.
(5) The data switches are responsible for distributing (pop/extract) data from the ”Data Buffers” and set it for load into the data warehouse, which can be a single-node or a parallel data warehouse depending on configured parameters (e.g., load time, query performance).
(6) The data warehouse can be in a single node, or parallelized by many nodes. If parallelized, the ”Data Switch” nodes will manage data placement according to configurations (e.g., replication and distribution). Each node of the data warehouse loads the data independently from batch files.
(7) Queries (7) are rewritten and submitted to the data warehouse nodes for computation. The results are then merged, computed and returned.

The main concepts, we propose are the individual ETL+Q scalability mechanisms of each part of the ETL+Q pipeline. By offering the solution to scale each part independently, we provide a solution to obtain configurable performance. Then, in future work based on user configuration parameters, a framework using these components, scales automatically the ETL+Q when necessary.

## IV. Scaling algorithms

In this section, we describe the algorithms which allow the framework to scale-in and scale-out each part of the ETL and Query process. For each part that we design for later, automatic scale in and out we explain the scaling algorithms.

### *A. Extraction & data distributors - Scale out*

Depending on the number of existing sources and data generation rate and size, the nodes that process the extraction of the data from the sources might need to scale. The addition of more ”extraction & data distributors” (2) depends on if the current number of nodes is being able to extract and process the data with the correct period and inside the maximum extraction time (without delays). For instance, if the extraction period is specified as every 5 minutes and the extraction duration is 10 seconds, every 5 minutes the ”Extraction & Data distributor” nodes cannot spend more than 10 seconds extracting data, if so, a scale-out is needed. By scaling out the extraction, nodes will have fewer data to extract/process and more concurrent extraction, leading to a performance improvement. Listing 1 pseudo-code describes the algorithm used to scale, independently of the used extraction method.

Listing 1. Extraction scalability

```
startTime = getCurrentTime();
source = requestSourceToExtractData();
size = requestSizeToExtract();
data = requestSourceExtraction(source, size);
endExtractionTime = getCurrentTime();
extractionTime.add(endExtractionTime−startTime);
for (all in extractionTime()){
    if (extractionTime > getMaxExtractionTime()){
        processScaleOut();
        break;
    }
}
for (all in extractionTime()){
    if (extractionTime > getExtractionPeriod()){
        processScaleOut();
        break;
    }
}
```

### B. Extraction & data distributors - Scale in

To save resources when possible, the nodes that perform the data extraction from the sources can be set in standby or removed. This decision is made based on the last execution times. If previous execution times of at least two or more nodes are less than half of the configured maximum, one of the nodes is set on standby or removed, and the other one takes over. Listing 2 pseudo code describes the used algorithm.

Listing 2. Extraction and data distribution

```
lowLoadNodes = 0;
nodesID = null;
for (all nodes){
    if (processingTime() < getExtractionFrequency() / 2){
        lowLoadNodes++;
        nodesID.add(nod eID);
    }
    if (extractionTime() < getMaxExtractionTime() / 2){
        lowLoadNodes++;
        nodesID.add(nodeID);
    }
}
if (lowLoadNodes >= 2){
    setNodeToSandBy(nodesID.getFirst());
}
```

### C. Transform - Scale-out

The transformation process is critical. If the transformation is running slow, data extraction at the refereed rate may not be possible, and information will not be available for loading and querying when necessary. The transformation step has an important queue, used to determine when to scale the transformation phase. If this queue reaches a limit size (by default 50%) then it is necessary to scale, because the actual transformer is not being able to process all data that is arriving. Another mechanism used to scale the transformation process is the user-configured maximum transformation execution time. If this time is exceeded then, the transformation must be scaled-out. Listing 3 pseudo code describes the used algorithm.

Listing 3. Transformation scale-out

```
limitSize = getLimitSize(); //by default 50%;
for (all nodes){
    currentQueueSize = queue.getSize();
    if (currentQueueSize > limitSize){
        addTrasnformerNode();
        break;
    }
    if (transformationTime > getMaxTransformTime()){
        addTrasnformerNode();
        break;
    }
}
```

### D. Transform - Scale in

The size of all queues is analyzed periodically. If this size at a specific moment is less than half of the limit size for at least two nodes and the average transformation time of at least two nodes is half of the specified then, one of those nodes is set on standby or removed, and another one of the low load nodes takes over. Listing 4 pseudo code describes the used algorithm.

Listing 4. Transformation scale-in

```
limitSize = getLimitSize() / 2; //by default 25%;
maxTransformTime = getMaxTransformTime();
count = 0;
for (all nodes){
    currentQueueSize = node.queue.getSize();
    currentTransformTime = node.getAvgTransformTime();
    if (currentQueueSize <= limitSize &&
      currentTransformTime <= maxTransformTime/2){
        count++;
        if (count >= 2){
           setNodeToSandBy(nodeID);
           break;
        }
    }
}
```

### E. Data buffer - Scale

The data buffer nodes scale-out based on the incoming memory queue size and the storage space available to hold data. Low data warehouse load frequency will require data buffers with storage space to hold the data until the scheduled load time. Thus, the data buffers scale dynamically as more storage space is necessary. Another scale-out situation is when the available incoming memory queue becomes above a certain threshold (by default 50%). This means that the data ingress rate is higher than the data swap speed, thus, nodes must scale-out in order to not lose data. By user request, the data buffers can also scale-in. In this case, the system will allow it if the data from any data buffer can be fitted inside other data buffer.

### F. Data switch - Scale

These nodes scale based on configured data rate limits. If after a data load process occurs the average limit extraction data rate is equal or above a certain limit, then these nodes are set to scale. The data switches can also scale-in. In this case, the system will allow it if the average data rate from the previews load period is less than the maximum supported by each data switch.

### G. Data Warehouse - Scale

The data warehouse scalability is detected after each load process. The loading process might include among other operations: destroy indexes, load data, update materialized view, and rebuild indexes. The data warehouse load process has a limit time to be executed every time it starts. If that limit time is exceeded then, the data warehouse must scale. Listing 5 pseudo code describes the used to scale the data warehouse when the load process occurs.

Listing 5. Data warehouse scale

```
startTime = getCurrentTime();
data = getData(size);
preLoadTask();
process(data);
posLoadTask();
startLoad();
endTime = getCurrentTime();
if (endTime - startTime > getMaxLoadTime(){
    dataWarehouseScaleOut();
}
```

The data warehouse scalability is not only based on the load & integration speed requirements, but also on the queries desired maximum execution time. The faster queries need to execute, more nodes will be necessary. Listing 6 pseudo code describes the used algorithm:

Listing 6. Data warehouse scale, integration speed

```
execute(queries);
avgTime = getQueryAverageExecutionTime();
desiredExecutionTime = getDesiredExecutionTime();
if (avgTime < deisredExecutionTime){
    dataWarehouseScaleOut();
}
```

Because it is a computationally expensive operation, when an alarm is raised (the data warehouse needs to scale) the data warehouse nodes scale-in and scales-out, can only be triggered by user request and iff the average query execution time and the average load time respect the conditions 1 and 2 (where *n* represents the number of nodes):

$$\frac{(n-1) \times avgQueryTime}{n} \leq desiredQueryTime \quad (1)$$

and

$$\frac{(n-1) \times avgLoadTime}{n} \leq maxLoadTime \quad (2)$$

Every time the data warehouse scales-out or scales in the data inside the nodes needs to be re-balanced. The default re-balance process to scale-out is based on the phases:

- Replicate dimension tables;
- Extract information from nodes;
- Load the extracted information into the new nodes.

## V. Experimental setup and results

In this section, we test the ability of the proposed auto-scale framework to automatic scale-out the ETL process when more performance is necessary to provide the desired results. For the purpose of these tests, we simulated the launch of an ETL system only concerning a single server machine. In this setup the considered ETL process consists on converting the TPC-H [1] benchmark data generator into the Star Schema Benchmark (SSB) [5], and execute the SSB queries. For all tests, we used equal nodes, with intel i5 3.00GHz, 16 GB of RAM and 1TB of disk. By applying the described algorithms, we observed how it scales to provide the desired performance. In the next sections, we demonstrate how each part of the ETL and Query execution scales-out.

### A. Data extraction nodes scalability

Considering that, we have data sources and extraction nodes to extract data. When the data flow is too high a single data node can not handle all ingress data. In this section, we study how the extraction nodes scale to handle different data rates. The extraction process uses an on-demand approach to extract data, where an ”automatic scaler” process orders the nodes to extract data from sources. There is a configured maximum allowed extraction time and a extraction frequency, represented by the Equations 3 and 4. If any of them is not respected the system is set to scale-out.

$$\underset{extractionTime}{\max} < \underset{desiredExtractionTime}{\max} \quad (3)$$

$$\underset{extractionTime}{\max} < ExtractionFrequency \quad (4)$$

Figure 2 shows: In the left Y axis is the average extraction time in seconds; In the right Y axis is the number of nodes; The X axis is the data-rate; Black line represents the extraction time; Grey line represents the number of nodes; The maximum allowed extraction time was set to 1 second maximum extraction time, with periodic extraction of 5 seconds.

As we can conclude from Figure 2 experimental results, every time the maximum allowed extraction time has exceeded the system requested an additional extraction node to improve the extraction performance.

### B. Transformation scalability

During the ETL process after data extraction, it is set for the transformation. In our tests the transformation consists

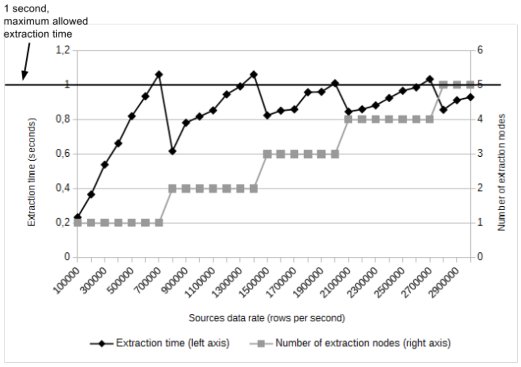


Figure 2. Extraction scalability

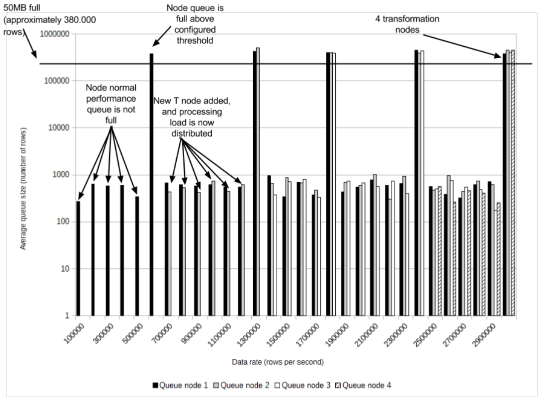


Figure 3. Automatic transformation scalability. For different data rates, for 60 minutes processing per data rate. Transformation threshold set to 50MB, approximately 380.000 rows

on converting the TPC-H dataset into SSB format. Because this process is computationally heavy it is often required to scale the transformation nodes. Each transformation node has an entrance queue for ingress data and an automatic scale monitors all queues. Once it detects that a queue is full above a certain (configured) threshold it starts the scaling process, this means that $Rate_{extract} \geq Rate_{transform}$.

Figure 3 shows: Y axis, average queue size in number of rows; X axis, the data rate in rows per second; Each plotted bar represents a node queue size, up to 4 nodes; The limit queue size to trigger the scale mechanisms was set at 50MB, approximately 380.000 rows; The maximum transformation time for each row was set to 1 second. Each measure is the average queue size of 60 seconds run.

During the experimental tests, as depicted in Figure 3, the maximum allowed transformation time was never exceeded. However, the queues size increased while increasing the data rate, allowing to show that the proposed approach is efficient to scale the transformation nodes.

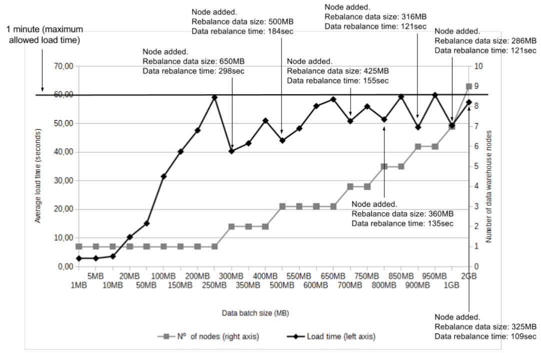


Figure 4. Data warehouse load scalability

### C. Data Buffer nodes

These nodes hold the transformed data until it is loaded into the data warehouse. During all our tests, we used a single machine with 16GB memory and 1TB disk, all available to be used. If the available storage space becomes full then, the automatic scale sets the system to scale the Data Buffer node (add one more node). However, during our tests, we never had the necessity to do so since all transformed data could fit into memory until the next load (into the data warehouse) period.

### D. Data warehouse scalability

In this section, we test the data warehouse scalability, which can be triggered or by the load process (because it is taking too long), or because of the queries time (they are taking more time than the desired execution time). If the maximum configured load time is exceeded, the data warehouse is set to scale.

Experimental results from Figure 4 show: Left Y axis, average load time in seconds; Right Y-axis, number of data warehouse nodes; X axis, data batch size in MB; The maximum allowed load time, set to 60 seconds; Each time a data warehouse (scales) node is added, we show the data size that was moved into the new node and the required time in seconds (re-balance time).

Based on our experimental results, we conclude that the proposed method to scale the data warehouse when the bottleneck is related to the load time is efficient, improving the overall load performance. Note that, every time a new node was added the data warehouse required to be re-balanced (data distributed by the nodes evenly). This process requires 3 steps, first extract (in parallel) the data from the existent nodes, second load the data into the new node, third load the new data (distributed and parallel) in batch and check if the load time is lower than the maximum allowed load time.

### E. Query scalability

When running queries, if the maximum desired query execution time (i.e. configured parameter) is exceeded then,

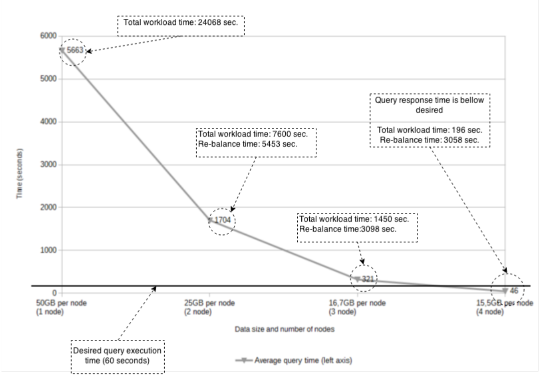


Figure 5. Data warehouse scalability, workload 1

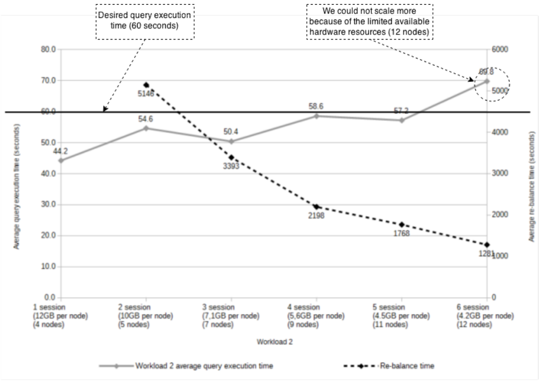


Figure 6. Data warehouse scalability, workload 2

the data warehouse is set to scale in order to offer more query execution performance. The following workloads were considered to test the proposed system:

- Workload 1;
  - 50GB total size;
  - Execute Q1.1, Q2.1, Q3.1, Q4.1 randomly chosen;
  - Desired execution time per query: 5 minutes (300 seconds).
- Workload 2 (as workload 1 but with more sessions);
  - **1 to 8 sessions;**

Workload 1 studies how the proposed mechanisms scale the data warehouse when running queries. Workload 2 studies the scalability of the system when running queries and the number of simultaneous sessions (e.g., the number of simultaneous users) increases. Both workloads with the objective to deliver the configured execution time per query (300 seconds).

### *F. Query scalability - Workload 1*

Figure 5 shows: The experimental results for workload 1; Y axis, average execution time in seconds using a logarithmic scale; X axis the data size per node and the current number of nodes; The horizontal line over 300 seconds represents the desired query execution time;

The results from Figure 5, show that the proposed system can detect and scale the data warehouse nodes until the average query execution time is the desired.

### *G. Query scalability - Workload 2*

Figure 6 shows: The experimental results for workload 2; Y axis, average execution time in seconds using a logarithmic scale; X axis the number of sessions, the data size per node and the number of nodes; The horizontal line over 300 seconds represents the desired query execution time; The last result does not respect the desired execution time because of the limited resources for our tests, 12 nodes.

In Figure 6, we show that while the number of simultaneous sessions increases the system scales the number of nodes in order to provide more performance, thus, the query average execution time follows the configured parameters. As our experimental results show the proposed system scales efficiently to provide the desired performance.

## VI. Conclusions & Future work

In this work, we propose mechanisms and algorithms to achieve automatic scalability for complex ETL+Q, offering the possibility to the users to think solely in the conceptual ETL+Q models and implementations for a single server. The tests demonstrate that the proposed techniques are able to scale-out. Future work will investigate an auto-scale framework for scale-out and scale in any ETL+Q and, at the same time, providing data freshness and support for near-real-time data stream processing.

## References

[1] T. P. P. Council. Tpc-h benchmark specification. *Published at http://www. tcp. org/hspec. html*, 2008.

[2] R. Halasipuram, P. M. Deshpande, and S. Padmanabhan. Determining essential statistics for cost based optimization of an etl workflow. In *EDBT*, pages 307–318, 2014.

[3] A. Karagiannis, P. Vassiliadis, and A. Simitsis. Scheduling strategies for efficient etl execution. *Information Systems*, 38(6):927–945, 2013.

[4] L. Muñoz, J.-N. Mazón, and J. Trujillo. Automatic generation of etl processes from conceptual models. In *Proceedings of the ACM twelfth international workshop on Data warehousing and OLAP*, pages 33–40. ACM, 2009.

[5] P. O'Neil, E. O'Neil, X. Chen, and S. Revilak. The star schema benchmark and augmented fact table indexing. In *Performance Evaluation and Benchmarking*, pages 237–252. Springer, 2009.

[6] A. Simitsis, K. Wilkinson, U. Dayal, and M. Castellanos. Optimizing etl workflows for fault-tolerance. In *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*, pages 385–396. IEEE, 2010.

# Informed Consent and Privacy of De-Identified Information and Estimated Data

## Lessons from Iceland and the United States in an Era of Computational Genomics

Donna M. Gitter
Department of Law
Baruch College, City University of New York
55 Lexington Avenue
New York, New York 10010
USA
e-mail: Donna.Gitter@baruch.cuny.edu

*Abstract*—**Advances in bioinformatics and computational genomics necessitate reexamination of the principles of privacy and informed consent. The law of informed consent requires that research subjects give their consent to participation in biomedical research. In the current age of bioinformatics and computational genomics, however, researchers are in many cases able to use genetic and genealogical data from research subjects who did agree to participate in genetic testing, in order to make educated guesses about the genetic profile of the subjects' relatives, who did not volunteer to participate. The law of informed consent does not address the use of estimated data, given that it was not possible before the advent of computational genomics to conduct "in silico" research. In considering whether to extend informed consent protection to those to whom "estimated data" is extrapolated, it is useful to consider currently proposed changes to the law of informed consent in the U.S. These proposed changes arise from the notion that biospecimens are increasingly considered intrinsically identifiable, and therefore individuals ought to be asked for their informed consent before the use of even de-identified specimens. Moreover, the recently revised Genomic Data Sharing (GDS) Policy of the U.S. National Institutes of Health (NIH) goes even further to require informed consent not only for use of biospecimens and identifiable private information, but also for genomic or other data, even if it is de-identified. It follows logically that those who do not agree to participate in biomedical research, but from whom estimated data are gleaned, ought to be asked for their informed consent.**

***Keywords-bioinformatics; computational genomics; privacy; informed consent.***

## I. INTRODUCTION

Advances in bioinformatics and computational genomics necessitate reexamination of the principles of privacy and informed consent. Since the formation of the Nuremberg Code, which developed as a result of the Nazi War Crimes Tribunal and was the first internationally recognized code of research ethics, medical researchers must recognize protections for human research subjects. The primary tenet of the Nuremberg Code is that "The voluntary consent of the human subject is absolutely essential." In the current age of bioinformatics and computational genomics, however, researchers are in many cases able to use genetic and genealogical data from research subjects who did agree to participate in genetic testing, in order to make educated guesses about the genetic profile of the subjects' relatives, who did not volunteer to participate. This estimated data can then be combined with health records of the non-volunteers in order to conduct genetic research, often termed "in silico" biology, without their informed consent. Researchers use these technologies to calculate the probability that an individual carries a particular genetic variant, without sequencing that person's deoxyribonucleic acid (DNA), thereby developing estimated data for inclusion in research databases.

Section II of this paper considers the use of computational genomics in Iceland to conduct research using estimated data from individuals without their informed consent, noting that this conflicts directly with a legal trend toward enhanced recognition of the privacy rights and autonomy of research participants, as reflected in proposed changes to the law and policy of informed consent in the United States. Section III then considers proposed changes in the U.S. enhancing informed consent protection for research with de-identified materials, and advocates for the same level of protection for estimated data, in keeping with traditional norms of informed consent.

## II. THE USE OF COMPUTATIONAL GENOMICS IN ICELAND

Controversial methods of computational genomics, particularly the use of estimated genetic data, are particularly effective in Iceland, an island nation with detailed genealogical records and a population of approximately 320,000 citizens who are considered to be genetically homogeneous. The intimacy of this small country is made evident by the existence of a smart-phone app in Iceland that permits individuals to determine whether they are related to another person whom they are considering dating.

In light of Iceland's genetic homogeneity and the availability of detailed genealogical information, in 1996 Icelander Dr. Kari Steffansson founded the company deCODE Genetics in order to use Iceland's population to pioneer genetic population studies. In 1999, the Icelandic government granted deCODE an exclusive 12-year license to build a Health Sector Database to hold centralized health records of its entire population [1]. The plan incited much controversy due to the presumption that citizens of Iceland would be deemed to consent to participate unless they actively opted out. In November 2003, the Supreme Court of Iceland disrupted deCODE's plans by ruling in favor of Ragnhildur Gudmundsdottir, an eighteen-year-old student, holding that she could prevent the transfer to the database of her deceased father's health records. The court held that the records in the database might allow her to be identified as an individual at risk of a heritable disease, even though the data would be anonymous and encrypted. The court noted that this risk was heightened by the fact that the Health Sector Database would allow information to be linked with data from other genetic and genealogical databases [1].

DeCODE then pursued another strategy, using estimated data to create a research database to find genetic sequences linked to diseases. Using DNA and clinical data from more than 120,000 research volunteers, deCODE analyzed their DNA sequences for a selection of slight variations called single nucleotide polymorphisms (SNPs), which are the most common genetic variations among individuals and some of which may prove important in the study of human health.

Using a relatively new technique, deCODE geneticists calculate the probability that an individual carries a particular genetic variant without actually sequencing that person's DNA. For example, deCODE was able to use its whole genome sequencing of the DNA of approximately 2,500 research participants in order to extrapolate the genomes of many more individuals. When deCODE identified a genetic variant of interest among the 2,500 whole genomes, the company used the more limited SNP data that it had amassed from its 120,000 volunteers in order to impute, with 99 per cent accuracy, whether any among the 120,000 also carried the mutations [6]. As noted by one source, "if your mother had been in the hospital for a stroke and agreed to participate in a clinical study, while her brother had volunteered his DNA, deCODE would be able to predict *your* likelihood of a genetic disposition for stroke [5]."

While other researchers are using the same technique as deCODE, the company's unique approach is to combine the known and estimated genotypes for its research participants with its genealogical database, thereby permitting deCODE to estimate what it calls the "in silico" genotypes of close relatives of the volunteers whose SNPs were analyzed. This permits deCODE to infer data about 200,000 living and 80,000 deceased Icelanders, who have not consented to participate in deCODE's studies. Further, it could give the company genotypes for the largely consanguineous population of 320,000 people in its entirety. Researchers can then determine whether a variant in a DNA sequence found by fully sequencing the DNA of a small group likewise appears in a larger population in the same proportion [6].

The company has used these estimated genotypes for individuals as controls in its studies and also combined them with health records for patients who were involved in a disease study in Iceland but whose DNA has not been sampled. Using estimated data, deCODE published six papers between 2011 and 2013 in the prestigious journals *Nature*, *Nature Genetics*, and the *New England Journal of Medicine*, linking specific genetic mutations to risks of diseases. DeCODE's drug discovery efforts were less successful, however, and the company declared bankruptcy in 2009. In December 2012, Amgen purchased the company for $415 million [6].

In 2012, deCODE planned to use its strategy as part of a new study. Having imputed the genotypes of the close relatives of the volunteers whose SNPs had been fully catalogued, deCODE intended to collaborate with Iceland's National Hospital to link these relatives to certain hospital records for individuals, such as surgery codes and prescriptions. On May 28. 2013, Iceland's Data Protection Authority (DPA) denied this request, on the grounds that it would violate the relatives' privacy unless they gave their informed consent. The DPA gave deCODE until November 2013 to demonstrate that it obtained consent [10].

DeCODE ultimately found a means of working around the requirement of informed consent, describing it in a November 5, 2013 letter to the DPA. DeCODE confirmed that it had deleted all data registers containing imputed genotypes for individuals from whom consent was lacking. However, deCODE also presented the DPA with a proposal, according to which genotype data from research participants (who had consented) would be linked with genealogy data in a way that would generate statistical results as strong as those formerly achieved. According to the Iceland DPA, this would entail that a genetic imputation for those who had not consented would be generated "in a split [] second in the processing memory of a computer. However, this imputation would then cease to exist and would never be accessible to anyone in any form. The only accessible data would be the aforementioned statistical results, which would not in any way be traceable to individuals [10]." The DPA confirmed in a letter dated 26 November 2013 that this proposal did not give rise to objections if "all the aforementioned prerequisites were met [10]."

Most recently, deCODE published a series of papers in the journal *Nature Genetics* in March 2015 that described sequencing the genomes of 2,636 Icelanders, the largest collection ever analyzed in a single human population. Using the imputation technique, deCODE claims that it was able to combine the full genomes it has for about 10,000 Icelanders and the partial genetic information on 150,000 more to generate a report for genetic disease on every person in Iceland. For example, the firm can identify every Icelander with the well-known BRCA2 mutation, which raises the risk of breast and ovarian cancer, even if the individuals have not submitted to genetic testing themselves.

Dr. Steffánsson of deCODE contends that his company's research methods do not violate patient privacy because the company is not actually sequencing the citizens' DNA, but rather devising "conjectures" or "hypotheses" about them, rather than obtaining personal information. He notes that estimated DNA sequences, unlike directly measured sequences, are not very accurate for individuals, though they are valuable at the group level. Moreover, Steffánsson emphasizes that, until now, both the DPA and Iceland's national bioethics committee have approved the use of estimated genotypes for the two-thirds of Icelanders who have not consented to its research [6].

Geneticists disagree as to whether deCODE must obtain informed consent. Jón Jóhannes Jónsson, a geneticist with the University of Iceland, observes that deCODE is not truly doing anything new, given that geneticists routinely infer whether relatives who are not part of a particular study carry a genetic mutation. What is different about deCODE's strategy is that it invokes the DNA sequences of the entire Icelandic population. Jónsson concedes that deCODE's plan to use estimated data supplemented by hospital records presents a difficult case. Daniel MacArthur, a geneticist at Massachusetts General Hospital in the United States, suggests that although deCODE did not actually violate the privacy of individuals, from an ethics points of view the researchers should at least attempt to obtain informed consent. MacArthur laments that blocking deCODE from using its estimated data present a "tragedy" not only for the company, but the wider "complex disease genetics community [6]."

On the other hand, DeCODE's promise to delete individuals' data once it has calculated statistical results remains problematic, given the increasing proliferation of easy, cheap, and powerful reidentification technologies. [8] Erlich and Narayanan, experts in computational biology and computer information systems, have deemed deCODE's actions a "breach" of "genetic privacy" of the sort increasingly common in the last few years as the range of techniques to carry out such privacy breaching "attacks" has expanded. In particular, they term deCODE's method a "completion technique," meaning the use of known DNA data "to enable prediction of genomic information when there is no access to the DNA of the target." There have been several high profile breaches of privacy whereby an "attacker" has been able to infer, from the known genome of one individual, the genomes of his or her relatives [3].

Erlich and Narayanan note that deCODE's approach is an advanced version of the completion technique, given that deCODE has access to the genealogical and genetic information of several relatives of the target, and permits genotypes of distant relatives to be inferred. They explain that it is possible to develop an algorithm that finds relatives of a "target" who donated their DNA to the reference panel and who share a "unique genealogical path that includes the target, for example, a pair of half-first cousins when the target is their grandfather [3]." A shared DNA segment between the relatives indicates that the target has the same segment. By studying more pairs of relatives that are connected through the target, it is possible to collect more genomic information on the target without any access to his or her DNA, and, more importantly, without his or her informed consent [3]. This conflicts directly with a legal trend toward enhanced recognition of the privacy rights and autonomy of research participants, as reflected in proposed changes to the law and policy of informed consent in the United States.

## III. Proposed Changes to to the Law and Policy of Informed Consent in the U.S.

The September 8, 2015 Notice of Proposed Rulemaking (NPRM) published by the U.S. Department of Health and Human Services in the Federal Register, entitled *Human Subjects Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay, and Ambiguity for Researchers*, reflects the emerging recognition of the dangers of re-identification of research participants [4].

In the summary of its major provisions, the NPRM provides that "informed consent would generally be required for secondary research with a biospecimen (for example, part of a blood sample that is left over after being drawn for clinical purposes), even if the investigator is not being given information that would enable him or her to identify whose biospecimen it is [4]." The NPRM describes the changes in technology driving this proposed change, noting that "[n]ew methods, more powerful computers, and easy access to large administrative datasets produced by local, state, and federal governments have meant that some types of data that formerly were treated as non-identified can now be re-identified through combining large amounts of information from multiple sources," including publicly available sources. In light of this change, "the possibility of fully identifying biospecimens and some types of data from which direct identifiers had been stripped or [which] did not originally include direct identifiers has grown, requiring vigilance to ensure that such research be subject to appropriate oversight [4]." "Most importantly", according to the NPRM, "[a] growing body of survey data shows that many prospective participants want to be asked for their consent before their biospecimens are used in research [4]." Thus, the NPRM clearly prioritizes an individual's right to elect or decline participation in research. This notion aligns with recognition of the right of informed consent for individuals who participate via in silico biology, though the use of their estimated data.

Moreover, the U.S. National Institutes of Health (NIH) have recently revised their Genomic Data Sharing Policy (GDS) to set forth the expectation that investigators will obtain participants' consent not only for the use of their biospecimens and identifiable private information, but also for the use of their genomic data. This will be true even if the cell lines or clinical specimens used to generate the data are de-identified [7]. By requiring informed consent for genomic data, the GDS goes even further than the NPRM is recognizing the risks of re-identification and an individual's right to informed consent for research participation.

There are many reasons that individuals may object to the use of their de-identified information, even if it is estimated data. First, individuals may decline on ethical, religious or other personal grounds to participate in certain controversial forms of research, such as somatic nuclear cell transfer, stem cell research, and germ-line gene therapy. As noted in the Human Subjects Research NPRM, "a more participatory research model is emerging in social, behavioral, and biomedical research, one in which potential research subjects and communities express their views about the value and acceptability of research studies [4]." Second, research participants may object to commercial exploitation of discoveries developed through the use of their de-identified information. Largely in response to some highly publicized lawsuits in which research participants have sued researchers for revenue earned from using their information and biospecimens, it has become common for researchers to present research participants with informed consent documents that disclaim any economic interest in possible commercial applications flowing from the research. Research using de-identified records is highly problematic in that there is no informed consent and therefore no disclaimer.

Just as there are many valid arguments in favor of expanding informed consent protections for research participants, there are numerous reasons why the research community is likely to oppose the extension of research protections, whether for de-identified biospecimens or information, or estimated data. First, it is not feasible to contact each individual from whom materials have been gathered in order to request that person's informed consent. Even if it were possible, it would be very time-consuming and costly. Each individual's contribution to the research is so small, perhaps as to be dispensable, yet would require the full process of informed consent. Most importantly, and flowing from these reasons, the necessity of such informed consent might delay and perhaps even preclude altogether the development and introduction of medical advances. Furthermore, it is not only researchers, but also patient advocacy groups, who warn of these dangers. As noted by these critics, in the context of requiring informed consent for the use of de-identified biospecimens and identifiable private information, requiring such consent "might inappropriately give greater weight to the [] principle of autonomy over the principle of justice, because requiring consent could result in lower participation rates in research by minority groups and marginalized members of society," though "most of the comments from individual members of the public strongly supported consent requirements for use of their biospecimens, regardless of identifiability [4]."

Indeed, it can undermine trust in the medical establishment when individuals learn that their biospecimens or information, whether de-identified or estimated, are used without their consent. Indeed, the Human Subjects Research NPRM states that "the failure to acknowledge and give appropriate weight to this distinct autonomy interest in research using biospecimens could, in the end, diminish public support for such research, and ultimately jeopardize our ability to be able to conduct the appropriate amount of future research with biospecimens [4]."

It is clear that the trend, as evidenced by the Human Subjects Research NPRM and the revised NIH GDS, is toward the requirement of informed consent for the use of de-identified biospecimens and genetic information. The question then arises whether there is a meaningful distinction between de-identified biospecimens and information, on the one hand, and estimated data, on the other, in terms of the need for informed consent. It should be noted that neither de-identified information nor estimated data requires any direct interaction with the individual about whom it is gathered. Indeed, the Common Rule specifies that human subject research occurs when an investigator conducting research obtains "data through intervention or interaction with the individual", or obtains "identifiable private information" from any source [2]. The regulation further provides that "Private information must be individually identifiable (i.e., the identity of the subject is or may readily be ascertained by the investigator or associated with the information) in order for obtaining the information to constitute research involving human subjects [2]." It is this condition of individual identifiability that deCODE Genetics seeks to avoid when it declares to the Icelandic Data Protection Authority that the data will be individually identifiable only for a split second and then deleted from the computer memory. This argument fails, however, if data are as easily identifiable as Yaniv and Erlich have described.

The main difference between de-identified biospecimens and identifiable private information, on the one hand, and estimated data, on the other, is that the latter are not accurate at the individual level, but only at the group level**.** While this fact may adequately address the privacy issue, it does not resolve the issue of autonomy, meaning individuals' ability to decline to participate in research, either totally or as a means of rejecting the specific research proposed.

## IV. Conclusion and Future Work

Biospecimens are increasingly viewed as intrinsically identifiable. What is more, armed with bioinformatics and computational genomics techniques, along with public and private databases, researchers can accurately impute the genetic sequence information of individuals without their informed consent. While this can yield new discoveries and vital data for improving diagnostics, it also raises complex questions regarding the need to obtain informed consent from research participants about whom data is imputed via in silico research. The law of informed consent, codified before the development of powerful current technologies, does not address issues arising from the use of estimated data.

Proposed changes to U.S. regulations would provide enhanced protection for research subjects by requiring informed consent for the use of their biospecimens and identifiable private information, whether clinical or from prior research. Presently, researchers can use these specimens without consent by stripping them of identifiers. The newly revised NIH GDS goes even further by requiring informed consent for the use of genomic data, even if it derives from de-identified sources. These changes reflect the current view that researchers ought to respect the privacy and autonomy of research participants in an era where re-identification of research subjects has become easier to achieve. While a liberal reading of the proposed federal rule changes and the new NIH policy support the notion that those from whom estimated data is gathered and used are entitled to the same rights of informed consent, privacy, and autonomy as conventional research subjects, the proposed rule changes contemplate for the moment only research subjects who contribute biospecimens or identifiable private information, whether wittingly or not. This article contends that individuals who contribute estimated data are similarly entitled to be asked for their informed consent for their research participation.

The next steps in this research will be an investigation of the “right not to know” the results of one’s genetic risks. Paradoxically, while the law provides increasing protection for the right of informed consent, there is an emerging view that genetic incidental findings ought to be gathered and returned to individuals, even absent their informed consent. Indeed, deCode declares that it ought to be able to contact Icelanders to inform them of the genetic risks of which deCode learned when studying their estimated data. This raises the troubling specter of individuals who have given consent neither for the use of their estimated data, nor the return of incidental findings to them, having their data used for research and then being contacted with researchers’ incidental findings. This paternalistic approach conflicts deeply with the longstanding norms of biomedical ethics.

## Acknowledgment

The author thanks for their helpful comments on this work the hosts of and participants at the Law and Ethics of Big Data Research Colloquium held at University of Indiana at Bloomington on April 18, 2015. The author also expresses appreciation to the Zicklin School of Business, Baruch College, City University of New York, for research support for this work.

## References

[1] A. Abbott, “Icelandic database shelved as court judges privacy in peril,” Nature, vol. 429, p. 118, May 13, 2004, doi:10.1038/429118b.

[2] Code of Federal Regulations, 45 C.F.R. § 46.102(f) (2009).

[3] Y. Erlich and A. Narayanan, “Routes for breaching and protecting genetic privacy,” Nature Reviews Genetics, vol. 15, pp. 409-421, May 8, 2014, doi:10.1038/nrg323.

[4] Federal Register, Federal Policy for the Protections of Human Subjects: Proposed Rules, Vol. 80, No. 172, 53,933 – 54,060 (September 8, 2015).

[5] R. Goldin, “Privacy and our genes: is deCODE’s DNA project ‘Big Brother’ or the gateway to a healthier future,” Genetic Literacy Project, June 24, 2013, available at https://www.geneticliteracyproject.org/2013/06/24/privacy-and-our-genes-is-decodes-dna-project-big-brother-or-the-gateway-to-a-healthier-future/, retrieved: January, 2016.

[6] J. Kaiser, “Agency nixes deCODE’s new data-mining plan,” Science, vol. 340, pp. 1388-1389, June 21, 2013, doi: 10.1126/science.340.6139.1388.

[7] National Institutes of Health, NIH Genomic Data Sharing Policy, Notice Number NOT-OD-14-124 (Aug. 27, 2014), http://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html, retrieved: January, 2016.

[8] P. Ohm, “Broken promises of privacy: responding to the surprising failure of anonymization,” UCLA Law Review, vol. 57, pp. 1701-1777, 2010.

[9] T. Sveinsson, Iceland Data Protection Authority, E-mail message to Donna M. Gitter, Professor of Law, Baruch College (Oct. 20, 2014), unpublished.

[10] K. Yandell, “All Icelandic women with the BRCA2 gene can be found in the database,” News of Iceland, May 13, 2013.

# Social Sentiment Indices Powered by X-Scores

Brian Davis[*], Keith Cortis[†], Laurentiu Vasiliu[‡], Adamantios Koumpis[†], Ross McDermott[*] and Siegfried Handschuh[†]
[*]INSIGHT Centre for Data Analytics, NUI Galway, Ireland
Email: name.surname@insight-centre.org
[†]Universität Passau, Passau, Germany
Email: name.surname@uni-passau.de
[‡]Peracton, Dublin, Ireland
Email: laurentiu.vasiliu@peracton.com

***Abstract*—Social Sentiment Indices powered by X-Scores (SSIX) seeks to address the challenge of extracting relevant and valuable financial signals in a cross-lingual fashion from the vast variety of and increasingly influential social media services, such as Twitter, Google+, Facebook, StockTwits and LinkedIn, and in conjunction with the most reliable and authoritative newswires, online newspapers, financial news networks, trade publications and blogs. A statistical framework of qualitative and quantitative parameters called X-Scores will power SSIX. This framework will interpret financially significant sentiment signals that are disseminated in the social ecosystem. Using X-Scores, SSIX will create commercially viable and exploitable social sentiment indices, regardless of language, locale and data format. SSIX and X-Scores will support research and investment decision making for European SMEs, enabling end users to analyse and leverage real-time social media sentiment data in their domain, creating innovative products and services to support revenue growth with focus on increased alpha generation for investment portfolios.**

***Keywords–social sentiment index; cross-lingual; social media analytics; sentiment analysis; big social and news data.***

## I. INTRODUCTION

The emerging use of social media data as part of the investment process has seen a rapid increase in uptake in recent years, as by examined Greenfield [1]. The lag between Social Media Monitoring and Social Media Analytics - "Brand Analytics" and Finance specific analytics applications has narrowed. The Social Finance Analytics sector has built on the base developed by Brand Analytics and has evolved the ecosystem to focus on investment decision-making. The growth of trading specific social networks like StockTwits has also provided highly valuable structured social data on trading discussions, which was not accessible previously on general social media communities. This new data source has provided a vital pipeline of thoughts, words and decisions between people; connecting and interacting as never before. This collective pulse of conversations and emotional attitudes acts as a gauge of opinions and ideas on every aspect of society. Finance specific social media applications provide asset managers, equity analysts and high frequency traders with the ability to research and evaluate subtle real-time signals, such as sentiment volatility changes, discovery of breaking news and macroeconomic trend analysis. These data streams can be incorporated into current operating models as additional attributes for executing investment decision-making, with a goal to increase alpha and manage risk for a portfolio.

The European research project Social Sentiment Indices powered by X-Scores (SSIX - http://ssix-project.eu/), seeks to assist in this challenge of incorporating relevant and valuable social media sentiment data into investment decision making by enabling X-Scores metrics and SSIX indices to act as valid indicators that will help produce increased growth for European Small and Medium-sized Enterprises (SMEs). X-Scores provide actionable analytics in the shape of unique metrics calculated out of the Natural Language Processing (NLP) output. SSIX will extract meaningful financial signals in a cross-lingual fashion from a multitude of social network sources, such as Twitter, Google+, Facebook, StockTwits and LinkedIn, and also authoritative news sources, such as Newswires, Bloomberg, Financial Times and CNBC news channel; transforming these signals into clearly quantifiable sentiment metrics and indices regardless of language or locale. Financial services' SMEs can customise SSIX indices enabling them to provide meaningful domain specific insight to design more efficient systems, test trading and investment strategies, better understand risk and volatility behaviour of social sentiment and identifying new investment opportunities.

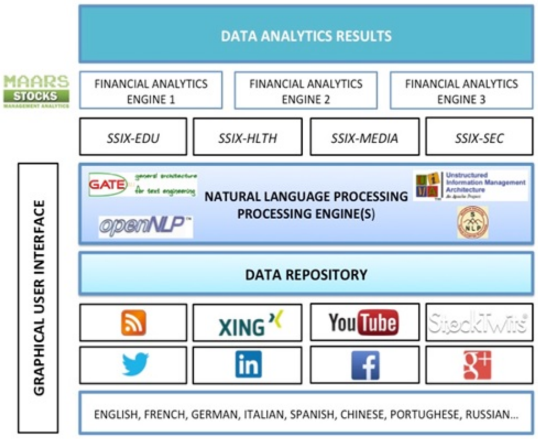

DATA ANALYTICS RESULTS
MAARS
STOCKS
MANAGEMENT ANALYTICS
FINANCIAL ANALYTICS ENGINE 1
FINANCIAL ANALYTICS ENGINE 2
FINANCIAL ANALYTICS ENGINE 3
SSIX-EDU
SSIX-HLTH
SSIX-MEDIA
SSIX-SEC
GRAPHICAL USER INTERFACE
GATE
general architecture for text engineering
openNLP
NATURAL LANGUAGE PROCESSING
PROCESSING ENGINE(S)
Unstructured Information Management Architecture
An Apache Project
NLP
DATA REPOSITORY
XING
YouTube
StockTwits
ENGLISH, FRENCH, GERMAN, ITALIAN, SPANISH, CHINESE, PORTUGHESE, RUSSIAN...

Figure 1. SSIX platform architecture

SMEs can exploit the open source SSIX tools and methodologies to provide financial analytics services or alternatively resell custom SSIX Indices as valuable financial data products to third parties, thus leading to growth and increased revenue for SSIX industry partners within the consortium and beyond. Beyond the financial application, the SSIX approach and methodologies can have broader impact across geopolitical and socio-economic domains, generating multifaceted and multi-domain sentiment index data for commercial exploitation. Figure 1 presents the overall SSIX platform architecture design.

The objectives of the SSIX project are to:

1) **Develop the "X-Scores" statistical framework**, which will analyse metadata from indexed textual sources to capture the signature of social sentiment, generating a sentiment score. Statistical methods will include regression, covariance and correlation analysis. These X-scores will be used to create the custom SSIX Indices.
2) **Create an open-source template for generating custom SSIX indices** that can be tailor-made with domain specific data parameters for specific analysis objectives, such as Economics, Trading, Investing, Government, Environmental or Risk profiling.
3) **Create a powerful, easy to implement and low latency "X-Scores API"** to distribute the raw sentiment data feed and/or custom SSIX Indices that will allow end users to easily integrate SSIXs sentiment data into their own systems.
4) **Enable end users to do cross-lingual target and aspect oriented sentiment analysis** over any significant social network using user defined dedicated SSIX Index.
5) **Enable various public/private organisations and institutions to create a SSIX Index** and integrate them with their proprietary tools in an easy to use manner.
6) **Explore the domain of SSIX Indices and X-Scores beyond its primary focus of Finance applications.** Research has shown there is a positive correlation between social media sentiment and a financial securities performance, but it is more difficult to measure a broad topic such as, welfare of a region or community. X-Scores will seek to provide metrics which can filter out the noise and provide real quantifiable data, which can give insight via a custom SSIX Index into domains diverse as Education (SSIX-EDU), Media trust (SSIX-MEDIA), Economic sociology (SSIX-ECOSOC), Security (SSIX-SEC) and Health (SSIX-HLTH).
7) **Empower and equip SMEs within the emerging Big Data Financial News sector** to better compete with established industry players via technology transfer involving stable, mature and scalable open source semantic and content analysis technologies.
8) **Trigger, nurture and maintain a SSIX and X-Scores commercial ecosystem within and beyond the project lifecycle.**
9) **Pierce language barriers with respect to untapped and siloed multilingual financial sentiment** content by harvesting cross lingual Big Social Media and News Data.

By number crunching news text and social networks data feeds regarding a company, product or various financial products (such as, stocks, funds, exchange-traded funds (ETFs), bonds etc.) in a mathematical and statistical way, our approach will allow investors and traders to combine SSIX generated indices with their own proprietary tools and methodologies. We envisage empowering the end-user, such as financial data providers, financial, institutions, investment banks, wealth management houses, asset management professionals, online brokers, professional traders and individual investors with the ability to make more informed and better and safer financial decisions. Finally, SSIX could help in identifying unwanted or dangerous trends that could be signalled to financial regulators in advance in order to take appropriate measures, potentially preventing unhealthy and toxic trading behaviour, thereby safeguarding economic growth and prosperity.

The remainder of the paper discusses related work in Section II. Information about SSIX Templates is provided in Section III, whereas Section IV discusses Big Social and News Data Management for SSIX. Details about Natural Language Processing Services and Analysis is presented in Section V. A Business Case Study about Investment and Trading is discussed in Section VI, before providing some concluding remarks in Section VII.

## II. Related Work

### A. *Sentiment Analysis on Financial Indices*

In [2], Bormann defines several psychological definitions about feelings, in order to explain what might be meant by "market sentiment" in literature on sentiment indices. This study is very relevant to SSIX, since it relates short and long term sentiment indices to two distinct parts of sentiments, namely emotion and mood; and extracts two factors representing investor emotion and mood across all markets in the dataset.

The FIRST project [3] provides sentiment extraction and analysis of market participants from social media networks in near real-time. This is very valuable towards detecting and predicting financial market events. This project is relevant to SSIX, since the tool consists of a decision support model based on Web sentiment as found within textual data extracted from Twitter or blogs, for the financial domain. The relationship between sentiment and trading volume can provide the end-user with important insights about financial market movements. It can also detect financial market abuse, e.g., price manipulation of financial instruments from disinformation. Unlike SSIX, only social networking services are used for extracting and analysing sentiment, whereas the developed tool cannot be easily customised to support media sources, target specific companies or select the required language. In this respect, SSIX provides a template methodology and source code to create in a consistent manner the sentiment index for any type of financial product and financial derivatives. Also the outcome is easily integrated within other analytics tools as a data stream with values between 0 and 100 that will define the ranges of that specific sentiment.

Mirowski et al. [4] presents an algorithm for topic modelling, text classification and retrieval from time-stamped documents. It is trained on each stage of its non-linear multi-layer model in order to produce increasingly more compact representations of bags-of-words at a document or paragraph level, hence performing a semantic analysis. This algorithm has been applied to predict the stock market volatility using financial news from Bloomberg. The volatility considered is estimated from daily stock prices of a particular company. On a similar level, in [5] the authors present StockWatcher through a customised, aggregated view of news categorised by different topics. StockWatcher performs sentiment analysis on a particular news messages. Each message can have either a positive, negative or neutral effect on the company. This tool enables the extraction of relevant news items from RSS feeds concerning the NASDAQ-100 listed companies. The sentiment of the news messages directly affects a company's respective stock price. SSIX, will extract meaningful financial signals from multilingual heterogeneous (micro-blogging and conventional) content sources and not just news items.

Gloor et al. introduces a novel set of social network analysis based algorithms for mining unstructured information from the Web to identify trends and the people launching them [6]. This work is relevant, since the result of a three-step process produces a “Web buzz index” for a specific concept that allows for an outlook on how the popularity of the concept might develop in the future. A possible application of this system might be for financial regulators who try to identify micro- and macro-trends in financial markets, e.g., showing the correlation between fluctuations in the Web buzz index for stock titles and stock prices. Similarly, the Financial Semantic Index estimates the probability that on a particular day, an article in the financial press expresses a positive attitude towards financial markets. This is measured through the emotional tone of the mentioned article [7]. It is relevant to SSIX, since it provides a certain viewpoint of the media environment the market participants consume. In the case of SSIX, it targets to transform the extracted information into multiple clearly quantifiable social financial sentiment indices regardless of language and data format. This will improve the trading and investment accuracy through the combination of various fundamental and technical parameters together with sentiment ones.

### B. *Cross-lingual mining of information*

The MONNET project provides a semantics-based solution for integrated information access amongst language barriers [8]. MONNET is relevant for SSIX, since one of its major innovations is the provision of cross-lingual ontology-based information extraction techniques for semantic-level extraction of information for text and (semi) structured data across languages by using multilingual localised ontologies. It provides real-life applications that demonstrate the exploitation potential in several areas, such as financial services. In fact, one of the project's use-cases deals with searching and querying for financial information in the user's language of choice. On the other hand, it focused on cross-lingual domain, thus failed to target other important aspects, e.g., mining the extracted information. SSIX will help identify unwanted/dangerous trends that could be signalled to financial regulators in advance, in order to potentially prevent unhealthy trading behaviour. Hence, SSIX indices can be used as ‘early warning’ signals for traders, investors and regulator agencies, such as European Central Bank, EU states national banks and rating agencies.

TrendMiner, another European project [9], presents an innovative and portable open-source real-time method for cross-lingual mining and summarisation of large-scale social media streams, such as weblogs, Twitter, Facebook, etc. One high profile case study was a financial decision support (with analysts, traders, regulators and economists). In terms of novelty, a weakly supervised machine learning algorithm is utilised for automatic discovery of new trends and correlations, whereas a cloud-based infrastructure is used for real-time text mining from stream media. This project is relevant to SSIX given that it provides several multilingual ontology-basedsentiment extraction methods.

The main goal of the LIDER project [10] is to create a Linguistic Linked Data (LLD) cloud that is able to support content analytics tasks of unstructured multilingual cross-media content. This will help in providing an ecosystem for a new Linked Open Data based ecosystem of free, interlinked and semantically interoperable language resources (e.g., corpora, dictionaries, etc.) and media resources (e.g., image, video, etc.). It also aims to make an impact on the ease and efficiency with which LLD is exploited in processes related to content analysis with several use cases in multiple industries within the areas of social media, financial services and other multimedia content providers and consumers. One limitation is that LIDER aims to make an impact on the LOD cloud and not to further transform any extracted signals into clearly quantifiable social sentiment indices, as in the case of SSIX. Such indices are targeted to any equities, stock indices or derivatives.

The AnnoMarket project has delivered a cloud-based platform for unstructured data analytics services, in multiple languages [11]. This text annotation market is delivered via annomarket.com and has been in public beta as of April 2014. The services being offered can be adopted and applied for many business applications, e.g., large-volume multi-lingual information management, business intelligence, social media monitoring, customer relations management. It includes several text analytics services that would be of benefit to the SSIX project. Similarly, OpeNER will provide a number of ready to use tools in order to perform some NLP tasks (entity mentions detection and disambiguation, sentiment analysis and opinion detection) that can be freely and easily integrated in the workflow of SMEs [12]. this project aims to have a semi-automatic generation of generic multilingual (initially for the English, French, German, Dutch, Italian and Spanish languages) sentiment lexicons with cultural normalisation and scales through the reuse of existing language resources. SSIX goes beyond text analysis on unstructured data, since an “X-Scores” statistical framework will be implemented to capture the signature of social sentiment from indexed textual sources. These scores will help create custom SSIX Indices that can be tailored for a particular domain depending on specific data parameters. This will provide a meaningful insight to drive trading, investment decisions and strategies, and create new investment opportunities.

## III. SSIX Templates

SSIX templates will empower both the public and private sectors to develop innovative disruption-enabling mobile and cloud services and products, to leverage the massive amount of sentiment data that is constantly produced and published on various social media networks within multiple domains such as Finance, Economy, Government, Politics and Health.

The SSIX templates will be able to gauge the actual voiced sentiment from social media conversations, specifically emotional attributes, such as (but not restricted to) optimism and pessimism. These sentiment signals can be analysed to evaluate their influence on real world financial/economic/social/political outcomes and can act as valid indicators. An ideal paradigm that can benefit from the integration of SSIX templates is the field of investment decisions. Traditionally, research on securities, such as stocks, fixed income and foreign exchange relied on applying a Fundamental and/or Technical Analysis approach to determine the most efficient and lowest risk investment decision for a given amount of expected return. In this scenario, market sentiment is derived from the aggregation of a variety of these two disciplines (Fundamental and Technical analysis), including attributes, such as price action, price history, economic and financial reports/data, market valuation indicators, fund flows, sentiment surveys (e.g., ZEW Indicator of Economic Sentiment - A Leading Indicator for the German Economy), commitment of traders report analysis, analysis of open interest from the futures market, seasonal factors and national/world events. As a consequence, it is difficult to get a reliable and easy to interpret measure of a securities sentiment score without using a selection bias and almost impossible to measure a niche sector efficiently; this type of sentiment classification tends to be a lagging indicator to price movement but can act as confirmation.

The growth of social media APIs and the application of news analytics has provided a new method allowing sentiment analysis from a social media perspective to be carried out on financial securities, which has been proven to show a positive correlation to price performance ("Twitter is now a leading indicator of movement (up and down) of specific stocks - we can prove it.", Social Market Analytics). This data can be analysed to gain a greater understanding of sentiment behaviour and its correlation to price volatility for an individual security/sector or the entire market. By using this new sentiment data source, SSIX can deliver unique sentiment indices using X-Scores (a statistical framework of qualitative and quantitative parameters, such as regression, covariance and correlation analysis), such as the 'Social Sentiment Index for Healthcare' - SSIX_Health or the 'Social Sentiment Index for Technology' - SSIX_Tech, which will show the sentiment levels for their corresponding sectors, quantifying how market participants feel. X-Scores metrics can used in conjunction with industry standard technical parameters to analyse securities, such as Moving Average Convergence-Divergence (MACD), Relative Strength Index (RSI), Moving Averages (MA), Exponential Moving Average (MVA), Pivots Points, etc. SSIX X-Scores will provide real quantifiable data and tools to anticipate volatility and to analyse past performance, which will help develop alternative and more efficient approaches to reduce risk. SSIX can be used to identify trading signals, helping to make more informed investment decisions, resulting in a more efficient use of capital while reducing any associated risk. SMEs will be able to integrate the SSIX framework data into their own models for use in any area of application where sentiment analysis is used.

## IV. Big Social and News Data Management

Data retrieved from digital social networking and news sources provides significant data samples to the NLP component of SSIX. The entire process is developed through the following steps:

- Data download and gathering from different digital platforms (social networks, blogs, news sites, etc.) with different techniques (API usage, CSV download, Web scraping, etc.);
- Data cleaning and filtering to isolate significant information;
- Data processing to produce analysed and enriched data (smart data);
- Data sampling to extract pieces of smart data intended to be used by NLP component.

### A. *Big Data Challenges*

In SSIX, multiple kinds of data are constantly collected, which process is continuous for the duration of the project. The following are types of data in question:

- Public available data from social networks
- Datasets part of the Linking Open Data (LOD) cloud
- LLD Cloud resources
- Public data available from domain-specific SMEs
- Survey data collected from independent events, such as technology summits, conferences, etc., or organised events, such as workshops, focus groups, etc.
- Financial and Economic trends outlined by the SSIX framework from analysis/mining of data
- Language Resources (LRs) either automatically acquired or reused from SentiWordNet (LR for opinion mining) and EuroSentiment (EU Project that provides a marketplace for LRs and Services dedicated to Sentiment Analysis).

Several challenges also arise due to the diverse nature of the gathered data. SSIX is able to deal with the three main challenges coming from the big data field namely, high volume, high velocity and high variety.

- High volume: constant growing of the data repository is managed through adoption of scalable technologies and architectures. The space required for the storage can be easily increased on request, while the technologies used are suitable to manage big quantities of data (e.g., Cassandra, Hadoop).
- High velocity: big stream of data is collected and managed with specific technologies and adequate processing capabilities. The project adopts high-performing servers with possibility to scale the computing power.
- High variety: the gathered data comes from multiple sources. In this case, each data source is treated separately. When required, an unstructured data model is implemented, in order to store information that can vary over time.

### B. *From Big Data to Smart Data*

Figure 2 illustrates the flow that all the data will follow before entering the SSIX platform for further NLP and analysis, which process transforms the data retrieved into smart data.

Each process is explained in more detail as follows:

- BIG DATA: indicates all the information available on different external platform in form of data sources (e.g., social networks, blogs, news sites, etc.)
- DOWNLOADER: the data are gathered from the different data sources using techniques, such as API usage, CSV download and parsing, web pages scraping, etc.

- DATA FILTERING → FILTERED DATA: a first process of noise removal and data processing that will produce a layer of filtered data.
- DATA PROCESSING → SMART DATA: in this phase of the process, all the data will be parsed and transformed into smart data.
- DATA SAMPLING → SAMPLES FOR NATURAL LANGUAGE PROCESSING: the last step will consist in the extraction of significant data samples destined for NLP.

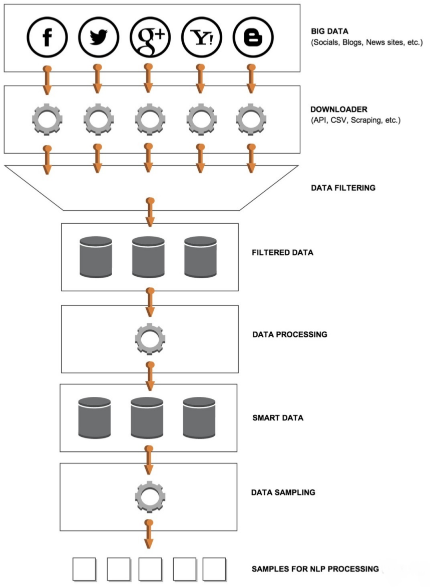

BIG DATA
(Socials, Blogs, News sites, etc.)
DOWNLOADER
(API, CSV, Scraping, etc.)
DATA FILTERING
FILTERED DATA
DATA PROCESSING
SMART DATA
DATA SAMPLING
SAMPLES FOR NLP PROCESSING

Figure 2. SSIX platform data-flow

All the smart data will be archived into a high performing repository. A cluster of servers will produce significant samples retrieved from the smart data repository that will be taken and streamed to the SSIX platform by and End Point component. The first prototype will use three physical servers to implement the architecture presented in Figure 3.

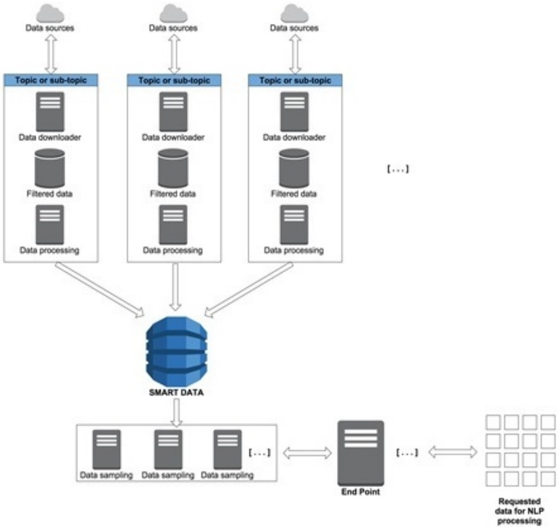


Figure 3. SSIX platform data architecture

The schema defined in Figure 3 illustrates the ideal architecture delegated to retrieve data from the identified data sources, in order to process it and to create data samples for the NLP phase. The business case studies that will be executed in the duration of the project (such as the one discussed in Section VI) will be managed by a cluster of machines that will include: i) a software component that will interface with the different data sources, which will retrieve the data from them; ii) a repository of filtered data; and iii) a software component for data processing.

## V. Natural Language Processing Services and Analysis

Analysing trends in social media content results in the process of a very large number of comparably short texts in near real-time. Therefore, the major challenge for the implementation of the NLP pipeline is in the orchestration of the different analysis components in a way that is potentially scalable in a cluster of servers that is able to handle hundreds of messages per second. Special care has to be taken to provide the NLP process as a distributed near real-time computation system that can reliably process unbounded streams of data. SSIX implements this process based on Apache Storm. Apache Storm is a framework that offers the foundations of distributed stream processing and is also fault-tolerant. Moreover, SSIX addresses the following major objectives:

- Automatic execution planning of NLP analysis processes: based on the descriptions of existing analysis components, available input and infrastructure, and desired output, SSIX automatically computes an appropriate execution plan (“topology” in Apache Storm);
- Standardised API for analysis components: a common problem in NLP processing is that there are many components for different, but related tasks, but they all implement completely different APIs, making it hard to combine them efficiently in a process. SSIX provides a standardised API and a standardised component description format to simplify integration of existing and additional analysis components.
- Sufficient collection of initial components: a big challenge in building this pipeline is to provide a sufficient collection of initial components so that we can (1) validate our execution model and API, and also provide examples for developers, (2) provide a process for real-time analytics, and (3) integrate with queuing and database technologies provided by SSIX. Figure 4 provides an overview architecture of the NLP pipeline.

### A. *Multilingual Language Resource Acquisition and Management*

The multilingual language resource acquisition and management occurs in two phases:

1) Identification and resource of existing language resources for adaption for SSIX business cases (one business case example is discussed in more detail in Section VI), i.e. exploitation of multilingual sentiment and domain specific lexica from European projects, such as EuroSentiment [13] –which provides a shared language resource pool for fostering sentiment analysis from a multilingual, quality and domain coverage perspective– or the adaptation of

LLD resources and carry out any necessary localisation of monolingual resources where target language equivalents are scarce, such as Asian languages.

2) Exploration of unsupervised and/or semi-automatic corpus based methods for acquisition of multilingual lexica to support entity and sentiment analysis tasks.

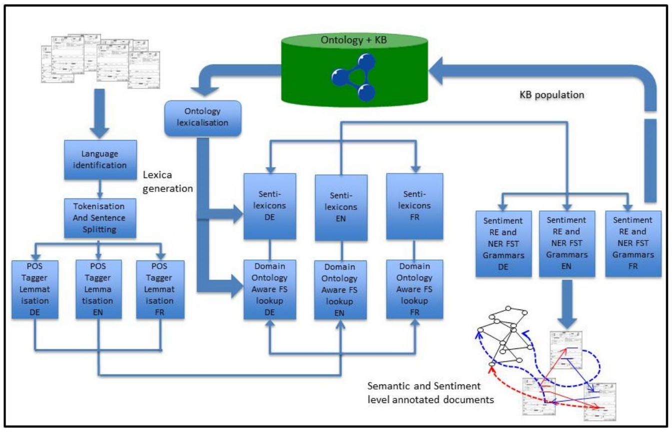

Ontology + KB
KB population
Ontology lexicalisation
Language identification
Lexica generation
Tokenisation And Sentence Splitting
POS Tagger Lemmat isation DE
POS Tagger Lemma tisation EN
POS Tagger Lemmat isation FR
Senti-lexicons DE
Senti-lexicons EN
Senti-lexicons FR
Domain Ontology Aware FS lookup DE
Domain Ontology Aware FS lookup EN
Domain Ontology Aware FS lookup FR
Sentiment RE and NER FST Grammars DE
Sentiment RE and NER FST Grammars EN
Sentiment RE and NER FST Grammars FR
Semantic and Sentiment level annotated documents

Figure 4. SSIX platform Knowledge-based NLP pipeline

## VI. Business Case Study: Investment and Trading

The SSIX sentiment index template will be used to determine social sentiments on stocks and then incorporate them as independent parameters within Peracton's MAARS platform (www.peracton.com/maars), for complex evaluation together with other financial parameters. Performance analysis will be made over historic data and real time, to determine the impact of SSIX indices. This case study will be made of the following 5 phases:

**Phase 1: Establish data sources and targets in order to generate unique SSIX Indices**

Establish data sources and targets in order to generate unique SSIX Indices In this phase we will identify the suitable data sources available on various social networks (Twitter, Facebook, LinkedIn, Google+). It is estimated that approximately 6,000+ stocks will be traced individually on US exchanges, such as NASDAQ, AMEX and NYSE.

**Phase 2: SSIX indices generation and storage**

Once the data sources are established, SSIX engine will be instantiated to generate 6,000+ unique sentiment indices that trace 6,000+ US stocks. Such indices will be uniquely identified, such as SSIX_AAPL, SSIX_YHOO, SSIX_LNKD, SSIX_FB, etc. Once instantiated, the SSIX indices values will be generated first for every day and stored accordingly and then for every minute (if this will be technically feasible).

**Phase 3: SSIX indices integration within MAARS**

The 6,000+ generated index sentiment values, stored every day (and every minute) will be integrated within MAARS analytics and attached to the existing financial data stocks that are already stored within MAARS cloud.

**Phase 4: Trading and Investment with SSIX indices**

As sentiment data starts to be updated within MAARS analytics, simulations tests of investing and trading will be performed. There will be trading and investment tests with no SSIX sentiment data (control tests) and then in parallel, same investment and trading tests involving sentiment data.

**Phase 5: Feedback** Based upon Phase 4 tests, feedback will be provided to the performance and changes in results of investment / trading exercise due to using sentiment data.

## VII. Conclusion

SSIX seeks to extract and measure meaningful financial sentiment signals in a cross-lingual fashion, from a vast multitude of social network sources, such as Twitter, Facebook, StockTwits, LinkedIn and public media outlets, such as Bloomberg, Financial Times and CNBC. It will generate custom X-Scores powered index for a given sentiment target or aspect, i.e. company or financial product. The primary domain is finance although SSIX has scope for Environment, Health, Technology, Geopolitics and beyond. The X-scores will be used by the industrial partners and bundled with their financial analytics, in order to increase the accuracy of their output combined with either end of day financial data or, real time data feeds. SSIX will adapt existing mature, proven and scalable open source text mining tools in order and circumvent language barriers with respect to unexploited multilingual financial sentiment content by harvesting cross lingual Big Social Media and News Data. Semantic Analytics will be employed to generate SSIX indices.

## Acknowledgment

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme ICT 2014 - Information and Communications Technologies under grant agreement No. 645425.

## References

[1] D. Greenfield, "Social media in financial markets: The coming of age..." GNIP, GNIP Whitepaper, 2014. [Online]. Available: http://stocktwits.com/research/social-media-and-markets-the-coming-of-age.pdf

[2] S.-K. Bormann, "Sentiment indices on financial markets: What do they measure?" Kiel Institute for the World Economy, Economics Discussion Paper 2013-58, 2013. [Online]. Available: http://www.economics-ejournal.org/economics/discussionpapers/2013-58

[3] "FIRST - large scale inFormation extraction and Integration infrastructure for SupporTing financial decision-making," *http://project-first.eu/*, 2013.

[4] P. Mirowski, M. Ranzato, and Y. LeCun, "Dynamic auto-encoders for semantic indexing," in NIPS 2010 Workshop on Deep Learning, Proceedings.

[5] A. Micu, L. Mast, V. Milea, F. Frasincar, and U. Kaymak, "Financial news analysis using a semantic web approach," in Semantic Knowledge Management: an Ontology-based Framework, Paolo Ceravolo, Ernesto Damiani, Gianluca Elia, Antonio Zilli (Eds.), November 2008, pp. 311–328.

[6] P. A. Gloor, J. Krauss, S. Nann, K. Fischbach, and D. Schoder, "Web Science 2.0: Identifying Trends through Semantic Social Network Analysis," in International Conference on Computational Science and Engineering. CSE '09., pp. 215–222.

[7] Ontology2, "FSI: Financial Semantic Index," *http://financialsentimentindex.com/fsi/*, 2012.

[8] "MONNET - Multilingual Ontologies for Networked Knowledge," *http://cordis.europa.eu/fp7/ict/language-technologies/projectmonnet_en.html*, 2013.

[9] "TrendMiner," *http://www.trendminer-project.eu/*, 2014.

[10] "LIDER: "Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe"," *http://www.liderproject.eu/*, 2015.

[11] "AnnoMarket," *https://annomarket.eu/*, 2014.

