UNIVERSITÀ DI NAPOLI L'ORIENTALE
Dipartimento Asia, Africa e Mediterraneo

Studi Africanistici
Serie Egittologica 6

Ancient Egypt
New Technology

Edited by
Stefania Mainieri & Rosanna Pirelli

UniorPress

On the cover: digital reproduction of the south wall
of the King's Chamber in the temple of Ramesses III at Medinet Habu, Luxor.
Image from the article in this volume, *The King's Chamber: A Digital Publication Prototype*,
by A. Singer, O. Murray, and A. Pantos.
Photos: Epigraphic Survey Photo Archive;
black and white: Henry Leicher; color: Owen Murray.

UniorPress

Università di Napoli L'Orientale
Dipartimento Asia, Africa e Mediterraneo

Studi Africanistici

Serie Egittologica
6

UNIVERSITÀ DI NAPOLI L'ORIENTALE
Dipartimento Asia, Africa e Mediterraneo

Studi Africanistici

Serie Egittologica
6

Ancient Egypt New Technology

Edited by
Stefania Mainieri & Rosanna Pirelli

UniorPress
Napoli 2026

UniorPress
Via Nuova Marina, 59 - 80133, Napoli
uniorpress@unior.it

ISBN 978-88-6719-368-4

Il presente volume è stato sottoposto al vaglio di revisori anonimi

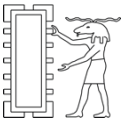

Table of Content

A digital archive for religious texts from the Nile Valley and beyond

Federico Maria Avano, Angela Bosco, Andrea D'Andrea,
Gilda Ferrandino & Zied Mnasri

Abstract

Within the project ITSERR, granted by Recovery Plan, the WP 10 ReTINA aims to create an environment that defines and optimizes guidelines for the digitization of a diverse range of sources, considering the challenges posed by very different languages and types of scriptural media. The dataset includes religious texts from sites in the ancient world of the Nile Valley and beyond. The contents of these texts are very diverse and varied: descriptions of religious ceremonies, funerary texts, prayers, incantations, lists of offerings and temple personnel. So far, several attempts have been made to achieve automatic textual analysis from digitized religious texts. However, three main problems still hinder this task: firstly, the lack of sufficient material, which results in small data sets, mostly related to a single site; secondly, the variety of languages and writing forms; and finally, the lack of dedicated methods, intentionally developed to analyze and process such ancient religious texts.

This paper intends to bridge the gap between the state of the art of linguistic text analysis on the one hand and image processing applied to historical texts on the other, starting from a review of AI and metadata projects in the field of Egyptian language text analysis. The paper aims to: a) study the datasets, methods and tools that enable the restoration of missing text fragments from scanned images of religious texts from the Nile Valley, including funerary inscriptions on tombs and other types of related materials such as papyri, parchments, wood, etc.; b) apply image processing and linguistic analysis to obtain transcription and possibly analysis and restoration of other religious manuscripts, covering a wide range of languages. The survey intends to provide a comprehensive review of the state of the art of this research topic, including three main aspects, which will be the focus of attention during the implementation of the project: i) datasets, including digitized images and manuscripts, ii) open data repositories, with associated metadata standards, taxonomies and thesauri, iii) the state of the art projects that has already been realized, either for Egyptological or other archaeological data annotation and 3D visualization and iv) artificial intelligence methods and tools, mainly for image processing and linguistic analysis, commonly used for this type of problems. Finally, a novel idea for applying artificial intelligence methods to the related data types is proposed in the framework of the project ITSERR.

Keywords: *Artificial Intelligence; Egyptian religious texts; Metadata; Image processing; Deep and Machine learning*.

1. Introduction

This paper is a preliminary review of Artificial Intelligence (AI) and metadata projects starting from the main goals of ITSERR (Italian Strengthening of the ESFRI RI RESILIENCE) project.[1] The main questions we asked are related to what data are used for the AI applications and how they are managed in the open data repositories, what are the projects currently exist and how image recognition can be used with 3D. On the base of this preliminary survey and the objectives of the ITSERR project we have to define our strategies of research.

[1] Funded by the European Union- Next Generation EU, Mission 4 Component 2 CUP B53C22001770006.

The proposal is part of Work Package 10 of the ITSERR project, an Italian interdisciplinary infrastructure for religious studies. ITSERR, funded under the Italian Recovery Plan, aims to define, and optimize guidelines for digitizing a wide range of data, including religious texts from ancient world sites in the Nile Valley and beyond. The contents of these texts are very diverse and varied: descriptions of religious ceremonies, funerary texts, prayers, incantations, and lists of offerings. The project fits into the RESILIENCE (Religious Studies Infrastructure: Tools, Innovation, Experts, Connections and Centers) European network infrastructure.[2]

In 2021, the European Strategy Forum on Research Infrastructures (ESFRI) Roadmap identified RESILIENCE as a new project belonging to the area of social and cultural innovation. Led by the Foundation for Religious Studies (FSCIRE, Bologna-Palermo, Italy), RESILIENCE is an interdisciplinary research infrastructure for religious studies that builds a high-performance platform, providing tools and services to scholars studying religions in their different forms and in their diachronic and synchronic variety. The digital archives are dedicate to complex, rare and/or endangered religious texts written in various media (stone, papyrus, etc.). The user needs analysis conducted by RESILIENCE enables ITSERR to effectively address the scholarly community at the European and national levels. Among many goals, RESILIENCE aims to develop semantic models and multilingual data and metadata standards for religious studies research. Furthermore, different digitization approaches, and 3D technologies will be tested to identify ways to digitally acquire these materials. Our team is currently involved in the analysis of the existing AI projects and open data repositories in archaeology for understanding how it is possible to extract information from shared knowledge, as well as drawing up Best Practice for digitization approaches and 3D technologies.

2. Preliminary survey

In recent years, there has been significant growth in AI, with machine learning algorithms being increasingly published and made accessible for use. Simultaneously, vast amounts of data continue to be generated, inundating the web. Data serves as the fuel for training powerful machine learning programs. Given the pivotal role of data and code in powering AI, prioritizing the quality of digital data, particularly in terms of consistent labelling, is paramount.

For decades, numerous domain specialists have focused on the collection, annotation, and archiving of data to ensure the long-term preservation of valuable digital assets. The overarching goal has been to facilitate their discovery and reuse for further investigations and applications, with AI being the latest beneficiary of these efforts.

Various projects have invested in data management infrastructures, particularly in fields like archaeology where several digital repositories have been established. Resources are stored in different structures such as aggregators, digital libraries, or simple databases. The main part of the repositories aims to promote data quality to facilitate open-ended reuse, often adhering to FAIR principles of Findability, Accessibility, Interoperability, and Reusability.

[2] https://www.resilience-ri.eu/

In Italy, the CulturaItalia project endeavours to aggregate data related to the country's cultural heritage. The project's portal collects and organizes information from participating providers, with data described according to metadata standards by the entities managing the resources. Since 2008, CulturaItalia has served as an accredited aggregator for Europeana,[3] a European initiative promoting digital cultural heritage. Noteworthy cloud-based projects such as the UK Archaeology Data Service (ADS) provide infrastructure to ensure the FAIRness of archaeological research data in the long term under payment. ADS offers data on sites and monuments, grey literature, fieldwork reports, and research archives. Additionally, ADS collaborates with ARIADNE[4] to integrate existing archaeological research data infrastructures. Another notable project is Open Context[5] by the Alexandria Archive Institute.

While these projects contribute to digitizing museum collections and other resources, data descriptions often fall short in accuracy or detail, impacting data quality. Sometimes, data can only be explored through brief textual descriptions and locations, lacking access to images or thumbnails, posing a significant challenge for AI applications, especially image recognition. In the realm of textual data, two projects stand out for their implementation of good practices to enhance data quality: the ERC project PAThs[6] and the Beta Maṣāḥəft project.[7] PAThs aim to provide a comprehensive understanding and representation of the geography of Coptic literary production, particularly focusing on religious texts produced in Egypt between the 3rd and 13th centuries. The main scientific outcome of PAThs is the Archaeological Atlas of Coptic Literature.

Similarly, the Beta Maṣāḥəft project seeks to compile manuscripts, literary and documentary texts, along with reference authority files for historical figures and places. These resources adhere to FAIR principles, with each having a persistent identifier and described with rich metadata standards, ensuring accessibility, interoperability, and reusability for further applications.

3. AI for Archaeology

The most common applications of AI algorithms in archaeology aim to simplify and speed up the processes of analysis and documentation (examples include projects for the automatic recognition of ceramic fragments, or those supporting the re-composition of decorative elements, e.g. ArchAide and RePair[8]); to support the scholar in research and typological classification (this is the case of on-line catalogues that exploit AI as a real super search engine, e.g. Cleo[9]); to identifying new archaeological sites (there are many projects that use neural networks in the field of remote sensing,[10] obtaining incredible results, especially in terms of time savings on searches, e.g. the Kalam project[11]) and deciphering unknown languages or interpreting incomplete.

