ARCHEO.FOSS XIV 2020

Open software, hardware, processes, data and formats in archaeological research

Proceedings of the 14th International Conference
15-17 October 2020

edited by

Julian Bogdani, Riccardo Montalbano,
and Paolo Rosati

ArcheoFOSS 14 | 2020

edited by

Julian Bogdani,
Riccardo Montalbano,
and Paolo Rosati

Image cover: Theatrum Pompei and surrounding areas, from SITAR - Sistema Informativo Territoriale Archeologico di Roma, Soprintendenza Speciale Archeologia Belle Arti e Paesaggio di Roma (Ministero della Cultura)

All contributions of this volume have undergone a double peer-review process.

This book is available in print and as a free download from www.archaeopress.com

Contents

Use, Application and Development of Free/Libre and Open-Source (FLOS) Tools in Archaeology

Creation, use and Promotion of Open Data and Open Formats in Archaeology

Appendix

Analysis and comparison of open and non-open spatial formats for archaeological research

Andrea D'Andrea, Francesca Forte

Andrea D'Andrea, University of Naples 'L'Orientale', dandrea@unior.it
Francesca Forte, University of Naples 'L'Orientale', fforte@unior.it

Abstract

This contribution attempts an overview on how to manage the spatial information of archaeological data available online in a non-open format and how they can efficiently be reused in different applications to extend the domains of knowledge: a key-role in this process is played by open standards, of which a review of comparison with a non-open format is provided in this study. A practical test has been carried out to demonstrate the benefits of using open formats for the scientific research and dissemination and to concretely verify the advantages of using formats, such as GeoJSON.

Keywords: OPEN SCIENCE; BIG GEOSPATIAL DATA; SPATIAL INFRASTRUCTURE; GEOJSON; OGC.

Introduction

This paper provides a short analysis of the current management and sharing of spatial information in the archaeological field according to the principles of Open Science (OS) and Big Data (BD). Up to now the topic has been treated mainly from the point of view of the implementation of digital infrastructures with scarce attention to standard and format. This issue has already been partially analysed in a previous contribution on comparisons of the most used spatial data in archaeology (Forte 2019: 95–111).

Our attention here focuses on the concept of Big Geospatial Data (BGD)[1] and its relationship with archaeological research. The paper deals with the geo-locational dataset available online and describes a test for re-using accessible archives in a non-open format. Furthermore, it shows how it is possible to integrate archaeological datasets through open systems and web applications, eventually to address the need to manage the growing size of archaeological geospatial datasets. The paper does not provide a solution from among the large availability of open and not-open formats; it offers some suggestions on how to choose the spatial data online to be integrated in the spatial analysis.

---

[1] BGD are all types of data objects or elements with geographical information unrelated to a specific data model (raster, vector), to a software, or format. Due to volume variety or velocity, BGD overcome the capacity of commonly used spatial systems.

Open Science and Big Geospatial Data

The term 'Open Science' identifies 'an approach to the scientific process that focuses on spreading knowledge as soon as it is available using digital and collaborative technology'.[2] One of the goals of research and innovation policy is to share knowledge and data as early as possible in the research process, making publications available in open-access and providing spatial datasets as open as possible; only in this way the scientific research improves its quality, efficiency and responsiveness.

Over the past few years, the growing diffusion of satellite networks and the emerging of the Internet Of Things (IoT) have allowed users to track and correlate the accurate positions of people and objects; furthermore, thanks to the spread of technologies that are becoming more and more accessible, a large amount of Geospatial Data has been produced.[3] Within the flow of data that is collected daily, Geospatial Data occupy a strategic role in the perspective of BD. Therefore, it is possible to use the term 'Big Geospatial Data' to define datasets that include locational information and exceed the capacity of widely available hardware, software, and human resources and require specialized effort to work with: without an appropriate methodology to manage them, the strong potentiality of BD dramatically decreases.

Unfortunately, most of the geospatial data are kept in personal repositories, formalised according to international standards, restricting the dissemination and the enrichment of the scientific knowledge. The adoption of FAIR (Findable, Accessible, Interoperable and Re-usable data) principles[4] and open data should strengthen the diffusion of the results of the scientific research.