[12] "OpeNER," *http://www.opener-project.org/*, 2014.

[13] "EuroSentiment," *http://eurosentiment.eu/*, 2014.

# Stream Mining Revisited

Sayaka Akioka
School of Interdisciplinary Mathematical Sciences
Meiji University
Tokyo, Japan 164–8525
Email: akioka@meiji.ac.jp

***Abstract*—Big data applications have become popular in recent years. Stream mining is one of the major data mining methodologies, which are frequently used in big data applications. Stream mining differenciates itself from the other big data applications for its severe requirement, and is also known for its changing behaviros according to the characteristics of input data. The problem is, however, the parameters, or methodologies for data characterization are not clearly defined yet. There is no study investigating explicit relationships between the characteristics of input data, and the behaviors of stream mining applications. Therefore, the current optimization methodology for stream mining is basically heuristic. This paper provides comprehensive survey on modeling stream mining to seek the strategy for this modeling problem.**

***Keywords*—*stream mining; modeling; characterization.***

## I. Introduction

Big data applications have become popular in recent years. These big data applications are supposed to collect gigantic amount of data from various data sources, analyze these data from several points of view, uncover new findings, and then provide totally new values. Compared to the conventional applications, big data applications need to handle extremely huge amounts of data, and this situation leads high, and increasing demand for the computational environment, which accelerates, and scales out big data applications. The serious problem here, however, is that the behaviors, or characteristics of big data applications are not clearly defined yet. There is no established model for big data applications.

Big data applications can be classified into several categories depending on the characteristics of data usage. Among these big data applications, this paper has special focus on stream mining applications. A stream mining application is such an application that analyzes data in a line. That is, the target data arrive one after another in chronological order. A stream mining application differenciates itself from the other big data applications for its severe requirement. A stream mining application needs to finish the analysis on the fly. In many cases, there is no chance to save the target data somewhere to revisit the data later. Algorithms specialized for stream mining applications (stream mining algorithms) are intensively studied [1]–[30], and Gaber et al. gave an excellent survey report on these algorithms [31].

High performance computing community has been investigating data intensive applications, which analyze huge amount of data as well. Raicu et al. pointed out that data intensive applications, and stream mining applications are fundamentally different from the viewpoint of data access patterns. Therefore, the strategies for speed-up of data intensive applications, and stream mining applications have to be radically different [32]. Many data intensive applications often reuse input data, and the primary strategy of the speed-up is locating the data close to the target CPUs. Stream mining applications, however, rarely reuse input data, and the strategy for data intensive applications does not work in many cases.

Modern computational environment has been evolving mainly for speed-up of benchmarks such as Linpack [33], or SPEC [34]. These benchmarks are relatively scalable according to the number of CPUs. Stream mining applications are not scalable to the number of CPUs in many cases. Current computational environment is not necessarily ideal for stream mining applications for this reason. Additionally, many researchers from machine learning domain, or data mining domain point out that the behavior, or execution time of a stream mining application varies according to the characteristics, or features of input data. The problem is, however, the parameters, or the methodologies for data characterization are not clearly defined yet. There is no study investigating explicit relationships between characteristics of input data, and behaviors of stream mining applications. Therefore, the current optimization methodology for stream mining is basically heuristic.

The major purpose of this paper is to provide a comprehensive survey on modeling stream mining applications. This paper focuses on generic models for stream mining applications, but does not cover the details of execution models of existing middlewares, or frameworks for stream mining applications. The primary purpose of this paper is to find keys to generalize stream mining applications, and clues to connect characteristics of input data, and behaviors of stream mining applications. This paper also tries to give some considerations on the strategy to address this modeling problem based on the survey. The rest of this paper is organized as follows. Section 2 introduces conventional proposals for stream mining algorithms. Section 3 discusses possible strategies, or directions for a stream mining application model. Section 4 concludes this paper.

## II. Models of Stream Mining Algorithms

### A. A Three-layer Model

Junghans et al. proposed a three-layer model, which is illustrated as the shaded part in Figure 1. They argued that most stream mining algorithms follow this three-layer model [35]. First, the filter component filters incomping data as necessary for the purpose of sampling, or load shedding. Secondly, the online mining component analyzes the original incoming data stream, or the filtered stream. Thirdly, the results of the online mining component will be stored in the synopsis, which is the

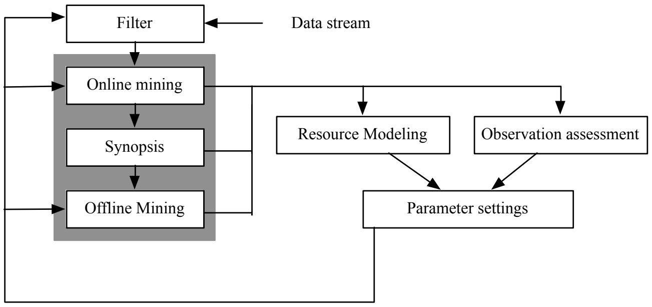


Fig. 1. Extended three-layer model.

second layer of the three-layer model. Here, synopsis indicates sketches, windows, or other dedicated data structures such as a pattern tree; those are often utilized in stream mining algorithms. Finally, the offline mining component answers user queries by accessing information stored in the synopsis. Here, the offline mining component does not need to fulfill the one pass requirement of stream mining.

As Junghans et al. developed their stream mining model with a background of embedded devices, they put an asumption that stream mining is conducted on limited available resources, such as limited number of CPUs, or limited amount of memory. Stream mining algorithms are often optimized for the better performance even under these constraints. Junghans et al., therefore, extended their three-layer model to include these optimization functionalities. Junghans et al. also extended the three-layer model for the better quality of the results by stream mining. The principle of this extension is to optimize the influential parameters in stream mining algorithms. This optimization contributes to the relaxed resource requirements, or the better quality of the mining results. Figure 1 illustrates this extended three-layer model. The shaded part of the figure is the original three-layer model as already described above, and the right part of the figure is the extension. The resource monitoring, and the observation assessment component collect information about the current system state. Based on the monitoring by the resource monitoring, and the observation assessment, the parameters are decided whether they should be adapted, or not. Then, the new parameters are set, and the stream mining algorithm run with the updated parameters.

### B. Stream Mining and Data Dependencies

Akioka et al. proposed another stream mining model [36], and the modeling put the focus on data dependencies. Figure 2 illustrates the overall model of stream mining algorithm. The model shown in Figure 2 is quite similar to the model by Junghans et al, while Figure 3 illustrates the detailed model of stream mining algorithms. Figure 3 depicts data dependencies, and control dependencies, and these dependencies lie among threads, or processes in one stream mining algorithm.

In Figure 2, a stream mining algorithm consists of two parts, stream processing part, and query processing part. The stream processing part consists of stream processing modules, sketches, and analysis modules. First, the stream processing module in the stream processing part picks the target data unit, and executes a quick analysis over the data unit. The quick analysis can be a preconditioning process such as a morphological analysis, or a word counting. Second, the stream processing module in stream processing part updates the data in one or more sketches. After this update, the data in the sketch(es) contains the latest results of the execution by the stream processing part. That is, the sketch(es) keeps the intermediate analysis, and the stream processing module updates the analysis incrementally as more data units are processed. Third, the analysis module in stream processing part reads the intermediate analysis from the sketch(es), and extracts the essence of the data in order to complete the quick analysis in the stream processing part. Finally, the query processing part receives this essence for further analysis, and the whole process for the target data unit is completed.

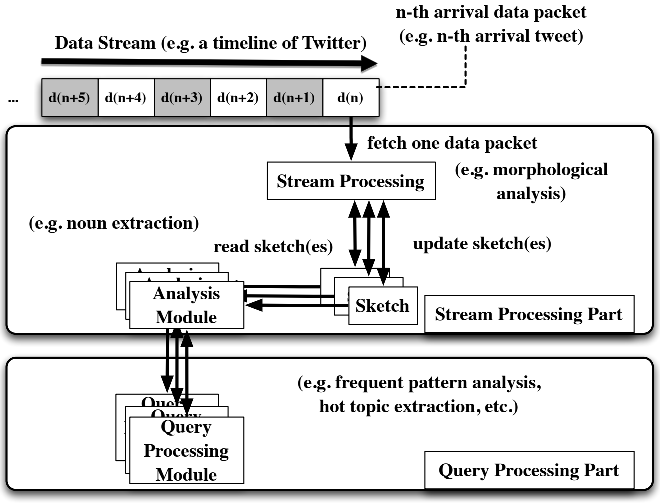


Fig. 2. A model of stream mining algorithms.

Based on the model shown in Figure 2, the major responsibility of the stream processing part is to preprocess each data unit for the further analysis, and that the stream processing part has the huge impact over the latency of the whole process. Therefore, the stream processing part also needs to finish the preconditioning of the current data unit before the next data unit arrives. Otherwise, the next data unit will be lost as there is no storage for buffering the incoming data. The query processing part takes care of the offline part of the analysis, and does not suffer from the strict requirement of stream mining.

Figure 3 focuses only on stream processing part as this is the part which impacts the overall performance. The figure illustrates data dependencies between two processes analyzing data units in line, and data dependencies inside each process. The assumption is that each process analyzes its own (different) data unit. The left top flow represents the stream processing part of the preceding process, and the right bottom flow represents the stream processing part of the successive process. Each flow consists of six stages; read from sketch(es), read from input, stream processing, update sketch(es), read from sketch(es), and analysis. An arrow represents a control flow, and a dashed arrow represents a data dependency. There are three data dependencies in total. These data dependencies are introduced by control flow for the correct executions, and the summary for these dependencies is as follows.

- The processing module in the preceding process should finish updating the sketch(es) before the processing module in the successive process starts reading the sketch(es) (Dep.1 in Figure 3).
- The processing module should finish updating the sketch(es) before the analysis module in the same process starts reading the sketch(es) (Dep.2 in Figure 3).
- The analysis module in the preceding process should finish reading the sketch(es) in the successive process starts updating the sketch(es) (Dep.3 in Figure 3).

### *C. Stream Mining with Multiple Data Streams*

Wu et al. pointed out that many of the existing researches on stream mining assume that there is one data stream, and they proposed formal definition of mining over multiple data streams [37]. In actual situations, the assumption with multiple data streams is more realistic than the single data stream. Therefore, the formal definition of the problems with multiple data streams is more practical, and reasonable.

Accordng to Wu et al., multiple data stream mining should be approached in a separate way from the way for the single data stream mining. First, multiple data streams are from many local data sources to generate distributed data streams independently. These data sources are not capable of processing more than simple data preconditioning, or saving all the generated data. Second, multiple data stream mining is supposed to process the mining across the data streams, not only on one single data stream. Third, these multiple data streams are not modeled as one single huge data stream with different attributes. Timestamp on each data is not under uniform criteria in many cases. Sampling rate of the data is different. The format of the generated data, or privacy concern is not controlled. There is no reason to handle these data streams as if one single data stream.

Wu et al. represent each data in a data flow as a quadruple of the form $(s, t, f, v)$, where $s$ is the identification of the place, $t$ is the time or sequence number identifying the event, $f$ is a function, and $v$ is a value vector of the output. Here, event refers whether data generation, or some other data processing. Each flow is a set of the quadruples, and fulfills the following properties.

- Each source specifies a single function to generate a single flow;
- For any pair of events, $e_1$ and $e_2$, that occur at the same source, if the two events have the same function invocation, and $e_1$ occurs before $e_2$, the value $t$ of $e_1$ is smaller than that of $e_2$;
- For any pair of events, $e_1$ and $e_2$, that occur at different sources, there is no function or rule between $e_1$ and $e_2$.

In addition to these properties, flows can have some additional properties:

- Homogeneous or heterogeneous: A pair of flows is said to be homogeneous (or heterogeneous) if the respective sources at which the two flows generate specify the same (or different) function(s), which are checked in terms of initial conditions and output domain;
- Relational: A pair of flows, indicated by $f_1$ and $f_2$, is said to be relational if the value vectors of $f_1$ and value vectors of $f_2$ satisfy some relationship $r$ (the relationship refers to values; events are independent).

Wu et al. also gave some considerations comparing multiple data streams, and other data stream models.

- Single stream with one dimension: This is the simplest model, and usually generated at a simple data stream application.
- Single stream with multiple dimensions: This applies to the applications in which there are multiple parameters, or attributes to be collected, and observed for each event occuring at a single source. The main difference between this model and multiple data streams is that single stream with multiple dimensions handles events of the same function invocation at a single source basically, while multiple data streams can invoke multiple functions distributed on different sources.
- Multiple data streams: This model is applicable to many real applications with multiple sources. These sources can be the same kind of devices, which are distributed at geometrically scattered locations. Basically, the multiple data sources can be viewed as a set of one or more dimensional single data streams.

## III. Discussions

### *A. Comparison of the Three Models*

Wu et al. defined multiple data streams, and mining multiple data mining (we refer to their model as MDS). The definition itself is beneficial in order to clarify the problem. Considering multiple data streams is more realistic as well. The point is, however, multiple data streams are still a set of single data streams as pointed out in their paper. Therefore, the priority for addressing this modeling problem should be the solid methodology for modeling the stream mining with one single data stream. How to superpose several models of stream mining models of a single data stream will be the next step.

Junghans et al. proposed their stream mining model in the context of embedded devices, such as sensors activated by batteries, and connected to the network by wireless (we refer to their model as Three-layer model). On the other hand, Akioka et al. proposed their stream mining model in the context of high performance computing (we refer to their model as DAP). Both of them proposed quite similar generic models for stream mining algorithms, and the discussions for the restrictions, and requirements for general stream mining are also in the same direction. These approximate models, however, do not deeply contribute for the strategy of scaling out stream mining algorithms. For the better choice of computational environment, size, allocations, preliminary estimations for resource requirements are indispensable. In this context, Akioka et al. proposed a model with data dependencies, and control dependencies. This model is quite similar to a task

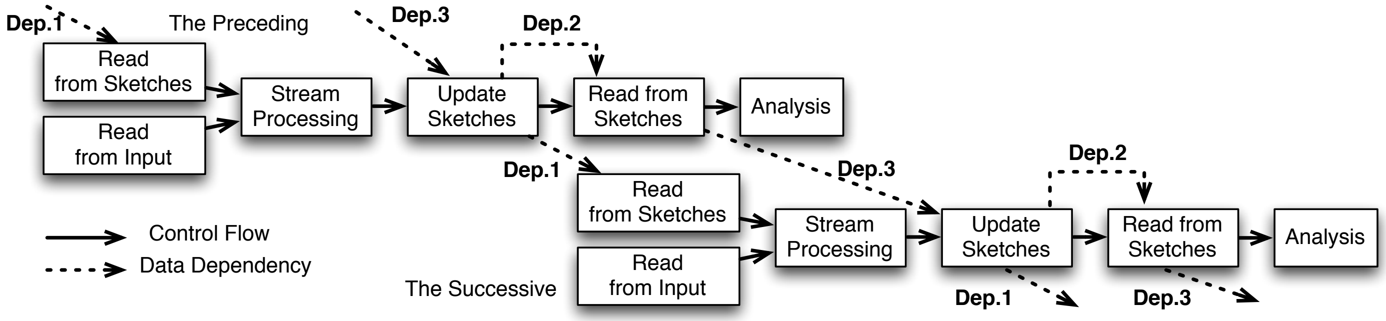


Fig. 3. Data dependencies of the stream processing part in two processes in line.

graph, which often used to solve scheduling problems. The problem is, however, there is no model for estimations of resources, or durations.

Junghans et al. put monitoring, and parameter update functionality into their model. They also picked up main memory, CPU cycles, bandwidth, and battery power as resources, and discussed stream properties (the stream rate, and characteristics of its individual elements such as value range, distribution, and size), input parameters, and query parameters influence the resource requirements of any stream mining algorithms. The solid models for these resources, or elements mentioned above are not proposed, however. The actual estimations are based on heuristics.

Table I summarizes the discussions in this section comparing the three models. As summarized in the table, there is no solid model proposal to fulfill the requirements for solving the load balancing problem, or scheduling problem of general stream mining applications. Here, we would like to remind you that a stream mining application changes its behavior according to the characteristics of input data, and that there is no model for input data. That is, everything is heuristic now for stream mining applications. Table I clearly shows that there is no successful project to give a solution on input data problem, and behavior characterizaion. These left problems are unavoidable for a direct solution for optimum execution of stream mining applications. We also need to be careful on the blanks regarding resource estimation. Currently, all of the modelings here heavily rely on a heuristic way to estimate the required resource, including the number of CPUs, or the duration of each stage of a stream mining application. There is no model here as well. Of course, as the application changes its behavior with input data, this problem is heavily connected to the input data problem. We still need to remember, however, that we have to prepare task graphs for all the stream mining applications without resource estimation models.

### B. Things to be Considered

We discussed how to understand, and model stream mining applications above. Here, another option for the optimum execution of stream mining applications would be a large-scaled cloud computing environment. If there is a good way to migrate whole, or some parts of running stream mining applications from one cloud environment to another environment, the restriction for the resource environment becomes loose. The challenges for this option will include the following items.

TABLE I. COMPARISON OF THE THREE MODELS.

| | MDS | Three-layer model | DAP |
|---|---|---|---|
| generic model (single stream) | no | yes | yes |
| data/control dependencies | - | no | yes |
| resource estimation | - | no | no |
| input data characterization | - | no | no |
| input/behavior characterization | - | no | no |
| generic model (multiple streams) | yes | no | no |
| data/control dependencies | no | - | - |
| resource estimation | no | - | - |
| input data characterization | no | - | - |
| input/behavior characterization | no | - | - |

- Practical cloud computing environment with task migrations: This option requires any part of the implementation of stream mining to be ready for migration on the fly. Although there are many researchers, or products that enable task migrations, however, the question is how much they are practically usable. The time when those migration techniques are intensively studied, applications with migrations were implemented by people with background of computer architecture, parallel computing, or optimization techniques of programs. On the contrary, the major implementers of big data applications are more various. They are not necessarily with the detailed background of computer science. The point is how much those implementers accept, and happily utilize the migration techniques.

- Consistency and preservation of the data and results: As repeated in this paper, stream mining is a continuous, one-way, one-pass application. There is no way to save input data, or intermediate output without stopping the current execution. Once you stop the execution, you will lose both the input data, and the expected results while you are suspending the problem. Even if you resume the program, you will not be able to acquire intrinsic results for a while, as many of stream mining algorithms rely on the results from the previous data input. How to preserve the whole flow of stream mining should be an inescapable problem.

- Migration management: Even if a good methodology for migration is established to solve the problem mentioned above, there is another problem. The problem is the strategy for migration. There should be an algorithm to decide which part of the whole program should be where, and when. This is basically load balancing problem, or scheduling problem. Therefore, we face the same modeling problem again. This time, we need to model both the behavior of stream mining applications, and current computational environment.

- Geographical placement: Except the case when the whole process (collection of input data, analysis with stream mining algorithms, and acquisition of the results) is performed in house, data source, and the actual computational environment are geographically scattered. Actually, this situation is quite common. Additionally, many of the current cloud computing services are employed in one place, or similar. This situation means that only a few points in the Internet accept almost all the input data collected all over the world. The inbound network load will be immeasurable, and have serious impact over performance of each of stream mining applications.
- Data privacy: Data privacy is one of the serious problems in recent concerns. Once you start migrating the whole, or some parts of stream mining applications, they will hop around with the data. One of the reasons why big data applications have became attractive rapidly is that many organizations have their own huge data without significance. The big data boom suggested that there is possible value in the huge sleeping data. The story will be different, however, once the data start moving around in the cloud, and it is difficult to protect the data from sniffing. People will become more conscious, and the boom will shrink.

## IV. Conclusions

This paper provided a survey on generic modeling of stream mining. The results of the survey suggested that there is no successful solid modeling to address the problems surrounding stream mining applications. Even though there are several research projects sharing the same problem, however, the level of modeling is more abstract than the level for practical use. Currently, no project is free from heuristic.

The last half of the discussion in this paper argued another possible approach for the better environment for stream mining applications, beside direct modeling of stream mining. Although the discussion part might show some other directions, however, we could not find a brilliant strategy to solve the problem. We keep seeking the solution with a broader view.

## Acknowledgment

This work was supported by JSPS KAKENHI, Grant-in-Aid for Young Scientists (B), and its Grant Number is 15K21423.

## References

[1] B. Babcock, M. Datar, R. Motwani, and L. O'Callaghan, "Maintaining variance and k-medians over data stream windows," in *Proceedings of the Twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, ser. PODS '03. New York, NY, USA: ACM, 2003, pp. 234–243. [Online]. Available: http://doi.acm.org/10.1145/773153.773176

[2] N. Tatbul, U. Çetintemel, S. Zdonik, M. Cherniack, and M. Stonebraker, "Load shedding in a data stream manager," in *Proceedings of the 29th International Conference on Very Large Data Bases - Volume 29*, ser. VLDB '03. VLDB Endowment, 2003, pp. 309–320. [Online]. Available: http://dl.acm.org/citation.cfm?id=1315451.1315479

[3] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom, "Models and issues in data stream systems," in *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, ser. PODS '02. New York, NY, USA: ACM, 2002, pp. 1–16. [Online]. Available: http://doi.acm.org/10.1145/543613.543615

[4] A. C. Gilbert, Y. Kotidis, S. Muthukrishnan, and M. Strauss, "Surfing wavelets on streams: One-pass summaries for approximate aggregate queries," in *Proceedings of the 27th International Conference on Very Large Data Bases*, ser. VLDB '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 79–88. [Online]. Available: http://dl.acm.org/citation.cfm?id=645927.672174

[5] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," in *Proceedings of the 29th International Conference on Very Large Data Bases - Volume 29*, ser. VLDB '03. VLDB Endowment, 2003, pp. 81–92. [Online]. Available: http://dl.acm.org/citation.cfm?id=1315451.1315460

[6] ——, "A framework for projected clustering of high dimensional data streams," in *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, ser. VLDB '04. VLDB Endowment, 2004, pp. 852–863. [Online]. Available: http://dl.acm.org/citation.cfm?id=1316689.1316763

[7] ——, "On demand classification of data streams," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '04. New York, NY, USA: ACM, 2004, pp. 503–508. [Online]. Available: http://doi.acm.org/10.1145/1014052.1014110

[8] G. Cormode and S. Muthukrishnan, "What's hot and what's not: Tracking most frequent items dynamically," *ACM Trans. Database Syst.*, vol. 30, no. 1, pp. 249–278, Mar. 2005. [Online]. Available: http://doi.acm.org/10.1145/1061318.1061325

[9] P. Yu, J. Pei, J. Han, H. Wang, G. Dong, and L. V. Lakshmanan, "Online mining of changes from data streams: Research problems and preliminary results," in *Proceedings of the 2003 ACM SIGMOD Workshop on Management and Prcessing of Data Streams*, 2003.

[10] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan, "Clustering data streams: Theory and practice," *IEEE Trans. on Knowl. and Data Eng.*, vol. 15, no. 3, pp. 515–528, Mar. 2003. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2003.1198387

[11] M. Charikar, L. O'Callaghan, and R. Panigrahy, "Better streaming algorithms for clustering problems," in *Proceedings of the Thirty-fifth Annual ACM Symposium on Theory of Computing*, ser. STOC '03. New York, NY, USA: ACM, 2003, pp. 30–39. [Online]. Available: http://doi.acm.org/10.1145/780542.780548

[12] P. Domingos and G. Hulten, "Mining high-speed data streams," in *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '00. New York, NY, USA: ACM, 2000, pp. 71–80. [Online]. Available: http://doi.acm.org/10.1145/347090.347107

[13] G. Hulten, L. Spencer, and P. Domingos, "Mining time-changing data streams," in *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '01. New York, NY, USA: ACM, 2001, pp. 97–106. [Online]. Available: http://doi.acm.org/10.1145/502512.502529

[14] L. O'Callaghan, N. Mishra, A. Meyerson, S. Guha, and R. Motwani, "Streaming-data algorithms for high-quality clustering," in *Data Engineering, 2002. Proceedings. 18th International Conference on*, 2002, pp. 685–694.

[15] C. Ordonez, "Clustering binary data streams with k-means," in *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, ser. DMKD '03. New York, NY, USA: ACM, 2003, pp. 12–19. [Online]. Available: http://doi.acm.org/10.1145/882082.882087

[16] E. Keogh and J. Lin, "Clustering of time-series subsequences is meaningless: Implications for previous and future research," *Knowl. Inf. Syst.*, vol. 8, no. 2, pp. 154–177, Aug. 2005. [Online]. Available: http://dx.doi.org/10.1007/s10115-004-0172-7

[17] H. Wang, W. Fan, P. S. Yu, and J. Han, “Mining concept-drifting data streams using ensemble classifiers,” in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’03. New York, NY, USA: ACM, 2003, pp. 226–235. [Online]. Available: http://doi.acm.org/10.1145/956750.956778

[18] V. Ganti, J. Gehrke, and R. Ramakrishnan, “Mining data streams under block evolution,” *SIGKDD Explor. Newsl.*, vol. 3, no. 2, pp. 1–10, Jan. 2002. [Online]. Available: http://doi.acm.org/10.1145/507515.507517

[19] S. Papadimitriou, A. Brockwell, and C. Faloutsos, “Adaptive, hands-off stream mining,” in *Proceedings of the 29th International Conference on Very Large Data Bases - Volume 29*, ser. VLDB ’03. VLDB Endowment, 2003, pp. 560–571. [Online]. Available: http://dl.acm.org/citation.cfm?id=1315451.1315500

[20] M. Last, “Online classification of nonstationary data streams,” *Intell. Data Anal.*, vol. 6, no. 2, pp. 129–147, Apr. 2002. [Online]. Available: http://dl.acm.org/citation.cfm?id=1293986.1293988

[21] Q. Ding, Q. Ding, and W. Perrizo, “Decision tree classification of spatial data streams using peano count trees,” in *Proceedings of the 2002 ACM Symposium on Applied Computing*, ser. SAC ’02. New York, NY, USA: ACM, 2002, pp. 413–417. [Online]. Available: http://doi.acm.org/10.1145/508791.508870

[22] G. S. Manku and R. Motwani, “Approximate frequency counts over data streams,” in *Proceedings of the 28th International Conference on Very Large Data Bases*, ser. VLDB ’02. VLDB Endowment, 2002, pp. 346–357. [Online]. Available: http://dl.acm.org/citation.cfm?id=1287369.1287400

[23] P. Indyk, N. Koudas, and S. Muthukrishnan, “Identifying representative trends in massive time series data sets using sketches,” in *Proceedings of the 26th International Conference on Very Large Data Bases*, ser. VLDB ’00. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, pp. 363–372. [Online]. Available: http://dl.acm.org/citation.cfm?id=645926.671699

[24] Y. Zhu and D. Shasha, “Statstream: Statistical monitoring of thousands of data streams in real time,” in *Proceedings of the 28th International Conference on Very Large Data Bases*, ser. VLDB ’02. VLDB Endowment, 2002, pp. 358–369. [Online]. Available: http://dl.acm.org/citation.cfm?id=1287369.1287401

[25] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, “A symbolic representation of time series, with implications for streaming algorithms,” in *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, ser. DMKD ’03. New York, NY, USA: ACM, 2003, pp. 2–11. [Online]. Available: http://doi.acm.org/10.1145/882082.882086

[26] D. Turaga, O. Verscheure, U. V. Chaudhari, and L. Amini, “Resource management for networked classifiers in distributed stream mining systems,” in *Data Mining, 2006. ICDM ’06. Sixth International Conference on*, Dec 2006, pp. 1102–1107.

[27] D. S. Turaga, B. Foo, O. Verscheure, and R. Yan, “Configuring topologies of distributed semantic concept classifiers for continuous multimedia stream processing,” in *Proceedings of the 16th ACM International Conference on Multimedia*, ser. MM ’08. New York, NY, USA: ACM, 2008, pp. 289–298. [Online]. Available: http://doi.acm.org/10.1145/1459359.1459398

[28] B. Thuraisingham, L. Khan, C. Clifton, J. Maurer, and M. Ceruti, “Dependable real-time data mining,” in *Object-Oriented Real-Time Distributed Computing, 2005. ISORC 2005. Eighth IEEE International Symposium on*, May 2005, pp. 158–165.

[29] N. K. Govindaraju, N. Raghuvanshi, and D. Manocha, “Fast and approximate stream mining of quantiles and frequencies using graphics processors,” in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. ACM, 2005, pp. 611–622.

[30] K. Chen and L. Liu, “He-tree: a framework for detecting changes in clustering structure for categorical data streams,” *The VLDB Journal*, vol. 18, no. 6, pp. 1241–1260, 2009. [Online]. Available: http://dx.doi.org/10.1007/s00778-009-0134-5

[31] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy, “Mining data streams: A review,” *SIGMOD Rec.*, vol. 34, no. 2, pp. 18–26, Jun. 2005. [Online]. Available: http://doi.acm.org/10.1145/1083784.1083789

[32] I. Raicu, I. T. Foster, Y. Zhao, P. Little, C. M. Moretti, A. Chaudhary, and D. Thain, “The quest for scalable support of data-intensive workloads in distributed systems,” in *Proc. the 18th ACM International Symposium on High Performance Distributed Computing (HPDC’09)*, 2009.

[33] J. Dongarra, J. Bunch, C. Moler, and G. Stewart, *LINPACK Users Guide*, SIAM, 1979.

[34] S. P. E. Corporation. Spec benchmarks. [Online]. Available: http://www.spec.org/benchmarks.html

[35] C. Junghans, M. Karnstedt, and M. Gertz, “Quality-driven resource-adaptive data stream mining?” *SIGKDD Explor. Newsl.*, vol. 13, no. 1, pp. 72–82, Aug. 2011. [Online]. Available: http://doi.acm.org/10.1145/2031331.2031342

[36] S. Akioka, H. Yamana, and Y. Muraoka, “Data access pattern analysis on stream mining algorithms for cloud computation,” in *Proc. the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA2010)*, 2010.

[37] W. Wu and L. Gruenwald, “Research issues in mining multiple data streams,” in *Proc. the First International Workshop on Novel Data Stream Pattern Mining Techniques (StreamKDD’10)*, 2010.

# Data Quality Centric Application Framework for Big Data

Venkat N. Gudivada*, Dhana Rao [†], and William I. Grosky [‡]
*Department of Computer Science, East Carolina University, USA
[†]Department of Biology, East Carolina University, USA
[‡]Department of Computer and Information Science, University of Michigan - Dearborn, USA
email: gudivadav15@ecu.edu, raodh16@ecu.edu, and wgrosky@umich.edu

***Abstract*—Risks associated with data quality in Big Data have wide ranging adverse implications. Current research in Big Data primarily focuses on the proverbial harvesting of low hanging fruit and applications are developed using the Hadoop Ecosystem. In this paper, we discuss the risks and attendant consequences emanating from data quality in Big Data. We propose a data quality centric framework for Big Data applications and describe an approach to implementing it.**

***Keywords*—*Big Data*; *Data Quality*; *Data Analytics*; *Application Framework.***

## I. Introduction

Big Data, which has emerged in the last five years, has wide ranging implications for the society at large as well as individuals at a personal level. It has the potential for groundbreaking scientific discoveries and power for misuse and violation of personal privacy. Big Data poses research challenges and exacerbates data quality problems [1][2].

Though it is difficult to precisely define Big Data, it is often described in terms of five Vs – Volume, Velocity, Variety, Veracity, and Value. *Volume* refers to the unprecedented scale and velocity refers to the speed at which the data is being generated. The heterogeneous nature of data – unstructured, semi-structured, and structured – and associated data formats refers to the *variety* dimension. Typically Big Data goes through several data transformations from its inception before reaching the consumers. *Veracity* refers to data trustworthiness and *data provenance* is one way to specify veracity. Finally, the *value* dimension refers to unusual insights and actionable plans that are derived from Big Data through analytic processes.

### A. Apache Hadoop Ecosystem

Currently, most of Big Data research is focused on issues related to volume, velocity, and value. These investigations primarily use the Hadoop Ecosystem which encompasses Hadoop Distributed File System (HDFS), a high-performance parallel data processing engine called Hadoop MapReduce, and various tools for specific tasks. For example, *Pig* is a declarative language for ad hoc analysis. It is used to specify dataflows to extract, transform, and load (ETL) process and analyze large datasets. Pig generates MapReduce jobs that perform the dataflows and thus provides a high level abstract interface to MapReduce. The *Pig Latin* enhances the Pig through a programming language extension. It provides common data manipulation operations such as grouping, joining, and filtering. *Hive* is a tool for enabling data summarization, ad hoc query execution, and analysis of large datasets stored in HDFS-compatible file systems. In other words, *Hive* serves as a SQL-based data warehouse for the Hadoop Ecosystem.

The other widely used tools in the Hadoop Ecosystem include Cascading, Scalding, and Cascalog. *Cascading* is a popular high-level Java API that hides many of the complexities of MapReduce programming. *Scalding* and Cascalog are even higher level and concise APIs to Cascading, accessible from Scala and Clojure, respectively. While Scalding enhances Cascading with matrix algebra libraries, Cascalog adds logic programming constructs.

Hadoop Ecosystem tools for the velocity dimension are the Storm and Spark. *Storm* provides a distributed computational framework for event stream processing. It features an array of spouts specialized for receiving streaming data form disparate data sources, incremental computation, and computing metrics on rolling data windows in real-time. Like Storm, Spark supports real-time stream data processing and provides several additional libraries for database access, graph algorithms, and machine learning.

Amazon Web Services (AWS) Elastic MapReduce (EMR) is a cloud-hosted commercial offering of the Hadoop Ecosystem from Amazon. Microsoft's StreamInsight is a commercial product for stream data processing with focus on complex event processing applications.

### B. Industry Driven Big Data Research

Big Data research is primarily industry driven. Naturally, the focus is on the proverbial harvesting of the low-hanging fruit due to economic considerations. Research investigations in variety (aka data heterogeneity) and *veracity* dimensions are in the initial phases and tools are yet to emerge. However, data heterogeneity has been studied since 1980s in the database context,

ISBN: 978-1-61208-457-2

where the focus has been on database schema integration and distributed query processing. These investigations assumed that the data is structured and all the component databases use one of the three data models – relational, network, and hierarchical. However, in the Big Data context, the data is predominantly unstructured, and semi-structured and structured data is derived using a processing framework such as the Hadoop MapReduce.

Typically Big Data is obtained from multiple vendor sources. Maintaining data provenance [3] – record of original data sources and subsequent transformations applied to the data – plays a central role in data quality assurance. Veracity investigations are beginning to appear under the umbrella term *data provenance* [4][5].

### C. Data Quality Issues in Big Data

The *value* facet of Big Data critically depends on upstream data acquisition, cleaning, and transformation (ACT) tasks. Especially with the emerging Internet of Things (IoT) technologies, more and more data is machine generated. While some of the IoT data is generated under controlled conditions, most of it is created in environments which are subjected to fluctuations and thereby the quality of data can vary in an unpredictable manner. For example, the operating environment of wireless sensor and smart camera networks is subject to weather.

Data quality in ACT tasks has a direct bearing on volume, velocity, variety, and veracity facets of Big Data. Though data quality has been studied for over two decades, these investigations focused on database and information systems. Data quality assurance is the ultimate biggest challenge for Big Data management. Currently, data quality assurance requires intensive manual cleaning efforts. This is neither feasible nor economically viable.

In this paper, we discuss data quality issues in the Big Data context. We propose a data quality centric reference framework for developing Big Data applications. We also describe how this framework can be implemented using open source tools.

The remainder of the paper is structured as follows. Risks and implications of data quality are discussed in Section II. Data quality centric application framework for Big Data is described in Section III. Considerations for implementing this framework are discussed in Section IV. Finally, Section V concludes the paper.

## II. Data Science, Risks and Implications of Data Quality

The ability to effectively process massive datasets has become integral to a broad range of scientific investigations. Data Science is a new interdisciplinary academic area with data driven approaches to problem solving as the foundation. Big Data has the potential to fundamentally affect all walks of life.

### A. Data Science

Big Data is a double-edged sword. On the one hand, it enables scientists to overcome problems associated with small data samples. For example, it enables relaxing the assumptions of theoretical models, avoids over-fitting of models to *training data*, effectively deals with noisy training data, and provides ample *test data* to validate models.

Halevy, Norvig and Pereira [6] argue that the accurate selection of a mathematical model ceases its importance when compensated by *big enough* data. This insight is particularly important for tasks that are ill-posed for mathematically precise algorithmic solutions. Such tasks abound in natural language processing including language modeling, part-of-speech tagging, named entity recognition, and parsing. Ill-posed tasks also occur in applications such as computer vision, autonomous vehicle navigation, image and video processing.

Big Data enables a new paradigm for solving ill-posed problems by managing the complexity of the problem domain through building simple but high quality models by harnessing the power of massive data. For example, in the imaging domain, an image can be recovered by simply averaging successive image frames that are highly corrupted by a normally distributed Gaussian noise.

### B. Big Data Risks

On the flip side, Big Data poses several challenges for personal privacy. It provides opportunities for using it in ways that are different from the original intention [7]. Cases of Big Data misuse abound. González et al. [8] describe how individual human mobility patterns can be accurately predicted. Their study tracked position history of 100,000 anonymized mobile phone users over a six-month period. Contrary to the existing theories, this study found that human trajectories show a high degree of temporal and spatial regularity and individual travel patterns collapse into a single spatial probability distribution. They conclude that despite the diversity of peoples' travel history, they follow simple reproducible travel patterns. In other words, there is a correlation between spatio-temporal history and a person's identity.

Another case in point is how the retailer Target used its customers' purchase history and other information to accurately predict their shopping needs. This information is used to issue relevant coupons and improve sales [9]. Davis describes how to balance the benefits of big data innovation with the risk of harm from unintended consequences in [10].

Big Data also entails negative and in some cases even disastrous results, if the insights discovered are based on poor quality data. As we go about performing our daily activities, we are inevitably generating a data trail – for example, location tracking though GPS in mobile phones. This data is captured and stored persistently along with meta data, such as temporal information. It is possible to reconstruct a person's life history by fusing together seemingly disparate data acquired from multiple sources. This information can be further enhanced to gain insight into the behavior and activities of people, which in turn can be used to create new products and services. With only a little effort, personal data can be acquired, cleaned, aggregated, analyzed, sold, and repurposed [10].

### C. Data Quality Issues

As indicated earlier, data quality issues have been studied for over two decades in the context of corporate data governance, enterprise data warehousing, and the Web. However, in the Big Data context, the following issues are either unique to Big Data or their severity is more pronounced: streaming data, disparate data types and multiple data vendors, preponderance of machine generated unstructured data, and integration difficulties. The problem of flight delay prediction illustrates the case in point. Solutions to this problem consider historical data, current weather, departure time, departing city, and other concurrent flights. These data can be obtained from multiple sources including FlightView Flight Tracker, Flight Aware, Flightwise Tracker Pro, and Orbitz. These sources use different terminology and their data conflicts with each other. Data quality issues that arise in Web data in the context of two applications – Stock Markets and Airline Flights – is investigated in [?].

Many organizations acquire massive datasets from diverse data vendors to complement their internally generated data. Usually, the data acquired from the vendors is produced without any specific application or analysis context. Therefore, the perceived meaning of the data varies with the intended purpose [11]. This necessitates defining data validity and consistency in the context of intended use. Big Data life cycle is relatively long. Another issue raised by this is the inconsistency between the recent copy of the vendor supplied data and the previous version copy of the same data. The latter has been modified to conform to intended use-specific validity and consistency checks.

In summary, the grand challenge for Big Data applications is developing automated tools for resolving data quality issues. Currently, Big Data analysis requires significant manual cleansing of input data.

## III. Data Quality-centric Framework for Big Data Applications

We have investigated the requirements of telemedicine, environmental monitoring, and agriculture domains to help us define the data quality-centric framework for extracting value from Big Data. These requirements necessitate the framework to feature high performance computing cluster driven data analytics and knowledge extraction capabilities to harvest value from data. Data analytics and knowledge extraction involves managing complex and heterogeneous data and using advanced data fusion and information extraction algorithms. More specifically, automated tools are needed for processing and analyzing structured, semi-structured, unstructured data.

The structure of the framework is shaped by the tasks that are canonical across Big Data applications. Figure 1 shows task chain activities. We refer to this structure as Data Quality-centric Framework (DQF).

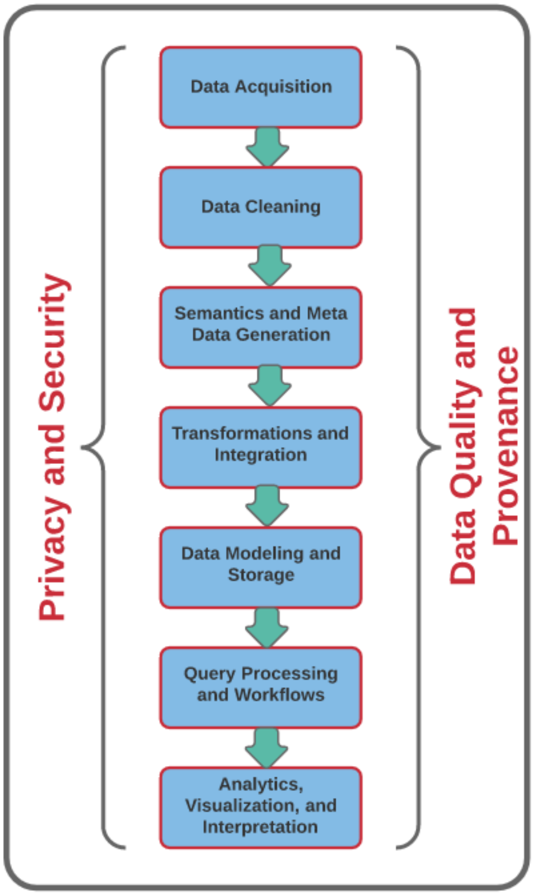

Privacy and Security
Data Acquisition
Data Cleaning
Semantics and Meta Data Generation
Transformations and Integration
Data Modeling and Storage
Query Processing and Workflows
Analytics, Visualization, and Interpretation
Data Quality and Provenance

Figure 1. Workflows in data quality-centric framework for Big Data applications

### A. Data Acquisition

Data acquisition devices can vary across the spectrum ranging from IoT, to wireless camera and sensor networks. Some of these devices are very simple in that they simply transmit the quantized data from the sensors. In addition to the sensed data, some devices will add meta data such as spatio-temporal and provenance data. Other devices may be much more sophisticated in that they employ sampling at Nyquist rate or variable sampling depending on the environmental conditions. Some devices may even apply real-time in-situ processing to detect anomalies and outliers and transmit only that data that has significance for intended data use.

Some data capture devices compress and transmit data using lossy or lossless data compression algorithms. Another important aspect of data acquisition is the scale dimension. Measurement theory specifies four levels or *scales* for assigning values to variables – nominal, ordinal, interval, and ratio. The chosen scale determines the type of processing that can be performed on the data.

### *B. Data Cleaning*

This is one of the most investigated areas of data quality and provides approaches and algorithms for inferring missing data, resolving conflicting and inconsistent data, detecting integrity constraint violations, and detecting and resolving outliers. Duplicate detection and elimination are also important data cleaning tasks.

Ganti and Sarma [12] describe a set of data cleaning tasks in an abstract manner to enable developing solutions for the common data cleaning tasks. They also discuss a few popular approaches for developing such solutions. They take an operator-centric approach for developing a data cleaning platform. The operators are customizable and serve as building blocks for data cleaning solutions. Other works in this direction include [13], [14].

### *C. Semantics and Meta Data Generation*

Bulk of the Big Data is unstructured in the form of video, images, audio, graphics, tweets, blogs, and natural language text. Information extraction techniques are used to turn unstructured data into semi- and structured data. For example, word boundary and sentence detection in spoken text, parts-of-speech tagging, parsing, named entity recognition, and coreference resolution are fundamental tasks in generating semi-structured representation from spoken and written text [15], [16].

### *D. Data Transformations and Integration*

Once the unstructured data is transformed into semi- and structured representations, associated data from multiple sources is fused to link related data. This task is referred to by various names including record linking, entity resolution, and data matching. For example, recognizing various pictures of the same person generated under different conditions as one and the same is entity resolution in image data. Record linking may lead to unexpected privacy violations. For example, various pieces of information about an individual viewed in isolation may not entail privacy violation. However, if these pieces are fused together using record linking techniques, this may lead to serious privacy violations.

Traditionally, Extract, Transform, and Load (ETL) tools [17] have been used for transformations task that involve structured data. ETL tools enable rule-based data transformations in batch processing mode and are capable of transforming data formats and detecting anomalies and outliers.

### *E. Data Modeling and Storage*

In relational database systems, data is uniformly modeled as relations. However, in the Big Data context, such a simple data model does not suffice. An assortment of new data models are used for Big Data and solutions based on such models are named NoSQL systems [18]. These systems can be grouped into classes, and each one meets the needs of a Big Data application category. Several data models for NoSQL systems have emerged during the last few years [19][20]. The new data models include key-value [21], column-oriented relational [22], column-family [23], document-oriented [24], and graph-based [25] [26]. The data modeling challenge in Big Data context is how to model heterogeneous data which requires multiple data models. It is more natural and practical to model the heterogeneous data using a collection of data models. Database-as-a-Service model [27] integrates a collection of data models and provides a unified interface.

Of late, the concept of data lakes is gaining popularity. A data lake is a storage repository that holds a vast amount of raw data in their native formats. Hadoop HDFS is often used for implementing data lakes since it is inherently better suited for storing large volumes of heterogeneous data with varying data formats.

### *F. Query Processing and Workflows*

This component of the DQF addresses query processing and optimization, programmatic interfaces, and defining and executing workflows. MapReduce and distributed computing principles are used in realizing this component.

NoSQL databases for Big Data are expected to provide five basic operations: Create (insert), Read (retrieve), Update (modify), Delete, and Search (CRUDS). The read operation retrieves data based on a precise match as in relational databases. In contrast, the search operation provides functionality similar to a Web search engine — the query is often imprecise and incomplete and the retrieved results are based on similarity measures and full-text search. For example, document data model based NoSQL systems provide full-text search by integrating with search engines and libraries such as Solr, Lucene, and EaslticSearch.

NoSQL systems' query languages vary a spectrum from procedural to declarative. Query languages and client interfaces are influenced by the data model and underlying storage engine. For example, ad hoc queries in MongoDB (a document-oriented NoSQL system) are expressed as *map* and *reduce* functions written in Javascript. On the other hand, Cassandra (a column family NoSQL system) provides a SQL-like query language called Cassandra Query Language (CQL).

Big Data workflows are similar to the conventional workflows. However, Big Data workflows introduce additional complexity which arises from disparate data types, semi-structured and unstructured data, and dramatic increase in storage and processing capacities. Given the volume and velocity of data, it should be possible to resume a failed workflow rather than starting all over.

### G. Analytics, Visualization, and Interpretation

This is the final component of the DQF and features functionality for Big Data analytics to discover and visualize actionable insights. It involves automatic hypothesis generation and testing and visual analytics. The latter facilitates analytical reasoning through interactive exploration using visual interfaces with human in the loop.

Big Data analytics enables several functions including hypothesis testing, population inferencing, inferencing about individuals in the population, profile construction, and outlier discovery. Hypotheses are statements that need validation. Hypotheses are formulated based on predictions from theory, heuristics, or hunches. In some cases, though controversial, hypotheses are automatically generated. The goal of hypothesis testing is to determine whether a hypothesis is supported by the available data.

Attributes characterize entities in the population. *Population inferencing* determines whether or not correlations exist between certain attributes among entities in the population. *Inferencing about individuals* may reveal, for example, whether or not an individual has been exhibiting consistent behavior over a period of time.

*Profile construction* is used to identify key characteristics that describe population classes. For example, what are the key characteristics of impulsive buyers?

*Outliers* are those entities in the population whose characteristics are drastically different from rest of the population. One may simply discard outliers assuming that they have risen due to data quality errors, or they may signify a new trend. Explaining an outlier often leads to valuable insights into the problem domain.

In all of the above tasks, data quality plays a critical role. The quality of inferences obtained is affected by the upstream activities - data acquisition, cleaning, semantics and meta data generation, transformations and integration.

### H. Privacy, Security, Data Quality and Provenance

These four facets pervade all the components of the DQF. *Differential privacy* is essential for Big Data. Just like authorization of entitlements in an application, differential privacy provides user access to data based on their job roles. Protecting the rights of privacy is a tremendous challenge. In 2013 alone, there were more than 13 million identity thefts in the United States [28]. Encryption in both hardware and software, and round the clock monitoring of security infrastructure are critical to protecting privacy.

A related issue is the notion of *personally identifiable information*, which is difficult to define precisely. Furthermore, as data goes through various transformations, it becomes even more difficult to identify and tag personally identifiable data elements. It has been shown that even anonymized data can often be re-identified and attributed to specific individuals [29].

*Data perturbation* is a technique for privacy preservation mainly used for electronic health records (EHR). It enables data analytics without compromising privacy requirements. It is considered as a more effective approach for privacy preservation of EHR compared to de-indentification and re-identification procedures.

Two types of data perturbation methods are suitable for EHR – *probability distribution* and *value distortion* approaches. In the first approach, the original data is replaced by data which is taken from the same distribution sample or from the distribution itself. In the second approach, the data is perturbed by adding randomly generated multiplicative or additive noise.

*Security* is implemented using access control and authorization mechanisms. *Access control* refers to ways in which user access to applications and databases is controlled. Databases limit access to those users who have been authenticated by the database itself or through an external authentication service, such as Kerberos [30]. *Authorization* controls what types of operations an authenticated user can perform. Access control and authorization capabilities of relational database systems have evolved over four decades. In contrast, data management for Big Data applications is provided by NoSQL systems [18]. Some systems provide limited security capabilities and others assume that the application is operating in a trusted environment and provide none.

As data goes through various processing steps, provenance is tracked and managed. Data provenance [31] is an issue that has received little or no attention from a security standpoint. As various transformations are applied to the data, metadata associated with provenance grows in complexity. The size of provenance graphs increases rapidly which makes analyzing them computationally expensive [4].

## IV. Implementing the Framework

Shown in Figure 2 is a reference architecture for implementing the DQF of Figure 1. This modular architecture enables mix and match best of the breed components for its implementation. The architecture is generic, configurable, and lends itself for implementation using stable and field-tested open source software components. We refer to this as Data Quality-centric Framework Architecture (DQFA).

### A. Remotely Deployed Wireless Sensor and Camera Networks

These are input capture devices which are connected though a wireless network. They are deployed in remote rural and mountainous communities that are not served by broadband networks. Wireless networks data is

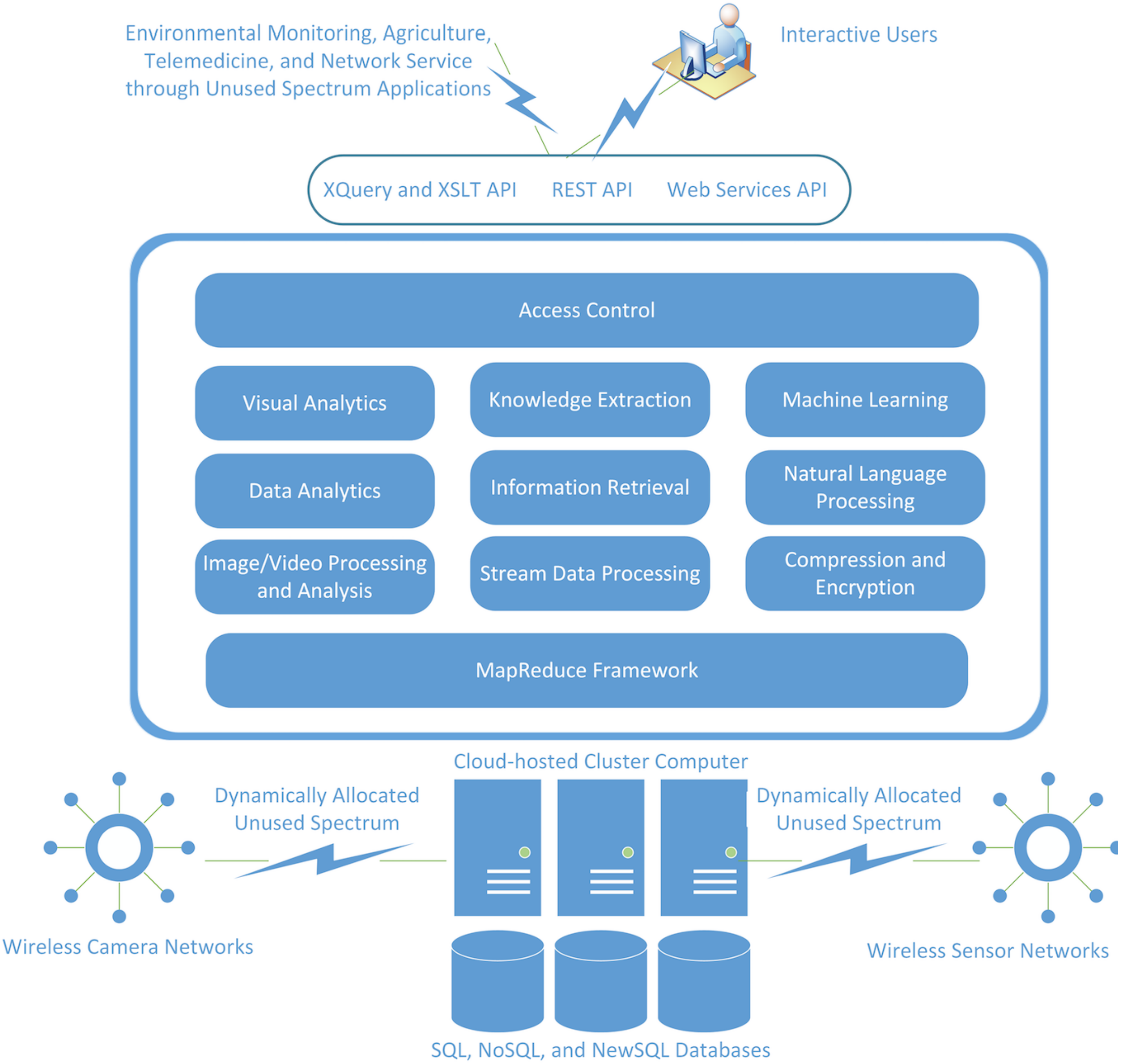

Environmental Monitoring, Agriculture, Telemedicine, and Network Service through Unused Spectrum Applications
Interactive Users
XQuery and XSLT API
REST API
Web Services API
Access Control
Visual Analytics
Knowledge Extraction
Machine Learning
Data Analytics
Information Retrieval
Natural Language Processing
Image/Video Processing and Analysis
Stream Data Processing
Compression and Encryption
MapReduce Framework
Cloud-hosted Cluster Computer
Dynamically Allocated Unused Spectrum
Dynamically Allocated Unused Spectrum
Wireless Camera Networks
Wireless Sensor Networks
SQL, NoSQL, and NewSQL Databases

Figure 2. Implementing the data quality-centric framework architecture

transmitted to the cloud-hosted cluster computers using a dynamically allocated unused spectrum.

Innovative uses of unused spectrum (aka white spaces) is gaining momentum in the US [32]–[34]. In addition to its current primary use as rural broadband, other uses of white spaces include connectivity Web for IoT, monitoring of oil and gas exploration and drilling, utilities monitoring [35], and smart grids [36], [37].

### *B. Cloud-hosted Cluster Computer*

Because of huge data volumes and the need for both batch and interactive processing, a cloud-hosted cluster computing platform will be used for implementing the DQFA. Each node in the cluster is self-contained and acts independently to remove single point of resource contention or failure. In this shared-nothing architecture, nodes share neither memory nor disk storage.

### *C. SQL, NoSQL, and NewSQL Databases*

Until recently, Relational Database Management Systems (RDBMS) were the mainstay for managing all types of data. Underlying the RDBMS is the relational model for structuring data and SQL query language for data manipulation and retrieval. Though RDBMS are a perfect

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-457-2

fit for many applications, they maybe less suitable or an expensive for certain applications. An array of new systems for data management have emerged in recent years to address the needs of such applications.

Currently there are over 300 systems for data management and new ones are introduced routinely. They are referred to by various names including NoSQL, NewSQL, Not Only SQL, and non-RDBMS. By design, these new database systems do not provide all the RDBMS features. They principally focus on providing near real-time reads and writes in the order of billions and millions, respectively. The DQFA will leverage these advances for data storage and processing.

### *D. MapReduce Framework*

MapReduce is computational paradigm for computing arbitrary functions on massive datasets in parallel if the computation fits a three-step pattern: map, shard and reduce. The *map process* is a highly parallel one comprised of several processes. Each one processes a different segment of data and produces (key, value) pairs. The *shard process* collects the generated pairs, sorts and partitions them. Each partition is assigned to a different *reduce* process, which produces one result. The DQFA infrastructure will feature MapReduce framework to speed up both batch and interactive jobs.

### *E. Compression and Encryption*

Massive data volumes require compression as a means to reduce storage requirements. Encryption converts data into unreadable form to ensure data integrity and confidentiality. Some DQFA-driven applications require compression and encryption to meet regulatory compliance enforced by government and industry standards organizations. Tools we will consider for this task include Basic Compression Library, Google's Zopfli Compression Algorithm, LZ4, and LZ4_HC.

### *F. Stream Data Processing*

The ubiquity of networked sensors is leading to sensorization of the real world. For example, some environmental monitoring applications generate streaming data. A concomitant effect is the emergence of many novel monitoring and control applications that require high-volume and low-latency processing.

Due to tremendous data volume, stream data is not stored in its entirety. First, the data is analyzed to determine which subset of it meets specified patterns and anomalies. Only such data is stored and processed further. The DQFA will provide an engine for stream data processing. Open source software to consider for this task include Apache Storm and Apache Spark.

### *G. Image/Video Processing and Analysis*

Smart camera networks generate image and video data streams. It is not practical to store all streaming data. Techniques such as statistical sampling and abnormal event detection are required to identify image and video data that has informational value. Open source libraries to consider for this task include NIH's ImageJ, OpenCV, ImageMagick, CImg, Scipy, and Java Advanced Imaging (JAI).

### *H. Data Analytics*

The set of tools required for data analytics is vast and varied. They encompass domain-independent descriptive and inferential statistics, as well as domain-specific processes and tools. Open source libraries to consider for this task include GNU Scientific Library (GSL), Computational Geometry Algorithms Library (CGAL), NumPy and SciPy.

### *I. Visual Analytics*

Visual Analytics is an emerging area which integrates the analytic capabilities of the computer and the abilities of the human analyst [38], [39]. It is the science of analytical reasoning facilitated by visual interactive interfaces. High performance computing is the backbone of Visual Analytics.

Visual Analytics offers great potential for uncovering unexpected and hidden insights in heterogeneous healthcare data, which may lead to ground-breaking discoveries and profitable innovation [40]–[42]. Open source libraries to consider for this task include Flare, Gephi, Google Vis, Graph Viz, IVTK, D3.js, and JGraph.

### *J. Information Retrieval*

Information Retrieval (IR) deals with modeling and retrieving of information from semi-structured and unstructured documents [43]. They provide full-text indexing and support various types of search including Boolean search and document-structure based search. They also rank the search results.

IR capability is essential for the DQFA to enable information fusion for knowledge extraction. Open source libraries to consider for this task include Apache OpenNLP, Stanford NLP, NLTK, Apache Lucene, Apache Solr, ElasticSearch, and Splunk.

### *K. Natural Language Processing*

Natural Language Processing (NLP) based querying complements IR search [16]. NLP tools are available for part-of-speech tagging, named entity recognition, parsing, abstracting and summarization, text and speech generation, and machine translation.

NLP tools will be used in DQFA to extract information from unstructured documents and to enable knowledge extraction. These tools will also be used for providing flexible and natural interfaces for user interaction. Open source libraries to consider for this task include Natural Language Toolkit (NLTK), Stanford CoreNLP, WordNet, SRILM, Apache Lucene, MontyLingua, and tm.

### L. Machine Learning

Machine learning algorithms are central to knowledge extraction. They include algorithms for basic statistics, feature extraction and transformation, classification and regression, clustering, dimensionality reduction, and optimization [44]. Machine learning libraries that we will consider include PyML, Apache Mahout, MLib, dlibml, WEKA, and scikit-learn.

### M. Knowledge Extraction

Knowledge extraction is a domain-dependent task [45]. It involves creating knowledge from disparate sources of data and information represented in forms such as structured relational databases, semi-structured document databases, text corpora, image and video collections, semantic annotations, XML, RDF, and ontologies. Open source tools that we will consider for this task include AIDA, AlchemyAPI, Apache Stanbol, DBPedia Spotlight, FOX, FRED, and NERD.

### N. Data Provenance

As computing has become distributed, the need to ensure security and privacy of data has increased greatly. In the proposed DQFA, data is created, processed, propagated, and consumed by diverse domain scientists, belonging to different security domains.

Secure data provenance is a key technology to ensure analysis results are reproducible by recording the lineage of data and information transformation processes. Tools that we will consider for this task are Pentaho Kettle, eBioFlow, PLIER, and SPADE.

### O. Access Control

This component is used for user rights management. Access control provides two important functions. First, the identity of entities accessing the DQFA is confirmed. This step is referred to as *authentication*. Second, the confirmed entities are restricted to perform only those functions that are authorized. This step is referred to as *authorization*. We will consider tools such as OpenDJ, OpenIDM, OpenAM, DACS, and Shibboleth for implementation of this component.

### P. Interfaces, Application Programming and User Access

We envision DQFA to provide several interfaces to promote flexible access to its services.

*1) Interactive Users:* Interactive users engage in exploratory style interaction with the system and expect real-time response. For example, visual analytics requires active participation of the domain scientists to discover patterns and formulate hypotheses.

*2) Web Services API:* This API will be designed to expose DQFA services to other applications. It enables applications to communicate and exchange data without concern for programming language, operating system, and network protocol issues.

*3) REST API:* Representational State Transfer (REST) is a minimal overhead Hypertext Transfer Protocol (HTTP) API for interacting with DAIS infrastructure. REST uses four HTTP methods – GET (for reading data), POST (for writing data), PUT (for updating data) and DELETE (for removing data).

*4) XQuery and XSLT API:* XQuery provides a declarative means for querying, updating, and transforming semi-structured and unstructured data mostly in the form of hierarchically structured XML documents. XQuery contains a superset of XPath expression syntax to address specific parts of an XML document. An extension to the XQuery/XPath language specifies how full-text search queries be specified as XQuery functions. XSLT is another declarative language for specifying how to transform an XML document into another.

XPath, XQuery, XQuery/XPath Full-text Search are all W3C standards. XSLT 3.0 has W3C Last Call Working Draft status. XQuery and XSLT API entail several advantages to the DAIS infrastructure. They include reduced applications development time through the use of standards, performance gain through elimination of data mappings between application layers by using the same data model, and enabling nontechnical staff to perform development and maintenance work.

## V. Conclusions

Data quality plays a critical role in Big Data applications. As data goes through various transformations and meanders from upstream to downstream applications, data quality errors propagate and accumulate. These error have the potential to cause detrimental consequences for an organization or individual. The data quality centric application framework model we proposed is intended to serve as a reference model to promote data quality research in Big Data context. Our future research direction is to implement this framework.

## References

[1] V. Gudivada, D. Rao, and V. Raghavan, *Big Data Analytics*. Elsevier, 2015, ch. Big Data Driven Natural Language Processing Research and Applications, pp. 203 – 238.

[2] V. Gudivada, R. Baeza-Yates, and V. Raghavan, “Big data: Promises and problems,” *IEEE Computer*, vol. 48, no. 3, pp. 20–23, Mar. 2015.

[3] P. Buneman, J. Cheney, W.-C. Tan, and S. Vansummeren, “Curated databases,” in *Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, ser. PODS ’08. New York, NY, USA: ACM, 2008, pp. 1–12.

[4] Y.-W. Cheah, “Quality, retrieval and analysis of provenance in large-scale data,” Ph.D. dissertation, Indianapolis, IN, USA, 2014.

[5] J. Cheney, P. Buneman, and B. Ludäscher, “Report on the principles of provenance workshop,” *SIGMOD Rec.*, vol. 37, no. 1, pp. 62–65, Mar. 2008.

[6] A. Halevy, P. Norvig, and F. Pereira, “The unreasonable effectiveness of data,” *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8 – 12, 2009.

[7] M. R. Wigan and R. Clarke, “Big data’s big unintended consequences,” *Computer*, vol. 46, no. 6, pp. 46–53, Jun. 2013.

[8] M. C. González, C. A. Hidalgo, and A.-L. Barabási, “Understanding individual human mobility patterns,” *Nature*, vol. 453, no. 7196, pp. 779–782, Jun. 2008.

[9] C. Duhigg. How companies learn your secrets. [retrieved: December, 2015]. [Online]. Available: http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?_r=0

[10] K. Davis, *Ethics of Big Data: Balancing Risk and Innovation.* O'Reilly Media, Inc., 2012.

[11] D. Loshin. Understanding big data quality for maximum information usability. [retrieved: December, 2015]. [Online]. Available: http://www.dataqualitybook.com

[12] V. Ganti and A. D. Sarma, *Data Cleaning: A Practical Perspective*, ser. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2013.

[13] J. W. Osborne, *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data.* SAGE Publications, 2012.

[14] Q. E. McCallum, *Bad Data Handbook: Cleaning Up The Data So You Can Get Back To Work.* O'Reilly Media, 2012.

[15] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.* O'Reilly Media, 2009.

[16] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 2nd ed. Prentice Hall, 2009.

[17] M. Casters, R. Bouman, and J. van Dongen, *Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration.* Wiley, 2010.

[18] V. Gudivada, D. Rao, and V. Raghavan, "Renaissance in data management systems: Sql, nosql, and newsql," *IEEE Computer, forthcoming.*

[19] solid IT. Knowledge base of relational and NoSQL database management systems. [retrieved: December, 2015]. [Online]. Available: http://db-engines.com/en/ranking

[20] A. Schram and K. M. Anderson, "Mysql to nosql: Data modeling challenges in supporting scalability," in *Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity*, ser. SPLASH '12. New York, NY, USA: ACM, 2012, pp. 191–202.

[21] R. Gandhi, A. Gupta, A. Povzner, W. Belluomini, and T. Kaldewey, "Mercury: Bringing efficiency to key-value stores," in *Proceedings of the 6th International Systems and Storage Conference*, ser. SYSTOR '13. New York, NY, USA: ACM, 2013, pp. 6:1–6:6.

[22] Z. Liu, S. Natarajan, B. He, H.-I. Hsiao, and Y. Chen, "Cods: Evolving data efficiently and scalably in column oriented databases," *Proc. VLDB Endow.*, vol. 3, no. 1-2, pp. 1521–1524, Sep. 2010.

[23] A. Lakshman and P. Malik, "Cassandra: A structured storage system on a p2p network," in *Proceedings of the Twenty-first Annual Symposium on Parallelism in Algorithms and Architectures*, ser. SPAA '09. New York, NY, USA: ACM, 2009, pp. 47–47.

[24] P. Murugesan and I. Ray, "Audit log management in mongodb," *2014 IEEE World Congress on Services*, pp. 53–57, 2014.

[25] R. Angles, "A comparison of current graph database models," *2014 IEEE 30th International Conference on Data Engineering Workshops*, pp. 171–177, 2012.

[26] I. Robinson, J. Webber, and E. Eifrem, *Graph Databases.* O'Reilly, 2013.

[27] D. Agrawal, A. El Abbadi, F. Emekci, and A. Metwally, "Database management as a service: Challenges and opportunities," in *Data Engineering, 2009. ICDE '09. IEEE 25th International Conference on*, March 2009, pp. 1709–1716.

[28] United Credit Service. Identity theft; will you be the next vitcim? [retrieved: December, 2015]. [Online]. Available: https://ucscollections.wordpress.com/2014/03/06/identity-theft-will-you-be-the-next-victim/

[29] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," *2013 IEEE Symposium on Security and Privacy*, vol. 0, pp. 111–125, 2008.

[30] S. T. F. Al-Janabi and M. A. S. Rasheed, "Public-key cryptography enabled kerberos authentication," in *Developments in E-systems Engineering (DeSE), 2011.* IEEE Computer Society, Dec 2011, pp. 209–214.

[31] U. Braun, A. Shinnar, and M. Seltzer, "Securing provenance," in *Proceedings of the* $3^{rd}$ *Conference on Hot Topics in Security*, ser. HOTSEC'08, 2008, pp. 4:1–4:5.

[32] P. Bahl, R. Chandra, T. Moscibroda, R. Murty, and M. Welsh, "White space networking with wi-fi like connectivity," *SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 4, pp. 27–38, Aug. 2009.

[33] R. Chandra, T. Moscibroda, P. Bahl, R. Murty, G. Nychis, and X. Wang, "A campus-wide testbed over the tv white spaces," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 15, no. 3, pp. 2–9, Nov. 2011.

[34] R. Chandra, "White space networking beyond the tv bands," in *Proceedings of the Seventh ACM International Workshop on Wireless Network Testbeds, Experimental Evaluation and Characterization*, ser. WiNTECH '12, 2012, pp. 1–2.

[35] C.-S. Sum, H. Harada, F. Kojima, Z. Lan, and R. Funada, "Smart utility networks in tv white space," *IEEE Communications Magazine*, vol. 49, no. 7, pp. 132–139, 2011. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5936166

[36] G. Nychis, B. DeBruhl, and H. Tang, "Demo: Tv white space networking capabilities and potential with an embedded & open-api platform," in *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '14, 2014, pp. 309–312.

