Archeologia e Calcolatori
37.1, 2026, 387-398
doi 10.19282/ac.37.1.2026.20

THE OPEN DIGITAL ARCHAEOLOGY HUB: ADVANCEMENTS AND FUTURE DIRECTIONS

1. Introduction

This paper aims to present the most recent advances related to ArchaeoHub – the Open Digital Archaeology Hub, https://open-archaeohub.cnr.it/ – a platform developed as a pilot service within the framework of the PNRR IR H2IOSC project (Caravale, Moscati, Rossi 2026). The platform, already described in previous publications, is designed to facilitate integrated access to editorial content, datasets, interactive resources, images, and scientific bibliographies related to the field of digital archaeology (Caravale *et al.* 2025). Bringing together these heterogeneous resources, ArchaeoHub provides users with a unified point of access and a simplified framework for consultation and exploration (Mancuso *et al.* 2025). At present, ArchaeoHub exposes 1406 open access papers, 16,230 bibliographic references, 7129 images, 137 datasets, 85 interactive resources and 136 research projects.

The ingestion and organisation of such a large body of metadata, although significantly supported by the use of structured sources such as the «Archeologia e Calcolatori» (A&C) repository, has required considerable effort in terms of human resources. For this reason, part of the work carried out after the launch of the platform, and further discussed in this contribution, has focused on the progressive automation of metadata creation and data preparation workflows, with the aim of easing the workload on staff and enabling a more streamlined, long-term sustainable management of the platform.

Particular attention has been paid to the geographic component, which plays a key role in enabling access to and exploration of archaeological information, as well as to the possibility of connecting ArchaeoHub to further structured data sources, such as Omeka S, through more scalable and semi-automated procedures. Attention has likewise been devoted to the bibliographic component and, more specifically, to the parsing process necessary to bulk transform unstructured references into coherent machine-readable metadata, ready to be exposed by ArchaeoHub through BiDiAr (Bibliography of Digital Archaeology, https://bidiar.cnr.it). Finally, the paper will briefly outline future directions for the platform, with reference to new forms of data interaction and to the potential offered by network graph-based visualisations for browsing and contextualising the relationships between resources.

G.M., N.P., A.D.

2. Toponym recognition through Large Language Model (LLM)

The geographic dimension of ArchaeoHub's content has been addressed through the development of a dedicated semi-automated pipeline for toponym extraction and geographic disambiguation, applied to the full corpus of A&C (Fig. 1). This work can be situated within a broader line of experimentation in which LLMs are used, with different objectives, to extract, organise, and interpret geo-knowledge from complex textual corpora (Hu *et al.* 2026a, 2026b; Pisetta *et al.* 2026). In the present case, the objective is more specifically operational: the production of reliable and reusable geographic metadata to support the indexing, browsing, and integration of editorial content within ArchaeoHub[1].

The process begins with the systematic collection of article metadata – titles, abstracts, and existing editorial annotations – through the A&C REST API (Paraciani 2024). For reasons related to the archaeological representativeness the pipeline deliberately focuses on titles and abstracts rather than on full texts. Extending the extraction process to entire articles would certainly increase the number of detected toponyms, but it would also capture many place names mentioned only for comparison and broader contextualisation. While such an approach could be valuable for specific research purposes, it would be less effective for the creation of geographic metadata intended to describe the primary spatial focus of each contribution.

Toponym extraction is performed through a Named Entity Recognition (NER) step based on a large language model running locally via Ollama (an open-source software application to manage LLMs). The use of a locally deployed model, specifically Qwen2.5 32B, rather than a cloud-based service, responds to both reproducibility and data sovereignty requirements. The model is prompted with the article title and abstract and instructed to identify ancient and historical place names, returning a structured JSON response that includes, for each toponym, the raw form as it appears in the text, a normalised scholarly form, the entity type – such as site, region, river, or settlement – a verbatim context fragment, and a confidence assessment. The use of an LLM at this stage is particularly helpful because it allows the pipeline to distinguish archaeologically relevant place names from expressions that contain geographic terms, but do not correspond to the spatial focus of the article, such as institutional names or affiliations – for instance, 'Normale di Pisa' or 'Sapienza Università di Roma'. This reduces the risk of introducing false geographic associations into the metadata and helps preserve the semantic relevance of the resulting

[1] This pipeline, whose code is publicly available on a GitLab repository hosted at https://baltig.cnr.it/giacomo.mancuso/AeC_geoparser to ensure transparency and replicability, is structured both as a single Python file and as a sequence of six Jupyter notebooks. It is designed to be applicable to any journal or editorial corpus sharing a comparable metadata structure; most of the code was vibe-coded with Anthropic Claude Code.