---

[3] https://www.europeana.eu/
[4] https://ariadne-infrastructure.eu/
[5] https://opencontext.org/
[6] http://paths.uniroma1.it/
[7] https://www.betamasaheft.uni-hamburg.de/
[8] http://www.archaide.eu; https://www.repairproject.eu
[9] https://cleo.aincient.org/pages/en/
[10] Argyrou, Agapiou, 2022.
[11] https://site.unibo.it/kalam/en

Almost all projects are based on automatic image recognition methods. Automatic imagery detection is a data-based method that is highly dependent on data, and the trend towards automation results from the availability of more complex and high-quality data. Data collection involves the extraction and storage of data from various data catalogues.

Despite the great evolution of these systems, which generally exploit deep learning Convolutional Neural Networks methods, the accuracy of the results is highly dependent on the classes defined in the training phase. Indeed, there is no missing case of completely meaningless results.

Different approaches of Artificial Intelligence methodologies have been used for ancient texts: application of OCR systems for scripts that have already been deciphered and classified (this is the case, for example, of the OCR - PT - CT Project on Egyptian hieroglyphics[12]); applications on signs that are difficult to interpret, also due to the writing media, such as clay tablets that make sign recognition operations more complex texts. Such cases require precise annotations and extensive image segmentation work (e.g. DeepScribe Project,[13] which refers to a database of 5,000 images of annotated tablets and 100,000 bounding boxes of cuneiform signs from the Persepolis Fortification Archive); and finally, applications that provide virtual interpretation and restoration of inscriptions, like the Ithaca project[14] on ancient Greek inscriptions.

It is well known that epigraphy cannot avoid a combined analysis of medium and text. The intrinsic connection of the written text with its support is often lost, or at least more blurred, at the digitalisation stage. 3D acquisition methodologies with high photographic resolution, which have recently had great development and application in the field of Cultural Heritage, could in the near future, combined with deep learning algorithms, overcome this gap in the proper digital documentation of these kinds of artefacts.

4. AI and 3D

Three-dimensional data processing has achieved very high levels of development. It is now possible to create a digital twin of the archaeological asset, thanks to the integrated use of tools, software and sensors. This has made it possible to create a data sharing environment, even in real time.[15]

Unfortunately, the absence of a standard workflow, which can be defined as an 'open method', does not allow the extraction of objective information from the shared models. Interesting examples of these immersive environments are developed, following an internally standardized workflow, even by Italian researchers but still at a visualization level.[16] Nevertheless, the three-dimensional model is no longer to be considered as a "simple" visualization tool, it must be intended as an innovative field of research.

Following these examples, limited to the workflow for acquisition, it is certainly possible to create unified standards in data collection which must not be influenced by interpretative schemes. This kind of approach would open up new and promising

[12] https://www.mortexvar.com/ocr-pt-ct. See also in the present volume the article by Gracia Zamacona *et al.* 2026.
[13] https://datascience.uchicago.edu/research/deciphering-cuneiform-with-artificial-intelligence/
[14] https://ithaca.deepmind.com
[15] Demetrescu *et al*. 2020.
[16] Fanini *et al*. 2021.

horizons in the field of AI applications to archaeological 3D models helping researchers to converge on a generally recognized method.

5. Project goals and prospected methods

These premises are some of the key points of our survey. This work is useful to define strategies that help us to develop the ITSERR objectives for providing services to RESILIENCE platform. The project is then subdivided into three main tracks, each taking care of a different side of the project, as described in the following. Besides, the data materials and computational tools that need to be collected and used will are thoroughly detailed.

5.1 Projected works

5.1.1 Track 1: Image-based hieroglyphic character recognition

The first goal of this project to create an Image Based Hieroglyphic Character Recognition that allows anyone curious about the meaning of hieroglyphs to utilize an algorithm to match the hieroglyphs to a well-known language, such as English. In the field of image processing, such as a software should be primarily developed with Optical Character Recognition (OCR) and machine learning tools.

The projected software should be able to provide, from scanned images of hieroglyphics, a translation in a modern language such as English, using either image recognition, or automatic hieroglyphic symbol translation.

As required materials, a large collection/corpus of hieroglyphic texts and images is required. The projected deliverable is an open-source software for online/offline hieroglyphic character recognition and prospectively providing and English translation.

To implement such a software, the following tasks need to be fulfilled:

- Collection/acquisition of a large corpus of hieroglyphics
- Image scanning and quality control
- Applying image pre-processing: denoising, segmentation and edge detection
- Development of a model of pattern recognition for hieroglyphics, based on machine learning tools such as Deep Neural Networks (DNN), Hidden Markov Models (HMM) and/or Support Vector Machines (SVM).
- Development of a machine translation tool, that allows converting the recognized hieroglyphic symbols to text in the language selected, e.g., English.
- Software test and validation

5.1.2 Track 2: Restoring ancient text using deep learning

One of the most significant issues with epigraphic inscriptions is that they are frequently destroyed over time, necessitating the restoration of unintelligible portions of the text by specialists known as epigraphists. As a result, ancient text restoration uses machine learning methods to recover lost characters from a damaged text input. Long-term context information should be handled by the algorithm, as well as missing or corrupted letter and word representations.

Therefore, a software able to provide from old scripts the missing parts of the text is projected. The input can be either a script or a scanned image. In the latter case, a character recognition step is required.

As required material, a corpus of ancient texts, preferably mixed (with and without spontaneously missing parts) is necessary. The deliverables consist in an open-source software that can be implemented following these steps:

- Collection/acquisition of a large corpus of epigraphy in a selected language, e.g., ancient Greek, Latin, Hebrew, Arab, etc...
- OCR scanning and quality control
- Development of a model able to guess the missing words/letters in the text taken from the epigraphy. Such a model should be based either on statistical N-gram models, or on machine learning models, such as word2vec and skip-gram that may provide a word embedding model. The latter is used to suggest the missing word/character.
- Software test and validation

5.1.3 Track 3: Lexical and semantic analysis of ancient sacred texts using machine learning and Natural Language Processing (NLP)

The practice of examining and analysing vast amounts of text data to extract high-quality information based on patterns and trends in the data is known as text mining. The study of similarity measures, as well as opinion mining or emotive analysis represented in the texts, are examples of patterns and trends. Text data mining uncovers hidden relationships in one or several text data sets. Applied to ancient religious texts, text mining may provide more insight and information about the cultural and religious context, and the hypothetic interconnection between neighbouring or contemporaneous religions.

The outcome is a software able to analyse ancient religious texts, to show the degree of similarity, influence and interconnection. Therefore, a corpus of ancient religious texts, preferably categorized by historic era and geographic location is required to provide input data. As a deliverable, an open-source software is projected. It should be implemented as follows:

- Text pre-processing: including tokenization, tagging, chunking, stemming, and lemmatization.
- Text normalization: including sentence extraction, HTML escape sequences, expanding contractions, lemmatizing text, removing special characters, stop words, unnecessary tokens, stems, and lemmas
- Text summarization and information extraction
- Clustering of texts, using similarity measures (metrics such as distance measures)
- Semantic and sentiment analysis model
- Software test and validation
- Finally, once the projected software parts are tested and validating, it will be possible to deploy them, either as online tools, or also as portable applications, either separately or jointly in an integrated system.

5.2 Prospected computational methods

The tasks required to implement the aforementioned applications can be classified into standard and specific ones. Standard tasks include text and image scanning and quality control, image enhancement and denoising, segmentation and edge detection,

text normalization, lemmatization and tokenization, and finally software test and validation. However, other tasks are more specific and require using or developing special computational methods or training novel models based on the input data. Such tasks are for example pattern recognition and automatic symbol-to-word translation (for track 1), text analysis and word prediction, cf. Track 2, and text summarization, information extraction and clustering, and sentiment analysis through semantic models, as required for track 3. To fulfil such special tasks, data-driven models, based either on machine learning or deep learning are highly recommended.

5.2.1 Deep learning methods

Deep Neural Networks (DNN) have shown notable success in several classification, recognition, and prediction problems, related to several areas such as image, text, speech, and natural language processing, in comparison to other "shallow" neural networks like the Perceptron architecture. The "deep" portion of "deep learning" refers to a neural network with a high number of layers together with extra weights and biases to boost the neural network's capacity to approximate more difficult tasks. There are many distinct types of neural networks for various goals in the complex field of deep learning. Novel DNN versions can be either recurrent or convolutional. Recurrent Neural Networks (RNN) are generally used to handle dynamic/time-varying signals, while Convolutional Neural Network (CNN) are usually used to train static/spatial data, such as pictures.

5.2.2 Machine learning methods

Despite the novelty of deep learning methods, pattern recognition and particularly hand-written character recognition have a long history of probabilistic modeling, using especially the Hidden Markov Models (HMM). This powerful tool has proved its ability to predict hand-written characters with a high precision, thanks to its sound mathematical formulation, based on a state-based model and the estimation of transition and emission probabilities. Thus, the features extracted from the text, or the image represent the observations, whereas the targets to be predicted are the characters or the patterns. Therefore, HMM can be used as a complementary tool, when DNN for example fails to achieve accurately the pattern recognition task; or to extract useful information that can be used to boost the performance of DNN. For instance, HMM-based transition and emission probabilities can be used as input features for the DNN. Conversely, a DNN can be used to cluster HMM models at the final prediction stage, instead of the standard classification trees. Particularly, this choice has been proved to improve speech recognition task and has been adopted in several speech and language processing tasks.[17]

Another powerful machine learning method is the Support Vector Machines (SVM) that were mainly applied for Optical Character Recognition (OCR).[18] However, their classification and regression abilities have been assessed in several problems, such as Historical Document Processing and Handwritten Character Recognition.[19] SVM are

[17] Zangar *et al*. 2021.
[18] Scholkopf *et al*. 1999.
[19] Nasien *et al*. 2010.

also based on a sound mathematical formulation and have the advantage of being highly explainable and interpretable. They are based on approximating a discrimination function that divides the feature space. Thus, it can be applied either to binary, multi-class or multi-label classification.