[37] S. W. Oh, F. Chin, and S. G. Kerk, "Tv white-space for smart grid," in *Proceedings of the 4th International Conference on Cognitive Radio and Advanced Spectrum Management*, ser. CogART '11, 2011, pp. 56:1–56:5.

[38] P. Alzamora, Q. V. Nguyen, S. Simoff, and D. Catchpoole, "A novel 3d interactive visualization for medical data analysis," in *Proceedings of the 24th Australian Computer-Human Interaction Conference*, ser. OzCHI '12. New York, NY, USA: ACM, 2012, pp. 19–25.

[39] T. E. Hansen, J. P. Hourcade, A. Segre, C. Hlady, P. Polgreen, and C. Wyman, "Interactive visualization of hospital contact network data on multi-touch displays," in *Proceedings of the 3rd Mexican Workshop on Human Computer Interaction*, ser. MexIHC '10. San Luis Potos, S.L.P. Mexico, Mxico: Universidad Politcnica de San Luis Potos, 2010, pp. 15–22.

[40] J. J. Caban and D. Gotz, "2011 workshop on visual analytics in healthcare: understanding the physician perspective," *SIGHIT Rec.*, vol. 2, no. 1, pp. 29–31, Mar. 2012.

[41] F. Fischer, F. Mansmann, and D. A. Keim, "Real-time visual analytics for event data streams," in *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, ser. SAC '12. New York, NY, USA: ACM, 2012, pp. 801–806.

[42] D. Gotz and J. Sun, "Ieee visweek workshop on visual analytics in health care 2010," *SIGHIT Rec.*, vol. 1, no. 1, pp. 31–32, Mar. 2011.

[43] C. D. Manning, P. Raghavan, and H. Sch́utze, *Introduction to Information Retrieval.* Wiley, 2010.

[44] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning From Data*, 2012.

[45] R.-S. Chen, C.-C. Chang, and I. Chi, "Ontology-based knowledge extraction-a case study of software development," in *Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2006. SNPD 2006. Seventh ACIS International Conference on*, Jun. 2006, pp. 91–96.

# Combining “Small Data” from Surveys and “Big Data” from Online Experiments at Pinterest

Jolie M. Martin
Quantitative User Experience Research
Pinterest
San Francisco, USA
jolie@pinterest.com

*Abstract*— **Running experiments while logging detailed user actions has become the standard way of testing product features at Pinterest, as at many other Internet companies. While this technique offers plenty of statistical power to assess the effects of product changes on behavioral metrics, it does not often give us much insight into *why* users respond the way they do. By combining at-scale experiments with smaller surveys of users in each experimental condition, we have developed a unique approach for measuring the impact of our product and communication treatments on user sentiment, attitudes, and comprehension.**

***Keywords: Experiments; Methodology; Surveys***

## I. Experiments at Pinterest

The foundation for our mixed methodology research at Pinterest is a solid experimental framework and process that we have adapted from our forerunners like Google [1], Yahoo! [2], and Facebook [3]. Due to our smaller size and capacity, though, Pinterest experiments do not aim to study generic individual or social decision-making, but rather the context-dependent decisions of our users. The product variations we test via experimentation can be as imperceptible to users as the re-ranking of recommended Pins, or as major as a complete redesign of the Pin close-up view. The unique challenges we face, distinct from those of more established companies, are in helping users to understand the value propositions of the service (discovering and saving personally relevant content) despite lower awareness in both the U.S. and globally.

Our experiments – as at other technology companies – aim to measure the impacts of product changes on the user experience before launching these changes to everyone. An experiment will usually be exposed to around 1% of users for a period of several weeks. Of course, there are experiments where only particular subsets of the user population are even eligible, such as restricting tooltips about search to those who have never searched on the site before. On the other hand, there are features with network effects (e.g., communication tools) that cannot be captured unless they are rolled out to a broader set of users at once. We try to clearly define our criteria for success prior to running an experiment so that the point at which to end the experiment and what action to take (usually, “launch” or “do not launch”) are straightforward.

## II. Surveys at Pinterest

One shortcoming of a purely experimental approach, however, is that we often want to learn something more broadly about our product and users than just about the specific experimental arms tested. Since we clearly cannot run every variation on the seemingly infinite set of possible conditions, we need alternative means to discover the fundamental reasons for observable behavioral differences. Surveys provide some of this insight, and enable us to include the quality of user experience – as opposed to behavioral metrics alone – in our launch criteria. In these surveys, we simply ask users what their perceptions are about some aspect of their experience on Pinterest. From their responses, we aim to extrapolate the underlying causes of behavioral differences across experimental arms that will then suggest the most promising future iterations of the same experiment, and in some cases, even unrelated experiments.

We typically survey just a relatively small subset of users pulled randomly from each experimental arm since detecting differences in multiple-choice responses requires a much lower sample size than detecting very subtle behavioral changes, such as propensity to click-through to the origin website of a Pin. The rule of thumb we employ is to survey as few users as possible to discern the distribution of responses and correlate them with behavior. Although the primary goal of a survey is not to provide a feedback forum, we do attempt to be minimally disruptive and retain the Pinterest “voice” by avoiding tedious or robotic questions, as well as following all of the other best practices for running surveys.

## III. Merging “Big” and “Small” Data

Until recently, we operationalized surveys as emails to users and panel samples that select for Pinterest usage, and sometimes this is still the best way of reaching those who rarely visit the site. As an alternative, we have created a set of technical tools and documented guidelines for inviting users to surveys directly within the Pinterest product. The benefits of in-product invites are multifold: (1) accessing a more representative set of users, including those who are less likely to respond to email surveys, as evidenced by far higher response rates for in-product survey invites, (2) providing context to respondents about the parts of Pinterest we reference in our questions, and (3) tracking user actions immediately preceding and following survey responses.

Despite these benefits, it is worth noting that there are some inherent complexities involved with running surveys in conjunction with experiments. Aside from the engineering challenge of ensuring that surveys trigger for the intended users, we need to take into account any systematic biases in that sample. For example, if a survey invite appeared only the fifth time a user landed on their Pinterest home feed, it would clearly be skewed toward a more active sample. In addition, the wording of questions needs to be as specific as possible while still making sense for users in different experimental arms. If some of these users have recently experienced a change in the product due to the experimental treatment, we want to ensure that they understand the version to which we refer. On the other hand, for more subtle experiments, we cannot expect users to have noticed any difference at all.

Thus, our combined experimental and survey approach should be employed only in consideration of the research questions at hand and the users being targeted. One instance where the benefits outweighed the drawbacks was a study of the new user signup flow. The experimental arms varied in the education users received about Pinterest as they created an account. The survey they received immediately following asked where they first heard about Pinterest, what prompted them to sign up, their perceived relevance of content on the site (previewed to them in the education), and expected future use. We then correlated these responses with first-day actions so that we could draw inferences about the attitudes of new users outside of the survey sample solely from their logged actions as a means of segmentation. We also measured interactions between attitudes and experimental treatment in predicting engagement over time to assess which signup conditions increased retention for different segments of users. This type of analysis allows us to customize the product to accommodate distinct groups of users, or in some cases, to keep the product homogeneous yet better understand how changes impact different groups of users.

While the effort of such surveys is not justified for all research questions we wish to answer, they help us to better understand user self-reported satisfaction and comprehension in instances where an experiment's behavioral findings could be attributed not only to the functionality of a feature, but to some combination of other explanations such as awareness, understanding, or privacy concerns. Teasing these apart via surveys then guides not only the actions we take directly as a result of the experiment, but also our design of future experiments and product iterations.

## REFERENCES

[1] Y. Chen, T. H. Ho, and Y. M. Kim, "Knowledge market design: A field experiment at Google Answers," Journal of Public Economic Theory, vol. 12 (4), pp. 641-664, 2010.

[2] M. Ostrovsky and M. Schwarz, "Reserve prices in internet advertising auctions: A field experiment," Proc. of the 12th ACM Conference on Electronic Commerce, ACM, June 2011, pp. 59-60.

[3] R. M. Bond, C. J. Fariss, J. J. Jones, A. D. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler, "A 61-million-person experiment in social influence and political mobilization," Nature, vol. 489 (7415), pp. 295-298, 2012.

# Unsupervised aircraft trajectories clustering: a minimum entropy approach

Florence Nicol

Ecole Nationale de l'Aviation Civile
7, Avenue Edouard Belin,
F-31055 Toulouse FRANCE
Email: nicol@recherche.enac.fr

Stephane Puechmorel

Ecole Nationale de l'Aviation Civile
7, Avenue Edouard Belin,
F-31055 Toulouse FRANCE
Email: stephane.puechmorel@enac.fr

*Abstract*—**Clustering is a common operation in statistics. When data considered are functional in nature, like curves, dedicated algorithms exist, mostly based on truncated expansions on Hilbert basis. When additional constraints are put on the curves, like in applications related to air traffic where operational considerations are to be taken into account, usual procedures are no longer applicable. A new approach based on entropy minimization and Lie group modeling is presented here, yielding an efficient unsupervised algorithm suitable for automated traffic analysis. It outputs cluster centroids with low curvature, making it a valuable tool in airspace design applications or route planning.**

***Keywords–curve clustering; probability distribution estimation; functional statistics; minimum entropy; air traffic management.***

## I. INTRODUCTION

Clustering aircraft trajectories is an important problem in Air Traffic Management (ATM). It is a central question in the design of procedures at take-off and landing, the so called sid-star (Standard Instrument Departure and Standard Terminal Arrival Routes). In such a case, one wants to minimize the noise and pollutants exposure of nearby residents while ensuring runway efficiency in terms of the number of aircraft managed per time unit.

The same question arises with cruising aircraft, this time the mean flight path in each cluster being used to optimally design the airspace elements (sectors and airways). This information is also crucial in the context of future air traffic management systems where reference trajectories will be negotiated in advance so as to reduce congestion. A special instance of this problem is the automatic generation of safe and efficient trajectories, but in such a way that the resulting flight paths are still manageable by human operators. Clustering is a key component for such tools: major traffic flows must be organized in such a way that the overall pattern is not too far from the current organization, with aircraft flying along airways. The classification algorithm has thus not only to cluster similar trajectories but at the same time make them as close as possible to operational trajectories. In particular, straightness of the flight segments must be enforced, along with a global structure close to a graph with nodes corresponding to merging/splitting points and edges the airways.

## II. PREVIOUS RELATED WORK

Several well established algorithms may be used for performing clustering on a set of trajectories, although only a few of them were eventually applied in the context of air traffic. The spectral approach relies on trajectories modeling as vectors of samples in a high dimensional space, and uses random projections as a mean of reducing the dimensionality. The huge computational cost of the required singular values decomposition is thus alleviated, allowing use on real recorded traffic over several months. It was applied in a study conducted by the Mitre corporation on behalf of the Federal Aviation Authority (FAA) [1]. The most important limitation of this approach is that the shape of the trajectories is not taken into account when applying the clustering procedure unless a resampling procedure based on arclength is applied: changing the time parametrization of the flight paths will induce a change in the classification. Furthermore, there is no mean to put a constraint on the mean trajectory produced in each cluster: curvature may be quite arbitrary even if samples individually comply with flight dynamics.

Another approach is taken in [2], with an explicit use of an underlying graph structure. It is well adapted to road traffic as vehicles are bound to follow predetermined segments. A spatial segment density is computed then used to gather trajectories sharing common parts. For air traffic applications, it may be of interest for investigating present situations, using the airways and beacons as a structure graph, but will misclassify aircraft following direct routes which is quite a common situation, and is unable to work on an unknown airspace organization. This point is very important in applications since trajectory datamining tools are mainly used in airspace redesign. A similar approach is taken in [3] with a different measure of similarity. It has to be noted that many graph-based algorithms are derived from the original work presented in [4], and exhibit the aforementioned drawbacks for air traffic analysis applications.

An interesting vector field based algorithm is presented in [5]. A salient feature is the ability to distinguish between close trajectories with opposite orientations. Nevertheless, putting constraints on the geometry of the mean path in a cluster is quite awkward, making the method unsuitable for our application.

Due to the functional nature of trajectories, that are basically mappings defined on a time interval, it seems more appropriate to resort to techniques based on times series, as surveyed in [6], [7] or functional data statistics, with standard references [8], [9]. In both approaches, a distance between pairs of trajectories or, in a weaker form, a measure of similarity must be available. The algorithms of the first category are based on sequences, possibly in conjunction with dynamic time warping [10] while in the second samples are assumed to come from an unknown underlying function belonging to a given Hilbert space. However, it has to be noticed that apart from this last assumption, both approaches yield similar end algorithms, since functional data revert for implementation to usual finite dimensional vectors of expansion coefficients on a suitable truncated basis. For the same reason, model-based clustering may be used in the context of functional data even if no notion of probability density exists in the original infinite dimensional Hilbert space as mentioned in[11]. A nice example of a model-based approach working on functional data is funHDDC [12].

## III. Dealing with curve systems: a paradigm change

When working with aircraft trajectories, some specific characteristics must be taken into account. First of all, flight paths consist mainly of straight segments connected by arcs of circles, with transitions that may be assumed smooth up to at least the second derivative. This last property comes from the fact that pilot's actions result in changes on aerodynamic forces and torques and a straightforward application of the equations of motion. When dealing with sampled trajectories, this induces a huge level of redundancy within the data, the relevant information being concentrated around the transitions. Second, flight paths must be modeled as functions from a time interval $[a, b]$ to $\mathbb{R}^3$ which is not the usual setting for functional data statistics: most of the work is dedicated to real valued mappings and not vector ones. A simple approach will be to assume independence between coordinates, so that the problem falls within the standard case. However, even with this simplifying hypothesis, vertical dimension must be treated in a special way as both the separation norms and the aircraft maneuverability are different from those in the horizontal plane.

Finally, being able to cope with the initial requirement of compliance with the current airspace structure in airways is not addressed by general algorithms. In the present work, a new kind of functional unsupervised classifier is introduced, that has in common with graph-based algorithms an estimation of traffic density but works in a continuous setting. For operational applications, a major benefit is the automatic building of a route-like structure that may be used to infer new airspace designs. Furthermore, smoothness of the mean cluster trajectory, especially low curvature, is guaranteed by design. Such a feature is unique among existing clustering procedures. Finally, our Lie group approach makes easy the separation between neighboring flows oriented in opposite directions. Once again, it is mandatory in air traffic analysis where such a situation is common.

In the first section the notion of entropy of a curve system is introduced. The modeling of trajectories with a Lie group approach is then presented. The next two sections will show how to estimate Lie group densities and to cluster curves in this new setting. Finally, results on a synthetic example are briefly given and a conclusion is drawn.

## IV. The entropy of a system of curves

Considering trajectories as mappings $\gamma\colon [t_0, t_1] \to \mathbb{R}^3$ induces a notion of spatial density as presented in [13]. Assuming that after a suitable registration process all flight paths $\gamma_i, i = 1, \dots, N$ are defined on the same time interval $[0, 1]$ to $\Omega$ a domain of $\mathbb{R}^3$, one can compute an entropy associated with the system of curves using the approach presented in [14]. Let a system of curves $\gamma_1, \dots, \gamma_N$ be given, its entropy is defined to be:

$$E(\gamma_1, \dots, \gamma_N) = -\int_\Omega \tilde{d}(x) \log\left(\tilde{d}(x)\right) dx,$$

where the spatial density $d$ is computed according to:

$$\tilde{d}\colon x \mapsto \frac{\sum_{i=1}^N \int_0^1 K\left(\|x - \gamma_i(t)\|\right) \|\gamma_i'(t)\| dt}{\sum_{i=1}^N l_i}. \qquad (1)$$

In the last expression, $l_i$ is the length of the curve $\gamma_i$ and $K$ is a kernel function similar to those used in nonparametric estimation. A standard choice is the Epanechnikov kernel:

$$K\colon x \mapsto C\left(1 - x^2\right) 1_{[-1,1]}(x),$$

with a normalizing constant $C$ chosen so as to have a unit integral of $K$ on $\Omega$.

Since the entropy is minimal for concentrated distributions, it is quite intuitive to figure out that seeking for a curve system $(\gamma_1, \dots, \gamma_N)$ giving a minimum value for $E(\gamma_1, \dots, \gamma_N)$ will induce the following properties:

- The images of the curves tend to get close one to another.
- The individual lengths will be minimized: it is a direct consequence of the fact that the density has a term in $\gamma'$ within the integral that will favor short trajectories.

Using a standard gradient descent algorithm on the entropy produces an optimally concentrated curve system, suitable for use as a basis for a route network. Figure 2 illustrates this effect on an initial situation given in Figure 1.

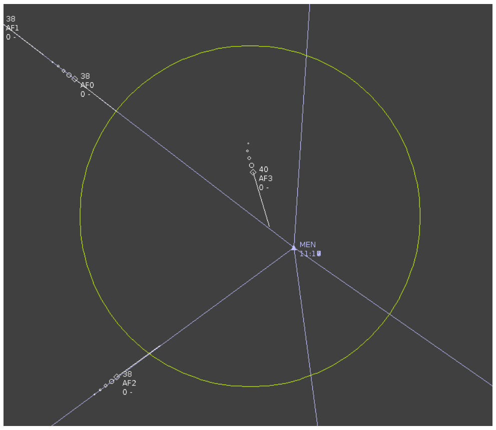


Figure 1. Initial flight plan.

The displacement field for trajectory $j$ is oriented at each point along the normal vector to the trajectory, with norm given by:

$$\int_\Omega \left.\frac{\gamma_j(t) - x}{\|\gamma_j(t) - x\|}\right|_{\mathcal{N}} K'\left(\|\gamma_j(t) - x\|\right) \log \tilde{d}(x) dx \|\gamma_j'(t)\| \qquad (2)$$

$$-\left(\int_\Omega K\left(\|\gamma_j(t) - x\|\right) \log \tilde{d}(x)) dx\right) \left.\frac{\gamma_j''(t)}{\|\gamma_j'(t)\|}\right|_{\mathcal{N}} \qquad (3)$$

$$+\left(\int_\Omega \tilde{d}(x) \log(\tilde{d}(x)) dx\right) \left.\frac{\gamma_j''(t)}{\|\gamma_j'(t)\|}\right|_{\mathcal{N}}, \qquad (4)$$

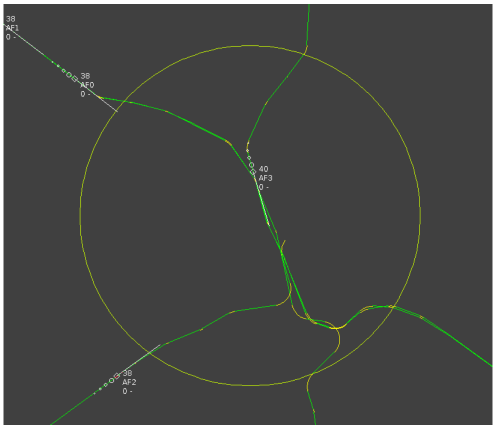


Figure 2. Entropy minimal curve system from the initial flight plan.

where the notation $v_{|\mathcal{N}}$ stands for the projection of the vector $v$ onto the normal vector to the trajectory. An overall scaling constant of:

$$\frac{1}{\sum_{i=1}^{N} l_i},$$

where $l_i$ is the length of trajectory $i$, has to be put in front of the expression to get the true gradient of the entropy. In practice, it is not needed since algorithms will adjust the size of the step taken in the gradient direction.

## V. A Lie group modeling

While satisfactory in terms of traffic flows, the previous approach suffers from a severe flaw when one considers flight paths that are very similar in shape but are oriented in opposite directions. Since the density is insensitive to direction reversal, flight paths will tend to aggregate while the correct behavior will be to ensure a sufficient separation in order to prevent hazardous encounters. Taking aircraft headings into account in the clustering process is then mandatory when such situations have to be considered.

This issue can be addressed by adding a penalty term to neighboring trajectories with different headings but the important theoretical property of entropy minimization will be lost in the process. A more satisfactory approach will be to take heading information directly into account and to introduce a notion of density based on position and velocity.

Since the aircraft dynamics is governed by a second order equation of motion of the form:

$$\begin{pmatrix} \gamma\prime(t) \\ \gamma''(t) \end{pmatrix} = F\left(t;\ \begin{matrix} \gamma(t) \\ \gamma'(t) \end{matrix}\right),$$

it is natural to take as state vector:

$$\begin{pmatrix} \gamma(t) \\ \gamma'(t) \end{pmatrix}.$$

The initial state is chosen here to be:

$$\left(\frac{0_d}{e_1}\right),$$

with $e_1$ the first basis vector, and $0_d$ the origin in $\mathbb{R}^d$. It is equivalent to model the state as a linear transformation:

$$0_d \otimes e_1 \mapsto T(t) \otimes A(t)(0_d \otimes e_1) = \gamma(t) \otimes \gamma'(t),$$

where $T(t)$ is the translation mapping $0_d$ to $\gamma(t)$ and $A(t)$ is the composite of a scaling and a rotation mapping $e_1$ to $\gamma'(t)$. Considering the vector $(\gamma(t), 1)$ instead of $\gamma(t)$ allows a matrix representation of the translation $T(t)$:

$$\left(\frac{\gamma(t)}{1}\right) = \left(\begin{array}{c|c} Id & \gamma(t) \\ \hline 0 & 1 \end{array}\right)\left(\frac{0_d}{1}\right).$$

From now, all points will be implicitly considered as having an extra last coordinate with value 1, so that translations are expressed using matrices. The origin $0_d$ will thus stand for the vector $(0, \ldots, 0, 1)$ in $\mathbb{R}^{d+1}$. Gathering things yields:

$$\left(\frac{\gamma(t)}{\gamma'(t)}\right) = \left(\begin{array}{c|c} T(t) & 0 \\ \hline 0 & A(t) \end{array}\right)\left(\frac{0_d}{e_1}\right). \qquad (5)$$

The previous expression makes it possible to represent a trajectory as a mapping from a time interval to the matrix Lie group $\mathcal{G} = \mathbb{R}^d \times \Sigma \times S\mathbb{O}(d)$, where $\Sigma$ is the group of multiples of the identity, $S\mathbb{O}(d)$ the group of rotations and $\mathbb{R}^d$ the group of translations. Please note that all the products are direct. The $A(t)$ term in the expression (5) can be written as an element of $\Sigma \otimes S\mathbb{O}(d)$. Starting with the defining property $A(t)e_1 = \gamma'(t)$, one can write $A(t) = \|\gamma'(t)\| U(t)$ with $U(t)$ a rotation mapping $e_1 \in \mathbb{S}^{d-1}$ to the unit vector $\gamma'(t)/\|\gamma'(t)\| \in \mathbb{S}^{d-1}$. For arbitrary dimension $d$, $U(t)$ is not uniquely defined, as it can be written as a rotation in the plane $\mathcal{P} = \text{span}(e_1, \gamma'(t))$ and a rotation in its orthogonal complement $\mathcal{P}^{\perp}$. A common choice is to let $U(t)$ be the identity in $\mathcal{P}^{\perp}$ which corresponds in fact to a move along a geodesic (great circle) in $\mathbb{S}^{d-1}$. This will be assumed implicitly in the sequel, so that the representation $A(t) = \Lambda(t)U(t)$ with $\Lambda(t) = \|\gamma'(t)\|\text{Id}$ becomes unique.

The Lie algebra $\mathfrak{g}$ of $\mathcal{G}$ is easily seen to be $\mathbb{R}^d \times \mathbb{R} \times$ **Asym**$(d)$ with **Asym**$(d)$ is the space of skew-symmetric $d \times d$ matrices. An element from $\mathfrak{g}$ is a triple $(u, \lambda, A)$ with an associated matrix form:

$$M(u, \lambda, A) = \left(\begin{array}{cc|c} 0 & \multicolumn{1}{|c|}{u} & \\ 0 & \multicolumn{1}{|c|}{0} & 0 \\ \hline \multicolumn{2}{c|}{0} & \lambda Id + A \end{array}\right). \qquad (6)$$

The exponential mapping from $\mathfrak{g}$ to $\mathcal{G}$ can be obtained in a straightforward manner using the usual matrix exponential:

$$\exp((u, \lambda, A)) = \exp(M(u, \lambda, A)).$$

The matrix representation of $\mathfrak{g}$ may be used to derive a metric:

$$\langle (u, \lambda, A), (v, \mu, B) \rangle_{\mathfrak{g}} = \textbf{Tr}\left(M(u, \lambda, A)^t M(v, \mu, B)\right).$$

Using routine matrix computations and the fact that $A, B$ being skew-symetric have vanishing trace, it can be expressed as:

$$\langle (u, \lambda, A), (v, \mu, B) \rangle_{\mathfrak{g}} = n\lambda\mu + \langle u, v \rangle + \textbf{Tr}\left(A^t B\right). \qquad (7)$$

A left invariant metric on the tangent space $T_g\mathcal{G}$ at $g \in \mathcal{G}$ is derived from (7) as:

$$\langle\!\langle X, Y, \rangle\!\rangle_g = \langle g^{-1}X, g^{-1}Y \rangle_{\mathfrak{g}},$$

with $X, Y \in T_g\mathcal{G}$. Please note that $\mathcal{G}$ is a matrix group acting linearly so that the mapping $g^{-1}$ is well defined from $T_g\mathcal{G}$ to $\mathfrak{g}$. Using the fact that the metric (7) splits, one can check that geodesics in the group are given by straight segments in $\mathfrak{g}$: if

$g_1, g_2$ are two elements from $\mathcal{G}$, then the geodesic connecting them is:

$$t \in [0,1] \mapsto g_1 \exp\left(t \log\left(g_1^{-1} g_2\right)\right).$$

where $\log$ is a determination of the matrix logarithm. Finally, the geodesic length is used to compute the distance $d(g_1, g_2)$ between two elements $g_1, g_2$ in $\mathcal{G}$. Assuming that the translation parts of $g_1, g_2$ are respectively $u_1, u_2$, the rotations $U_1, U_2$ and the scalings $\exp(\lambda_1), \exp(\lambda_2)$ then:

$$d(g_1, g_2)^2 = (\lambda_1 - \lambda_2)^2 + \tag{8}$$

$$\mathbf{Tr}\left(\log\left(U_1^t U_2\right)\log\left(U_1^t U_2\right)^t\right) + \|u_1 - u_2\|^2. \tag{9}$$

An important point to note is that the scaling part of an element $g \in \mathcal{G}$ will contribute to the distance by its logarithm.

Based on the above derivation, a flight path $\gamma$ with state vector $(\gamma(t), \gamma'(t))$ will be modeled in the sequel as a curve with values in the Lie group $\mathcal{G}$:

$$\Gamma\colon t \in [0,1] \mapsto \Gamma(t) \in \mathcal{G},$$

with:

$$\Gamma(t).(0_d, e_1) = (\gamma(t), \gamma'(t)).$$

In order to make the Lie group representation amenable to statistical thinking, we need to define probability densities on the translation, scaling and rotation components that are invariant under the action of the corresponding factor of $\mathcal{G}$.

## VI. Nonparametric estimation on $\mathcal{G}$

Since the translation factor in $\mathcal{G}$ is the additive group $\mathbb{R}^d$, a standard nonparametric kernel estimator can be used. It turns out that it is equivalent to the spatial density estimate of (1), so that no extra work is needed for this component. As for the rotation component, a standard parametrization is obtained recursively starting with the image of the canonical basis of $\mathbb{R}^d$ under the rotation. If $R$ is an arbitrary rotation and $e_1, \ldots, e_d$ is the canonical basis, there is a unique rotation $R_{e_1}$ mapping $e_1$ to $Re_1$ and fixing $e_2, \ldots, e_d$. It can be represented by the point $Re_1 = r_1$ on the sphere $\mathbb{S}^{d-1}$. Proceeding the same way for $Re_2, \ldots Re_d$, it is finally possible to completely parametrized $R$ by a $(d-1)$-uple $(r_1, \ldots, r_{d-1})$ where $r_i \in \mathbb{S}^{i-1}$, $i = 1, \ldots, d$. Finding a rotation invariant distribution amounts thus to construct such a distribution on the sphere.

In directional statistics, when we consider the spherical polar coordinates of a random unit vector $u \in \mathbb{S}^{d-1}$, we deal with spherical data (also called circular data or directional data) distributed on the unit sphere. For $d = 3$, a unit vector may be described by means of two random variables $\theta$ and $\varphi$ which respectively represent the co-latitude (the zenith angle) and the longitude (the azimuth angle) of the points on the sphere. Nonparametric procedures, such as the kernel density estimation methods are sometimes convenient to estimate the probability distribution function (p.d.f.) of such kind of data but they require an appropriate choice of kernel functions.

Let $X_1, \ldots, X_n$ be a sequence of random vectors taking values in $\mathbb{R}^d$. The density function $f$ of a random $d$-vector may be estimated by the kernel density estimator [15] as follows:

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_H\left(x - X_i\right),\ x \in \mathbb{R}^d,$$

where $\mathcal{K}_H(x) = \mid H \mid^{-1} \mathcal{K}(H^{-1}x)$, $\mathcal{K}$ denotes a multivariate kernel function and $H$ represents a $d$-dimensional smoothing matrix, called bandwidth matrix. The kernel function $\mathcal{K}$ is a $d$-dimensional p.d.f. such as the standard multivariate Gaussian density $\mathcal{K}(x) = (2\pi)^{d/2}\exp\left(-\frac{1}{2}x^T x\right)$ or the multivariate Epanechnikov kernel. The resulting estimation will be the sum of "bumps" above each observation, the observations closed to $x$ giving more important weights to the density estimate. The kernel function $\mathcal{K}$ determines the form of the bumps whereas the bandwidth matrix $H$ determines their width and their orientation. Thereby, bandwidth matrices can be used to adjust for correlation between the components of the data. Usually, an equal bandwidth $h$ in all dimensions is chosen, corresponding to $H = hId$ where $Id$ denotes the $d \times d$ identity matrix. The kernel density estimator then becomes:

$$\widehat{f}(x) = \frac{1}{nh^d}\sum_{i=1}^{n} \mathcal{K}\left(h^{-1}(x - X_i)\right),\ x \in \mathbb{R}^d.$$

In certain cases when the spread of data is different in each coordinate direction, it may be more appropriate to use different bandwidths in each dimension. The bandwidth matrix $H$ is given by the diagonal matrix in which the diagonal entries are the bandwidths $h_1, \ldots, h_d$.

In directional statistics, a kernel density estimate on $\mathbb{S}^{d-1}$ is given by adopting appropriate circular symmetric kernel functions such as von Mises-Fisher, wrapped Gaussian and wrapped Cauchy distributions. A commonly used choice is the von Mises-Fisher (vMF) distribution on $\mathbb{S}^{d-1}$ which is denoted $\mathcal{M}(m, \kappa)$ and given by the following density expression [16]:

$$K_{VMF}(x; m, \kappa) = c_d(\kappa)\, e^{\kappa m^T x},\ \kappa > 0,\ x \in \mathbb{S}^{d-1}, \tag{10}$$

where

$$c_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)} \tag{11}$$

is a normalization constant with $I_r(\kappa)$ denoting the modified Bessel function of the first kind at order $r$. The vMF kernel function is an unimodal p.d.f. parametrized by the unit mean-direction vector $\mu$ and the concentration parameter $\kappa$ that controls the concentration of the distribution around the mean-direction vector. The vMF distribution may be expressed by means of the spherical polar coordinates of $x \in \mathbb{S}^{d-1}$ [17].

Given the random vectors $X_i$, $i = 1, \ldots, n$, in $\mathbb{S}^{d-1}$, the estimator of the spherical distribution is given by:

$$\begin{aligned}\widehat{f}(x) &= \frac{1}{n}\sum_{i=1}^{n} K_{VMF}(x; X_i) \\ &= \frac{c_d(\kappa)}{n}\sum_{i=1}^{n} e^{\kappa X_i^T x},\ \kappa > 0,\ x \in \mathbb{S}^{d-1}.\end{aligned}$$

Please, note that the quantity $x - X_i$ which appears in the linear kernel density estimator is replaced by $X_i^T x$ which is the cosine of the angles between $x$ and $X_i$, so that more important weights are given on observations close to $x$ on the sphere. The concentration parameter $\kappa$ is a smoothing parameter that plays the role of the inverse of the bandwidth parameter as defined in the linear kernel density estimation. Large values of $\kappa$ imply greater concentration around the mean direction and lead to undersmoothed estimators whereas small values provide oversmoothed circular densities [18]. Indeed, if $\kappa = 0$, the vMF kernel function reduces to the uniform circular distribution on the hypersphere. Note that the vMF kernel function is convenient when the data is rotationally symmetric.

The vMF kernel function is a convenient choice for our problem because this p.d.f. is invariant under the action on the sphere of the rotation component of the Lie group $\mathcal{G}$. Moreover, this distribution has properties analogous to those of multivariate Gaussian distribution and is the limiting case of a limit central theorem for directional statistics. Other multidimensional distributions might be envisaged, such as the bivariate von Mises, the Bingham or the Kent distributions [16]. However, the bivariate von Mises distribution being a product kernel of two univariate von Mises kernels, this is more appropriate for modeling density distributions on the torus and not on the sphere. The Bingham distribution is bimodal and satisfies the antipodal symmetry property $K(x) = K(-x)$. This kernel function is used for estimating the density of axial data and is not appropriate for our clustering approach. Finally, the Kent distribution is a generalization of the vMF distribution, which is used when we want to take into account of the spread of data. However, the rotation-invariance property of the vMF distribution is lost.

As for the scaling component of $\mathcal{G}$, the usual kernel functions such as the Gaussian and the Epanechnikov kernel functions are not suitable for estimating the radial distribution of a random vector in $\mathbb{R}^d$. When distributions are defined over a positive support (here in the case of non-negative data), these kernel functions cause a bias in the boundary regions because they give weights outside the support. An asymmetrical kernel function on $\mathbb{R}^+$ such as the log-normal kernel function is a more convenient choice. Moreover, this p.d.f. is invariant by change of scale. Let $R_1, \dots, R_n$ be univariate random variables from a p.d.f. which has bounded support on $[0; +\infty[$. The radial density estimator may be defined by means of a sum of log-normal kernel functions as follows:

$$\widehat{g}(r) = \frac{1}{n} \sum_{i=1}^{n} K_{LN}(r; \ln R_i, h), r \geq 0, h > 0,$$

where

$$K_{LN}(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$$

is the log-normal kernel function and $h$ is the bandwidth parameter. The resulting estimate is the sum of bumps defined by log-normal kernels with medians $R_i$ and variances $(e^{h^2} - 1)e^{h^2} R_i^2$. Note that the log-normal (asymmetric) kernel density estimation is similar to the kernel density estimation based on a log-transformation of the data with the Gaussian kernel function. Although the scale-change component of $\mathcal{G}$ is the multiplicative group $\mathbb{R}^+$, we can use the standard Gaussian kernel estimator and the metric on $\mathbb{R}$.

## VII. UNSUPERVISED ENTROPY CLUSTERING

The first thing to be considered is the extension of the entropy definition to curve systems with values in $\mathcal{G}$. Starting with expression from (1), the most important point is the choice of the kernel involved in the computation. As the group $\mathcal{G}$ is a direct product, choosing $K = K_t.K_s.K_o$ with $K_t, K_s, K_o$ functions on respectively the translation, scaling and rotation part will yield a $\mathcal{G}$-invariant kernel provided the $K_t, K_s K_o$ are invariant on their respective components. Since the translation part of $\mathcal{G}$ is modeled after $\mathbb{R}^n$, the epanechnikov kernel is a suitable choice. As for the scaling and rotation, the choice made follows the conclusion of section VI: a log-normal kernel and a von-Mises one will be used respectively. Finally, the term $\|\gamma'(t)\|$ in the original expression of the density, that is required to ensure invariance under re-parametrization of the curve, has to be changed according to the metric in $\mathcal{G}$ and is replaced by $\langle\langle \gamma'(t), \gamma'(t) \rangle\rangle_{\gamma(t)}^{1/2}$. The density at $x \in \mathcal{G}$ is thus:

$$d_{\mathcal{G}}(x)) = \frac{\sum_{i=1}^{N} \int_0^1 K\left(x, \gamma_i(t)\right) \langle\langle \gamma_i'(t), \gamma_i'(t) \rangle\rangle_{\gamma_i(t)}^{1/2} dt}{\sum_{i=1}^{N} l_i} \quad (12)$$

where $l_i$ is the length of the curve in $\mathcal{G}$, that is:

$$l_i = \int_0^1 \langle\langle \gamma_i'(t), \gamma_i'(t) \rangle\rangle_{\gamma_i(t)}^{1/2} dt \quad (13)$$

The expression of the kernel evaluation $K\left(x, \gamma_i(t)\right)$ is split into three terms. In order to ease the writing, a point $x$ in $\mathcal{G}$ will be split into $x^r, x^s, x^o$ components where the exponent $r, s, t$ stands respectively for translation, scaling and rotation. Given the fact that $K$ is a product of component-wise independent kernels it comes:

$$K\left(x, \gamma_i(t)\right) = K_t\left(x^t, \gamma_i^t(t)\right) K_s\left(x^s, \gamma_i^s(t)\right) K_o\left(x^o, \gamma_i^o(t)\right)$$

where:

$$K_t(x^t, \gamma_i^t(t)) = \text{ep}\left(\|x^t - \gamma_i^t(t)\|\right) \quad (14)$$

$$K_s(x^s, \gamma_i^s(t)) = \frac{1}{x^s \sigma \sqrt{2\pi}} \exp\left(-\frac{\left(\log x^s - \log \gamma_i^s(t)\right)^2}{2\sigma^2}\right) \quad (15)$$

$$K_o(x^o, \gamma_i^o(t)) = C(\kappa) \exp\left(\kappa \mathbf{Tr}\left(x^{ot} \gamma_i^o(t)\right)\right) \quad (16)$$

with $C(\kappa)$ the normalizing constant making the kernel of unit integral. Please note that the expression given here is valid for arbitrary rotations, but for the application targeted by the work presented here, it boils down to a standard von-mises distributions on $\mathbb{S}^{d-1}$:

$$K_o(x^o, \gamma_i^o(t)) = C(\kappa) \exp\left(\kappa x^{ot} \gamma_i^o(t)\right)$$

with normalizing constant as given in (11). In the general case, it is also possible, writing the rotation as a sequence of moves on spheres $\mathbb{S}^{d-1}, \mathbb{S}^{d-2}, \dots$ and the distribution as a product of von-Mises on each of them, to have a vector of parameters $\kappa$: it is the approach taken in [19] and it may be applied verbatim here if needed.

The entropy of the system of curves is obtained from the density in $\mathcal{G}$:

$$E(d_{\mathcal{G}}) = -\int_{\mathcal{G}} d_{\mathcal{G}}(x) \log d_{\mathcal{G}}(x) d\mu_{\mathcal{G}}(x) \quad (17)$$

with $d\mu_{\mathcal{G}}$ the left Haar measure. Using again the fact that $\mathcal{G}$ is a direct product group, $d\mu$ is easily seen to be a product measure, with $dx^t$, the usual Lebesgue measure on the translation part, $dx^s/x^s$ on the scaling part and the lebesgue measure $dx^o$ on $\mathbb{S}^{d-1}$ for the rotation part. It turns out that the $1/x^s$ term in the expression of $dx^s/x^s$ is already taken into account in the kernel definition, due to the fact that it is expressed in logarithmic coordinates. The same is true for the Von-Mises kernel, so that in the sequel only the (product) lebesgue measure will appear in the integrals.

Finding the system of curves with minimum entropy requires a displacement field computation as detailed in [14]. For each curve $\gamma_i$, such a field is a mapping $\eta_i \colon [0,1] \to T\mathcal{G}$ where at each $t \in [0,1]$, $\eta_i(t) \in T\mathcal{G}_{\gamma_i(t)}$.Compare to the original situation where only spatial density was considered, the computation must now be conducted in the tangent space to $\mathcal{G}$. Even for small problems, the effort needed becomes prohibitive. The structure of the kernel involved in the density can help in cutting the overall computations needed. Since it is a product, and the translation part is compactly supported, being an epanechnikov kernel, one can restrict the evaluation to points belonging to its support. Density computation will thus be made only in tubes around the trajectories.

Second, for the target application that is to cluster the flight paths into a route network and is of pure spatial nature, there is no point in updating the rotation and scaling part when performing the moves: only the translation part must change, the other two being computed from the trajectory. The initial optimization problem in $\mathcal{G}$ may thus be greatly simplified.

Let $\epsilon$ be an admissible variation of curve $\gamma_i$, that is a smooth mapping from $[0,1]$ to $T\mathcal{G}$ with $\epsilon(t) \in T_{\gamma_i(t)}\mathcal{G}$ and $\epsilon(0) = \epsilon(1) = 0$. We assume furthermore that $\epsilon$ has only a translation component. The derivative of the entropy $E(d_\mathcal{G})$ the t curve $\gamma_i$ is obtained from the first order term when $\gamma_i$ is replaced by $\gamma_i + \epsilon$. First of all, it has to be noted that $d_\mathcal{G}$ is a density and thus has unit integral regardless of the curve system. When computing the derivative of $E(d_\mathcal{G})$, the term

$$-\int_\mathcal{G} d_\mathcal{G}(x)\frac{\partial_{\gamma_i} d_\mathcal{G}(x)}{d_\mathcal{G}(x)} d\mu_\mathcal{G}(x) = -\int_\mathcal{G} \partial_{\gamma_i} d_\mathcal{G}(x) d\mu_\mathcal{G}(x)$$

will thus vanish. It remains:

$$-\int_\mathcal{G} \partial_{\gamma_i} d_\mathcal{G}(x) \log d_\mathcal{G}(x) d\mu_\mathcal{G}(x)$$

The density $d_\mathcal{G}$ is a sum on the curves, and only the $i$-th term has to be considered. Starting with the expression from (12), one term in the derivative will come from the denominator. It computes the same way as in [14] to yield:

$$\left.\frac{\gamma_i^{t\prime\prime}(t)}{\langle\!\langle \gamma_i'(t), \gamma_i'(t)\rangle\!\rangle_\mathcal{G}}\right|_\mathcal{N} E(d_\mathcal{G}) \tag{18}$$

Please note that the second derivative of $\gamma_i$ is considered only on its translation component, but the first derivative makes use of the complete expression. As before, the notation $|_\mathcal{N}$ stands for the projection onto the normal component to the curve.

The second term comes from the variation of the numerator. Using the fact that the kernel is a product $K^t K^s K^o$ and that all individual terms have a unit integral on their respective components, the expression becomes very similar to the case of spatial density only and is:

$$-\left(\int_\mathcal{G} K\left(x, \gamma_i(t)\right) \log d_\mathcal{G}(x) d\mu_{\mathcal{G}(x)}\right) \left.\frac{\gamma_i^{t\prime\prime}(t)}{\langle\!\langle \gamma_i'(t), \gamma_i'(t)\rangle\!\rangle_\mathcal{G}^{1/2}}\right|_\mathcal{N} \tag{19}$$

$$+\int_{\mathbb{R}^d} e(t) K^{t\prime}\left(x^t, \gamma_i^t(t)\right) \log d_\mathcal{G}(x) \langle\!\langle \gamma_i'(t), \gamma_i'(t)\rangle\!\rangle_\mathcal{G}^{1/2} dx^t \tag{20}$$

with:

$$e(t) = \left.\frac{\gamma_i^t(t) - x^t}{\|\gamma_i^t(t) - x^t\|}\right|_\mathcal{N}$$

## VIII. Results

Only partial results are available for the moment and several traffic situations are still to be considered. On simple synthetic examples, the algorithm works as expected, avoiding going to close to trajectories with opposite directions as indicated on Figure 3.

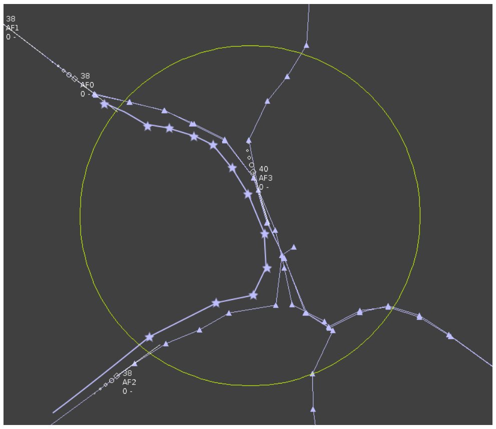


Figure 3. Clustering using the Lie approach

In a more realistic setting, the arrivals and departures at Toulouse Blagnac airport were analyzed. The algorithm performs well as indicated on Figure 4. Four clusters are identified, with mean lines represented through a spline smoothing between landmarks. It is quite remarkable that all density based algorithms were unable to separate the two clusters located at the right side of the picture, while the present one clearly show a standard approach procedure and a short departure one.

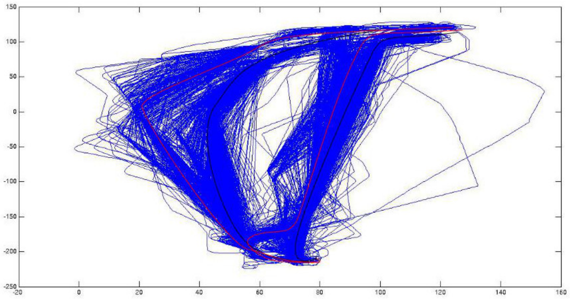


Figure 4. Bundling trajectories at Toulouse airport

An important issue still to be addressed with the extended algorithm is the increase in computation time that reaches 20 times compared to the appoach using only spatial density entropy. In the current implementation, the time needed to cluster the traffic presented in Figure 3 is in the order of 0.01s on a XEON 3Ghz machine and with a pure java implementation. For the case of Figure 4, 5 minutes are needed on the same machine for dealing with the set of 1784 trajectories.

## IX. Conclusion and future work

The entropy associated with a system of curves has proved itself efficient in unsupervised clustering application where shape constraints must be taken into account. For using it in aircraft route design, heading and velocity information must be added to the state vector, inducing an extra level of complexity. The present work relies on a Lie group modeling as an unifying approach to state representation. It has successfully extended the notion of curve system entropy to this setting, allowing the heading/velocity to be added in a intrinsic way. The method seems promising, as indicated by the results obtained on simple synthetic situations, but extra work needs to be dedicated to algorithmic efficiency in order to deal with the operational traffic datasets, in the order of tens of thousand of trajectories.

Generally speaking, introducing a Lie group approach to data description paves the way to new algorithms dedicated to data with a high level of internal structuring. Studies are initiated to address several issues in high dimensional data analysis using this framework.

## References

[1] M. Enriquez, “Identifying temporally persistent flows in the terminal airspace via spectral clustering,” in ATM Seminar 10, FAA-Eurocontrol, Ed., 06 2013.

[2] M. El Mahrsi and F. Rossi, “Graph-based approaches to clustering network-constrained trajectory data,” in New Frontiers in Mining Complex Patterns, ser. Lecture Notes in Computer Science, A. Appice, M. Ceci, C. Loglisci, G. Manco, E. Masciari, and Z. Ras, Eds. Springer Berlin Heidelberg, 2013, vol. 7765, pp. 124–137.

[3] J. Kim and H. S. Mahmassani, “Spatial and temporal characterization of travel patterns in a traffic network using vehicle trajectories,” Transportation Research Procedia, vol. 9, 2015, pp. 164 – 184, papers selected for Poster Sessions at The 21st International Symposium on Transportation and Traffic Theory Kobe, Japan, 5-7 August, 2015.

[4] M. Ester, H. peter Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise.” AAAI Press, 1996, pp. 226–231.

[5] F. N., K. J. T., S. C. E., and S. C.T., “Vector field k -means: Clustering trajectories by fitting multiple vector fields,” in Eurographics Conference on Visualization (EuroVis), Preim, P. Rheingans, and H. Theisel, Eds., 2013.

[6] T. W. Liao, “Clustering of time series data - a survey,” Pattern Recognition, vol. 38, 2005, pp. 1857–1874.

[7] S. Rani and G. Sikka, “Recent techniques of clustering of time series data: A survey,” International Journal of Computer Applications, vol. 52, no. 15, August 2012, pp. 1–9, full text available.

[8] F. Ferraty and P. Vieu, Nonparametric Functional Data Analysis: Theory and Practice, ser. Springer Series in Statistics. Springer, 2006.

[9] J. Ramsay and B. Silverman, Functional Data Analysis, ser. Springer Series in Statistics. Springer New York, 2006.

[10] W. Meesrikamolkul, V. Niennattrakul, and C. Ratanamahatana, “Shape-based clustering for time series data,” in Advances in Knowledge Discovery and Data Mining, ser. Lecture Notes in Computer Science, P.-N. Tan, S. Chawla, C. Ho, and J. Bailey, Eds. Springer Berlin Heidelberg, 2012, vol. 7301, pp. 530–541.

[11] A. Delaigle and P. Hall, “Defining probability density for a distribution of random functions,” The Annals of Statistics, vol. 38, no. 2, 2010, pp. 1171–1193.

[12] C. Bouveyron and J. Jacques, “Model-based clustering of time series in group-specific functional subspaces,” Advances in Data Analysis and Classification, vol. 5, no. 4, 2011, pp. 281–300.

[13] S. Puechmorel, “Geometry of curves with application to aircraft trajectory analysis.” Annales de la facult des sciences de Toulouse, vol. 24, no. 3, 07 2015, pp. 483–504.

[14] S. Puechmorel and F. Nicol, “Entropy minimizing curves with application to automated flight path design,” Lecture notes in computer science, Geometric Science of Information 2015” in MDPI Entropy, 2015.

[15] D. Scott, Multivariate Density Estimation: Theory, Practice, and Visualization, ser. A Wiley-interscience publication. Wiley, 1992.

[16] K. Mardia and P. Jupp, Directional Statistics, ser. Wiley Series in Probability and Statistics. Wiley, 2009.

[17] K. V. Mardia, “Statistics of directional data,” Journal of the Royal Statistical Society. Series B (Methodological), vol. 37, no. 3, 1975, pp. 349–393.

[18] E. García-Portugués, R. M. Crujeiras, and W. González-Manteiga, “Kernel density estimation for directional–linear data,” Journal of Multivariate Analysis, vol. 121, 2013, pp. 152–175.

[19] P. E. Jupp and K. V. Mardia, “Maximum likelihood estimators for the matrix von mises-fisher and bingham distributions,” Ann. Statist., vol. 7, no. 3, 05 1979, pp. 599–606.

# Modelling Support for a Linked Data Approach to Tool Interoperability

Jad El-khoury, Didem Gurdur, Frederic Loiret, Martin Törngren
Department of Machine Design
KTH Royal Institute of Technology
Stockholm, Sweden
email:{jad, dgurdur, floiret, martint}@kth.se

Da Zhang, Mattias Nyberg
Scania CV AB
Södertälje, Sweden
email: {da.zhang, mattias.nyberg}@scania.com

***Abstract*— Linked Data is increasingly being adopted for the integration of software tools, especially with the emergence of the Open Services for Lifecycle Collaboration (OSLC) standard on tool interoperability. In this paper, we present a modelling approach – with accompanying tool support – for the specification of Linked Data resources, focusing on the particular needs of tool-chain development. The approach provides graphical models for the specification of constraints on resources being shared in the tool-chain. Moreover, it aims to maintain a centralized understanding and management of the overall information model being handled in the federated tool-chain architecture. This is achieved through an integrated set of modelling views that cover the early phases of tool-chain development.**

***Keywords-Linked data modelling; OSLC; resource shapes; tool integration; information modelling.***

## I. Introduction

Over last few decades, the ongoing trend of adopting the Model-Driven Engineering (MDE) approach to product development promised an improvement in the quality and efficient access to product and process information, given that such information becomes managed through explicitly defined meta-models. However, the heterogeneity and complexity of modern industrial products requires the use of many engineering software tools, needed by the different engineering disciplines (such as mechanical, electrical, embedded systems and software engineering), and throughout the entire development life cycle (requirements analysis, design, verification and validation, etc.).

So, while MDE is a step in the right direction, unless interoperability mechanisms are developed to connect information across the model-based engineering tools, MDE may lead to isolated "islands of information", given the natural distribution of information across the many tools and data sources involved.

As an example from the automotive industry, the functional safety standard ISO 26262:2011 [1] mandates that requirements and design components are developed at several levels of abstraction; and that a clear traceability exists between requirements from the different levels, as well as between requirements and system components. The earlier practice, in which development artefacts are handled as text-based documentation, rendered such traceability ineffective – if not impossible. Even with the adoption of model-driven engineering, it remains a challenge to trace between the artefacts being created by the various engineering tools, in order to comply with the standard.

In summary, current development practices need a faster shift from the localized document-based handling of artefacts, towards a Federated Information-based Development Environment (F-IDE), where the information from all development artefacts is made accessible, consistent and correct throughout the development phases, disciplines and tools.

In this paper, we advocate the use of the Linked Data principles as a basis for such an F-IDE (See [4] for Tim Berners-Lee's four principles of Linked Data.). Yet, when applying these principles for parts of the development environment at the truck manufacturer Scania AB, certain challenges were encountered that needed to be addressed. We here describe our approach on how these challenges were tackled. In the next section, after a short motivation for adopting the Linked Data principles, we present a case study that will be used in the remaining paper. We then further elaborate on the challenges experienced during our case study. In Section III, we describe the overall modelling approach taken to solve these challenges, followed in Section IV by detailed descriptions of the supporting models. Reflections on applying the modelling approach on the case study are then discussed in Section V.

## II. Problem Formulation

### A. Background

One can avoid the need to integrate the information islands, by adopting a single platform (such as PTC Integrity [2] or MSR-Backbone [3]) through which product data is centrally managed. However, large organizations have specific development needs and approaches (processes, tools, workflow, in-house tools, etc.), which lead to a wide landscape of organization-specific and customized development environments. This landscape moreover needs to organically evolve over time, in order to adjust to future unpredictable needs of the industry. Contemporary platforms, however, offer limited customization capabilities to tailor for the organization-specific needs, requiring instead the organization to adjust itself to suite the platform. So, while they might be suitable at a smaller scale, such centralized platforms cannot scale to handle the complete heterogeneous set of data sources normally found in a large organization.

A more promising approach to deal with this challenge is to adopt the concepts of Linked Data to integrate the information from the different engineering tools - without relying on a centralized integration platform. To this end, OASIS OSLC [5] is an emerging interoperability open standard that adopts the architecture of the Internet to achieve massive scalability and flexibility. OASIS OSLC is based on the W3C Linked Data initiative and follows the Representational State Transfer (REST) architectural pattern. It provides for tool- and platform-neutral usage of these web technologies to create high cohesion between tools, while reducing the need for one tool to understand the deep data of another (low coupling). This lends itself well to the distributed and organic nature of the F-IDE being desired.

When developing such a federated OSLC-based F-IDE, there is however an increased risk that one loses control over the overall product data structure that is now distributed and interrelated across the many tools. This risk is particularly aggravated if one needs to maintain changes in the F-IDE over time. In this paper, we present a modelling approach to F-IDE development that tries to deal with this risk. That is, how can a distributed architecture – as promoted by the Linked Data approach - be realized, while maintaining a somewhat centralized understanding and management of the overall information model handled within the F-IDE?

### B. *Case Study Description*

Typical of many industrial organizations, the development environment at the truck manufacturer Scania consists of standard engineering tools, such as Jira and CAD drawing tools; as well as a range of propriety tools that cater for specific needs in the organization. Moreover, much product information is managed as generic content in office productivity tools, such as Microsoft Word and Excel.

As a subset of a larger case study, five propriety tools and data sources were to be integrated using OSLC:

1. *Code Repository* – A version-control system in which all software code resides, and from which parsers reconstruct the vehicle software architecture, based on an analysis of source code.

2. *Communication Specifier* – A tool that centrally defines the communication network of all vehicle architectures.

3. *ModArc* – A database that defines all hardware entities and their interfaces.

4. *Diagnostics Tool* - A tool that specifies the diagnostics functionality of all vehicle architectures.

5. *Requirements Specifier* - A propriety tool that allows for the semi-formal specification of system requirements.

As a first step, the data that needed to be communicated between the tools was analyzed. This was captured using a Class Diagram (Figure 1), as is the current state-of-practice at Scania for specifying a data model. For the purpose of this paper, it is not necessary to have full understanding of the data artefacts. It is worth highlighting that color-codes were used to define which tool managed which data artefact. Also, it is important to note that the model focuses on the data that needs to be communicated between the tools, and not necessarily all data available internally within each tool.

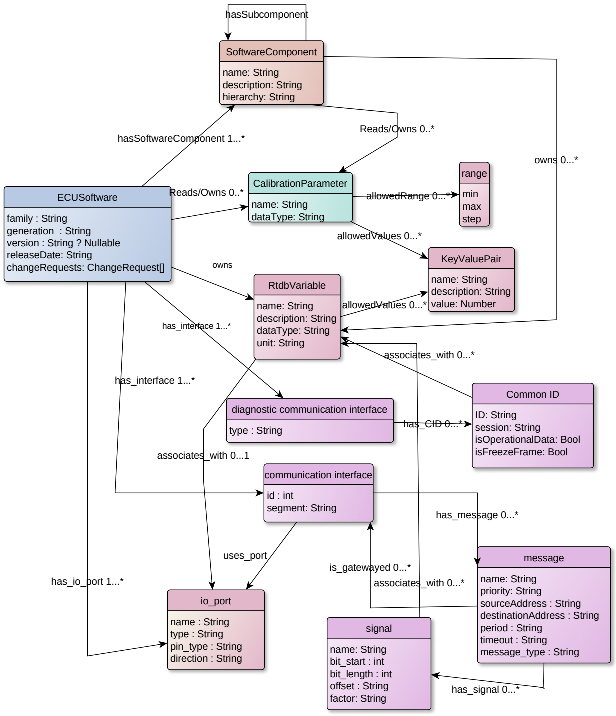


Figure 1. a UML class diagram of the resources shared in the F-IDE.

### C. *Identified Needs and Shortcomings*

In this paper, we focus on the initial stages of specifying and architecting the desired OSLC-based F-IDE. We elaborate on the needs and shortcomings experienced by an architect during these stages:

**Information specification** – there is a need to specify the information to be communicated between the tools. For pragmatic reasons, a UML class diagram was adopted to define the entities being communicated and their relationships. Clearly, the created model does not comply with the semantics of the class diagram, since the entities being models are not objects in the object-oriented paradigm, but resources according to the Resource Description Framework (RDF) graph data model. Since the information model is to be maintained over time, and is also intended for communication among developers, using a class diagram - while implying another set of semantics – may lead to misunderstandings. A specification that is semantically compatible with the intended implementation technology (of Linked Data, and specifically the OSLC standard) is necessary.

**Tool ownership** – For any given resource being shared in the F-IDE, it is necessary to clearly identify the data source (or authoring tool) that is expected to manage that resource. That is, while representations of a resource may be freely shared between the tools, changes or creations of such a resource can only occur via its owning tool. Assuming a Linked Data approach also implies that a resource is owned by a single source, to which other resources link. In practice, it is not uncommon for data to be duplicated in multiple sources, and hence mechanisms to synchronize data between tools are needed. For example, resources of type *Communication Interface* may be used in both *Communication Specifier* and *ModArc*, with no explicit decision on which of the tools defines it. To simplify the case study, we chose to ignore the *ModArc* source, but in reality one needs to synchronize between the two sources, as long as it is not possible to make one of them redundant.

**Domain ownership** – Orthogonal to tool ownership, it is also necessary to group resource definitions into domains (such as requirements engineering, software, testing, etc.). Domains can be generic in nature. Alternatively, such domain grouping can reflect the organization units that are responsible to manage specific parts of the information model. For example, the testing department may be responsible to define and maintain the testing-related resources, while the requirements department manages the definition of the requirements resources. Dependencies between the responsible departments can then be easily identified through the dependencies in the information models.

**Avoid mega-meta-modelling** – Information specifications originate from various development phases and/or development units in the organization. The resulting information models may well overlap, and would hence need to be harmonized. Hence, there is a need to harmonize the information models – while avoiding a central information model. Earlier attempts at information modeling normally resulted in large models that can easily become harder to maintain over time. The research project CESAR presents in [6] a typical interoperability approach in which such a large common meta-model is proposed. It is anticipated that the Linked Data approach would reduce the need to have such a single centralized mega information model. The correct handling of information through Domain and Tool Ownership (see above) ought to also help in that direction.

In summary, in architecting an F-IDE, there is a need to support the data specification using Linked Data semantics, while covering the two ownership aspects of tools (ownership from the tool deployment perspective) and domains (ownership from the organizational perspective).

## III. APPROACH

We take an MDE approach to F-IDE development, in which we define models that support the architect with the needs identified in the previous section. Concretely, we present a modelling tool for the graphical definition of Linked Data resource types, based on the Linked Data constraint language of Resource Shapes [15]. Resource Shapes is a mechanism to define the constraints on RDF resources, whereby a Resource Shape defines the properties that are allowed and/or required of a type of resource; as well as each property's cardinality, range, etc.

We define the model using two views: (1) domain ownership and (2) tool ownership; with each view covering the corresponding ownership needs identified in the previous subsection.

Even though our current case study focuses on the specification and architectural design phases of F-IDE development – and in particular on information specification – we aim for an approach that can be seamlessly extended to cover the complete F-IDE life cycle, and include additional integration aspects, such as control and presentation integration [7]. Towards this, we introduce a third modelling view that supports the detailed design phase of each tool interface in the F-IDE. This view definition is made compliant with an existing code generator of tool interfaces [8]. Besides being a practical feature for the developers of tool interfaces, by ensuring that the specification model (with its three views) can lead to the generation of working code, one can validate the model's completeness and correctness with respect to the Resource Shape constraints.

The Eclipse-based modelling prototype is developed based on the Ecore meta-model of the EMF [9] project. It is important to note that adopting the close-world metamodeling approach of Ecore does not necessarily contradict the open-world view of Linked Data. The information being shared across the F-IDE remains loyal to the open-world view, within the constraints specified through the Resource Shapes mechanism. Ecore is only necessary to develop the supporting tool to define these mechanisms.

## IV. THE MODEL

In this section, we present the F-IDE specification model and its three views.

**Domain Specification View** From this perspective, the architect defines the types of resources, their properties and relationships, using mechanisms compliant with the OSLC Core Specification [10] and the Resource Shape constraint language [15].

Figure 2 shows the Domain Specification diagram for the resources needed in our case study. The top-level container, *Domain Specification*, groups related *Resources* and *Resource Properties*. Such grouping can be associated with a common topic (such as requirements or test management), or reflects the structure of the organization managing the F-IDE. This view ought to support standard specifications, such as Friend of a Friend (FOAF) [11] and RDF Schema (RDFS) [12], as well as propriety ones. In Figure 2, three domain specifications are defined: *Software*, *Communication* and *Variability*, together with a subset of the standard domains of Dublin Core and RDF.

As required by the OSLC Core, a specification of a *Resource* type must provide a *name* and a *Type URI*. The *Resource* type can then also be associated with its allowed and/or required properties. These properties could belong to the same or any other *Domain Specification*. A *Resource Property* is in turn defined by specifying its cardinality, optionality, value-type, allowed-values, etc. Figure 3 illustrates an example property specification highlighting the available constraints that can be defined. A *Literal Property* is one whose value-type is set to one of the predefined literal types (such as string or integer); while a *Reference Property* is one whose value-type is set to either "resource" or "local resource". In the latter case, the *range* property can then be used to suggest the set of resource types the *Property* can refer to.

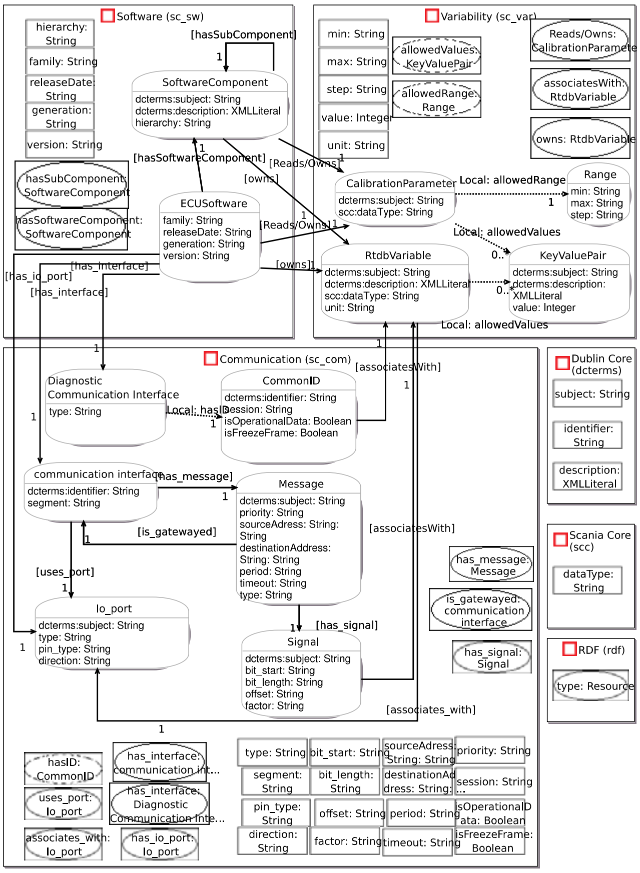
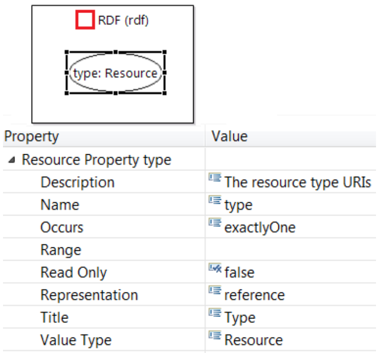


Figure 2. Domain Specification View

Borrowing from the typical notation used to represent RDF graphs, *Resource* types are represented as ellipses, while *Properties* are represented as rectangles (A *Reference Property* is represented with an ellipse within the rectangle.). In addition, *Resource Properties* are represented as first-class elements in the diagram.



| Property | Value |
|---|---|
| Resource Property type | |
| Description | The resource type URIs |
| Name | type |
| Occurs | exactlyOne |
| Range | |
| Read Only | false |
| Representation | reference |
| Title | Type |
| Value Type | Resource |

Figure 3. The specification of the rdf:type predicate, in the Domain Specification View

The association between a *Resource* type and its corresponding *Properties* is represented by arrows. However, in many cases, it becomes inconvenient to view all relationships from *Resources* to *Properties*. This is particularly the case for common *Literal Properties*, such as dcterms:subject, which can be associated to many resources across many domains. As a convenience, one can choose to hide *Resource to Literals* associations, and instead list *Literal Properties* within the *Resource* ellipse representation. It is this latter alternative that is being presented in Figure 2.

**Resource Allocation View** is where architect allocates resources to data sources. It gives the architect an overview of where the resources are available in the F-IDE, and where they are consumed. For each data source, the architect defines the set of resources it exposes; as well as those it consumes. These resources are graphically represented as *"provided"* (outwards arrows) and *"required"* (inwards arrows) ports on the edge of the Tool element, as illustrated in Figure 4. For example, the *Communication Specifier* tool exposes the *Message* resource, which is then consumed by the *Requirements Specifier* tool.

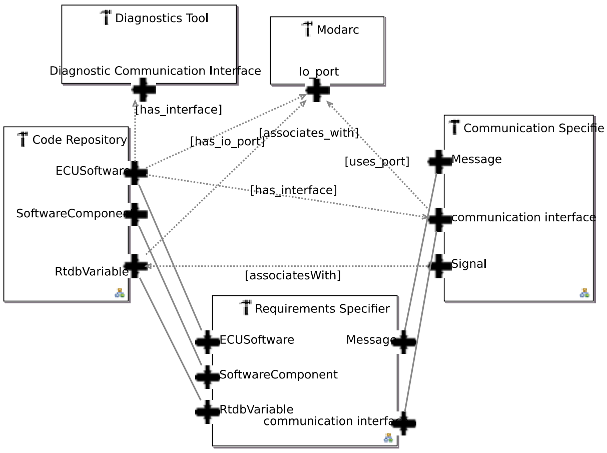


Figure 4. Resource Allocation View

In the Resource Allocation view, the interaction between a provider and consumer of a given resource is presented as a solid edge between the corresponding ports. In addition, any dependencies between resources that are managed by two different data sources are also represented in this model – as a dotted edge. For example, the resource ECUSoftware, managed by the Code Repository, has a property has_io_port that is a reference to resource IO_port; which is in turn managed through the data source Modarc. Hence, for a consumer of ECUSoftware, it is beneficial to identify the indirect dependency on the Modarc tool, since any consumption of an ECUSoftware resource, is likely to lead to the need to communicate with Modarc in order to obtain further information about the property has_io_port.

**Adapter Design View** is where the architect (or tool interface developer) designs the internal details of the tool interface – according to the OSLC standard. This can be performed for any of the *Tool* entities in the *Resource Allocation* view. Sufficient information is captured in this view, so that an almost complete interface code, which is compliant with the OSLC4J software development kit (SDK) can be generated, based on the Lyo code generator [8].

The OSLC Core Specification defines the set of resource services that can be offered by a tool. As illustrated in Figure 5, “OSLC Services are accessible via a Service Provider that describes the Services offered. Each Service can provide Creation Factories for resource creation, Query Capabilities for resource query and Delegated UI Dialogs to enable clients to create and select resources via a web UI.”[10]. The *Adaptor Design* view is a realization of the OSLC concepts in Figure 5. An example from our case study is presented in Figure 6, in which the *Core Repository* provides query capabilities and creation factories on all three resources.