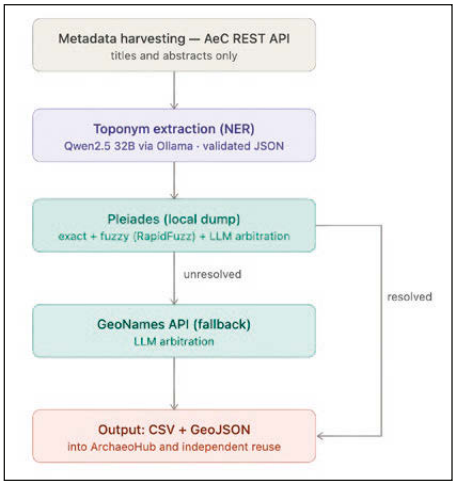

Fig. 1 – Conceptual diagram of the toponym recognition pipeline.

spatial index. The structured output format is enforced through JSON schema validation, ensuring consistency across the entire corpus.

The disambiguation step matches each extracted toponym against two geographic gazetteers in sequence. Pleiades (https://pleiades.stoa.org; Bagnall *et al.* 2026) is queried first through a locally indexed copy of its bulk data dump, using a combination of exact string matching, fuzzy matching via the Python RapidFuzz library, and, in cases of ambiguity, a second LLM-based arbitration step. Toponyms that remain unresolved after exhausting the Pleiades index are passed to the GeoNames API as a fallback, with a further LLM arbitration step to assess whether the returned candidates correspond to archaeologically relevant locations. Each resolved toponym is assigned a Pleiades URI or a GeoNames identifier, together with geographic coordinates and a record of the disambiguation method applied. In a sense, and with all due differences, this part of the workflow mirrors the approach already adopted by the Pelagios project (Vitale *et al.* 2021). In this case, the final output consists of a structured CSV table and a GeoJSON file, both suitable for updating the data source ingested into ArchaeoHub and for independent reuse, such as in a GIS (Fig. 2).

A comparative evaluation of the pipeline results against the existing manual annotations available for a subset of the A&C corpus is also provided[2], offering a quantitative assessment of the method's precision and

[2] This process was initially carried out on the A&C issues of the last five years (2021-2025), in order to index them on ArchaeoHub (Mancuso *et al.* 2025).

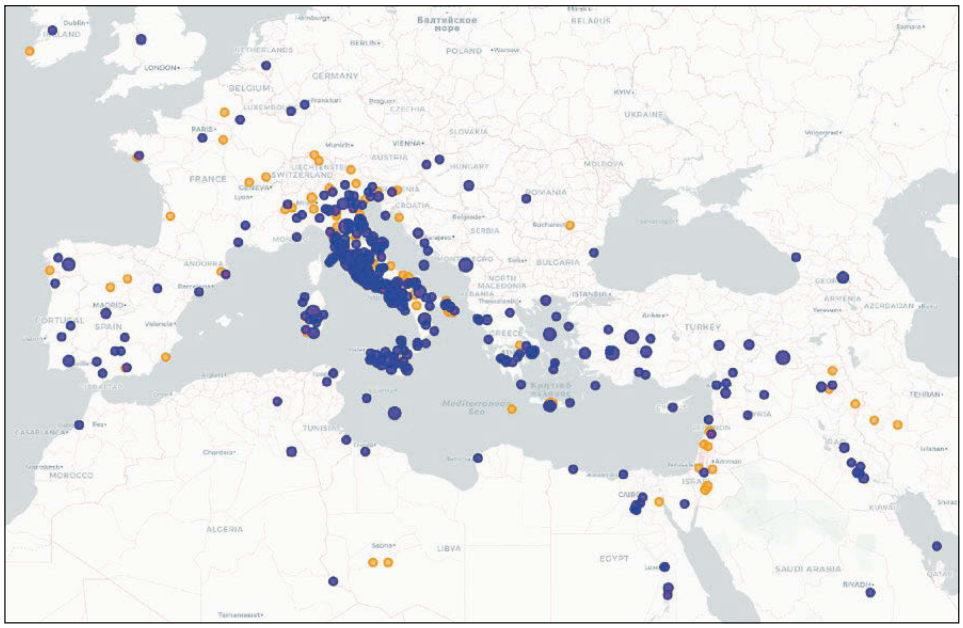

Fig. 2 – Map of all the toponyms identified and parsed to a Pleiades (blue) or GeoNames (orange) entity.