The first steps in this direction were made by OGC (Open Geospatial Consortium)[5] in 2017 with the publication of a White Paper[6] aimed at defining more efficient approaches for the treatment of BGD. The goal of the paper was to understand what actions the OGC should undertake for the correct management of spatial data and the improvement of the capabilities of BGD. The White Paper is a survey including open standards needed to guarantee interoperability, efficiency, innovation, and cost effectiveness. No specific recommendation is given, but the paper focuses on the need to adopt an analytical process to better exploit the opportunities offered by BGD.

Big Geospatial Data in archaeology

Geospatial Data in archaeology has an extraordinary importance for the reconstruction of the past, by allowing the identification of territorial patterns. Technological advancements have positively affected archaeology, therefore currently archaeologists create, manage, and share geospatial archaeological data much more than they did before. It is specifically worth noting that today archaeological datasets cannot be defined as BGD, since the volume of data archaeologists work with rarely exceeds the capacity of the available resources. However,

---

[2] https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science_en (accessed 7/7/2021).
[3] Geospatial Data is any information relating to the relative position of things on the earth's surface.
[4] https://www.go-fair.org/fair-principles/ (accessed 7/7/2021).
[5] https://www.ogc.org/ (accessed 7/7/2021).
[6] http://docs.opengeospatial.org/wp/16-131r2/16-131r2.html (accessed 7/7/2021).

it must be highlighted that although the volume of archaeological data is still manageable, BGD will increase significantly in the future. The debate on these aspects (McCoy 2017: 74–94; Cooper and Green 2016: 271–304; Gattiglia 2015: 113–124) is very active, considering the amount of data constantly produced by rescue archaeology, by public and private organizations that hold excavation concessions, by universities, and by local authorities in charge of safeguarding cultural heritage.

The challenge is to model the growing geospatial datasets before it becomes necessary to find specific solutions for the integration. From this point of view, it is important to identify the best practices for the creation, publication, and preservation of geospatial data also through the adoption of open standard formats. This approach guarantees the reliability of the data, allows their correct divulgation and their valid reuse in different application domains.

The last decade has witnessed a proliferation of standards and formats for the interchange of data: up to now, OGC has approved 69 formats.[7] There is a growing literature on the topic (Previtali and Valente 2019: 17–27; Richards-Rissetto and Landau 2019: 120–135; Carlisle *et al.* 2014), but a specific guideline for orienting researchers among the multitude of formats is still missing; the lack of best practices makes it difficult to encourage the interoperability and the interaction with a broad variety of sources. Notwithstanding, there is a great potential in using standard formats for the online publication of spatial databases.

The state of the art

A key approach for understanding in which way to proceed leads to the analysis of how the current scenario in Italy responds to the process of standardizing the data formats available online (geoportals and digital repositories) and how they can be integrated. Firstly, the institutional experiences in the field of management geospatial data for cultural heritage was examined. An important push for the homogenization of the spatial data came from the Ministry of Cultural Heritage with the publication of the Circular 30/2019.[8] This document describes how Geospatial Data must be represented (both for raster and vector data) within the documentation that must be delivered at the end of excavation, and indications relating to the publication and dissemination of research data. The formats admitted are the ESRI Shapefile,[9] OGC GeoPackage[10] and DXF,[11] all with specific geometry (*multipolygon*) and reference system (WGS 84 – EPSG: 4326).

The ESRI Shapefile is one of the most popular formats used to encode geographic data and even if it is not included in the OGC list, it has become a standard *de facto*. It is worth noting that in the above list, only the OGC Geopackage is a standard approved by OGC. It is one of the most recent standards (2014) designed to store and transfer complex and voluminous data (raster and vectors), without losing the style of vectors among the data elements. The documentation, organised through these *criteria*, will then populate the National Geoportal

---

[7] https://www.ogc.org/docs/is (accessed 7/7/2021).
[8] http://www.ic_archeo.beniculturali.it/getFile.php?id=494 (accessed 7/7/2021).
[9] https://doc.arcgis.com/it/arcgis-online/reference/shapefiles.htm (accessed 7/7/2021).
[10] https://www.ogc.org/standards/geopackage (accessed 7/7/2021).
[11] https://documentation.help/AutoCAD-DXF/ (accessed 7/7/2021).

