Archeologia e Calcolatori
35.1, 2024, 543-562
doi 10.19282/ac.35.1.2024.32

A RESOURCE HUB FOR INTEROPERABILITY AND DATA INTEGRATION IN HERITAGE RESEARCH: THE H-SETIS DATABASE*

1. The Humanities and Cultural Heritage Italian Open Science Cloud (H2IOSC) Project

H2IOSC is a project led by the Consiglio Nazionale delle Ricerche (CNR), actively involving several of its Institutes (https://www.h2iosc.cnr.it/). Its main objective is to create a federated and inclusive cluster of the Italian nodes of the four European Research Infrastructures (RIs) in the field of Humanities and Cultural Heritage: DARIAH for Humanities, CLARIN for language sciences, OPERAS for scientific communication in the field of Humanities and Social Sciences, and E-RIHS for Heritage Science (HS). Their nature is very heterogeneous, including both physical instrumentation and repositories such as archives and databases, computing and communication systems that are essential for research purposes. The entire H2IOSC Project activity aims to support data-driven research and the digital transformation of the cultural and creative industries sectors. H2IOSC indeed promotes a data-centric approach, with data made accessible through an integrated digital environment designed according to FAIR principles (Findable, Accessible, Interoperable, Reusable; Wilkinson *et al.* 2016).

The Istituto di Scienze del Patrimonio Culturale (CNR-ISPC) is directly involved in the project as part of the E-RIHS network; the Milan branch of ISPC, as leader of Task 4.10 'Resources interoperability: DIGILAB resources (E-RIHS)' within WP4 'RIs Nodes and Resources Interoperability', is in charge of a general survey of semantic tools for the Heritage sector, of designing strategies to assure data interoperability, and of the integration of digital resources within the Heritage domain and the wider H2IOSC common semantic framework.

2. Heritage Science and semantic tools

2.1 *Heritage Science: a long-standing and yet novel discipline*

Heritage Science (HS) is an interdisciplinary research field that combines social sciences and natural sciences applied to cultural and natural Heritage. It is a research area whose scope and objectives are well known, but whose formalization as an autonomous sector has not yet been completed. This

[*] Both authors have equally contributed to the content of this paper.

definition of HS is quite recent, as it was jointly developed in 2019 by E-RIHS and the International Centre for the Study of the Preservation and Restoration of Cultural Property (ICCROM)[1], while the term itself was defined by the Science and Technology Select Committee of the British House of Lords and dates back to 2006 (House of Lords Science and Technology Select Committee 2006; Kennedy 2015, 214-215; Strlič 2018, 7260; Kennedy *et al.* 2024). The new definition aims to overcome disciplinary barriers, especially between the fields of Cultural Heritage Conservation, which has always focused on the technical aspects of assessing and controlling degradation, restoration, and protection of tangible Cultural Heritage, and that of social sciences, which focuses on the study of material evidence and the relationship between humans and the environment from a historical perspective, such as archaeology, art history, and anthropology, among others (Carman, Sørensen 2009).

In recent decades, the increasing use of scientific techniques involving remote data acquisition has enhanced the analytical potential and application of HS diagnostic methodologies. This is because the fragility of the materials under study often discourages or prohibits more invasive interventions and direct sampling. The chance to acquire data without physically contacting the object of study has expanded the number of measurements and the fields of investigation, increasing also scientific results (Kennedy 2015, 220-221).

The identification of the HS domain aims to shift the focus from individual disciplines, each with its specificities, to the object of research, namely ‘Heritage’. By changing the perspective, there is greater integration among researchers and contextual improvements compared to individual research projects, thanks to the contribution of multiple expertise as well as a coordinated approach. The need for significant shared data repositories and appropriate management tools has been emphasized by several authors (Kennedy 2015, 224-225; Bordalo, Bottaini, Candeias 2020; Castelli, Felicetti, Proietti 2021). Although this is a common requirement for the entire scientific community, it plays an even more significant role in HS due to the complex articulation of possible fields of investigation and its multidisciplinary nature.

2.2 *Assessing the use of semantic technologies for the Heritage field*

Building on the emerging focus on collaboration and overcoming disciplinary divides within the HS domain, the surge in specialized software and data management tools becomes even more crucial. By leveraging semantic technologies, these tools can bridge the gap between disciplines and facilitate the contribution of multiple areas of expertise to a common object of research.

[1] https://www.e-rihs.eu/e-rihs-in-a-nutshell/.

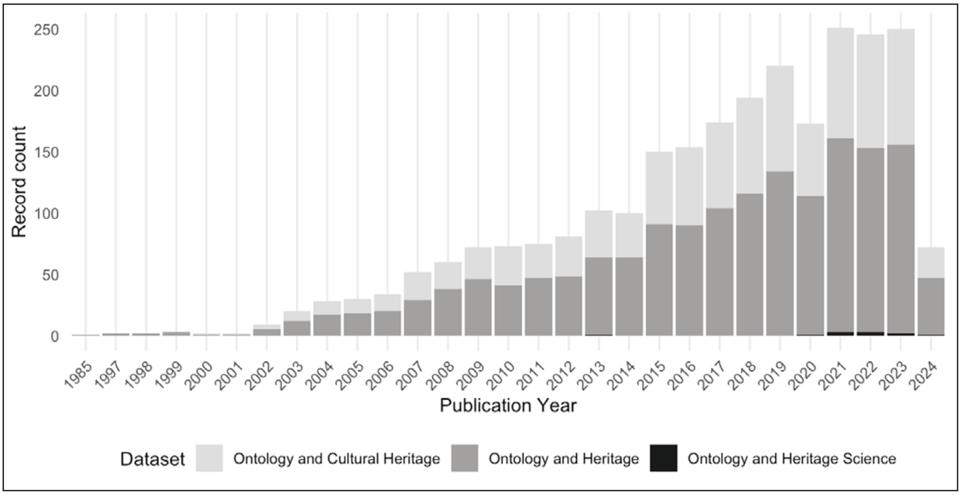

Fig. 1 – Annual distribution of research products (articles, conference proceedings, book chapters) related to the combination of keywords indicated in the legend (data source: Scopus).

However, as the field embraces these advancements, ensuring accessibility and reusability of the data and tools themselves becomes paramount.

With this regard, an initial analysis of relevant literature revealed a rapidly changing landscape (Fig. 1)[2]. Since 2010, controlled vocabularies, taxonomies, ontologies, and specialized software designed for this field became popular, and the use of semantic technologies improved the integration of interoperability principles. Within the domain of semantic tools, which extends beyond Heritage practitioners, a consensus is emerging regarding the necessity for such tools to comply with accessibility and reusability standards, as many resources, particularly ontologies and vocabularies, still fall short of these principles. The recently published report within the FAIR-IMPACT project (Le Franc *et al.* 2020, 11-12), which is part of the initiatives undertaken within the framework of the implementation plan of the European Open Science Cloud (EOSC), highlights the main issues that characterize, for example, the development of ontologies (Garijo, Poveda-Villalón 2020). Often, they lack proper documentation, version control, and are not published and maintained following Linked Data (https://www.w3.org/DesignIssues/LinkedData.html) and FAIR data principles.