5.2.3 Unsupervised methods

The methods mentioned above are based on supervised learning, i.e. where all datasets are labeled, and the task consists in learning a model that fits the input features and the target labels. However, large datasets are either not totally labels, or require a huge amount of time and human resources to accurately label all samples, e.g. categorizing all images and analyzing texts, with the underlying costs and risk of errors. Therefore, another approach can be used when using large unlabeled datasets, includes unsupervised modeling. These models are based on clustering methods, such as K-means, that aim at grouping data into clusters, based on similarity or distance measures. Thus, the aim of training is to minimize the distance measure through the computation of a cost function. Once the cost function is optimized, the clustering model can be applied to all samples. However, it should be noted that such a method does not provide the desired labels as output, and therefore, a further step is required to match clusters with labels.[20]

6. Projected applications

The aforementioned methods can be used to establish analytic models or learn data-driven ones, based either on a probabilistic approach, e.g. for HMM, a statistical one for DNN, or on explicit formulation for SVM. However, their use depends on the final task. Therefore, through some examples, we explain how these methods can be used to fulfil the projected tasks.

6. 1 Image segmentation

Several techniques for interpreting images need segmentation. It entails segmenting pictures into several items, regions or patterns. There are several uses for image segmentation, such as medical image analysis, video surveillance, and augmented reality. In the specialized field of archaeology, this method has been used for item identification and geographic image analysis.

The task of segmenting an image may be approached in two ways: either by classifying pixels with semantic labels (semantic segmentation) or by splitting distinct objects (instance segmentation). While image classification predicts a single label for the whole picture, semantic segmentation does pixel-level labeling for each image pixel using a collection of item categories (e.g., human, car, tree, sky) for each pixel. As such, it is often a more difficult task. By identifying and separating each pertinent component in the image, instance segmentation broadens the scope of semantic segmentation, e.g., splitting separate items.[21]

---

[20] Ahmed 2015.
[21] Minae *et al*. 2021.

6. 2 Image classification

Finding similar attributes amongst images is the first step in categorizing them. These qualities, which are frequently referred to as descriptors or features, may be overt or covert. Classification may be accomplished by a variety of machine learning techniques, which fall into two main categories: supervised techniques (like DNN) and unsupervised techniques (like K-means). The latter are dependent on whether the original labels are available during the learning process. Since labels are unknown a priori in the latter scenario, the word clustering fits the data better than classification.

Typically, each visible feature is carefully extracted or selected to ensure that it matches a physical description of the image. However, additional statistical methods such as Principal Component Analysis (PCA) or even more sophisticated machine learning methods like Deep/Variational Autoencoders (DAE/VAE) are used to retrieve latent features. Latent features are believed to be more discriminating but less explicable than visible characteristics. Despite their lack of interpretability, deep features make the classification model more accurate.[22]

6.3 Character recognition

Even if printed works and ancient manuscripts differ, optical character recognition and handwriting recognition may be used to handle text extraction challenges in a similar fashion. Text recognition either extracts keywords from a line of text or creates a verbatim transcription of it after pre-processing and segmentation.

Whether the material was written or printed, the primary objective is to accurately translate the words in the document picture into digital text. The expected regularity of the space between letters and words serves as the foundation for optical character recognition's classification. The character is the fundamental unit of recognition. Because words and their constituent characters may be separated frequently and accurately, optical character recognition classifiers are able to recognize and produce a transcription using individual character glyphs. Consistent letter and word spacing, however, is insufficient for handwriting identification due to the quirks of human handwriting.[23]

6.4 Language modelling

Systems will often utilize a statistical language model to improve optical character recognition and handwritten text recognition accuracy if the document language is known ahead of time. A language model makes sure that the words that the computer recognizes are consistent with the grammar and even the known vocabulary of the language. For optical character identification and handwriting recognition, deep learning techniques based on neural networks as well as traditional machine learning techniques can be applied.[24]

---

[22] Coates, Ng 2011.
[23] Fischer 2011.
[24] Frinken 2013.

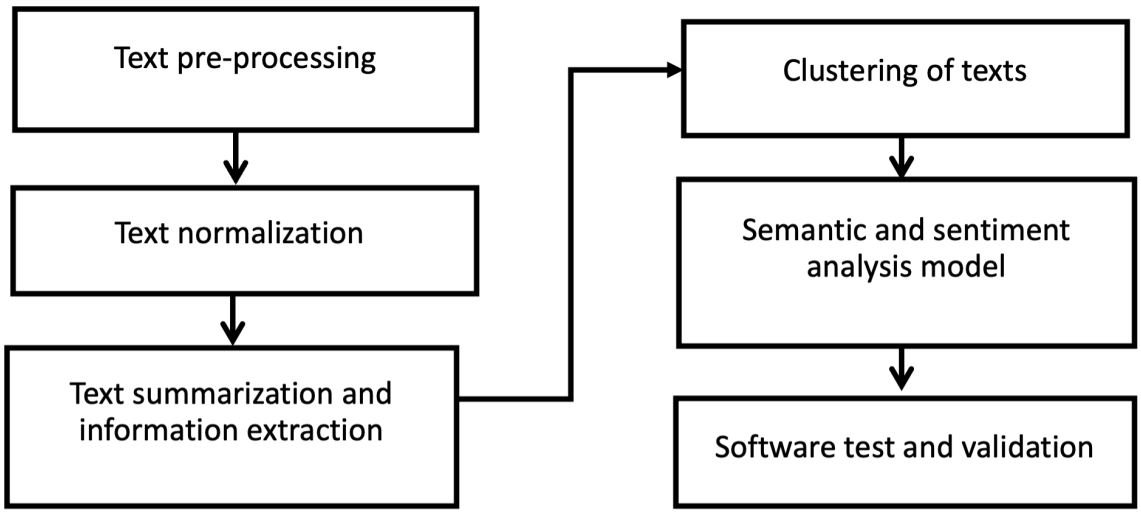

Fig. 1: Generic flowchart of automatic text processing techniques.

7. Conclusion and future directions

The ITSERR project, under the RESILIENCE European network infrastructure, endeavours to create a comprehensive digital archive for religious texts from the Nile Valley and beyond. This initiative faces several challenges, including the diversity of languages and writing forms, scarcity of material, and the absence of dedicated methods for analysing ancient religious texts. This paper serves as a bridge between the state-of-the-art linguistic text analysis and image processing, focusing on AI and metadata projects in the field of Egyptian language text analysis. By reviewing existing projects and methodologies, the paper aims to develop strategies for applying AI to the analysis and restoration of ancient religious texts.

This work outlines three main tracks for the ITSERR project, each addressing different aspects of text analysis and restoration. These tracks include image-based hieroglyphic character recognition, restoring ancient text using deep learning, and lexical and semantic analysis of ancient sacred texts using machine learning and NLP.

To achieve the goals outlined in the project tracks, various computational methods are proposed, including deep learning, machine learning, and unsupervised methods. These methods will be employed for tasks such as image segmentation, character recognition, language modelling, and semantic analysis.

The projected applications of these computational methods offer promising avenues for advancing the digitization and analysis of ancient religious texts. By leveraging AI and metadata standards, researchers aim to enhance the accessibility, interoperability, and reusability of these invaluable cultural artefacts for scholars and the broader community alike.

References

A. Argyrou, A. Agapiou, 2022. A Review of Artificial Intelligence and Remote Sensing for Archaeological Research, in *Remote Sensing 14 (23), 6000.* [https://doi.org/10.3390/rs14236000, accessed April, 2025].

N. Ahmed, 2015. Recent review on image clustering. *IET Image Processing* 9 (11), 1020-1032.

A. Coates, A.Y. Ng, 2011. The importance of encoding versus training with sparse coding and vector quantization, in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 921-928.

E. Demetrescu, E. d'Annibale, D. Ferdani, B. Fanini, 2020. Digital replica of cultural landscapes: An experimental reality-based workflow to create realistic, interactive open world experiences, *JCH 41*, 125-141.

B. Fanini, D. Ferdani, E. Demetrescu, S. Berto, E. d'Annibale, 2021. ATON: An Open-Source Framework for Creating Immersive, Collaborative and Liquid Web-Apps for Cultural Heritage, in *Applied Sciences* 11, 22. [https://doi.org/10.3390/app112211062, accessed April, 2025].

A.I. Fischer, 2011. HMM-based alignment of inaccurate transcriptions for historical documents, in *International Conference on Document Analysis and Recognition (ICDAR)*, 53-57. [doi: 10.1109/ICDAR.2011.20, accessed April, 2025].