recall, together with a discussion of systematic discrepancies. Overall, the results can be considered highly encouraging. The system proved effective in identifying and distinguishing archaeologically relevant toponyms, with no false positives observed in the evaluated sample. Compared with the existing manual metadata, however, the pipeline produced a broader set of geographic entities. This difference is largely due to the fact that the manual annotation process had systematically excluded certain categories of place names, such as regions, physical geographic features, rivers, and other broader spatial entities, whereas the automated workflow retained them when they could be matched to gazetteer records. This discrepancy should not necessarily be interpreted as an error, but rather as an indication of the different levels of geographic description that may be relevant depending on the intended use of the metadata. For ArchaeoHub, whose primary objective is to provide meaningful geographic access points to archaeological resources, a more selective treatment of these entities was made through the entity type values recorded in the NER step. Future refinements of the pipeline could therefore include more explicit instructions for filtering or weighting broader geographic entities at an earlier stage of the workflow, as well as a more fine-grained classification of the types of places for which geographic resolution should be attempted.

In this sense, the experiment represents a first and promising level of automation in the preparation of geographic metadata for ArchaeoHub. While human supervision remains necessary, especially for defining the scope and granularity of geographic indexing, the pipeline shows the potential to significantly reduce the manual effort required for the processing of large editorial corpora and to support more scalable workflows for the integration of archaeological information into the platform.

G.M.

3. Adaptation of the ArchaeoHub for Omeka S

The ArchaeoHub is essentially a data aggregator and an indexing platform, since it collects data from external sources via REST APIs and only keeps an index in its internal database, along with minimal metadata and references to the original record identifiers, rather than fully replicating the data with its own copies. This collection (or harvesting) activity is mostly automated, being performed by a custom command that runs on the server as a scheduled daily task. For this to work correctly, the command needs to have access to a pre-compiled list of data sources with their respective JSON endpoints (in the form of URLs), associated with the corresponding resource types that it will have to index (articles, images, datasets, interactive resources, etc.).

In the earliest development phase, ArchaeoHub's data sources consisted of A&C's and IADI's REST APIs for journal articles and images, respectively, and the first implementation of DHeLO (Digital Heritage Landscaping platform – https://dhelo.cnr.it) for datasets resulting from H2IOSC landscaping activities (Mancuso, D'Eredità 2024). When DHeLO was transferred to Omeka S in the context of the CHLOE platform (Cultural Heritage Linked Open data Environment – https://chloe.cnr.it) and consequently extended (Mancuso 2025), there was a need to transition the ArchaeoHub to a new data source, that is, the JSON-LD REST APIs provided by Omeka S. These are structured according to a standard metadata format intended for usage in a Linked Open Data (LOD) context, while the JSON document provided by A&C and IADI (Interactive Atlas of Digital Images – https://iadi.archcalc.cnr.it), as well as DHeLO, adopted a custom (arbitrary) schema, so the ArchaeoHub was developed to process only this particular schema when importing data.

As mentioned above, Omeka S uses JSON-LD as its exchange format, which is essentially a way to serialise the RDF data model using an 'ordinary' JSON document, given the latter's almost ubiquitous presence in web-based services (Lanthaler, Gütl 2012). Thus, the schema adopted by Omeka S is standardised and relies on well-known contexts, defined by their corresponding ontologies, to represent the data. This allows clients consuming the APIs to leverage the Linked Data approach, since full identifiers (IRIs) can be derived

```
{
  "@context": "https://chloe.cnr.it/api-context",
  "@id": "https://chloe.cnr.it/api/items/322",
  "@type": [
    "o:Item",
    "dctype:Dataset"
  ],
  "o:id": 322,
  "o:is_public": true,
  "o:owner": {
    "@id": "https://chloe.cnr.it/api/users/2",
    "o:id": 2
  },
  "o:resource_class": {
    "@id": "https://chloe.cnr.it/api/resource_classes/24",
    "o:id": 24
  },
  "o:resource_template": {
    "@id": "https://chloe.cnr.it/api/resource_templates/8",
    "o:id": 8
  },
  "o:thumbnail": null,
  "o:title": "Modello 3D del Sarcofago degli Sposi",
  "thumbnail_display_urls": {
    "large": null,
    "medium": null,
    "square": null
  },
  "o:created": {
    "@value": "2025-02-05T16:25:06+00:00",
    "@type": "http://www.w3.org/2001/XMLSchema#dateTime"
  },
  "o:modified": {
    "@value": "2026-03-24T08:50:22+00:00",
    "@type": "http://www.w3.org/2001/XMLSchema#dateTime"
  },
```