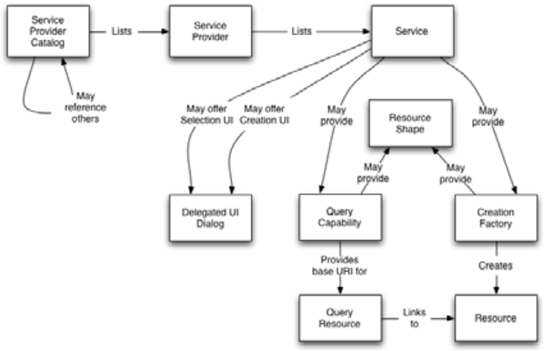


Figure 5. OSLC Core Specification concepts and relationships [10]

The *Adaptor Design* view also models its consumed resources (In Figure 6 no consumed resources are defined.). Note that the provided and required resources - as defined in this view - remain synchronized with those at the interface of the *Tool* entity in the *Resource Allocation* view.

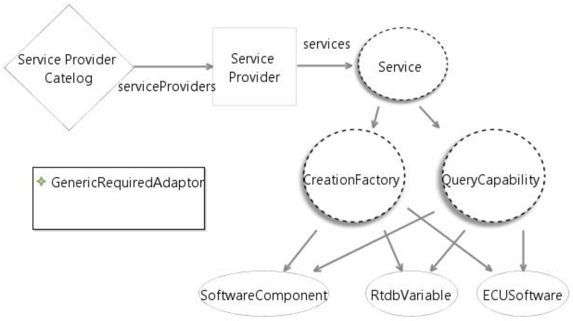


Figure 6. Adaptor Design View

There is no particular ordering of the above views, and in practice, the three views can be developed in parallel. Consistency between the views is maintained since they all refer to the same model. For example, if the architect removes a resource from the *Adaptor Design* view, the same resource is also removed from the *Resource Allocation* view.

## V. Reflections

Compared to the original approach of using a UML class diagram (See Figure 1) to represent the F-IDE resources, the proposed model may seem to add a level of complexity by distributing the model information into three views. However, upon further investigation, it becomes clear that the class diagram was actually used to superimpose information for both the *Domain Specification* and *Resource Allocation* views into the same diagram. For example, classes were initially color-coded to classify them according to their owning tool. However, the semantics and intentions behind this classification soon become ambiguous, since the distinction between tool and domain ownership is not identified explicitly. In the original approach, different viewers of the same model could hence draw different conclusions when analyzing the model, depending on their implicit understanding of the color codes.

Moreover, the usage of a class diagram is not compatible with the open-world view of Linked Data, nor is it suitable to specify all necessary information according to the OSLC standard. This became apparent when the detailed specification and design of a tool’s interface needed to be defined. No complements to the UML class diagram can provide such support. Instead, a dedicated domain-specific language (DSL) that follows the expected semantics can be better used uniformly across the whole organization. We here illustrate two examples where our DSL helped communicate the correct semantics, which were previously misinterpreted or not used:

- A *Resource Property* is a first-class element that can be associated with multiple *Resource* types. For example, the same allowedValues property (with range *KeyValuePair*) is a property used for both the *CalibrationParameter* & *RtdbVariable* resources. Previously, two independent properties were unnecessarily defined.
- Certain resources (such as Range) can only exist within the context of another parent resource, and hence ought not to have their own URI. Our DSL helped communicate the capability of defining *Local Resources*.

By breaking the model into two views, and by structuring each view along the managing domains and tools respectively, the information model is not expected to be developed in a top-down and centralized manner. Instead, a more distributed process is envisaged, in which resources are defined within a specific domain and/or tool. Only when necessary, such sub-models can then be integrated, avoiding the need to manage a single centralized information model.

Finally, the two orthogonal views of the F-IDE allow the architect to identify dependencies within the F-IDE, form both the organizational as well as the deployment perspective:

- In the Resource Allocation View of the model, the architect can obtain an overview of the coupling/cohesion of the tools of the F-IDE. One could directly identify the direct producer/consumer relations, as well as the indirect dependencies, as detailed in Section IV.
- In the Domain Specification view, the architect views the dependencies between the different domains (irrespective of how the resources are deployed across tools). Such dependencies reveal the relationship between the organizational entities involved in maintaining the overall information model. This explicit modelling of domain

ownership helps lift important organizational decisions, which otherwise remain implicit.

While the need for a dedicated DSL is convincing, the proposed views are not necessarily final, and there remains room for improvements. For example, while the possibility to represent *Properties* as first-class elements was appreciated, it was experienced that they (The squares in Figure 2) cluttered the overall model, and did not make efficient usage of the available modeling space. Similarly, the relationships between *Resources* and their associated *Literal Properties* (not shown in Figure 2) cluttered the model. Currently, filtering mechanisms are available to support different representations that suite different users, while maintaining the same underlying model. In Figure 2, the filter that hides the arrows representing relationships between *Resources* and *Literal Properties* is activated. However, the filter that hides all *Property* elements is not activated. This makes the view seem almost similar – visually - to the class diagram of Figure 1, yet the more appropriate Linked Data semantics lie behind this view.

## VI. Related Work

There exists a large body of research that in various ways touches upon information modeling and model integration. (See for example [13] and [14]). Our work - and the related work of this section - is delimited to the Linked Data paradigm. The work in this paper builds upon the Resource Shape constraint language suggested in [15], by providing a graphical model to specify such constraints on RDF resources.

The most relevant work found in this area is the Ontology Definition Metamodel (ODM) [16]. ODM is an OMG specification that defines a family of Meta-Object Facility (MOF) metamodels for the modelling of ontologies. ODM also specifies a UML Profile for RDFS [12] and the Web Ontology Language (OWL) [17], which can be realized by UML-based tools, such as Enterprise Architect's ODM diagrams [18]. However, as argued in [15], OWL and RDFS are not suitable candidates to define and validate constraints, given that they are designed for another purpose - namely for reasoning engines that can infer new knowledge.

Earlier work by the authors has also resulted in a modelling approach to tool-chain development [19]. In this earlier work, even though the information was modelled targeting an OSLC implementation, the models were directly embedded in the specific tool adaptors, and no overall information model is readily available. The models did not support the tool and domain ownership perspectives identified in this paper.

## VII. Conclusion

In this paper, an MDE approach to F-IDE development based on the Linked Data principles is presented. A prototype modelling tool has been developed that allows for the modelling of the information model for a complete F-IDE, based on the Resource Shapes constraint language [15]. The model is defined through three views focusing on the specification and design stages of F-IDE development. It is envisaged however that the modelling support will be extended to cover the complete development life-cycle, specifically supporting the requirements analysis phase, as well as automated testing. The current focus on data integration needs to be also extended to cover others other aspects of integration, in particular control integration [7].

The Eclipse-based prototype is to be released as an open-source contribution, yet this has not been done at the time of writing this article.

## References

[1] Road vehicles - functional safety, ISO standard 26262:2011, 2011.

[2] (2015, Dec.) PTC Integrity. [Online]. Available: http://www.mks.com/platform/

[3] B. Weichel, and M. Herrmann, "A backbone in automotive software development based on XML and ASAM/MSR.", SAE Technical Papers, 2004, doi:10.4271/2004-01-0295.

[4] T. Berners-Lee. (2015, Dec.) Linked data design issues. [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.html

[5] (2015, Dec.) OASIS OSLC. [Online]. Available: http://www.oasis-oslc.org/

[6] A. Rossignol, "The reference technology platform" in CESAR - Cost-efficient methods and processes for safety-relevant embedded systems, A. Rajan and T. Wahl, Eds. Dordrecht: Springer, pp. 213-236, 2012.

[7] A. I. Wasserman, "Tool integration in software engineering environments", the international workshop on environments on Software engineering environments, 1990, pp. 137-149.

[8] (2015, Dec.) Eclipse Lyo Code Generator. [Online]. Available: http://wiki.eclipse.org/Lyo/AdaptorCodeGeneratorWorkshop

[9] (2015, Dec.) Eclipse EMF. [Online]. Available: https://eclipse.org/modeling/emf/

[10] OSLC Core Specification, OSLC standard v2.0, 2013.

[11] (2015, Dec.) FOAF Vocabulary Specification. [Online]. Available: http://xmlns.com/foaf/spec/

[12] RDF Schema 1.1, W3C Recommendation, 2014.

[13] M. Törngren, A. Qamar, M. Biehl, F. Loiret, and J. El-khoury, "Integrating viewpoints in the development of mechatronic products.", Mechatronics (Oxford), vol. 24, nr. 7, 2014, pp. 745-762.

[14] R. Basole, A. Qamar, H. Park, C. Paredis, and L. Mcginnis, "Visual analytics for early-phase complex engineered system design support.", IEEE Computer Graphics and Applications, vol. 35, nr. 2, 2015, pp. 41-51.

[15] A. G. Ryman, A. Le Hors, and S. Speicher, "OSLC resource shape: A language for defining constraints on linked data.", CEUR Workshop Proceedings, Vol.996, 2013.

[16] Ontology Definition Metamodel, OMG standard, document number: formal/2014-09-02, 2014.

[17] OWL 2 Web Ontology Language, W3C Recommendation, 2012.

[18] (2015, Dec.) Enterprise Architect ODM MDG Technology. [Online]. Available: http://www.sparxsystems.com/enterprise_architect_user_guide/9.3/domain_based_models/mdg_technology_for_odm.html

[19] M. Biehl, J. El-khoury, F. Loiret, and M. Törngren, "On the modeling and generation of service-oriented tool chains.", Software & Systems Modeling, vol. 13, nr 2, 2014, pp. 461-480.

# Reasoning with Place Information on the Linked Data Web

Alia I Abdelmoty, Khalid M. Al-Muzaini

Cardiff School of Computer Science & Informatics
Cardiff University
Wales, UK
Email: {A.I.Abdelmoty, Almuzainiko}@cs.cf.ac.uk

***Abstract*—The Linked Data Web (LDW) is an evolution of the traditional Web from a global information space of linked documents to one where both documents and data are linked. A significant amount of geographic information about places are currently being published on this LDW. These are used to qualify the location of other types of datasets. This paper examines the limitations in the nature of location representation in some typical examples of these Resource Description Framework (RDF) resources, primarily resulting from the simplified geometric representation of location and the incomplete and random use of spatial relationships to link place information. The paper proposes a qualitative model of place location that enforces an ordered representation of relative spatial relationships between places. The model facilitates the application of qualitative spatial reasoning on places to extract a potentially large percentage of implicit links between place resources, thus allowing place information to be linked and to be explored more fully and more consistently than what is currently possible. The paper describes the model and presents experimental results demonstrating the effectiveness of the model on realistic examples of geospatial RDF resources.**

***Keywords–qualitative place models; spatial reasoning; geospatial web.***

## I. INTRODUCTION

One of the 'Linked Data Principles'[1] is to include links to connect the data to allow the discovery of related things. However, identifying links between data items remains a considerable challenge that needs to be addressed [2]. A key research task in this respect is identity resolution, i.e., to recognise when two things denoted by two URIs are the same and when they are not. Automatic linking can easily create inadequate links, and manual linking is often too time consuming [3]. Geo-referencing data on the LDW can address this problem [4], whereby links can be inferred between data items by tracing their spatial (and temporal) footprints. For example, the BBC uses RDF place gazetteers as an anchor to relate information on weather, travel and local news [5].

Yet, for geospatial linked data to serve its purpose, links within and amongst the geographic RDF resources need themselves to be resolved. That is to allow place resources to be uniquely identified and thus a place description in one dataset can be matched to another describing the same place in a different dataset. A scheme that allows such links between place resources to be discovered would be a valuable step towards the realisation of the LDW as a whole.

In this paper, location is used as a key identifier for place resources and the question to be addressed is how location can be used to define a *linked* place model that is sufficient to enable place resources to be uniquely identified on the LDW. Several challenges need to be addressed, namely, 1) location representation of RDF place resources is simple; defined as point coordinates in some resources, detailed; defined with extended geometries in others, and sometimes missing all together, 2) coordinates of locations may not match exactly across data sources, where volunteered data mapped by individuals is mashed up with authoritative map datasets, 3) non-standardised vocabularies for expressing relative location is used in most datasets, e.g., in DBpedia, properties such as *dbp:location*, *dbp-ont:region* and *dbp-ont:principalarea* are used to indicate that the subject place lies inside the object place.

Towards addressing this problem, a linked place model is proposed that uses qualitative spatial relationships to describe unique place location profiles. The profiles don't rely on the provision of exact geometries and hence can be used homogeneously with different types of place resources. They can be expressed as RDF statements and can thus be integrated directly with the resource descriptions. The rationale behind the choice of links to be modelled is primarily twofold: to allow for a sensible unique description of place location and to support qualitative spatial reasoning over place resources. The value of the linked place model is illustrated by measuring its ability to make the underlying RDF graph of geographic place resources browsable. Samples of realistic geographic linked datasets are used in the experiments presented and results demonstrate significant potential value of the methods proposed.

The paper is structured as follows. An overview of related work on the representation and manipulation of place resources on the LDW is given in section II, In section III the proposed relative location model is presented and in section IV, its application on two different realistic datasets is evaluated. Conclusions and an overview of future work is given in section V.

## II. RELATED WORK

Here related work on the topics of representing place resources and reasoning with them on the LDW are reviewed.

### A. Representing RDF place Resources on the LDW

Sources of geographic data on the LDW are either volunteered (crowdsourced) resources, henceforth denoted Volunteered Geographic Information (VGI), created by individuals with only informal procedures for validating the content, or authoritative resources produced by mapping organizations, henceforth denoted Authoritative Geographic Information (AGI). Example of VGIs are DBpedia (dbpedia.org), GeoNames (geonames.org), and OpenSreetMaps (linkedgeodata.org)[6] and examples of AGIs are the Ordnance Survey linked data [7] and the Spanish linked data [8].

The volume of VGI resources is increasing steadily, providing a wealth of information on geographic places and creating detailed maps of the world. DBpedia contains hundreds of thousands of place entities, whose locations are represented as point geometry. GeoNames is a gazetteer that collects both spatial and thematic information for various place names around the world. In both datasets, place location is represented by a single point coordinates. While DBpedia does not enforce any constraints on the definition of place location (e.g., coordinates may be missing in place resources), reference to some relative spatial relationships, and in particular to represent containment within a geographic region, is normally maintained. GeoNames places are also interlinked with each other by defining associated parent places.

In [9], the LinkedGeoData effort is described where OSM data is transformed into RDF and made available on the Web. OSM data is represented with a relatively simple data model that captures the underlying geometry of the features. It comprises three basic types, nodes (representing points on Earth and have longitude and latitude values), ways (ordered sequences of nodes that form a polyline or a polygon) and relations (groupings of multiple nodes and/or ways). Furthermore, [10] presented methods to determine links between map features in OSM and equivalent instances documented in DBpedia, as well as between OSM and Geonames. Their matching is based on a combination of the Jaro-Winkler string distance between the text of the respective place names and the geographic distance between the entities. Example of other work on linking geodata on the Semantic Web is [11], which employs the Hausdorff distance to establish similarity between spatially extensive linear or polygonal features.

In contrast to VGI resources that manages geographic resource as points (represented by a coordinate of latitude and longitude), AGI resources deal with more complex geometries as well, such as line strings. AGIs tend to utilise well-defined standards and ontologies for representing geographic features and geometries. Ordnance Survey linked data also demonstrates the use of qualitative spatial relations to describe spatial relationships in its datasets. Two ontologies, the Geometry Ontology and the Spatial Relations Ontology, are used to provide geospatial vocabulary. These ontologies describe abstract geometries and topological relations (equivalent to RCC8 [12]) respectively.

In summary, the spatial representation of place resources in VGI datasets is generally limited to point representation, and is managed within simple ontologies that encode non-spatial semantics and in some cases limited spatial relationships. On the other hand, place data provided as AGI tend to present more structured and detailed spatial representations, but is also limited to specific types and scales of representation. Use of some qualitative spatial relationships has been demonstrated for capturing the spatial structure in some example datasets. The model proposed in this paper offers a systematic and homogenous representation of place location that can be consistently applied to VGIs or AGIs and demonstrates the value of heterogenous qualitative spatial relations in representing place information on the LDW.

### *B. Manipulating and Querying RDF place resources on the LDW*

Recently, much work has been done on extending RDF for representing geospatial information through defining and utilising appropriate vocabularies encoded in ontologies to represent space and time. The work capitalises on specification of standards, defined by the Open Geospatial Consortium (OGC)(opengeospatial.org), for modeling core concepts related to geospatial data. Prominent examples are GeoSPARQL, an OGC standard [13] and stRDF/stSPARQL [14]. Both proposals provide vocabulary (classes, properties, and functions) that can be used in RDF graphs and SPARQL queries to represent and query geospatial data, for example *geo:SpatialObject*, which has as instances everything that can have a spatial representation and *geo:Geometry* as the superclass of all geometry classes. In addition, geometric functions and topological functions are offered for performing computations, such as *geof:distance* and for asserting topological relations between spatial objects, e.g., *dbpedia:Cardiff geo:sfWithin dbpedia:Wales*.

Qualitative spatial representation and reasoning (QSRR) are established areas of research [15], whose results have influenced the definition of models of spatial relationships in international standards, e.g., the OGC models, and commercial spatial database systems (for example, in the Oracle DB system). $RCC8$, a QSRR model, has been recently adopted by GeoSPARQL [13], and there is an ever increasing interest in coupling QSR techniques with Linked Geospatial Data that are constantly being made available [14]. On the other hand, Semantic Web reasoning engines have been extended to support qualitative spatial relations, e.g., Racerpro [16] and PelletSpatial [17]. Scalability of the spatial reasoning is recognised and reported challenge. Scalable implementations of constraint network algorithms for qualitative and quantitative spatial constraints are needed, as RDF stores supporting Linked Geospatial Data are expected to scale to billions of triples [14]. Lately, promising results have been reported by [18], who proposed an approach for removing redundancy in RCC8 networks and by [19], who examined graph-partitioning techniques as a method for coping with large networks; in both cases leading to more effective application of spatial reasoning mechanisms. Finally, qualitative methods were used to complement existing quantitative methods for representing the geometry of spatial locations. In [20], heterogenous reasoning methods are proposed that combine calls between a spatial database system and a spatial reasoning engine implemented in *OWL2 RL* to check the consistency of place ontologies. In [21], Younis et al described query plans that make use of a combination of qualitative spatial relationships associated with place resources in DBpedia and detailed representations of geometry maintained in a spatially indexed database for answering complex queries. In both cases, qualitative reasoning was limited by the fragmented and scarce availability of spatial relationships to work on. The qualitative scheme of representation of place location proposed in this paper addresses this issue and provides a novel method for defining spatial relationships that is designed to support and facilitate the effective use of qualitative spatial reasoning on the LDW.

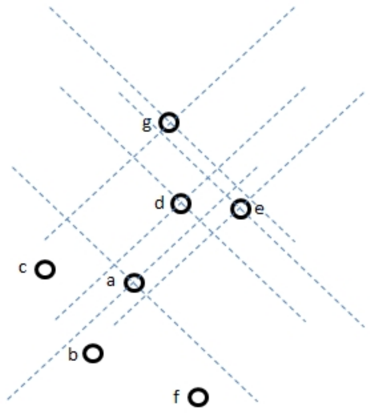
## III. A Linked Place Model for the Linked Data Web

A Relative Location model (*RelLoc*) is proposed here to capture a qualitative representation of the spatial structure of place location. Two types of spatial relations are used as follows.

1) Containment relationships, to record that a parent place directly contains a child place; i.e., one step hierarchy. For example, for three places representing a district, a city and a country, the model will explicitly record the relationships: inside(district, city) and inside(city, country), but not inside(district, country).
2) Direction-proximity relationships, to record for every place the relative direction location of its nearest neighbour places. The direction frame of reference can be selected as appropriate. For example, for a 4-cardinal direction frame of reference, a place will record its relative direction relation with its nearest neighbour in four directions.

For a given set of places $Pl$, let $DirPr$ be the set of all direction-proximity relations between instances of places in $Pl$ as defined above, and let $Con$ be the set of containment relations between instances of places in $Pl$ as defined above. Then, $RelLoc(Pl)$ is defined as a tuple $RelLoc(Pl) := (Pl, D, C)$, where: $D \in DirPr$ and $C \in Con$. $R_{nn}(x, y)$ is used to denote that $x$ is the nearest neighbour from the direction $R$ to object $y$. For example, $N_{nn}(pl_1, pl_2)$ indicates that $pl_1$ is the nearest neighbour from the north direction to $pl_2$, etc.

To illustrate the model, consider the scene in Figure 1 that consists of a set of places, $a$ to $f$, with a 4-cardinal direction frame of reference overlaid for some places in the scenes. A representative point is used to define the place location. It is further known that places represented as points $a, b, c, e$ are inside $d$ and places $d, f$ are inside $g$. The full set of relationships used to model the scene are given in the table in Figure 1(b). Note that in some cases, no relation can be found, e.g., there are no neighbours for object $c$ from the west direction in Figure 1(a).



(a)

| Set of spatial relations to model relative location |
|---|
| $N_{nn}(d,a)$, $S_{nn}(b,a)$, $W_{nn}(c,a)$, $E_{nn}(e,a)$ |
| $N_{nn}(g,d)$, $S_{nn}(a,d)$, $W_{nn}(c,d)$, $E_{nn}(e,d)$ |
| $N_{nn}(g,c)$, $S_{nn}(b,c)$, $E_{nn}(a,c)$ |
| $N_{nn}(a,b)$, $E_{nn}(f,b)$ |
| $N_{nn}(a,f)$, $W_{nn}(b,f)$ |
| $N_{nn}(g,e)$, $S_{nn}(b,e)$, $W_{nn}(d,e)$ |
| $S_{nn}(d,g)$ |
| $in(a,d)$, $in(b,d)$, $in(c,d)$, $in(e,d)$, |
| $in(d,g)$, $in(f,g)$ |

(b)

Figure 1. (a) An example map scene with a set of places represented as points.(b) Set of direction, proximity and containment relations chosen to representative relative location in the proposed model.

### A. Spatial Reasoning with the Relative Location Model

We can reason over the relative location model to infer more of the implicit spatial structure of place location. Qualitative spatial reasoning (QSR) tools can be utilised to propagate the defined relationships and derive new ones between places in the scene. QSR takes advantage of the transitive nature of the partial or total ordering of the quantity space in order to infer new information from the raw information presented. In particular, the transitive nature of some spatial relationships can be used to directly infer spatial hierarchies, for example, containment and cardinal direction relations. The scope of the model is deliberately focussed on general containment relationships and ignores other possible topological relations, such as overlap or touch. Hence, building containment hierarchies is straightforward using the transitivity rules: $inside(a, b) \wedge inside(b, c) \rightarrow inside(a, c)$ and $contains(a, b) \wedge contains(b, c) \rightarrow contains(a, c)$.

In the case of direction relationships, more detailed spatial reasoning can be applied using composition tables. Table I shows the composition table for a 4-cardinal direction frame of reference between point representations of spatial objects.

TABLE I. COMPOSITION TABLE FOR 4-CARDINAL DIRECTION RELATIONSHIPS.

| | $N$ | $E$ | $S$ | $W$ |
|---|---|---|---|---|
| $N$ | $N$ | $N \vee E$ | $All$ | $N \vee W$ |
| $E$ | $N \vee E$ | $E$ | $S \vee E$ | $All$ |
| $S$ | $All$ | $S \vee E$ | $S$ | $S \vee W$ |
| $W$ | $W \vee N$ | $All$ | $W \vee S$ | $W$ |

In considering the entries of the composition tables, some of those entries provide definite conclusions of the composition operation, i.e., the composition result is only one relationship (emboldened in table), other entries are indefinite and result in a disjunctive set of possible relationships, e.g., the composition: $N(a, b) \wedge E(b, c) \rightarrow N(a, c) \vee E(a, c)$.

Spatial reasoning can be applied on the linked place model using different strategies. The most straightforward is through deriving the algebraic closure, i.e., completing the scene by deriving all possible missing relationships between objects. Path-consistency algorithms for deriving the algebraic closure has been been implemented in various tools, e.g., in the SparQ spatial reasoning engine [22]. Table II shows the result of this operation for the example scene in Figure 1. Explicit relations are shown in bold and the remaining relation are inferred by spatial reasoning. As can be seen in the table, using the 19 relationships defined for the model in Figure 1(b), reasoning was able to derive a further 19 definite relationships, completing over 90% of the possible relations in the scene.

TABLE II. RESULT OF REASONING WITH CARDINAL RELATIONS FOR THE PLACE MODEL IN FIGURE 1.

| | $a$ | $b$ | $c$ | $d$ | $e$ | $f$ | $g$ |
|---|---|---|---|---|---|---|---|
| $a$ | - | **N** | **E** | **S** | $W$ | **N** | $S$ |
| $b$ | **S** | - | **S** | $S$ | **S** | **W** | $S$ |
| $c$ | **W** | $W$ | - | **W** | $W$ | $N \vee W$ | $S \vee W$ |
| $d$ | **N** | $N$ | $E$ | - | **W** | $N$ | **S** |
| $e$ | **E** | $N$ | $E$ | **E** | - | $N$ | $S$ |
| $f$ | $S$ | **E** | $S \vee E$ | $S$ | $S$ | - | $S$ |
| $g$ | $N$ | $N$ | **N** | **N** | **N** | $N$ | - |

## B. *Applying the Relative Location Place Model on the LDW*

The underlying structure of any expression in RDF is a collection of triples, each consisting of a subject, a predicate and an object. A set of such triples is called an RDF graph, in which each triple is represented as a node-arc-node link and each triple represents a statement of a relationship between the subjects and objects, denoted by the nodes, that it links. The meaning of an RDF graph is the conjunction (logical AND) of the statements corresponding to all the triples it contains.

The *RelLoc* place model can be interpreted as a simple connected graph with nodes representing place resources and edges representing the spatial relationships between places. Thus a realisation of the place model for a specific RDF document of place resources is a subgraph of the RDF graph of the document. The *RelLoc* RDF graph is completely defined if RDF statements are used to represent all spatial relationships defined in the model, e.g., for the scene in Figure 1, 25 RDF statements are needed to encode the cardinal (19) and containment (6) relationships in the table in Figure 1(b).

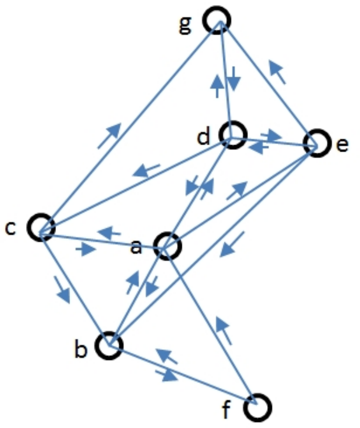
Let $Pl$ be a finite set of place class resources defined in an RDF data store and $DirPr(Pl)$ defines cardinal direction relations between members of $Pl$ and $Con(Pl)$ describes the containment relations between members of $Pl$ as defined by the relative location model above.

A *RelLoc* subgraph $\mathbb{G}_{\mathbb{L}} = (V_{\mathbb{L}}, E_{\mathbb{L}})$ is a simple connected graph that models $Pl$, where: $V_{\mathbb{L}} = Pl$ is the set of nodes, $E_{\mathbb{L}} = \{DirPr(Pl) \cup Con(Pl)\}$ is the set of edges labelled with the corresponding direction and containment relationships.

Note that there exists a subgraph of $\mathbb{G}_{\mathbb{L}}$ for every place $pl \in Pl$, which represents the subset of direction-proximity and containment relationships that completely define the relative location of $pl$. Thus, a *location profile* for a particular place $pl \in Pl$ can be defined as $\mathbb{L}_{pl} = \{(DirPr_{pl}, Con_{pl}\}$. $\mathbb{L}_{pl}$ is the restriction of $\mathbb{L}$ to $pl$, where $DirPr_{pl}$ and $Con_{pl}$ defines direction proximity and containment relations respectively between $pl$ and other places in $Pl$, as specified by our model.

For example the location profile for place $a$ in Figure 1 is the set of statements describing the relations: $N(d, a), S(b, a), W(c, a), E(e, a), in(a, d)$.

The *RelLoc* graph can be represented by a matrix to register the adjacency relationship between the place and its nearest neighbours. The scene in Figure 1 is shown as a graph with nodes and edges in Figure 2(a) and its corresponding adjacency matrix is shown in (b). The fact that two places are neighbours is represented by a value (1) in the matrix and by a value (0) otherwise. Values of (1) in the matrix can be replaced by the relative orientation relationship between the corresponding places as shown in Figure 2(c) and the resulting structure is denoted *Adjacency-Orientation Matrix*.



(a)

| | $a$ | $b$ | $c$ | $d$ | $e$ | $f$ | $g$ |
|---|---|---|---|---|---|---|---|
| $a$ | - | 1 | 1 | 1 | 1 | 0 | 0 |
| $b$ | 1 | - | 0 | 0 | 0 | 1 | 0 |
| $c$ | 1 | 1 | - | 0 | 0 | 0 | 1 |
| $d$ | 1 | 0 | 1 | - | 1 | 0 | 1 |
| $e$ | 0 | 1 | 0 | 1 | - | 0 | 1 |
| $f$ | 1 | 1 | 0 | 0 | 0 | - | 0 |
| $g$ | 0 | 0 | 0 | 1 | 0 | 0 | - |

(b)

| | $a$ | $b$ | $c$ | $d$ | $e$ | $f$ | $g$ |
|---|---|---|---|---|---|---|---|
| $a$ | - | $N$ | $E$ | $S$ | 0 | $N$ | 0 |
| $b$ | $S$ | - | $S$ | 0 | $S$ | $W$ | 0 |
| $c$ | $W$ | 0 | - | $W$ | 0 | 0 | 0 |
| $d$ | $N$ | 0 | 0 | - | $W$ | 0 | $N$ |
| $e$ | $E$ | 0 | 0 | $E$ | - | 0 | 0 |
| $f$ | 0 | $E$ | 0 | 0 | 0 | - | 0 |
| $g$ | 0 | 0 | $N$ | $N$ | $N$ | 0 | - |

(c)

Figure 2. (a) A graph representing the sample map scene from Figure 1. (b) Adjacency matrix for the location graph representing nearest neighbour relationships. (c) Adjacency-orientation matrix representing nearest neighbour and direction relationships.

## IV. Application and Evaluation

The main goals of the Linked Place model is to provide a representation of place location on the LDW that allows for place information to be linked effectively and consistently. The effectiveness of the proposed model can be evaluated with respect to two main aspects; whether it provides a sound definition of place location, that is to test the correctness of the place location profiles, and whether it provides a complete definition of place location, that is whether a *complete* relative location graph can be derived using the individual place location profiles.

The soundness of the location profiles is assumed as it essentially relies on the validity of the computation of the spatial relationships. Issues related to the complexity of this process are discussed in the next section.

Here, we evaluate the completeness aspect of the model. An individual place location profile defined using the model

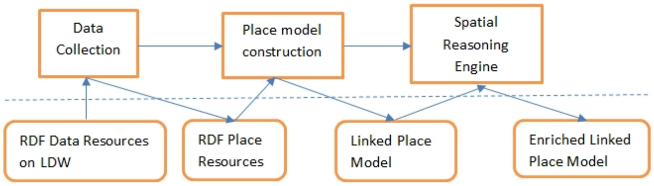


Figure 3. Components of the developed system to implement the linked place model.

```
prefix d: <http://dbpedia.org/ontology/>
prefix :<http://dbpedia.org/resource/>
prefix prop: <http://dbpedia.org/property/>
prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>

select ?place  (MAX(?lat) as ?lat)(MAX(?long) as ?long)
where{
?place ?ontology ?resource.
?place a d:Place.
?place geo:lat ?lat.
?place geo:long ?long.
filter ( ?resource = :Wales or ?resource = "Wales"@en )
}
group by ?place
order by ?place
```

Figure 4. SparQL query used to extract place data from DBpedia.

represents a finite set of spatial relationships between a place and its nearest neighbours and direct parent. Completeness of the model can be defined as the degree to which these individual profiles can be used to derive implicit links between places not defined by the model. The model is entirely complete if a full set of links between places can be derived using automatic spatial reasoning, i.e., the model can produce a complete graph, where there is a defined spatial relationship between every place in the dataset and every other place.

A system was developed that implements the Linked Place model and further builds an enriched model using spatial reasoning for evaluation purposes as shown in Figure 3. Two datasets were used in this experiment, DBPedia and the Ordnance Survey open data [7]. These were chosen as they exhibit different representations of place resources on the LDW and are typical of VGIs and AGIs respectively. A description of the datasets used is presented below, along with the results of the application of spatial reasoning over the constructed linked place models.

### DBpedia DataSet

A sample dataset containing all Places in Wales, UK, has been downloaded from DBpedia using the sparQL query in Figure 4.

A total of 489 places were used, for which a relative location graph of 2751 direction-proximity relations was constructed. Completing the graph resulted in 116403 total number of relations, out of which 50340 relations are definite (defining only one possible relationship).

Note that of the indefinite relationships some are a disjunction of 2 relations, e.g., $\{N, NW\}$ or $\{E, SE\}$ and some are a disjunction of 3 relations, e.g., $\{N, NE, NW\}$ or $\{NE, E, SE\}$. In both cases, relations can be generalised to a "coarser" direction relation, for example, $\{NE, E, SE\}$ can be generalised to general $East$ relationship. These results are considered useful and thus are filtered out in the presentation.

TABLE III. RESULTS OF REASONING APPLIED ON THE DBPEDIA DATASET.

| Defined | Definite | 2-Relations | 3-Relations | Others |
|---|---|---|---|---|
| 2751 | 50340 | 63148 | 28 | 136 |
| 2.36% | 43.24% | 54.22% | 0.02% | 0.12% |

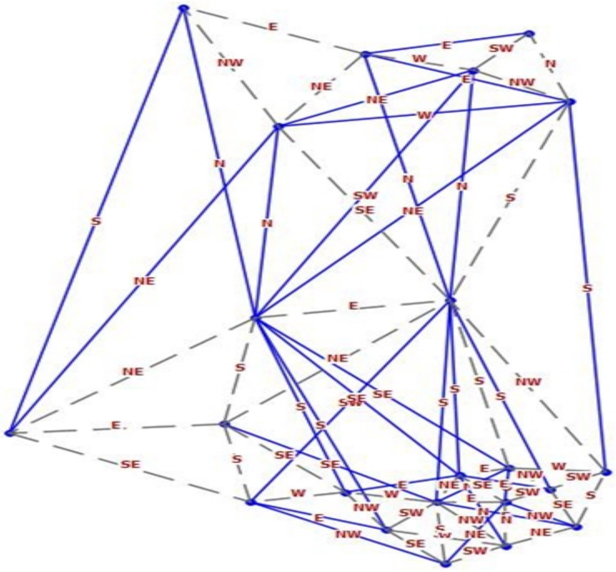


Figure 5. (a) Linked Place Graph for the Unitary Authorities in Wales from the Ordnance Survey dataset.

The remaining results are disjunctions of unrelated directions, e.g., $\{N, NE, E\}$, and are thus considered to be ambiguous. A summary of the results is shown in table III. Using the Linked Place model we are able to describe nearly half the possible relations precisely $(45.6\%)$, as well as almost all of the rest of the scene $(54.22\%)$ with some useful generalised direction relations.

### Ordnance Survey DataSet

The Boundary-line RDF dataset for Wales was downloaded from the Ordnance Survey open data web site [7]. The data gives a range of local government administrative and electoral boundaries.

Figure 5 shows the relative location graph constructed for the Unitary Authority dataset for Wales. Dashed edges are used to indicate that relationships (and inverses) are defined both ways between the respective nodes, but only one relation is used to label the edge in the Linked Place model. The set contains 22 regions, for which 73 direction-proximity relations were computed. Reasoning applied on this set of relations produces the results shown in Table IV.

We can use the above results to describe the effectiveness

TABLE IV. RESULTS OF REASONING APPLIED ON THE ORDNANCE SURVEY DATASET.

| Defined | Definite | 2-Relations | 3-Relations | Others |
|---|---|---|---|---|
| 73 | 94 | 64 | 0 | 0 |
| 31.6% | 40.69% | 27.7% | 0 | 0 |

TABLE V. SUMMARY OF THE EXPERIMENT RESULTS.

| | $\frac{Defined}{Definite}$ | $\frac{Defined}{Useful}$ |
|---|---|---|
| $DBpedia$ | 0.054 | 0.024 |
| $OS$ | 0.78 | 0.32 |

of the linked place model in terms of the information content it was able to deduce using the ratio of the number of defined relations to the number of deduced relations. A summary is presented in table V.

## V. Discussion and Conclusions

Data on geographic places are considered to be very useful on the LDW. Individuals and organisations are volunteering data to build global base maps enriched with different types of traditional and non-traditional semantics reflecting people's views of geographic space and place. In addition, geographic references to place can be used to link different types of datasets, thus enhancing the utility of these datasets on the LDW. This work explores the challenges introduced when representing place data using the simple model of RDF, with different geometries to represent location and different non-standardised vocabularies to represent spatial relationships between locations.

A linked place model is proposed that injects certain types of spatial semantics into the RDF graph underlying the place data. Specific types of spatial relationships between place nodes are added to the graph to allow the creation of individual place location profiles that fully describe the relative spatial location of a place. It is further shown how the enriched relative location graph can allow spatial reasoning to be applied to derive implicit spatial links to produce even more richer place descriptions.

The results obtained from the initial evaluation experiments demonstrate possible significant value in the proposed model. Further work need to be done to explore the potential utility of the proposal. Some of the interesting issues that we aim to explore in the future are described below.

- Simple methods and assumptions were used to compute the direction relationships between places. Further study need to be carried out to evaluate whether more involved representations are useful.
- No distinction between the types of place nodes are made when creating the graph. Can place semantics be utilised to guide this process further?
- Applications of spatial reasoning need to be considered further. Describing the complete graph is not a practical (nor a useful) option. Can spatial reasoning be selectively applied, for example, as part of query processing on the location graph.
- Further evaluation is required to understand the scalability of the proposals to much larger RDF triple stores.

## References

[1] T. Berners-Lee. Linked data. http://www.w3.org/DesignIssues/LinkedData.html. Accessed: 2009-12-15.

[2] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data - the story so far," *International Journal on Semantic Web and Information Systems*, vol. 5, no. 3, pp. 1–22, 2009.

[3] J. Goodwin, C. Dolbear, and G. Hart, "Geographical linked data: The administrative geography of great britain on the semantic web," *Transactions in GIS*, vol. 12, pp. 19–30, 2008.

[4] G. Hart and C. Dolbear, *Linked Data: A Geographic Perspective*. CRC Books, 2013.

[5] C. Henden, P. Rissen, and S. Angeletou. Linked geospatial data, and the bbc. http://www.w3.org/2014/03/lgd/papers/lgd14_submission_28. Accessed: 2009-12-15.

[6] M. Hackley, "How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets," *Environment and Planning B*, vol. 37, no. 4, pp. 682–703, 2010.

[7] Ordnance survey linked data platform. http://data.ordnancesurvey.co.uk/. Accessed: 2009-12-15.

[8] Geolinkeddata.es. http://geo.linkeddata.es/. Accessed: 2009-12-15.

[9] S. Auer, J. Lehmann, and S. Hellmann, "Linkedgeodata- adding a spatial dimension to the web of data," in *Proc. of 7th International Semantic Web Conference (ISWC)*, vol. LNCS 5823. Springer, 2009, pp. 731–746.

[10] C. Stadler, J. Lehmann, K. Hffner, and S. Auer, "Linkedgeodata: A core for a web of spatial open data," *Semantic Web Journal*, vol. 3, no. 4, pp. 333–354, 2012.

[11] J. Salas and A. Harth, "Finding spatial equivalences accross multiple rdf datasets," in *Terra Cognita Workshop,co-located with ISWC*, R. Grtter, D. Kolas, M. Koubarakis, and D. Pfoser, Eds., 2011, pp. 114–126.

[12] A. Cohn, B. Bennett, J. Gooday, and N. Gotts, "Qualitative spatial representation and reasoning with the region connection calculus," *Geoinformatica*, vol. 1, pp. 1–44, 1997.

[13] Geosparql - a geographic query language for rdf data. http://www.opengeospatial.org/standards/geosparql. Accessed: 2009-12-15.

[14] M. Koubarakis, M. Karpathiotakis, K. Kyzirakos, C. Nikolaou, and M. Sioutis, "Data Models and Query Languages for Linked Geospatial Data," in *Reasoning Web. Semantic Technologies for Advanced Query Answering,Invited tutorial at the 8th Reasoning Web Summer School 2012 (RW 2012)*, T. Eiter and T. Krennwallner, Eds., vol. LNCS 7487. Springer, 2012, pp. 290–328.

[15] A. G. Cohn and J. Renz, *Qualitative Spatial Representation and Reasoning*, ser. Handbook of Knowledge Representation. Elsevier, 2008, pp. 551–596.

[16] What is racerpro. http://www1.racer-systems.com/products/racerpro/index.phtml. Accessed: 2009-12-15.

[17] M. Stocker and E. Sirin, "Pelletspatial: A hybrid rcc-8 and rdf/owl reasoning and query engine," in *6th Intern. Workshop on OWL: Experiences and Directions (OWLED2009)*. Springer-Verlag, 2009, pp. 2–31.

[18] M. Sioutis, S. Li, and J.-F. Condotta, "On redundancy in linked geospatial data," in *2nd Workshop on Linked Data Quality (LDQ)*, ser. CEUR Workshop Proceedings, A. Rula, A. Zaveri, M. Knuth, and D. Kontokostas, Eds., no. 1376, 2015. [Online]. Available: http://ceur-ws.org/Vol-1376/#paper-05

[19] C. Nikolaou and M. Koubarakis, "Fast consistency checking of very large real-world rcc-8 constraint networks using graph partitioning," in *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, 2014, pp. 2724–2730.

[20] A. Abdelmoty, P. Smart, and B. El-Geresy, "Spatial reasoning with place information on the semantic web," *International Journal on Artificial Intelligence Tools*, vol. 23, no. 5, p. 1450011, 2014.

[21] E. Younis, C. Jones, V. Tanasescu, and A. Abdelmoty, "Hybrid geo-spatial query methods on the semantic web with a spatially-enhanced index of dbpedia," in *Geographic Information Science*, ser. Lecture Notes in Computer Science, N. Xiao, M.-P. Kwan, M. Goodchild, and S. Shekhar, Eds., vol. 7478. Springer Berlin Heidelberg, 2012, pp. 340–353.

[22] D. Wolter and J. Wallgrn, "Qualitative spatial reasoning for applications: New challenges and the sparq toolbox," in *Qualitative Spatio-Temporal Representation and Reasoning: Trends and Future Directions*, S. Hazarika, Ed., 2012, pp. 336–362.

# Graph-based Data Integration in EUDAT Data Infrastructure

Vasily Bunakov[*], Paolo D'Onorio De Meo[†], Stephan Kindermann[‡], Anna Queralt[§] and Jedrzej Rybicki[¶]
[*]Science and Technology Facilities Council (STFC),
Harwell Campus, Didcot OX11 0QX, UK
Email: vasily.bunakov@stfc.ac.uk
[†]CINECA,
Via dei Tizii 6, 00185 Rome, Italy
Email: p.donoriodemeo@cineca.it
[‡]Deutsches Klimarechenzentrum GmbH (DKRZ),
Bundesstrasse 45a, 20146 Hamburg, Germany
Email: kindermann@dkrz.de
[§]Barcelona Supercomputing Center (BSC),
Carrer de Jordi Girona 29, 08034 Barcelona, Spain
Email: anna.queralt@bsc.es
[¶]Juelich Supercomputing Center (JSC),
Wilhelm-Johnen-Strasse, 52425 Juelich, Germany
Email: j.rybicki@fz-juelich.de

***Abstract*—European Data Infrastructure (EUDAT) is a distributed research infrastructure offering generic data management services to the research communities. The services deal with different phases of the data life cycle, some of them are tailored to account for special needs of the individual communities or replicated to increase the availability and resilience. All that leads to scattering of the large and heterogeneous data across service landscape limiting discoverability, openness, and data reuse. In this paper, we show how graph database technology can be leveraged to integrated the data across service boundaries. Such an integration will facilitate better cooperation among the researchers, improve searching and increase the openness of the infrastructure. We report on our work in progress, to show how better user experience and enhancement of the services can be achieved by using graph algorithms.**

***Keywords*–*Data Integration; Graph Databases; Designing for Open Data; Linked Data.***

## I. Introduction

Nowadays, it is widely accepted that public data, and in particular research results, should be made accessible to society, facilitating better, more efficient science and innovation. In line with the open data and open access movements, EUDAT [1] is a pan-European initiative building a sustainable cross-disciplinary and cross-national data infrastructure providing a set of shared services for accessing and preserving research data. The EUDAT services work with digital collections comprised of data objects. The term data object in EUDAT is pretty broad and encompasses structured data, text, multimedia binaries, binary output of scientific simulations, and much more. In this paper, we will use terms data object, digital object, and object interchangeably. EUDATs vision is to enable European researchers and practitioners from any research discipline to preserve, find, access, and process data in a trusted environment, as part of a Collaborative Data Infrastructure (CDI).

The problem with a generic infrastructure like EUDAT is that it must fulfill a lot of expectations at the same time. The expectations come from different communities or usage scenarios. The usual way of dealing with different community requirements is to add new services to the infrastructure portfolio or tailor the existing ones accordingly. It is a strength and weakness at the same moment. The cost of the flexibility is the complexity of the service landscape. It is further amplified by the geographical distribution used to increase the scalability and resilience of the infrastructure. There are many instances of the same service created at different locations to serve different groups of users. Users use different services to tackle different problems or phases of data life-cycle. Altogether, this leads to fragmentation of the content: some data objects are uploaded to one service, others to other service. In extreme cases, it can even happen that the same data object is uploaded to many services as there is no way of finding out if and where it was previously stored. External identifiers as used by some services, for instance in form of handles (like [2]), do not necessary help. They are opaque, hash-based values generated independently of the content of the object. To cope with this heterogeneity a much more expressive model of the data stored in the infrastructure is required.

In computer science, every decent software design starts with an analysis of the domain model [3]. This approach is not directly applicable to the EUDATs case. The reason for that is the heterogeneity the project has to deal with. As a resource and service provider it is not in the position to define a common domain model to account for all the special use cases originating from the communities. It rather tries to account for the domain models coming from different communities and map them on services in generic CDI. In this paper, we show how we provide the communities with a unified view of the infrastructure and the data that are already stored in the existing EUDAT services. Such integrated view will enable better understanding of the data, make a first step towards data interoperability, increasing openness, and potentially facilitate data reuse. We show how we plan to establish and store such integrated model of the different data sources, and how it allows for new features and service extensions.

ISBN: 978-1-61208-457-2

It is good to offer tailored services to attract users but it is at least equally important to use content collected in the infrastructure as an attractor. Researchers can be interested in using the CDI solely based on the content it stores. The abundance of content might lead to a situation where it is hard to find or even be aware of all the data objects relevant for given scientific endeavor. The challenge, which is not unique to EUDAT, is to make the collected content visible, and searchable in ways going far beyond the currently supported keyword-based searches or faceted searches. Application of graph-based algorithms [4] revolutionized the way the Internet search engines work and how people engage in social interactions [5]. We believe that such algorithms might not be directly applicable for the scientific communities and data (e. g., most popular data set might not be the most attractive for the researchers). It would be, however, beneficial to offer graph-based descriptions of the content so that individual users can work on their own searching algorithms or just explore the content in an interactive way. The graph abstraction is already used to successfully tackle Big Data challenges [6].

Our goal is to create a generic infrastructure service to integrate the content gathered from different sources. As a service provider we are not in a position to impose a common domain model on all the communities we serve. Therefore, we provide a flexible service to describe single use cases or domains as interactive graphs. This is an abstraction that is well tested, easy understandable and quite powerful at the same time.

The rest of the paper is structured as follows. We present our design in Section II. We proceed with a short description of the implementation approach. Subsequently, an overview of the use cases currently worked on is given. We conclude this work with a summary and a list of future challenges in Section V.

## II. Design

The core services offered by EUDAT CDI are shown in Figure 1. B2DROP is a service for storing, synchronizing, and exchanging dynamic research data with colleagues or team members. B2SHARE provides an easy way to upload, tag and share research data, which is made citable via persistent identifiers (PID). B2SAFE enables an automatic, rule, and policy-driven replication of data across a federation of data centres. B2STAGE allows data to be staged into and out of the CDI to, for instance, external high-performance computing services to process the data. Finally, B2FIND exposes a metadata catalog through a user-friendly, web-based search portal and a standard API. The authentication and authorization infrastructure (AAI) is orthogonal to all these services, and controls access to the infrastructure.

To improve the discoverability of the content scattered across different services and locations, we aim at providing a unified, expressive view of all the data items collected. We decided to include relations between objects to add flexibility to the model and allow for exploration of the content by just following those links. In other words, we create a graph describing the infrastructure and integrating the content collected across many services.

A valid approach for data integration and an often prerequisite for further analysis are so called “data lakes”. They are collections populated by the data extracted from all the services in an infrastructure. Although the approach is valid it is also controversial. Especially the need for replicating the data might render it prohibitively expensive. Therefore, we decided not to duplicate the content but just include the metadata representation of data objects. In our graph, we model them as nodes with properties describing details like object name, creation date, etc. Graph nodes are also used to model further entities like service instances, people, or metadata objects. To model all kinds of dependencies between digital objects we use relations (edges in graph).

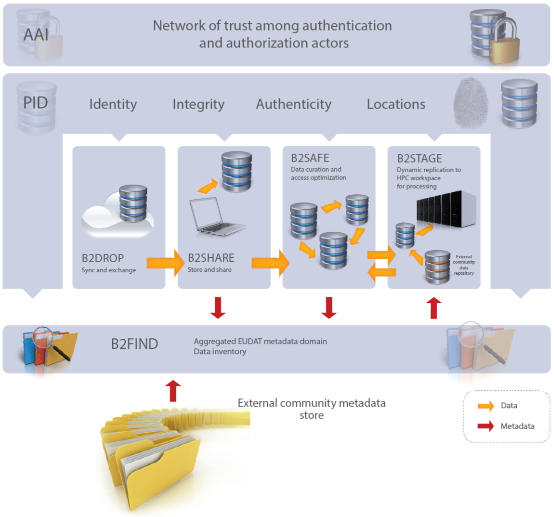

AAI
Network of trust among authentication and authorization actors
PID
Identity
Integrity
Authenticity
Locations
B2DROP
Sync and exchange
B2SHARE
Store and share
B2SAFE
Data curation and access optimization
B2STAGE
Dynamic replication to HPC workspace for processing
External community data repository
B2FIND
Aggregated EUDAT metadata domain
Data inventory
External community metadata store
Data
Metadata

Figure 1. EUDAT Service Landscape

When tackling data integration one can follow bottom-up or top-down approach. In case of bottom-up, the data are gathered from services with help of specialized spiders, cleansed (if required), and then uploaded to a common repository to provide complementary, integrated view of the content. Top-down approach, on the other hand, promotes the repository to the single user-facing service with just one view of the data. During the upload of the data, the individual users describe the object with help of graph semantics. From there, the data are propagated to individual back end services. Both approaches have their advantages and drawbacks. Since we are still in an exploratory phase of implementing the service, we decided to follow the bottom-up approach: Gather as much data as possible, provide alternative view of the infrastructure and data, evaluate the benefits of such data integration and (in case of positive result) promote the service. At least for some services also an intermediate step would be possible: Graph database could be used to substitute the existing relational back end.

The bottom-up approach produces graphs describing domains of single services or domains of single communities. In the process of data integration, those graphs shall be merged together. To this end, integration points (graph overlaps) have to be identified. In general, there are two kinds of graph overlaps: common nodes in two or more graphs and relations connecting nodes originating from different graphs. An obvious candidate for a common node is a person: the same user can own data objects across multiple services. Also, metadata nodes describing people like affiliation, community, or research interests can constitute good integration points. Another type of graph overlaps are the digital objects. It is, for instance, possible to have replicas of an object stored in different places or a set of objects derived from a given root object. Some EUDAT services assign external persistent identifiers to the managed object, so this could be clearly used to identify the same object across services. As stated above we are not storing the actual content of the digital objects, thus it is not possible to define content-based identity of any given objects. We do, however, store metadata describing objects. This metadata are either technical metadata (like checksum), or community-provided semantic metadata like provenance description or keywords. Some of the metadata will create common nodes across services but metadata can be used to identify similar objects across service boundaries. Such a similarity can be modeled as a relation (graph edge) crossing service boundaries. In the future, we plan to incorporate new services for extracting even more features from digital objects and store those features in the common graph. This can be based for instance on Linked Data AppStore [7] and would certainly help to identify commonalities between different domains.

The high-level goal of our design is “about making links, so that a person or machine can explore the web of data”, which is a quote from the seminal Tim Berners-Lees note on Linked Data [8]. We try to incorporate as many good design principles from the world of linked data as possible. There are, however, some implementation details which differ from the usual way in which linked data is implemented. First of all, some of the services in EUDAT CDI do not offer HTTP(S) URIs for accessing the data. Secondly, we are in sought of benefits from exploring the graph and applying graph algorithms. Therefore, we decided not to use the RDF [9] end point but rather upload the data to a graph database where people can interact with it. In other words, we squashed together the steps of collecting the data and exploring the data. In the future we can expose the collected data as RDF and SPARQL interface to account for more sophisticated use cases and enable better integration with other infrastructures.

## III. Implementation

In this section, we describe some of the implementation details of our work in progress on graph-based data integration. As already explained we follow a bottom-up approach and in the first step extract data from different EUDAT services to create distinct graph models in those bounded contexts. In the next step we integrate the data by connecting the single graphs. To manage graphs we use graph database neo4j [10], a native graph database available under GPLv3 license. It supports full Atomicity, Consistency, Isolation, Durability (ACID) consistency model, and it offers an interactive graphical interface, clients in many different programming languages, and a ReST API, leaving us with many options with regard to integration with other services as well as offering access to end users. neo4j uses property graph as internal graph model. It means that nodes in the graph can have properties and each node can be labeled with (multiple) labels. Labels can be used to divide the entities in the graph into different “abstraction classes”. Properties, on the other hand, describe particular entities. Relations in property graph can have properties and names, they are also directed. neo4j offers quite a flexibility with respect to properties. It is not required that all the nodes have the same properties. Even if they share the same label it is still possible to introduce the heterogeneity. An example would be a graph with nodes representing people where some of the nodes have a property called address (if people decided to share their address) while others do not. This kind of flexibility in the graph model is really useful for evolving the model over time, e. g., when new features are added or more data are collected we can simply add properties to newly created nodes while keeping old nodes valid and potentially update them later. This kind of evolution is proven to be hard in relational databases. For an extensible explanation and comparison of current graph data models we refer the reader to [11].

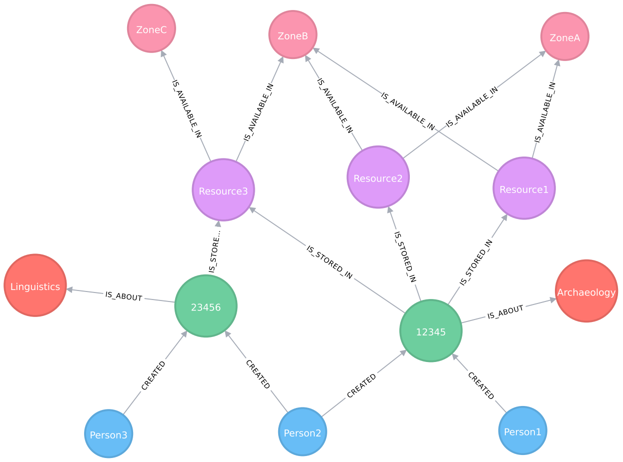


Figure 2. Example of graph data integration

## IV. Use cases

This section describes the use cases we are currently implementing to showcase the advantages and possibilities that arise from the graph representation of the EUDATs data.

B2FIND stores metadata about the digital objects, including their authors, language, and discipline, among others. By structuring this information as a graph, it is easy to infer new relationships from the already existing ones. For instance, if two persons are authors of the same object, then we can assume that they know each other. And if a person recently created objects belonging to a discipline, we can infer that this person works in this discipline. This information can be used, for instance, to look for collaborators with a particular expertise, as well as how to reach them through co-authorship relationships by means of a shortest path query. We plan to further integrate this information with the data coming from the authentication and authorization service, which also stores affiliation of the users. In this way, we can restrict the searches, for instance by finding only experts from a given institution or country. A clear application of this use case is to propel collaboration between researchers based on the identified social-network-like links. But also more technical benefits can be obtained, for instance the location of the data object can be changed based on the expected usage to optimize the access times.

A rather more technical than social use case is fed with the data from the B2SAFE service. There the data objects are registered, replicated to avoid data lost, and made referenceable via globally unique persistent identifiers managed by the corresponding administrative domains. The PID can be also used to locate the copies (replica) of data objects across different federations. A graph database is used to model the ownership of the data, actual replication paths of data objects, and store technical metadata describing the objects. The model also include collections (with metadata descriptions) to extend the limited functionality of the actual B2SAFE back end. There are at least two benefits of this data. First of all, it gives the data owner a good view of the infrastructure and status of their data and thus improve the trust in the CDI. Secondly, the graph database could relate a person to all of its PID and replicas across all federations allowing for better data accessibility.

Both aforementioned use cases can be used to understand the actual data integration we are sought after. Let us consider a graph as the one shown in Figure 2, where the digital objects, identified by their persistent identifiers (12345 and 23456), are gathered from the different EUDAT services. The resources in which an object is replicated, as well as the zones in which a resource is available, are obtained from the B2SAFE service. The authors of a digital object and the discipline related to it can be collected from B2FIND.

The B2SHARE use case is pretty close to the B2SAFE case. There are, however, two important differences. The content in B2SHARE is currently not replicated and there are community-provided metadata descriptions available. The model we are currently using to store this information is pretty straight-forward. For each data object we have an uploader, set of metadata (currently modeled as a single graph node) and set of keywords (each keyword is a separate node). An interesting application of this model is to provide the users with their individual “universe” composed of data objects, keywords, and people. The universe is generated with a breadth-first search of given depth and includes the “most important” objects from the domain. This feature can be incorporated into B2SHARE in the future, combined with the social-like features described in the first use case.

A different kind of use case is implemented by a representative of European Network for Earth System Modeling (ENES), which is one of the EUDAT communities. This climate research community is developing comprehensive Earth system models capable of simulating natural climate variability and human-induced climate change. The use case concentrates on modeling the distributed ENES data federation: the organization of datasets in collections served by data services hosted by data servers. The services come both from ENES community and EUDAT. The sole existence of such an overview contributes to better mutual understanding between a community (ENES) and provider of generic services (EUDAT), potentially resulting in a better usage of services. Since the model includes information about data objects harvested from EUDAT and ENES worldwide data collections, there are some interesting overlaps. EUDAT cataloged data collections are from a later phase of the data life cycle (published archived objects with a DOI assigned). The data are still worked on, thus newer versions of the same collections (or subparts of the collections) are accessible in the ENES data federation. By connecting those two worlds an integrated view of a life cycle of a digital object can be derived and the provenance of single objects and collections can be better tracked and understood.

Finally, the semantic annotations service developed in EUDAT, called B2NOTE is yet another use case for the graph database integration. The goal of B2NOTE is to provide a plug-in to the graphic interfaces of other EUDAT services for human annotation, as well as text mining tools for the automated annotation in the back end. So, the graph database could successfully address and handle the annotation provenance records: who, when, in what EUDAT service and by what tool or machine agent has produced the annotation. The annotations will be then available across all services.

## V. Conclusion and Future Work

In this paper, we reported on our work in progress showing how the graph database technology allows EUDAT to integrate the data across services boundaries to provide features originally missing. We are still in a preliminary phase but the first use cases implementation already lead to some improvements, for example an easier way to walk through relations inside and across the separate services. We design our experiment in such a way that the main focus was laid on the use cases and not on, e. g., selecting the best graph database or best service and data integration scenario. This later subjects remain open until the potential of the approach is positively verified.

In our work, we have identified some challenges. Some of them will follow from our decision not to clone content inside the graph database, but only include metadata. We will have to provide a means to keep the metadata up-to-date and efficiently retrieve the data from all services from all federations. On the higher abstraction layer, we will have to work on identifying integration points between services. Such points will have to be non-intrusive, as we are not going to impose anything on the data models of single services nor communities. To this end, more effort will be made in the data cleansing process. Lastly, although this is not yet a formal requirement, we are considering an offering of an RDF endpoint, for instance, to facilitate data exchange with other infrastructures.

Among the most promising improvements identified so far, is the potential to offer better user experience by making the borders between services less visible and less relevant to the users. After a successful integration of the data, all the information will be available in all the services. The better user experience is also given by the possibility to add social-network-like features to the existing services and offer links to explore the data domain of EUDAT. In particular, a much more powerful and customizable searching functionality can be implemented based on the data collected in the graph database. Finally, the integrated information can be used to better understand the community domains, access patterns and use cases, and the EUDAT CDI itself to tune it accordingly.

## Acknowledgment

The work has been supported by EUDAT2020, funded by the European Union under the Horizon 2020 programme - DG CONNECT e-Infrastructures (Contract No. 654065).

## References

[1] W. Gentzsch, D. Lecarpentier, and P. Wittenburg, “Big data in science and the EUDAT project,” in SRII Global Conference, Apr. 2014, pp. 191–194.

[2] R. Kahn and R. Wilensky, “A framework for distributed digital object services,” International Journal on Digital Libraries, vol. 6, no. 2, Apr. 2006, pp. 115–123.

[3] E. Evans, Domain-Driven Design. Addison-Wesley, 2004.

[4] S. Brin and L. Page, “The anatomy of a large-scale hypertextual Web search engine,” Computer Networks and ISDN Systems, vol. 30, no. 1-7, Jul. 1998, pp. 107–117.

[5] J. Weaver and P. Tarjan, “Facebook Linked Data via the Graph API,” Semantic Web – Interoperability, Usability, Applicability, vol. 4, no. 3, 2013, pp. 245–250.

[6] V. N. Gudivada, S. Jothilakshmi, and D. Rao, “Data management issues in big data applications,” in ALLDATA 15: The 1st International Conference on Big Data, Small Data, Linked Data and Open Data, Apr. 2015, pp. 16–21.

[7] R. Dumitru et al., “The linked data AppStore,” in Mining Intelligence and Knowledge Exploration, ser. Lecture Notes in Computer Science. Springer International Publishing, 2014, vol. 8891, pp. 382–396.

[8] T. Berners-Lee. Linked data. [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.html [retrieved: Dec., 2015]

[9] G. Schreiber and Y. Raimond. RDF 1.1 primer. [Online]. Available: http://www.w3.org/TR/rdf11-primer/ [retrieved: Dec., 2015]

[10] J. Webber, “A programmatic introduction to Neo4j,” in SPLASH ’12: 3rd ACM Annual Conference on Systems, Programming, and Applications: Software for Humanity, Oct. 2012, pp. 217–218.

[11] R. Angles, “A comparison of current graph database models,” in ICDEW ’12: 28th IEEE International Conference on Engineering Workshops, Apr. 2012, pp. 171–177.

# IPR Issues in Data Sharing via Linkage of Platforms and Apps

Iryna Lishchuk, Marc Stauch, Nikolaus Forgó
Institut für Rechtsinformatik
Leibniz Universität Hannover
Hannover, Germany
e-mail: {lishchuk, stauch, forgo}@iri.uni-hannover.de

***Abstract*—Data exchange systems have made it possible to link platforms, apps, wearables and share the users´ data. Apart from personal data, collected by the platforms, also user generated content may be shared. However, sharing the content brings various intellectual property (IP) issues into play – on top of privacy matters. So, sharing creative content requires authorization from the content owner. But who is the content owner in respect of content shared online? From whom and how can a service provider obtain rights on use of the content? In this article, we explore some legal issues associated with content sharing and provide solutions on how these issues may be resolved.**

***Keywords-user-generated content; Intellectual Property Rights (IPR); IP protected content; content license; copyright.***

## I. INTRODUCTION

Many platforms, like Facebook, Fitbit and Twitter, release their application programming interfaces (API) for the purposes of data sharing. By doing so, they offer third party service providers a technical possibility to access and share users´ data. Through a platform API, a third party service can connect to the platform and exchange data with it. Apart from the user´s personal data, such as name, date of birth, etc., the platform may also store content generated by the user. Such content, if produced by intellectual creation, may be protected by intellectual property rights (IPR).

While processing and sharing the users´ personal data is subject to the law on data protection and may require consent of the user, the processing of IP protected content will be subject to intellectual property law and, unless exceptions apply, require authorization of the right holder (who is not necessarily the user). In the case of personal data, the user behind the data is normally identified or at least identifiable. Hence, the user to whom the data relate has the right to decide with whom and how to share his data. However, in the case of user-generated content, which is freely shared among the networks, the question about who is the right holder and may decide on its exploitation is often complicated by the unclear origin of the work. Creative content may be produced by a user who uploads such content, or it may also be created by a group of users (who would share copyright in it together), or it may be a result of post-processing of someone´ else work (requiring permission by the latter), etc. When a user posts some creative content to the platform services, it does not mean that the user is the copyright owner or even has the right to upload such content and share it with the public. On the other hand, almost all forms of online communication require copying and distributing creative content, thus becoming copyright-related [1]. This makes it necessary for service providers, engaged in spreading user generated content, to obtain a copyright license. However, a problem arises when the holder of rights in creative content is not that easy to identify. In such a case, how should the service provider go about obtaining the relevant license?

In this paper, we explore legal issues such as these, associated with content sharing and content licensing and suggest some options as to how service providers may get rights in order to carry creative content and provide their services to their users.

This paper is organized as follows. Section II provides insight into data sharing in clinical research. Section III describes the types of data collected by the platforms. Section IV then deals with IP rights in content and IPR implications by content sharing. Data sharing via API exchange systems follows in Section V. Section VI deals with licensing implications. The overall findings are summarized in Section VII.

## II. DATA SHARING IN CLINICAL RESEARCH

Data sharing is widely used now as a way to increase service functionality. Many service providers offer an option to share content via Facebook, Twitter, etc. Increasingly, too, the research community is looking into data sharing as a potential resource for expanding research.

One such ICT research project funded under the EU 7th Framework Program is ‘MyHealthAvatar’ (full name “A Demonstration of 4D Digital Avatar Infrastructure for Access of Complete Patient Information”, abbreviated to “MHA”) This aims at creating a platform, . “*…that offers access, collection and sharing of long term and consistent personal health status data through an integrated digital representation of an in silico environment, which helps to deliver clinical analysis, prediction, prevention and treatment tailored to the individual citizen*.” [2]. Various possibilities are being investigated to allow, inter alia, connecting the avatar to hospital records, and to third party social networks (e.g. Facebook, Twitter, etc.), to extend the population of medical data within the platform.

## III. PLATFORM DATA

Because the MHA project is engaged in clinical research, lifestyle platforms Fitbit [3], Withings [4] and Moves [5] are the primary sources for data sharing. Also, the social platform Twitter is being explored for linking, but rather because of the role of Twitter in spreading the content, than the nature of the data, which Twitter stores.

The information stored on these platforms may have different value and quality in terms of the law. Personal data and user-generated content are the two major categories of information processed by the service providers. The processing of these two types of information is subject to different legal rules.

### A. Personal Data

Most social platforms collect in one way or another data related to the user. Usually, platforms ask the user to provide some personal information, such as name, date of birth, e-mail, etc., when creating a user account.

Data, which may be associated to a particular user, who is identified or may be identified by some parameters or features, will qualify as personal data. Personal data comprises "*any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified , directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity*" [6]. Processing of personal data is subject to the law on data protection.

Data collected by wearable devices, such as Moves, Fitbit and Withings, would normally have the status of personal data Moves collects data from a mobile application Moves App, which records walking, cycling and running, which the user does while the application is on [5]. The major part of Fitbit data also comes from wearables, which track the physical activity of the user. The collected data may include the number of steps taken by the user, heart rate, calories burnt, etc. "*Every time Fitbit owners walk by their wireless base station (Bluetooth dongle and computer), data from their Fitbit device is silently uploaded in the background to fitbit.com.*" [7]. The user can also enter certain data to fitbit.com manually, such as sleep logs, food logs, other activity logs.

According to European data protection law, data concerning the user´s health falls into a special category of sensitive personal data. Processing of sensitive data has to comply with more stringent legal requirements than those applicable to processing of personal data as such [6]. The sharing of sensitive health related data among the platforms raises multiple legal issues and privacy concerns, which have been described elsewhere [8], [9] and are outside the scope of the present paper.

The data recorded by the tracking devices or data entered by the user manually, which does not involve any creative input would normally qualify as personal data (with a higher or lower degree of sensitivity). Another quality may be attributed to data generated by the user himself, such as comments or images taken by the user. These types of data may expose certain parameters relating to a particular user (such as when a user is marked as the author), thus falling into the category of personal data. At the same time, this data may comprise creative input invested by the user (or another person), thus qualifying as a copyright work. Creative content, related to a particular person, such as marked as author or captured on a picture, would be subject to both legal regimes: the law of copyright and data protection at once.

### B. User Generated Content

Most online platforms, either lifestyle or social, allow their users to submit their own content, like text, photographs or other data and information. Any data, which is produced and supplied to service providers in digital form constitutes "digital content" [10]. Such content may include "*computer programs, applications, games, music, videos or texts, irrespective of whether they are accessed through downloading or streaming, from a tangible medium or through any other means.*" [10]. Digital content created and provided to the online services by the users and made by such means accessible directly to the public is commonly referred to as "user generated content" [1]. If produced by intellectual effort, such content may be protected by IP rights (IP protected content).