V. Frinken, A. Fischer, C-D. Martínez-Hinarejos, 2013. Handwriting recognition in historical documents using very large vocabularies, in *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing*, 67-72. [doi: 10.1145/2501115.2501116, accessed April, 2025].

C. Gracia Zamacona, D. Pizarro Pérez, Á. Hernández Alonso, S. Palazuelos Cagigas, R. Nieto Capuchino, D. Fuentes Jiménez, *et. al.*, 2026. Digital projects on Earlier Egyptian mortuary texts at the University of Alcalá, in S. Mainieri, R. Pirelli (eds.), *Ancient Egypt – New Technology. Proceedings of the International Conference (2nd edition): 5-7 July 2023, University of Naples 'L'Orientale'*. Serie Egittologica VI, 103-112 Napoli.

S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, D. Terzopoulos, 2021. Image segmentation using deep learning: A survey. IEEE transactions on pattern analysis and machine intelligence 44.7, 3523-3542. [https://doi.org/10.48550/arXiv.2001.05566, accessed April, 2025].D. Nasien, H. Haron, S.S. Yuhaniz, 2010. Support Vector Machine (SVM) for English handwritten character recognition, in *2010 Second international conference on computer engineering and applications* 1, 249-252.

B. Schölkopf, R.C. Williamson, A. Smola, J. Shawe-Taylor, J. Platt, 1999. Support vector method for novelty detection, in S.A. Solla, T.K. Leen, K.-R. Müller (eds.) *Advances in neural information processing systems (NIPS) 12*, 582-588.

I. Zangar, Z. Mnasri, V. Colotte, D. Jouvet, 2021. Duration modelling and evaluation for Arabic statistical parametric speech synthesis, in *Multimed Tools Appl.* 80 (6), 8331–8353. [doi: 10.1007/s11042-020-09901-7, accessed April, 2025].

Online Resources

ArchAide Project	http://www.archaide.eu
ARIADNE	https://ariadne-infrastructure.eu/
Beta Maṣāḥəft Project	https://www.betamasaheft.uni-hamburg.de/
Cleo	https://cleo.aincient.org/pages/en/
DeepScribe Project	https://datascience.uchicago.edu/research/deciphering-cuneiform-with-artificial-intelligence/
Europeana	https://www.europeana.eu/

Ithaca Project	https://ithaca.deepmind.com
Kalam Project	https://site.unibo.it/kalam/en
Mortexvar	https://www.mortexvar.com/ocr-pt-ct
Open Context	https://opencontext.org/
PAThs Project	http://paths.uniroma1.it/
RePair Project	https://www.repairproject.eu
RESILIENCE	https://www.resilience-ri.eu/

Archaeological data and digital society

Andrea D'Andrea

Abstract

Speaking in 2006 at the annual conference of the Association of National Advertisers, British mathematician Clive Robert Humby coined the famous phrase "*Data is the new oil*". The next 17 years proved that Humby was right about the dominant role that data would play in business as well as in culture and research. However, unlike oil, which is a non-renewable source, data is not limited, it is renewable and reusable, and its quantity will continue to grow in the future. Like oil, data cannot be used in a raw way; it must be refined and transformed to have value. Artificial intelligence and machine learning are certainly very hungry for data that must be in some form structured according to a standard schema with defined types and relationships. Today's challenge lies in the ability to transform Big Data resources into valuable training models for machine-learning based applications. The increasing use of data contributes to a paradigm shift in science increasingly characterized by the navigation of an almost infinite sea of data, capable of generating new knowledge. Starting from this premise, the paper illustrates how in the world of archaeological research, data has progressively become a key aspect not only in the reconstruction of the past, but also in practices dedicated to sharing archives.

Keywords: *Data-sharing; Semantic Technologies; Big-Data; Artificial Intelligence; Repository.*

1. Introduction

Today's society is characterized by the deployment of a technology that influences every economic, scientific, and cultural relation. Therefore, it is necessary to reflect on the correct use of these innovations to avoid losing control of them. Recent discussions on the perspectives of artificial intelligence-based applications and their potential positive or negative effects on humanity confirm the widely shared need to discuss the value of technologies and their effective use. Therefore, rather than focusing on the efficiency of digital applications in archaeology, it is better to investigate the role that data play in historical research today, including highlighting the ambiguity in the use of terms related to the archaeological record (data, information, digital documents, information resources, etc.). The risk, which exists in the misuse of technology, often stems from the choice of overly rigid formal structures that nullify all possible uncertainty and variability to the advantage of consistency on the purely IT side.[1]

Many scholars pointed out that we live in a world of data captured by multiple sensors, instruments, devices, automatically or manually, consciously, or unconsciously. But what does the term data means? Data is a collection of discrete values describing quantity, quality, facts, basic units of meanings or simply a sequence of symbols that may be further interpreted. Data are stored into structures and used as variables in a computational process. Data, moreover, may represent abstract or practical concepts and they are commonly used in scientific research and, virtually, in every form of human organizational activity. It has been estimated that more than 2.5 quintillion bytes of data are being produced every day. We often use the term information as a synonym for data; on a linguistic level, the two terms

[1] Huggett 2020a.

are equivalent, but from a computational point of view, confusing data, and information, can generate errors difficult to be discovered when the structure of the digital document is inaccessible. But what data can be used for? In the wisdom hierarchy, data is at the lower level. Data are entities that carry information but are not information on its own, while information derived from data through the process of interpretation and analysis.

Data can be aggregated to increase their information content; there are corpora that include heterogeneous sources, datasets that organise coherent resources, and repositories that integrate digital data with metadata, providing a range of search, aggregation, and preservation operations. Finally, we have the infrastructures that, in addition to the requirements of repositories and aggregators, add services, tools and best practices for researchers.

Given the value that data are taking on in archaeological research, this paper aims to explore how digital data are produced, interpreted and, most importantly, shared. At this stage, characterized by the still experimental use of artificial intelligence, a process of knowledge standardization is making data *machine-understandable* and *machine-readable*.

2. Data from oil to soil and a fourth paradigm

The growing importance of data in the world economy prompted the British mathematician, Clive Robert Humby, to declare in 2006 that "*Data is the new oil*".[2] Speaking at the annual conference of the Association of National Advertisers, the scholar pointed out the dominant role that data would play in business as well as in culture and research. Data in the 21st century is like oil in the 18th century: those who understand the fundamental value of data, and learn how to extract and use it, will be able to best develop their business. In the digital economy, data is the key to growth, and without it, progress would stop. The following years proved that Humby was right.

All the statistics about the largest companies by market cap show how much the value of data industry has grown at the expense of the oil industry in the last years. Today the most important players are all Internet service providers, thus suppliers of digital data. Not only the players have ousted the oil producers in the ranking of most lucrative business, but they have also significantly increased their revenues over the oil industry. The digital knowledge industry has developed primarily with the refinement of data, as raw data is not usable on its own. Value enhancement is possible when data is collected, accumulated, and linked with other relevant and meaningful data.

Nevertheless, unlike oil, which is a non-renewable source, data is not limited, it is reusable, and its quantity will continue to grow in the future. Despite being a sustainable source, data must be protected and maintained. Not only does it contains patterns, indicators, and information, but it is also a renewable supply without limitation. You can repeatedly use the same data for different purposes and applications. Data not only needs to be refined but also needs to be cultivated and curated to be fertile. For this reason, other scholars highlighted that data is not oil but rather "a new soil". According to British journalist David McCandless,[3] data is a reusable resource for producing new forms of value that, in turn, feed new data in a virtuous cycle. McCandless sees data as fertile soil that can be cultivated and reused over time (unlike oil). Data is a valuable

[2] Palmer 2006.
[3] McCandless 2010.

resource that makes new products blossom and improves the user experience. Data takes root and multiplies in an evolving ecosystem for the benefit of tomorrow's users. Like soils that are not all equal, data to be transformed into fertile resources require the careful and competent work of scholars/farmers.

The increasing use of data is contributing to a paradigm shift in science, which is more and more characterized by navigating an almost endless sea of data generating new knowledge. Data scientists are using advanced statistical and machine learning methods to produce more detailed information for various users, including decision makers. Data organisation, and especially its standardization through the use of shared data-models, becomes the benchmark for data integration, access and reuse for advanced research. The goal of the new paradigm is to aggregate data to discover meaningful information for decision making. Therefore, predictive analysis is a technique used to estimate the probability of an outcome based on past data. Because data science models today are computer simulations, different possible scenarios can be easily tested, enabling stakeholders to make an informed decision. But data science is not a mere application of solving complex mathematical calculations. Data analysts must possess domain knowledge to evaluate specific conditions in datasets and their aggregation. Value extraction requires solutions that allow searching for data from various sources and types. These sources can include applications, databases, web sites, and cloud environments and can come in all shapes and sizes, structured, semi-structured, and unstructured. Therefore, data must be standardized using coding models acknowledged by the relevant community before being processed and analysed.