Fig. 3 – Excerpt of a JSON-LD document as provided by Omeka S for a dataset record.

from the JSON-LD '@context' and '@id' properties and can thus be followed to (optionally) import additional data. Fig. 3 shows a listing (excerpt) of a JSON-LD document from DHeLO as an example. The '@context' property at line 2 refers to a specific Omeka S endpoint that provides the list of all relevant vocabularies corresponding to the namespaces found in the JSON, including Omeka's own custom vocabulary (with the 'o' prefix) and the 'geo' vocabulary to represent georeferenced data, according to the WGS 84 standard.

Adapting the ArchaeoHub so it could harvest JSON-LD data was important, both for better compliance with standards and in the view of including more data sources from other providers within the digital archaeology domain. ArchaeoHub imports data following a two-step process: first, it updates its internal configuration by building an index of references for each resource provider, then it queries the individual HTTP endpoints associated with these references and stores the relevant information in its database. When it comes to Omeka S sources, ArchaeoHub parses the JSON-LD to extract the following metadata: 1) Resource type; 2) Resource name and DHeLO identifier; 3) Geospatial information.

The procedure applied for point 3 above is an example of the LOD approach: the import process looks for the 'dcterms:spatial' property, which is a list of objects, in the main record document returned by Omeka S. If it finds it, along with a valid value for the '@id' property of the first object in the list, it follows the corresponding URI, which points to a Location resource in DHeLO. Then, by opening and parsing the returned JSON-LD document, it proceeds to look for Pleiades URIs in the 'dcterms:identifier' property[3], which are then followed to fully retrieve geospatial metadata, including place names as reported by Pleiades. It should be noted that, to simplify the importing process, point coordinates – latitude and longitude – for locations are included directly in the 'geo:lat_long' property that is returned by Omeka S in the original document. This property is not public, so it is not shown in DHeLO's web pages, although it is used to show georeferenced points on its web map. ArchaeoHub can retrieve it by authenticating to the REST API via private key credentials. Geospatial information, as mentioned above, is necessary for the ArchaeoHub to properly assign items to thematic collections, as well as showing them in the contextual map.

Regarding bibliographic references, currently ArchaeoHub refers only to A&C's APIs, since they include the full HTML bibliography for a given article[4] as retrieved directly from the Zotero collection linked to BiDiAr, via collection keys and Zotero's own APIs, and formatted according to the journal's citation style, itself deposited to Zotero's official repository. To avoid querying the Zotero APIs each time the bibliography of a record needs to be retrieved, which would be cumbersome, A&C caches it by persisting the string in its internal database.

N.P.

4. Bibliographic metadata through semi-automated parsing

One of the distinctive features of ArchaeoHub is that it exposes not only the articles themselves but also the bibliographic metadata of the reference lists that accompany them (Mancuso *et al.* 2025). These data are harvested from BiDiAr, an autonomous platform for bibliographic references exposed both through a Zotero shared Library (https://www.zotero.org/groups/5293298/bidiar) and an independent website (https://bidiar.cnr.it), here acting as a data-provider for ArchaeoHub (Caravale *et al.* 2025).

BiDiAr grew out of the citations extracted from the articles published in A&C, turning the journal's citational heritage into a structured, queryable dataset

[3] This is also a list (JSON array) of object values. Pleiades and GeoNames URIs are reported in the '@id' property of the corresponding objects in the list.