### C. Copyrighted Content

A number of items uploaded to the platform services, including images, melodies, videos, commentaries, etc., by showing a certain degree of creativity may relate to original intellectual creations in the literary or artistic domain and constitute works protected by copyright [11]. A comment where the user "*through the choice, sequence and combination of those words ... may express his creativity in an original manner and achieve a result which is an intellectual creation"* [12] would qualify as a copyright work and be protected as such. Also, a picture taken by a user exercising free and creative choices thus stamping a picture with his personal touch [13] should be copyright protected.

However, just as data protection law has certain requirements for processing of personal data, so too will the use of copyrighted content on the digital services need to comply with the rules of copyright law. We look further at the substance of these rules below.

## IV. IPR Issues In Content Sharing

### A. Protected Rights

Whereas reading a book or listening to music does not create a copyright relevant action, the upload of a photo to online services, sharing music online or streaming may produce a copyright relevant action. The reason is that, in contrast with the case in which there is simple perception of the work by a viewer or hearer, technical actions of this kind involve a degree of copying or communication to the public.

Reproduction and communication to the public may also be carried out by service providers in the course of providing their services. Thus, transmission of content items between and/or on behalf of the user, the upload and hosting of content items on the platform facilities, or making the content items available to the others may qualify as one or another copyright relevant action and, unless exceptions apply, require authorization by the right owner.

Because the "fair use" doctrine [1] and exhaustion of copyright do not apply to the digital content commonly shared via online services [15], service providers who deal with the user generated content would normally require a copyright license from the user.

### B. Content License

Platforms usually obtain such a license on use of IP protected content when the user registers for a platform account and agrees to the platform terms. As a rule, a content license is incorporated into the platform terms of use and constitutes part of the agreement between the provider and the user.

Normally, platforms acquire a complete copyright license, which allows them to perform any actions with the user´s content as required to provide their services. A typical content license is granted on a royalty-free, non-exclusive, perpetual, worldwide basis and includes the sublicensable right of reproduction, modification, distribution, communication and making the content available to the public. For instance, Fitbit users grant Fitbit a "*perpetual, irrevocable, non-exclusive, worldwide, royalty-free license, with the right to sublicense, to reproduce, distribute, transmit, publicly perform, publicly display, digitally perform, modify, create derivative works of, and otherwise use and commercially exploit any text, photographs or other data and information you submit to the Fitbit Services (collectively, "User Generated Content") in any media now existing or hereafter developed, including without limitation on websites, in audio format, and in any print media format.*" [16]. Similar content license conditions may be found in the terms of other platforms, like Twitter and Withings.

## V. DATA SHARING VIA API EXCHANGE SYSTEMS

As mentioned, most platforms, which collect information from their users, be it personal data or digital content, allow the sharing of such information via API exchange systems. However, in allowing third party services to use an API, platforms normally do not allow the use of creative content, which they store.

### A. API Exchange Systems

API stands for application programming interface and is an element through which software interact and exchange information with each other. The use of a platform API allows external applications to communicate with the platform and access the platform data (if a platform allows this). In legal terms, an API can be defined as an element of a computer program, which provides for "*a logical and, where appropriate, physical interconnection and interaction ... to permit all elements of software and hardware to work with other software and hardware and with users in all the ways in which they are intended to function*." [17]. For example, when a word processor sends a document to a printer, the word processor talks to the printer driver via API [18]. Although an element necessary for interoperability, API is a constituent part of a platform and usually released into use under an API license. This then allows software developers to use platform APIs in order to develop compatible apps designed to interact with a platform and exchange users data.

### B. API License

As may be observed, when platform operators release platform APIs, they enable third party services to connect to the platform and share the data. Therefore, a typical API license is generally limited to the purpose of data sharing. For instance, Fitbit allows use of Fitbit API "*to develop Applications designed to interact with and enhance the Fitbit Platform, to retrieve or post Fitbit Data, subscribe to User Data-feeds and render and display information in external applications according to these Terms of Service"* [19]. As a rule, a personal, non-exclusive, non-transferrable license is granted.

Whereas an API license allows use of API for data sharing, it is, as previously noted, mostly the case that rights on the use of content itself are not included, unless such rights are expressly granted. In these circumstances, third party service providers, who intend to carry user generated content on their services, need to get the content license by themselves. The ways in which service providers may do this are described below.

## VI. LICENSING IMPLICATIONS

There are several options how a service provider may obtain rights on use of content. One is to get the rights from the platform. Another possibility is to obtain a content license directly from the user himself. However, both of these options carry further legal implications. These implications can relate to copyright ownership, validity and survival of rights, applicable contract type, form requirements, etc, and may vary from jurisdiction to jurisdiction. Some key points in this respect, which are relevant to cases of content licensing by linking, are discussed below.

### A. Content Sublicense from the Platform

Platforms, which have a sublicensable content license, have a right to sublicense their rights, which, however, they rarely make use of.

In fact, out of the considered platforms, it appears that only Twitter, when licensing its API, grants developers rights to use the content. In particular, Twitter accords the developer "*a non-exclusive, royalty free, non-transferable, non-sublicensable, revocable license...to...Copy a reasonable amount of and display the Content on and through your Services to End Users...; Modify Content only to format it for display on your Services"* [20].

First, Twitter has the right to sublicense its rights in content because the user grants to Twitter "*a worldwide, non-exclusive, royalty-free license (with the right to sublicense) to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute such Content in any and all media or distribution methods (now known or later developed)."* [21]. As explained by Twitter, the user authorizes not only Twitter to make the user´s Tweets available to the public, but also "*let others do the same."* [21]. Second, the user generated content is included into the term "Content", which may be shared via Twitter API. In the Twitter API license agreement, the term "Content" covers

"*Tweets, Tweet IDs, Twitter end user profile information, and any other data and information made available to you through the Twitter API or by any other means authorized by Twitter, and any copies and derivative works thereof.*" [20].

In contrast, as follows from the API license terms of other platforms, in particular Fitbit and Withings (which also carry some user generated content), express grant of rights on use of the content is not included. If we consider Fitbit, and also Withings, it is rather unclear whether they allow connecting services to use and re-post the user´s content on their services or not.

Thus, dealing with Fitbit first, it also hosts some user generated content and has a sublicensable content license [16]. Fitbit allows data sharing via its API, but 'User Generated Content', as defined in the Fitbit Terms of Use [16], is not included in the scope of Fitbit Data, which may be shared via Fitbit API. The Fitbit Data, as defined in the Fitbit API license agreement, covers "*the user data collected from the Fitbit Tracker and made available to you through the API"* [19]. Fitbit allows a third party service provider to "*use the API to retrieve or post Fitbit Data, subscribe to User Data-feeds and render and display information in external applications"* [19], but Fitbit is silent on the rights to re-post the User Generated Content. At the same time, Fitbit is rather clear in not allowing developers to "*upload or otherwise transmit any content that you do not have a right to transmit under any law or under contractual relationships*" [19]. However, whether the term "content", as used in the Fitbit API license agreement, covers also "User Generated Content", as defined in the Fitbit Terms of Use (and which a third party service provider may not transmit on its services ) is for Fitbit to answer.

A similarly unclear content licensing practice is pursued by the lifestyle platform Withings. Withings also provides a function to submit users´ comments and opinions to Withings website. Withings also obtains "*a sub licensable, right on a worldwide basis to represent and reproduce your commentary and/or opinion in whole or in part, in a lineal manner or not on any media, such as the Website, press review or advertising, presentation or any physical or digital media as long as the rights shall enjoy legal protection*" [22]. By licensing its API for data sharing, Withings allows use of API to "*exchange data concerning you, Withings or Withings' Products and Services Users* [23]. As long as a commentary or opinion can be related to a Withings user (for instance, when the user is marked as the author), then the user´s comments may be considered as relating to the user and included into the scope of data, which may be shared via Withings API. On the other hand, if Withings expects to make use of the user´s comments or opinion, such as in an advertisement or third party website, Withings should contact the user. In cases where it is unable to reach the user, or upon the user´s request, it also reserves the right to use the user´s commentary or opinion without identifying the user as the author [22]. At the same time, this wording and practice of Withings makes it questionable whether a third party service designed to interact and share data with Withings may re-post the user´s comments on its services or not.

In the absence of an express term, an implied license on use of copyrighted content may be presumed. However, an implied license may not be regarded as a reliable instrument for getting the rights because of varying interpretation rules and the copyright licensing implications, which we consider next.

### *B. Implied Copyright License*

The legal strength of an implied license as a basis for using copyrighted content is relative and depends on the rules on interpretation of agreements and court practice. The rules on interpretation of agreements vary from jurisdiction to jurisdiction and relevant domestic case law is rather scarce. If an agreement is to be interpreted by purpose, as is typically the case under the German or English law, then in the absence of an express term an implied license may be presumed. Hence, a UK court might depending on the facts of the case accept an implied copyright license where such license is "*necessary to give business efficacy to the contract"* [24]. If accepted, an implied license would be limited to the purpose of contract. In the context of an API license agreement limited to the purpose of data sharing, it may be argued that a developer might be entitled to an implied personal, non-assignable and non-sublicensable, royalty free copyright license to access, copy and re-display the user´s content on its service as necessary to provide a service to the user. Such implied copyright license might be considered as justifiable for data sharing with a platform whose data assets subsist for the most part in the user generated content, like Twitter, for example. Otherwise, i.e. in the absence of such an express content license, the principal goal of using the API for data sharing would be lost.

In contrast, most data assets of lifestyle platforms Fitbit or Withings come from the tracking devices. It is obvious that exactly such lifestyle data is a target for data sharing. Some smaller part of Fitbit and Withings data may comprise creative content produced by the users, such as comments, opinions or photographs. But, in comparison to the volume and value of the lifestyle data, it is hardly arguable that such user generated content would constitute the primary goal for data sharing. Under these circumstances, it is doubtful how far a content license in the Fitbit or Withings API license agreement is "*necessary to give business efficacy to the contract*". Hence, the chance that an implied content license as granted by Withings or Fitbit under an API license agreement would be accepted by the court is arguably fairly low. Under the rules of verbal interpretation of agreements (as may be the case under the Russian law [25], for example), the prospects for an implied content license may also be assessed as negative. According to oral interpretation, a right, which is not expressly granted is to be considered as not granted at all.

### *C. Content License from the User*

Alternatively, as noted, a service provider may get a license on use of the content from the user. The core legal issue here is that the user, who introduces creative content to the platform, is not necessarily the author with the right to make such content available to the public.

According to the rule of first ownership in copyright, it is the original author, who created a work and who owns copyright in it [26]. Creative content may be generated by a group of people, sharing co-authorship and copyright respectively. Such content may also be produced by re-using and/or transforming pre-existing copyright works [1]. The latter type of content might fall into the category of derivative works. The use and sharing of such derived content would be legitimate if permission on transformation of the prior work and making such derivative work available to the public is obtained from the original copyright owner.

Some platforms try to address this situation by making the user guarantee that he has the rights to share the content, which he introduces. Such a provision, by which a user represents and warrants that he has obtained all necessary rights and licenses required to allow posting of any content posted by the user [16], may be found in the terms of some platforms. However, though such clause may have effect and be enforceable under the US law, it may not survive the control of general terms and conditions provided for under the German law [27]. Instead, the inclusion of such a clause into the terms of a service provider established in Germany (or also in other jurisdictions) needs to be considered on a case-by-case basis.

As noted earlier, platform operators usually get a content license from the user by incorporation of such content license into the platform terms. The user, at the time of registering an account or using the platform services, accepts the terms - and by so doing grants rights on use of his content to the platform - [16]. In the absence of other plausible options, this approach may also be extended to other service providers who intend to carry the user´s content on their services.

Regarding the scope of the license, as we saw, platform operators typically acquire the rights, which they consider necessary to provide their services. As a rule, a non-exclusive, royalty free, worldwide, non-assignable license to copy, reproduce, display, transmit, distribute, post, publish, modify, produce derivative works, make the content available to the public in the media, in the form and via distribution methods, known and later developed is specified [16]. Such scope of rights may also be considered as sufficient for third party services.

However, it is not advisable to copy this scope of license verbatim, because a license term, which can have validity for the US based platform, may have no legal effect for a service provider established elsewhere. It may be noted, that the terms of most platforms, including Twitter, Fitbit and Withings, are governed by US law. Whereas the content license, which allows exploitation of content "*in any media now existing or hereafter developed*" [16] granted in this form, i.e. via clicking the "Accept" button, may have effect and be enforceable under the US law, this may not be the case under the national law of some EU member states. Thus, the German Copyright Act, Article 31a, requires that contracts concerning unknown types of exploitation be made in writing [28]. Also, under UK copyright law agreements as to future ownership of copyright would only be enforceable if evidenced in writing [24]. In this regard, the UK Copyright, Designs and Patents Act, section 91 (1), provides that agreements in relation to future copyright be made in writing. *"Future copyright"* in this context means "*copyright which will or may come into existence in respect of a future work or class of works or on the occurrence of a future event"* [29]. Thus, for such a license to be enforceable in Germany or the UK, it would need to be signed via handwritten or e-signature by authorized representatives of the parties, which can hardly be expected in a license agreement concluded online.

From this observation, it may be noted, that a form, in which one or another type of copyright license needs to be obtained in order to have legal effect, should be considered on a case- by-case basis and depending on the jurisdiction where the service provider is established.

## VII. Summary and Recommendations

In this paper, we have described some core legal issues associated with content sharing on digital services and by linking the platforms and apps, in particular.

To summarize the main points, a service provider carrying some user-generated content on its platform needs to have an IP license to do so. Platform operators, who host and transmit user-generated content, typically acquire a copyright license from the user who uploads the content. Such a content license is normally included into the platform terms, which the user accepts (and thereby grants a content license to the platform) when the user signs up for the platform services. Normally, it is non-exclusive, non-assignable, worldwide royalty free license with the right to sublicense. The scope of rights normally covers the whole spectrum of copyright relevant actions, which a platform may need to perform for providing its services. The basic rights of reproduction, distribution and communication to the public are typically included.

Third party service providers who intend to exchange data with social platforms via API exchange systems, such as via apps designed to communicate with a platform via API, may obtain the rights on use of the content from the platform or from the user. In cases where the rights on use of the user generated content are incorporated into and granted under the API license agreement (such as is done by Twitter), an external service provider may rely on the content license from the platform (subject to its validity) and does not necessarily have to obtain a separate content license from the user.

In the absence of content license from the platform (and due to the absence or weak legal strength of other alternatives), the remaining option would be to obtain a license on use of the content from the user himself. In this case, a service provider may follow the practice of platform operators, i.e. include the content license into the service terms and make acceptance of the terms by the user a pre-requisite of using the service. However, when following this practice, re-use of the content license verbatim is not encouraged. First, the terms in question may themselves be copyrighted and not be reproduced without authorization of the right holder. Second, a license granted on the terms and in the form, which have legal effect in one jurisdiction may be challengeable and subject to the risk of being declared invalid by the court in another.

As we have seen, there are multiple copyright issues, which are inherent to the sharing of creative content and which service providers need to handle. However, as is also apparent, the methods and means of dealing with such issues may vary depending on the legal and technical background. The issues, which need to be looked at include the following: what type of data is stored and is to be shared with the platform?  Will the user generated content be stored by the platform? Does the platform grant the rights on use of the content via API license agreement or not? What scope of rights is needed for provision of the service? And in what jurisdiction is the service provider established? Therefore, there is no hard rule, which may be considered as applicable and advisable to all service providers. Rather, these matters will need to be assessed as part of arriving at a satisfactory legal solution in each particular case.

## ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union Seventh Framework Program FP7/2007-2013 under grant agreement No 600929.

## REFERENCES

[1] L. Dobusch, “Need for new regulation to enhance creativity in the digital age: The case of user generated content and cultural heritage institutions”, input paper to Baku Conference, First Council of Europe Platfrom Excahnge on Culture and Digitisation “Creating an enabling environment ofr digital culture and for empowering citizens”, July 2014, Baku, Azerbaijan.

[2] MyHealthAvatar, Project, Concept <http://www.myhealthavatar.eu/?page_id=927> 2015.10.26.

[3] Fitbit <http://www.fitbit.com/uk> 2015.12.28.

[4] Withings <http://www.withings.com/eu/de/> 2015.12.28.

[5] Moves App <https://www.moves-app.com/> 2015.12.28.

[6] Directive 95/46/EC of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, Official Journal of the European Communities, No L 281 /31, 23.11.95.

[7] Fitbit, WEB API Documentation <https://dev.fitbit.com/docs> 2015.12.28.

[8] S. Jandt and C. Hohmann, “Fitness- und Gesudheits-Apps – Neues Schutzkonzept für Gesundheitsdaten?”, Kommunikation und Recht, pp. 694-701, November 2015.

[9] A. Dahi, N. Forgó, S. Jensen, and M. Stauch, “Using patient avatars to promote health data sharing applications: perspectives and regulatory challenges”, European Journal of Health Law, available from: <http://www.brill.com/european-journal-health-law> 2016.01.11. Accepted for publication in March 2016.

[10] Directive of 25 October 2011 on consumer rights, amending Council Directive 93/13/EEC and Directive 1999/44/EC of the European Parliament and of the Council and repealing Council Directive 85/577/EEC and Directive 97/7/EC of the European Parliament and of the Council, OJEU, L 304/64, 22.11.2011.

[11] Berne Convention for the Protection of Literary and Artistic Works of 9 September 1886.

[12] CJEU, Judgment of 16 July 2009, Case C 5/08, Infopaq International A/S v Danske Dagblades Forening, Recital 45.

[13] CJEU, Judgment of 7 March 2013, Case C 145/10 REC, Eva-Maria Painer v. Standard VerlagsGmbH, Axel Springer AG, Süddeutsche Zeitung GmbH, Spiegel-Verlag Rudolf Augstein GmbH & Co. KG, Verlag M. DuMont Schauberg Expedition der Kölnischen Zeitung GmbH & Co. KG, Recital 94.

[14] Directive 2001/29/EC of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society, Official Journal of the European Communities, L 167/10, 22.6.2001.

[15] M. Schmidt-Kessel, “Verträge über digitale Inhalte – Einordnug und Verbracuherschutz”, Kommunikation und Recht, pp. 475-483, July-August 2014.

[16] Fitbit Terms of Use <http://www.fitbit.com/uk/terms> 2015.12.28.

[17] Directive 2009/24/EC of 23 April 2009 on the legal protection of computer programs, Official Journal of the European Union, L 111/16, 5.5.2009, Recital 10.

[18] C. McSherry, “Dangerous Decision in Oracle v. Google: Federal Circuit Reverses Sensible Lower Court Ruling on APIs”, Electronic Frontier Foundation, 9 May 2014, available from: <https://www.eff.org/deeplinks/2014/05/dangerous-ruling-oracle-v-google-federal-circuit-reverses-sensible-lower-court> 2016.01.05.

[19] Fitbit API Terms of Service <https://dev.fitbit.com/terms> 2015.12.28.

[20] Twitter Developer Agreement <https://dev.twitter.com/overview/terms/agreement> 2015.09.18.

[21] Twitter Terms of Service <https://twitter.com/tos?lang=en> 2015.12.28.

[22] Withings Website Terms of Use <http://www.withings.com/uk/en/legal/legal-information#/uk/en/legal/website-terms-of-use> 2015.12.28.

[23] Withings API Terms of Use <http://www-media-cdn.withings.com/wysiwyg/legal/2015-Withings-API-Terms-of-Use-VUS.pdf?_ga=1.254361352.1215146296.1418739123> 2015.12.28.

[24] D.Rowland, U.Kohl, and A.Charlesworth, “Information Technology Law“, 4th edition, Routledge, Taylor&Francis Group, pp. 400, 2012.

[25] Civil Code of Russian Federatrion N 51-ФЗ of 30.11.1994 /Гражданский кодекс Российской Федерации (ГК РФ) от 30.11.1994 N 51-ФЗ <http://pravo.gov.ru/proxy/ips/?docbody=&nd=102033239&intelsearch=%C3%F0%E0%E6%E4%E0%ED%F1%EA%E8%E9+%EA%EE%E4%E5%EA%F1+%D0%EE%F1%F1%E8%E9%F1%EA%EE%E9+%D4%E5%E4%E5%F0%E0%F6%E8%E8> 2015.12.28.

[26] C.Reed and J.Angel, “Computer Law”, Sixth Edition, Oxford University Press, pp.352-353, 2007.

[27] German Civil Code in the version promulgated on 2 January 2002 (Federal Law Gazette, p.42, 2909, 2003, p.738, last amended by Article 4 para 5 of the Act of 1 October 2013 (Federal Law Gazette, p.3719), Artricle 307.

[28] Copyright Act of 9 September 1965 (Federal Law Gazette Part I, p. 1273), as last amended by Article 8 of the Act of 1 October 2013 (Federal Law Gazette Part I, p. 3714).

[29] United Kingdom, Copyright, Designs and Patents Act 1988.

# Modelling the Cost of Open Data

Jolon Faichney, Bela Stantic
Yasaman Moaven, Sanjeev Hiremath
School of Information and Communication Technology
Griffith University
Gold Coast, Australia
email: {j.faichney, b.stantic}@griffith.edu.au
email: {yasaman.moaven, sanjeev.hiremath}@griffithuni.edu.au

John Galvin
Organisational Services
City of Gold Coast Council
Gold Coast, Australia
email: jgalvin@goldcoast.qld.gov.au

***Abstract*—The basic principle of *Open Data* is that data should be freely available to the public to use it without restrictions from copyright or other mechanisms of control. Open Data has benefits including improvements in transparency, productivity, integrity, and accountability. However, at what cost do these benefits come? Relatively little work has been done in quantifying the costs of Open Data in comparison to quantifying the benefits. In this paper we provide a case study on the Open Data initiatives within the City of Gold Coast council. We provide a detailed analysis and description of the processes and people involved in opening data sets and provide estimates for the time involved for each participant in the process. We also explore methods to reduce the time and costs involved through the use of automation. By providing cost models for the Open Data process, organisations will be better equipped to formulate and budget for Open Data strategies.**

***Keywords-open data; case study; cost modelling.***

## I. Introduction

Open Data is a broad term that has been described by the Open Data Institute as “accessible at marginal cost and without discrimination, available in digital and machine-readable format, and provided free of restrictions on use or redistribution” [1]. Even though the term, Open Data, is used to describe all forms of Open Data, it is commonly associated with Government Open Data [2].

Open Data provides both economic and non-economic benefits. By making data openly available to the public, there is more transparency within the government providing the potential for reduced levels of corruption [3]. In 2007, $3.2 billion of misused funds were detected in Canada through the use of Open Data [4].

A report by McKinsey Global Institute [5] found that Open Data can unlock $3 trillion in economic value annually across seven sectors including: education, transportation, consumer products, electricity, oil and gas, health care, and consumer finance. In the United Kingdom, publishing data on cardiac arrests has estimated to have reduced mortality rates, which in turn has an economic value of £400 million per annum, an example of both economic and non-economic benefits [6].

Despite the many benefits of Open Data, the processes involved in making data openly available come at a cost. Given the recentness of Open Data, little work has been done in capturing the cost. However, it is important for organisations to understand the costs involved to make strategic decisions in their Open Data strategies in terms of what data will be made available, how it will be published, and how frequently it will be updated.

In this paper, based on an ongoing collaboration between Griffith University and the City of Gold Coast, we investigate the processes involved in opening data and provide a model for estimating the costs involved.

In Section II, we describe the requirements of Open Data in more detail providing an understanding of the deliverables of an Open Data process. We also explore existing attempts at quantifying the cost of Open Data. In Section III, we specifically focus on the City of Gold Coast’s Open Data strategies which we have been working closely with since its inception. In Section IV, we consider the drivers generating demand for Open Data. In Section V, we describe the current process used by the City of Gold Coast to make its data open. In Section VI, we attempt to capture the costs involved in activities, actors, and time in the Open Data process. In Section VII, we look at ways to reduce the cost of the Open Data process through automation. In Section VIII, we discuss the results from our investigation into the cost of Open Data. In Section IX, we provide conclusions and directions for future work as a result of this study.

## II. Background

In this section, we describe state-of-the-art definitions and standards of Open Data. The requirements of Open Data have an impact on the processes involved in producing it, and hence the cost. The definition of Open Data first begins with the definition of ‘Open’.

### A. Open Definition

The Open Knowledge Foundation provides the Open Definition, now at version 2, as “*Knowledge is open if anyone is free to access, use, modify, and share it  subject, at most, to measures that preserve provenance and openness*” [7]. The Open Definition does not describe how the data is to be made available, but focuses on the policies of the availability of the data. Existing organisations often have a culture where data is not open by default. Therefore, part of the Open Data process is to adopt new policies around openness and educating data custodians to adopt a new culture around Open Data.

### B. Sunlight Foundation Open Data Principles

In 2010, the Sunlight Foundation defined 10 principles of Open Data (extending the previous 8 Sebastopol Principles): Completeness, Primacy, Timeliness, Ease of Physical and

Electronic Access, Machine readability, Non-discrimination, Use of commonly Owned Standards, Licensing, Permanence, and Usage costs [8].

Many of the Sunlight Foundation principles are now covered in the Open Definition 2.0, specifically the last five principles listed above. The first five principles however introduce a burden on the data custodians to ensure that the data they provide is in formats that machines can understand. Providing data in raw, primal, machine-readable form may at first appear simple, however rarely do organisations simply export their data in raw format. For example, much data today is stored in relational tables and simply exporting it would introduce problems such as interpreting the internal schema and exposing private fields. In reality database views must be constructed to produce the Open Data. However, if the data is already made available publicly, for example in PDF form, it is possible that the database views used to generate the data in the PDF will already exist and can be used for the export.

### *C. 5-Star Linked Data*

Based on our experience, raw, unprocessed data can make Open Data less accessible [9]. Tim Berners-Lee introduced the 5-star Linked Open Data framework with an emphasis on technical accessibility [10]. Each level makes the data more accessible to applications. The five levels of the Linked Open Data framework are shown below:

1) Make the data available on the web in any format with an open license.
2) Make it available as structured, computer-readable data (not in image or PDF formats).
3) Use non-proprietary formats such as CSV and XML.
4) Use URIs within data so that other websites can point to resources
5) Link data to other data to provide context.

Berners-Lee's focus on linked data is related to his work on the semantic web [11]. The requirement to provide URIs within data which point to other resources and provide context creates another burden for Open Data providers.

### *D. Open Data Accessibility Framework*

Based on their work with the City of Gold Coast, Faichney and Stantic [9] proposed the Open Data Accessibility Framework (ODAF), which can be seen as an expansion of the third level of the 5-star Linked Data. In our experience it is more useful for Open Data consumers to improve the technical accessibility of Open Data than providing linked data. The ODAF is described using the following six criteria:

1) Resource Naming.
2) Data Coalescing.
3) Data Filtering.
4) Data Consistency.
5) Data Formats.
6) API Accessibility.

The above criteria improve usability of the Open Data for Open Data *consumers* but places an extra burden on the Open Data *providers*.

### *E. ODI Certificates*

The Open Data Institute (ODI) has developed the Open Data Certificates [12] which combine the Sunlight Foundation Principles and 5-star Linked Data frameworks into four levels of Open Data access, which are:

> **Raw –** A great start at the basics of publishing open data.
> **Pilot –** Data users receive extra support from, and can provide feedback to the publisher.
> **Standard –** Regularly published open data with robust support that people can rely on.
> **Expert –** An exceptional example of information infrastructure.

The Expert level technical requirements can be summarised as follows:

- Provide database dumps at dated URLs,
- provide a list of the available database dumps in a machine readable feed,
- statistical data must be published in a statistical data format,
- geographical data must be published in a geographical data format,
- URLs as identifiers must be used within data,
- a machine-readable provenance trail must be provided that describes how the data was created and processed.

### *F. Quantifying the Cost of Open Data*

As can be seen in the previous subsections a lot of work has been done in determining the requirements of Open Data, and providing mechanisms to evaluate and rate the quality of Open Data, primarily with the Open Data consumer in mind. However, how much will it cost the Open Data producers to fulfil the preceding requirements?

The Open Data Institute has identified that there are costs associated with technical work, administration and governance, and building skills capacity [13]. However, no attempts were made at quantifying the costs.

The Transit Co-operative Research Program (TCRP) conducted a survey of 60 respondents working with transit data and reported a broad range of hours required to work on Open Data [14]. The survey identified the following types of costs associated with Open Data:

- Staff time to update, fix, and maintain data as needed
- Internal staff time to convert data to an open format
- Staff time needed to validate and monitor the data for accuracy
- Staff time to liaise with data users/developers
- Web service for hosting data
- Publicity/marketing
- Consultant time to convert data to an open format

## III. Case Study: City of Gold Coast

In this paper we investigate the costs of opening data through a case study with the City of Gold Coast Council, located within the state of Queensland, Australia. The City of Gold Coast is the second largest council in Australia. In this section we provide an overview of the City of Gold Coast's Open Data strategy.

In 2013, the City of Gold Coast appointed an Enterprise Architect with the purpose of implementing an Open Data strategy. Their commitment to Open Data was also demonstrated by sponsoring the GovHack Gold Coast competition in 2013. GovHack is a national hackathon organised by the federal government. The City of Gold Coast have since sponsored GovHack in 2014 and 2015.

In addition to implementing an Open Data strategy the City of Gold Coast supported and sponsored three apps developed by Griffith University which utilise Open Data: Access GC, GC Dog Parks, and GC Heritage. The three apps all utilise geospatial Open Data integrated with other data sets. Griffith University's work with City of Gold Coast Open Data led to the development of the ODAF presented in the previous section.

In 2015 a new Enterprise Architect for Open Data was appointed initiating increased collaboration with external organisations. For example they are active participants of the ODI Queensland branch and hold regular Open Data Working Groups for the Gold Coast region. The City of Gold Coast's philosophy is *Open by Default*, a concept promoted by ODI. The work in Open Data is broadening to now include Smart Cities, recently signing a Letter of Intent with the Open and Agile Smart Cities initiative.

The City of Gold Coast Open Data is published on the data.gov.au national data portal hosted by the federal government. The City of Gold Coast has published 61 data sets and is ranked 5th in Australia according to the Open Data Census [15].

In the following sections we detail the processes involved to make a data set open and then identify the costs associated with the process.

## IV. Sources of Demand

The concept of Demand-Driven Open Data (DDOD) has recently been promoted by the US Department of Health and Human Services (HHS) as the main driver for opening data [16]. The purpose of DDOD is to create value for the 'customer'. The process is managed with *use cases*, which define a clear and concise definition of a desired outcome.

In the City of Gold Coast, three sources of demand for Open Data have been identified, as shown in Figure 1:

1) External Entities
2) Business Users
3) Open Data Team

As in DDOD, external entities may make requests for Open Data. However, so far this has represented only a small portion of requests for Open Data. The majority of requests have come from the Open Data Team themselves. The Open Data Team conducted a survey where participants indicated their interest in data sets listed in the information register of publicly available data sets and ranked the data sets by interest. The information register of publicly available assets existed before the Open Data initiative within the council, however it is worth noting that even though the data was 'publicly' available, it was not necessarily 'open data' in terms of being available electronically and in a machine readable formats. The Open Data Team has been progressively releasing data based on the demand indicated from the survey.

Finally the Business Users, i.e., people within a Business Unit within the organisation, may make a request for Open Data themselves. This may be motivated by a reduction in costs associated with the existing process of other entities requesting data. By making the data open, the costs of managing that process will reduce.

## V. High-Level Open Data Process

The process for opening data is outlined in Figure 2. The Open Data process begins with the Business User making a request for a data set to be opened. Note that the Business User's request may have been initiated by one of the three sources of demand in the previous section. It is also important to note that the Business User is the custodian of the data.

### A. Request Data Publication

The Business User begins by making a request for a data set to be opened. This may either occur electronically via email to the Open Data Team or may involve interaction with a Business Relationship Officer. It is important to capture the possible interaction of the Business Relationship Officer in the request process in terms of determining the cost of Open Data.

The Open Data Team receives a request for Open Data which includes details on the types and forms of data required. After reviewing the data set the Open Data Team may elect to agree to publish the data.

The remaining flow is determined by how the data is stored:

1) Relational Database
2) Geospatial Information System
3) Spreadsheet or other file format

### B. Opening Relational Data

Data in relational databases is relatively easy to publish once a database view has been created. Creating the database view however may be challenging as it can involve complex SQL queries that may cross multiple tables and databases. In the City of Gold Coast a large portion of the business data is stored within SAP databases.

### C. Opening GIS Data

In the City of Gold Coast there currently isn't a mechanism to create a 'view' of GIS data in the same way as relational databases. As a result the data must be exported from the GIS system as a file in a common GIS format such as KML, SHP, or GeoJSON.

### D. Opening File Data

Other data may be stored in individual files, such as spreadsheets. These files can be published as is. However, if they need to be modified this will be a resource intensive phase, generally more so than the publication of database or GIS data.

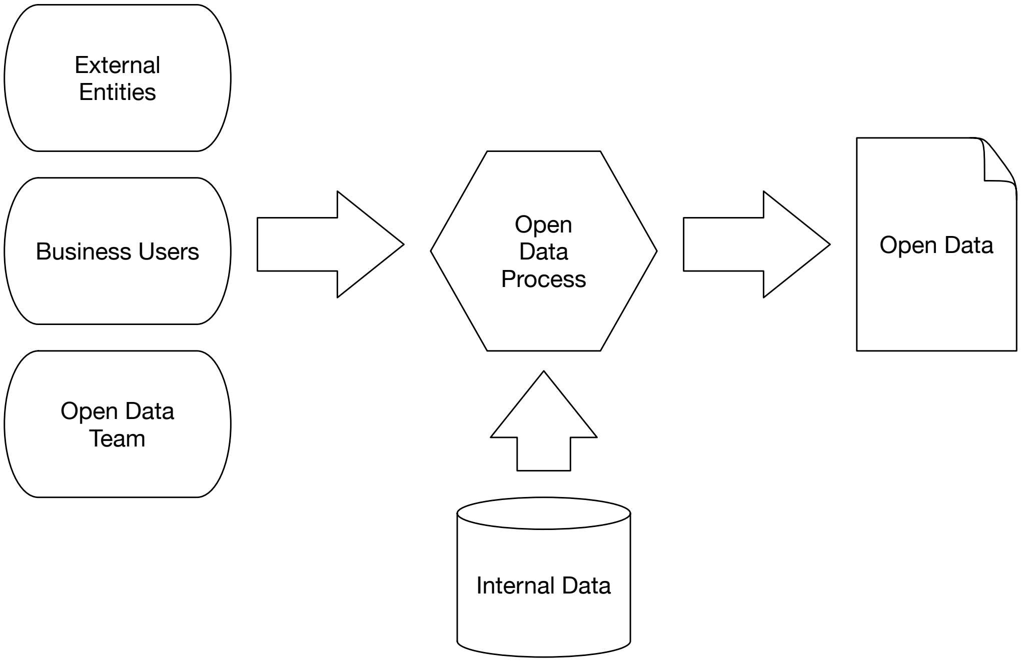


Figure 1. Sources of demand in the Open Data process.

### E. Privacy Concerns

Data may need to be de-identified to ensure privacy policies are not breeched. This may involve the removal columns or tags from data sets, or only releasing aggregate data views.

### F. Test Data Review

Before publication the Business User is sent a sample extract of data to be published. The Business User confirms whether the extracted data is correct.

### G. Automation

If data is to be released periodically, an automation process can be established. Currently in the City of Gold Coast data publication is only automated if it is updated more frequently than yearly. Database views are relatively simple to automate. GIS and file-based data currently still involves human intervention. Automation is discussed in more detail in Section VII.

### H. Approve for Publishing

Once the automation process is implemented, the Business User may approve the data for publishing. Data is currently published to the national Open Data portal data.gov.au. It is made available on data.gov.au initially with private access to ensure the processes are working correctly. Once the Business User approves the publication of data, the data set is made public by the Open Data Team.

## VI. Cost of Open Data

In this section we aim to model the cost of the Open Data process. The main cost in the Open Data process is staff time. Since the cost of staff time varies between organisations, cities, and countries, we will model our costs as a proportion of a staff member's time. Table I shows our estimate for the number of full-time equivalent (FTE) days spent for a single data set. Note that the total of 6.5-16 days is not the wall clock time required to release a data set as some work can be performed in parallel and multiple data sets can be released simultaneously if multiple staff members exist on the team, likewise the wall clock time may be longer if there are delays in the process such as organising meetings at a future date.

The City of Gold Coast has one member in the Open Data Team being the Enterprise Architect for Open Data. The Enterprise Architect has other responsibilities, such as developing strategies and policies for Open Data and engaging with the community. This limits the amount of time dedicated to releasing new data sets. Approximately 40% of the Enterprise Architect's time is dedicated to releasing data sets. The above table indicates that 1-3 days are required to release a data set, which approximately correlates with the current output of around 40 data sets a year by the City of Gold Coast.

The ranges provided indicate variations in complexity. Data sets which involve more files will take longer to publish. Each data set published to the data portal requires the data set to be registered and a unique key provided which is used for subsequent updates to the file. The complexity will also depend on the data. Database views are generally the most complex to formulate. GIS data exports are often simpler as the required

TABLE I. Days required to release a single data set.

| Actor | FTE Days/Data Set |
|---|---|
| Open Data Team | 1-3 days |
| Business Users | 2-5 days |
| Business Relationship Officer | 0.5-1 day |
| Analyst | 2-5 days |
| Integration Analyst | 1-2 days |
| **Total** | **6.5-16 days** |

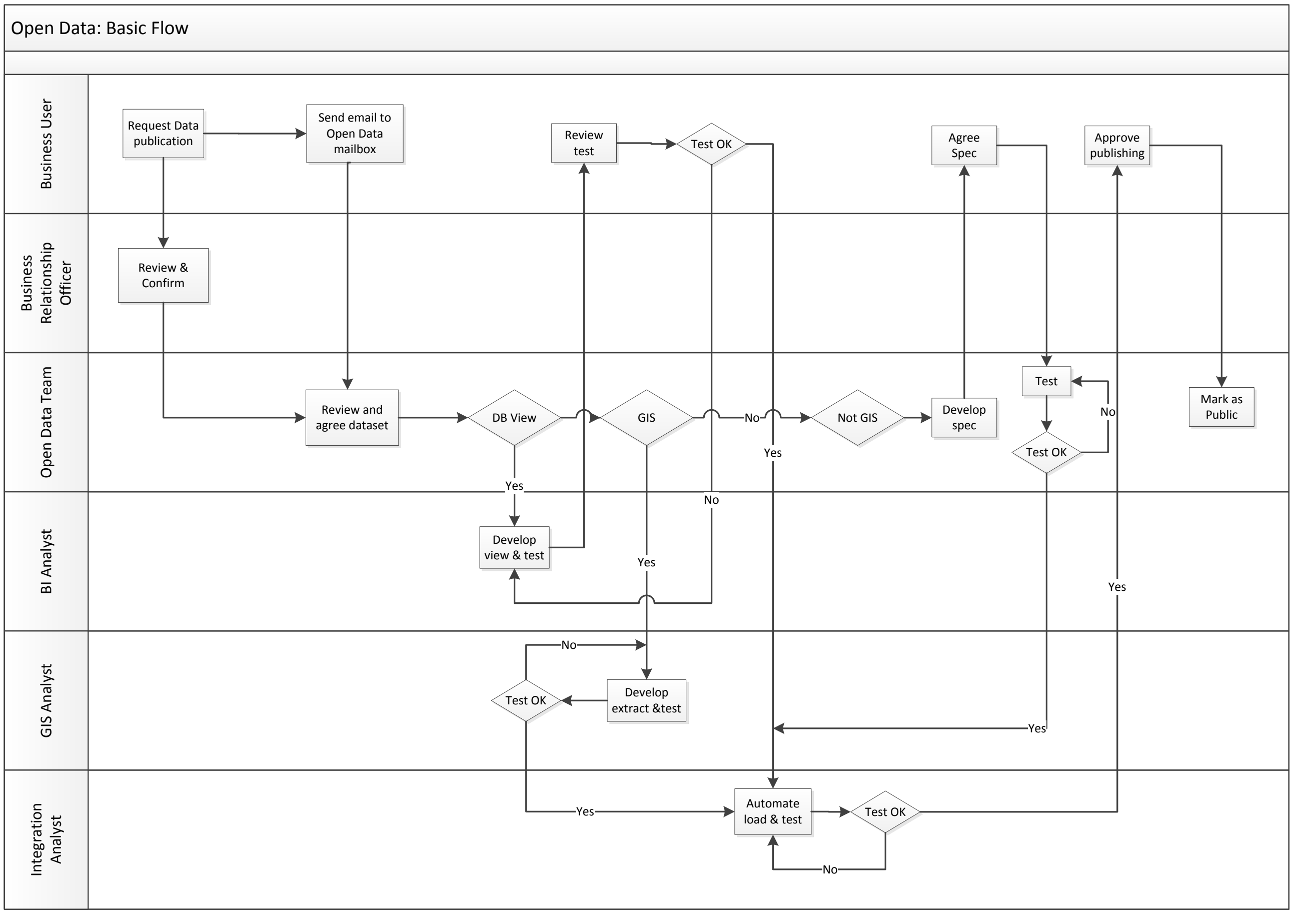


Figure 2. Process used by the City of Gold Coast to open existing data sets.

data exists in a layer and can be exported in its entirety.

## VII. Automation

Automation can be utilised to reduce the cost of releasing periodic data and can also reduce human errors that can be introduced when repeatedly releasing data. The City of Gold Coast has two primary mechanisms for automating the release of data:

1) Database Views
2) Automatic File Upload

Both approaches upload data to the data portal at regular intervals. The data portal utilises the CKAN content management system. The first approach is completely automated. Database views are generated and the results uploaded to the data portal via a script. GIS and other file data is currently produced manually resulting in a file to be uploaded to the data portal. To simplify this process, a network directory is monitored, when a file is copied to the network directory it is automatically uploaded to the respective section of the data portal. This reduces the time required by the staff that administer the data portal.

Automation is able to reduce the cost of releasing data on an ongoing basis. However it will require further time upfront to establish the automation process. This additional time requires the Integration Analyst to implement and test the automation procedure and the Business User to confirm that it is working.

Some of the automation procedures can be reduced across data sets. The City of Gold Coast spent 20 days building their current automation system.

## VIII. Discussion

As can be seen in the previous sections, Open Data has a cost. Do the benefits of releasing Open Data outweigh the costs? The literature so far indicate that the benefits of Open Data outweigh the costs, this conclusion has been determined by estimating the overwhelming benefits without providing finer grained analysis of the processes and costs involved in releasing Open Data. In this paper, we have looked at staff time which correlates with a financial cost and can be evaluated against a financial benefit. However there are other non-financial benefits to releasing Open Data such as transparency and social benefit. Can we evaluate the non-financial benefits against the financial costs? We don't think it is necessary to draw a connection between the financial cost of Open Data and the non-financial benefits. Modelling the cost of Open Data is sufficiently important for organisations to help plan their Open Data strategies.

## IX. Conclusions and Future Work

In this paper we have reported our collaboration with the City of Gold Coast in capturing the costs involved with releasing Open Data. Some studies have reported the macro-economic benefits of Open Data, but relatively little has been done in capturing the cost. In this paper we report the processes currently used by the City of Gold Coast and estimate the roles involved in opening data and the FTE time required by staff. This will help organisations budget and plan for Open Data rollouts and transitions.

Further work will investigate in greater detail the causes of variations in complexity in releasing Open Data, such as the type of data (database, GIS, file), the number of files within the data set, additional data processing, and the time required to deal with cultural resistance to releasing data openly.

Additional work will also investigate how various automation techniques can be used to further reduce the cost of Open Data, particularly in the cases of GIS and spreadsheet-based data.

## Acknowledgment

The authors would like to thank the Open Data Working Group of the City of Gold Coast Council for their collaboration and support during this project.

## References

[1] M. Heimstädt, F. Saunderson, and T. Heath, "From toddler to teen: Growth of an open data ecosystem." eJournal of eDemocracy & Open Government, vol. 6, no. 2, 2014, pp. 123–135.

[2] J. Kloiber, "Open government data - between political transparency and economic development," Master's thesis, Utrecht University, 2012.

[3] N. Rajshree and B. Srivastava, "Open government data for tackling corruption-a perspective," in Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012, pp. 21–24.

[4] D. Eaves, "Case study: How open data saved Canada $3.2 billion," 2012, retrieved: January, 2016. [Online]. Available: http://eaves.ca/2010/04/14/case-study-open-data-and-the-public-purse/

[5] J. Manyika et al., "Open data: Unlocking innovation and performance with liquid information," McKinsey Global Institute, Tech. Rep., 2013.

[6] Deloitte, "Market assessment of public sector information," UK Department for Business Innovation and Skills, Tech. Rep., 2013.

[7] Open Knowledge Foundation, "Open definition 2.0," 2014, retrieved: January, 2016. [Online]. Available: http://opendefinition.org/od

[8] Sunlight Foundation, "Ten principles for opening up government," 2010, retrieved: January, 2016. [Online]. Available: http://sunlightfoundation.com/policy/documents/ten-open-data-principles/

[9] J. Faichney and B. Stantic, "A novel framework to describe technical accessibility of open data," in The First International Conference on Big Data, Small Data, Linked Data and Open Data (ALLDATA), 2015, pp. 52–57.

[10] T. Berners-Lee, "Is your linked open data 5 star?" 2010, retrieved January, 2016. [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.html

[11] T. H. C. Bizer and T. Berners-Lee, "Linked data – the story so far," International Journal on Semantic Web and Information Systems, 2009, pp. 1–22.

[12] The Open Data Institute, "Open data certificates," 2013, retrieved: January, 2016. [Online]. Available: https://certificates.theodi.org

[13] ——, "Estimating the cost of a government open data initiative," 2014, retrieved: January, 2016. [Online]. Available: https://theodi.org/blog/estimating-the-cost-of-a-government-open-data-initiative

[14] C. L. Schweiger, "Open data: Challenges and opportunities for transit agencies, a synthesis of transit practice," Tech. Rep., 2015.

[15] O. K. Foundation, "Local open data census for Australia," 2015, retrieved: January, 2016. [Online]. Available: http://au-city.census.okfn.org

[16] D. Portnoy, "Identifying and harnessing demand to drive open data," 2015, retrieved: January, 2016. [Online]. Available: http://www.hhs.gov/idealab/2015/03/16/identifying-harnessing-demand-drive-open-data/

# Link Detection Based on Named Entity Keywords in Turkish News Corpus

Hamid Ahmadlouei, Hayri Sever
{hamid,sever}@hacettepe.edu.tr
Department of Computer Engineering
Hacettepe University
Ankara, Turkey

Erhan Mengusoglu
emengusoglu@thk.edu.tr
Computer Engineering
University of Turkish Aeronautical Association
Ankara, Turkey

***Abstract*— In this study, we investigate the influence of Named Entities (NEs) on the task of Story Link Detection (SLD), which is one of the important subtasks in Topic Detection and Tracking (TDT). TDT aims at developing algorithms for either clustering documents, e.g., online news, and then tracking new ones with respect to a predetermined topic or otherwise detecting a new topic. Furthermore, SLD focuses on determining whether the stories are about the same topic. Vector Space Model (VSM) was used as a base method in this work for "All-words" and Named Entities (NE) separately. We also investigated the effect of controlled entity intersection on the performance of previous VSM based methods. Combining these methods provided improvement in correctly estinating whether the stories are linked or not.**

***Keywords*— story link detection; topic detection and tracking; vector space model; information retrieval; named entity.***

## I. INTRODUCTION

In recent years, there have been a growing number of online news sources. Having many options might be attractive for the user, but, on the other hand, the user might spend a substantial amount of time surfing the Internet in search for the needed information. If proper information retrieval (IR) techniques are used, helping the user reach the needed information becomes an easier task. So, in order to manage the huge increase in news articles, grouping similar articles and linking similar stories are indispensable steps for IR tasks.

An information retrieval system is composed of a corpus of documents, and access functions with capability of comparing the words of the query terms in user queries, and the terms of the documents in the corpus to determine the relevant documents. At this point, the basic function of information retrieval systems is to access all relevant documents in the corpus and comb out non-relevant ones in order to meet the information needs of users [1]. TDT is one of the most important tasks in the study of IR. Hence, recent academic studies on the Web IR systems mainly focus on the TDT program.

TDT studies aim to organize, identify and follow all kinds of stories published on the Web [2]. To accomplish this goal, TDT studies are divided into five main tasks: story segmentation, topic detection, topic tracking, first story detection and story link detection:

- Story Segmentation: determines story boundaries,

- Topic Detection: determines the subject of the story,

- Topic Tracking: follows a pre-determined story,

- First Story Detection: identifies stories not encountered previously,

- Story Link Detection: determines if two stories are linked or not,

SLD tasks are reported as the most important sub-tasks of TDT studies [3]. The purpose of SLD is to determine whether two independent stories are on the same subject or not [4]. "Story" in this paper is defined as a piece of news about a single "topic" in TDT problems. "Topic" is a seminal event or an activity along with all directly related events and activities. [2]. "Event" is a specific thing that happens at a specific time and place.

In this paper, we analyze story link detection and investigate word based and named entity based techniques. Our purpose is to detect whether two documents are linked or not. We present the performance of two techniques and show the improvement in the performance of a link detection system using a combination of these techniques. After performing the SLD task, the results of this sub-task can be applicable to other sub-tasks of TDT. In this respect, successful determination of whether two stories are on the same topic or not by using SLD, is expected to solve many problems for TDT [5].

This paper presents a combination of different techniques to improve the performance of link detection. A combination of methods in some cases provides an improvement in estimating whether two stories are linked or not. The work is evaluated on the Turkish news corpus, and the experimental results indicate that story link detection using a combination of methods can help obtain a better performance.

We used VSM as the base model in this work. Our experimental results indicate the result of VSM which is inspired by the co-sine similarity concept, is better than alternatives. Word-based (WB) and entity-based (EB) tests of this method are carried out separately.

We also control the intersection of named entities between two stories. Intersection checking is used in order to determine which two different news are on the same topic. Experiments are designed to assess how VSM performance is affected when entity intersection checking is used. We analyzed OR and AND logical combination of VSM with the Named Entity Intersection (NEI) method. Word-based and entity-based scenarios of VSM, OR/AND-logical combination with NEI are carried out separately.

Inspired by the study presented in [6], we defined a simple Named Entity Resemblance Function (NERF) in order to give more importance to the named entities between two stories. To enhance the effectiveness of named entities by simple normalization on naming entities between news articles, the similarity score between two news stories is calculated using the function in [6]. This method was not successful in Turkish news tasks, but it was successful in a Chinese study.

This paper is organized as follows. In Section 2, we give an outline of the related work within the topic. In Section 3, we talk about the methodology used in the paper. In Section 4, we give details of the tests carried out during our experiments and present their results. In Section 5, we present the conclusion by summarizing our contribution. In Section 6, we briefly note future studies that can be carried out as a follow-up for this work.

## II. Related Work

Academic studies in the field of TDT story link detection task require identifying pairs of linked stories. In the story link detection systems which have been developed so far, the best technology for link detection relies on the use of cosine similarity between document terms vectors with Term Frequency - Inverse Document Frequency (TF-IDF) term weighting.

Some academic studies in the SLD field show that relevance models (RM) produce better results than other IR methods [7]. Some works on SLD based on VSM use the cosine similarity measurement between the data streams [8] [9] [10]. UMass has examined a number of similar measures in the link detection task, such as weighted sum and language modeling, and found that the cosine similarity produced the best results [10]. Additionally, a different study by the same authors showed that VSM performed better on the SLD task based on Turkish news [4].

Another point that information retrieval researchers focus on is how to select terms representing documents and weight them effectively. Document representation is an extremely important step in traditional IR systems as well as in TDT studies [11]. Depending on the study areas, word-based and entity-based methods are usually used for the representation of the documents [6] [12] [13].

In SLD task, many methods are used to compare the quantity of the overlapping words within two stories. Large numbers of overlapping words between two stories means there is a higher probability that they discuss the same topic. This approach formed the basis of all the methods that use vector space models [14].

Some studies on TDT task also claim that document representation using only words is not enough [15]. Experiments reported in [16] compared two different stories using named entities taking into account: person (who), location (where), time (when) and action (what) words. In a similar study, name, location and time information in the news are expressed in separate vectors and named entities such as name, place and time are extracted by automatic inference methods. Using the named entities to identify the most recent (newest) news provides a significant performance increase [6]. Researchers in both of those studies proposed similarity metrics based on intersection, especially while comparing the time and place [17].

The use of entity names on the Turkish corpus in order to improve SLD performance was not studied so far. Literature emphasizes that more in-depth studies should be done in this regard [18].

In some studies, a combination of different methods provides improvement for estimating whether two stories are linked. Furthermore, it is also reported in the literature that access performance is increased by using a combination of different methods [4] [6] [19].

Named entities are also used in determining news similarity in order to perform SLD task [18]. Similar studies for Turkish corpus, mainly by using machine-learning methods, that extract named entities (name, place, organization e.g.) automatically from texts are reported in [20] [21].

## III. Methodology

### A. TDT Test Collection

We used new event detection and topic tracking test collection (BilCol-2005), which was developed by the information retrieval group at Bilkent University. The Bilkent information retrieval group aims to develop effective and efficient information retrieval tools, with an emphasis on the Turkish language. The BilCol-2005 test collection comprises news stories from five different Turkish news sources on the Web (both broadcast news and daily news articles): CNN Turk, Haber 7, Milliyet, TRT, and Zaman. More information about BilCol-2005 is provided in [18].

In the BilCol-2005 corpus, 5,883 news stories were classified under 80 different topic titles, while the rest (203,442) has yet to be classified. In this study, tests were carried on the classified news stories.

Most of the studies on Named Entity Recognition (NER) subfield have been done for English, Chinese and Spanish texts. Studies for Turkish language texts have only started recently. Thereby, automatic NER methods in Turkish are still immature. So, in the preparation phase of the dataset, tagging of named entities was carried out manually. As in many IR systems, most studies focus on which words to select as named entities or keywords and how weighting of these keywords has to be done as well as how these weighted keyword will most effectively be compared [22] [23]. In this context, named entities in the dataset are tagged with the following labels: *"Person"*, *"Location", "Organization", "Time", "Date", "Percentage", "Money", "Unknown"*. The label "*unknown*" is used for tagging all entities in the text when tagging with the other labels above was not possible.

Here is an example of a labeled entity:

*<Person>Shakespeare<Person>*

*<Location>Ankara<Location>*

*<Organization>Galatasaray<Organization>*

*<Date>1992<Date>*

### B. Evaluation Methodology

The performance is measured by obtaining *precision*, *recall* and *F-measure* values for each test. *Recall* is the proportion of retrieved relevant documents to total related documents and *precision* is the proportion of accessed relevant documents to total accessed documents. *F-measure* identifies the harmonic mean of precision and recall [24]. These three values are expressed mathematically in the following equations:

$$\Pr ecision = \frac{number-of-accessed-relevant-document}{number-of-accessed-document} \tag{1}$$

$$\text{Re}call = \frac{number-of-accessed-relevant-document}{total-number-of-relevant-document} \tag{2}$$

$$F-Measure = \frac{2 * \Pr ecision * \text{Re}call}{\Pr ecision + \text{Re}call} \tag{3}$$

In this paper, we assumed that high *precision* and high *recall* or higher *F-measure* values represent better results.

Studies on the combination of different methods generally increases the values of recall, yet, at the same time, retrieves a lot of unrelated documents, thereby decreasing precision values and degrading the overall system performance. Therefore, it is extremely important to develop combined models that would provide the best possible values for both precision and recall.

### C. SLD Methods Used In The Study

Different types of documents may have different retrieval characteristics. Text retrieval methods are typically designed to find documents relevant to a query based on some criterion, such as cosine similarity. We used vector space model (VSM) as a base method for the Turkish corpus. The vector space model developed in the late 1960s is still a very popular approach and is commonly used in IR systems as a retrieval function [25]. Although this method has been widely used for SLD task of TDT studies, there is, to the best of our knowledge, no study that was carried out to apply it on a Turkish corpus [2][3].

VSM calculates the similarity between compared documents based on common term conflicts. So, we analyze the word-based and entity-based approaches in the first two steps of the experiments.

The steps used in our experiment are:

1- *VSM Word Based (WB)*

2- *VSM Entity Based (EB)*

3- *Named Entity Intersection (NEI)*

4- *VSM (WB) OR NEI*

5- *VSM (WB) AND NEI*

6- *VSM (EB) OR NEI*

7- *VSM (EB) OR NEI*

8- *Named Entity Resemblance Function (NERF)*

VSM is used with words and entity based approach as an access function. In information retrieval systems, which use this method, each document is shown as a vector of a collection of $t_1$, $t_2 ... t_n$ single words. Coefficient values of $t_1$, $t_2, ... t_n$, are determined based on the number of times that related word appears in the collection ($t_i$).

In traditional IR methods, a general approach for representing the vector coefficients is identified as the *idf-weighted cosine coefficient* and is shown as *tf.idf* (*term frequency*, *inverse document frequency*). Similarity between the two vectors (*a* and *b*) is calculated by applying Equation 1 where $tf_a\ (w)$ represents the frequency of word *w* in the document *a*, $tf_b\ (w)$ represents the frequency of word *w* in document *b*, and *idf (w)* represents the frequency of word *w* in all documents in the corpus.

$$sim(a,b) = \frac{\sum_{w=1}^{n} tf_a(w).tf_b(w).idf(w)}{\sqrt{\sum_{w=1}^{n} tf_a^2(w)}.\sqrt{\sum_{w=1}^{n} tf_b^2(w)}} \tag{4}$$

In the second step, by using the determined entity names, we created entity vectors for each article document. By applying Equation 4, we can calculate the similarity between the entity vectors. With this method, the biggest challenge is that some documents may not have enough named entities for creating a good quality entity vector. If the vector created is very sparse, it will not be good enough to make comparisons.

To resolve this problem, in the third step, we controlled the named entity intersection between the stories. When two news stories are compared and even if only one entity intersection is found, then these two news stories were determined as linked. In this step, the entity intersection control is examined by determining whether these two news stories are on the same topic or in different topics.

In the first three steps of the experiments, we illustrated the realization task of SLD with independent decisions of these methods separately. However, in the next four steps (*4,5,6,7*) we made judgments using the OR-AND logical operators to obtain combined decision results for these methods. Thus, we had the chance to catch the relevant missed documents with

VSM by using the Named Entities methods [34]. So, in the fourth step, independent decisions about VSM word-based and NE intersection were carried out and coupling was done with OR logical operator. Following this, in the next step (fifth) we performed AND-logical combination of VSM (WB) and NE intersection methods. In the sixth and seventh steps, similar to two previous steps (*4,5*) experimental tests were carried out between entity-based VSM and NE intersection checking method.

In the final step, we used a resemblance function to calculate the similarity between two news stories using only named entity. Actually, by normalizing common entities between the stories, this function calculates the similarity score between news stories based on named entities. In order to reach the similarity score between news stories based on named entities, and also to emphasize the importance of the named entities in comparison, we used the resemblance function [5]. The resemblance function between two documents *a* and *b* is defined as follows:

$$f(a,b) = \frac{|a \bigcap b|}{|a \bigcup b|} \qquad (5)$$

In Equation 5, the numerator is the number of entities common in *a* and *b* and the denominator is the number of all entities in *a* and *b*.

## IV. Testing and Results

The dataset is divided into training (one third of the news stories) and test (two thirds of news stories) sets. The news stories in the training set are used to determine the threshold parameter value. In this respect, the threshold value of the VSM method was defined as the optimum point where recall and precision become equal. Tests were carried on the corpus which includes 3,922 news stories with known topic titles that were not used in the training set. During testing, news stories with known topic titles were compared with the rest of the news stories in the test set. Furthermore, logical OR/AND operators were applied to the results obtained using VSM and NE methods so that the effects of OR/AND operators on precision and recall measures could be evaluated. To determine the overall performance, precision, recall and F-measure values were also calculated. In the course of testing, Turkish stop-words were removed and also stemming was applied.

## V. Discussion and Conclusion

This paper analyzes how well the performance of the VSM is in the SLD task. Our purpose in this work was to detect whether two documents are linked or not. This paper presents a combination of different word-based and entity-based techniques to improve the performance of link detection. A combination of methods is shown to provide improved estimation performance in some cases.

The findings obtained using all methods are presented in Table I (*P: Precision, R: Recall, F: F-Measure, T: Threshold*). As the results indicate, the combinations using Boolean OR operator of VSM word-based and VSM entity-based with Named Entities intersection methods resulted in substantial improvements in performance.

The highest performance is obtained using the OR combination of VSM entity-based and NE intersection which was obtained with an F-measure value of 0.90 (recall: 0.84 and precision: 0.98). This combination achieved a substantial 30% increase in performance compared to the best case with VSM which resulted in an F-measure value of 0.60 (recall: 0.56 and precision: 0.65).

In this work we aimed at developing methods that provide higher precision and recall simultaneously. So, when we analyze the results of Named Entity Intersection (NEI) with a precision value of 0.13 and recall value of 0.81, we understand that this method was not very successful for the SLD task. But, these results can also be interpreted differently as it is possible to conclude that this method is able to determine that two articles are on different topics with a rate of 81%.

TABLE I. TEST RESULTS

| **Method** | **P** | **R** | **F** | **T** |
|---|---|---|---|---|
| VSM (WB) | 0.61 | 0.58 | 0.59 | 0.05 |
| VSM (EB) | 0.65 | 0.56 | 0.60 | 0.02 |
| NEI | 0.13 | **0.81** | 0.23 | - |
| VSM (WB) *OR* NEI | 0.77 | 0.75 | **0.76** | 0.04 |
| VSM (WB) *AND* NEI | 0.53 | 0.51 | 0.52 | 0.05 |
| VSM (EB) *OR* NEI | 0.98 | 0.84 | **0.90** | 0.01 |
| VSM (EB) *AND* NEI | 0.65 | 0.56 | 0.60 | 0.02 |
| NERF | 0.98 | 0.13 | 0.22 | 0.02 |

In this work, SLD that drew special attention within the TDT research is applied for the first time on a Turkish corpus using Named Entities method using named entities extracted from the text manually. The results clearly show that the VSM performance is substantially affected by the combination of NE methods, in identifying the similarities of news stories.

## VI. FUTURE WORK

Further studies should investigate the effect of Named Entities individually and in different combinations on the retrieval performance for identification. Actually, controlling named entities and identifying individual intersection in order to determine which named entities lead to better performances are two different approaches that can be investigated.

Furthermore, named entities with binary (person and location), triple (person, location, and date) or quad (person, location, date and organization) combination intersections between news can be analyzed. By analyzing these named entities, we were able to detect which combination of named entities is better for the SLD task. In some studies, event based methods are used for SLD detection. “Date” and “Location” named entities are used to extract events. Extraction of events will enable the use of event word-based methods.

In this study, we do not use “query expansion” which is a well-established technique in IR. An important performance parameter in IR applications is the query length, i.e., the number of words used in queries. The effect of query length on retrieval performance based on named entities can also be analyzed for IR application environments. “Query expansion” technique based on non-entity words (all-words) was carried out by the same authors previously in [26] and [27].

## ACKNOWLEDGMENTS

This Work Is Partially Supported By The Scientific And Technical Research Council Of Turkey (TÜBİTAK) Under The Grant Number 111k030.

## REFERENCES

[1] Tonta, Y., Bitirim, Y., and Sever, H. (2002). Türkçe arama motorlarında performans değerlendirme. Total Bilişim.

[2] Allan, J. (2002). Introduction to topic detection and tracking. In *Topic detection and tracking* (pp. 1-16). Springer US.

[3] Lavrenko, V., Allan, J., DeGuzman, E., LaFlamme, D., Pollard, V. and Thomas, S. (2002, March). Relevance models for topic detection and tracking. In *Proceedings of the second international conference on Human Language Technology Research* (pp. 115-121). Morgan Kaufmann Publishers Inc..

[4] Kose, G., Tonta, Y., Ahmadlouei, H., and Polatkan, A. C. (2013, November). Story Link Detection in Turkish Corpus. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on* (Vol. 1, pp. 154-158). IEEE.

[5] Allan, J., Carbonell, J. G., Doddington, G., Yamron, J. and Yang, Y. (1998). Topic detection and tracking pilot study final report.

[6] Letian Wang and Fang Li, Story Link Detection Based on Event Words, Springer-Verlag Berlin Heidelberg 2011

[7] Allan, J., Lavrenko, V. and Swan, R. (2002), Explorations Within Topic Tracking and Detection, Topic Detection and Tracking: Event-based Information Organization, J. Allan, Ed., Kluwer Academic Publishers, pp. 197-224.

[8] Chen, F., Farahat, A., Brants, T., Multiple similarity measures and sourcepair information in story link detection. Presented in the Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004), Boston, Massachusetts, pp. 313–320, 2004.

[9] Allan, J., Lavernko, V., Nallapati, R. UMass at TDT2002. Presented in the Proceedings of the Topic Detection and Tracking Workshop, 2002.

[10] Allan, J., Lavrenko, V., Frey, D., Khandel Wal, V. UMass at TDT2000. Presented in the Proceedings of Topic Detection and Tracking Workshop, 2000.

[11] Chirag, S. and Koji E., (2009). Improving Document Representation for Story Link Detection by Modeling Term Topicality. IPSJ Online Transactions

[12] Shah, C., Croft, W. B. and Jensen, D. (2006). "Representing Documents with Named Entities for Story Link Detection (SLD)," a poster presentation at the ACM Fifteenth Conference on Information and Knowledge Management (CIKM) 2006, Arlington VA, November 6-11, 2006.

[13] Tadej Š. and Marko Grobelnik, (2009) Story Link Detection With Entity Resolution. ACM, Madrid, Spain

[14] Schultz, J. M. And Liberman, M. Y. (2002). Towards a “Universal Dictionary” for multi-language information retrieval applications. In *Topic detection and tracking* (pp. 225-241). Springer US.

[15] Makkonen, J., Ahonen-Myka, H. and Salmenkivi, M. (2003). Topic Detection and Tracking with Spatio-Temporal Evidence. In Proceedings of 25th European Conference on Information Retrieval Research (ECIR 2003). 251-265.

[16] Kumaran, G. and Allan, J. (2005). Using names and topics for new event detection. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 121-128). Association for Computational Linguistics.

[17] Makkonen, J., Ahonen-Myka, H. and Salmenkivi, M. (2002). Applying Semantic Classes in Event Detection and Tracking. Proc. International Conference on Natural Language Processing (ICON'02). 175-183.

[18] Can, F., Kocberber, S., Baglioglu, O., Kardas, S., Ocalan, H. C. and Uyar, E. (2010). New event detection and topic tracking in Turkish. *Journal of the American Society for Information Science and Technology*, *61* (4), 802-819.

[19] Yang, Y., Carbonell, J., Brown, R., Lafferty, J., Pierce, T. Ault, T. (2002). Multi-strategy learning for topic detection and tracking. In J. Allan (Ed.), Topic Detection and Tracking: Event-based Information Organization (pp. 85-114). Norwell, MA: Kluwer Academic Publishers.

[20] Dalkiliç, F.E., Gelisli, S. and Diri, B. (2010). “Türkçe Kural Tabanli Varlik Ismi Tanima”, (Turkish Rule Based Assets Recognition) 18. Sinyal Isleme ve Uygulama Kurultayi, Diyarbakir, (22-24 Nisan) 2010.

[21] Küçük, D. and Yazici, A. (2010). A Hybrid Named Entity Recognizer for Turkish with Applications to Different Text Genres. In Proceedings of the 25th International Symposium on Computer and Information Sciences (ISCIS). London, UK. E. Gelenbe et al. (Eds.): Computer and Information Sciences, LNEE 62, pp. 113-116.

[22] Xianshu Z. and Tim O., (2013), Finding News Story Chains Based on Multidimensional Event Profile.OAIR2013, Lisbon, Portugal

[23] Hua Zhao1 and Tiejun Zhao, (2009). Applying Dynamic Co-occurrence in Story Link Detection. Journal of Computing and Information Technology

[24] Rennie, J. D. M. (2008). Derivation of the F-measure, 2004. URL http://people. csail. mit. edu/jrennie/writing/fmeasure. pdf. accessed 15 May 2013.

[25] Nomoto, T. (2010). Two-tier similarity model for story link detection. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 789-798). ACM.

[26] Raghavan, V. V. and Sever, H. (1995). On the Reuse of Past Optimal Queries, Proceedings of 18th ACM International Conference on Research and Development in Information Retrieval (SIGIR'95), Seattle, WA, USA, July 1995, pp. 344-351.

[27] Manmatha, R. and Sever, H (2002). A Formal Approach to Score Normalization for Metasearch, Human Language Technology Conference (HLT'02), March 24-27, 2002, San Diego, CA.