The world of science has changed. The new approach involves data being acquired from instruments or generated by simulations before being processed by any software. Only then is the information or knowledge extracted by the data stored in computers. Scientists can only examine their data at a late stage in this process. The techniques and technologies for this data-intensive science are so diverse that it is worth distinguishing data-intensive science from computational science as a new, fourth paradigm for scientific exploration.[4] This scenario is possible if we use appropriate mathematical methods and if data are encoded in a homogeneous way. People now no longer look through telescopes. They look through complex, large-scale instruments that transmit data to data centres, and only then researchers look at the information on their computers. This new approach has pushed toward open data and more collaborative and transparent science. If data is the fuel for any professional, economic, or research activity, then analytical and aggregation techniques are the tools for extracting new knowledge.[5]

Starting from these premises, 10 years after Humby, scientist Andrew Ng[6] has uttered a new phrase that complements the previous one: "if data is oil, then artificial intelligence is the new electricity". Just as electricity has changed the world by disrupting transport, manufacturing, agriculture, and healthcare, artificial Intelligence will have a similar impact, even if it is not running as fast as we would expect for two reasons: the absence or scarcity of good data and the lack of specific skills capable of adapting the application of algorithms to contexts. Artificial intelligence is certainly very hungry for data, which

[4] Hey, Tansle, Tolle 2009.
[5] Huggett 2020b.
[6] Ng 2017.

should be structured according to a standard knowledge representation scheme to be properly digested. Today's challenge lies in the ability to transform Big Data resources into valuable training models for machine-learning based applications.[7]

If data is the soil where products grow, Big Data should represent the harvest, while artificial intelligence should be the engine that reinforces soil fertility. Scholars-farmers should know when to sow, how to eliminate weeds, when to water, how to harvest the product and how to transport it, still fresh, to the scientific market.

Data collection and description are central to having good models and good harvest. The first issue that one encounters in this process is that very often data and information are confused with serious effects for the process of machine-learning. If, for example, a scholar includes dating and/or function or interpretation in data description, the machine will take these parameters into account when analyzing other similar data, thus influencing the result. For these reasons, the researchers need to spend about 50% of their research time to curate and prepare data before their use.[8]

Data and information are not interchangeable terms:[9] data are individual entities, while information represents the structure and interpretation of those data. In other words, data are the building blocks of the house, while information represents the organisation of spaces. Data may come in the form of text, observations, figures, pictures, numbers, graphs, or symbols. Data is a raw form of knowledge and has no meaning or purpose. In other words, you must interpret the data to give it a meaning. Data can be simple and may even seem useless until it is analysed, organized, and interpreted. There are two main types of data: quantitative, provided in numerical form, such as weight, volume, or dimensions; and qualitative. Information, on the other hand, is the result of analysing and interpreting pieces of data. But what is the situation in archaeological research? Data has progressively become a key aspect not only in the reconstruction of the past, but also in practices dedicated to sharing archives.

The growth of user-friendly computer tools, as well as the increased availability of digital data and archives, is driving archaeologists to a greater understanding of digital techniques and methods also for the purpose of promoting the results of their own research. This has prompted scholars to think about promoting a new discipline called Digital Archaeology,[10] the development of which can be traced back to the 1950s and 1970s.

Digital archaeology encompasses numerous fields of activity with related application areas that highlight some specific cross-cutting issues:

- The digital representation of the archaeological record.
- The normalized descriptions.
- The formalization of reasoning.
- The implementation of procedures for the validation of the archaeological discourse.

The extensive digitisation of data and of archaeological techniques themselves, such as surveying, photography and data processing is highlighting several critical areas. Despite the changes that computer science has brought to archaeological methodologies, there are some areas where a closer look at the relationship between archaeology and computer science is

[7] The expression Big Data primarily identifies a data set that is too large or complex to be handled by traditional data processing software.
[8] Marsolek *et al*. 2023.
[9] Zins 2007.
[10] Evans, Daly 2006.

needed. I will list just a few of them: the ambiguity and bias of data that make the integration of archives difficult; the ways of long-term preservation of legacy data and their reuse; and, finally, the development of 3D technologies aimed at creating BIM platforms according to the emergence of the so-called digital twin. Therefore, archaeologists must place, at the heart of the debate, the issue of the correct formalisation and digitisation of data by reflecting on the nature of the archaeological record, which today includes old paper and digital data.[11]

3. Data-sharing

The first area of criticism and reflection must concern the data-sharing and the databases, the most basic, but also the oldest, form of data digitisation.

The figure 1 shows the record of a possible digital archive concerning the Professors at the University "L'Orientale". The structure, typical of a database, has columns with attributes and rows containing values. It is a simple record, with some values such as a name (Andrea D'Andrea), an academic affiliation (the University "L'Orientale"), a location (Naples) and two disciplines (Digital Humanities and Archaeology). As the structure is flat and the knowledge is implicit, the lack of cross-references could create potential ambiguity; we do not know the relationship between Andrea and the University of Naples, just as we do not know the relationship with the city or the various disciplines. The author of the database can understand and process data, but a computer is not able to solve the ambiguity of the statements. Therefore, this dataset can be linked only to other data with a similar structure. To avoid implementing a database that meets all possible needs of every archaeologist, we can use the semantic technologies to reconstruct the representation of the record by eliminating ambiguities and inserting links and relations to overcome the uncertainty.

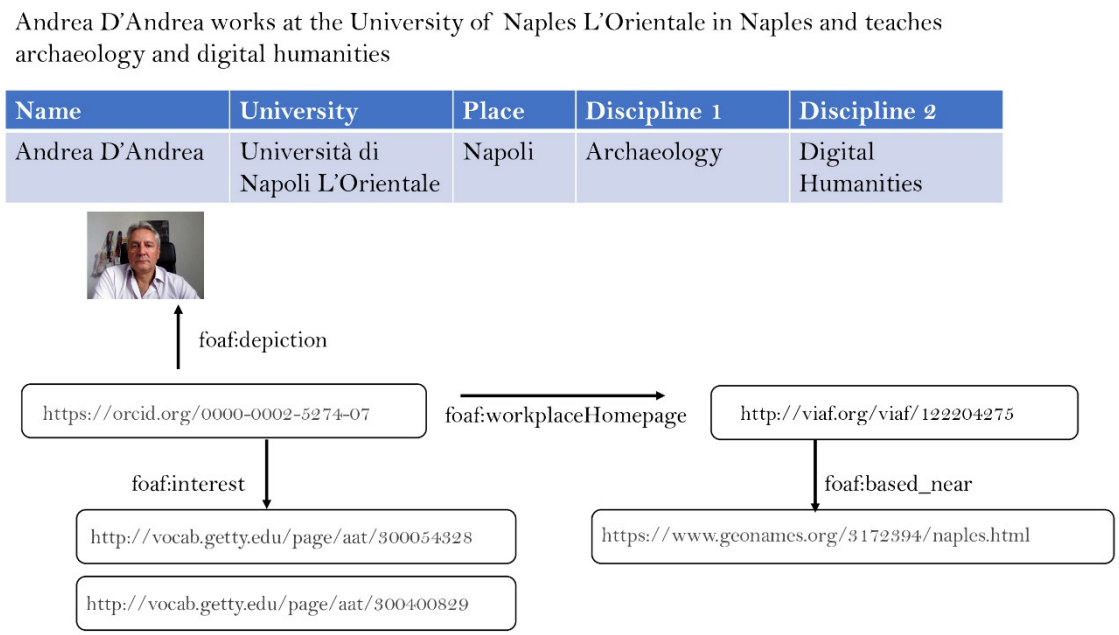

Fig. 1: On the top a simple text; in the middle the formalization of the text into a database; on the bottom the graphical schema of the semantic description of the text.

[11] Castelli, Felicetti, Proietti 2021; Huggett 2020a, 2022; Kansa *et al*. 2019.

To make my statement shareable and comprehensible by a machine, I must use a language based on a standard encoding with unambiguous terminology according to LOD (Linked Open-Data) principles.[12] In the example my name is replaced by my ID of ORCID[13] which deleted issues with other Andrea D'Andrea researchers. Instead of the University of Naples "L'Orientale" I entered the URI (Uniform Resource Identifiers)[14] of the public register VIAF;[15] if the name of the university changes, I do not need to modify the record. For the location I use the list of places published by GeoNames,[16] thus avoiding writing the city in English or in another language. Finally, to define the disciplines taught, I use the Getty AAT thesaurus.[17] By adding some properties as *depiction*, *workplaceHomepage*, *interest*, *based_near* between the statements, I can enrich the description and make more explicit the record. By using a machine-readable ontology, codified in RDF (Resource Description Framework),[18] it is possible to construct any kind of statement in form of triples that a computer can understand. If we use URIs instead of values, multiple connections and relationships can be potentially discovered. A truly open and collaborative science requires the use of metadata and thesauri to make data homogeneous and to enable the integration of archives.

To support the creation of good data for open science, some researchers developed in 2016 the FAIR (Findable, Accessible, Interoperable, Reusable) principles.[19] These recommendations, mandatory for those receiving European funding for data-driven projects, should encourage scholars to open their data to achieve four goals: making data findable, accessible, interoperable, and reusable. FAIR Data simplifies the integration of repositories and ensures the integrity of the digital data and the long-term preservation. The rules of the FAIR principles enable data protection and, more importantly, can support a data-intensive approach by recognising a clear license in case of reuse by third parties.