Figure 1: Resources available for 'Archaeological Monuments' (https://data.yorkopendata.org/dataset/archaeological-monuments, accessed 13/10/2021).

for Archaeology,[12] which will collect data already partially structured. At present, the platform hosts some local, regional, or national portals dealing with spatial data. The list, particularly useful for finding spatial data already available and accessible, includes some specific archaeological initiatives or thematic sites. Unfortunately, a search engine has still not been developed, and users must look for data in every single archive. Furthermore, there is no available information on licensing and the re-use of data, and on how archaeologists should structure their data to make them completely interoperable.

The awareness that archaeologists, above all for those not super-technical, have to face some problems when they deal with spatial data, leads to such questions as: among all these standards which could be the best to use? How are they going to ensure the effective interoperability of their data with others? How can we shape our knowledge in the perspective of BGD?

Searching for answers based on consolidated frameworks, it is important to look at two useful English experiences: York Open Data[13] and the Archaeology Data Service[14] have represented in recent years a reference point for open data and public sharing of knowledge in archaeology.

---

[12] The project, still ongoing, is aimed at the creation of an online service for the archiving, research and knowledge of data relating to the Italian archaeological heritage (http://www.ic_archeo.beniculturali.it/it/222/il-geoportale-nazionale-per-l-archeologia-gna, accessed 7/7/2021).
[13] https://www.yorkopendata.org/ (accessed 7/7/2021).
[14] https://archaeologydataservice.ac.uk/ (accessed 7/7/2021).

Data Type	Preferred File Format	Accepted File Format	Metadata Template Download Type	Example
Collection-level Metadata			Microsoft Word Open Office Document	
3D Models, Visualisation, and Virtual Reality	Virtual Reality Modelling Language .vrml Wavefront OBJ File .obj (+ .mtl + jpg textures)	Adobe Portable Document Format (3D) .pdf This is accepted for dissemination purposes only, it is not suitable for preservation of 3D data. STL .stl	Microsoft Excel Open Office Spreadsheet	
GIS (Geodatabases)		Delimited text and ESRI Shapefile .csv + .shp GeoJSON .geojson	Microsoft Excel Open Office Spreadsheet csv	GIS PDF
GIS (Raster)	Geo-referenced TIF Image .tif (+ world file: .tfw) or GeoTIFF	ERDAS Imagine files .img (+ .rrd, aux.xml, img.xml) ESRI GRID ascii .asc/.grd ESRI GRID binary .adf JPG World .jpg + .jpw (.rrd, .aux, .xml) Keyhole Markup Language .kml PNG World .png + .pgw (.rrd, .aux, .xml)	Microsoft Excel Open Office Spreadsheet csv	GIS PDF

Figure 2: Table of all file types accepted for GIS by ADS (https://guides.archaeologydataservice.ac.uk/, accessed 13/10/2021).

The city of York offers several downloadable datasets about the city, e.g. urban assets, the environment, archaeological monuments, events. The resources are divided by the geometry, easily downloadable through three non–proprietary open formats: CSV, KML and GeoJSON (Figure 1). It is possible to download the datasets with no legal restrictions on their use. It is also possible to easily read and understand what information each dataset contains, by the simple use of a text editor, and the data are both human readable and machine-parseable. These data can be re-used with no problems relating to technical specifications.

The cutting-edge ADS project has updated (July 2020) its guideline for depositors with a list of all file types and formats they accept, providing also a template for the required metadata associated with each data-type. ESRI Shapefile and CSV or the GeoJSON file are admitted for GIS-data (Figure 2).