[4] Returned as a text string in one of the JSON properties.

that stands alongside, and complements, the article metadata already indexed in its repository (Mancuso, D'Eredità 2024). Before BiDiAr the reference lists in A&C existed only as free-form text, so turning them into usable data demanded a long and painstaking effort of digitisation and structuring. At first this was done by hand, with entries keyed in manually or through the Zotero Connector[5]; a slower route that allowed close control over the data and its quality. However, as the corpus grew, manual entry became untenable, and automation – again drawing on AI-based tools – turned from convenience to necessity. The tool used for this process is AnyStyle (https://anystyle.io/), a bibliographic parser built on Conditional Random Fields (CRF), chosen above all for its flexibility (Cioffi, Peroni 2022): the same engine can be used interactively, through a web interface well suited for testing, or driven at scale from the command line. Around it, a sequential pipeline was built (Fig. 4), in which each stage performs a well-defined task based on free open-source tools: 1) text extraction from PDF/InDesign files; 2) parsing with AnyStyle and export to BibTeX; 3) import BibTeX into Zotero BiDiAr library; 4) export to BiDiAr on Omeka S.

Text extraction came first: wherever possible the InDesign files were used for typesetting, kindly supplied by the publisher, while articles predating 2005, which survive only as low definition PDFs, were recovered through high-resolution scanning and OCR. The result was a separate file for each article's bibliography, later fed one by one into the web version of AnyStyle. Its interface moves through three stages: 1) an automatic parse of the pasted text; 2) a manual pass in which the segmentation can be corrected by hand; 3) a save step that handles the export, always with the 'Use my changes to train the parser!' option selected, so that each correction feeds back into the model. Internally, AnyStyle tokenises the text and tags each token with a semantic label output – author, title, date, container-title, volume, pages, DOI, and so on – producing structured references that can then be exported as a cumulative BibTeX file for compatibility with Zotero. Trained largely on English-language references, the base model stumbled in predictable ways over the A&C style: it confused author with date in the 'Surname N. year' pattern, missed the guillemets (« ») that mark journal titles as container-titles, mis-segmented edited volumes in the 'Author (ed.), Title' construction, and treated multilingual references inconsistently. Iterative training reduced these errors substantially, though it never dispensed with manual revision altogether – grey literature, archival sources, and online resources, whose origin and date are often untraceable, still had to be checked by hand. The BibTeX files were then loaded into Zotero and filed within a system of sub-collections that mirrors the harvesting logic of ArchaeoHub, with each article's citations

[5] A browser plugin that can automate metadata ingestion into the Zotero desktop application (https://www.zotero.org/download/).

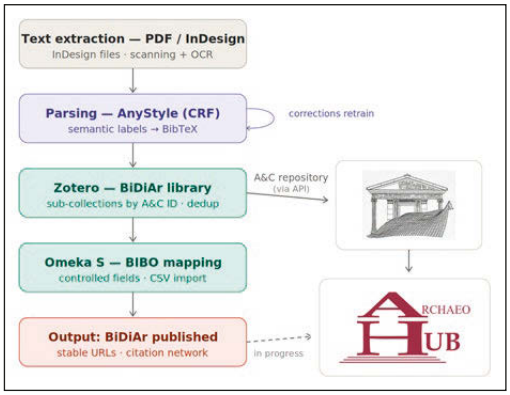

Fig. 4 – Conceptual diagram of the bibliographic parsing pipeline and its outputs.

gathered under a collection named by its ID in the A&C repository. Batch import was followed by a systematic review on three fronts: 1) correcting the fields the parser had mis-assigned; 2) filtering out the duplicate records that inevitably arise when the same work is cited across several articles (through Zotero's native deduplication tool); 3) reconciling author names against a controlled list.

The final stage, still underway, is the import into Omeka S. A dedicated Omeka S module – the Zotero Importer – can draw records directly from a Zotero library through a custom API key, but its granularity proved insufficient for our needs: several fields were left poorly disambiguated, and duplicates were awkward to resolve on the platform. Thus, the over 4000 references already on the platform were mapped first onto the BIBO ontology (https://www.dublincore.org/specifications/bibo/) and then batch imported from a CSV file. The work was then divided into two passes, one for the bibliographic metadata and one for the citation relationships. The first brings in only the records not yet present in Omeka S, matched and skipped by their Zotero ID, which is retained throughout; the second runs across every article, linking each record to the works it cites, once again by Zotero ID. Because that identifier stays with the record, the system can be configured to generate stable URLs automatically, letting the reader move directly from an Omeka record to the corresponding Zotero library without losing the connection between the two platforms. The citations themselves are expressed through the bibo:cites property and its reciprocal bibo:citedBy, which preserves the direction of each citation and open the possibility to further explore and analyse this wide network.