Metadata schema, Open-Data, Linked Open-Data, Ontologies, and FAIR principles are the conceptual tools of the Semantic Web[20] that make discovery open, complete, and easily reproducible. Accelerating the change to a more open and shared space requires the adoption of behaviours compatible with the Open Science scenario.[21] If the semantic technology facilitates the use of transparent, collaborative, and integrable workflows, the repositories, and the research infrastructures host systems able to map specific metadata or data models or with a mapping process between a local metadata system and a standard one.

---

[12] Linked Open Data (LOD) is a way of publishing structured data that, through the use of appropriate technologies and open web standards, allows resources to be linked together. LODs are readable by machines that can directly interpret the information on the web.
[13] https://orcid.org
[14] A URI is a unique sequence of characters that identifies an abstract or physical resource.
[15] The Virtual International Authority File, https://viaf.org
[16] https://geonames.org
[17] Getty AAT (Art & Architecture Thesaurus) https://www.getty.edu/research/tools/vocabularies/aat
[18] RDF is the basic tool proposed by the World Wide Web Consortium (W3C) for encoding, exchanging and reusing metadata by enabling semantic interoperability between applications sharing information on the Web.
[19] Wilkinson *et al.* 2016, 2019; Hermon, Niccolucci 2021.
[20] The Semantic Web is an extension of the World Wide Web aimed at making data, found on the Web, machine-readable.
[21] Huggett 2015.

Repositories are the first level of the Open Science perspective. They store a significant amount of information and perform several operations, including protecting, classifying, processing, searching, and duplicating documents. Resource management is centralized and implemented in an environment accessible from multiple hardware machines. Repositories ensure proper management of document flow through metadata and template-data that are ccompliant with international standards. Moreover, they facilitate classification and search mechanisms for data retrieval and visualization. OpenContext,[22] tDAR,[23] and ADS[24] offer the best solutions to implement a simple archaeological repository in a secure cloud that deploy metadata compliant with FAIR principles. An aggregation of different archives is Ariadne,[25] a European research infrastructure aimed at aggregating archaeological data through a schema able to integrate digital data stored in different repositories. Ariadne manages more than 3.000.000 metadata and offers some services as mapping tool, guidelines, and recommendation for making data open and FAIR, and a platform to upload 3D data or images.

4. Legacy Data

A second critical area of so-called legacy data is that of archive management and maintenance.[26] Unlike the issues addressed in the previous section, which seem to have been solved by the evolution of semantic technologies, the area of legacy data involves the development of a largely manual approach. The term legacy data refers to a wide range of digital documents that cannot be (re)utilised without substantial migration to formal structures that have been updated to the latest technological and information processing innovations. A non-exhaustive list of legacy data includes:

- Records acquired in a predigital era or with early devices and stored in embryonic database forms.
- Records, often inaccessible, not usable in more modern computer applications.
- Records produced by Digital Archaeology and soon to become obsolete due to the rapid speed in the evolution of digitisation processes.
- Records belonging to abandoned research approaches.
- Records to be reused with substantial restructuring of data representation forms.

In short, these are digital resources, stored somewhere in our computers, but which need to be revised to be reused. There is a wide variety of legacy data, but the most widespread are the outcomes of the massive digitisation that took place in the 1990s in archaeology without any standards or long-term preservation perspective.

Archaeologists frequently come across digital data or archives created at the dawn of digital archaeology. One of the most complex aspects to deal with concerns the transfer of old archives into new digital formats and structures without changing or forcing their content. If one wants to introduce old excavation data into a database that records new stratigraphic information, it is not correct to adapt the previous data to the new excavation and documentation methodologies, but,

[22] https://opencontext.org
[23] The Digital Archaeological Record https://core.tdar.org
[24] Archaeology Data Service https://archaeologydataservice.ac.uk
[25] https://www.ariadne-research-infrastructure.eu
[26] Allison 2008.

on the contrary, to safeguard the different approaches used, which must remain distinct. This solution may not allow research or comparisons to be made on all the data acquired, but forcing the nature of the legacy data, to align all records, could lead to an alteration of the original data and a falsification of the results.

In a recent revision of the documentation of the investigation of the medieval city of al-Balid, ancient Zafar, in Oman, it was necessary to migrate the old CAD archive, created in the 1990s, to a GIS application.[27] Although CAD has long been the most widely used drawing programme to produce plans and sections, even in archaeology, the need to integrate a multiplicity of georeferenced spatial information sources has driven many scholars towards the adoption of GIS systems to ensure better management of the data associated with geometries. Despite the common basis of geometric primitives to represent any spatial information, the conversion of CAD data into GIS is not exclusively an automatic process.

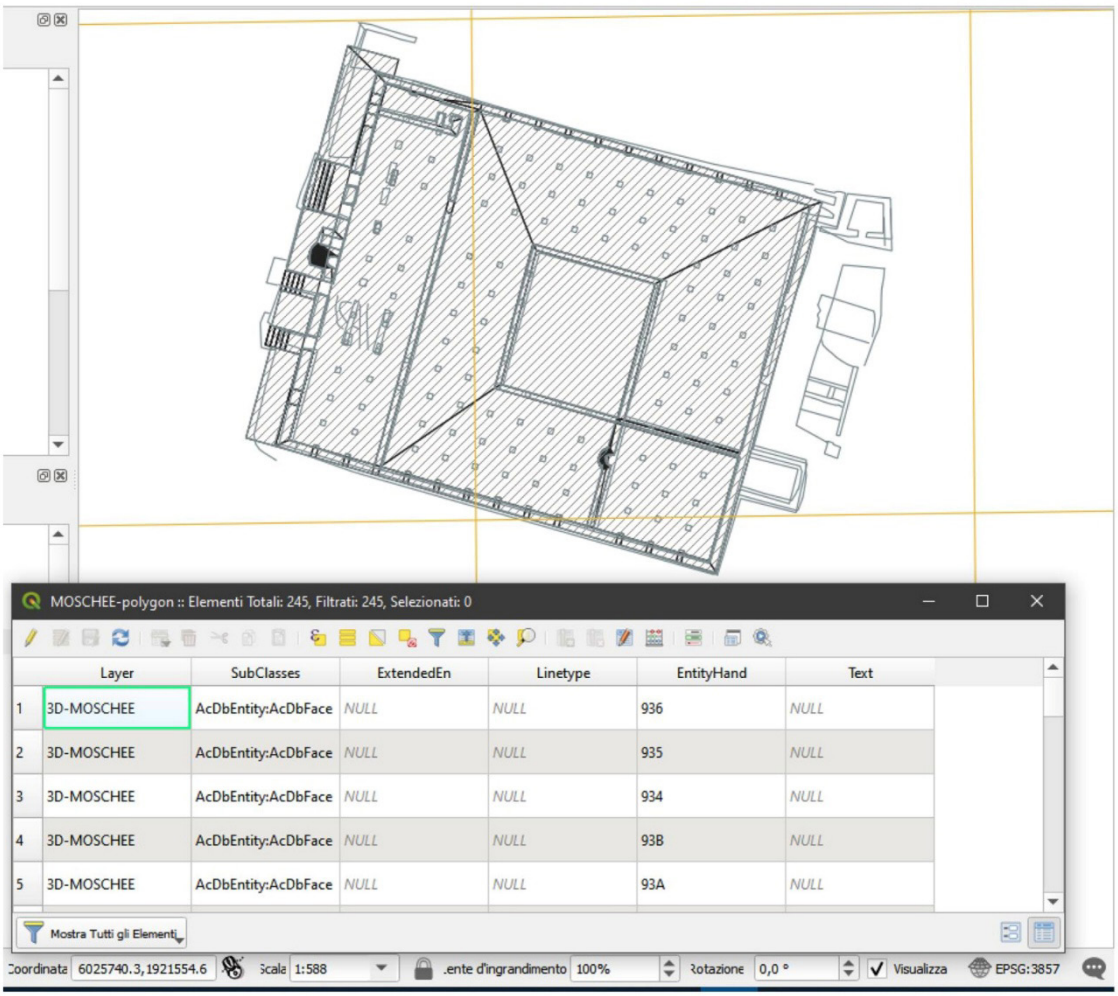

Fig. 2: On the top the plan of the Great Mosque visualized into QGIS. On the bottom the table associated to the 245 geometrical entities imported into QGIS from the original CAD file.

[27] D'Andrea 2021.

Figure 2 shows the result of importing a CAD file of the Great Mosque into QGIS. While the drawing appears geometrically correct, the various structural and architectural parts of the building are not defined in layers or semantic objects, but rather in individual entities, thus making the automatic import of vector objects impossible.[28] Probably the simplest and most correct way would be to re-draw the spatial information, but this would raise a problem with the re-arrangement of the imported geometric data.