A practical example

It is noteworthy that there is a common thread that connects the experiences mentioned above, identifiable in the GeoJSON format,[15] a format largely used to encode data for browser-based web applications. It is useful to compare it with another widely used format, the ESRI Shapefile, by a practical example of the analysis and management of an archaeological dataset, and to demonstrate the benefits of using open formats for the scientific research and dissemination. The aim was to concretely verify the advantages of using a GeoJSON file instead of an ESRI Shapefile through a simulation of a survey on restrictions in urban development planning, using data available online. The pipeline was the following:

---

[15] https://geojson.org/ (accessed 7/7/2021). GeoJSON has become a standard (RFC 7946) in 2014, although the specification was finalized in June 2008.
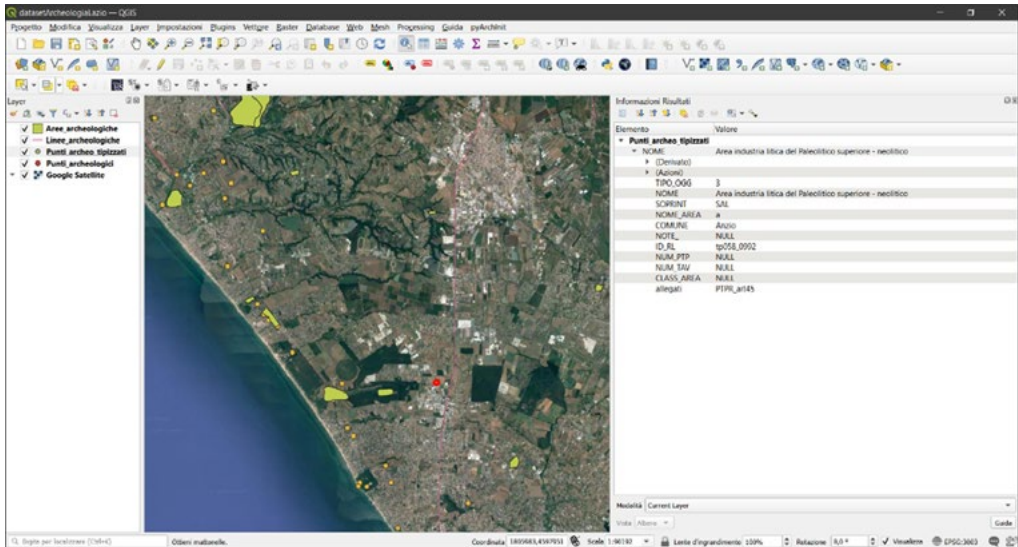
Figure 3: Different geometries and attribute tables of archaeological area in Lazio Region, shown in QGIS interface.

- Downloading a set in Shapefile format from a geoportal.
- Loading the dataset into QGIS.
- Exporting it in GeoJSON.
- Comparing the two datasets encoded in SHP and GeoJSON formats.

Uploading the GeoJSON file into a web and a desktop application and analyse if the whole process, from a non-open format to an open one, worked out.

The dataset for archaeological areas in the Lazio territory was downloaded by the Open Data portal of Lazio Region,[16] in the section Piano Territoriale Paesistico Regionale (PTPR), made of points, lines and polygons in shapefile format. The data were uploaded into the open-source GIS desktop application QGIS 3.10,[17] to visualize the geometries and attribute tables linked to each resource (Figure 3).

The next phase involved the conversion of the original Shapefiles in a standard and open format, i.e. GeoJSON, to simplify the management of data. The original data of the Shapefile were projected in ED50/33N (EPSG:23033) coordinate reference system, and were necessarily reprojected into WGS84 (EPSG:4326), according to the syntax of GeoJSON.[18] The subsequent comparison between the two data sets (Figure 4) has clearly shown:

---

[16] http://dati.lazio.it/catalog/it/dataset?tags=PTPR (accessed 7/7/2021).
[17] https://www.qgis.org (accessed 7/7/2021).
[18] An 'Internet Standards Track Document' made by the Internet Engineering Task Force (IETF) provides additional information about the CRS for the GeoJSON Format (https://datatracker.ietf.org/doc/html/rfc7946#section-4, accessed 7/7/2021).
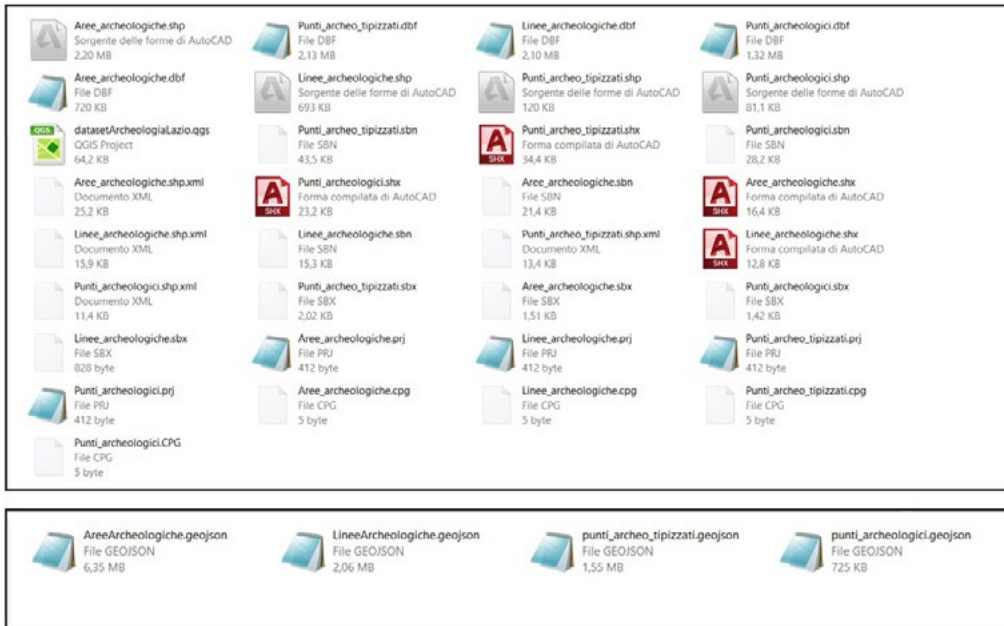
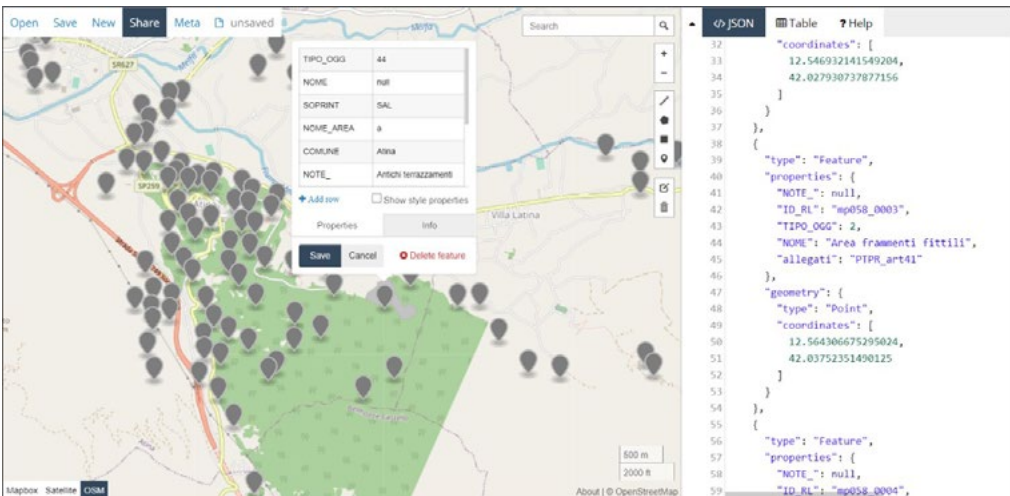Figure 4: Comparison between the two datasets in shapefile (on top) and GeoJSON (down).
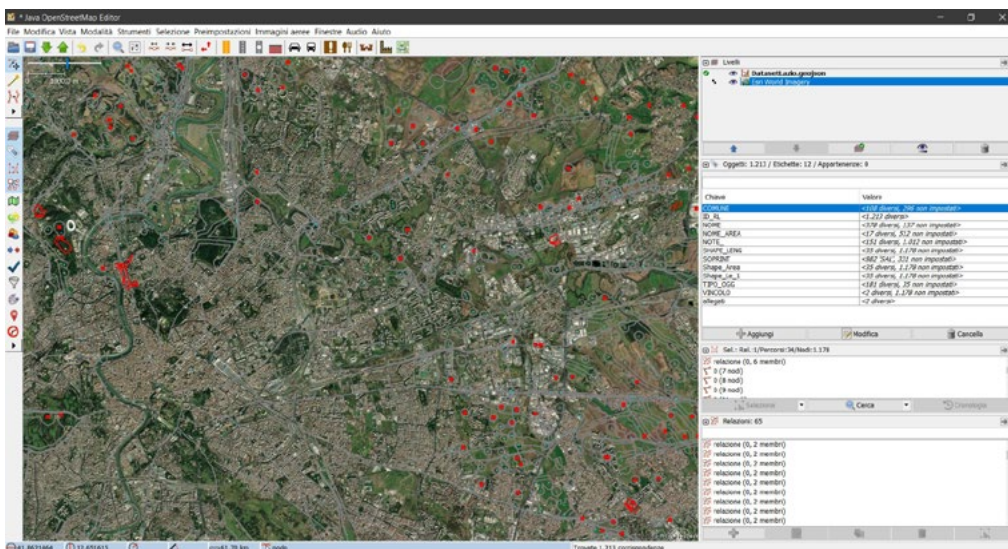