A.D.

5. Future developments

The sum of the processes described above – the integration of Omeka S as a data provider (§3) and the progressive BiDiAr data consolidation on the same platform (§4) – provides the ArchaeoHub with the means for moving beyond the simple display of bibliographic references associated with a resource. Alongside the traditional textual list of citations, ArchaeoHub can be configured to dynamically generate citation networks (in a graph), allowing users to explore the relationships among publications from multiple perspectives. Such visualisations may be organised according to different criteria, including chronological arrangements that make it possible to follow the diachronic development of scholarly debates. Users can therefore navigate not only the works cited by a selected publication, but also the subsequent publications that cite it, thereby opening new forms of literature exploration and contextualisation.

From a technical perspective, current experiments are focusing on the adoption of the JavaScript library D3.js, a widely used framework for producing interactive and data-driven visualisations directly in web browsers (Bostock, Ogievetsky, Heer 2011). Through a series of queries to the Omeka S APIs, it is possible to retrieve bibliographic records together with their citation relationships and transform them into dynamic graph structures displayed within dedicated HTML components (Fig. 5). The flexibility of D3.js makes it possible to create interactive and highly configurable network visualisations. Citation graphs can be filtered according to different criteria, such as the degree of connectivity of individual publications or the type and direction of citation relationships. At the same time, each node can maintain direct links to the corresponding BiDiAr records in Omeka S, establishing a seamless connection between ArchaeoHub and the underlying bibliographic infrastructure. This process could be facilitated by the fact that ArchaeoHub can already retrieve data from Omeka S, as described in §3. When importing items from their JSON-LD records, ArchaeoHub could store an intermediate representation of bibliographic references extracted by indexing the lists of objects found under the relevant 'bibo' properties, optimised for consumption by D3.js. Therefore, when an article item is viewed in the ArchaeoHub, the system retrieves this representation[6], along with the rest of the metadata, and builds the graph dynamically in the page's visual context.

These developments represent an important step towards transforming ArchaeoHub from a platform primarily devoted to the aggregation and indexing of digital archaeological resources into an environment that also supports the exploration of the conceptual relationships that connect those resources. More generally, they demonstrate the potential of combining Linked Open Data

[6] Typically, a JSON object, natively interpretable via JavaScript.

Fig. 5 – Citation network of a single article with depth 1 (left) and depth 2 (right) connections.

technologies with interactive visualisation techniques to create new modes of discovery, navigation, and interpretation within the digital archaeology domain.

G.M., N.P.

Giacomo Mancuso, Nicolò Paraciani, Antonio D'Eredità
Istituto di Scienze del Patrimonio Culturale - CNR
giacomo.mancuso@cnr.it, nicolo.paraciani@cnr.it, antonio.deredita@cnr.it

Acknowledgements

This research was supported by H2IOSC Project - Humanities and cultural Heritage Italian Open Science Cloud funded by the European Union NextGenerationEU - National Recovery and Resilience Plan (NRRP) - Mission 4 'Education and Research' Component 2 'From research to business' Investment 3.1 'Fund for the realization of an integrated system of research and innovation infrastructures' Action 3.1.1 'Creation of new research infrastructures strengthening of existing ones and their networking for Scientific Excellence under Horizon Europe' - Project code IR0000029 – CUP B63C22000730005. Implementing Entity CNR. The authors also wish to thank the publisher All'Insegna del Giglio for kindly granting access to the InDesign files used for text extraction in the bibliographic parsing pipeline described in §4.

REFERENCES

Bagnall R.S., Talbert R.J.A., Becker J.A., Elliott T., Gillies S., Horne R. 2026, *Pleiades: A Community-Built Gazetteer and Graph of Ancient Places. Institute for the Study of the Ancient World/Ancient World Mapping Center* (https://pleiades.stoa.org).

Bostock M., Ogievetsky V., Heer J. 2011, *D³ Data-Driven Documents*, «IEEE Transactions on Visualization and Computer Graphics», 17, 12, 2301-2309 (https://doi.org/10.1109/TVCG.2011.185).