[2] Data for all charts and graphs included in the article can be found on Zenodo (https://doi.org/10.5281/zenodo.11388725).

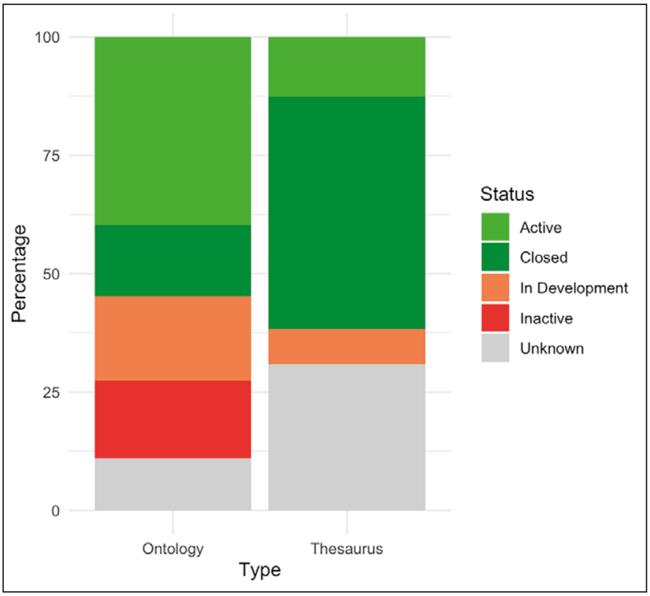

Fig. 2 – Percentage distribution of semantic artefacts cataloged in H-SeTIS according to their curation status (i.e., completeness and maintenance).

More concerningly, many resources lack essential metadata regarding their creation, purpose, usage, and maintenance. This absence of information significantly hinders both understanding and reusability. Garijo, Poveda-Villalón (2020, 7) recommend including 23 metadata elements within an ontology's documentation, with 12 of these being optional. Notably, a significant portion of these belong to the Dublin Core standard, including dc:title, dc:author, dc:contributor, and dc:description. Unfortunately, most resources fail to provide metadata related to creation and modification dates, namespaces, and bibliographic references. Versioning, a fundamental tool in digital tool development, allows tracking a resource's evolution over time and identifying stable versions. However, versioning information is often missing from most of the resources' URIs and, when present, it rarely adheres to semantic versioning principles (https://semver.org/).

The Heritage sector mirrors this broader trend: in-depth documentation exceeding essential metadata remains scarce for the ontologies and the tools cataloged thus far. Most only present traditional documentation: the scientific development process for a semantic resource and the final product are often published in academic journals or conference proceedings. However, the lack of proper documentation and the inability to locate them using unique identifiers (URIs) significantly hinders or even impedes their reuse. It is worth noting that among the roughly 200 semantic tools cataloged so far, a significant portion have a 'negative' status, indicating incompleteness, lack of updates, or irretrievability (Fig. 2).

Recognizing the critical need for well-documented and accessible tools, the H-SeTIS database (Heritage - Semantic Tools and Interoperability Survey, see below §3) aims to create a collection of existing semantic and interoperable resources for the Heritage domain and will serve as an up-to-date toolkit for developing similar tools. Such overview plays a crucial role in creating a knowledge model specifically designed to integrate E-RIHS data into the H2IOSC semantic framework, thereby promoting interoperability and data integration within the Heritage domain.

This state-of-the-art review informed key decisions about the H-SeTIS database design and data entry process. The adoptions of platforms like the Linked Open Vocabularies (LOV) catalog and OntoPortal software are hampered by the limitations of the data itself. LOV (Vandenbussche *et al.* 2017) offers robust functionalities for visualizing semantic vocabularies. OntoPortal (Jonquet *et al.* 2023) focuses on creating comprehensive catalogs of diverse semantic resources. Both leverage semantic technologies to automatically generate metrics and statistics. However, the fragmented nature of information associated with many cataloged resources in H-SeTIS hinders their effectiveness in this context.

In addition, H-SeTIS goes beyond simply cataloging existing semantic tools in the Heritage sector. It also aims to capture information and data about the scientific process behind their creation, even if these do not meet the minimum quality requirements of accessibility and reusability previously mentioned. This additional objective provides valuable insights into the development process of these tools, even if they may not be fully functional or well-maintained.

2.3 *Geospatial analysis of institutions involved in the research of semantic tools for Heritage*

As a preliminary overview of the international scientific community involved in the development of semantic tools within the field of Heritage, a geospatial analysis focused on the involved institutions was carried out using available repositories.

The initial bibliographic dataset was obtained from Web of Science (WoS), a repository selected for the completeness of the raw data made available to the user. The query compiled in the 'Topic' field, which includes titles, abstracts, and author keywords, is as follows: TS=(heritage) AND (TS=(ontolog*) OR TS=("semantic web")); the term 'heritage' was selected to maximize results related to Heritage (Huang 2024): the English term stresses the importance of inheritance passed down through generations, encompassing both Cultural and Natural Heritage. The query therefore returned bibliographic records in which the term 'heritage' was associated with 'ontolog*' (which includes both 'ontology' and 'ontologies') or alternatively with 'semantic web' in titles, abstracts, or keywords. Associating these terms was necessary to disambiguate queries that could have been misleading. For example, the term 'ontology' alone returns hundreds of results dealing with ontologies from a strictly philosophical point of view or, conversely, with computer ontologies applied to any possible theme. The association with the terms 'heritage' and 'semantic web' allowed a more relevant dataset to be obtained, increasing the focus on the Cultural Heritage domain – the main focus of the H2IOSC Project – and retrieving those contributions that also deal with semantic technologies for Heritage, resulting in a final dataset of 1248 bibliographic records[3].

The obtained dataset was used to carry out a quantitative network analysis focused on the institutions involved in the research on ontologies and semantic tools for the Heritage. This dataset was initially preprocessed using VOSviewer software, specifically designed for network analysis and clustering of bibliometric data (van Eck, Waltman 2010). Institutions were selected based on authors' affiliations, with the requirement that they had relations with each other and appeared at least twice to refine the results further. The institutions thus filtered count 128. VOSviewer automatically processed also the related network of relations, connecting institutions based on their appearance in scientific contributions with authors from different affiliations.

To conduct more in-depth analyses, the output data from VOSviewer were subsequently post-processed using Gephi, R, and QGIS software. In Gephi, the network of relations was regenerated, refining it by combining institutions that were considered separate by VOSviewer due to discrepancies in the input data; names were then normalized to overcome any residual ambiguities. With the tidygeocoder package in R (Cambon *et al.* 2021), it was possible to automatically retrieve the geographical coordinates of individual institutions, except for a few cases where manual intervention was necessary. The obtained data were then loaded into a GIS environment, where the initially generated network from VOSviewer was georeferenced, allowing for further advanced analyses (Fig. 3).

[3] The access to large bibliographic repositories is of great assistance to research, with the awareness, however, that the provided data are not always complete or entirely accurate. A comparison of results from various queries has highlighted how difficult it is to extract completely reliable datasets, except perhaps in extremely delimited disciplinary sectors, and how the search system itself likely applies optimizations on results that are hardly controllable. The query used here is the one that has proved to be more flexible compared to others that are more restrictive and accurate, albeit at the cost of omitting a large amount of data. However, it is possible that not all resulting institutions are actually engaged in the field of ontologies and semantic web for Cultural Heritage.