# Using Domain Ontologies for Classification and Semantic Interpretation of Documents

Samia Iltache
UMMTO
Tizi Ouzou, Algeria
e-mail: s_iltache@hotmail.com

Catherine Comparot
IRIT, Université de Toulouse,
CNRS, INPT, UPS, UT1, UT2J
France
e-mail: Catherine.Comparot@irit.fr

Malik Si Mohammed
UMMTO
Tizi Ouzou, Algeria
e-mail: m_si_mohammed@esi.dz

Pierre-Jean Charrel
IRIT, Université de Toulouse,
CNRS, INPT, UPS, UT1, UT2J
France
e-mail: Charrel@univ-tlse2.fr

*Abstract*—**The work presented in this paper addresses the problem of interpretation and semantic classification of documents. One of the issues faced by natural languages is related to the presence, in glossaries, of words with similar morphologies and different meanings. Our approach is based on the use of domain ontologies for nouns disambiguation. We begin our process with a global disambiguation, by linking the considered document to a semantic domain (represented by an ontology) which we select among several candidate ones. We define a candidate domain as any domain in which at least one significant word of the text can be considered and makes sense. We then perform a local disambiguation by using the selected ontology and finally build a semantic representation of the content of the document as a conceptual graph.**

***Keywords-Domain ontology; semantic interpretation; disambiguation; classification; conceptual graph.***

## I. INTRODUCTION

A document is represented by a set of words that expresses its global meaning. In conventional approaches, a document is represented by the lemmas of words describing its contents. To these lemmas is assigned a weight indicating their importance in the document. This weight combines local weighting linked to the document itself and a global weighting based on the considered corpus.

Semantic approaches aim to give meaning to the terms of the document to address the shortcomings of conventional indexing based on single words.

The issue of words with similar morphologies and different meanings is faced in all languages. If the assignment of adequate meaning to a word is easily done by a human being, because he uses his knowledge, this process is made difficult for an application using the textual content of the documents based on the morphological appearance of words.

Our approach aims to achieve an interpretation and semantic classification of textual content of documents. We propose to use the knowledge represented by domain ontologies as a basis for our process. In fact, we consider that concepts of an ontology can allow to give the appropriate meaning to the words of the document. We first perform a global disambiguation through a classification process. This is to determine, among several domain ontologies, which one is the best to be considered, in order to obtain the correct semantic of document content; it is determined by the overall context of the document. In the second step, a local disambiguation is performed if some of the terms can be associated to several concepts within the retained ontology. The process, thus defined, allows to respond to the problem of polysemy and synonymy.

Our approach allows to thematically group the documents and to obtain a semantic representation of their content. A document can then be represented by a conceptual graph extracted from the ontology to which the document has been attached.

This paper is organized as follows. First, in Section II and Section III, we present a brief state of the art by introducing some works related to our problem. Then, in Section IV, we focus on the different steps of our process. In Section V, we present some examples to illustrate our approach. In the last Section, we conclude on the usefulness of our approach and give the prospects for its use and evolution.

## II. SEMANTIC INDEXING, CONCEPTUAL INDEXING

To represent the meaning conveyed by the textual content of a document, several approaches use thesauri or ontologies to annotate the document. The semantic annotation is usually accompanied by a disambiguation process.

In order to find the appropriate meaning of an ambiguous word occurrence, endogenous approaches use its context in the document and all the documents of the corpus [1]. Exogenous approaches exploit external linguistic resources such as digital dictionaries or Machine

ISBN: 978-1-61208-457-2

Readable Dictionary (MRD) [2], thesauri [3], or ontologies [4].

WordNet [5] is a linguistic resource. Its lexical database covers almost the entire English language. A concept in WordNet, which is called a synset, is represented by a set of synonyms. Synsets are connected by hyponym - hypernym (specialization - generalization) relations and meronymy - holonymy (part - all) relations. WordNet is a widely used resource, particularly in information retrieval. To represent a document, Baziz [6] defines a semantic core. The semantic content of a document is obtained by projecting the terms of the document on WordNet to extract the most representative synsets. The links between these synsets are weighted based on the semantic proximity (semantic similarity) between these synsets. The choice of synsets is based on two criteria: the co-occurrence called *cf.idf* and the semantic similarity used to disambiguate the synsets. Kolte [7] also uses WordNet to find the synsets corresponding to content of a document. He uses the various relationships defined in WordNet, as well as links, such as “ability link”, “function link” and “capability link” to disambiguate the ambiguous words. For each word or group of words in a document *d*, Wang [8] constructs a matrix *Uc* for each candidate synset *c*, extracted from WordNet, corresponding to *d*. The rows and columns of *Uc* represent the words, *di (i=1,n*) forming *c*. The row *i* of *Uc* gives the probability that a word *di* and a word *dj, (j=1,n)* appear simultaneously in *d*. The matrix *Uc* denotes the relevance of *d* with a synset *c*. In WordNet, domains are assigned to synsets to define the different meanings they may have. Kolte [9] uses these domains to find the correct meaning of a synset depending on other terms appearing together with it in the same sentence. Fauceglia [10] disambiguates verbs by exploiting information about the verbs that appear in similar contexts. His approach is applied in the Event Mention Detection task (EMD) to classify event types. He uses a database of the meaning of the verbs and no structure highlighting a relationship between the meanings of the verbs is used as it is the case in WordNet.

## III. Automatic classification of documents

The automatic text classification aims to organize documents into categories. One or more labels (classes, categories) are thus assigned to a document according to its text content.

Approaches dealing with supervised classification assign documents to predefined classes [11][12][13] while unsupervised classification approaches automatically define classes, called clusters [14].

In supervised classification, classifiers use two collections of documents: A collection containing learning documents to determine the features (terms) for each category and a collection containing new documents to be automatically classified. The classification of a new document depends on features retained for each category. A document is represented by a vector whose dimension is equal to the number of features selected to represent the different categories and no relationship between these features is highlighted. The vector document is then represented as a "bag of words".

Some classifiers create a "prototype" class from the learning collection [11]. This class is represented by the average vector of all vectors of the documents in the collection. Only certain features are retained, which represents a loss of information.

Other approaches replace the learning collection composed of selected documents for each category, by data extracted from the "world knowledge" as Open Directory Project (ODP) [15]. Other approaches use thesauri [16] and domain ontologies [17] with conventional classifiers such as Support Vector Machine (SVM), Naïve Bayes, K-means, etc.) and represent a document by a vector of features represented by concepts or by a combination of terms and concepts.

The representation of the features by a vector assumes their independence from one another. The different approaches face the problem caused by the large size of the document vector, which reduces their performance. A step for restricting the features is thus performed.

## IV. Proposed approach

Our approach aims to build, for a document, a graph whose nodes and arcs are respectively represented by concepts and relations between concepts. Our process is based on a global disambiguation step based on a classification of documents using several domain ontologies and a local disambiguation based on a domain ontology.

The documents classification allows grouping documents according to the knowledge domain defined by their content. This grouping identifies a global similarity expressed by the context in which the document has a coherent sense. This classification determines the concepts to retain for the document through its global context.

The classification that we implement is a semantic classification because unlike conventional approaches, we take into account the link between terms with their context of appearance in the document and we extract concepts corresponding to these terms from domain ontologies.

The classification allows to project the content of a document on several domain ontologies to determine which one best expresses its content. Synonyms and polysemic terms are assigned to concepts representing their appropriate sense. We consider the following facts:

- Someone can use the same terms to describe different knowledge. Thus, a term may have several meanings depending on the context in which it is used. The same term *ti* extracted from a document *d* can then be assigned to several concepts which belong to different ontologies.

$$t_i^d = \{c_{\theta 1}, c_{\theta 2,} ......\} \qquad (1)$$

  $C_{\theta i}$ represents the concept extracted from the ontology $\theta i$.

- A term can match with several concepts of the same ontology.
- The theme discussed in a document depends on the terms used in its content and the way these terms are grouped together in sentences and paragraphs.

### A. *Projection, extraction of terms and candidate concepts.*

The “projection” of a document on different ontologies allows to associate meaning to the terms of the document with respect to concepts belonging to these ontologies, and to select the candidate concepts. The notion of concept gives a meaning to a term relative to the domain in which this concept is defined.

We divide the whole document into sentences. Each sentence is browsed from left to right from the first word. We project the words of each sentence on different domain ontologies to extract the longer phrases (groups of words called "terms") that denote concepts. This choice is determined by: 1) the concepts are often represented by labels consisting of several words, 2) long terms are less ambiguous.

Several concepts belonging to the same domain ontology may be candidates for a given term.

### B. *Local disambiguation.*

The disambiguation process is used to select for a term *t* the most appropriate concept among several candidates belonging to the same ontology. To do this, we consider the context of occurrence of the term *t* in the document.

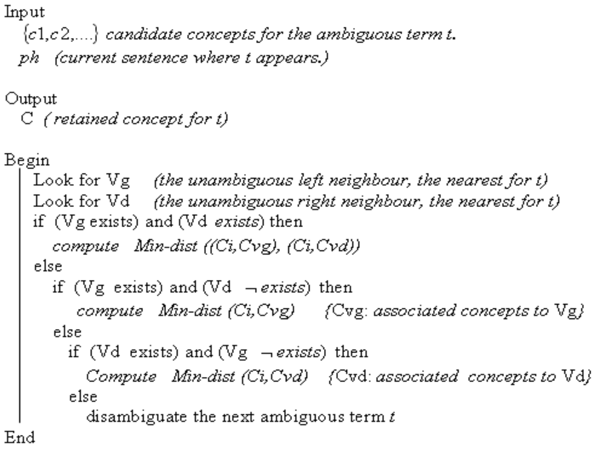
We consider the following assumptions:

- We assume that the semantic link between the terms depends on the distance between these terms within the document. The shorter the distance, the greater the semantic link. The semantic link decreases when passing from sentence to paragraph and also from one paragraph to another.
- We choose the appropriate concept for the term *t*, taking into account both the semantic distance between the term *t* with neighboring terms, (i.e. which occur in its context), and the semantic distance between concepts associated with the term *t* and the concepts corresponding to the neighboring terms in the ontology considered.
- The meaning of a term *t* in a document is determined by its nearest neighbors terms. *t* will then be disambiguated by its nearest neighbor on the left or by its nearest neighbor on the right. In case the left and right neighbors exist simultaneously, they will both be taken into consideration.

The disambiguation process is then done in three levels, starting at the sentence level. For each sentence, the ambiguous terms are disambiguated considering their left and right neighbors in the sentence. Any disambiguated term helps to move forward in the process of disambiguation of next terms. This process is repeated in case ambiguous terms still remain, considering in a second step the paragraph level, and finally, if necessary, the document level.

The disambiguation of a term *t* at sentence level is represented in Figure 1.

The disambiguation process at sentence level considers neighboring terms, unambiguous, that have associated concepts in the ontology considered, surrounding *t*: it retrieves C$v_g$ and C$v_d$, corresponding respectively to $v_g$, the nearest neighbor on the left of *t* and $v_d$, the nearest neighbor on the right of *t*.



```
Input
  {c1,c2,....} candidate concepts for the ambiguous term t.
  ph   (current sentence where t appears.)

Output
  C  ( retained concept for t)

Begin
  Look for Vg    (the unambiguous left neighbour, the nearest for t)
  Look for Vd    (the unambiguous right neighbour, the nearest for t)
  if (Vg exists) and (Vd exists) then
     compute  Min-dist ((Ci,Cvg), (Ci,Cvd))
  else
     if (Vg exists) and (Vd ¬exists) then
        compute  Min-dist (Ci,Cvg)    {Cvg: associated concepts to Vg}
     else
        if (Vd exists) and (Vg ¬exists) then
           Compute  Min-dist (Ci,Cvd)   {Cvd: associated concepts to Vd}
        else
           disambiguate the next ambiguous term t
End
```

Figure. 1. Local disambiguation process, sentence level.

The appropriate concept for the term *t* among candidate concepts is the semantically nearest concept of C$v_g$ or C$v_d$. This amounts to browsing the ontology and calculating the minimum distance between each concept associated with *t* and candidate concepts C$v_g$, C$v_d$. Several existing metrics in the literature are used to calculate this minimum distance.

### C. *Classification: global disambiguation*

While Kolt [9] determines the meaning of an ambiguous word with the most represented domain identified by the terms appearing with it in the same sentence, we seek to determine the context defined by a document. We propose to represent it not by words but by a set of concepts.

We rely on Wang's approach [8] which operates the occurrence of words within paragraphs to determine which concept to assign to a term of a document. We extend the process to the classification in order to determine the importance of all concepts extracted from different ontologies relative to the terms of the document.

At the end of the preceding steps, a document *d* is represented by several set of concepts extracted from domain ontologies $\theta i$ on which it has been projected.

$$d = \left\langle \begin{array}{l} \theta_1^d = \{c_{11}, \mathrm{c}_{21}....., c_{n1}\} \\ \\ \theta_i^d = \{c_{1i}, \mathrm{c}_{2\mathrm{i}}....., c_{ni}\} \\ ... \end{array} \right. \qquad (2)$$

The classifier needs to conclude the relevance of a document relative to a given context and to choose among the different ontological representations, which one best corresponds to its context. To do this, associating different domain ontologies to classes, the classifier will make the classification of a document relative to a single domain ontology.

The words used to describe a particular idea are not arbitrarily chosen. They are semantically related and are chosen with a common sense guided by this idea. However, it is almost impossible to find a document or a text in which all used terms refer exclusively to a same domain.

Recall that the previous steps are used to extract the concepts corresponding to the terms in the document. The extracted concepts can be related to multiple ontologies.

The classification we define in this work aims to determine, for each ontology *θi*, the semantic weight of each concept extracted for the document *d*. This determines the importance of a concept relative to a document. The evaluation of this weight is performed at two levels: paragraph level and document level.

*Paragraph level*: We calculate the weight of each concept *Ci* based on the other concepts appearing with it in a paragraph.

*Document level*: We calculate the total weight of each concept *Ci* throughout the document. This weight is obtained by adding the weights obtained for the concept *Ci* in the various paragraphs of the document *d*.

For each ontology and for each document we associate a matrix such as (3).

$$M_{\theta_i}^{d} = \begin{pmatrix} lc_1c_1 & lc_1c_2 \;..... & lc_1c_n \\ & & \\ lc_nc_1 & lc_nc_2 \;..... & lc_nc_n \end{pmatrix} \qquad (3)$$

The rows and columns of this matrix represent all concepts extracted from ontology *θi* for the document *d*.

*Ci* is any concept extracted from the ontology *θi* after the projection of the document *d* on *θi; lcicj* represents the weight of the link between the concept *Ci* and the concept *Cj (i≠j)*. This weight is calculated as follows:

- The matrix is initialized to zero
- If a term *ti* and a term *tj* appear together within the same paragraph of the document *d* and concepts *Ci* and *Cj* correspond to terms *ti* and *tj* respectively, then the weight *lcicj =1.*
- The weight *lcicj* is updated each time terms *ti* and *tj* appear together in the same paragraph.
- The weight *lcici* corresponds to the appearance of term *ti* in the paragraph. It is equal to 1.
- The weight *lcicj* is updated for all paragraphs in the document *d*.

Each row of the matrix represents the total weight of a concept extracted from the ontology *θi* relative to a document *d*. This weight assesses the importance of the concept *Ci* in *d*.

The total weight of all the extracted concepts of an ontology relative to document *d*, measures how well each ontology represents this document. The highest score will determine the ontology candidate which will be chosen to represent the document *d*.

## V. IMPLEMENTATION AND EXAMPLES

We implemented our approach using both WordNet and WordNet Domains resources. In WordNet Domains, several knowledge domains are used, such as medicine, computer science, economy etc, and each synset is annotated with one or more domains in which it has a meaning.

To achieve our classification, we have assimilated these domains to domain ontologies. To evaluate the distance between two synsets in WordNet we used Rita similarity metric [18].

The words within sentences are tagged with their type (noun, verb, adverb, adjective, etc.) by Stanford Part-Of-Speech Tagger (POS Tagger) [19].

To illustrate our approach, we apply it on the three following examples:

- Txt1: *The role of banks in the economy was clear and well established as the financial markets were underdeveloped because they were the only ones to provide liquidity and credit to businesses and households. The unprecedented development of financial markets, driven by the late 1970s in the Anglo-Saxon countries, has led some economists to question the specificity of bank financing compared with direct funding and the survival of traditional banks. Several arguments have been advanced.*
- S1: *Banks use their networks to exploit economies of scale between activities (collection of savings, management of means of payment, exchange, offer insurance products, securities investment services….*
- S2: *The player throws the baseball and he improves the score…*

### A. *Example 1: the Txt1 case*

*1) Global disambiguation:* We consider four domains and we apply the classification process to determine the domain that represents best the content of the text *Txt1*. It determines the synsets to retain for the text through its global context. Table I shows the result of the projection of *Txt1* on four ontologies and the score obtained by each ontology.

The selected domain is *Economy* because it has obtained the highest score.

*2) Local disambiguation, sentence level:* In the domain retained, the term *economy* has two synsets. So this is an ambiguous term. A local disambiguation is performed to determine what synset to retain for this term. It is performed at sentence level. The nearest unambiguous neighbor of the term *economy* is only on the right: it is the term *credit*. There is no path between *economy* and *credit* in WordNet. Another nearest neighbor is sought in the sentence. This is the term *business* that is on the right of *economy*.

The distance between *business* 07485368- n and *economy* 00182005-n is: 1.0.

The distance between *business* 07485368-n and *economy* 07857433-n is: 0.8333333.

The synset retained for *economy* is 07857433-n.

The text *Txt1* is represented by the following synsets (economy 07857433-n, credit 12616435-n, business 07485368-n, economist 09401295-n).

TABLE I. BREAKDOWN BY ONTOLOGIES OF SYNSETS ASSOCIATED TO TERMS OF TXT1.

| Ontologies | Terms | Synsets | Score |
|---|---|---|---|
| **economy** | economy | 00182005-n<br>**07857433-n** | 16 |
| | credit | 12616435-n | |
| | business | 07485368-n | |
| | economist | 09401295-n | |
| **finance** | bank | 12599211-n | 1 |
| **enterprise** | business | 01033295-n<br>01031794-n<br>07571175-n | 6 |
| | financing | 01036077-n | |
| | funding | 01036077-n | |
| **banking** | bank | 02690337-n<br>07909067-n | 8 |
| | credit | 12620638-n | |
| **Synsets** | **Definitions (Glosses of WordNet)** | | |
| 00182005-n | an act of economizing; reduction in cost. | | |
| 07857433-n | the system of production and distribution and consumption. | | |
| 07485368-n | business concerns collectively.<br>“Government and business could not agree” | | |
| 01033295-n | the volume of business activity;<br>“business is good today” | | |
| 01031794-n | commercial_enterprise, business_enterprise<br>the activity of providing goods and services involving financial and commercial and industrial aspects. | | |
| 07571175-n | business_organisation<br>a commercial or industrial enterprise and the people who constitute it. | | |

### B. *Example 2: the S1 case*

*S1* is an extract of a sentence belonging to a text classified in the domain *Economy*. Table II summarizes the synsets associated with its terms.

*Payment* is an ambiguous term since it has two synsets. A local disambiguation is realized at sentence level. The nearest unambiguous neighbors for *payment* are *means*, which is on the left and *exchange* which is on the right.

The distance between *exchange* 01045967-n and *payment* 01056649-n is: 0.4.

The distance between *exchange* 01045967-n and *payment* 12522505-n is: 1.0.

The distance between *means* 12596703-n and *payment* 01056649-n is: 1.0.

The distance between *means* 12596703-n and *payment* 12522505-n is: 0.85714287.

The shortest distance is given by the term *exchange* that is on the right of *payment*. The synset retained for *payment* is 01056649-n.

### C. *Example 3: the S2 case*

We consider an extract from the sentence *S2* belonging to a text classified in the domain *Play*. Table III summarizes the synsets associated with its terms.

*Baseball* is ambiguous. It has two neighbors but only one is unambiguous. This is the term *player* that is on the left. So *Player* disambiguates *baseball*.

The distance between *player* 09762180-n and *baseball* 02701461-n is: 0.75.

TABLE II. SYNSETS ASSOCIATED TO TERMS OF S1

| Terms | Synsets | Definitions (Glosses of WordNet) |
|---|---|---|
| economy_of_scale | 00182453-n | |
| saving | 00182005-n | |
| means | 12596703-n | |
| payment | **01056649-n**<br>12522505-n | the act of paying money.<br>a sum of money paid. |
| exchange | 01045967-n | |
| security | 12592487-n | |
| investment | 12576508-n | |

TABLE III. SYNSETS ASSOCIATED TO TERMS OF S2

| Terms | Synsets | Definitions (Glosses of WordNet) |
|---|---|---|
| player | 09762180-n | |
| baseball | **02701461-n**<br>00447188-n | a ball used in playing baseball.<br>a ball game played with a bat and ball between two teams of 9 players. |
| score | 00176295-n<br>12829162-n | |

The distance between *player* 09762180-n and *baseball* 00447188-n is: 1.0

The synset retained for *baseball* is: 02701461-n.

## VI. CONCLUSION

In this paper, we proposed an approach to extract semantics from documents by using domain ontologies with a disambiguation process. This process combines local and global disambiguation. The first one finds an appropriate concept for a term with several meanings in a single domain ontology, the second one retrieves the appropriate concept for a term that has multiple meanings in different knowledge domains. Throughout the disambiguation process, we took into account the context of appearance of the ambiguous terms in the document. The quality of the disambiguation process of course depends on domain ontologies, since they must cover the entire vocabulary of the represented domain.

The major problem conventional classifiers suffer from is the vector representation of a document in a high dimensional space. Indeed, the size of the vector equals the number of features that represent all classes used by the classifier. This dimension, very large, lowers the performance of classifiers. Moreover, characteristics representing a document are independent of each other.

Our approach has the advantage of responding to the problem of polysemy and synonymy engendered by the terms of the document. The document is not represented by a vector of high dimension but by a conceptual graph where concepts correspond only to the terms describing its contents. We believe that the use of ontologies in our classification process is a more stable base that the use of a set of learning documents, in which the choice of such learning documents affects the result of the classification.

We have conducted tests on a first set of short documents. Even if the obtained results are very encouraging, we have obviously to confirm the interest of our approach by considering larger collections, with more candidate domains. This is what we plan to do in order to assess the effectiveness of our approach in comparison to the existing ones.

## REFERENCES

[1] H. Schütze and J. Pedersen, “Information retrieval based on word senses,” In Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval, 1995, pp. 161-175.

[2] J.A. Guthrie, L. Guthrie, Y. Wilks and H. Aidinejad, “Subject-dependant cooccurrence and word sense disambiguation,” In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkley, CA, 1991, pp.146-152.

[3] D. Yarowsky, “Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora,” Proceedings of the 14th international Conference on Computational Linguistics (COLING-92). Nantes, France, August. 1992, pp.454-460.

[4] P. Resnik, “Disambiguating noun groupings with repect to WordNet senses,” 3thWorkshop on Very Large Corpora, 1995, pp.54-68.

[5] G. A. Miller, “WordNet: A Lexical Database for English,” Communications of the ACM Vol. 38, No. 11, 1995, pp. 39-41.

[6] M. Baziz, M. Boughanem and N. Aussenac-Gilles, “Conceptual Indexing Based on Document Content Representation,” CoLIS 2005, pp. 171-186.

[7] S. G. Kolte and S. G. Bhirud, “Exploiting links in WordNet hierarchy for word sense disambiguation of nouns,” International Conference on Advances in Computing, Communication and Control, (ICAC3’09), 2009.

[8] H. Wang, Y. Guo and X. Shi, “Research of the conceptual representing of documents based on light ontology,” 9th International Conference on Fuzzy Systems and Knowledge Discovery, (FSKD, 2012), 2012.

[9] S. G. Kolte and S. G. Bhirud, “WordNet: A Knowledge Source for Word Sense Disambiguation,” International Journal of Recent Trends in Engineering, Vol 2, No. 4, November. 2009.

[10] N. R. Fauceglia, Y.C. Lin, X. Ma and E. Hovy, “Word sense disambiguation via propstore and ontonotes for event mention detection,” In Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation, Denver, Colorado, june. 2015, pp.11–15.

[11] T. Joachims, “A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization,” Proceedings of ICML-97, Tennessee, 1997, pp.143-151

[12] Y. Yang and X. Liu, “A re-examination of text categorization methods,” 22nd Annual International SIGIR, Berkley, August. 1999, p. 42–49.

[13] S. Jaillet, A. Laurent and M. Teisseire, “Sequential patterns for text categorization”. Intelligent Data Analysis, IOS Press, 2006.

[14] A. Hotho, A. Maedche and S. Staab, “Ontology-based Text Document Clustering,” KI 16 (4), 2002, pp. 48-54.

[15] E. Gabrilovich and S. Markovitch, “Feature Generation for Text categorization Using World Knowledge,” IJCAI 2005: the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30-August 5.2005, pp. 1048-1053

[16] A. Hotho, S. Staab and G. Stumme, “Ontologies Improve Text Document Clustering,” ICDM:3rd IEEE International Conference on Data Minin 2003, pp. 541-544

[17] H. H. Tar and T.T. Soe.Nyunt, “Ontology-Based Concept Weighting for Text documents,” International Conference on Information Communication and Management IPCSIT vol.16, IACSIT Press, Singapore, 2011.

[18] D. C. Howe, “RiTa: creativity support for computational literature,” In Proceedings of the seventh ACM conference on Creativity and cognition (C&C '09). ACM, New York, NY, USA, 2009, pp. 205-210.

[19] K. Toutanova, D. Klein, C. Manning, and Y. Singer, “Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network”, In Proceedings of HLT-NAACL, 2003, pp. 252-259.

# Extraction of Business Information on the Web to Supply a Geolocated Search Service

Armel Fotsoh Tawofaing$^*$, Christian Sallaberry$^\dagger$, Annig Le Parc - Lacayrelle$^\dagger$, Tanguy Moal$^\ddagger$
$^*$LIUPPA - Cogniteev
Email: aftawofaing@univ-pau.fr
$^\dagger$LIUPPA, University of Pau, France
Email: christian.sallaberry@univ-pau.fr, annig.lacayrelle@univ-pau.fr
$^\ddagger$Cogniteev, Bordeaux, France
Email: tanguy@cogniteev.com

***Abstract*—Searching information about local businesses is not a trivial problem to address. Most of existing services are supplied with manually recorded data. Based on the observation that more and more businesses are referenced on the web, we propose a new approach, which consists to extract companies' targeted information (addresses, activities, jobs, products, emails, fax, phone numbers) from websites, to supply a local business search service. The information retrieval module combines thematic, spatial and full-text criteria.**

***Keywords–Information Extraction; Geographic Information Retrieval***

## I. Introduction

Identifying specific information on the web according to a spatial localization is a topic increasingly explored. For example, a geolocated search service dedicated to emergency facilities is presented in [1]. Our research goal is to crawl the web and extract data in order to build specific geolocated entities, like businesses, events or persons.

This contribution presents the architecture of a service dedicated to local business search. As opposed to traditional search engines that relies on full document indexation, business local search services, mostly relies on companies descriptors provided by some specialized organisations. Far from adequate to build efficient systems, the basic descriptors must be supplemented by new ones. Therefore, we propose a service crawling the web, extracting information about companies (activity fields, practised jobs, commercialized or manufactured products, postal addresses, emails, phone and fax numbers) and storing them in indexes. The ultimate goal is to process a user need containing a thematic part (jobs, activities or products) and a spatial one, in order to query the indices and to get relevant results according to such different criteria.

The proposal relies on a model of business entity that is composed of two different parts. The first one is constituted of business basic data (official name, registration category, identifier, etc.) collected by crawling some specific administrative directories. The second part is composed of extended data extracted from companies' websites by using knowledge resource and pattern based extraction approaches. The retrieval system relies on the corresponding business entities and geolocated data.

The rest of this paper is organized as follow. Section II presents some related work; Section III describes the architecture of the proposed approach used to build our service; Section IV presents the implementation of a first version of a prototype and Section V concludes the paper and presents some prospects.

## II. Related work

Some research works focus on the extraction of information related to businesses on the web. For example, Ahlers [2] develops a system which analyses web pages content in other to enrich a Yellow Pages [3] data provider. Analysed web pages are identified using Directory Mozilla (DMOZ) [4]. Extracted data here is addresses, phone numbers, emails, commercial and tax information of businesses. This data is used to consolidate and enrich the one contained in the Yellow Pages database. It is important to note that, in the French context in particular, the proportion of businesses registered in DMOZ remains very slight. Thus enriched data will also be limited.

Bootstrapping targeted data from Internet is a topic increasingly explored by several research works. Rae and al. [5] propose an approach for bootstrapping the web in order to identify Points Of Interest (POIs) websites. Their proposal uses Wikipedia data to list a POI information which is used to query Bing API as to retrieve the most relevant website corresponding to this POI.

Furthermore, many services dedicated to business information retrieval are available on the web. We organize these services into three main categories: (i) data providers like Factual [6] which collect and commercialize business data; (ii) directories like Yellow Pages, Google Maps [7] which contain a database of business information available online; (iii) social networks like Yelp [8] or Foursquare [9], which are services used in information sharing and reviews about businesses, places or events. Data supplying these services mostly come from manual recordings (users and employees), partners companies and open data. Thereby, if a company is not registered in the databases supplying those services or if its data is not updated frequently, it will lead us to have missing or out of date information.

Extraction of geographical information in web pages is a well-explored area of research. Some specific works focus on automatic extraction of addresses in unstructured web pages. An ontology-based approach for recognizing, extracting and geocoding Brazilian addresses in web pages is presented in [10]. A pattern-based approach [11] is used by Ahlers and Bool in [12] to present a technique of extraction and validation of

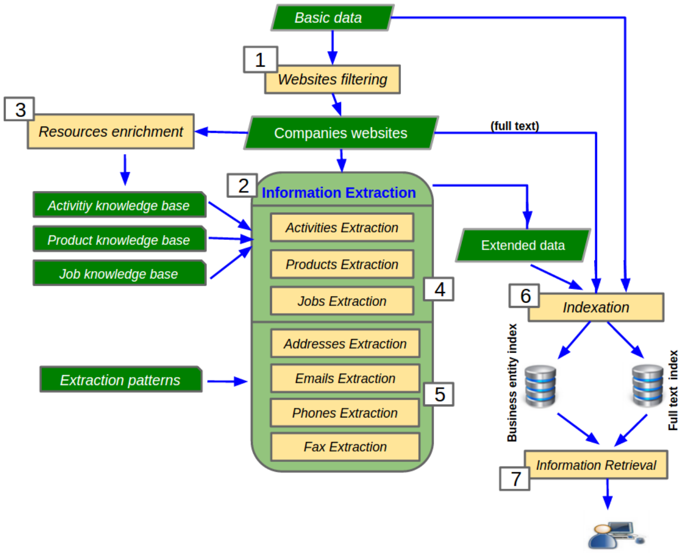


Figure 1. Processing chain

German addresses from web pages. City name or ZIP code are the two entry points of the extraction process. Moreover, [12] uses a street extraction process based on a gazetteer containing all the German street names.

Exploitation of knowledge resources is also a technique increasingly used for the extraction of thematic information in text, especially with significant progress in semantic web field these last years. Structuring and formalizing the knowledge of a specific domain in an ontology allows to annotate concepts [13] and semantic relations [14] in the text. Therefore, it is possible to perform semantic reasoning on the extracted information.

## III. Proposition

We propose an architecture of a local business search service, supplied with web data. Indeed, the contribution describes an approach that aims companies' websites identification on the web, and extraction of location and thematic information from these websites by combining some information extraction techniques. This proposal improves the existing services named in the state of the art, because it aims construction of low cost and up to date data. Differently from [2], we extract data in a web pages corpus constituted by filtering companies websites based on a heuristic. Besides of contact information and location data, we also extract activities, products and jobs in companies' web pages, using knowledge resources.
The process flow of our service is composed of four main steps (Figure 1).

### A. Preprocessing

In this first step, one of our goals is to constitute the corpus of companies' websites in which we want to extract information. For this purpose, we have to bootstrap the web in order to filter those websites. We use a similar approach to the one described in [5] for our corpus constitution task. In fact, a directory containing license information of companies [15] is crawled to collect business basic data. These basic data, especially the name of the company and its localization, are used to query automatically Google and identify company website when it exists (Figure 1, Process 1). Business directories, social and professional networks are stored into a stop list in order to be filtered during the website search process. This helps to reduce not relevant results in the filtering process. Websites list identified during this step constitutes the input data to the extraction process.

Besides, we have constructed three core knowledge resources describing business activities, products and job positions respectively. These resources are based on hierarchical organisations defined by the French National Statistics Institute called INSEE [16] for the first two ones and by the French organisation for work, Pole Emploi [17] for the third one. We transformed all these resources in order to represent them as OWL ontologies. Furthermore, the Latent Dirichlet Allocation (LDA) clustering algorithm [18] is applied to companies' websites of each activity group, in order to extract the corresponding vocabulary. Indeed, this learning algorithm enriches our core knowledge resources (Figure 1, Process 3).

### B. Information extraction in websites

One of the most difficult challenges in web pages analysis is that information on Internet is mostly unstructured. Indeed, only few websites use metadata or microformats to publish structured information. In our proposal, we want to analyse companies' websites and extract thematic and spatial information in order to fill in business extended data.

*1) Thematic information extraction:* Extraction of activities, jobs and products relies on an ontology-based approach. The three knowledge resources built and enriched in the preprocessing step, are used to annotate automatically the website corpus (Figure 1, process 4). Each term or phrase in pages content which corresponds to a resource category is tagged with the corresponding identifier.

Emails, phone and fax numbers are extracted by using a pattern-based approach. We wrote extraction rules by observing patterns used for the writing of the targeted information in a sample of French companies' websites. Phone numbers follow a specific pattern, depending if it has a country telephone code or not. Our proposal makes a distinction between mobile and landline phone numbers based on French standards. Fax number are landline phone numbers introduced by special keywords (”Télécopie”, ”Télécopieur”, ”Fax”, etc.). An email is a phrase which follows this pattern (Table I corresponds to the legend of extraction patterns) :

$$\begin{aligned}\textbf{email} \rightarrow \ & Login \quad (\text{"@"} \ | \text{"}(at)\text{"}) \quad Domain\_Name \\ & (\text{"."} \ | \text{"}(dot)\text{"}) \quad Domain\_Extension\end{aligned}$$

The entry point of the extractor is the identification of a domain extension ($Domain\_Extension$) and ”@” or ”(at)” symbol. We use a gazetteer of domain extensions for this purpose.

*2) Spatial information extraction:* In the literature, models are proposed to represent entities like addresses. Schema.org has a module dedicated to address modelling. In our business model, location representation is based on the one defined by the French governmental data lab named Etalab [19]. Our goal is to extract addresses in web pages up to the street number granularity level according to the Etalab addresses model.

TABLE I. LEGEND OF EXTRACTION PATTERNS

| | |
|---|---|
| A ? | A is optional |
| A \| B | A or B |
| A B | A and B |
| A(n) | A is repeated n times |
| [i-j] | element in set i, ..., j with $i < j$ |

We also want to extract information like address supplements (building name, floor number, etc.), postal boxes numbers and letter numbers (it is a number used by postal services to facilitate postal mail transmission, e.g., ”CEDEX 07”).

Several difficulties are related to addresses extraction from text in general and French ones in particular. In fact, there are many address formats used and standards about the order of the various components of an address do not exist (a council name is not always following a ZIP code). Besides, just a few web publishers use address keyword introducer or specific tagging to facilitate address automatic identification. Furthermore, in the French context, there is no a complete and available gazetteer containing all street names. Thereby, the approach proposed by [12] has to be adapted in our study case.

We propose an approach to automatically extract French addresses in web pages. The observation of a sample of 160 companies' websites allowed us to identify some frequent patterns of addresses formats in French companies' websites. In this sample, the ZIP code was always present in the identified addresses. Table II describes the different components used in the expression of a French address. Each one is extracted using gazetteers or specific rules. For example, ZIP Code is extracted using the following rules:

$$\textbf{ZC} \rightarrow\ ''F-''\quad [0-9](5) \qquad (1)$$
$$\textbf{ZC} \rightarrow\ [0-9](5) \qquad (2)$$
$$\textbf{ZC} \rightarrow\ [0-9](2)\quad [0-9](3) \qquad (3)$$

Priorities are associated to each of these rules. The rule (1) has the highest invocation priority, rule (2) afterwards. More complex rules are used to extract fields like street name (SNa) and address supplement(AS).

The address extraction patterns observed have been summarized into three main rules. They are described below:

**Extraction rule 1**

$$\textbf{Address} \rightarrow\ AS?\quad ((PB\quad SC) \mid (SC\quad PB) \mid PB \mid SC)?\quad SNu?\quad SNa\quad AS?\quad ((PB\quad SC) \mid (SC\quad PB) \mid PB \mid SC)?\quad ((ZC\ C) \mid (C\ ZC))\quad LN?\quad D?\quad Co?$$

**Extraction rule 2**

$$\textbf{Address} \rightarrow\ AS?\quad ((PB\quad SC) \mid (SC\quad PB) \mid PB \mid SC)?\quad ((ZC\ C) \mid (C\ ZC))\quad LN?\quad D?\quad Co?$$

**Extraction rule 3**

TABLE II. ADDRESS COMPONENTS

| Field names | Symbols | Examples |
|---|---|---|
| Address Supplement | AS | Résidence Rigaud |
| Postal Box | PB | BP 1167 |
| Special Course | SC | CS 2587 |
| Street Number | SNu | 10 ter |
| Street Name | SNa | Avenue de l'université |
| ZIP Code | ZC | 64000 |
| City Name | C | Pau |
| Letter Number | LN | CEDEX 01 |
| Department | D | Pyrénées-Atlantiques |
| Country | Co | France |
| Street Introduce | SI | Avenue |

$$\textbf{Address} \rightarrow\ AS?\quad ((PB\quad SC) \mid (SC\quad PB) \mid PB \mid SC)?\quad SNu?\quad SNa\quad AS?\quad ((PB\quad SC) \mid (SC\quad PB) \mid PB \mid SC)?\quad ZC\quad LN?\quad D?\quad Co?$$

In the first pattern, the street name, ZIP code and city name are required (e.g., ”10 Rue du Maréchal Foch 49000 Angers”). This pattern represents about 75% of the addresses of our dataset. In the second pattern, only the ZIP code and city name are required, street name must be omitted. (e.g., ”Résidence Rigaud 33350 Mouliets-et-Villemartin”). This pattern represents about 14% of the observed addresses. In the third pattern, street name and ZIP code are required and the city name must be omitted ((e.g., ”10 rue du Maréchal Foch F-33500”). This last case represents less than 4% of the detected addresses. Others components like the address supplement and postal box might complete the extracted addresses.

Algorithm 1 details the conditions of the different address extraction rules triggering. The entry point of the algorithm is the ZIP Code which is identified by using rules defined previously. For each sentence in which a potential ZIP Code is identified, if a council name and a street introducer are detected, pattern 1 is triggered. Otherwise, if a country name only is detected, pattern 2 is triggered. Finally if a street introducer only is detected, pattern 3 is triggered. In all the other cases, there is no address in the sentence. Let us note that SI detection is identified using a gazetteer.

**Algorithm 1** Address Extraction Algorithm

```
for each sentence do
    if (ZC & C & SI) then
        trigger extraction rule 1
    else
        if (ZC & C) then
            trigger extraction rule 2
        else
            if (ZC & SI) then
                trigger extraction rule 3
            else
                No Address
            end if
        end if
    end if
end for
```

*3) Full-text extraction:* The content of each web page of the corpus is processed by eliminating all the HTML tags in page content as well as metadata. JavaScript scripts and CSS code are also removed. Only the full-text is kept.

### C. Indexation

For each company, basic and extended data are merged to build a complete business entity according to the proposed model. Every extracted address is geocoded using Etalab tool. The business entities so constructed are stored in an index. In parallel, the full-text corresponding to each web page is stored in another index (Figure 1, process 6).

### D. Information Retrieval

The indices built in the previous stage are used to answer user needs (Figure 1, process 7). The information retrieval process analyses each query to handle separately or accordingly its spatial and thematic parts. Furthermore, the retrieval process integrates the querying of the web pages whole content. This retrieval stage combines spatial, thematic and full-text querying criteria.

## IV. Implementation

A first prototype of our approach has been implemented, dealing with companies of the South West region of France. We worked on 22,000 companies websites representing 212 business activity fields defined by INSEE. The related corpus is constituted by crawling all these websites with Apache Nutch [20] framework. This generates a corpus of 550,000 web pages. Both of the pattern-based and the ontology-based extraction approaches, described in the extraction process, are performed on this corpus using GATE [21] platform. We connected GATE with Apache HADOOP [22] framework in order to process the large volume of web pages. The use of HADOOP Map Reduce framework with two machines working in parallel, has reduced by more than 50% the time of web pages annotation. The information extracted in these web pages has been indexed using Elasticsearch [23]. The business entity and full-text indices have a total size of 3 GB.

”Oak beams in south of Bordeaux” is an example of user need. The system has to return business entities of the ”carpentry work” activity located in the south of Bordeaux. This information need is processed according to the Elasticsearch syntax and submitted to our prototype. The execution of the preview query retrieves a list of relevant business entities (the first one is http://www.belles-toitures-girondines.com).

## V. Conclusion

Our service uses a modular structure. It combines several research areas and techniques, which become complementary in the construction of business entities. This solution leverages learning techniques to enrich knowledge resources. It also performs information extraction (spatial and thematic) using a pattern based approach and knowledge resources. Moreover, our service relies on a new model of business entities, combining administrative data and those extracted from websites. Finally, the service addresses spatial, thematic and full-text information retrieval.

Future work will focus on the evaluation of website filtering and information extraction processes. A definition of an efficient information retrieval model to combine such criteria and support natural language queries will also be carried out in future research. Moreover, an evaluation of the entire architecture with a representative set of queries will also be performed in future work.

## References

[1] W. Li, M. F. Goodchild, R. L. Church, and B. Zhou, “Geospatial data mining on the web: Discovering locations of emergency service facilities,” 2012, pp. 552–563.

[2] D. Ahlers, “Business entity retrieval and data provision for yellow pages by local search,” in IRPS Workshop@ ECIR2013, 2013.

[3] “http://www.pagesjaunes.fr,” 2015.10.29.

[4] “http://www.dmoz.org,” 2015.10.29.

[5] A. Rae, V. Murdock, A. Popescu, and H. Bouchard, “Mining the web for points of interest,” in Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR ’12. New York, NY, USA: ACM, 2012, pp. 711–720. [Online]. Available: http://doi.acm.org/10.1145/2348283.2348379

[6] “http://www.factual.com,” 2015.10.29.

[7] “http://www.google.fr/maps,” 2015.10.29.

[8] “http://www.yelp.fr,” 2015.10.29.

[9] “http://fr.foursquare.com,” 2015.10.29.

[10] K. A. V. Borges, A. H. F. Laender, C. B. Medeiros, and C. A. Davis, Jr., “Discovering geographic locations in web pages using urban addresses,” in Proceedings of the 4th ACM Workshop on Geographical Information Retrieval, ser. GIR ’07. New York, NY, USA: ACM, 2007, pp. 31–36. [Online]. Available: http://doi.acm.org/10.1145/1316948.1316957

[11] S. Blohm, Large-scale pattern-based information extraction from the world wide web. KIT Scientific Publishing, 2011.

[12] D. Ahlers and S. Boll, “Retrieving address-based locations from the web,” in Proceedings of the 2Nd International Workshop on Geographic Information Retrieval, ser. GIR ’08. New York, NY, USA: ACM, 2008, pp. 27–34. [Online]. Available: http://doi.acm.org/10.1145/1460007.1460015

[13] S. Nešić, F. Crestani, M. Jazayeri, and D. Gašević, “Concept-based semantic annotation, indexing and retrieval of office-like document units.” CID, 2010.

[14] A. Royer, C. Sallaberry, A. Le Parc-Lacayrelle, and M.-N. Bessagnet, “Extraction automatique de relations sémantiques définies dans une ontologie,” in Actes RISE 2015, 2015, pp. 30–42.

[15] “http://www.societe.com,” 2015.10.29.

[16] “http://www.insee.fr/fr,” 2015.10.29.

[17] “http://www.pole-emploi.fr,” 2015.10.29.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” the Journal of machine Learning research, vol. 3, 2003, pp. 993–1022.

[19] “http://www.etalab.gouv.fr,” 2015.10.29.

[20] “http://nutch.apache.org,” 2015.10.29.

[21] “http://gate.ac.uk,” 2015.10.29.

[22] “http://hadoop.apache.org,” 2015.10.29.

[23] “http://www.elastic.co,” 2015.10.29.

# A Semantic Annotation Model for Syntactic Parsing in Ontology Learning Task

Maria Pia di Buono, Mario Monteleone
DISPSC
UNISA
Fisciano (SA), Italy
{mdibuono,mmonteleone}unisa.it

***Abstract*—In this paper, we present an on-going research on a semantic annotation model, which aims at creating a system for syntactic parsing apt to achieve Ontology Learning (OL) tasks. Our approach is based on a deep linguistic analysis, derived from predicate-argument structures and established on the notion *de prédicat sèmantique* (semantic predicate), following Maurice Gross' approach.**

***Keywords - Ontology Learning; Deep Linguistic Analysis; Syntactic Parsing; Lexicon-Grammar.***

## I. INTRODUCTION

Semantic annotation process of sentence structures is one of the most challenging Natural Language Processing (NLP) tasks. Usually, applied techniques may be distinguished by the use of shallow or deep semantic analysis [1]. Besides, some recent approaches apply hybrid methods [1], in order to exploit the benefits derived from both for each step of the process. Shallow linguistic processing concerns the achievement of specific NLP tasks, even if it does not aim at accomplishing an exhaustive linguistic analysis. Systems based on a shallow approach are generally oriented to tokenization, part-of-speech tagging, chunking, named entity recognition, and shallow sentence parsing. Due to the improvement of such systems, in the last years the capability of text analysis achieved by shallow techniques has increased. Still, in terms of efficiency and robustness, shallow technique results are not comparable to deep system ones.

On the other hand, deep linguistic processing mainly refers to approaches, which apply linguistic knowledge to analyze natural languages. Such linguistic knowledge is encoded in a declarative way, namely in formal grammars, yet neither in algorithms nor in simple database. Thus, a formal grammar becomes an expression of both a certain linguistic theory and some operations, which are used to check consistence and to define information fusing. For this reason, deep linguistic processing is usually defined as a rule-based approach. Due to the fact that statistical methods may also be applied to deep grammars and systems, this does not mean that rule-based approaches are opposite to statistical methods. In deep linguistic processing, rules state constraints, based on a linguistic theory, which drives correct syntax of linguistic entities, while words are encoded in a specific lexicon. Syntax rules are not only related to grammatical correctness, on the basis of which a sentence is grammatically approved or rejected , but they may also describe semantic representations. Thus, syntax seems to be able to express both linguistic levels, namely the grammatical level and the meaning one.

The rest of the paper is structured as follows. In Section II, we present the most widely used annotated corpora, the Penn TreeBank and the PropBank. Then, in Section III, we introduce the main goals of OL, highlighting the challenging aspects in the achievement of such task. Consequently, in Section IV, we propose our framework to develop a system for syntactic parsing. Finally, in Section V, we analyze the proposed annotation process, which is based on a semantic tagging.

## II. RELATED WORKS

The most well-known annotated corpora are the Penn TreeBank [2] and the PropBank [3].

The Penn TreeBank (1989-1996), is a parsed corpus, syntactically and semantically annotated, which produces:

- 7 million words of part-of-speech tagged text,
- 3 million words of skeletally parsed text,
- over 2 million words of text parsed for predicate-argument structure,
- and 1.6 million words of transcribed spoken text annotated for speech disfluencies [2].

Its corpus is composed of texts, derived from different sources, e.g., Wall Street Journal articles, IBM computer manuals, etc., and also from transcribed telephone conversations. Data produced by the Treebank are released through the Linguistic Data Consortium (LDC)[1]. The annotation process is achieved through a two-step procedure, which involves an automatic tagging and a human correction.

PropBank (Proposition Bank) is a project of Automatic Content Extraction (ACE), which aims at creating a text corpus annotated with information on basic semantic propositions. Predicate-argument relations were added to the syntactic trees of the Penn Treebank. Thus, PropBank offers

[1] https://www.ldc.upenn.edu/.

predicate-argument annotations, presenting a single instance, which contains information about the location of each verb, and the location and identity of its arguments [3].

## III. Linguistic analysis and Ontology Learning

The main aim of ontology learning is  retrieving from the knowledge extracted concepts and relationships among concepts. To develop adequate tools for this task, shallow and deep semantic analysis approaches are used.

Generally speaking, “the shallow semantic analysis measures only word overlap between text and hypothesis” [4]. Starting from tokenization and lemmatization of text and hypothesis, this analysis uses Web documents as a corpus and assigns to each entry inverse document frequency as a weight in the hypothesis. Thus, we have a higher score for less occurring words, which means that we assign more importance to less frequent words. Shallow analysis needs tagged corpora as training resources. This technique may be applied at both syntactic and semantic level. Shallow approach is largely used in various tasks of ontology learning:

- Term Extraction: terms are extracted using chunkers. Outputs, as nominal phrases, may be included in the basic vocabulary of the domain. Usually, in order to evaluate weight of extracted terms with respect to the corpus, statistical measures of Information Extraction (IE), such as Term Frequency for Inverse Document Frequency (TF*IDF) algorithm [5], are applied.
- Taxonomy Extraction: this task is related to the extraction of hierarchical relations among ontology classes or individuals [6] [7]. The hierarchy is usually extracted using lexical and syntactic patterns, expressed by means of regular expressions.
- Relation Extraction: using shallow parsing, it is possible to extract only limited reliable relations, i.e., simple patterns such as Noun Phrase + Verb Phrase + Noun Phrase. This analysis does not address complex sentence structures in which there are discontinued dependencies or other language ambiguities. Obviously, this limit does not allow axiom learning, obtainable only with deeper syntactic methods.

While shallow NLP covers syntactic steps as for the learning process, various methods are also applied to generate a shallow semantic parsing (semantic tagging). These methods are more useful in ontology population procedure than in learning tasks, because they govern an extraction approach relied on conceptual structures. Such approach is extremely different from the one based only on texts and syntactic NLP. Indeed, to extract entities and semantic relationships, semantic parsing requires the identification of the structures presented in the corpus. Thus, in order to discover instances of these resources, population process relies on a set of knowledge resources, such as frame, templates or roles [8]. According to [9], these resources may include role taxonomies, lists of named entities and also lexicons and dictionaries. For these reasons, shallow semantic parsing necessitates word sense disambiguation process, useful to assign to a given word the correct meaning or concept. This procedure is also applied to recognize particular semantic relationships, such as synonyms, meronyms, or antonyms using predefined patterns.

While shallow semantics may adequately respond to some ontology learning steps, the  results are inadequate for more complex tasks. Shallow methods do not guarantee a fine-grained linguistic analysis. For instance, as for anaphora resolution, or quantifier scope resolution, extracting rich domain ontologies requires text processing. Considering that deep NLP allows to work not only on concepts and relations but also on axioms, such approach seems more appropriate for understanding the meaning of sentences and discourses. Indeed, if shallow methods focus only on text portions, deep ones reach a fine-grained analysis working on the whole meaning of a sentence or a discourse.

Deep methods represent a useful approach to extract representations and to infer on the basis of such representations. It means that this kind of analysis may contribute to inferencing and reasoning capabilities of machines through textual Web resources representation based on a machine-readable standard ontological language. Due to the need of applying an ontological language, in order to process textual resources, it is necessary to use grammar rules. Such set of grammar rules may be applied by a syntactic parser, “the first essential component for a deep analysis of texts” [8].  Indeed, syntactic parsing uses a set of grammar rules, known as syntactic grammars, in order to assign parse trees to sentences.

A formal language and its syntactic grammar rely also on a vocabulary, which includes all the acceptable combinations of characters of a specific alphabet. Such predefined vocabulary may be used in parsing sentences. Another way to create the lexical knowledge base useful to parse a sentence is based on training sets of hand-labeled sentences. This methodology represents the foundation of statistical parsers [10].

Parsing produces outputs represented in the form of phrase structure trees or dependency parses. “Phrase structure parses associates a syntactic parse in the form of a tree to a sentence, while dependency parses creates grammatical links between each pair of words in the sentence” [8].  By most syntactic theories, both formalizing methods are applied as complementary and not as opposite approaches. Many scholars [11] [12], also in shared tasks in Conference on Natural Language Learning  (CoNLL), apply dependency parsing because it allows to model predicate-argument structures  in a more intuitive way. Indeed, using predicate-argument structures for IE paradigms enables high quality IE results.

Various researches aim at establishing a correspondence between predicate-argument structure and first order predicate logic, even if this goal seems problematic. Also, according to [13], “the predicate/argument system of natural language is more complex than that of first order predicate logic”.

## IV. Framework

Most of ontology learning methodologies apply syntactic parsing, based on patterns or machine learning, to improve extraction of relevant structures. It means that syntactic parsing may allow a fine-grained analysis, guaranteeing also the extraction both of Atomic Linguistic Units (ALUs) and relations and axioms learning. The method applied must be adequate to the particular task we want to perform: extracting a whole ontology or only a constituents of such ontology, i.e., classes, relations or axioms.

Actually, various methodologies have been applied to increase retrieval and extraction system performance in different knowledge domains. The common aim is to process unstructured texts and, through semantic annotation procedures, formalize them in a structured representation. This step of converting texts represents the way in which we move to machine-readable language to systemize, manage and extract knowledge from the amount of data. Subtasks, involved in the formalizing process, concern entities and relations between them and their attributes. It means that in a text we have to analyze not only subjects and objects, which take part in a specific situation, that is discourse and sentence contexts, but also identify which kind of relation exists among them. Reconstructing the network of relations and attributes among entities lead us to reconstruct Aristotelian definition process of a concept. Thus, we get close to understand the meaning expressed in a text, which may be analyzed through a precise formalization of natural language, based on linguistic studies rather than on the development of stochastic algorithms. Due to such considerations, we apply Lexicon-Grammar (LG) methodologies in order to create Linguistic Resources (LRs) semantically annotated. LG, set up by Maurice Gross [14] during the ‘60s, is based on a language formalization achieved through a deep lexical analysis.

## V. Semantic Tagging

As presented in [15], our semantic annotation process is structured into a two-step procedure: first, we tag electronic dictionary entries, and then we develop Finite State Automata/Transducers (FSA/FSTs) in order to recognize and annotate predicate-argument structures. The utility of assigning semantic labels to words is strictly linked both to the definition of semantic predicates, and to the creation of FSA/FSTs for coherent text processing. Gross’ definition starts from the fact that, for each given language analyzed, LG can establish sets of lexical-syntactic structures, on the basis of the semantic features of each verb. These features are made explicit directly by the application of the rules of co-occurrence and selection restriction, through which verbs semantically select their arguments to construct acceptable simple sentences. Also, the arguments selected by each verb are given the value of attants (subjects included). Therefore, we may have semantic predicates expressing the intuitive notion of “exchange” (i.e., “Transfer Predicates” as “to give” or “to receive”), “motion” (“Movement Predicates” as “to go” or “to move”) or “production” (“Creation Predicates” as “to build”, “to assemble”). Each set of semantic predicates will select only and exclusively those arguments, which have with them compatible semantic roles. For instance, “Transfer Predicates” will select a “giver”, an “object to transfer” and a “receiver”, as in:

Max$_{\text{(giver)}}$ gives a present$_{\text{(object to transfer)}}$ to John$_{\text{(receiver)}}$ (1)
John$_{\text{(receiver)}}$ receives a present$_{\text{(object to transfer)}}$ from Max$_{\text{(giver)}}$ (2)

“Movement Predicates” will select an “agent of motion”, eventually an “object to move”, and a “locative name”, as in:

Max$_{\text{(agent of motion)}}$ goes to Rome$_{\text{(locative name)}}$ (3)
Max$_{\text{(agent of motion)}}$ moves the table$_{\text{(object to move)}}$ from the living room to the kitchen$_{\text{(locative names)}}$ (4)

On such basis, electronic dictionary nouns may also be labeled predicting their likelihood of becoming arguments of (a specific set of) semantic predicates. However, the list of semantic tags likely to be used is not easily identifiable, due to the polysemy of simple nouns. In fact, from the above examples, it can be seen that “agent of motion” and “creator” are sub-classes of the class “Hum”, and that “object to move”, and “creation” are sub-classes of the class “Conc” (concrete objects). Moreover, the words abbey and train can be selected by both “Motion” and “Creation”:

Max (entered + built) the (abbey + train) (5)

and also occur as human nouns:

The (abbey + train) laughed at Max’s joke (6)

On this basis, ‘abbey’ and ‘train’ could be labelled as follows:

abbey,N+FLX=APPLE+Conc
abbey,N+FLX=APPLE+Hum
abbey,N+FLX=APPLE+Loc
train,N+FLX=APPLE+Conc
train,N+FLX=APPLE+Hum
train,N+FLX=APPLE+Loc

An attempt to define a comprehensive set of semantic tags is currently in progress for the Italian DELAS-DELAF and has produced the list, presented in Table I.

TABEL I. LIST OF SEMANTIC CATEGORIES AND LABELS.

| | |
|---|---|
| NAbb: clothing article | NLud: game/sport |
| NAlimE: edible substance | NMal: illness/disease |
| NAlimP: potable substance | NMass: mass |
| NAnim: animal/animate | NMat: material |
| NArr: (piece of) furniture | NMec: object with parts to assemble |
| NAst: abstract | NMeta: non-physical/metaphysical |
| NAtmo: weather event | NMis: unit of measure |
| NBot: botanical | NMon: currency |
| NChim: chemical | NMus: musical instrument |
| NColl: collective human | NNum: numeric |
| NConc: concrete | NPc: body-part |
| NCosm: cosmetic | NPcOrg: human and/or |

| | |
|---|---|
| | animal organism |
| NCreat: creation | NPsic: psychic/psychological state |
| NDisp: device | NQual: positive/negative quality |
| NEComOr: oral communication | NQuantD: defined quantifier |
| NEComScr: written communication | NQuantI: undefined quantifier |
| NEdi: construction | NSostG: gaseous-state substance |
| NFarm: drug or medication | NSostL: liquid-state substance |
| NFig: figurative | NSostS: solid-state substance |
| NGramm: grammatical, morphological, syntactic | NStrum: mechanical tool/object |
| NLin: tongue, dialect, jargon | NTmp: defined/undefined period of time/event |
| NLiq: non-potable substance | NUm: human |
| NLoc: locative | NVeic: vehicle |

In terms of sets, the semantic features specified in this list overlap to a variable extent, especially with regard to all the possible subclasses of concrete nouns. For this reason, during the labeling of nouns, it will be possible to assign more than one tag to a single name.

## VI. CONCLUSIONS AND FUTURE WORKS

This system of semantic classification is still in an embryonic state, but it could simplify and make even more efficient the building of NooJ FSA/FST grammars [16], allowing for verbs and nouns the insertion into nodes of one or more tags, which could be used to identify classes of words instead of single words. In this sense, it could also be possible to build large-coverage grammars for single sets of semantic predicates, as the one presented in Figure 1, which accounts for 105 simple sentences of the following kind:

(He + Max) (draws + is drawing) (a picture + pictures) (7)

(We + Paul and John) (outline + are outlining) (a drawing + drawings) (8).

## REFERENCES

[1] J. Bos and K. Markert. "Combining shallow and deep NLP methods for recognizing textual entailment", in Proceedings of the First PASCAL Challenges Workshop on Recognising Textual Entailment, Southampton, UK, 2005, pp. 65-68.

[2] A. Taylor , M. Marcus, and B. Santorini, "The Penn TreeBank: an overview", in Treebanks, Springer Netherlands, 2003, pp. 5-22.

[3] M. Palmer, D. Gildea, and P. Kingsbury, “The Proposition Bank: A Corpus Annotated with Semantic Roles”, Computational Linguistics Journal, 31:1, 2005.

[4] J. Bos and K. Markert, "Combining shallow and deep NLP methods for recognizing textual entailment", in Proceedings of the First PASCAL Challenges Workshop on Recognising Textual Entailment, Southampton, UK, 2005, pp. 65-68.

[5] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval", Information processing & management 24, no. 5, 1988, pp. 513-523.

[6] S. Staab and A. Maedche. "Knowledge portals: Ontologies at work." AI magazine 22, 2001, no. 2: 63.

[7] P. Cimiano and J. Völker. "Towards large-scale, open-domain and ontology-based named entity classification." In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP). 2005, pp. 166–172.

[8] A. Zouaq and R. Nkambou. "Building domain ontologies from text for educational purposes." Learning Technologies, IEEE Transactions on no. 1, 2008, pp. 149-62.

[9] A. Giuglea and A. Moschitti, “Shallow Semantic Parsing Based on FrameNet, VerbNet and PropBank”, in Proceedings of 17th European Conference on Artificial intelligence, IOS Press , 2006, pp. 563-56.

[10] D. Klein and C. D. Manning, "Accurate unlexicalized parsing", in Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1, Association for Computational Linguistics, 2003, pp. 423-430.

[11] T. Briscoe, and J. Carroll, "Evaluating the accuracy of an unlexicalized statistical parser on the PARC DepBank", in Proceedings of the COLING/ACL on Main conference poster sessions, Association for Computational Linguistics, 2006 pp. 41-48.

[12] S. Kübler, R. McDonald, and J. Nivre, "Dependency parsing", Synthesis Lectures on Human Language Technologies 1, no. 1, 2009, pp. 1-127.

[13] E. Luuk, "The noun/verb and predicate/argument structures", Lingua 119, no. 11, 2009, pp. 1707-1727.

[14] M. Gross, “Empiric Bases of semantic-predicate notion” (*“Les bases empiriques de la notion de prédicat sémantique*”), in Langages, 15$^{e}$ année, n°63, 1981, pp. 7-52.

[15] S. Vietri and M. Monteleone, “The NooJ English dictionary”, in S. Koeva, S. Mesfar and M. Silberztein (eds.) Formalising Natural Language with NooJ 2013: selected papers from the NooJ 2013 International Conference, Cambridge Scholars Publishing, Newcastle upon Tyne, UK, 2013, pp. 69-86.

[16] M. Silberztein, “Formalizing languages: the NooJ approach” (“*La formalisation des langues : l'approche de NooJ*”). ISTE: London, 2015.

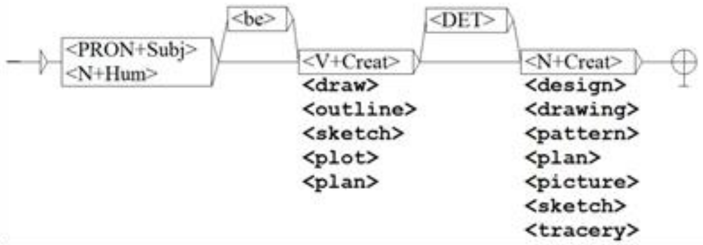


Figure 1. Sample Local Grammar for Creation of Semantic Predicates.

# Semantic Annotation to Support Decision-Making

Francesca Parisi
Dipartimento di Ingegneria Informatica, Modellistica, Elettronica e Sistemistica
Università della Calabria
Rende, Italy
francesca.parisi@dimes.unical.it

***Abstract*—Domain-specific organization management processes require particular operations and information exchange. A large amount of information concerning these tasks has to be reported and capitalized. Usually, people working in big organizations use Information Systems (IS) and other collaborative software producing a large number of information sources (data, documents, e-mails, images, etc.). The methodology presented in this paper concentrates on the contribution of semantic knowledge within textual information to improve processes carried out in domain-specific organizations. For example, activity logs, technical reports or contracts represent a part of structured textual knowledge. E-mails or texts describing activities are a part of unstructured (or semi-structured) knowledge and often contain important information to support decision-makers. Therefore, people involved in these processes need to analyze such documents to enrich their knowledge about said processes. Hence, the present methodology concerns the use of textual analysis approaches in order to evaluate their contribution to expert's activities. The experts are the human resources with specific competences involved in processes that require particular analyses. The methodology proposed is focused on an automatic semantic annotation technique considering the most important entities and their significance within specific domains. It works on the different kinds of textual knowledge (structured and unstructured) and allows to construct the most representative document classification label. In particular, the proposed method uses results of semantic annotation techniques to optimize document classification and, consequently, to support the decision-making process. This methodological proposal is the result of a work experience within a company operating in the energy sector.**

***Keywords-semantic annotation; document classification; corpus annotation; decision-making support.***

## I. INTRODUCTION

The proposal put forth in this paper was developed and refined during a case study carried out in the energy field and aims at analyzing the processes of any type of domain-specific company (e.g., companies working in the energy, manufacturing, industrial sector etc.) that needs to organize their working group, maximize task efficiency and minimize production costs. This kind of company has to solve many types of alert situations where it is necessary to be ready to act promptly to avoid negative consequences for both people and the environment.

Usually, this kind of company has to manage important problems (manual intervention, ordinary and extraordinary maintenance of technological machines) where an optimal resources intervention schedule is necessary. The experts need to be aware of all the information regarding activity sequence and the specific competences of the people involved in the process. The experts have to take decisions in a very short period of time and an error could bring negative consequences for company activities. It is necessary to explicate all knowledge typologies (structured, unstructured and semi-structured knowledge) to make information available when a particular situation is verified.

One of the most important activities for the experts is to build the entire history for each event. In domain-specific organizations, this need is stronger since there are many situations in which people have to promptly take a decision. When the experts have to analyze particular situations they need all the information about process activities and the people that are working on them. Often, a large part of information is stored in documents or reports and experts spend a lot of time looking for the relevant parts.

In particular, we consider the experts' need to rebuild the history of events in order to analyze or to find similar situations that have already been solved. This means that they have to search for both the same subject and a similar event that occurred for similar machines. This aspect is important in deciding on the significance for each domain entity to extract from texts.

The methodology proposed in this paper presents an approach to improve the search of relevant documents involved in the management of alert situations through textual semantic annotation techniques. It goes on to explain how this improvement can support decision makers and enrich the global performance of processes. It illustrates how an optimal document organization can support the experts during their analysis and how semantic annotation techniques could explicate a relevant part of important process knowledge contained in the text. The aim is to present a methodological approach usable for all kinds of texts and specific domains where analysis of the entities' semantic relatedness plays an important role.

This paper is structured as follows. Section II presents the state of art of related works. Section III describes the context and the problem statement of the methodological proposal. Section IV illustrates the aim and steps of the proposed methodology. Section V discusses future work and presents concluding observations.

## II. STATE OF ART

Text linguistic annotation consists in coding the linguistic information associated to textual data. In computational linguistics, text annotation has an important role that has been consolidated over the years. It allows computers or machines to extract, interpret and explore the linguistic structure of texts and gives an added value to single terms. From a linguistic point of view, textual data are arranged in different levels through a hierarchical organization that is often partially defined. As a consequence, text annotation is a delicate process as it gives different interpretations of the phenomena that have to be annotated based on annotation levels applied to the texts [1].

The proposed methodology uses semantic annotation techniques to extract concepts from the specific domain texts where the general meaning within a specific domain is important to explicate linguistic entities.

This section shows a short overview of the related work. In particular, let us refer to the semantic annotation techniques applied in different domains, for different aims and on different types of texts.

Due to the large amount of literature concerning Natural Language Processing (NLP) techniques, this section is focused on a representative set for the presented work. The aim is showing how NLP is used in different specific domains, mainly referring to the medical diagnosis and business processes. Concerning the methodology presented in this paper, this section aims at illustrating how it was defined and also how, with examples of NLP applications in other domains, the guidelines and criteria to follow were identified.

For example, in the medical domain NLP techniques have had a large application in structuring clinical information and making available codified diagnosis information so as to understand a natural specific domain language used in the text [2]. In another application in the medical domain, an annotated corpus (PhenoCHF) was created to better understand the medical sub-domain of congestive heart failure (CHF) [3]. The corpus focuses on the identification of phenotype information for a specific clinical sub-domain, congestive heart failure (CHF) and the annotation scheme, whose design was guided by a domain expert, includes both entities and relations pertinent to CHF. Extracting phenotype information can have a major impact on our deeper understanding of disease etiology, treatment and prevention. This is a case in which a corpus is annotated in order to complete the domain-specific vocabulary and to understand a possible evolution of the phenomena.

In general, there are a large number of NLP medical domain applications as the language and the context are more difficult to understand; there are linguistic and contextual factors to consider and language is subject to continuous evolution. Such as in the domain presented in this paper, the medical domain is an interesting similar application that offers many points for reflection to define the following methodology. The medical domain is similar to the domain considered in this paper for the importance of entities' semantic relatedness for analyzing diagnostic processes.

In [4], a platform named MeTAE (Medical Texts Annotation and Exploration) to extract medical entities and relationships from medical texts is presented. Determining the categories of the medical entities identified in the text is difficult. Such as in the domain presented in this paper, one of the most important obstacles is the high terminological variation to express the same concept (ex. Swine influenza = swine flu = pig flu). The semantic annotation application is also used to understand medical problems and to maintain the problems lists accurate and up-to-date. Indeed, in [5] an NLP application to extract medical problems from narrative text clinical documents is presented. The algorithm has extracted 80 different medical problems selected for their frequency of use. Within a set of documents, it identifies for each document a series of problems treated, selecting the medical entities from the sentences. This methodology also works on negation detection to explicate both what a patient has and what a patient does not have. The same disease could be part of a positive sentence (ex. "The patient is known for diabetes mellitus") but also in the negative sentence (ex. "No diabetes is reported in the patient's history") so the application may be able to recognize specific linguistic context.

As mentioned above, the medical domain offers an important starting point to reflect on the diagnostic processes analyzed in this paper. The diagnostics process for technological machines could be compared to the patients' disease diagnosis in the medical domain. The problems list is important also in the technical domain (such as the energy domain) and the same concept can be expressed in different linguistic forms. It is not sufficient to identify the specific entities but it is necessary to analyze their context in the sentences. This application on the texts in the specific domain could be an important starting point for improving technical vocabulary that often, mainly in the free texts (such e-mails, narrative description etc.), is not used in their typical forms. In addition, these NLP techniques in the medical domain play an important role also in clinical decision making support such as explained in [6].

