ATLAS: Towards a Knowledge Graph of International Scholarly Research on the Italian Digital Cultural Heritage

Sebastiano Giacomini[1,*], Alessia Bardi[2], Marina Buzzoni[3], Marilena Daquino[1], Riccardo Del Gratta[4], Angelo Mario Del Grosso[4], Franz Fischer[3], Chiara Martignano[3], Roberto Rosselli Del Turco[5], Giorgia Rubin[4] and Francesca Tomasi[1]

[1]*Department of Classical Philology and Italian Studies - University of Bologna, Bologna, Italy*
[2]*Institute of Information Science and Technologies "A. Faedo" - National Research Council, Pisa, Italy*
[3]*Department of Humanities - University of Venice, Venice, Italy*
[4]*Institute for Computational Linguistics "A. Zampolli" - National Research Council, Pisa, Italy*
[5]*Department of Humanities - University of Turin, Turin, Italy*

Abstract

In recent years, the abundance of available scholarly information has requested constant development and revision of standardized models and shared guidelines. Based on these frameworks, the Digital Humanities (DH) landscape features a variety of aggregators expected to enhance research data findability while promoting use and reuse. However, current semantic models fail to capture the specificity of DH research products, hindering data discovery and hampering the valorisation of Cultural Heritage. The ATLAS project addresses these key challenges by developing a unified framework for describing and aggregating scholarly outputs, particularly in the Italian Digital Cultural Heritage domain. This paper presents the initial versions of the ATLAS Ontology and Knowledge Graph, designed to model DH outcomes such as Digital Scholarly Editions, text collections, Linked Open Data, ontologies, and software. In so doing, ATLAS aims to enhance resource findability and reuse, paving the way for improved interoperability and future advancements in the field.

Keywords

Digital Humanities, Knowledge Graph, Semantic Web, Research Infrastructures, Italian Cultural Heritage

1. Introduction[1]

In recent years, the World Wide Web and its technologies have significantly changed how scholarly activities in the Digital Humanities (DH) domain are carried out, offering unprecedented opportunities for preserving, sharing, and reusing research outputs and publications [1, 2]. At the same time, the abundance of available scholarly information has requested constant development and revision of standardised models and shared guidelines. Such frameworks have become the foundation for data aggregators and exploratory environments, such as Europeana and OpenAIRE, designed to collect documents and data from various research settings, including those entirely or partially focused on DH. Many of such initiatives have embraced Semantic Web technologies, particularly Linked Open Data, to unravel the complex relations between scholarly endeavours and Cultural Heritage.

IRCDL 2025: 21st Conference on Information and Research Science Connecting to Digital and Library Science, February 20–21, 2025, Udine, Italy

*Corresponding author.
sebastiano.giacomin2@unibo.it (S. Giacomini); alessia.bardi@isti.cnr.it (A. Bardi); mbuzzoni@unive.it (M. Buzzoni); marilena.daquino2@unibo.it (M. Daquino); riccardo.delgratta@ilc.cnr.it (R. Del Gratta); angelo.delgrosso@ilc.cnr.it (A. M. Del Grosso); franz.fischer@unive.it (F. Fischer); chiara.martignano@unive.it (C. Martignano); roberto.rosselli delturco@unito.it (R. Rosselli Del Turco); giorgia.rubin@ilc.cnr.it (G. Rubin); francesca.tomasi@unibo.it (F. Tomasi)
0009-0007-7813-0939 (S. Giacomini); 0000-0002-1112-1292 (A. Bardi); 0000-0002-9306-6599 (M. Buzzoni); 0000-0002-1113-7550 (M. Daquino); 0000-0003-1867-8445 (R. Del Gratta); 0000-0002-4867-6304 (A. M. Del Grosso); 0000-0002-2162-5531 (F. Fischer); 0000-0002-8984-9574 (C. Martignano); 0000-0002-8945-9314 (R. Rosselli Del Turco); 0009-0008-0584-7290 (G. Rubin); 0000-0002-6631-8607 (F. Tomasi)

[1]Authors' responsibilities: Sebastiano Giacomini is responsible for Sections 2,4; all authors contributed to Sections 1,3,5.