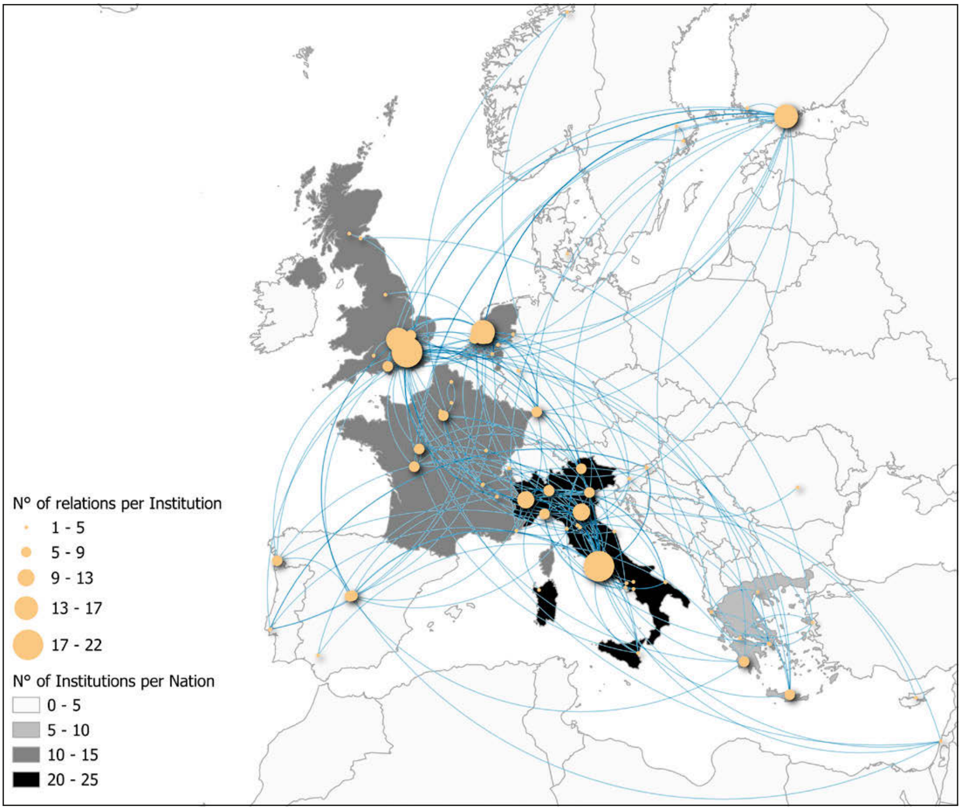
N° of relations per Institution
1 - 5
5 - 9
9 - 13
13 - 17
17 - 22
N° of Institutions per Nation
0 - 5
5 - 10
10 - 15
20 - 25

Fig. 3 – Spatial distribution and relations of the institutions involved in the research of ontologies and semantic web for Heritage; the size of nodes is proportional to the number of relations.

From the observation of the obtained result, several preliminary considerations can be drawn. The majority of research institutions involved in the development of ontologies for Cultural Heritage are located in Europe (about 84%)[4], while among the remaining countries, institutions from the United States, Australia, Israel, Qatar, Pakistan, China, Hong Kong, Taiwan, South Korea, and Vietnam are represented. Among European countries, the most involved one in research activities, as appreciable by the number of its involved institutions, is certainly Italy, with almost double the number of entities compared to the Netherlands, which is second, with France, the

[4] In this case, when referring to Europe, the continent is meant rather than the political Union, thus including countries such as the United Kingdom, Switzerland, Norway, and the Mediterranean basin with Israel.

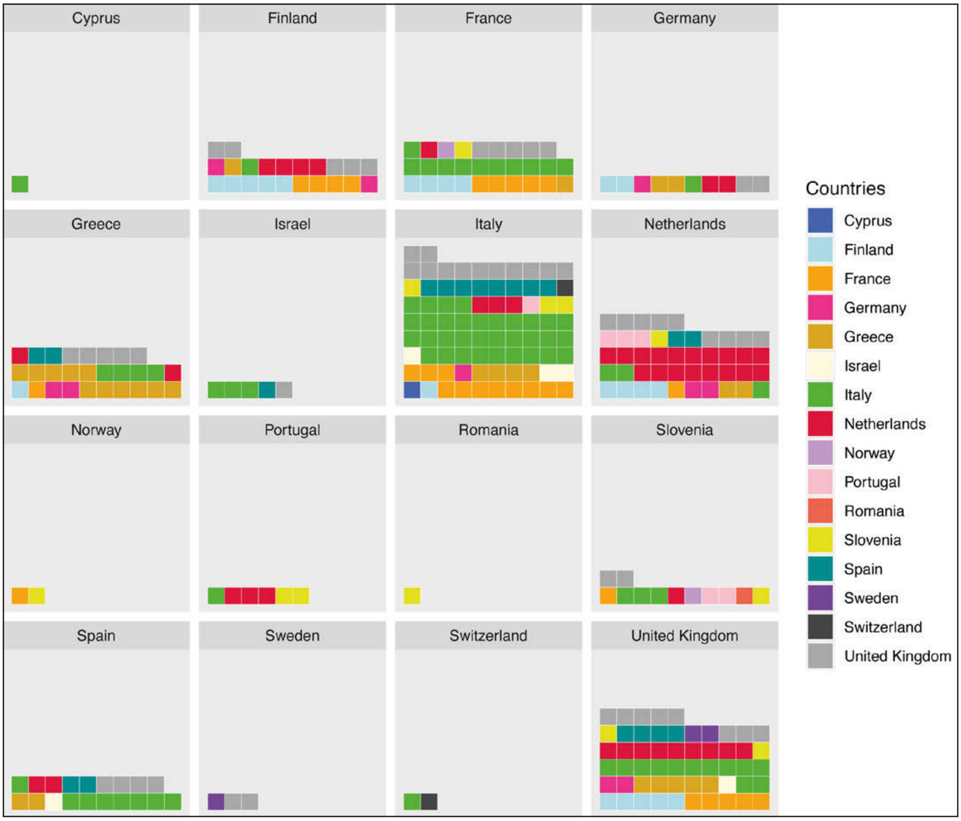

Fig. 4 – Waffle chart displaying national and international collaborations among the institutions mapped in Fig. 3.

United Kingdom, and Greece following. Interconnections between institutions from the same country, especially among those from Italy, the Netherlands, and the United Kingdom, are quite evident. In terms of relationships between individual institutions, the highest number of collaborations is achieved by the CNR, followed by the University College London, the Vrije Universiteit Amsterdam, the Open University (UK), and the Aalto University (Finland). The most significant collaborations at the level of individual countries are found between Italy and the United Kingdom, Italy and France, the United Kingdom and the Netherlands, and Italy and Spain.