When archaeologists identify a layer, they first define it in space by drawing it in plan. Often, however, due to the inconsistency and reworking of deposits, it is not possible to precisely indicate the boundaries of the layer; in such circumstances, the drawing shows dotted lines that suggest the uncertainty of the boundaries. From the planimetry, it is therefore not always possible to reconstruct the boundaries of the layers without having a text that clarifies their extent and spatial relationships with other stratigraphic units. Since GIS requires the association of a geometric datum with an attribute, without manual intervention it is not possible to import a CAD drawing and automatically create archaeological objects. Furthermore, only through the description of the relationships between the various deposits is it possible to reconstruct the topology. Conversion is therefore not based on the development of automatic procedures, but on patient work that consists of following the transition from graphic and geometric entities to a semantic digital system using the available archaeological documentation, even if it is largely incomplete and often lacking the necessary stratigraphic data. In short, GIS does not replace CAD. After developing an appropriate strategy for the migration of legacy data, good metadata must be associated with the spatial data to enable preservation and reuse of the spatial archive.[29]

5. 3D Replica

The third and final critical area is that of 3D replicas. The topic does not concern acquisition and restitution tools and methodologies. The 3D technology industry has made great strides in recent years and today we have low-cost tools that allow the acquisition of complex architectural structures in a short time and with great accuracy.[30] Many protocols have been developed to produce geometrically correct digital replicas and there are various equipment and methods to obtain good 3D reproductions. On the geometric information processing side, studies on semantic segmentation are opening new frontiers for the automatic management of large amounts of 3D data. Unlike semantic data and legacy data, which remain areas of study, 3D is now an integral part of professional and academic archaeology.

[28] Bibby, Ducke 2017.
[29] Chapman 2001.
[30] Bosco 2022.

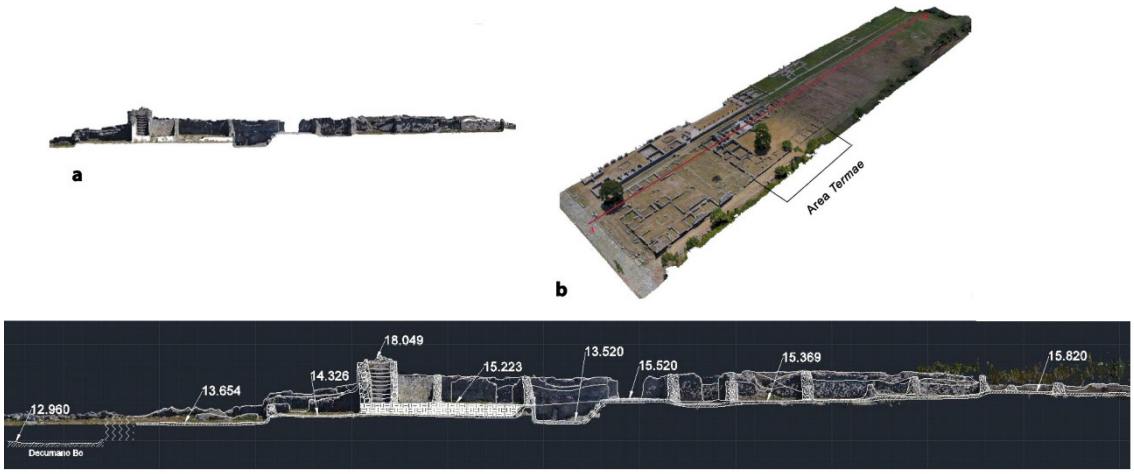

Fig. 3: The 3D replica of the Insula 4-6 of Paestum.

The introduction of BIM (Building Information Modelling) also in archaeology has made it possible, the entire 3D workflow has been reorganised and given a specific structure that goes beyond the production of simple and beautiful images or interactive models. BIM allows traditional graphic documents such as sections, elevations and plans to be easily extracted from 3D. In Figure 3 one can see the result of a recent archaeological investigation in a block of Paestum.[31] The model was generated by integrating different 3D spatial technologies, such as aerial and terrestrial photogrammetry and laser scanning. Georeferencing was performed using points measured with GPS. BIM greatly facilitated the work of rendering and analysing geometric information, but, in my opinion, this is not what makes the software or the BIM approach truly innovative.

The most important feature of BIM is the creation of libraries that can be easily shared. Analytical descriptions of walls, materials, binders, and construction techniques can be used to identify certain parametric values characteristic of a specific masonry work. This data can, in turn, be reused to reconstruct missing architectural and construction components of other buildings, even those belonging to other archaeological contexts. The library can be encoded in a standard language and format and imported into other projects, constituting a special form of shared knowledge that can be used by different specialists.

BIM is certainly a second-hand technology, as the software was originally developed for civil engineering. However, today it is also widely used in archaeological research for the study and understanding of architectural techniques and historical buildings. Unlike three-dimensional reconstructions, created for scientific or popular reasons, in which the choices made by the scholar are invisible to the user, BIM records every logical-functional step of the monument reconstruction process, from the import of the 3D replica to the modelling and definition of attributes, allowing any formal verification of the process.

[31] Bosco *et al.* 2020.

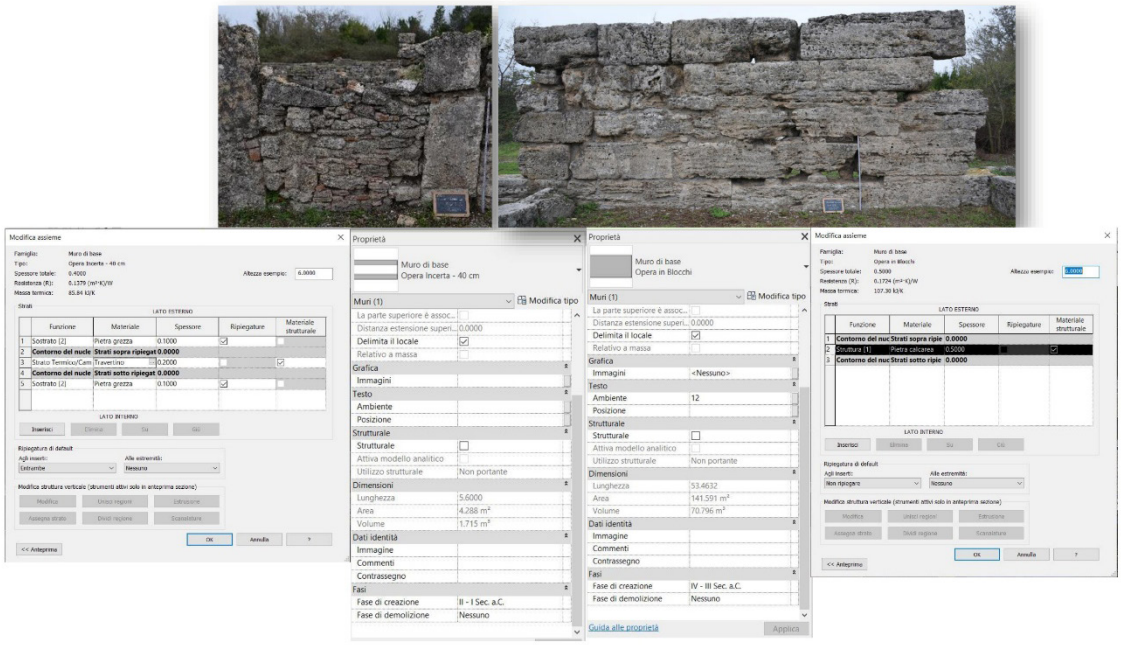

Fig. 4: Different construction techniques described in BIM.

Figure 4 shows the description of two masonry techniques encoded in a structure very similar to a database connected to geometric data. The libraries function like Lego bricks that can be used to reconstruct other contemporary architectures and structures. The possibility of sharing libraries makes BIM a tool of great impact on the study of ancient architecture that shows a minimum of standard structuring to a certain extent that can be parameterised. In another example, developed for the solar temple of Niuserra,[32] we demonstrated the strength of an approach that makes the reconstruction process transparent based on assumptions that can always be verified.

The archaeological object, decomposed and organised into different semantic levels, can be easily interrogated. The digital replica becomes the basis for the formulation of alternative reconstructive hypotheses and for simulations aimed at ensuring the preservation of the monument. This interaction, guaranteed by BIM between the physical and virtual world, is one of the distinguishing features of the Digital Twin, which is not only a geometrically correct replica of an object, but also a metaphor for a digital sphere in which the researcher can perform different types of tasks.

The Digital Twin is a computational representation of an abstract or real physical product, system or process that acts as a replica for practical purposes such as simulation, integration, monitoring, and maintenance. The digital replica enables continuous feedback between real and virtual that enriches the knowledge of the monument and improves its preservation. The Digital Twin is not simply a digital and static replica, perhaps with more detail, but a digital document that can be viewed and interrogated dynamically from multiple disciplinary perspectives. The virtual model can become the

[32] Bosco *et al.* 2019.

pivot point of a possible metaverse that starts from a digital replica and is enriched through content on the web.

In a recent project aimed at digitising the Catacombs of San Gennaro in Naples,[33] the final product was uploaded onto the Sketchfab[34] platform and enriched with annotations and links that refer to further descriptions or details of the monument acquired with different techniques. The currently available network of relations and references was implemented manually; in the future these connections may be generated automatically. In this perspective, it will be necessary to have good models and 3D replicas, but also good metadata associated with digital objects published on the web. The metadata schemes currently available in the repositories mentioned above are not able to correctly describe the 3D digitisation protocol.