Figure 5: The dataset uploaded on geojson.io.

Figure 6: The GeoJSON file uploaded into OpenStreetMap Editor.

- Shapefile is a multifile format, it uses at least three files (*.shp, *.dbf, *.shx), making it impossible to share just one file; furthermore only one geometry type (point, line, polygon and others) is supported by a single file, and mixed geometry features are not possible. As a result, 33 files are needed to describe the four geometries of archaeological features of Regione Lazio. Users could zip all the files into one archive and unzip them on the other end of the distribution chain, but this is error-prone and limits interoperability. In addition, there is no way to describe topological relations in the format or to store 3D data with textures or parametric objects.
- GeoJSON data, exported through the specific QGIS function, has just four files to describe the four different geometries; the advantages of having the same data but with a smaller number of files is unarguable, even though less complexity makes it easier to store and exchange data, thus promoting interoperability. It can handle complex vector data features and build complex hierarchical data models. GeoJSON is recommended as a shapefile replacement for data interchange, particularly for web services, such as the ones used in this procedure.

To check the functionality of the tested datasets, it was necessary to deal with a single file containing all the geometries. To achieve this purpose, an open web tool was used for editing GeoJSON data on the internet: geojson.io.[19] It enables editing raw GeoJSON through a map interface, therefore it has been used to merge the four geometric GeoJSON files, thus combining the 7,363 resources; this stage confirmed that in the fusion process all the linked attributes have been preserved (Figure 5). It has also been integrated with GitHub,[20] allowing users to edit their data in Git repositories and GitHub Gists, making data accessible and downloadable.

---

[19] http://geojson.io/ (accessed 7/7/2021). This is an open-source software under the permissive MIT license. That license applies to the software and not the data created with geojson.io (http://geojson.io/about.html).
[20] https://github.com/mapbox/geojson.io (accessed 7/7/2021).

A useful tool to upload, visualize, edit, and query different types of data is the open Desktop Application *Java OpenStreetMap Editor*.[21] The merged GeoJSON file was uploaded to JOSM Editor to check that all the attributes and relations had been preserved without losing any part of the information. With this tool it has been possible to search any single information independently from the geometries[22] (Figure 6).

Conclusion

An approach such as the one illustrated in this paper, with the coding of spatial data structured through shared and common standards approved by the scientific community, fits perfectly into the theme of the production and consumption of BGD, an issue that is going to be faced soon within the scope of the implementation of web infrastructures able to integrate resources and open-data. According to a document approved by W3C and OGC in March 2021[23] on the best practices for the publication of spatial data on the Web, there are only two reference geometry formats widely used in the geospatial and Web communities: GML for 3D data and GeoJSON for 2D geometries. As GeoJSON is serialized in JSON, this format is ready for browser-based Web applications.