From this analysis, it appears quite evident that Italy stands out as the country that contributes the most, at least in terms of number of institutions and collaborations, to research in the ontological field for the

Cultural Heritage domain. The debate itself appears, at present, substantially Eurocentric. It is also possible to assert that there are countries where the number of collaborations among national institutions is considerably high, as is the case with Italy, the Netherlands, and Greece, while for other nations this ratio is more skewed towards relationships with foreign institutions (Fig. 4).

2.4 *Future steps: towards an ontology for Heritage Science*

At present, HS is focused on tangible Cultural Heritage of any nature and scale, from individual objects to landscapes; however, future expansion to intangible Heritage is not to be excluded (Skublewska-Paszkowska *et*

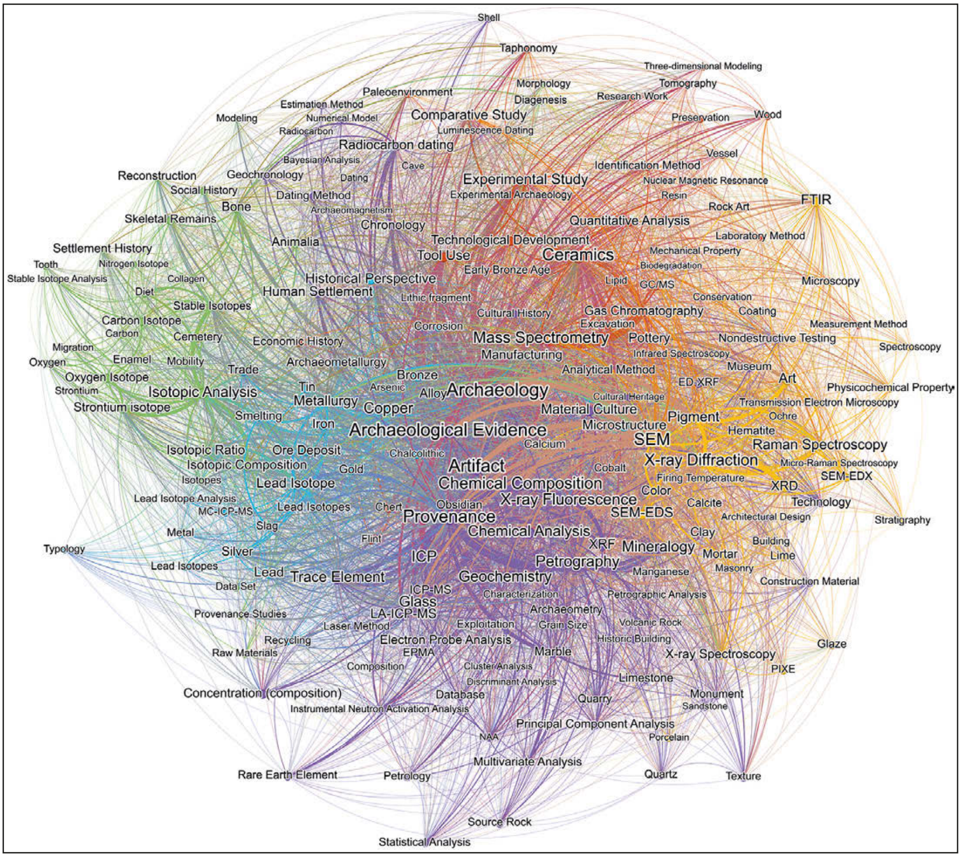

Fig. 5 – Keyword network for the journal «Archaeometry» (2242 articles, 1958-2024), generated with VOSviewer and reprocessed with Gephi, spatialized using Force Atlas 2. Color is based on VOSviewer clustering, while node size is based on Eigenvector centrality factor.

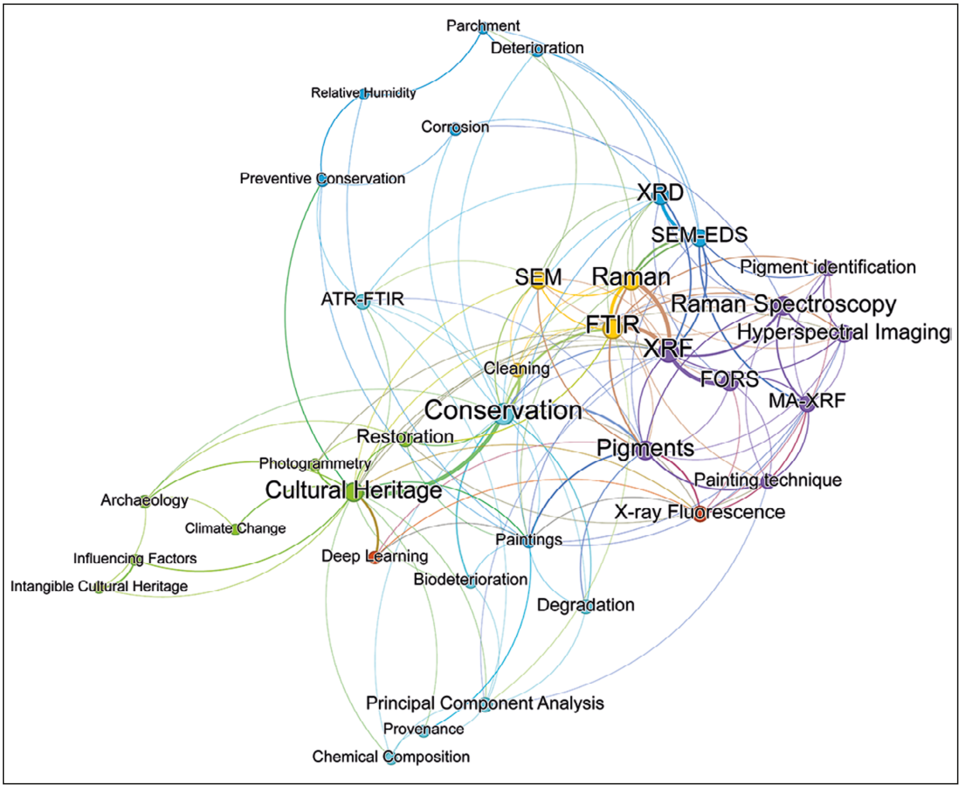

Fig. 6 – Keyword network for the journal «Heritage Science» (1264 articles, 2013-2024), generated with VOSviewer and reprocessed with Gephi, spatialized using Force Atlas 2. Color is based on VOSviewer clustering, while node size is based on Eigenvector centrality factor.

al. 2022). The more theoretical aspects of HS are still under development (Strlič 2018); therefore, the design of a dedicated ontology can significantly contribute to the formal definition of the discipline itself. As already observed, there is still no conceptual model that formalizes the domain related to diagnostics for cultural Heritage, and the data produced are often non-interoperable (Castelli, Felicetti, Proietti 2021, 280). There are resources from related fields and disciplines that participate in HS, although they are not specific to this particular domain, which can be reused, such as geographic and spatial resources, but none of these present specific features for this domain.

The lack of semantic web technology-related keywords in relevant literature underscores the current state of the field. Analyses of author-chosen keywords from two prominent journals, «Archaeometry» (Fig. 5) and «Cultural

Heritage» (Fig. 6), indicate a complete absence of semantic technologies. The dataset, gathered from Scopus using the ISSN of each journal, was visualized through network processing by VOSviewer with Gephi.