A solution to this critical issue was provided by the European project 3D-ICONS,[35] which developed a specific metadata schema based on digital provenance and paradata. Provenance describes the technical process of digitising the replica, i.e. the tools used to capture and process the data, while paradata should be used to indicate the reasons why the replica was created. Paradata can also contain information that makes the process of interpreting a monument transparent by clarifying the assumptions followed by the author in the reconstruction. The metadata developed in the 3D-ICONS project can capture the semantics related to the physical object and the methods and tools used to achieve the digital replica. The scheme, designed to provide 3D replicas of archaeological monuments to the Europeana library,[36] reuses part of the CIDOC-CRM ontology[37] and its extension CRMdig.[38] Furthermore, where possible, for some fields the values are expressed as URIs to facilitate possible further connections.

6. Conclusions

The paper examined the importance of data not only for scientific research but also for society more generally. Next, some critical areas in archaeological data management that would require more attention from scholars were highlighted and described. First, the knowledge and databases are not shared, thus making it less easy to create added value for interdisciplinary research. In addition, the excess of data virtually available on the web complicates the work of analysis and information retrieval. Finally, considering the increasingly complex tasks of researchers, the future challenge will be to create datasets that can be queried and aggregated by automated agents. Researchers must, therefore, be aware that not only physical assets must be protected, but also digital archives that risk being underutilized or, at worst, dispersed. While technology enables innovative strategies for long-term preservation, the lack of standardization often makes it difficult to extract data from interconnected archives. The challenge in the coming years will be increasingly related to the need to produce quality digital data associated with metadata that also records the processes of digitisation of sources.

---

[33] Bosco, Minucci 2020.
[34] https://sketchfab.com/GlobalDigitalHeritage/collections/catacombs-of-san-gennaro-italy
[35] http://3dicons-project.eu/
[36] https://www.europeana.eu/it
[37] CIDOC Conceptual Reference Model (CRM) https://www.cidoc-crm.org
[38] CRM Digital https://www.cidoc-crm.org/crmdig

Archaeology as a science has developed in parallel with the transformations of a society that today appears increasingly immersed in the digital domain. If scholars must invest part of their working time to the curation of data, it is necessary at the same time to ensure the transparency and reproducibility of digital processes thus increasing the credibility of archaeological research. The quality and integrity of archives play a central role at the data reuse stage and in the independent evaluation of final research results. The sharing and accessibility of archives should encourage researchers to create quality data, while the adoption of open science principles and practices should promote inclusiveness by removing those financial, institutional, and cultural barriers that prevent researchers from freely exchanging knowledge, methods, and data.

The future appears to be shaped by the development of a 5.0 society, centred on the balance between economic progress and problem solving within a space that is not only physical but also cybernetic. Digital ecosystems will concretely support this transformation by connecting people, things, data, and knowledge through the spread of semantic technologies, artificial intelligence algorithms and Digital Twins. Technological changes in the coming years will progressively free scholars from very demanding manual tasks by innovating the methodologies of archaeological research and outlining new perspectives for the study and reconstruction of the ancient world.

References

P. Allison, 2008. Dealing with Legacy Data - An introduction, in *Internet Archaeology* 24. [https://doi.org/10.11141/ia.24.8, accessed April, 2025].

D. Bibby, B. Ducke B., 2017. Free and Open-Source Software Development in Archaeology. Two interrelated case studies: gvSIG CE and Survey2GIS, in *Internet Archaeology*, 43. [https://doi.org/10.11141/ia.43.3, accessed April, 2025].

A. Bosco, 2022. *3D Surveying Methods and Digital Information Management for Archaeological Heritage*. BAR International Series 3091, Oxford. [https://doi.org/10.30861/9781407359731, accessed April, 2025].

A. Bosco, E. Minucci, 2020. Rendering RTI ed editing d'immagine per elaborazioni SFM: confronto tra tecniche e strumenti di visualizzazione per la documentazione di graffiti in contesto archeologico. *Newsletter di Archeologia CISA* 11, 43-66.

A. Bosco, L. Carpentiero, A. D'Andrea, E. Minucci, R. Valentini, 2022. A parametric model to manage archaeological data, in *IMEKO TC-4 International Conference on Metrology for Archaeology and Cultural Heritage* Trento, Italy, October 22-24, 2020, 220-225. [https://www.imeko.org/publications/tc4-Archaeo-2020/IMEKO-TC4-MetroArchaeo2020-042.pdf, accessed April, 2025].

A. Bosco, A. D'Andrea, M. Nuzzolo, P. Zanfagna, 2019. A Bim Approach for the Analysis of an Archaeological Monument, in *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences.* XLII-2/W9, 165–172. [https://isprs-archives.copernicus.org/articles/XLII-2-W9/165/2019/, accessed April, 2025].

L. Castelli, A. Felicetti, F. Proietti, 2021. Heritage Science Heritage Science and Cultural Heritage: standards and tools for establishing cross-domain data interoperability, in *International Journal on Digital Libraries* 22 (3), 279-287. [https://doi.org/10.1007/s00799-019-00275-2, accessed April, 2025].

H. Chapman, 2001. Understanding and Using Archaeological Topographic Surveys – The 'Error Conspiracy', in Z. Stančič, T. Veljanovski T (eds.), *Computing Archaeology for Understanding*

the Past, CAA 2000, Computer Applications and Quantitative Methods in Archaeology, BAR International Series 931, 19–24, Oxford.

A. D'Andrea, 2021. Reconsidering the topography of al-Balid: a preliminary review of the graphical documentation. *AIOO* 81, 39-50. [https://doi.org/10.1163/24685631-12340110, accessed April, 2025].

T.L. Evans, P.T. Daly, 2006. *Digital Archaeology: Bridging Method and Theory*, London and New York.

T. Hey, S. Tansley, K. Tolle (eds.) 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Washington.

S. Hermon, F. Niccolucci, 2021. FAIR Data and Cultural Heritage Special Issue Editorial Note, in *International Journal on Digital Libraries* 22, 251–255. [https://doi.org/10.1007/s00799-021-00309-8, accessed April, 2025].

J. Huggett, 2015. 2 Digital Haystacks: Open Data and the Transformation of Archaeological Knowledge, in A. T. Wilson, B. Edwards (eds.), *Open-Source Archaeology*, 6-29. Warsaw. [https://doi.org/10.1515/9783110440171-003, accessed April, 2025].

J. Huggett, 2020a. Capturing the Silence in Digital Archaeology Knowledge, in *Information* 11(5), 278. [https://doi.org/10.3390/info11050278, accessed April, 2025].

J. Huggett, 2020b. Is Big Digital Data Different? Towards a New Archaeological Paradigm. *JFA* 45 (sup.1), S1, S8-S17. [https://doi.org/10.1080/00934690.2020.1713281, accessed April, 2025].

J. Huggett, 2022, Data Legacies, Epistemic Anxieties, and Digital Imaginaries in Archaeology, in *Digital* 2(2), 267-295. [https://doi.org/10.3390/digital2020016, accessed April, 2025].

W.S. Kansa, A. Levent, E.C. Kansa, R.H. Meadow, 2019. Archaeological Analysis in the Information Age: Guidelines for maximizing the Reach, Comprehensiveness, and Longevity of Data, in *Advances in Archaeological Practice* 8(1), 1-13. [https://doi.org/:10.1017/aap.2019.36, accessed April, 2025].

W. Marsolek, S.J. Wright, H. Luong, S.M. Braxton, J. Carlson, S. Lafferty-Hess, 2023. Understanding the value of curation: A survey of researcher perspectives of data curation services from six US institutions, in *PLoS One* 18(11): e0293534. [https://doi.org/10.1371/journal.pone.0293534, accessed April, 2025].

D. McCandless, 2010. *The beauty of data visualization*. Video July 2010 [https://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization, accessed April, 2025].

Ng Andrew, 2017. *Artificial Intelligence and the New Electricity*. Video January 2017 [https://www.youtube.com/watch?v=21EiKfQYZXc, accessed April, 2025].

M. Palmer, 2006. *Data is the New Oil*, Blog November 2006 [https://ana.blogs.com/maestros/2006/11/data_is_the_new.html, accessed April, 2025].

M.D. Wilkinson, M. Dumontier, I. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg *et al.*, 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship, in *Scientific Data* 3, 160018. [https://www.nature.com/articles/sdata201618, accessed April, 2025].

M.D. Wilkinson, M. Dumontier, I. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg *et al.* 2019. Addendum: FAIR Guiding Principles for Scientific Data Management and Stewardship, in *Scientific Data* 6, 6. [https://doi.org/10.1038/s41597-019-0009-6, accessed April, 2025].

C. Zins, 2007. Conceptual Approaches for defining Data, Information, and Knowledge, in *Journal of the American Society for Information Science and Technology* 58(4), 479–493. [https://doi.org/10.1002/asi.20508, accessed April, 2025].

Online Resources

ADS	https://archaeologydataservice.ac.uk
ARIADNE	https://www.ariadne-research-infrastructure.eu
CIDOC	https://www.cidoc-crm.org
CRM Digital	https://www.cidoc-crm.org/crmdig
EUROPEANA	https://www.europeana.eu/it
GeoNames	https://geonames.org
Getty AAT	https://www.getty.edu/research/tools/vocabularies/aat
OpenContext	https://opencontext.org
ORCID	https://orcid.org
Sketchfab	https://sketchfab.com/GlobalDigitalHeritage/collections/catacombs-of-san-gennaro-italy
tDAR	https://core.tdar.org
VIAF	https://viaf.org
3D-ICONS	http://3dicons-project.eu/