As already noted, NLP techniques have a wide variety of domain applications. In the business intelligence domain, the enterprise can use these textual techniques to capture information and opinions contained in different kinds of sources. An experiment to identify lexical, morpho-syntactic, and sentiment-based features derived from web sources is presented in [7]. The aim was to collect all opinions about the company and analyze what has been said about company. To do that, business analysts need to analyze textual sources and accordingly make decisions about business actions. The experiments use NLP techniques to identify if the sentences refer to positive or negative opinions. These techniques are inserted in the business lifecycle as a strategic monitoring phase.

Another example of NLP techniques applied to Business Intelligence is described in [8]. A system which allows extracting relevant information to be fed into models for analysis of financial and operational risk and other business intelligence applications is described.

Let us consider that NLP techniques and information extraction from texts have a strategic role in making textual knowledge available. This kind of knowledge is often important to understand domain evolution (in terms of vocabulary used, opinions, contexts), support the decision-makers that need to have available and formalized all kinds of knowledge, and also to predict future actions and strategies.

These examples are aimed at giving a general idea on NLP applications inserted in specific domains and useful for specific processes and objectives.

The proposal presented was inspired by this kind of experiences but with the consequence to optimize classification documents crucial for the diagnostic processes carry out in a specific domain. The methodology proposed in this paper refers to specific domains in which textual information has to be inserted in diagnostic process as the guidelines to promptly find relevant documents. It considers not only the domain specific entities present in the text and their relevance but also the relatedness within them (e.g., temperature – pressure, plant n°1- plant n°2, etc.). In this way, people involved in diagnostic processes can find the correct information and the information related whit their initial searches in the shortest time.

## *III.* METHODOLOGICAL PROPOSAL

### *A. Context*

The work proposed in this paper has an important application in domain-specific organizations. The proposed methodology can be applied in domain-specific organizations due to the important role that technological capital plays for these enterprises. In particular, this assumption is based on the diagnostic process provided in companies operating in the energy sector.

The present methodological proposal concentrates on diagnostic problem processes that require a strong expertise exchange within the experts involved and where there is a large part of textual knowledge to explicate. The diagnostic activities represent all operations carried out by experts to determinate the main causes and the nature of defects verified on technological machines or other support. Let us consider the diagnostic processes where there are two main typologies of human resources that communicate with each other: the technicians for the manual intervention and maintenance, and the experts with specific competences that have to analyze monitoring data and parameters (Fig. 1).

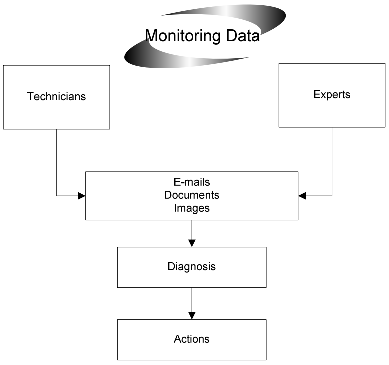


Fig. 1: Diagnostic process

With their expertise exchange, they produce unstructured information sources (e.g., e-mails, graphs, images) that have to be analyzed for formulating precise diagnosis quickly. In order to carry out these analyses and search for solutions, an optimal text classification could be an important aspect to improve complex diagnostic processes. The assumption is that document classification plays a fundamental role in experts' activities while searching for relevant information that sometimes is unstructured and difficult to obtain.

### *B. Problems Statement*

This section illustrates the problem statement and outlines the kind of processes in which the proposed methodology can be applied. Let us consider the diagnostic process to identify problems on technological machines and equipment or the maintenance process carried out to repair defects or malfunctions. They are two different kinds of processes; however, both impose the analytical phases.

Let us take, for example, the problems that may occur in a power plant or nuclear power plant where an in depth analysis is needed and a solution must be found quickly.

All pieces of these plants are controlled with specific monitoring tools and all particular situations are reported in specific documents (e.g., maintenance report, check-ups). The experts' activities concern the examination of monitoring values and the diagnostic problems after consulting with other actors involved in the process. When the experts verify alert situations in the parameters concerning the technological pieces of the energy plants, they begin an information exchange using specific information support (databases, e-mail exchange, documents, images etc.) and creating different kinds of information sources.

The methodology presented here concentrates solely on the textual information concerning particular events which is sometimes unstructured and difficult to find.

The semantic annotation approach presented in this paper gives a significance value to each domain-specific entity. In this way, the document classification labels will be formed of relevant domain entities and will better represent the texts' content. The general concept consists in assigning a significance indicator to each entity, consequently building the classification labels considering these criteria. The aim is to create a representative label with the entities based on a relevant importance indicator and not only on frequency.

Let us remember that the methodological aim is to create document classification labels to allow finding relevant documents for alert situations that could be verified in a specific domain. This relevance indicator could be established a-priori considering specific domain language and the specific aims, which will be explained in the next section.

To sum up, the assumed diagnostic processes are structured as follows:

1) *All pieces of energy plants or technological machines are always monitored. People with special competences have the role to verify potential anomalies.*

2) *Monitoring tools show the anomalies in the values of parameters or that machines do not function properly.*

3) *The experts begin the information exchange using specific information support. They produce specific documents such as reports or write e-mails to consult other actors involved in the process.*

ISBN: 978-1-61208-457-2

*4) This textual information is capitalized in specific document bases that should facilitate the users' search for relevant documents. In particular, documents are organized according to specific classification schema that has to explicit the real content of the texts.*

*5) When a particular situation is verified, the experts have to search relevant documents to rebuild the history of events.*

Therefore, the proposal outlined here, has the aim to improve document classification in order to find relevant documents in less time. In order to do so, the methodology proposes to use semantic annotation techniques to explicate the real contents of the texts with a domain-specific language.

## *IV.* METHODOLOGICAL APPROACH

Let us imagine a sample of documents (structured or unstructured texts) that need to be classified based on their content, and suppose that such documents are composed of terms used in a specific domain. Let us consider that a part of these documents already have a classification and are already organized with the appropriate metadata. Let us consider also that the terminology that has to be found in the documents is composed of single and composite terms and an associated importance indicator has to be determined for each entity.

In this paper, the list of domain vocabulary is referred to as "domain-specific terminology" and it is searched for in each text included in the corpus.

The paper proposes an application of semantic annotation techniques to detect representative features for each document and optimize their classification where their consultation is a strategic part of an important process. It works on domain-specific language contained in structured or unstructured texts and aims to identify the most representative features based on significance indicators. These indicators will allow working with a reduced number of features considering the terms' distributional semantics in analyzed context [9] and their semantic similarity [10]. Examples of methods for features reduction are presented in [11].

Specific semantic annotation rules are used to identify the composite terms and suggest how this annotation can facilitate document classification and consequently contribute to improving processes aimed at decision making. In this way, textual knowledge extracted through semantic annotation techniques could be considered an important support during processes that require all kinds of knowledge available and formalized. For this reason, the methodology is applied when textual information and document classification play an important role in the processes.

For the diagnostic processes presented, the specific documents and textual information produced contain a great deal of expertise exchange and fundamental rules to solve particular situations.

### *A. Text pre-processing*

In a first phase, it is necessary to identify a "bag of words" characterizing the domain-specific terminology. Let us consider that any structured language is used and any domain ontology is built for the analyzed domain. Let us suppose that a specific enterprise has a proper vocabulary (ex. particular name for machines and pieces) that sometimes is difficult to explicate for the domain inexpert.

The specific domain entities are identified considering the specific terms used within specific enterprises to carry out their activities. For example, the specific machine codes, place with a specific definition ("place name + plant number"), specific machines' piece names ("piece name + piece number"), etc. For each of these entities a significance indicator is determined for building a classification label. It is possible to associate a different significance indicator to the terms and build a classification label according to the final goal (e.g., it could be important to rebuild the history of the parameters or places but parameters are more important than places).

The corpus is annotated according to Part-of-speech techniques and specific annotation rules are created in order to find all the elements in the domain-specific terminology list. In particular, in a first phase the methodological proposal consists in applying semantic annotation techniques using Part-of-speech tagging tools.

Part-of-speech (POS) tagging is one of the most popular and thoroughly researched tasks in the field of natural language processing, particularly since it is a prerequisite for a wide variety of more complex tasks [12].

This analysis gives the grammatical category for each term contained in the texts based on the text language. Subsequently, the methodology provides the construction of a set of semantic annotation rules to identify word features of the specific domain.

Each term contained in the domain-specific terminology is searched for in the documents regardless if they are unstructured or structured. Hence, the proposed method allows the possibility to classify all kinds of textual information involved in processes with an important analytical phase. For each document, a word vector characterizing the texts' content is identified.

### *B. Features Selection*

The identified terms represent the features for each document that are subsequently used to form the classification label formed by terms having greater significance. A selection from the word vector associated to each document is carried out according to domain and entities relatedness. Let us consider finding a list of terms associated to a document; and let us suppose that the list of terms to find in the documents is divided into categories (e.g., parameters, places, machines etc.). According to the method presented, if there is more than one term belonging to the same category in the analyzed text, the most frequent term along with the one with the greatest semantic relatedness are chosen to represent the text content.

Indeed, the classification classes that the method presented here aims to create are based on the texts' content and related tags. The approach provides a textual analysis (in particular semantic annotation) to explicate the real content of documents involved in analytical phases of processes that require prompt actions. The application of this work could be an automatic and dynamic classification schema capable of identifying in real time the correct label of documents. This approach could improve classification schema precision, minimizing search time and facilitating the access to relevant documents. The results obtained could benefit decision support makers because of their need to have knowledge available of past and connected events.

The classification labels built with this methodological approach could help event trend analysis and explicate the important knowledge contained in the texts.

### C. Example Case Study

What follows is an illustration of an example case study in which the proposed methodology is being applied. A corpus of e-mails exchanged among the different actors involved in diagnostic or maintenance processes in a domain-specific organization has been analyzed. In particular, approximately 2000 e-mail conversations are considered; the POS tagging has been applied and the specific domain entities are in part identified.

This organization, just like many others, has a technological capital that has to be constantly monitored to repair defects and avoid serious damages.

Each e-mail text has to be classified based on its content so as to facilitate future access to information. In these unstructured texts, they explain particular situations that require a diagnostic analysis and how they solved past problems; this however, is not the only content. As already mentioned, the experts control the values of parameters of the technological machines through specific monitoring tools. When they verify a particular situation that requires a discussion, they start an information exchange via e-mail to find a correct diagnosis. During this phase they have to study the history of the event in question searching in previous documents or e-mails related to this event.

This expertise exchange is capitalized in a document base and each conversation (a set of e-mails that should be referred to the same subject) is classified with a label and metadata. Sometimes the e-mail content was not compatible with the metadata associated to the specific conversation because relevant elements could appear successively.

These e-mail texts already have a classification schema based on the e-mail subject or experts' personal evaluation that often do not represent the real content of the messages. Let us take into consideration the particular situations described in the texts with the widest variety of terms possible that could be found within them.

Consequently, the actors involved spend a lot of time searching for pertinent documents and a reduction in the time spent could bring an important improvement to the general diagnostic process. In particular, they have to find similar events in the past related on the same machine, all events concerning the particular piece analyzed or similar events concerning the same piece but belonging to another machine with the same technological structure.

Considering this example case study, the methodology proposed aims at creating a classification schema that allows to explicate the real contents and that considers all important elements in the text. For this reason, it is extremely important to determine the list of domain-specific terminology and their significance: to capture as many domain-specific terms as possible in the text Compound terms are identified subsequent to POS tagging and the parsing techniques [13] and through the automatic rules used in GATE (General Architecture for Text Engineering) software with the JAPE language [14]. With this platform it is possible to build specific rules starting from general categorization obtained with POS tagging. The JAPE language allows to determine a grammar for GATE to annotate not only the named entities but also the domain-specific items.

The extracted terms represent the candidates in setting classification labels. The label associated to each text will be formed considering the term frequency but also the importance indicators associated to each term.

To sum up, the main methodological steps are structured as follows:

*1) Identifying a process where documents search and consulting have a crucial role.*

*2) Determining the aims for semantic annotation and the relevant aspects to search in the documents*

*3) Structuring the set of specific domain items (with typical vocabulary used in the specific company) to search in the texts and determining for each of these a significance indicator for selecting the representative features.*

*4) Applying the POS tagging technique.*

*5) Building the specific rules to annotate the texts and find specific domain entities.*

*6) Verifying for each document the vector of words associated and selecting the entities with the greatest significance indicator.*

*7) Determining for each document the classification label sorted by their significance indicator and their respective class.*

## V. CONCLUSIONS AND PERSPECTIVES

This paper has outlined the most important aspects of semantic annotation techniques applied in a specific domain. In particular, it has presented the use of semantic annotation to support decision making within processes that require important analytical phases and where document consulting plays an important role in specific processes. The presented annotation method aims to improve document classification in order to help experts who need to find relevant information in a relatively short period of time. It has explained how this could be an important contribution in improving the global process that is being analyzed. In particular, it has based its assumption on the real need identified after an experience in a domain-specific organization.

In this work it would be worthwhile to use specific optimization algorithms in order to evaluate the classification schema found through semantic annotation techniques. The futures goals will be to test the classification quality using semi-supervised classification algorithms. In particular, the aim is to represent each document in the n-dimensional space where n is the dimension of bag of words. For each text, a vector of n elements will be created. It represents the absence or presence in the text of each word contained in the bag of words. Subsequently, the proposal is to verify the classification with a semi-supervised algorithm.

## REFERENCES

[1] A. Lenci, S. Montemagni, and V. Pirrelli, “Testo e computer: elementi di linguistica computazionale”, Carocci, May 2005, ISBN:88-430-3425-1.

[2] C. Friedman, H. George, "Natural language processing and its future in medicine", Academic Medicine 1999, 74.8, pp. 890-

895.

[3] N. Alnazzawi, P. Thompson, and S. Ananiadou, "Building a semantically annotated corpus for congestive heart and renal failure from clinical records and the literature", LOUHI 2014, pp. 69-74.

[4] A.B. Abacha, P. Zweigenbaum, “Automatic extraction of semantic relations between medical entities: a rule based approach”, J. Biomedical Semantics 2011, 2(S-5), S4.

[5] S. Meystre, P.J. Haug, “Natural language processing to extract medical problems from electronic clinical documents: performance evaluation”, Journal of biomedical informatics 2006, 39(6), pp. 589-599.

[6] C. Demner-Fushman, W. W. Chapman, and C.J. McDonald, “What can natural language processing do for clinical decision support?”, Journal of biomedical informatics 2009, 42(5), pp. 760-772.

[7] H. Saggion, A. Funk, “Extracting opinions and facts for business intelligence”, RNTI Journal 2009, E (17), pp. 119-146.

[8] D. Maynard, H. Saggion, M. Yankova, K. Bontcheva, and W. Peters, “Natural language technology for information integration in business intelligence”, Business Information Systems, Springer Berlin Heidelberg, January 2007, pp. 366-380.

[9] A. Lenci, "Distributional semantics in linguistic and cognitive research", Italian Journal of Linguistics 20.1, 2008, pp. 1-31.

[10] T. Cohen, D. Widdows, “Empirical distributional semantics: Methods and biomedical applications”, Journal of biomedical informatics, 42(2), 2009, pp. 390 -405.

[11] Y. Yang, J.O. Pedersen, “A comparative study on feature selection in text categorization”, ICML 1997, Vol. 97, pp. 412-420.

[12] H. Van Halteren, ed. “Syntactic word class tagging”, Kluwer 1999.

[13] G. Kennedy, “An introduction to corpus linguistics”, Routledge 2014.

[14] K. Bontcheva, V. Tablan, D. Maynard, and H. Cunningham, "Evolving GATE to meet new challenges in language engineering", Natural Language Engineering 2004, 10.3-4, pp. 349-373.

# An Ontology-based Method for Discovering Specific Concepts from Texts via Knowledge Completion

Céline Alec, Chantal Reynaud-Delaître, Brigitte Safar
LRI, Univ. Paris-Sud, CNRS, Université Paris-Saclay, Orsay cedex, F-91405, Email: firstname.lastname@lri.fr

***Abstract*—Heterogeneity in user's queries and data sources can easily cause problems in perceiving sufficient information to form correct answers. In this paper, we address this issue when data sources are unstructured short texts describing only key characteristics of concerned individuals but when keywords in user's queries are customized concepts. To bridge the gap between texts and user's concepts, we propose an ontology-based approach, named SAUPODOC (Semantic Annotation Using Population of Ontology and Definition of Classes), to discover formal definitions of specific concepts via population of property assertions. Property assertions are extracted from texts but the texts under our consideration are incomplete, i.e., information about the target concepts is missing. To solve this problem, we further propose a method to extract property assertions by exploiting LOD (Linked Open Data) datasets to deal with missing and multiple values. Experiments have been carried out in two application domains, whose results show a clear benefit of SAUPODOC over well-known classifiers.**

***Keywords*–discovering concepts from texts; ontology enrichment; ontology population.**

## I. INTRODUCTION

A lot of information is accessible over the Internet but discovering relevant information for users still poses today research challenges. This is mostly caused by heterogeneity in user's queries and in data sources. In this paper, we address this issue when data sources are unstructured documents describing individuals only by their key characteristics and when keywords in user's queries are customized concepts. We have to handle format heterogeneity but, most of all, content heterogeneity. One frequently used approach to semantic heterogeneity is to rely on Semantic Web techniques, particularly on ontologies. We are adopting this direction of research.

Ontologies, which are formal specifications of a domain of interest, specify the meaning of concepts in a semantically well-founded way and can be processed by machines. They are a key component to address our problem, which can be defined this way: "Given a customized concept $C_c$ only defined as being a specific concept of $C$, and given a set of individuals $I$ instances of $C$, find the subset of $I$ that groups instances of $C_c$. Moreover, if that set is empty, find the subset of $I$ that fulfills only partial properties of $C_c$". Solving this problem requires knowing definitions, i.e., the property restrictions of each customized concept and the property assertions for each considered individual. There are two reasons explaining the need for discovering precise definitions of customized concepts. The first is that it will find out individuals satisfying the definitions by exploiting the values of their properties. Secondly, definitions can be used by reasoning to provide users with partial satisfactory proposals. This point is crucial in our work. Consequently, the approach presented in the paper aims at enriching/populating a domain ontology both with formal definitions of concepts and with property assertions, which can be solved by ontology learning techniques at first glance.

However, no single approach in the ontology learning state-of-the-art addresses our problem. First, approaches dealing with generating formal definitions of concepts from texts describing generic domains [1] are not applicable because the content of our texts is not adapted. Second, property assertions in our texts are incomplete with respect to those required for defining customized concepts. Thus, extraction of property assertions from given texts only would not be enough to apply an approach at the instance level as in [2][3]. Third, the majority of ontology learning tools aiming at building a lightweight ontology extract only ontological elements [4] recognized by terms in texts. These techniques are inapplicable too. Our texts do not include names of concepts or instances. Finally, some ontology learning work deals with texts close to ours [5][6] in the sense that they involve properties of instances without naming the underlying concept. However, they assume that the precise definitions of concepts corresponding to individuals described in texts are known in advance. It is not the case in our work either.

In this paper, we investigate how several approaches can be combined in order to jointly contribute to address our problem. The proposed approach, called SAUPODOC, relies on a domain ontology relative to the field under study, which has a pivotal role, on its population by property assertions, and on automatic generation of formal concept definitions from the enhanced ontology. Our contributions are the following. We first design the SAUPODOC approach combining various tasks. Second, as property assertions extracted from texts are incomplete with respect to those required for defining customized concepts, we propose techniques to exploit LOD datasets in order to obtain further property assertions, which deal in particular with missing and multiple values. Obtaining appropriate values for all properties required to define customized concepts is indeed a critical issue solved in SAUPODOC. Finally, we experiment our approach in two application domains. We analyze the results and demonstrate the relevance of such a combined approach compared to well-known classifiers.

The remainder of the paper is organized as follows. Section II presents a motivating example of the approach. Section III exposes some related work. Section IV describes our approach. Section V presents experiments to evaluate the approach. Section VI concludes and outlines future work.

## II. A MOTIVATING EXAMPLE

Let us consider a Business to Consumer (B2C) company in the holiday destination domain, accepting products from several suppliers and proposing to users the most appropriate products according to their needs. Usually, the needs are expressed from a point of view significantly different from that of product suppliers and change over time. These companies must design applications making searching products easy, i.e., offering totally satisfactory products or, if that is impossible, partially satisfactory products. Such applications have to be designed for a wide range of products. These application requirements have motivated our work in the context of a collaboration with the Wepingo start-up.

In this settings, textual documents are descriptions of destinations extracted from advertising catalogs or websites. They describe the main features of destinations, praise their virtues and include hardly any negative expressions. Users may be interested in destinations where they can do water sports during winter ($DWW$), with nightlife ($DWN$), or in cultural destinations ($CD$). $DWW$, $DWN$ and $CD$ are what we call customized concepts. They are specific in regard to the general concept Destination. There is no in-line travel catalogs, or knowledge bases indexing Dominican Republic as $DWW$. As an illustration, here are two excerpts of the description of Dominican Republic coming from Thomas Cook website: "[...] especially loved by scuba divers. Over 20 existing diving sites and 3 old shipwrecks are waiting to be discovered" and "[...] get active with a range of water sports". These two description parts do not reveal whether Dominican Republic is a $DWW$ or not. One part includes the terms "scuba divers" and "diving", the other one mentions the term "water sport" but is it possible to practise water sport during winter? What is the weather in winter? There is nothing about that. Information in these texts is incomplete and can not be used to infer if the described entities are instances of $DWW$.

The SAUPODOC approach aims at discovering the specific concepts whose instances are described in textual documents. It requires property assertions, $pa$, for all considered individuals. Some $pa$ as practised activities can be extracted from textual documents and other required $pa$ as temperatures or precipitations about which the texts are silent will be extracted from another source. Based on the enhanced ontology, definitions of specific concepts can be then discovered. An example of definition for $DWW$ is "a destination hot enough in winter (a mean temperature exceeding 23°C) and with little precipitation (less than 70 mm) to practise water sport activities". Dominican Republic totally satisfies that definition. More generally, all destinations satisfying the formal $DWW$ definition will be part of the answer to a query with the keyword $DWW$. By contrast, Bali, which does not satisfy the constraint about precipitations, will be a partial satisfactory proposal delivered in the absence of totally satisfactory answers.

## III. RELATED WORK

We distinguish three categories of work focusing mainly on semantic knowledge extraction from unstructured texts.

The first addresses learning expressive ontologies in favor of applications based on ontology reasoning. Some work in this category applies on texts containing concepts but not instances. For instance, LExO [1] applies syntactic transformation rules to generate DL axioms from definitory natural language sentences. Ma and Distel investigate a way to learn concept definitions via a relation extraction based approach [7] and propose formal constraints in [8] to ensure the quality of the definitions. It is not appropriate to apply those approaches in our work because the content of the texts is not adapted. By contrast, two research works [2][3] seek to generate a logical description from instances. They rely on the inductive logic programming technique to find a new concept description from assertions of an ontology. [3] applies to expressive DL ontologies, where [2] propose a system for light-weight DL ontologies. The main drawback of those approaches is that they require a large amount of facts about individual entities when applying to real world ontologies. Compared to [2][3], our inputs are texts, not assertions in an ontology about individuals. Needed assertions are expressed in unstructured texts but incompletely.

The second category addresses generating lightweight ontologies limited in their expressiveness, which often consists of taxonomies. Research work investigates how to extract various ontological elements that are learned from textual resources [4]. In regards to concept extraction, a main step is extracting the relevant domain terminology [9] using different term weighting measures. Clustering techniques can then be applied to detect synonyms. An ontological class can be derived from each group of similar terms. Researchers have also investigated learning concept hierarchies from texts. They mainly apply unsupervised hierarchical clustering techniques in order to learn subclass relations and concepts at the same time [4]. Approaches where patterns are identified in the texts are other applied techniques [10] although they discover lexical relations between terms, not between concepts. Finally, when the ontology has not to be built from scratch and when a concept hierarchy already exists and has to be extended with new concepts, supervised methods become possible as well. Classifiers need to be trained for each concept in the existing ontology. The number of those concepts cannot be very large. Unsupervised approaches applied in this settings use an appropriate similarity measure to compare a new concept with those already in the ontology [11]. All these research work seeks to recognise terms denoting concepts (or instances) in texts, and then extract them. However, sometimes texts involve properties of instances without naming the underlying concept, as in our work. Other approaches (3rd category) are then necessary.

The third category includes work using reasoning as a partial replacement of the traditional techniques of information extraction. In the BOEMIE system [5], concepts are divided into primitive and composite concepts. Primitive concepts are populated by classical extraction tools. Instances of composite concepts can not be found in texts but rather their properties. Consequently, composite concepts, defined in terms of primitive ones in the ontology, are populated by reasoning over primitive instances. In [6], the authors extract facts from texts thanks to an ontology and natural language processing tasks. New facts, not explicitly mentioned in texts, are then derived from extracted facts and ontology knowledge. Reasoning is based on background knowledge and inference rules given in advance. Compared to [5][6], we work on texts with a close content but the point is that we have no definitions of concepts that have to be populated.

This state-of-the-art shows that none of these approaches taken in isolation is the solution. However, some can help provided they are adapted to our requirements as outlined in the following section.

## IV. THE SAUPODOC APPROACH

The tasks performed in SAUPODOC cooperate via an ontology defining the domain knowledge and populated/enriched little by little by property assertions, definitions and then individuals of customized concepts. As textual descriptions of entities are often incomplete in regards to required property assertions, collected data has to be complemented. We propose to exploit LOD datasets. The knowledge engineer has in charge to choose which properties can be populated from texts and those which can be populated from other sources. Definitions of customized concepts are then derived based on the populated ontology and applied to obtain their individuals. We present the general approach and each of the tasks that it comprises.

### A. *An ontology-based approach*

The ontology, an input of SAUPODOC, contains all elements defining entities in the application field. It is domain specific but approach independent. Indeed, the only constraints imposed by the approach are described in this subsection. The ontology can therefore be largely reused, or (semi-) automatically built, its creation is not the focus of the paper. More formally, the ontology $\mathcal{O}$ is an OWL ontology defined as a tuple ($\mathcal{C}$, $\mathcal{P}$, $\mathcal{I}$, $\mathcal{A}$) where $\mathcal{C}$ is a set of classes, $\mathcal{P}$ a set of (datatype, object and annotation) properties characterizing the classes, $\mathcal{I}$ a set of individuals and property assertions, and $\mathcal{A}$ a set of axioms including constraints on classes and properties: subsumption, equivalence, type, domain/range, characteristics (functional, transitive, etc.), disjunction. Figure 1 shows an excerpt of an ontology in the holiday destination domain. The classes Activity, Environment, FamilyType and Season are respectively the roots of a hierarchy, e.g., Environment expresses the natural environment (Aquatic, Desert, etc.) or its quality (Beauty, View). Some object properties represented on the figure have subproperties. Datatype properties are represented under their domain class. Individuals are not represented on the figure.

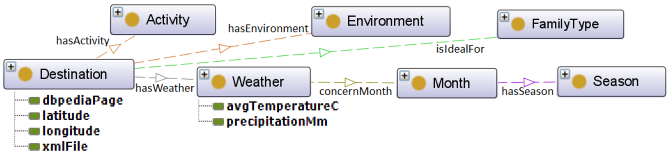


Figure 1. The structure of the destination ontology

Our approach is based on the fact that the names of customized concepts are known by the knowledge engineer but he has no precise definitions. The only thing that he knows is that they are more specific than more general ones including numerous individuals. What we are proposing in our approach is to introduce that knowledge in the ontology. Classes in the ontology are then distinguished as follows:

- ***the main class*** ($MC$) corresponds to the general type of entity considered.
- ***the target classes*** ($TC$) represent customized concepts. They have a name but no definition.
- ***the descriptive classes*** ($DC$) are all the other classes included in the ontology.

We add a subclass relationship between each $TC$ and $MC$. $\mathcal{I}$ initially contains individuals being instances of $DC$, and annotation property assertions.

The SAUPODOC approach will gradually complete the ontology as follows (cf. activity diagram Figure 2):

- by individuals of $MC$ representing the entities in the corpus, an individual being created for each considered document.
- by property assertions: either properties of individuals extracted from texts or from other (possibly several) sources;
- by definitions of $TC$ specifying their specificities in regard to all the properties of $MC$, this step being only necessary the first time documents of a given domain are addressed;
- by individuals of $TC$.

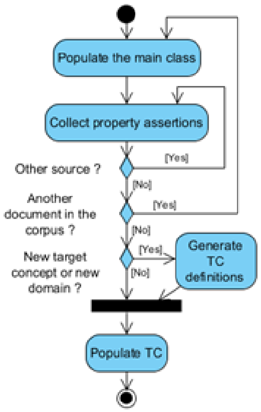


Figure 2. Activity diagram

### B. *Extraction of property assertions from texts*

This step consists in annotating texts given $\mathcal{O}$ where all concerned properties are represented and then in representing those annotations in $\mathcal{O}$ as property assertions. We assume that texts do not include negative expressions that could disrupt the process. Thus, a simple information extraction system is appropriate provided the system can take $\mathcal{O}$ as input.

A property assertion states that an individual is connected by a property to another individual or a literal. Property assertions added in $\mathcal{O}$ have to follow this format. The extraction step is guided by $\mathcal{O}$, which includes terms corresponding to labels of individuals and allows extraction of property assertions related to individuals. For instance, property assertions connected to "snorkeling" can be added because "snorkeling" is an individual in $\mathcal{O}$, associated with labels that are used in texts to refer to it. The introduction of property assertions into $\mathcal{O}$ is guided by object properties and their associated range constraints expressed in $\mathcal{O}$. For example, based on the statement <Destination, hasActivity, Activity>, which asserts that the range values of the property hasActivity must belong to the extension of the class Activity, if the description of an entity $e$ contains a match with an individual $a$ instance of Activity, then <$e$, hasActivity, $a$> is introduced in $\mathcal{O}$.

In our work, we use GATE [12], an open source software performing a lot of text processing tasks. The GATE resource OntoRoot Gazetteer can produce annotations over textual documents w.r.t. an ontology given as input, in combination with other generic GATE resources. The JAPE transducer applies JAPE rules in order to transform annotations to property assertions. Rules are automatically created from one pattern, a rule per object property having to be populated. The same pattern is used whatever the ontology.

### C. *Extraction of property assertions from the LOD*

Exploiting datasets coming from the LOD is interesting because number of those datasets can easily be queried thanks to SPARQL endpoints. Our goal is to build Construct SPARQL queries to get data from the LOD and add property assertions in our ontology. At first, the knowledge engineer has to find the most relevant datasets including information relative to the entities in the corpus. Then he has to recognize within those datasets data that corresponds to what is required. As the number of concerned properties is not too large, this task can be done manually. However, vocabularies differ in $\mathcal{O}$ and in the LOD datasets and correspondences may be very complex. We analyze this mapping process by giving first a few examples.

The first example illustrates a *1:n* correspondence. A property $prop$ characterizing $c$ in $\mathcal{O}$ may be associated to several equivalent properties but with a different syntax and having possibly a value expressed with a different unit of measurement. For example, the property "precipitationMm" for January in a given place in $\mathcal{O}$ is represented in DBpedia amongst others by dbp:janPrecipitationMm, dbp:janRainMm, dbp:janRainInch. The second example is also a *1:n* correspondence but the properties are not equivalent to each other and are all needed for calculating the value of $prop$. For example, the temperature for any month in a given location can be obtained in DBpedia with the average between the highest and the lowest temperatures during that month. The third example is about missing values. While DBpedia is a huge knowledge base, it is also incomplete. A lot of property assertions are missing. For example, a given destination may have no value for the highest temperature during a given month.

We have to consider all these various situations including multiple properties, missing values together with multivalued properties, i.e., one property with possibly different values. Although incompleteness is not a problem in Linked Data under assumptions of many research works, having complete information is essential in our case in order to achieve the best results as output of the entire process. Consequently, we define a model to specify all the possible correspondences between a property $prop$ characterizing $c$ in $\mathcal{O}$ and related properties in a RDF dataset. We define also a model to specify alternative access paths to properties in case of missing values.

*1) Specification of correspondences between properties:* The correspondences are the result of a matching process between a source ontology $\mathcal{O}_s$ and a target ontology $\mathcal{O}_t$, more precisely between an OWL ontology and another one providing access to a RDF data source. The correspondences are relationships between Property Expressions ($PE$) expressed as triples ($id$, $PE_s$, $PE_t$) where $id$ is the correspondence identifier, $PE_s$ is a property expression in $\mathcal{O}_s$ and $PE_t$ is a property expression in $\mathcal{O}_t$.

A property expression in $\mathcal{O}_s$ ($PE_s$) is either an object property ($op$) or a datatype property ($dp$) possibly including restriction constraints on its domain ($PE_s.Constr(d)$) using any ontological constraint from the time it can be expressed in OWL DL, cf definition 1.

**Definition 1.** $PE_s$ = $op$ | $dp$ | $PE_s.Constr(d)$

**Example** The datatype property "precipitationMm" with its domain "Weather" constrained by <Weather concernMonth January> is a $PE_s$ that matches the property dbp:janPrecipitationMm in $O_t$, i.e., precipitationMm.<Weather concernMonth January> $\in PE_s$.

A property expression in $\mathcal{O}_t$ ($PE_t$ cf definition 3) is either an elementary property ($p_e$) in $\mathcal{O}_t$, a mathematical, set-theoretic, transformation or agregation expression (f) using property expressions in $\mathcal{O}_t$. A $PE_t$ can include domain or range typing constraints ($PE_t.Constr$). An elementary property $p_e$ (cf definition 2) is either a property in $\mathcal{O}_t$ or its inverse.

**Definition 2.** $p_e$ = $op$ | $dp$ | $op^{-1}$

**Definition 3.** $PE_t$ = $p_e$ | *f*($PE_t$) | *f*($PE_t$, $PE_t$) | $PE_t.Constr$

**Example** AVG(UNION(dbp:janPrecipitationMm, dbp:janRainMm)) is a $PE_t$ whose value is the average of all the values of dbp:janPrecipitationMm and dbp:janRainMm.

The knowledge engineer defines correspondences between properties in $\mathcal{O}_s$ and $\mathcal{O}_t$ based on this model and on knowledge in $\mathcal{O}_s$. That way, if $PE_s$ in $\mathcal{O}_s$ is related to a functional property and its correspondence $PE_t$ in $\mathcal{O}_t$ is multivalued, an aggregation function is applied to obtain a single value.

*2) Specification of access paths:* The correspondence model requires access to $p_e$ values in the target dataset but some of them may be missing. For instance, no precipitation data exists in the Kefalonia page. To solve this problem, we define the notion of *i*th*-Order property* (cf definition 4).

**Definition 4.** *An i*th*-Order property ($p^i.p^{i-1}...p^1$) is a property that can be reached from the initial page through a path of length $i$ in the data graph.*

**Example** "janPrecipitationMm", a property of Abu Dhabi with the value 7, is a 1st-Order property ($p^1$) w.r.t. Abu Dhabi.

**Example** "country.capital.janRainMm" with "janRainMm" a property of Athens with the value 56.9, Athens being the capital of Greece, which is the country of Kefalonia, is a 3rd-Order property ($p^3.p^2.p^1$) with respect to Kefalonia.

Based on this notion, we define two kinds of access paths for $p_e$ w.r.t. a $PE_t$: *direct* if $p_e$ is accessed by a 1st-Order property, *combined* if $p_e$ is accessed by a nth-Order property (n>1).

Combined access paths are alternatives to access to property values. They allow to obtain approximate values, which are more or less good values obtained by composing properties. The knowledge engineer has to define all access paths (a maximum) for each $p_e$ involved in correspondences with properties in $\mathcal{O}_s$. In case of multiple paths of a given order, they must be ordered w.r.t. their relevance. Parts of combined access paths independent of $p^1$ can be reused.

In our work, we chose to work with DBpedia. We applied DBpedia Spotlight [13] on each document of the corpus to have an access to the DBpedia page corresponding to the described entity. Then, the model of correspondences and the specification of access paths are used as a support to write SPARQL queries in order to access DBpedia.

### *D. Deriving definitions of target concepts*

Discovering *target definitions* of $TC$ is a reasoning task based on the populated ontology, i.e., on all property assertions of individuals of $MC$. The knowledge engineer may not be able to express precise definitions of $TC$ but we assume that he is able to manually annotate a subset of documents describing instances as positive and negative examples of each class from $TC$. Manual annotations of entity descriptions can therefore be used by Machine Learning (ML) tools as positive and negative examples for each $TC$, in order to induce their definition. This will allow us to get an explicit formal definition for all $TC$. These definitions are then applied to get instances of $TC$.

We need ML tools capable of learning definitions of classes expressed in Description Logics from expert-provided examples. Specifically, definitions of $TC$ must be learned from (i) a populated OWL ontology including all the property assertions of all individuals of $MC$ and from (ii) positive and negative examples. Moreover, explicit specifications of relations (subsumption, object/datatype properties) between features as it is expressed in an OWL ontology have to be taken into account. For example, it could be able to learn what a $DWW$ is, given which destinations from the corpus are one (e.g., Dominican Republic) and which are not (e.g., Alaska).

We chose to use DL-Learner [14], an open source tool using inductive logic programming on Description Logics. The DL-Learner definitions are conjunctions and disjunctions of elements. An element can be a class (Destination) or an expression using object properties (hasActivity some Nightlife), numerical datatype properties (avgTemperatureC some double[>= 23.0]), or cardinality constraints (hasCulture min 3 Culture). Ranges are conjunctions and disjunctions of elements. For example, (Destination and (hasActivity some Watersport) and (hasWeather min 2 ((concernMonth some (hasSeason some MidWinter)) and (avgTemperatureC some double[>= 23.0]) and (precipitationMm some double[<= 70.0])))) is a definition for $DWW$ that can be learned by DL-Learner.

We developed a methodology from the conducted experiments. For each $TC$, 10 configurations are tested, each one with a different set of parameters, tuned using test experiments. For each test, we keep the highest ranked solution, which is the best one in terms of Accuracy and length. For each $TC$, we then choose the best definition from the 10 tests.

### E. *Ontology enrichment with individuals of target classes*

The discovered definitions are applied to retrieve all individuals that instantiate $TC$. This task has to be performed any time new entity descriptions are provided. We use FaCT++ [15], an efficient OWL-DL reasoner applicable to a large number of individuals unlike HermiT [16] and Pellet [17], according to our experiments. If there is no instances for a $TC$, some restrictions in the definition can be relaxed so that we obtain partial satisfactory proposals.

In conclusion to this section, it may be noted that adjustments simulating the Closed-World Assumption (CWA) were made for ensuring all tasks cooperate with each other. As the open world assumption is made for OWL ontologies, we disabled negation (NOT) and universal restrictions (ONLY) in the definition learning task. We adopted the Unique Name Assumption (UNA) in the ontology specifying that all individuals are different from each other. Otherwise, they are supposed to be possibly connected by owl:sameAs links and this can lead to problems when reasoning with definitions containing minimum cardinality restrictions generated by learning methods under a CWA. Furthermore, as inferences cannot be made on OWL with a definition having a maximum cardinality restriction, this type of restriction is ignored, i.e., we keep the highest ranked definition such that it does not contain a maximum cardinality restriction. The extraction tasks have also been adjusted. If one property assertion is not extracted, we presume that it is unlikely. For example, if a document about a destination does not mention beaches, it is assumed that there are no beaches, as documents are supposed to describe all the assets of destinations. Conversely, if properties are relevant for some types of individuals, values must be found for each of them. This explains why techniques presented in Section IV-C complete the missing property assertions in RDF data sources.

## V. Experimental evaluation

We compare Saupodoc with classification approaches, which are tools to discover concepts from texts even if they do not generate any definitions.

### A. *Materials*

We experiment our approach in two application domains chosen for their different characteristics.

*1) The holiday destinations field:* The corpus of holiday destinations is small (80 documents), which makes it possible to manually verify collected assertions. Each document has been automatically extracted from the Thomas Cook catalog (http://www.thomascook.com/) and describes a specific place (country, region, island or city). The documents are promotional, i.e., describe the qualities of destinations on a comprehensive basis and have hardly any negative expressions. The ontology includes one main class, Destination. Descriptive classes (161) characterize the nature of the environment (46 classes), the activities that can be done (102 classes), the kind of family that should go there, e.g., people with kids, couples, etc. (6 classes) and classes to define the weather like the seasons (7 classes). These descriptive classes contain individuals associated with terminological forms for facilitating their identification in texts. For example, the terms "archaeology, archaeological, acropolis, roman villa, excavation site, mosaic" are associated to the individual archaeology. 39 $TC$ are addressed.

*2) The film field:* The film corpus contains 10,000 documents, a significant number in order to check the applicability of the learning step with many individuals. It has been automatically built using DBpedia. Each document corresponds to a DBpedia page about a film. A document contains the DBpedia URI (no need to use DBpedia Spotlight to get the page) and its abstract describing the film (with hardly any negative expressions). The film ontology is basic (5 descriptive classes). It only contains the needed classes w.r.t. $TC$ of our experiments. 12 $TC$ corresponding to DBpedia categories (values of the property dcterms:subject) are addressed.

### B. *Evaluation of the* Saupodoc *approach*

*1) Experimental scenario:* Positive and negative examples for each $TC$ have to be given as input for all tested approaches. They are manually given by the knowledge engineer for destinations and automatically generated for films: a film $f$ is a positive example for a $TC$ corresponding to a category $c$ if <$f$ dcterms:subject $c$>, otherwise it is a negative example.

Saupodoc relies on an ontology but classifiers do not. The idea is then to considerate the domain terminology given by the knowledge engineer as a domain dictionary. Each document is represented as a vector (Vector Space Model). We use a bag-of-words method, where each element of the vector represents a word in the dictionary, which can be one or several keywords or keyphrases. Basically, if a document contains a word (lemmatization is performed), the value for its element is TF-IDF, otherwise 0. These vector representations are used as input of (i) a SVM classifier and (ii) a decision tree classifier. Both classifiers are tested with several parameters. We keep the best results.

For evaluation, documents are split into 2/3 for the training set and 1/3 for the test set. This means that learning is performed on 2/3 of data and that results are given on the rest of data. Several metrics are computed.

*2) Results:* First, we observe (see Table I) that the three approaches give satisfactory results in terms of Accuracy although slightly better results are obtained with ours. However, Accuracy is not the measure the best appropriate to our problem because each $TC$ has lots of negative examples and few positive ones: if a classifier predicts negative on all inputs (no instances of $TC$ found) then Accuracy is high (91.76% on average for film $TC$). True negatives and true positives are not of equal importance. Alternative metrics such as Precision, Recall and F-measure are needed to evaluate the prediction of positives, which is central in our problem. Table I shows the results with respect to those metrics. We can observe that our approach is the best in terms of Precision, Recall and F-measure on both domains.

$$Acc. = \frac{TP+TN}{TP+FP+TN+FN} \qquad F\text{-}measure = \frac{2\times Precision \times Recall}{Precision+Recall}$$

$$Precision = \frac{TP}{TP+FP} \qquad Recall = \frac{TP}{TP+FN}$$

TABLE I. Average results for $TC$

| Metric (%) | Accuracy | | | F-measure | | |
|---|---|---|---|---|---|---|
| Corpus | Us | SVM | Tree | Us | SVM | Tree |
| Destination (39 $TC$) | 95.89 | 84.52 | 86.23 | 72.23 | 54.14 | 63.22 |
| Film (12 $TC$) | 95.46 | 94.41 | 94.32 | 75.65 | 61.74 | 61.40 |

| Metric (%) | Precision | | | Recall | | |
|---|---|---|---|---|---|---|
| Corpus | Us | SVM | Tree | Us | SVM | Tree |
| Destination (39 $TC$) | 73.95 | 58.10 | 64.23 | 71.58 | 55.32 | 65.89 |
| Film (12 $TC$) | 76.27 | 69.90 | 67.72 | 77.76 | 57.59 | 58.99 |

These results evaluate the combined performance of the various tasks performed by SAUPODOC. The definition learning task allows a good classification but the tasks upstream of the learning process have an impact on the results too in the sense that they affect the quality of data used to learn definitions. In the following, we analyze the two extraction tasks on the destination domain. The corpus contains few documents. A manual validation can be conducted.

We analyze first the extraction of property assertions from texts. We notice 52 wrong property assertions (false positives) out of 2,375 (2.19% of noise). Precision reaches 97.81%. Recall is assumed to be close to 1. Indeed, if a property assertion is not mentioned in a text, then that property does not characterize the described individual since all main features are supposed to be given in textual descriptions. This way, the number of false negative assertions (missing assertions) should be extremely limited. This clearly shows that the quality of the extraction task from texts is good. Relevant assertions are introduced in the ontology with minimal noise.

With respect to the extraction task from the LOD, the proposed techniques dealing with multiple or multivaluated properties and missing values proved to be very useful. Only 29 from 80 destinations have weather data (tested on DBpedia 2014). Specification of access paths provided approximated values. For example, the weather for Boston has been given from the page Quincy_Massachusetts. Moreover, complex correspondences have been defined for all properties (26) having to be valued from DBpedia. Property expressions in DBpedia corresponding to properties in the ontology were quite complex, never elementary properties.

Finally, let us note that well-known classifiers do not result in explicit definitions. SVM classifiers create a model, which is not comprehensive for human. Decision tree classifiers are a bit more intelligible since trees can be seen as sets of rules. However, these rules deal with the TF-IDF number associated with a dictionary word, which is hard to interpret by humans. In SAUPODOC, definitions are comprehensive and could be refined if needed.

## VI. CONCLUSION AND FUTURE WORK

We proposed an ontology-based approach, SAUPODOC, to solve an issue of heterogeneity between customized concepts without a priori definitions and unstructured documents describing individuals in an incomplete way. The approach combines several tasks operating at different abstraction levels and exploits the LOD. We also proposed a model to specify complex correspondences between an ontology and LOD data sources and mechanisms to deal with incompleteness and multiple properties or values. Finally, experiments have been carried out. Results show the relevance of SAUPODOC and a better Precision, Recall and F-measure than well-known classifiers. Future work will address automatic generation of SPARQL construct queries to query the LOD based on the two models described in Section IV-C.

## ACKNOWLEDGEMENTS

This work has been funded by the PORASO project, in the setting of a collaboration with the Wepingo company.

## REFERENCES

[1] J. Völker, P. Hitzler, and P. Cimiano, “Acquisition of OWL DL Axioms from Lexical Resources,” in ESWC. Innsbruck, Austria: Springer, 2007, pp. 670–685.

[2] M. Chitsaz, “Enriching Ontologies through Data,” in Doctoral Consortium co-located with ISWC, Sydney, Australia. CEUR-WS.org, 2013, pp. 1–8.

[3] J. Lehmann and P. Hitzler, “Concept learning in description logics using refinement operators,” Machine Learning, vol. 78, no. 1-2, 2010, pp. 203–250.

[4] P. Cimiano, Ontology learning and population from text - algorithms, evaluation and applications. Springer, 2006.

[5] G. Petasis, R. Möller, and V. Karkaletsis, “BOEMIE: Reasoning-based Information Extraction,” in LPNMR. A Corunna, Spain: CEUR-WS.org, 2013, pp. 60–75.

[6] N. Yelagina and M. Panteleyev, “Deriving of Thematic Facts from Unstructured Texts and Background Knowledge,” in KESW, vol. 468. Springer, 2014, pp. 208–218.

[7] Y. Ma and F. Distel, “Learning Formal Definitions for Snomed CT from Text.” in Proc. of Artificial Intelligence in Medicine. Springer Berlin Heidelberg, 2013, pp. 73–77.

[8] ——, “Concept Adjustment for Description Logics,” in K-CAP. New York, NY, USA: ACM, 2013, pp. 65–72.

[9] P. Cimiano, J. Völker, and R. Studer, “Ontologies on Demand? - A Description of the State-of-the-Art, Applications, Challenges and Trends for Ontology Learning from Text,” Information, Wissenschaft und Praxis, vol. 57, no. 6-7, Oct. 2006, pp. 315–320.

[10] P. Cimiano, S. Handschuh, and S. Staab, “Towards the Self-annotating Web,” in WWW. New York, NY, USA: ACM, 2004, pp. 462–471.

[11] P. Cimiano and J. Völker, “Text2Onto - A Framework for Ontology Learning and Data-driven Change Discovery,” in NLDB, vol. 3513. Alicante, Spain: Springer, 2005, pp. 227–238.

[12] H. Cunningham et al., Text Processing with GATE. University of Sheffield Department of Computer Science, 2011.

[13] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer, “DBpedia Spotlight: Shedding Light on the Web of Documents,” in I-Semantics. NY, USA: ACM, 2011, pp. 1–8.

[14] J. Lehmann, “DL-Learner: Learning Concepts in Description Logics,” Journal of Machine Learning Research, vol. 10, 2009, pp. 2639–2642.

[15] D. Tsarkov and I. Horrocks, “FaCT++ Description Logic Reasoner: System Description,” in IJCAR. Berlin, Heidelberg: Springer, 2006, pp. 292–297.

[16] R. Shearer, B. Motik, and I. Horrocks, “HermiT: A Highly-Efficient OWL Reasoner.” in OWLED, vol. 432. CEUR-WS.org, 2008.

[17] E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz, “Pellet: A practical OWL-DL reasoner.” Journal of Web Semantics, vol. 5, no. 2, 2007, pp. 51–53.

# Analyzing the Retirement Satisfaction Predictors among Men and Women Using a Multi-Layer Feed Forward Neural Network and Decision Trees

Ehsan Ardjmand, Gary R. Weckman, Diana Schwerha
Department of Industrial and Systems Engineering
Ohio University
Athens, Ohio USA
Email: ardjmand@ohio.edu, weckmang@ohio.edu,
schwerha@ohio.edu

Andrew P. Snow
School of Information and Telecommunication Systems
Ohio University
Athens, Ohio USA
Email: snowa@ohio.edu

***Abstract*— In this article, we will analyze the effect of different retirement satisfaction predictors on each other and the retirement satisfaction level among men and women. The following factors will be used as predicators of retirement satisfaction: health; wealth; smoking and drinking habits; education; faith; income; impact of health on activities of daily living (ADL); frequency of activities; and the number of people in a household. A set of 858 retired men and 1179 retired women from a 2012 Health and Retirement Study database have been chosen and analyzed. A neural network was trained for each gender in order to predict retirement satisfaction; it also generated a decision tree that symbolizes the retirement satisfaction and its predictors. The results demonstrate that health, age, smoking habits, income, and wealth are the most significant predictors for both genders, while for men, education also plays an important role in retirement satisfaction.**

***Keywords- Retirement Satisfaction; Artificial Neural Networks; Multi-Layer Perceptron; Decision Tree***

## I. INTRODUCTION

As the population of retired people is growing, retirement satisfaction has become a significant issue in aging and retirement research. It is predicted that around 24 percent of the United States' work force in 2018 will be at least 55 years old [1]. In addition to positive changes in lifestyle, retirement—as a major alteration in life for the elderly—can be the source of many negative experiences, such as loneliness, anxiety, and sometimes even psychological disorders [2].

There is a large body of research on factors which may have an effect on retirement satisfaction—among which health and wealth, as the two most important predictors, have been shown to have a positive correlation with this kind of satisfaction [3-8]. A positive psychological condition is also shown to have a positive correlation with retirement satisfaction [6].

Sexuality is also another analyzed factor in literature. Although there are many studies focusing only on men or women in terms of retirement satisfaction, the studies show that there is no significant difference among men and women in this category [6, 9-15].

Voluntary retirement, engagement in social activities, higher educational level, and having a spousal partner also can have a positive effect on retirement satisfaction [8, 12, 15-21].

Although the retirement satisfaction factors have been analyzed extensively in literature, the inter-relational effect of these factors remains an unchallenged problem. For example, we know that wealth and health have a positive correlation with retirement satisfaction [5], but how will a high level of wealth and a low level of health affect retirement satisfaction simultaneously? Additionally, what level of each factor is the threshold at which retirement satisfaction may be altered?

In this paper, using the data of 858 retired men and 1179 retired women from the 2012 Health and Retirement Study database, we predict the retirement satisfaction level as a dependent variable and the health, wealth, smoking and drinking habits, education, faith, income, impact of health on instrumental and regular ADLs, frequency of activities, and number of people in a household as independent variables by using a multi-layer perceptron neural network. We then try to illustrate the effect of different levels of independent variables on retirement satisfaction simultaneously by using a decision tree for both men and women.

In Section 2, we explain the method and data we use for analysis. In Section 3, the results of analyzing retirement satisfaction as an outcome of predictor variables are presented for both men and women. In Section 4, the overall conclusion is stated.

## II. DATA AND METHODOLOGY

### A. Health and Retirement Study

The data for this research came from the 2012 Health and Retirement Study (HRS), which was launched in 1992. The total number of randomly considered retired people chosen from HRS for this study was 2037, which consisted of 858 men and 1179 women. Notice that only the respondents with no missing values in both dependent and independent variables were considered in this study.

The dependent variable is considered to be retirement satisfaction. If a person is reported to be retired in 2012 he/she is asked the G136 question, “All in all, would you say that your retirement has turned out to be very satisfying, moderately satisfying, or not at all satisfying?" The answer to this question is supposed to capture the retirement satisfaction level for retirees.

The independent variables in this research are the age (in months); years of education; belief in a higher power; self-report of health (based on a 5-point scale in which 1 shows excellent health and 5 shows very poor health); a binary variable which shows if the health limits the ability to work or not; level of difficulty in pursuing the ADLs (based on a 6-point scale in which 0 shows no difficulty and 5 shows someone is unable to perform ADL); mental health (based on a 9-point scale in which 0 is excellent and 8 is very poor); a set of binary variables that show if the person has blood pressure, diabetes, cancer, lung disease, heart problem and/or arthritis; frequency of vigorous, moderate, and light activity; a binary variable that shows if the person smokes or not; the number of alcoholic drinks consumed per week; wealth; income; and the number of people living in a household.

*B. Methodology*

In this research for modeling retirement satisfaction and other independent variables, we use a multi-layer feed forward neural network. For illustrating this relationship in a symbolic structure, we will use a decision tree technique proposed by Craven [22] and modified by Young [23].

*1) Artificial Neural Networks (ANN)*

ANNs are mathematical models that mimic the human brain. Besides being considered a “black-box” model, ANNs also have the limitation of requiring a large amount of training and cross-validation data, i.e., typically three times more training samples than network weights [24]. However, a systematic way of modelling complex non-linear patterns is proposed [25]. Since their resurgence in the 1980s, ANNs have been applied to a variety of problem domains such as speech recognition [26] and generation [27], symbolic learning [28], robotic design [29], medical diagnostics [30], game playing [31], healthcare systems [32], bankruptcy [33], credit cards [34], and estimation of functions as in forecasting [35-37]. Theoretically, it is possible to prove that a three-layered NN can estimate the value of a function with desirable accuracy [38, 39]. Since the relationship of retirement satisfaction and other independent variables is not necessarily linear and can be considered highly complex, feed forward neural networks can be a useful tool for predicting the value of retirement satisfaction.

There are many types of ANN topologies that have been comprehensively documented [40], and they range in their use and complexity. One of the most widely used neural networks (NN) is the feed forward neural network (FNN). For example, Figure 1 shows the general structure of a FNN. The network shown is fully connected, since each layer is connected via previous layers. The first hidden layer’s neurons are connected to the second hidden layer’s, and the second hidden layer’s neurons are connected with all of the output layer’s neurons.

There are two main paradigms of ANN training--supervised and unsupervised learning. The primary difference between the two learning schemes is that in supervised learning, known outputs, or--“targets”--are used to adjust the network’s weights. In unsupervised learning, there is not a known output, and the method functions as a clustering algorithm.

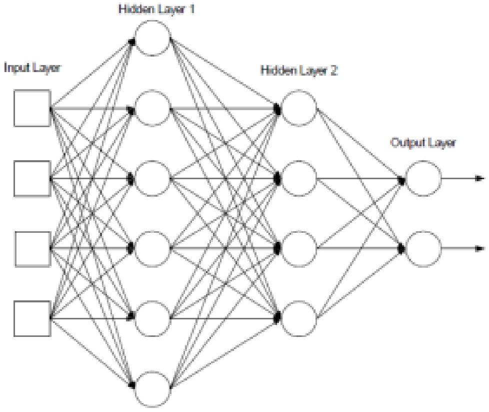


Figure 1. Feed Forward Neural Network [41].

*2) Decision trees and TREPAN*

One of the main drawbacks of neural networks is the lack of explanation capability [42]. In order to represent the knowledge about retirement satisfaction learned by a neural network, we use decision trees. Decision trees classify data through recursive partitioning of the dataset into mutually exclusive subsets [43], which best explain the variation in the dependent variable under observation [44]. A decision tree model consists of logical tests, which result in possible classifying consequences. Decision trees have been used to aid decision makers in many real-world problems. For example, Leech [45] applied a decision tree to a chemical nuclear power plant process involving continuous feedback systems. Another use led a manufacturing company to reduce inventory levels and improve processing efficiencies, which saved the company ten million dollars in operation expenses a year [46].

TREPAN is a novel rule-extraction algorithm [47] that utilizes the behavior of a trained ANN. Given a trained ANN, TREPAN extracts decision trees that provide a close approximation to the function represented by the network when there are issues of accurately calculating tree partitions, which are caused by limited sample sizes. TREPAN uses a concept of recursive partitioning similar to other decision tree algorithms; however, in contrast to the depth-first growth used by other decision tree algorithms, TREPAN expands using the “best first” principle. For conventional induction algorithms, the amount of training data decreases as a decision tree grows. Thus, there is less data at the bottom of the tree able to determine class labels accurately. In contrast, TREPAN uses an “oracle” to answer queries that determine decision tree splits better when sample instances are limited. One important aspect of this feature is the user-determined parameter called minimum sample. TREPAN ensures that splits are determined with a minimum number of sample instances. If the number of instances at a particular node, *m*, is less than the minimum sample allowed, TREPAN will make membership queries equal to the minimum sample from the ANN oracle in order to artificially create sample instances to meet the minimum sample requirement.

TREPAN uses an entropy-based criterion called “information gain” to determine the best position in which to partition the dataset. TREPAN uses M-of-N expressions as it splits upon the dataset. In this case, N rules are created. The algorithm also determines a value for M, which represents the minimum conditions that must be met, which in turn dictates the preceding node or final classification. This approach allows multiple features to be present in one node. To prevent testing of all the possible M-of-N combinations, TREPAN makes use of the heuristic "beam search" process. This process begins by selecting the best binary split at a given node based upon information gain. Additional splitting conditions are determined based on the initial rule's "complement" [48].

When sample instances are sparse, TREPAN interacts with an ANN oracle by means of membership queries. The goal of a membership query is to determine a new instance among a group of instances. To create appropriate sample instances, distributions of attribute values are created that conform to the decision tree constraints [49]. Once the ranges are determined, random pulls are made from the attributes' distribution in order for the oracle to accurately estimate the classification output label.

#### 3) Retirement Satisfaction Model

For this study, we train a feed forward neural network with two hidden layers. There are 15 processing units in the first layer and 10 processing units in the second hidden layer, as well as tangent hyperbolic and linear transfer functions for the hidden and output layers, respectively, that use back propagation algorithms in NeuroSolutions 6.20 software. The output of the network, i.e., retirement satisfaction – is a continuous number. In order to convert the output of the network into the categorical scale of retirement satisfaction, we divide the output into three categories of (-∞,1.66], (1.66,2.33], and (2.33,+∞], which are equivalent to not satisfying, moderately satisfying, and very satisfying. Notice that in the data we use the numbers 1, 2, and 3 to represent satisfying, moderately satisfying, and very satisfying, respectively.

## III. Results

Figure 2 and Figure 3 show the decision tree obtained for men and women regarding the relationships of the independent variables and retirement satisfaction. Notice that every rectangular shape in the decision tree shows a condition that, if met, the right branch should be followed. The left branch is for the case in which the condition is rejected. The oval shapes show the consecutive retirement satisfaction level in each branch.

As it is depicted in Figures 2 and 3, not all of the variables are involved in predicting retirement satisfaction. The reason is partially because of the low correlation of some independent variables and retirement satisfaction, as well as the overwhelming impact of these important variables on the latter that makes the other factors neutral. Another reason is the structure of the decision tree itself. By generating a decision tree, we are trying to extract the knowledge of the neural network, and the generated tree is formed in a way to represent the most possible knowledge in the form of rules according to the neural network, which can cause us to ignore some of the inputs.

### A. Comparison with Literature

All of the extracted rules in decision trees are consistent with the results in literature. Age has a positive correlation with retirement satisfaction [3]. This effect can be seen by following branches that point to older ages and comparing them to the other branches in Figures 2 and 3. High levels of mental and physical health correspond to higher retirement satisfaction [3, 4, 6-8]. Higher levels of wealth and income also correspond to higher retirement satisfaction [3, 5-7]. Years of education have a positive correlation with retirement satisfaction [20].

### B. New Findings

In addition to the result comparisons to previous literature, some new patterns can be deduced from the decision tree. Compared to women, the years spent in education for men is an important factor. In Figure 2, one of the parameters that affects the retirement satisfaction in men is education level. However, in Figure 3 the education level is not a condition in defining the retirement satisfaction, which shows that for women, it is not an important parameter.

Since for men the wealth appears in higher levels of the decision tree, it follows that, compared to women, wealth for men is a more important factor. Following the same logic, we can see that compared to men, mental health is a stronger predictor for women. In addition, for women with poor health, wealth is not a predictor at all. Despite this, for men with poor overall health, age cannot predict the retirement satisfaction.

Among all the health conditions analyzed, only diabetes plays a significant role in explaining retirement satisfaction. In both decision trees, i.e., men's and women's – having diabetes can cause lower retirement satisfaction, except where the income level is rather high. Although poor conditions of physical and mental health for both men and women can cause low retirement satisfaction, a high amount of wealth and income can ameliorate this situation.

## IV. Conclusion

In this paper, using the 2012 data of the Health and Retirement Study for 858 retired men and 1179 retired women, we trained a feed forward neural network to predict the retirement satisfaction, considering health, wealth, smoking and drinking habits, education, faith, income, impact of health on ADLs, frequency of activities, and the number of people in a household as independent variables. The knowledge of neural networks was represented in the form of a decision tree.

The results show a very high consistency with previous findings in literature. Additionally, some new knowledge regarding retirement satisfaction was also revealed in the form of rules in the decision tree. It was shown that, compared to men, years of education is more important to women in regards to retirement satisfaction. Under the condition of poor health, age is an important predictor of retirement satisfaction for women. Among all the health-related diseases, diabetes plays the most important role in terms of predicting retirement satisfaction. Additionally, a poor health condition can be negated by higher income or wealth.

To the best of our knowledge, the use of decision trees in retirement satisfaction is introduced for the very first time in this article. The results show that this technique can be a very powerful method for revealing hidden relationships between the various predictors of retirement satisfaction.

## References

[1] D. Schwerha, C. Ritter, S. Robinson, R. W. Griffeth, and D. Fried, "*Integrating ergonomic factors into the decision to retire,"* Human Resource Management Review, 2011. 21(3), pp. 220-227.

[2] F. J. Floyd, et al., "*Assessing retirement satisfaction and perceptions of retirement experiences,"* Psychology and Aging, 1992. 7(4), pp. 609.

[3] K. A. Bender, "*An analysis of well-being in retirement: The role of pensions, health, and 'voluntariness' of retirement,"* The Journal of Socio-Economics, 2012. 41(4), pp. 424-433.

[4] L. T. Dorfman, "*Health conditions and perceived quality of life in retirement,"* Health & Social Work, 1995. 20(3), pp. 192-199.

[5] C. W. Panis, "*Annuities and retirement well-being,"* Pension design and structure: New lessons from behavioral finance, 2004, pp. 259-74.

[6] C. A. Price and S. Balaswamy, "*Beyond health and wealth: Predictors of women's retirement satisfaction,"* The International Journal of Aging and Human Development, 2009. 68(3), pp. 195-214.

[7] M. Reis and D. P. Gold, "*Retirement, personality, and life satisfaction: A review and two models,"* Journal of applied Gerontology, 1993. 12(2), pp. 261-282.

[8] N. Schmitt, J. K. White, B. W. Coyle, and J. Rauschenberger, "*Retirement and life satisfaction,"* Academy of Management Journal, 1979. 22(2), pp. 282-291.

[9] T. M. Calasanti, "*Gender and life satisfaction in retirement: An assessment of the male model,"* The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 1996. 51(1), pp. S18-S29.

[10] E. Fouquereau, A. Fernandez, and E. Mullet, "*Evaluation of determinants of retirement satisfaction among workers and retired people,"* Social Behavior and Personality: an international journal, 2001. 29(8), pp. 777-785.

[11] K. Hanson and S. Wapner, "*Transition to retirement: Gender differences,"* The International Journal of Aging & Human Development, 1994.

[12] C. A. Price and E. Joo, "*Exploring the relationship between marital status and women's retirement satisfaction,"* The International Journal of Aging and Human Development, 2005. 61(1), pp. 37-55.

[13] K. Seccombe and G. R. Lee, "*Gender differences in retirement satisfaction and its antecedents,"* Research on Aging, 1986. 8(3), pp. 426-440.

[14] K. F. Slevin and C. Wingrove, "*Women in retirement: A review and critique of empirical research since 1976,"* Sociological Inquiry, 1995. 65(1), pp. 1-21.

[15] M. E. Szinovacz, "*Preferred retirement timing and retirement satisfaction in women,"* The International Journal of Aging & Human Development, 1987.

[16] E. Bonsang and T. J. Klein, "*Retirement and subjective well-being,"* Journal of Economic Behavior & Organization, 2012. 83(3), pp. 311-329.

[17] B. Butrica, "*Satisfaction and engagement in retirement,"* 2006.

[18] L. T. Dorfman and M. M. Moffett, "*Retirement satisfaction in married and widowed rural women,"* The Gerontologist, 1987. 27(2), pp. 215-221.

[19] E. Fouquereau, A. Fernandez, A. M. Fonseca, M. C. Paul, and V. Uotinen, "*Perceptions of and satisfaction with retirement: A comparison of six European Union countries,"* Psychology and Aging, 2005. 20(3), pp. 524.

[20] Y. Kremer, "*Predictors of retirement satisfaction: a path model,"* International journal of aging & human development, 1983. 20(2), pp. 113-121.

[21] C. Kupperbusch, R. W. Levenson, and R. Ebling, "*Predicting husbands' and wives' retirement satisfaction from the emotional qualities of marital interaction,"* Journal of Social and Personal Relationships, 2003. 20(3), pp. 335-354.

[22] M. W. Craven, "*Extracting comprehensible models from trained neural networks*." 1996, UNIVERSITY OF WISCONSIN–MADISON.

[23] W. A. Young, et al., "*An investigation of TREPAN utilising a continuous oracle model,"* International Journal of Data Analysis Techniques and Strategies, 2011. 3(4), pp. 325-352.

[24] M. M. Nelson and W. T. Illingworth, "*A practical guide to neural nets."* Vol. 1. 1991, Addison-Wesley Reading, MA.

[25] K.-L. Du, "*Clustering: A neural network approach,"* Neural Networks, 2010. 23(1), pp. 89-107.

[26] A. Waibel, "*Modular construction of time-delay neural networks for speech recognition,"* Neural computation, 1989. 1(1), pp. 39-46.

[27] T. J. Sejnowski and C. R. Rosenberg, "*NETtalk: A parallel network that learns to read aloud."* 1988, MIT Press.

[28] J. W. Shavlik, R. J. Mooney, and G. G. Towell, "*Symbolic and neural learning algorithms: An experimental comparison,"* Machine learning, 1991. 6(2), pp. 111-143.

[29] G.-T. Hsu and R. Simmons. "*Learning footfall evaluation for a walking robot,"* in *Proceedings of the Eigth International Workshop on Machine Learning*, 2014, pp. 303-307.

[30] M. Jabri, et al. "*ANN based classification for heart defibrillators,"* in *Advances in neural information processing systems*, 1992, pp. 637-644.

[31] G. Tesauro, "*Practical issues in temporal difference learning."* 1992, Springer.

[32] G. G. Towell and J. W. Shavlik, "*Extracting refined rules from knowledge-based neural networks,"* Machine learning, 1993. 13(1), pp. 71-101.

[33] D. K. Chandra, V. Ravi, and P. Ravisankar, "*Support vector machine and wavelet neural network hybrid: application to bankruptcy prediction in banks,"* International Journal of Data Mining, Modelling and Management, 2010. 2(1), pp. 1-21.

[34] N. Naveen, V. Ravi, and D. A. Kumar, "*Application of fuzzyARTMAP for churn prediction in bank credit cards,"* International Journal of Information and Decision Sciences, 2009. 1(4), pp. 428-444.

[35] R. S. Gutierrez, A. O. Solis, and S. Mukhopadhyay, "*Lumpy demand forecasting using neural networks,"* International Journal of Production Economics, 2008. 111(2), pp. 409-420.

[36] R. Kuo, "*A sales forecasting system based on fuzzy neural network with initial weights generated by genetic algorithm,"* European Journal of Operational Research, 2001. 129(3), pp. 496-517.

[37] G. Zhang, B. E. Patuwo, and M. Y. Hu, "*Forecasting with artificial neural networks:: The state of the art,"* International journal of forecasting, 1998. 14(1), pp. 35-62.

[38] K.-I. Funahashi, "*On the approximate realization of continuous mappings by neural networks,"* Neural networks, 1989. 2(3), pp. 183-192.