The commitment of the CNR to the development of a formal ontology for HS and its computer implementation, a challenging endeavor that is by its very nature ongoing, is based on its multi-decadal experience in the specific field, with coordination from the Italian node of E-RIHS. As previously stated in 2.3, CNR is already one of the mostly involved institutions in terms of national and international collaborations for what concerns this research field. In this regard, it is interesting to recall also the programmatic document of the Science and Technology Select Committee, which coined the term 'Heritage Science' in 2006 and pointed to CNR as a virtuous example of a research institute where basic and applied research are «(...) inextricably intertwined» (House of Lords Science and Technology Select Committee 2006, 24).

3. H-SeTIS database: the resource hub

3.1 *Preliminary conceptual structure*

The H-SeTIS database centers around five key objects: 'ontologies', 'metadata standards', 'thesauri', 'application profiles', and 'software'. While software itself does not constitute a 'semantic artefact' (see below), its inclusion within the cataloging framework is warranted due to its role in facilitating the implementation of the aforementioned four items. Collectively, these resources are referred to as 'semantic tools', reflecting their capacity to structure and enable the representation of knowledge in a machine-understandable format.

The term 'semantic artefact' emphasizes the ability of ontologies, metadata standards, thesauri, and application profiles to be processed by computers (Le Franc *et al.* 2020, 11-17). They represent the latest stage in the evolution of Knowledge Organization Systems (KOS). While many contemporary KOS applications are digital, the concept encompasses a broader range of tools, both physical and digital, designed to organize knowledge (Hodge 2000, 5). Classic examples of KOS include the Linnaean taxonomy for classifying animals and the Dewey Decimal Classification system, both predating modern computers. Despite their long history, the scientific community continues to debate the precise definitions and classifications of these tools, regardless of their machine readability (Souza, Tudhope, Almeida 2012). Hodge (2000, 4-5), for example, distinguishes between controlled vocabularies (authority files, glossaries, dictionaries, gazetteers), classifications and categorizations (subject headings and taxonomies), and lists of relationships (thesauri, semantic networks, ontologies). Hedden (2010, 1-15) prefers to use the single term 'taxonomies' distinguishing between controlled vocabularies, hierarchical taxonomies, thesauri, and ontologies.

For this reason, terms like taxonomy, thesaurus, ontology, and controlled vocabulary, although all falling under the umbrella of KOS, are frequently used interchangeably to describe various forms of knowledge representation. The classification and description of these models also exhibit significant variation in the literature due to the frequent entanglement of their characteristics, objectives, and specific use cases. Therefore, these entities can be conceptualized as distinct classification systems situated along a spectrum of rising complexity, each possessing unique characteristics and fulfilling diverse purposes (Souza, Tudhope, Almeida 2012, 183).

Drawing upon this overview, H-SeTIS employs a pragmatic and incremental approach to classifying semantic artefacts. This approach prioritizes simplicity and flexibility, enabling it to accommodate the diversity of these tools. This strategy facilitates the addition of new information and the integration of emerging semantic tools, ensuring the long-term scalability of the classification system. Aligned with FAIR principles and the Linked Data paradigm, H-SeTIS exposes its information through APIs and describes resource attributes by integrating standard metadata schemas such as Dublin Core and schema.org, thereby promoting data interoperability, reusability, and discoverability[5].

3.2 *The five semantic tools: an overview*

As mentioned above, H-SeTIS focuses on cataloging five key types of semantic tools: 'ontologies', 'metadata standards', 'thesauri', 'application profiles', and 'software' (Fig. 7). For each tool, information is provided to assess its compliance with the FAIR principles and its alignment with the Linked Data paradigm. References to research documenting each tool are listed within its record. The relevant bibliography is maintained through a public Zotero group (https://www.zotero.org/groups/5434475), which currently includes about 340 references: H-SeTIS utilizes Zotero's APIs to query the bibliographic data and leverages the unique identifiers provided by Zotero to manage and retrieve the individual references associated with each semantic tool. The project's bibliographic dataset will also be seamlessly integrated into the database's user interface.

Records for thesaurus-type artefacts encompass a diverse group of controlled lists with varying characteristics. An example of authority lists, a special type of controlled vocabulary that lists standardized forms of proper

[5] The numeric IDs referenced in the following notes correspond to the identifier that will appear in the URI of each resource within the H-SeTIS public website.

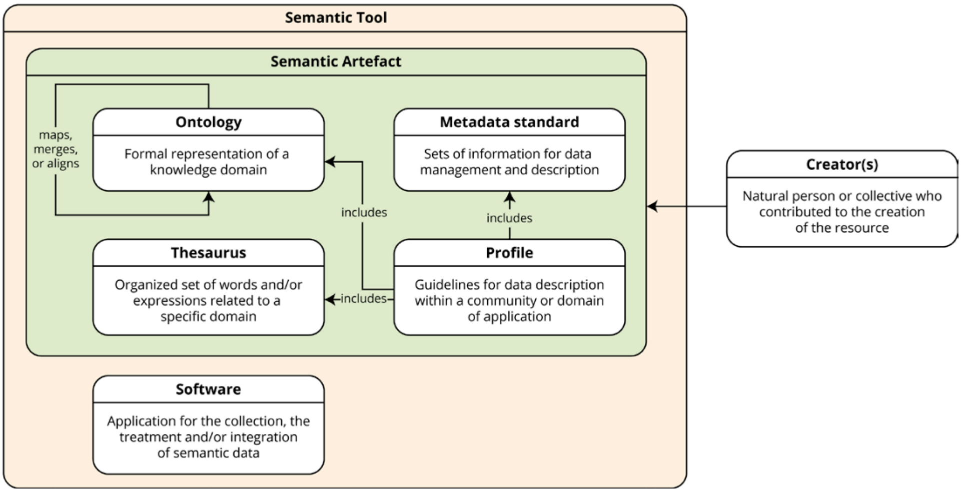
Semantic Tool
Semantic Artefact
maps, merges, or aligns
Ontology
Formal representation of a knowledge domain
Metadata standard
Sets of information for data management and description
includes
includes
Thesaurus
Organized set of words and/or expressions related to a specific domain
includes
Profile
Guidelines for data description within a community or domain of application
Software
Application for the collection, the treatment and/or integration of semantic data
Creator(s)
Natural person or collective who contributed to the creation of the resource

Fig. 7 – Preliminary Conceptual Structure of H-SeTIS.

nouns, is the *Liste d'autorités 'Auteurs'* (ID 109) used to populate the 'Auteur' field of the Joconde database (https://www.pop.culture.gouv.fr/search/list?base=%5B%22Collections%20des%20mus%C3%A9es%20de%20France%20%28Joconde%29%22%5D), which includes proper nouns referring to «une personne physique, un groupe de personnes physiques, une personne morale, une population, une civilisation, etc». Gazetteers, on the other hand, are generally limited to place names only. In this regard, the *Liste d'autorités 'Lieux'* (ID 131) used to populate the 'Lieux' field of the Joconde database falls into the category of gazetteers. As mentioned, it is possible to combine multiple characteristics of these artefacts, as in the case of the *Cairo Gazetteer* (ID 15): the latter presents a list of historical sites in Cairo hierarchically ordered by type and for which coordinates and descriptions are provided. The *Cairo Gazetteer* also presents semantic associations between the terms that compose it, a typical characteristic of thesauri.