One of the challenges to face is to understand how to preserve the topological information of the spatial data in relation to web mapping; an extension called TopoJSON[24] has been developed to express the geospatial topology encoded in GeoJSON, but a more in-depth analysis is necessary to understands how it fits with GeoJSON format.

In this context, we focused on the advantages of using specific formats to encode 2D data for web applications; however, further investigation is needed on 3D spatial data: an in-depth study for its excellent prospects is the standard open format Geopackage, the data container based on SQLite, which allows users to manage vectors, rasters, attributes, topology, and, thanks to its extensions, even 3D data. This is the case of the 3D tiles[25] format, even if it is not yet an official extension to the GeoPackage Encoding Standard, but the future is promising: 3D Tiles is an open specification for sharing, visualizing, fusing, interacting with, and analysing massive heterogeneous 3D geospatial content across desktop, web, and mobile applications.

It is also worth highlighting the importance of formats such as GeoJSON and Geopackage, which, differently from Shapefile, are transparent, encouraging interoperability. GeoJSON allows the easy reuse of spatial data, undoubtedly opening new opportunities for integrating data within the scenario of the semantic web; Geopackage does the groundwork for the management of 3D data with spatial coherence.

---

[21] https://josm.openstreetmap.de/ (accessed 7/7/2021).
[22] Differently, in QGIS the single GeoJSON file, once uploaded in the software, was divided in three layers, corresponding to the three geometries described in the file.
[23] https://w3c.github.io/sdw/bp/ (accessed 7/7/2021).
[24] https://github.com/topojson/topojson-specification (accessed 7/7/2021).
[25] http://www.opengis.net/doc/CS/3DTiles/1.0 (accessed 7/7/2021).

Bibliography

Carlisle, P.K., I. Avramides, A. Dalgity and D. Myers 2014. The Arches Heritage Inventory and Management System: a standards-based approach to the management of cultural heritage information. In *Proceedings of the CIDOC Conference Access and Understanding – Networking in the Digital Era, Dresden, Germany, 6–11 September 2014.* (http://archesproject.org/wp-content/uploads/2014/10/I-1_Carlisle_Dalgity_et-al_paper.pdf, accessed 13/10/2021).

Cooper, A. and C. Green 2016. Embracing the Complexities of 'Big Data' in Archaeology: the Case of the English Landscape and Identities Project. *Journal of Archaeological Method and Theory* 23: 271–304 (https://doi.org/10.1007/s10816-015-9240-4).

Forte, F. 2019. Informazioni spaziali e strumenti d'indagine semantica sui dati archeologici digitali. *Newsletter di Archeologia CISA* 10: 95–111.

Gattiglia, G. 2015. Think big about data: Archaeology and the Big Data challenge. *Archäologische Informationen* 38: 113-124 (https://doi.org/10.11588/ai.2015.1.26155).

Hiebel, G., G. Goldenberg, C. Grutsch, K. Hanke and M. Staudt 2020. FAIR data for prehistoric mining archaeology. *International Journal on Digital Libraries* (https://doi.org/10.1007/s00799-020-00282-8).

McCoy, M.D. 2017. Geospatial Big Data and archaeology: Prospects and problems too great to ignore. *Journal of Archaeological Science* 84: 74–94.

Previtali, M. and R. Valente 2019. Archaeological documentation and data sharing: digital surveying and open data approach applied to archaeological fieldworks. *Virtual Archaeology Review* 10 (20): 17–27 (https://doi.org/10.4995/var.2019.10377).

Richards-Rissetto, K. and H. Landau 2019. Digitally Mediated Practices of Geospatial Archaeological Data: Transformation, Integration, & Interpretation. *Journal of Computer Applications in Archaeology* 2 (1): 120–135 (https://doi.org/10.5334/jcaa.30).