Caravale A., D'Eredità A., Mancuso G., Moscati P. 2025, *An open system for textual, visual, and bibliographic resources: The Open Digital Archaeology Hub*, «Archeologia e Calcolatori», 36.1, 455-468 (https://doi.org/10.19282/ac.36.1.2025.25).

Caravale A., Moscati P., Rossi I. (eds.) 2026, *The H2IOSC Project and Its Impact on Digital Antiquity within the E-RIHS Infrastructure*, Firenze, All'Insegna del Giglio (https://doi.org/10.19282/h2iosc.2026).

Cioffi A., Peroni S. 2022, *Structured references from PDF articles: Assessing the tools for bibliographic reference extraction and parsing*, in G. Silvello, O. Corcho, P. Manghi, G.M. Di Nunzio, K. Golub, N. Ferro, A. Poggi (eds.), *26th International Conference on Theory and Practice of Digital Libraries, TPDL 2022 (Padua 2022)*, Springer International Publishing, 425-432 (https://doi.org/10.1007/978-3-031-16802-4_42).

Hu X., Kersten J., Klan F., Farzana S.M. 2026a, *Toponym resolution leveraging lightweight and open-source large language models and geo-knowledge*, «International Journal of Geographical Information Science», 40, 3, 670-697 (https://doi.org/10.1080/13658816.2024.2405182).

Hu X., Sun Y., Hecking T., Kersten J., Klan F. 2026b, *UniTopRank: A scalable and language-independent method for toponym resolution*, «International Journal of Geographical Information Science», 1-28 (https://doi.org/10.1080/13658816.2026.2645831).

Lanthaler M., Gütl C. 2012, *On using JSON-LD to create evolvable RESTful services*, presented to *Proceedings of the Third International Workshop on RESTful Design (Lyon, France)*.

Mancuso G. 2025, *Beyond monitoring. Reimagining DHeLO as a Linked Open Data infrastructure for Cultural Heritage research*, «Archeologia e Calcolatori», 36.1, 443-454 (https://doi.org/10.19282/ac.36.1.2025.24).

Mancuso G., Caravale A., D'Eredità A., Moscati P. 2025, *Merging knowledge and tools in Heritage Science and Digital Archaeology. Practices from H2IOSC WP2*, in S. Campana, D. Ferdani, H. Graf, G. Guidi, Z. Hegarty, S. Pescarin, F. Remondino (eds.), *Digital Heritage 2025 (Siena 2025)*, Goslar, The Eurographics Association (https://doi.org/10.2312/dh.20253320).

Mancuso G., D'Eredità A. 2024, *DHeLO and BiDiAr: New digital resources within the H2IOSC Project*, «Archeologia e Calcolatori», 35.1, 521-542 (https://doi.org/10.19282/ac.35.1.2024.31).

Paraciani N. 2024, *A new website for the journal «Archeologia e Calcolatori»*, «Archeologia e Calcolatori», 35.2, 471-482 (https://doi.org/10.19282/ac.35.2.2024.49).

Pisetta J., Souza F.A., Amorim J., Antonio N.D., Nunes D.M., Araki H., Camboim S.P. 2026, *Mapping with words: Integrating Large Language Models into geospatial practice*, «ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences», X-3/W4-2025, 285–291 (https://doi.org/10.5194/isprs-annals-X-3-W4-2025-285-2026).

Vitale V., Soto P.D., Simon R., Barker E., Isaksen L., Kahn R. 2021, *Pelagios. Connecting histories of place. Part I: Methods and tools*, «International Journal of Humanities and Arts Computing», 15, 1-2, 5-32 (https://doi.org/10.3366/ijhac.2021.0260).

ABSTRACT

This paper presents recent developments of ArchaeoHub, the Open Digital Archaeology Hub developed within the PNRR IR H2IOSC project. The work focuses on two main lines of automation. The first concerns a semi-automated pipeline for toponym recognition and disambiguation applied to the «Archeologia e Calcolatori» corpus. The second concerns the adaptation of ArchaeoHub to ingest Omeka S sources via JSON-LD. Finally, the paper outlines future developments related to integration with BiDiAr (another platform developed within the PNRR IR H2IOSC project), opening the way to dynamically generated citation networks via D3.js.