Thesaurus artefacts, in the strictest sense of the term, feature a more structured organization and provide detailed information on the relationships between terms. These relationships include hierarchical links, associations (i.e., related concepts), and equivalences (i.e., synonyms). One of the best-known is the *Getty Art & Architecture Thesaurus* (ID 2), in which the terms are hierarchically organized and the most common spelling among various synonyms indicated together with a definition. Another type of thesaurus is the *Digitizing Early Farming Cultures* (ID 23), developed by the Austrian Centre for Digital Humanities for their project and relating to the cataloging of Neolithic and Chalcolithic sites and finds in Greece and Anatolia (ca. 7000-3000 BCE; https://defc.acdh.oeaw.ac.at/). The DEFC presents a hierarchical organization of concepts and includes SKOS broader/narrower relationships, but also horizontal relationships between terms and definitions. It is therefore halfway between a taxonomy and a thesaurus. Compared to traditional taxonomies, thesauri place greater emphasis on the interconnections between terms, offering a more dynamic representation of semantic relationships.

Expanding on the previously mentioned types of thesauri, hierarchical taxonomies are also included in this category. Characterized by a tree-like structure, taxonomies feature broader terms encompassing more specific ones. Similarly to controlled vocabularies, taxonomies are domain-specific but generally simpler than thesauri: they omit equivalence and association relationships, focusing solely on presenting the preferred term chosen by the creators. Interestingly, despite their hierarchical structure, many of the resources described earlier often self-define as thesauri. This highlights the interchangeable use of these terms in practice.

Ontologies represent a further evolution in terms of complexity compared to other semantic artefacts. They incorporate logical relationships between terms to comprehensively represent a domain of knowledge. Consequently, they are classified as a separate artefact from thesauri. Gruber (2009, 1963) defines an ontology as «a set of representational primitives with which to model a domain of knowledge or discourse». Through inference, ontologies allow for reasoning about concepts, extracting new knowledge based on the encoded concepts, relationships, and rules.

Ontologies are also an annotation tool for which reusability is one of the distinguishing features. This promotes efficiency, consistency, and interoperability, and is primarily (but not only) realized through three modalities: aligning, merging, or mapping multiple ontologies. Alignment aims to identify correspondences between different ontologies in an automated or semi-automated manner. The merging method involves creating a new unified ontology by combining elements from multiple source ontologies, while mapping establishes relationships between concepts in different ontologies (Narula *et al.* 2018). In this regard, the H-SeTIS structure allows the ontology-type artefact to be related to itself in order to collect in a structured way the reuses that have occurred between various ontologies.

The CIDOC Conceptual Reference Model (CIDOC-CRM), a core ontology, is the most widely cited ontology in the Heritage sector. A ‘core ontology’ serves to express the basic concepts according to which a domain of knowledge is modeled. Due to this characteristic, it is scalable, meaning that it can be extended as needed. A ‘foundational’ ontology (also defined as ‘upper’ or ‘top-level’) instead models categories so general that they can be considered independent of any specific domain. This category also includes DOLCE

(Descriptive Ontology for Linguistic and Cognitive Engineering), created by the Istituto di Scienze e Tecnologie della Cognizione (CNR-ISTC) to reproduce the ontological categories of natural language and common sense (Gaio *et al.* 2010). The Architecture of Knowledge ontology (ArCO, ID 42) indirectly reuses two light versions of DOLCE, DOLCE-zero and DOLCE+DnS. EpiONT, a specialization of CIDOC-CRM for the epigraphic domain (ID 76), aims to achieve the same result (Cantone *et al.* 2019, 155). The ICON ontology also reuses some DOLCE classes (Sartini *et al.* 2023, 14-15).

As investigated by Moraitou *et al.* (2019, 623-624, Tab.2), CIDOC-CRM has been extended over the years through mapping, merging, or extension. A recent example of mapping is the adaptation of the Italian national standards for coding archaeological information, developed and maintained by the Istituto Centrale per il Catalogo e la Documentazione (ICCD), to CIDOC-CRM within the ARIADNE project (Felicetti *et al.* 2013; Moraitou *et al.* 2019, 617). CIDOC-CRM also has several official extensions, including CRM-archaeo, developed for the conceptual representation of the excavation process and related activities (Christaki *et al.* 2024).

In addition to Knowledge Organization Systems (KOSs), H-SeTIS also encompasses metadata standards, a distinct type of semantic artefact. While KOSs focus on vocabulary control and knowledge representation, metadata standards provide a structured and consistent way to describe various entities, from individual objects to entire databases, and offer a method for organizing, describing, tracking, and ultimately improving access to information (Gilliland 2008, 2-3). Similarly to thesauri and ontologies, the term 'metadata' can carry different meanings: within the Heritage domain, it typically refers to a set of supplementary information designed to organize, describe, track, and enhance access to information about a cultural asset and its associated physical collections. One of the pioneering examples is the schema developed by the Art Museum Image Consortium (AMICO), established by the Association of Art Museum Directors, aiming to standardize and regulate the reuse, distribution, and reproduction of digital images archived in the digital catalogs of various museums. Notably, AMICO, active from 1997 to 2005 was initially mapped by CIDOC-CRM (https://cidoc-crm.org/lrmoo/Resources/the-amico-data-model)[6].

Application profiles, on the other hand, are collections of practices, schemes, and guidelines adopted by a specific community or domain of application to describe a certain type of resource. Essentially, they provide instructions on how to effectively utilize metadata schemes within a particular domain (Baca 2008, 73): the Dublin Core Metadata Initiative Usage Board defines a profile as «a document (or package of documents) which describes

[6] See also the AMICO Data Specification (https://www.amico.org/AMICOlibrary/dataspec.html) and the AMICO Data Dictionary (https://www.amico.org/AMICOlibrary/dataDictionary.html).

a metadata application in order to facilitate broader reuse of its metadata» (https://www.dublincore.org/specifications/dublin-core/profile-review-criteria/).

An example of this type of semantic artefact is the Europeana Collection Profile, which integrates the Europeana Data Model (EDM) by reusing and extending existing classes and properties. As such, application profiles are positioned within H-SeTIS in relation to the other three semantic artefacts (ontologies, thesauri, metadata standards) to map the reuse of existing resources.

H-SeTIS also incorporates data standards alongside application profiles. Data standards function as comprehensive guidelines, defining the structure, formats, ontologies, and vocabularies employed for data management within a specific domain. Compared to application profiles, data standards are more general in scope: an example of a data schema is MIDAS Heritage, which aims to formalize data documentation for historical sites in the United Kingdom (https://historicengland.org.uk/images-books/publications/midas-heritage/midas-heritage-2012-v1_1/).