[39] K. Hornik, M. Stinchcombe, and H. White, "*Multilayer feedforward networks are universal approximators,"* Neural networks, 1989. 2(5), pp. 359-366.

[40] M. Kubat, "*Neural networks: a comprehensive foundation by Simon Haykin, Macmillan, 1994, ISBN 0-02-352781-7*." 1999, Cambridge Univ Press.

[41] G. R. Weckman, W. Young, S. Hernández, M. Rangwala, and V. Ghai, "*Extracting Knowledge from Carbon Dioxide Corrosion Inhibition with Artificial Neural Networks,"* International Journal of Industrial Engineering: Theory, Applications and Practice, 2010. 17(1).

[42] B. Baesens, R. Setiono, C. Mues, and J. Vanthienen, "*Using neural network rule extraction and decision tables for credit-risk evaluation,"* Management science, 2003. 49(3), pp. 312-329.

[43] G. Liepins, R. Goeltz, and R. Rush, "*Machine learning techniques for natural-resource data-analysis,"* Ai Applications in Natural Resource Management, 1990. 4(3), pp. 9-18.

[44] D. Biggs, B. De Ville, and E. Suen, "*A method of choosing multiway partitions for classification and decision trees,"* Journal of Applied Statistics, 1991. 18(1), pp. 49-62.

[45] W. Leech, "*A rule-based process control method with feedback,"* ISA transactions, 1986. 26(2), pp. 73-78.

[46] P. Langley and H. A. Simon, "*Applications of machine learning and rule induction,"* Communications of the ACM, 1995. 38(11), pp. 54-64.

[47] M. Craven and J. W. Shavlik. "*Using Sampling and Queries to Extract Rules from Trained Neural Networks,"* in *ICML*, 1994, pp. 37-45, Citeseer.

[48] D. Martens, J. Huysmans, R. Setiono, J. Vanthienen, and B. Baesens, "*Rule extraction from support vector machines: an overview of issues and application in credit scoring*," in *Rule extraction from support vector machines*. 2008, Springer, pp. 33-63.

[49] K. Krawiec, R. Słowiński, and I. Szczešniak. "*Pedagogical Method for Extraction of Symbolic Knowledge,"* in *Rough Sets and Current Trends in Computing*, 1998, pp. 436-443, Springer.

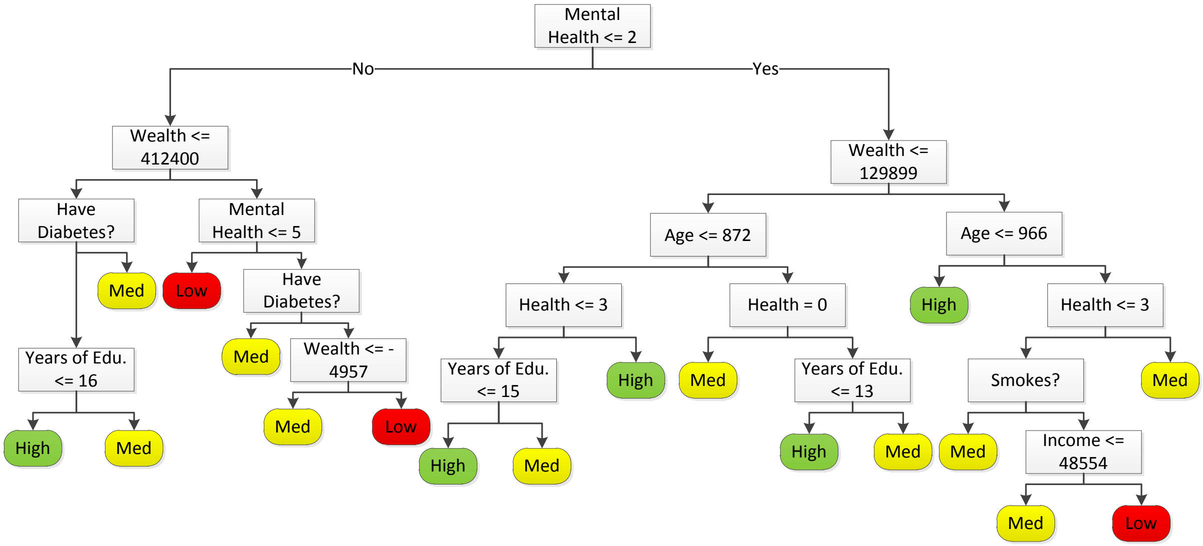


Figure 2. Decision Tree of Retirement Satisfaction for Men.

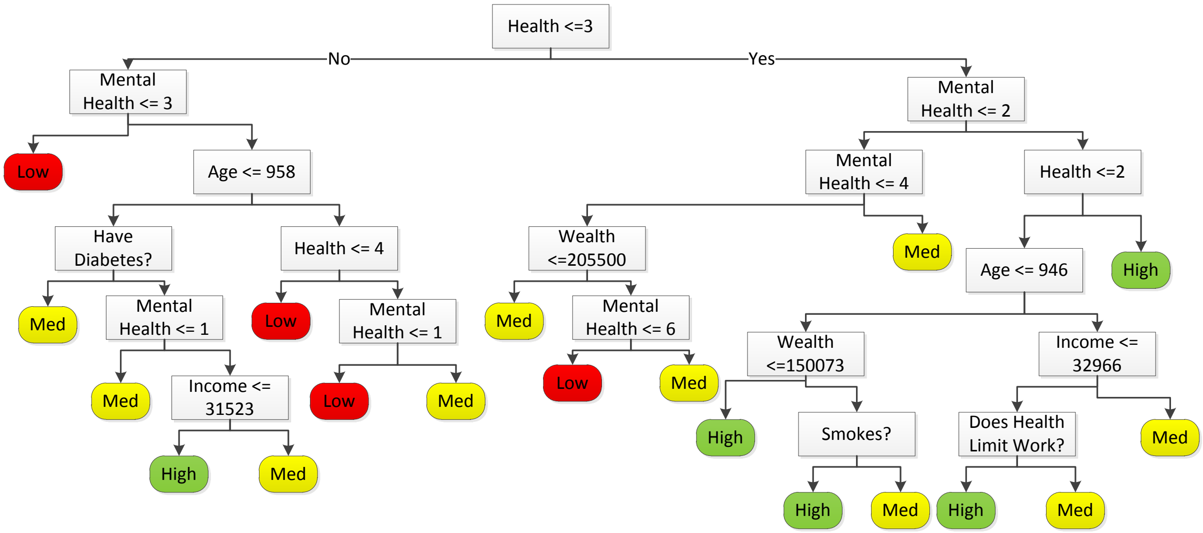


Figure 3. Decision Tree of Retirement Satisfaction for Women.

# Group Method of Data Handling: How does it measure up?

Poorani Selvaraj, Gary R. Weckman
Department of Industrial and Systems Engineering
Ohio University
Athens, Ohio USA
Email: ps365012@ohio.edu, weckmang@ohio.edu

Andrew P. Snow
School of Information & Telecommunication Systems
Ohio University
Athens, Ohio USA
Email: snowa@ohio.edu

*Abstract*— **Prediction is the method of determining future values based on the patterns deduced from a data set. This research compares various data mining techniques—namely, multiple regression analysis in statistics, Artificial Neural Networks (ANN), and the Group Method of Data Handling (GMDH), including both with and without feature selection. Currently, the literature suggests that GMDH, an inductive learning algorithm, is an excellent tool for prediction. GMDH builds gradually more complex models that are evaluated via a set of multi-input, single-output data pairs. ANNs are inspired by the complex learning that happens in the closely interconnected sets of neurons in the human brain and are also considered an excellent tool for prediction. This article is the beginning of a more detailed research project to investigate how well GMDH performs in comparison to other data mining tools.**

***Keywords- Group Method of Data Handling; Artificial Neural Networks; Prediction; Statistics; Feature Selection.***

## I. Introduction

In this day and age, a large amount of profitable information is being processed from the myriad of data that is being collected. Why is this information significant, who uses this information, and for what purposes is it being used? To be able to answer these questions, we will begin asking the very basic question: What is data? Data is information of any nature that, when quantized, can be meaningfully disseminated into useful knowledge [1].

There has been no discrimination in regards to the type of industries from which databases emerge. Disparate industries have understood the need to exploit the knowledge that can be extracted by scouring through these large repositories of data. Knowledge is a term that is commonly associated with data [1]. For example, the itemization of a grocery bill, along with its corresponding loyalty card number, is generally considered data. This data may be used by data scientists to estimate the number of people in the household, the age group, and so on. This is known as knowledge, which is extracted based upon data obtained from the bill. However, the biggest challenge that we face is disentangling the useful knowledge that is desegregated from all the noise, but that is also collected as part of the data in a repository[1][2]. Data mining helps in addressing this data overload problem that we face at a time when the world is progressing towards an era of digital information.

Data mining is the art of extracting understandable patterns that may be concealed within a vast repository of data and identifying potentially useful information that can be used to our advantage [3]. The choice of the proper data mining technique among those that are usually used depends on the kind of information we wish to extract from the data. Depending upon the type of knowledge we wish to gain, data mining usually involves six common classes of applications.

The process of detecting interesting relationships between attributes is known as association. This type of learning commonly explores large spaces of potential patterns and chooses those that may be of interest to the user, which are generally specified using a constraint [4]. This application of data mining is most commonly used in the business world, where they base most of their micro- and macro-business decisions off of these patterns

The use of a model to fit data into pre-categorized discrete classes is known as classification [5]. According to Zhang and Zhou, classification is the process of identifying common features that describe and distinguish common classes [6]. E.W.T. Ngai et al. suggest that the most commonly used techniques for the classification of data include neural networks, the naïve Bayes technique, decision trees, and support vector machines [5].

According to E.W.T. Ngai et al, dividing objects that are similar to each other in order to form groups that are conceptually meaningful—and, at the same time, very dissimilar to those from the other group--is called clustering [5]. Zhang and Zhou explain this as maximizing “intra-class” similarity and minimizing “inter-class” similarity [6]. The clusters are usually mutually exclusive and exhaustive. For a more complex and in-depth representation of the data, the model may be chosen to represent a hierarchical or overlapping category [7].

For estimation, we find an approximate value of a target or dependent variable using a set of predictor or independent variables [8]. Regression analysis is used for this purpose. Regression analysis is the generation of an equation that best represents the relationship between a continuous dependent output variable and one or more independent input variables [5]. It is mostly a statistics-based methodology that is used for estimation and forecasting. Based on the number of independent or predictor variables, a simple linear regression

ISBN: 978-1-61208-457-2

or a multiple linear regression may be performed. A well-fit model is one in which a strong correlation exists between the two variables.

Prediction, as the name suggests, is the method of determining future values based upon the patterns deduced from the data set. It is very similar to classification and estimation--however, a very fine line exists between them. According to E.W.T. Ngai et al, the most common techniques used for prediction analysis are neural networks and logistic model predictions [4].

In Section 2 and 3, we explain the techniques we use for analysis. In Section 4 and 5, we discuss the methodology and the database used. In Section 6, we discuss the results and the overall conclusion is stated.

## II. Artificial Neural Networks

ANNs are inspired by the complex learning that happens in the closely interconnected sets of neurons in the human brain [8]. An analogy between a neural network and a human brain can be drawn in the following manner [9], as illustrated in Figure 1.

An ANN always acquires its knowledge through learning, similar to a human brain, which is constantly learning through experiences. The ANN's knowledge is stored in its neurons, using weights associated in its inter-neuron connections, which are similar to the synaptic weights that we have in the neurons in a human brain.

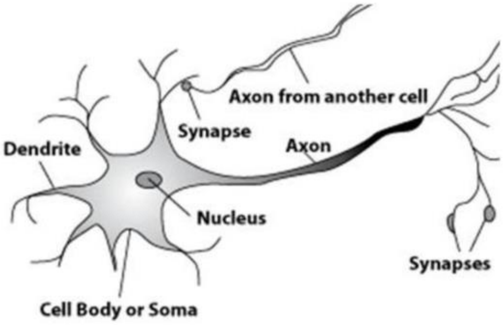


Figure 1. Biological neuron [9].

A generic structure of a neural network model can be explained as a mathematical equivalent, which is shown in Figure 2.

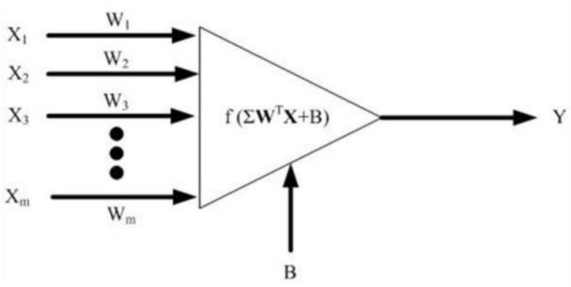


Figure 2. Arificial Perceptron [9]

Figure 2 displays a mathematical approximation of the biological model in Figure 1; for the mathematical model, a linear combination of weights and input values are passed through activation functions that process the data. For example, the perceptron in Figure 2 has independent inputs (*X1*, *X2*, *X3*, and *Xm*), connection weights (*W1*, *W2*, *W3*, and Wm), a bias (*B*), and a dependent variable (*Y*). During this operation, the perceptron computes a weighted sum as each input and bias passes through the neuron. Weights represent the strength of the association between independent and dependent variables, and can be positive, negative, or zero; the weighted sum is then processed by the neuron's activation function and sent through the rest of the network.

The mathematical expression of the perceptron neuron shown above is:

$$y_k = \varphi(\sum w_{km} * x_m + b_k) \tag{1}$$

Where $w_{km}$ =connection weight of source neuron *m* to target neuron *k;* $b_k = \text{bias}$ ; $y_k = \text{output;}$ $\varphi = \text{transfer function}$. One of the major advantages of ANN is their robust nature that allows them to represent and learn both linearly- and non-linearly-related data with ease [9].

A multilayer perceptron (MLP) is a modification of the simple linear perceptron, and can easily model highly nonlinear data of input vectors to accurately represent the new unseen data [11][12]. A simple MLP structure is illustrated in Figure 8 (end of article), and demonstrates how the perceptron fits into an ANN.

The architecture of an MLP can usually vary, but in general, it has input neurons which provide the input parameters to the network. There may be one or more hidden layers, which are made up of nodes connected to every other node in the prior and subsequent layers. The output signals from these nodes are the sum of the input signals to the node, which may be modified by an activation function or transfer function, which is generally non-linear. We make use of nonlinear functions to produce the outputs, which are then scaled using the weights associated with the nodes in order to be fed forward as the input to every node in the next layer of the network. This method of information processing, which is directed in the forward path of a network, makes the MLP a feed forward network [11].

By constant training, the MLP network has the ability to learn and to accurately model the input-output relationship. A set of training data with input vectors and target output vectors is used for the purposes of training. While training, the model constantly adjusts the weights of the nodes until the magnitude of the error of the MLP network is minimized. This is performed by constantly monitoring the error that arises as a difference between the predicted and the actual output, and adjusting the weights accordingly [11]. Thus, the MLP uses a supervised learning technique. The most commonly used algorithm for this purpose is the back-propagation algorithm. This training of the MLP is stopped when the performance of the network can accurately predict the target variable, which is compared to the testing data set.

## III. Group Method of Data Handling

Ivakhnenko in 1966 introduced the concept of GMDH as an inductive learning algorithm [13][14]. According to Ravisankar et al, GMDH builds gradually more complex models that are evaluated on a set of multi-input, single-output data pairs [15][16][17]. Dipti suggests that the complexities involved in other neural networks—such as determining the most important input variables, the number of hidden layers, and the neurons--are all circumvented by GMDH [14][18]. The need for prior knowledge to build models is eliminated by GMDH, and hence, it is a self-organizing feed forward neural network.

The neurons use competitive learning, as opposed to back propagation error correction. The competitive learning is based upon the way that the neurons compete with each other in order to respond to the input neurons. The overall methodology includes the following steps:

A data set with *n* observations for regression analysis is collected, with *m* independent variables $x_i$ and a dependent variable $y_j$; $i$=1, 2, 3….$m$; $j$=1, 2, 3….$n$

Step 1: The data set is divided into training and checking sets.

Step 2: A regression equation for each pair of independent variables is computed as follows :

$$y = A + Bx_i + Cx_j + Dx_i^2 + Ex_j^2 + Fx_{ij} \quad (2)$$

leaving us with $\binom{m}{2} = m(m-1)/2$ sets of regression polynomials, each made up of pairs of independent variables. We now have higher order variables predicting the output, as opposed to the original *m* variables $x_1$, $x_2$, $x_3$....$x_m$.

Step 3: Each of these regression surfaces will then be evaluated at all *n* data points. For example, a regression equation of the first two independent variables $x_1$ and $x_2$ is generated and then evaluated against all *n* data points as ($x_{11}$, $x_{12}$), ($x_{21}$, $x_{22}$), ($x_{31}$, $x_{32}$)…….($x_{n1}$, $x_{n2}$). This is now stored as a new variable $Z_1$. The remaining variables are computed in a similar manner. It is common knowledge that these new variables predict the output better than the original independent variables $x_1$, $x_2$....$x_m$.

Step 4: We now choose survivors, or those *Z* variables that best represent the output variable *y* by evaluating it against the checking set. The survivors are calculated by estimating the regularity criterion, which is usually the mean squared error ($r_{min}$), and arranging the values in increasing order of the regularity criterion for each *Z* variable. Based upon a pre-determined value *R*, those values of *Z* whose mean squared error is less than that of *R* are chosen as the survivors to replace the corresponding values of *x*.

The whole process is repeated, and we now have a regression polynomial of order four. In this way, the model builds gradually, complicating polynomial models of increasing order until the $r_{min}$ value of one model is no longer less than the previous model. This now implies that the $r_{min}$ value has reached its minimum and we can stop the process.

Feature selection is the process of using a smaller set of features, predictors, or independent variables to describe a sample in the measurement space. According to Guyon and Elisseeff, there are a number of potential benefits by this method of feature selection [19]:

- Helps in the visualization and better understanding of the data
- Feature selection enables reduced storage requirements
- Helps reduce the time to train the network
- Reduces the dimensionality of the network and improves the prediction performance.

For this research, GMDH is used to select the most significant features, depending upon their ability to have the best accuracy in their test data set. Ten-fold cross validation is performed, and the features which have the most frequency of occurrence in all of the 10 folds were selected based upon the percentages that were obtained from the GMDH Shell software.

## IV. Methodology

The flow chart below gives an outline of the methodology that has been followed for analyzing data for the purposes of comparing various data mining techniques.Namely, this includes multiple regression analyses in statistics, MLP, and GMDH, including those both with and without feature selection. See Figure 9.

To compute various statistical data sets usually for six sigma initiatives--Minitab is used. It is a very versatile and effective tool [20]. It provides tools and options for both basic and advanced data analysis.

To compute neural network models, NeuroSolutions [21], an easy-to-use neural network software package for Windows, is used. It provides an easy-to-use Excel interface and a user-friendly intuitive wizard with an icon-based network design interface, which are used to implement advanced artificial intelligence and learning algorithms [9].

GMDH Shell is used for the purpose of accurately forecasting time series, build classification, and regression models. It is also neural network-based software that allows for a full spectrum of parametric customization. However, the differentiating factor is that it is very fast, since it implements advanced parallel processing and has highly optimized the core algorithms. It can be used for any task from data sciences and financial analysis to inventory forecasting, demand forecasting, load forecasting, demand and sales forecasting, and stock market prediction [22][23].

In this broad range of knowledge discovery applications, the main idea is to train a subset of the population with known labels, and then make the predictions for a test subset with unknown labels [24]. When the learning algorithm is trained using only the training set, the algorithm looks for patterns in the training set which are depictive of the correlations that exist between the features and output of the data set. However, it is important to note that these patterns may be specific to only the training data set, i.e., they are valid only in the training set, and are not actually true for the

ISBN: 978-1-61208-457-2

general population of the database [24]. This will cause the algorithm to have a higher accuracy rate in the training set. It is not uncommon for the learning algorithm to become one hundred percent accurate with issues of over fitting. The tailoring of the algorithm to match the patterns that may be unique to only the training set--which might be due to randomness in how the training data was selected from the population--is called over fitting. In order to avoid these issues, a validation subset which is mutually exclusive of the training set is used to fine-tune the model.

- Training set: To fit the parameters, i.e., weights
- Validation set: To tune the parameters, i.e., architecture
- Test set: To assess the performance. i.e., generalization and predictive power

For the purpose of our research, we have divided 60 percent of the data set as a training set, 20 percent as cross validation, and 20 percent as a test set. The data set was initially randomized to avoid any discrepancies that may have occurred while experimenting.

## V. CASE STUDY

This case study is based on data collected by the National Oceanic and Atmospheric Administration (NOAA) study addressing the impacts of mussel recruitment on the bay's water quality via provided hydrological data (April through November, 1991–1996) [25]. The database includes the following variables: temperature (TEMP), Secchi depth (SECCHI), light attenuation (Kd), total suspended solids (TSS), TP, soluble reactive phosphorus (PO4 )3-P), nitratenitrogen (NO3 )-N), ammonia-nitrogen (NH4 +-N), silica (SiO2), particulate silica (PSiO2), particulate organic carbon (POC), chloride (CL), total photosynthetic radiation (TotPAR), visibility (VISIB), ambient temperature (TEMPAmb), and wind speed (WNDSpd). The dataset consists of 251 records [25].

Saginaw Bay (See Figure 3) is part of Lake Huron. Prior to the 1980s, excessive nutrient loading altered the water quality and resulted in expansive cyanobacterial blooms during the summer months. Bloom intensity and frequency decreased in the mid-1980s following the initiation of nutrient-abatement programs. Invasion of mussels occurred during the early 1990s, with larvae and adult mussels first being observed in 1991. During 1994–1996, their growth stabilized and populations became established (Nalepa et al. 1995). Coincident with mussel occurrence, blooms of toxic Microcystis reappeared during late summer months, and have remained annually recurrent throughout the bay [25].

Saginaw Bay is divided into two regions. Water quality in the shallow ''inner bay'' largely is influenced by nutrient-laden inflows of the Saginaw River. The river drains from agricultural, industrial, and urban areas. The mean circulation is weak, and water exchanges between the inner and outer bays occur along the northern shorelines.

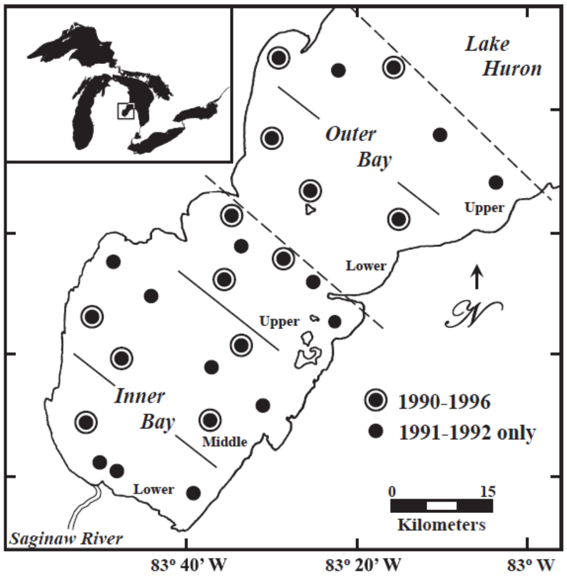


Figure 3. Location of sampling stations within Saginaw Bay [25].

## VI. RESULTS AND CONCLUSIONS

Many models were constructed, trained, and tested, as described earlier. The performance metric used is $R^2$--which is commonly called the coefficient of determination or the coefficient of multiple determination for multiple regression. The results of this analysis are summarized in Table 1:

TABLE 1: COMPARISON OF DATA MINING TECHNIQUES

| | Train | Cross Validation | Test |
|---|---|---|---|
| **Statistics** | 0.843 | | 0.685 |
| **GMDH** | 0.929 | 0.912 | 0.890 |
| **MLP** | 0.945 | 0.951 | **0.927** |
| **GMDH - FS** | 0.924 | 0.912 | 0.850 |
| **MLP - FS** | 0.977 | 0.949 | **0.922** |

As noted in the table, the MLP outperformed both the GMDH and basic statistics, whereas GMDH outperformed just statistics. In this case, the feature selection (FS) did not seem to have a benefit in MLP or GMDH, except in terms of reducing the number of attributes in the model. Figures 4 through 7 illustrate how well the model actually predicted the output values versus actual values. The perfect model would fit the 45-degree diagonal line. In reviewing the figure, it can be seen that the MLP model has the least variation from the diagonal line.

This research only looked at one database, so the conclusions are very limited in scope. Additional research is being performed, and the results of a more extensive study will be published in the near future.

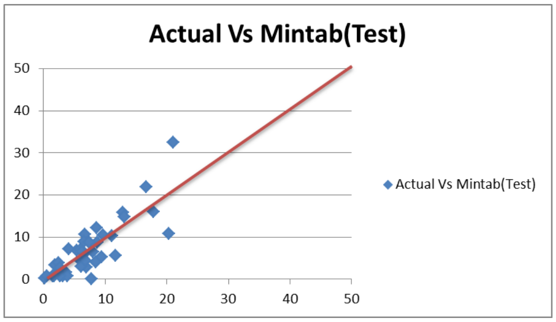


Figure 4. Comparison of Actual versus Minitab.

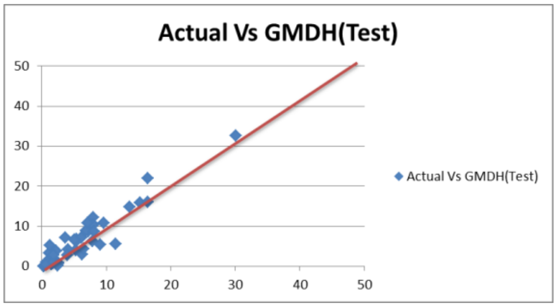


Figure 5. Comparison of Actual versus GMDH.

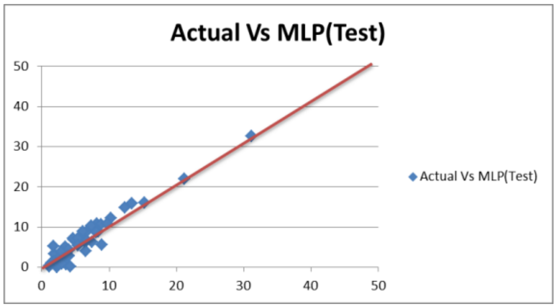


Figure 6. Comparison of Actual versus MLP.

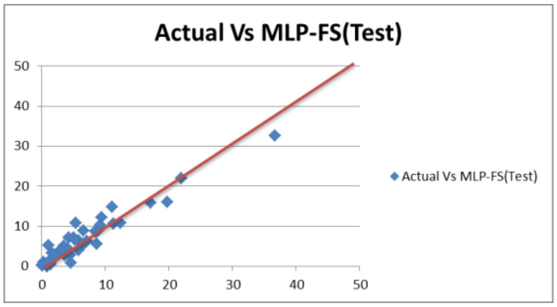


Figure 7. Comparison of Actual versus MLP with feature selection.

## REFERENCES

[1] G. Florin, Data mining: concepts, models and techniques. Berlin, Heilelberg: Springer-Verlag, 2011.

[2] L. A. Kurgan and P. Musilek, “A survey of Knowledge Discovery and Data Mining process models,” Knowl. Eng. Rev., vol. 21, no. 01, Mar. 2006, pp.1-24.

[3] U. M. Fayyad, P. Stolorz, “Data mining and KDD:Promise and challenges,” Future Gener. Comput. Syst., vol. 13, no. 2-3, Nov. 2007, pp. 99-115.

[4] G. I. Webb, “Discovering Significant Patterns,” Mach. Learn., vol. 68, no. 1, May 2007, pp. 1–33.

[5] E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, “The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature,” Decis. Support Syst., vol. 50, no. 3, Feb. 2011, pp. 559–569.

[6] D. Zhang and L. Zhou, “Discovering Golden Nuggets: Data Mining in Financial Application,” IEEE Trans. Syst. Man Cybern. Part C Appl. Rev., vol. 34, no. 4, Nov. 2004, pp. 513–522.

[7] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases,” AI Mag., vol. 17, no. 3, 1996, pp. 37-54.

[8] D. T. Larose and C. D. Larose, Wiley Series on Methods and Applications in Data Mining: Data Mining and Predictive Analysis, Second. John Wiley & Sons, Inc, 2015.

[9] W. A. Young II, W. S. Holland, and G. R. Weckman, "Determining Hall of Fame Status for Major League Baseball Using an Artificial Neural Network," Journal of Quantitative Analysis in Sports, vol. 4 : iss. 4, no. 4 Oct. 2008, pp. 1131-1135.

[10] D. J. Fonseca, D. O. Navaresse, and G. P. Moynihan, “Simulation metamodeling through artificial neural networks,” Eng. Appl. Artif. Intell., vol. 16, no. 3, Apr. 2003, pp. 177–183.

[11] M. W. Gardner and S. R. Dorling, “Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences,” Atmos. Environ., vol. 32, no. 14, Aug. 1998, pp. 2627–2636.

[12] T. Isokawa, H. Nishimura, and N. Matsui, “Quaternionic Multilayer Perceptron with Local Analyticity,” Information, vol. 3, no. 4, Nov. 2012, pp. 756–770.

[13] S. J. Farlow, Self-Organizing Methods in Modeling: GMDH Type Algorithms. CRC Press, 1984.

[14] D. Srinivasan, “Energy demand prediction using GMDH networks,” Neurocomputing, vol. 72, no. 1–3, Dec. 2008, pp. 625–629.

[15] P. Ravisankar and V. Ravi, “Financial distress prediction in banks using Group Method of Data Handling neural network, counter propagation neural network and fuzzy ARTMAP,” Knowl.-Based Syst., vol. 23, no. 8, Dec. 2010, pp. 823–831.

[16] P. Ravisankar, V. Ravi, G. Raghava Rao, and I. Bose, “Detection of financial statement fraud and feature selection using data mining techniques,” Decis. Support Syst., vol. 50, no. 2, Jan. 2011, pp. 491–500.

[17] M. Mottaghitalab, A. Faridi, H. Darmani-Kuhi, J. France, and H. Ahmadi, “Predicting caloric and feed efficiency in turkeys using the group method of data handling-type neural networks,” Poult. Sci., vol. 89, no. 6, Jun. 2010, pp. 1325–1331.

[18] G. C. Onwubolu, “Design of hybrid differential evolution and group method of data handling networks for modeling and prediction,” Inf. Sci., vol. 178, no. 18, Sep. 2008, pp. 3616–3634.

[19] I. Guyon and A. Elisseeff, “Empirical Inference for Machine Learning and Perception Department”, The Journal of Machine Learning Research, vol 3, Mar. 2003, pp. 1157-1182.

[20] “Review of Minitab 15 Software Program for Use in Six Sigma,” Brighthub Project Management. [Online]. Available: http://www.brighthubpm.com/software-

reviews-tips/33580-review-of-minitab-fifteen-for-six-sigma/. [Accessed: 15-Oct-2015].

[21] "Neurosolutions7". [Online]. Available: http://www.neurosolutions.com/neurosolutions/help/. [Accessed: 29-Sep-2015].

[22] "GMDH Shell | Binary Today." [Online]. Available: http://www.binarytoday.com/gmdh-shell/. [Accessed: 21-Oct-2015].

[23] "Data Mining Solution for Business", GMDH Shell Forecasting and data Mining Software." [Online]. Available: https://www.gmdhshell.com/data-mining-software. [Accessed: 18-Sep-2015].

[24] C. Elkan, "Evaluating classifiers," Univ. San Diego Calif. Retrieved 01-11-2012 Httpcseweb Ucsd Edu~ Elkan B, vol. 250, 2012.

[25] D. F. Millie et al, "An 'enviro-informatic' assessment of Saginaw Bay (Lake Huron USA) phytoplankton: characterization and modeling of Microcystis (Cyanophyta)," Journal of Phycology, vol. 47, no. 04, Aug. 2011, pp. 714-730.

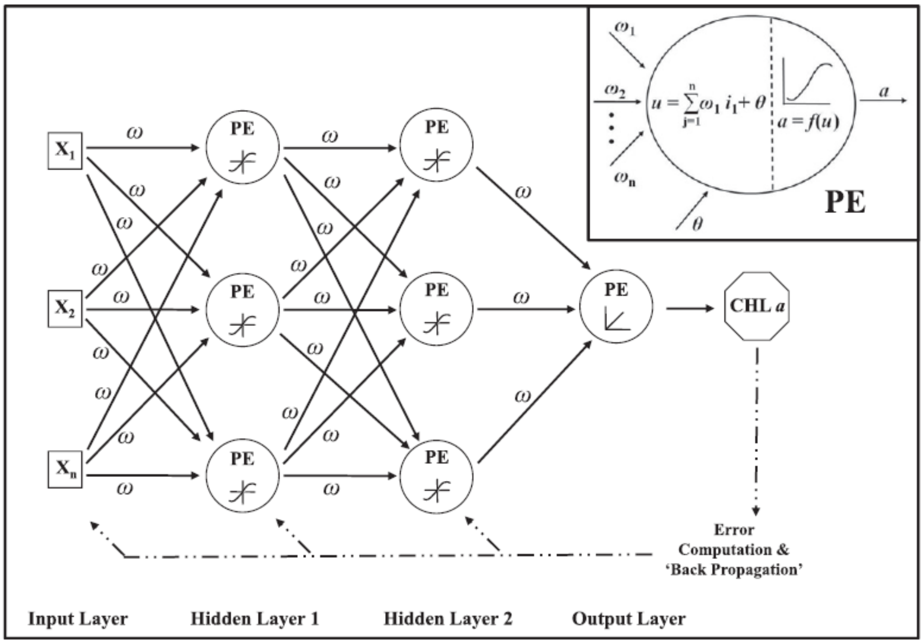


Figure 8. Artificial Neural Network [25].

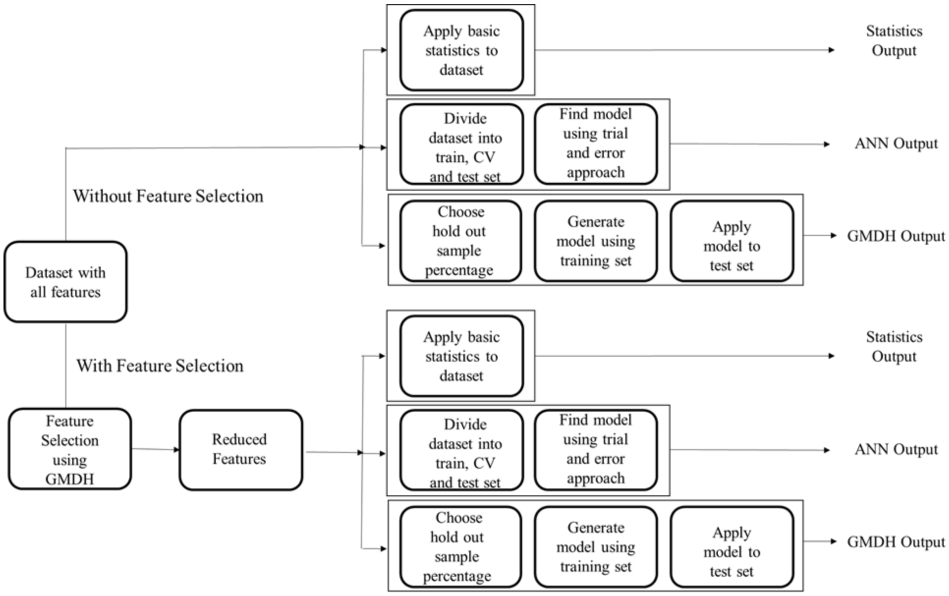

Dataset with all features
Without Feature Selection
Apply basic statistics to dataset
Statistics Output
Divide dataset into train, CV and test set
Find model using trial and error approach
ANN Output
Choose hold out sample percentage
Generate model using training set
Apply model to test set
GMDH Output
With Feature Selection
Feature Selection using GMDH
Reduced Features
Apply basic statistics to dataset
Statistics Output
Divide dataset into train, CV and test set
Find model using trial and error approach
ANN Output
Choose hold out sample percentage
Generate model using training set
Apply model to test set
GMDH Output

Figure 9. Methodology of Case Study.

# A Novel Event-based Method for Abducting Evolutional Motivation of Empirical Engineering Knowledge

Xinyu Li*[1], Zuhua Jiang[1], Lijun Liu[1], Puling Liu[1]
[1]Department of Industrial Engineering and Management, Shanghai Jiao Tong University
Shanghai, PR China
e-mail: lxy2003jacky@sjtu.edu.cn, zhjiang@sjtu.edu.cn, liulijun@sjtu.edu.cn, karrigen@sjtu.edu.cn

***Abstract*—In the engineering field, empirical engineering knowledge (EEK) accumulated from long-term engineering activities is the knowledge source for engineers to solve the innovative design and decision-making problems. Cognition and utilization of the mechanism and rules of EEK evolution over time are the real and urgent problems in knowledge management and seem to lack attention from researches. To deal with these problems, this paper proposes a novel method that abducts the motives of EEK evolution with the events related to the evolution of EEK field, completely and clearly finding the factors that influence the EEK evolution. An experiment in computer-aided design (CAD) is executed to verify the feasibility of the proposed method, and the result shows that the proposed method can effectively acquire the motives of EEK evolution and also be beneficial for engineers to deeply cognize the EEK evolution.**

**Keywords-*knowledge evolution; empirical engineering knowledge; evolutional motive; evolutional event; abductive reasoning.***

## I. INTRODUCTION

Driven by the rapidly emerging concepts, techniques, methodologies, experiences and activities, knowledge is fast maturing and mutating in this era of knowledge-driven economy. The effective management of such evolving knowledge is the key to maintain the competitiveness preponderance of the organizations and enterprises in creativity and adaptability [1]. Especially in the engineering field, empirical engineering knowledge (EEK), which is concluded and accumulated from engineering activities over years, is the knowledge source for engineers to solve the innovative design and decision-making problems [2]. Observing the evolutional process and then acquiring the rules and factors that motivate the process is an urgent problem that needs an answer in knowledge management. A proper knowledge management mechanism founded on the answer to this question will help the intellectual workers and practitioners obtain a deep cognition of the developments of the engineering field in a long period of time. They could also agilely adapt themselves to the changes of demands in the engineering activities, and more precisely forecast the future trends in the engineering field.

Therefore, the investigation on the motivation of EEK evolution is a task with high necessity, yet lacking attention from researches. To rectify this, based on the representation of network structure of EEK field, this paper proposes a novel event-based method for abducting the motives of the evolution of EEK. The evolutional patterns in the EEK evolution process are recognized and extracted in the first step, and then abductive reasoning is used for finding the factor events that influence the process of EEK evolution based on the construction of the archive of collected evolutional events. Because of the complete search for the possible explanation of evolutional patterns and because it offers the evolutional events in readable texts, the proposed method will help the engineers in cognizing the EEK evolution in depth.

The remainder of this paper is organized as follows. Section 2 introduces some related works of the proposed method. The general framework of the proposed method is designed in Section 3. Section 4 details the implementation of the proposed method by illustrating the evolutional patterns recognition and event-based abductive reasoning. The example of using the proposed method to acquire the motive events in the evolution of EEKs originated from computer-aided design (CAD) missions is presented in Section 5 and verifies the feasibility of the proposed method. The last section concludes the paper with some possible improvements.

## II. RELATED WORKS

The theory of evolution was initially designed for understanding and explaining the development of complex biological systems by Charles Darwin in 1842. In knowledge evolution, the fundamental hypothesis of a generalized evolution theory is that the mutating internal concepts of knowledge are chosen or eliminated in order to cope with the rapidly changing external environments, such as the demands or costs of engineering projects [10]. Although the practitioners in the domain could perceive the evolution process with the experience concluded and accumulated from engineering activities over a very long time, they have little understanding of the motives that may influence or even determine the development of the domain, as well as the degrees of impact brought by such motives. Taking accomplishing CAD missions as an example, the evolution of EEKs in CAD field and its benefits can be felt by the engineers through the increase of work efficiency in handling such missions. However, they know little about exactly what kind of new process or new approach that leads to such evolution, if without the complex experiments, measurements and analysis conducted by professional research institutions (see Figure 1).

There are few studies that focus on acquiring the motives of knowledge evolution. Existing related works can be categorized into three kinds: concluding empirical laws

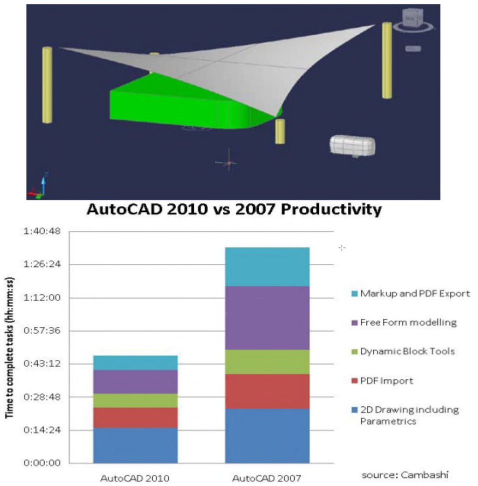


Figure 1. Work efficiency measurement: to find out the new features in AutoCAD 2010 that lead to a significant save of time in accomplishing a design task [11]

[7][8], using statistical analysis [4]-[6] and using complex networks analysis [3][9].

According to the observed relationships between the phenomena of knowledge evolution and related modifications in the field, some researches empirically proposed some reasonable explanations and hence summarize some laws of motivation of the evolution. Grebel [8] used four generic rules as behavioral assumptions and constructed a percolation model to empirically explain the motivations of the structural evolution of the network of research topics in basic science researches. Gross et al. [7] investigated the motivation of biology ontology evolution by detecting the changes in the results of statistical applications and analyzing corresponding modifications in the categories of ontology caused by new knowledge accumulation. Using such qualitative explanations and empirical laws that measure the effect of factors on the knowledge evolution, practitioners could obtain a global comprehension of the motivation of the evolution and easily forecast the future trend of the field, but the specific properties of the motivation (for example, involved concepts and occurrence time) may not be clearly elicited. The intellectual workers are still unaware of the motives of the special breaking points of the evolution process.

Some scholars also analyzed the statistics collected from the working process and environment to propose the factors that influence the knowledge evolution and their degree of impact. Erdil et al. [6] examined 14 statistics about employee interaction, information systems and organizational structure in the enterprises to measure the process of technological knowledge evolution. Johnson et al. [4] and De Noni et al. [5] investigated the factors in social nature of online communities and open source software communities separately to discuss some mechanisms that interact with the generation of new knowledge. Generally, motives tested and obtained from the statistical methods have rather high significance and strong persuasiveness, but they are often the external factors that have weak relations with the knowledge and the engineering fields, and therefore unable to reveal the internal factors of the motivation of the knowledge evolution.

With the proposal and implementation of the theories and tools of complex network analysis (CNA), measurements obtained from the knowledge networks were used to investigate the motives of the knowledge evolution. Modeling the existing knowledge as nodes with potentials in the network, Schumann et al. [9] concluded the motivation of evolution of the domain network with the interactions of the knowledge nodes, including splitting of high concentrated knowledge nodes and fusion of individual ones. Also modeling the knowledge sharing and diffusion with networks, Jiang et al. [3] utilized Exponential Random Graph Models (ERGMs) to examine the interactions and evaluate their impacts of network structure on a longitudinal data set that covered 1991-2010. However, in their works, knowledge is quantified to a node that contains little semantic information, leading to the imperfect integrity of their conclusions.

## III. PROPOSED METHOD

Oriented to acquire the motives of the EEK evolution, this paper proposed a novel event-based abducting method, based on the network-based representation of the structure of EEK field. Figure 2 presents the framework of this two-step method.

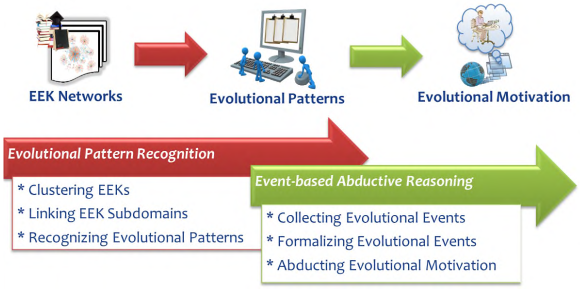

EEK Networks
Evolutional Patterns
Evolutional Motivation
Evolutional Pattern Recognition
* Clustering EEKs
* Linking EEK Subdomains
* Recognizing Evolutional Patterns
Event-based Abductive Reasoning
* Collecting Evolutional Events
* Formalizing Evolutional Events
* Abducting Evolutional Motivation

Figure 2. Framework of the proposed abducting method

**Evolutional pattern recognition:** Utilizing the EEK networks in continuous time intervals, the sorts and scales of the subdomains of the field of EEK are firstly acquired through clustering approaches; then the subdomains in neighboring time intervals are numerically and semantically compared to recognize two kinds of evolutional patterns; the patterns are formalized with the vectors in semantic space model for the convenience of subsequent calculations.

**Event-based abductive reasoning:** An archive of events is constructed with the events that are possibly related to the EEK evolution, and key information of each event is also refined; then the relationships of evolutional events and patterns are determined with cosine similarities of semantic vectors and occurrence time; using abducting algorithm, some events that obey the rules are chosen to form the motives chains, and finally used to explain the motivation of EEK evolution.

## IV. IMPLEMENTATION OF PROPOSED METHOD

### A. Evolutional pattern recognition

*1) Representing EEK subdomain with EEK clusters*

New ideas and concepts are often the consequences of the original ones [12]. In engineering field, such relationships are also helpful for generating the links among EEKs and establishing the EEK networks. Li et al. [13] proposed a corresponding modeling method. They firstly formalized EEK with seven kinds of attributes: *Engineering Problem*, *Problem Context*, *Problem Solution*, *Feature Association*, *Effectiveness*, *Contributor* and *Time*, and then determined the strength of relevance relationships of EEK pairs using the supervised fuzzy neutral networks, and finally used the pairs with high relationships to construct the EEK networks.

Although the networks established by Li et al. could fully consider the properties of EEK and precisely portray the structure of the EEK field, the networks are static and unable to reveal the evolution of EEK directly. To improve this, this paper firstly arrays the EEK with their time attribute, and categorizes EEKs with several continuous time intervals. Relevance relationship networks of each time intervals are separately established and the dynamic change of these networks are utilized for portraying the phenomena of EEK evolution.

Different from the approaches based on complex network analysis (CNA), when analyzing the phenomena of EEK evolution using the proposed method in this paper, EEK groups containing a bunch of strongly inter-related EEKs are focused, rather than some key nodes in the networks. This paper assumes that these groups could represent a category of EEKs that are accumulated from the engineering activities over a long time and verified by a large number of practitioners, dividing the engineering field into several subdomains. The variation of the sorts and scales of such subdomains could quantitatively illustrate the evolution of EEK. So, how to get the EEK groups in the networks is the initial problem for investigating the motives of the evolution. To solve this problem, the commonly used K-means clustering method is adopted to cluster the EEK nodes in the networks.

In order to guarantee that the representative EEK groups can be found and noise EEK nodes are filtered out, minimal number of members in the cluster $|C|_{\mathrm{threshold}}$ is set to reserve the EEK clusters with certain scale. A reserved cluster in time interval $T_n$ is denoted as $C_{n,i}$. Only the dynamic variations of these reserved clusters are analyzed in the following process of the proposed method.

*2) Linking corresponding EEK subdomains*

The sorts and scales of the subdomains in each time interval can be acquired directly from the clustered EEK network. However, the corresponding relations of subdomains in neighboring time intervals are unknown, and the development sequences of correspondingly same or similar subdomains are unavailable to extract.

This paper handles this unavailability with the calculation of semantic relations between clusters. Specifically, we firstly calculate the Term Frequency-Inversed Document Frequency (TF-IDF) weights of all the concepts contained in cluster $C_{n,i}$ of time interval $T_n$, and extract the key concepts with the highest weights. For a concept *NP*, its weight in $C_{n,i}$ is calculated as follows:

$$TFIDF(w) = \frac{Count(w)}{\sum_{w_k \in C_{n,i}} Count(w_k)} \times \log \frac{|C_{n,i}|}{|\{EEK \mid EEK \in C_{n,i}, w \in EEK\}|} \tag{1}$$

$$W(NP) = \frac{1}{|NP|} \sum_{w \in NP} TFIDF(w) \tag{2}$$

where *w* is a word, *Count(w)* is the count of occurrence of *w* in $C_{n,i}$, the sum of $Count(w_k)$ is the total count of words in $C_{n,i}$. $|C_{n,i}|$ is the count of containing EEKs, which is divided by the count of EEKs that contains *w*. $|NP|$ is the length of noun phrase of concept *NP*.

Key concepts and their weights are used to express the $C_{n,i}$ with a vector in semantic space model constructed by the concepts in the corpus, namely $\{W^{n,i}(NP_1), W^{n,i}(NP_2),\ldots, W^{n,i}(NP_k)\}$. *k* is the number of all noun phrases in the vocabulary of corpus. For two clusters $C_{n,i} = \{W^{n,i}(NP_1), W^{n,i}(NP_2),\ldots, W^{n,i}(NP_k)\}$ and $C_{n+1,j} = \{W^{n+1,j}(NP_1), W^{n+1,j}(NP_2),\ldots, W^{n+1,j}(NP_k)\}$ in neighboring time intervals $T_n$ and $T_{n+1}$, semantic similarity is computed with the cosine similarity of their semantic vectors:

$$\cos Sim(C_{n,i}, C_{n+1,j}) = \frac{\sum_{p=1}^{k} W^{n,i}(t_p) W^{n+1,j}(t_p)}{\sqrt{\sum_{p=1}^{k} \left(W^{n,i}(t_p)\right)^2} \sqrt{\sum_{p=1}^{k} \left(W^{n+1,j}(t_p)\right)^2}} \tag{3}$$

If the value of cosSim($C_{n,i}$,$C_{n+1,j}$) exceeds a pre-set threshold $CosSim_{threshold}$, which means the key concepts contained in $C_{n,i}$ and $C_{n+1,j}$ overlap to a certain degree, then two clusters are semantically similar, hence representing the same or similar EEK subdomains in neighboring time intervals.

*3) Recognizing evolutional patterns*

After the linking of corresponding EEK subdomains in all time intervals, the phenomena of EEK evolution can be quantitatively represented with the variation of scales. Three evolutional patterns can also be concluded: expansion, contraction and staying.

The knowledge expansion pattern is defined as the rapid raise of EEK numbers contained in EEK clusters, while the key concepts of the corresponding subdomain are not changed too much. The backgrounds of expansion patterns are often the emergence of some new concepts and approaches, or the sudden concentration on the existing original ones in the corresponding subdomains, which leads to the burst in adoption in related engineering missions and activities and abundant accumulation of empirical knowledge in the subdomains. The knowledge contraction pattern is a reversed pattern of knowledge expansion pattern. With the updating of the engineering field, obsolete experience and methods are gradually eliminated by the engineers, and the corresponding subdomains will also be marginalized or even disappeared. Both kinds of patterns will reflect the distinct changes in the evolution of the EEK field, while the other patterns not belonging to one of these two kinds are not considered in this paper.

According to the representations of the clusters extracted before, this paper formally defined two kinds of patterns as follows, and recognized them with (4) - (5).

**Knowledge Expansion Pattern:** if in neighboring time intervals $T_n$ and $T_{n+1}$, two clusters $C_{n,i}$ and $C_{n+1,j}$ are semantically related, and the size of $C_{n+1,j}$ are larger than $C_{n,i}$, namely:

$$\begin{cases} \cos Sim\left(C_{n,i}, C_{n+1,j}\right) \ge CosSim_{threshold} & \text{(a)} \\ \dfrac{\left|C_{n+1,j}\right|}{\left|C_{n,i}\right|} \ge Scale_{threshold} & \text{(b)} \end{cases} \qquad (4)$$

then an expansion pattern $P: C_{n,i} \xrightarrow{\underline{\text{KEP}}} C_{n+1,j}$ is recognized.

**Knowledge Contraction Pattern:** if in neighboring time intervals $T_n$ and $T_{n+1}$, two clusters $C_{n,i}$ and $C_{n+1,j}$ are semantically related, and the size of $C_{n+1,j}$ are smaller than $C_{n,i}$, namely:

$$\begin{cases} \cos Sim\left(C_{n,i}, C_{n+1,j}\right) \ge CosSim_{threshold} & \text{(a)} \\ \dfrac{\left|C_{n,i}\right|}{\left|C_{n+1,j}\right|} \ge Scale_{threshold} & \text{(b)} \end{cases} \qquad (5)$$

then a contraction pattern $P: C_{n,i} \xrightarrow{\underline{\text{KCP}}} C_{n+1,j}$ is recognized.

The degrees of scale variations of subdomains are judged by $Scale_{threshold}$, which is a positive number larger than 1. The sensitivity of recognizing evolutional patterns from linked clusters is affected by the setting of this threshold. The larger of $Scale_{threshold}$, the larger degree of changes are revealed in the evolutional patterns, yet the fewer sorts of subdomains are considered. For a recognized evolutional pattern $P$: $C_{n,i} \rightarrow C_{n+1,j}$, it can also be represented with the semantic vectors in semantic space model as $P=\{W^P(NP_1), W^P(NP_2),\ldots, W^P(NP_k)\}$, and $W^P(NP_q)$ in it is computed as:

$$W^{p}\left(NP_q\right) = \frac{\left|C_{n+1,j}\right| W^{n+1,j}\left(NP_q\right) - \left|C_{n,i}\right| W^{n,i}\left(NP_q\right)}{\left|C_{n+1,j}\right| - \left|C_{n,i}\right|} \qquad (6)$$

Besides that, the time of duration of the evolutional pattern is also considered with the involving clusters and valued with $T_n \cup T_{n+1}$.

### B. Event-based abductive reasoning

#### 1) Collecting and formalizing evolutional events

Although the recognized patterns could infer some information about the evolution process of EEK over a long period of time, it is difficult for engineers to understand the meanings since these patterns are expressed with concepts and weights, lacking readable explanation texts.

Therefore, this paper uses texts of events described with natural language to infer and explain the patterns and their motives. Such events are evolutional events, which are the facts that already happened at a certain time, strongly related to the knowledge evolution or directly lead to the evolution. They are derived from the news of tools updating, the investigations of authorized institutions, the summaries from experienced long-term practitioners, or other records of domain-related comments. Table I shows an illustrative record of an evolutional event, describing the event of adding parametric design tools in AutoCAD 2010. This event aroused strong repercussions of the users and finally result the evolution in EEKs in CAD field.

TABLE I. AN ILLUSTRATIVE RECORD OF EVOLUTIONAL EVENT

| |
|---|
| **Time:** 2009.329 |
| **Content:** The geometry in AutoCAD has always driven the dimensions. We draw a line the correct length and then dimension the line. What if you could drive the geometry from the dimensions? You change the value of the dimension and the geometry automatically updates! That is exactly what we now have in AutoCAD 2010. |

For these natural-language-described texts, Song et al. [14] proposed a processing method by selecting some key phrases from the texts to represent the events. They used Stanford Parser to find the noun phrases and chose those with large IDF weights in order to filter out the common words and reflect the characteristics of the texts.

This paper also maps the events to vectors in semantic space model, namely $E=\{W^E(NP_1), W^E(NP_2),\ldots, W^E(NP_k)\}$. $W^E(NP_r)$ is the IDF weights of $NP_r$, which is calculated with all the documents of evolutional events. Semantic similarities between events and patterns, and among events, can also be computed with (3). The occurrence of events can be acquired directly from the source of texts and denoted with $t_E$.

#### 2) Abducting evolutional motives

Abductive reasoning is a kind of logical inference which goes from an observation to a theory which accounts for the observation, seeking the possible explanations for the happened phenomena[15]. Abductive reasoning, accompanied by deductive reasoning and inductive reasoning, is an indispensable part of human cognitive activities[16]. We use the EEK of new features listed in Figure 1 as an example to illustrate the process of abductive reasoning. We observe the significant decrease of the cost of time when accomplishing the CAD missions, which response to the evolution of CAD field in shaping and modeling. And according to the work efficiency measurement, the EEK of new features will lead to such decrease. Therefore we construct the probable causal association that the EEK of new features is the motive of the evolution of CAD field if there are no other conflicting rules. Even though the modification of the measurement reports or the proposal of more persuasive surveys will vary the belief of this causal association, or even disconfirm the association, some interesting explanations may still be found and useful conclusions will be probably refined.

In abducting the motives of the evolution, the set of evolutional patterns $\{P\}$, the set of evolutional events $\{E\}$, and the set of rules $\{R\}$ are the inputs of the reasoning process. If an event $E$ is the motive of a pattern $P$ according to $\{R\}$ and denoted as M($E \rightarrow P$), it should satisfy two conditions:

- $P$ follows from $E$ according to $\{R\}$;
- $E$ is consistent with $\{R\}$.

Three reasoning rules are put into $\{R\}$. These rules constrain the explaining of unrelated or contradictory events for the motives of the evolutional patterns:

**Rule 1**: Evolutional event $E_i$ is a possible cause of event $E_j$, if $E_i$ and $E_j$ are semantically related and $E_i$ is happened earlier than $E_j$;

**Rule 2**: Evolutional pattern $P$ is a possible consequence of event $E$, if $P$ and $E$ are semantically related and $E$ is happened earlier than the end of $P$;

**Rule 3**: A motive chain M($E_i \rightarrow E_j \rightarrow P$) is constructed, if event $E_i$ is the possible cause of event $E_j$ and pattern $P$ is the possible consequence of event $E_i$ and $E_j$ simultaneously.

The process of abducting algorithm is listed as follows:

**Input:** evolutional pattern set $\{P\}$, evolutional event set $\{E\}$ and rule set $\{R\}$;
**Output:** the set of possible motives $\{M\}$;
**Process:**
(1) Create an empty motive set $\{M\}$;
(2) Choose an earliest begun and undiagnosed evolutional pattern $P$ from $\{P\}$;
(3) Choose a latest happened and unchecked evolutional event $E$ from $\{E\}$; find all possible causes of $E$ according to Rule 1, then construct *List*<*E*>;
(4) Choose a latest happened and unchecked $E_i$ in *List*<*E*>, add a motive chain M($E_i \rightarrow P$) into $\{M\}$ if $P$ and $E_i$ satisfy Rule 2;
(5) Repeat step 4, until all the events in *List*<*E*> are checked; merge the motive chains with Rule 3;
(6) Repeat step 3-5, until all the events in $\{E\}$ are checked; save $\{M\}$ for $P$; set all the events in $\{E\}$ unchecked;
(7) Repeat step 2-6, until all the patterns in $\{P\}$ are diagnosed; output $\{M\}$.

With the readable appendix texts of the evolutional events, it will be easier for engineers to understand the evolution process of EEKs with the clear and specific motives, hence is helpful for them to obtain a deep cognition of the knowledge evolution. Meanwhile, the output motive chains can also be further verified and evaluated by domain experts, escalating the relationship between evolutional events and evolutional patterns from the statistical correlation to more cogent logical correlation.

## V. CASE STUDY

From three professional virtual communities forums *autodesk.com*, *www.cadtutor.net* and *www.cadforum.cz*, 3276 EEKs of accomplishing computer-aided engineering design missions using AutoCAD software were elicited and formalized, ranging from February 2001 to September 2015. The evolutional patterns were recognized from the networks constructed by these EEKs. For the evolutional events related to the evolution of CAD field, *ReadMe* documents of each update and all software versions ranging from AutoCAD version 14.0 (AutoCAD R14, published in 1997, February) to version 20.1 (AutoCAD 2016, published in 2015, March) were downloaded from the official website *www.autodesk.com*, in which detailed the emergence of new tools and the modifications in original functions in AutoCAD software. Proposed by an authorized institution *HyperPics Consult Company*, open accessed documents of news of the software *AutoCAD What's New* were also collected. We also collected the long-term experienced user's summary *AutoCAD Tips & Tricks Booklets* written by Lynn Allen, who has used AutoCAD for over 25 years and served as Autodesk University emcee for over 10 years. 1080 records of evolutional events, as shown in table 1, were finally extracted.

The whole time span was divided into five time intervals: 2001-2003, 2004-2006, 2007-2009, 2010-2012, and 2013-2015. The EEKs were clustered in each EEK network, and Figure 3 shows the clusters in EEK network of time interval 2001-2003. The number of initial clusters $k$ in K-means was set to 30, and minimal number of cluster member $|C|_{\text{threshold}}$ was set to 4. Similar EEK subdomains in the five networks were linked with $CosSim_{threshold}$=0.5. 28 evolutional patterns were recognized when $Scale_{threshold}$=2, containing 17 knowledge expansion patterns and 11 knowledge contraction patterns. Using abducting algorithm, evolutional patterns were explained with the acquired motive chains. An evolutional pattern and its abducted motives chains are shown in Table II.

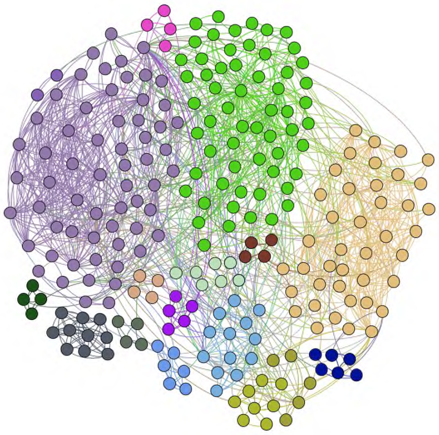


Figure 3. Clusters in CAD EEK network of time interval 2001-2003

TABLE II. AN ILLUSTRATIVE RECORD OF EVOLUTIONAL EVENT

| **Pattern:** $C_{3,11} \xrightarrow{\text{KEP}} C_{4,2}$ | | | | **Motive Chains** |
|---|---|---|---|---|
| ***C₃,₁₁*** (Size: 25 Time: 2007-2009) | | ***C₄,₂*** (Size: 66 Time: 2010-2012) | | |
| **Concepts** | **Weights** | **Concepts** | **Weights** | |
| object | 0.2078 | object | 0.2056 | |
| cursor | 0.1391 | angle | 0.1345 | **Chain 1:**<br>Adding Constraints → Inferring Geometric Constraint → P |
| angle | 0.0927 | constraint | 0.0824 | |
| parameter manager | 0.0637 | parameter manager | 0.0736 | |
| drawing | 0.0545 | drawing | 0.0563 | |
| dimension | 0.0499 | cursor | 0.0486 | |
| length | 0.0477 | distance | 0.0425 | |
| distance | 0.0383 | dimconstraint | 0.0424 | |
| vertex | 0.0315 | block | 0.0422 | |
| acad line | 0.0297 | plane | 0.0386 | **Chain 2:**<br>Changing to Annotational Dimensions → P |
| perpendicular | 0.0273 | direction | 0.0292 | |
| geometry | 0.0269 | polyline | 0.0260 | |
| object snap | 0.0260 | degree | 0.0252 | |
| degree | 0.0260 | object snap | 0.0239 | |
| intersection | 0.0259 | intersection | 0.0230 | |
| plane | 0.0249 | selection | 0.0225 | **Chain 3:**<br>Dynamic Block → P |
| polyline | 0.0238 | geomconstraint | 0.0222 | |
| grid | 0.0222 | vertex | 0.0217 | |
| polar point | 0.0216 | dynconstraint | 0.0211 | |
| selection | 0.0206 | polar point | 0.0184 | |

Parametric design is a milestone in the development process of CAD field. The fundamental principle of parametric design is using the geometric constrains and variable parameters to conveniently manipulate and rapidly modify the drawings, which significantly accelerate the speed in plotting and promotes the transition from design intent to design response [17]. The tools of parametric design were firstly added into AutoCAD 2010 in 2009, triggering a positive response from the majority of CAD engineers. They frequently applied dynamic blocks with geometric constraints and dimensional constraints to accomplish the engineering projects and accumulated abundant related experiences. According to the aforementioned report [11], it is the emergence of these parametric design tools that consequently motivated to the evolution of EEKs in subdomain of modeling and shaping in CAD field. The abducted motives chains in Table II are also consistent with the motivation concluded from this professional report.

These motive chains were also verified by the domain experts in order to prove their validities. Some motives chains in abducted results were deleted according to their evaluation. Finally motivations of 25 evolutional patterns in all 28 were firmly explained with these evolutional events, while the other 3 patterns did not obtain persuasive motivations.

## VI. DISCUSSION AND CONCLUDING REMARKS

Explaining the evolutional patterns with the evolutional events, this paper proposed a novel method for investigating the motivation of EEK evolution. Based on the networks representing the structure of the field of EEKs, clustering algorithm is adopted to divide the subdomains of the networks in each time intervals. EEK evolutional patterns are recognized with the scales and semantic informations of the linked subdomains in neighboring time intervals. Evolutional events related to EEK evolution are collected and used to abduct the motive chains, explaining the motivation of the evolution in depth. Evolution of EEKs in CAD is investigated and evaluated with the experts and practitioners, proving the feasibility and effectiveness of the proposed event-based abducting method in acquiring the clear and specific motivation of the EEK evolution.

The advantages of the proposed method are in three aspects. Firstly, semantic meanings of the EEKs are fully considered in the method. Therefore, our method can provide a more integrated motivation of the EEK evolution than most of traditional works. Secondly, our method uses the domain-related events to investigate the factors that impact the evolutional process, making our motivation more facilitated to those intellectual workers in the domain. At last, the utilizing of abductive reasoning is consistent with human cognitive activities. It will mine all possible motives according to the input event archives, which significantly promotes the discovery of new interesting explanations.

There are several possible improvements for our methods. First, according to the flowchart in Figure 3, the maximum computational complexity of the algorithm is $O(1/2|P||E|^2)$. In order to shorten the operation time when $|E|$ is larger, more compatible filtering rules should be added into rule set $\{R\}$. Second, although the motive chains are acquired in this paper, the quantitative degree of their impact on the semantic meaning and scales of the subdomains is less considered and will be paid attention in the future research.

## ACKNOWLEDGMENT

The research was supported by National Natural Science Foundation of China (No. 70971085, 71271133), Shanghai Science and Technology commission (No. 13111104500), and Shanghai Municipal Education Commission (13ZZ012).

## REFERENCES

[1] G. Dosi, M. Faillo, and L. Marengo, “Organizational capabilities, patterns of knowledge accumulation and governance structures in business firms: An introduction”, *Organization Studies*, vol. 29, no. 8-9, 2008, pp. 1165-1185.

[2] L. Liu, Z. Jiang, and B. Song, “A novel two-stage method for acquiring engineering-oriented empirical tacit knowledge”, *International Journal of Production Research*, 2014, pp. 1-22.

[3] S. Jiang, Q. Gao, H. Chen, and M.C. Roco, “The Roles of Sharing, Transfer, and Public Funding in Nanotechnology Knowledge-Diffusion Networks”, *Journal of the Association for Information Science and Technology*, vol. 66, no. 5, 2015, pp. 1017-1029.

[4] S. L. Johnson, S. Faraj, and S. Kudaravalli, “Emergence of power laws in online communities: the role of social mechanisms and preferential attachment”, *MIS Quarterly*, vol. 38, no. 3, 2014, pp. 237-795.

[5] I. De Noni, A. Ganzaroli, and L. Orsi, “The evolution of OSS governance: a dimensional comparative analysis”, *Scandinavian Journal of Management*, vol. 29, no. 3, 2013, pp. 247-263.

[6] A. Erdil and H. Erbiyik, “The effect of information technologies on employees”, *Proceedings of the 6th Knowledge Cities World Summit (KCWS 2013)*, 2013, pp. 733-756.

[7] A. Gross, M. Hartung, K. Pruefer, J. Kelso, and E. Rahm, “Impact of ontology evolution on functional analyses”, *Bioinformatics*, vol. 28, no. 20, 2012, pp. 2671-2677.

[8] T. Grebel, “Network evolution in basic science”, *Journal of Evolutionary Economics*, vol. 22, no. 3, 2012, pp. 443-457.

[9] C. Schumann and C. Tittmann, “Evolution Analysis of Knowledge Potentials by Pattern Matrices”, *Proceedings of the 11th European Conference on Knowledge Management, Vols 1 and 2*, 2010, pp. 892-900.

[10] D. J. Futuyma and T. R. Meagher, “Evolution, science and society: Evolutionary biology and the national research agenda”, *California Journal of Science Education*, vol. 1, no. 2, 2001, pp. 19-32.

[11] Cambashi Inc., “Autodesk whitepaper: Study on the efficiency of AutoCAD2010”, 2009

[12] P. C. Palvia, S. Palvia, and J. E. Whitworth, “Global information technology: a meta analysis of key issues”, *Information & Management*, vol. 39, no. 5, 2002, pp. 403-414.

[13] X. Li, Z. Jiang, B. Song, and L. Liu, “Network-based relevance relationship generating for empirical engineering knowledge”, *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, 2015, pp. 272.

[14] B. Song, Z. Jiang, and X. Li, “Modeling knowledge need awareness using the problematic situations elicited from questions and answers”, *Knowledge-Based Systems*, vol. 75, 2015, pp. 173-183.

[15] D. Poole, “Explanation and prediction: an architecture for default and abductive reasoning”, *Computational Intelligence*, vol. 5, no. 2, 1989, pp. 97-110.

[16] T. Menzies, “Applications of abduction: knowledge-level modeling”, *International Journal of Human-Computer Studies*, vol. 45, no. 3, 1996, pp. 305-335.

[17] J. Li and J. Huang, “Design and Implementation of the Parametric Designing System Based on ADO”, *Advanced Materials Research*, 2012, pp. 281-284.