H-SeTIS further aims to gather information on software that enables the utilization of semantic artefacts, with a preferential but not exclusive focus on software developed in the Heritage sector. Examples of such software include Arches, an open-source platform for managing cultural heritage data (https://www.archesproject.org/), and Omeka S (https://omeka.org/s/), a software for creating virtual exhibitions developed by the Digital Scholar project of the Roy Rosenzweig Center for History and New Media.

Finally, H-SeTIS also includes record types pertaining to resource creators, encompassing both individuals and collectives who contributed to the development of a particular tool. This type of record will enable the construction of an up-to-date overview of the key players actively involved in the creation of semantic resources for the Heritage domain.

3.3 *H-SeTIS preliminary keyword analysis*

Each semantic artefact cataloged in H-SeTIS has been manually associated with one or more keywords. Both the artefacts and keywords were analyzed using Gephi (Fig. 8). The greatest aggregation revolves around the concept of ‘Material Culture’, which is expected given that a large part of Heritage is tangible. The only area currently covered by intangible or immaterial Heritage is essentially music, to which few ontologies and independent thesauri are associated. ‘Material Culture’ also serves as a macro-keyword for other closely associated concepts that specify it, such as ‘Archaeology’, ‘Architecture’, ‘Chronology’, ‘Geography’, ‘History’, ‘Museum Collections’. It is noticeable how archaeology is well-represented at a semantic level, with about a fifth of thesauri and ontologies directly related to the discipline itself, as well as Archaeometry, Epigraphy, Egyptology, Numismatics.

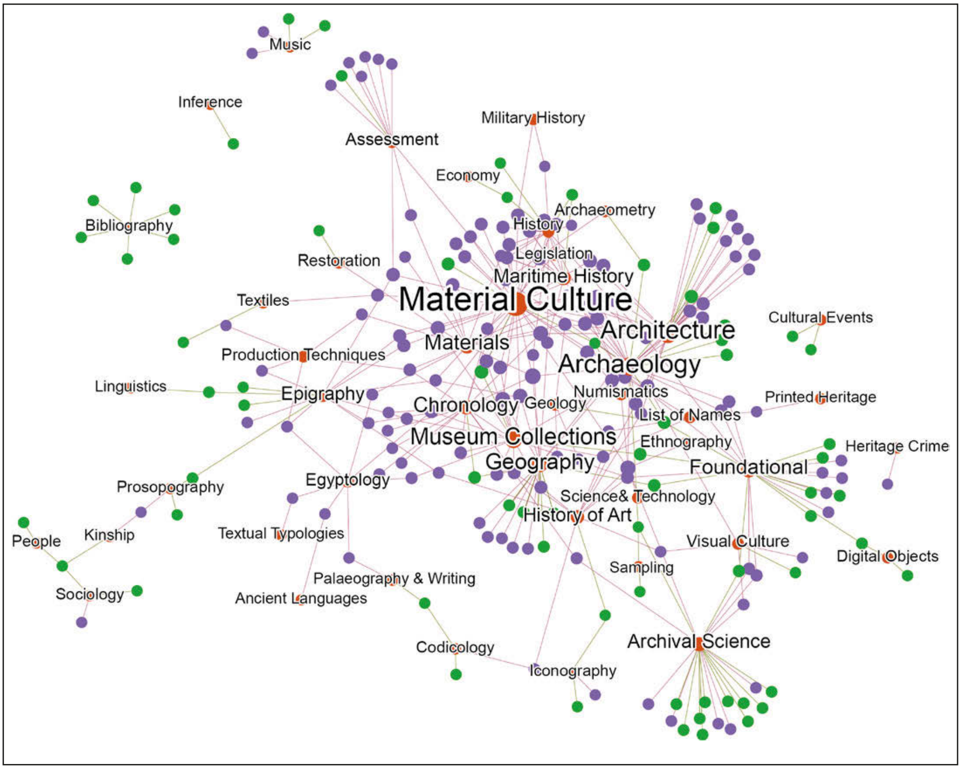

Fig. 8 – Keyword network for the first 200 records added to H-SeTIS. Green nodes refer to ontologies, purple nodes to thesauri, red nodes to keywords.

The clustered spatialization allows for the identification of at least three other subgroups: those of ‘Foundational’ artefacts, those of ‘Archival Sciences’, and those related to ‘Library Science’. The first ones pertain to resources that, despite belonging to different fields, are fundamental for the description of other contents, such as Dublin Core (https://www.dublincore.org/specifications/dublin-core/dcmi-terms/) or the GS84 Geo Positioning RDF schema (http://www.w3.org/2003/01/geo/wgs84_pos#). Resources related to archival sciences are noteworthy both for the number of artefacts and for their relative independence from other keywords, although they are not entirely isolated like bibliographic ones: this, as well as their proximity to ‘Foundational’ resources, is partly due to the reason that a significant portion of ontological and taxonomic activity has traditionally been the responsibility of archivists and librarians, among the categories most involved, especially at a practical level, in organizing knowledge.

4. Final remarks

The H2IOSC Project is going to have a significant impact on HS both theoretically, improving to define its boundaries, and practically, increasing the interoperability and the machine-readability of data. At a preliminary overview, the core institutions engaged in researching semantic artefacts for Heritage are located in Europe and namely in Italy, that plays an important role in the Heritage research as a whole: within this landscape, CNR is actively involved with its long-standing and multidisciplinary expertise on it.

The H-SeTIS database, a central deliverable of the H2IOSC Project, serves as a foundational digital repository for semantic resources in Heritage studies. This includes ontologies, metadata standards, thesauri, application profiles, and software. The database will not only ensure a continued monitoring of these resources but will also facilitate the development of a comprehensive Heritage studies ontology. In this regard, it will provide the essential knowledge for integrating and mapping these resources with existing ones, offering a clear picture of the areas already covered by existing semantic artefacts. From these initial research steps, a more comprehensive HS ontology will be made available to the scientific community, promoting FAIRer data and potentially leading to a wider impact for the entire discipline.

Erica Scarpa, Riccardo Valente
Istituto di Scienze del Patrimonio Culturale - CNR
erica.scarpa@cnr.it, riccardo.valente@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 H-SeTIS implementation benefited from the technical support of the IT specialist Matteo Gallo, who was entrusted with the software development service within this endeavor. The authors wish to express their gratitude to Irene Rossi and Chiara Salvador for their insightful comments and editorial assistance.

References

Baca M. (ed.) 2008, *Introduction to Metadata*, Los Angeles, Getty Research Institute.

Bordalo R., Bottaini C., Candeias A. 2020, *A framework design for information management in Heritage Science laboratories*, «Journal on Computing and Cultural Heritage», 14, 1, 1-14 (https://doi.org/10.1145/3417304).

Cambon J., Hernangómez D., Belanger C., Possenriede D. 2021, *Tidygeocoder: An R package for geocoding, R package version 1.0.5*, «Journal of Open Source Software», 6, 65, 3544 (https://doi.org/10.21105/joss.03544).

Cantone D., Nicolosi-Asmundo M., Santamaria D.F., Cristofaro S., Spampinato D., Prado F. 2019, *An EpiDoc ontological perspective: The epigraphs of the Castello Ursino Civic Museum of Catania via CIDOC CRM*, «Archeologia e Calcolatori», 30, 139-157 (https://doi.org/10.19282/ac.30.2019.10).

Carman J., Sørensen M.L.S. 2009, *Heritage Studies – An outline*, in J. Carman, M.L.S. Sørensen (eds.), *Heritage Studies. Methods and Approaches*, London, Routledge, 11-28.

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

Christaki E., Doerr M., Felicetti A., Hermon S., Hiebel G., Kritsotaki A., Masur A., May K., Ore C.-E., Ronzino P., Schmidle W., Theodoridou M., Tsiafaki D. 2024, *Definition of the CRMarchaeo. An Extension of CRMbase to support the archaeological excavation process* (https://www.cidoc-crm.org/crmarchaeo/sites/default/files/CRMarchaeo_v2.1%28site%29.pdf; last accessed 29/05/2024).

Felicetti A., Scarselli T., Mancinelli M.L., Niccolucci F. 2013, *Mapping ICCD archaeological data to CIDOC-CRM: The RA Schema*, in V. Alexiev, V. Ivanov, M. Grinberg (eds.), *CRMEX 2013 Practical Experiences with CIDOC CRM and its Extensions*, CEUR WS, 11-22.

Gaio S., Borgo S., Masolo C., Oltramari A., Guarino N. 2010, *Un'introduzione all'ontologia DOLCE*, «AIDA informazioni: Rivista di scienze dell'informazione», 28, 107-125 (https://doi.org/10.1400/212462).

Garijo D., Poveda-Villalón M. 2020, *Best Practices for Implementing FAIR Vocabularies and Ontologies on the Web* (https://doi.org/10.48550/arXiv.2003.13084).

Gilliland 2008, *Setting the stage*, in M. Baca (ed.), *Introduction to Metadata*, Los Angeles, Getty Research Institute, 1-19.

Gruber T. 2009, *Ontology*, in L. Liu, M.T. Özsu (eds.), *Encyclopedia of Database Systems*, Boston, Springer, 1963-1965 (https://doi.org/10.1007/978-0-387-39940-9_1318).

Hedden H. 2010, *The Accidental Taxonomist*, Medford, Information Today.

Hodge G.M. 2000, *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*, Washington, Digital Library Federation.

House of Lords Science and Technology Select Committee 2006, *Science and Heritage. Report with Evidence*, London, The Stationery Office Limited.

Huang Y. 2024, *Bibliometric analysis of GIS applications in heritage studies based on Web of Science from 1994 to 2023*, «Heritage Science», 12, 1, 57 (https://doi.org/10.1186/s40494-024-01163-y).

Jonquet C., Graybeal J., Bouazzouni S., Dorf M., Fiore N., Kechagioglou X., Redmond T., Rosati I., Skrenchuk A., Vendetti J.L., Musen M. 2023, *Ontology repositories and semantic artefact catalogues with the OntoPortal technology*, in T.R. Payne, V. Presutti, G. Qi, M. Poveda-Villalón, G. Stoilos, L. Hollink, Z. Kaoudi, G. Cheng, J. Li (eds.), *The Semantic Web – ISWC 2023*, Cham, Springer Nature Switzerland, 38-58 (https://doi.org/10.1007/978-3-031-47243-5_3).

Kennedy C.J. 2015, *The role of Heritage Science in conservation philosophy and practice*, «The Historic Environment: Policy & Practice», 6, 3, 214-228 (https://doi.org/10.1080/17567505.2015.1099925).

Kennedy C.J., Penman M., Watkinson D., Emmerson N., Thickett D., Bosché F., Forster A.M., Grau-Bové J., Cassar M. 2024, *Beyond Heritage Science: A review*, «Heritage», 7, 3, 1510-1538 (https://doi.org/10.3390/heritage7030073).

Le Franc Y., Parland-von Essen J., Bonino L., Lehväslaiho H., Coen G., Staiger C. 2020, *D2.2 FAIR Semantics: First recommendations* (https://zenodo.org/records/3707985; last accessed 29/05/2024).

Moraitou E., Aliprantis J., Christodoulou Y., Teneketzis A., Caridakis G. 2019, *Semantic bridging of Cultural Heritage disciplines and tasks*, «Heritage», 2, 1, 611-630 (https://doi.org/10.3390/heritage2010040).

Narula G.S., Wason R., Jain V., Baliyan A. 2018, *Ontology mapping and merging aspects in Semantic web*, «International Robotics & Automation Journal», 4, 1 (https://doi.org/10.15406/iratj.2018.04.00087).

Sartini B., Baroncini S., Van Erp M., Tomasi F., Gangemi A. 2023, *ICON: An ontology for comprehensive artistic interpretations*, «Journal on Computing and Cultural Heritage», 16, 3, 1-38 (https://doi.org/10.1145/3594724).

Skublewska-Paszkowska M., Milosz M., Powroznik P., Lukasik E. 2022, *3D technologies for intangible cultural heritage preservation. Literature review for selected databases*, «Heritage Science», 10, 1, 3 (https://doi.org/10.1186/s40494-021-00633-x).

Souza R.R., Tudhope D., Almeida A.M.B. 2012, *Towards a taxonomy of KOS: Dimensions for classifying knowledge organization systems*, «Knowledge Organization», 39, 3, 179-192 (https://doi.org/10.5771/0943-7444-2012-3-179).

Strlič M. 2018, *Heritage Science: A future-oriented cross-disciplinary field*, «Angewandte Chemie International Edition», 57, 25, 7260-7261 (https://doi.org/10.1002/anie.201804246).

van Eck N.J., Waltman L. 2010, *Software survey: VOSviewer, a computer program for bibliometric mapping*, «Scientometrics», 84, 2, 523-538 (https://doi.org/10.1007/s11192-009-0146-3).

Vandenbussche P.-Y., Atemezing G.A., Poveda-Villalón M., Vatant B. 2017, *Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the web*, «Semantic Web», 8.3, 437-452 (https://doi.org/10.3233/SW-160213).

Wilkinson M.D., Dumontier M., Aalbersberg Ij.J. *et al.* 2016, *The FAIR Guiding Principles for scientific data management and stewardship*, «Scientific Data», 3, 1, 160018 (https://doi.org/10.1038/sdata.2016.18).

ABSTRACT

This article explores the contributions of the Milan branch of CNR-ISPC to the Humanities and Cultural Heritage Italian Open Science Cloud (H2IOSC) Project, focusing on facilitating data integration within Heritage Science. Its primary objective is to ensure seamless interoperability between resources from multiple institutions by establishing a shared semantic framework. The multidisciplinary nature of Heritage Science underscores the necessity for shared data repositories and effective management tools. Recent literature highlights the importance of semantic technologies in improving data integration and interoperability. To this end, the H-SeTIS database is currently under development. H-SeTIS will function as a hub for the systematic surveying and description of various semantic tools relevant to the Heritage domain. Interestingly, a preliminary analysis of data within H-SeTIS reveals that many semantic resources specifically designed to address the unique requirements of the Heritage domain do not meet the minimum quality requirements of accessibility and reusability. This finding underscores a potential area for future development: the creation of H-SeTIS aims to support the ongoing development of a comprehensive ontology for Cultural Heritage, enhancing data FAIRness and the discipline's overall impact.