In particular, the DH landscape features a variety of aggregators, each focusing on different aspects of research activities. In so doing, they are expected to enhance data and metadata findability while promoting use and reuse [3]. By aggregating resources, such systems attempt to offer additional research value that conventional forms of retrieval and browsing cannot achieve [4]. However, to the best of our knowledge, despite recent efforts made by cultural institutions, the analysis of the Italian context reveals the lack of a unified research framework for Cultural Heritage and research data discoverability, as well as the lack of a comprehensive catalogue of DH scholarly data [5], and domain-dependent best practices to foster data findability and reusability, ultimately hindering resource discovery. In other terms, (1) representative services for aggregating DH research products are missing, and (2) domain-specific ontologies and vocabularies are not easily adaptable to describe the heterogeneous nature of Digital Cultural Heritage outputs (e.g. digital editions, text collections).

In this article, we present the initial results of the ATLAS project, including the ATLAS ontology, the ATLAS knowledge graph, and the technical requirements of the ATLAS platform. The ATLAS ontology has been developed to meet the main challenges posed by the description of DH research activities and products. These include Digital Scholarly Editions, text collections, Linked Open Data datasets, RDF vocabularies, and software. To populate the ontology and test the proposed model, an initial knowledge graph has been developed by extracting, structuring, and enriching high-quality data from potentially unstructured or semi-structured digital sources. To achieve this aim, the ATLAS project has extended the functionalities of CLEF[2] (Crowdsourcing Linked Entities via Web Form), a collaborative web platform for data entry that facilitates users in LOD collection and visualisation. Among the new features, the latest version of CLEF allows researchers to semi-automatically extract knowledge from various sources, including APIs, SPARQL endpoints, and static files (.csv, .json, and .xml formats) and populate the descriptive record of a research object. To support both the ontology design and the technical requirements of the ATLAS platform, a set of pilot projects on the Italian Digital Cultural Heritage was analysed, and ontological models for describing scholarly data have been reviewed and mapped to highlight classes and properties currently lacking.

The paper is structured as follows. Section 2 examines scholarly aggregators of DH research activities and outputs, with specific considerations on Italian Digital Cultural Heritage, as well as existing semantic models and their main properties, so as to highlight the motivation for our work. Section 3 describes the methodology and approach used to develop the ontology and populate it through a knowledge graph. Section 4 presents the initial versions of both the ATLAS Ontology and the related knowledge graph, including an illustrative example from the described pilot resources. Finally, Section 5 evaluates findings and limitations, and outlines future steps of the ATLAS project.

2. State of the Art

Over the last few years, GLAM institutions (Galleries, Libraries, Archives, and Museums) have increasingly promoted initiatives aimed at sharing their holdings across the web. While these efforts have significantly broadened access to invaluable Cultural Heritage resources, they have also resulted in the proliferation of new models, schemas, and vocabularies, leading to uncontrolled growth of metadata standards across the Web [6]. Amidst this complex and fragmented landscape, a number of aggregators have recently emerged, highlighting the fundamental role of such services in providing homogeneous access to heterogeneous (meta)data collections [7].

Within the Italian scenario, institutions have invested in digitising and aggregating cultural holdings, making them available as Linked Open Data collections. Projects like dati.culturaitalia[3], the Linked Open Data platform by the Italian Ministry of Culture, exemplify the recent commitment to making Italian Cultural Heritage data interoperable with some prominent digitisation efforts within the European landscape, including ARIADNE and Europeana [8]. Similarly, the ArCO[4] project has developed a

[2] https://polifonia-project.github.io/clef/.
[3] https://dati.culturaitalia.it.
[4] https://w3id.org/arco/.

Knowledge Graph from the General Catalog of Italian Cultural Heritage, offering reusable Linked Open Data collections based on the official institutional database of Italian Cultural Heritage [9]. Despite these efforts and other limited initiatives for collecting DH research data[5], there are either no representative, comprehensive catalogues tailored to DH projects, or they do not allow the retrieval of research products on the Italian Cultural Heritage. Additionally, no structured collections on DH projects and artefacts leveraging Semantic Web technologies are available [14]. The broader scholarly landscape presents several platforms that play a crucial role in providing persistent identification, long-term preservation, and enhanced findability of research data [15]. Prominent services include Zenodo[6] and OpenAIRE[7] [16]. The OpenAIRE network integrates several services, including community web portals like the Digital Humanities and Cultural Heritage gateway[8], which facilitate the discovery and sharing of research outcomes and Open Science practices.

However, despite targeted attempts to highlight DH research activities, aggregators like Zenodo and OpenAIRE serve as broad data collectors on various disciplines, often lacking references to the Cultural Heritage sources that drove the creation of DH scholarly data. In addition, the absence of domain-specific vocabularies hampers the identification of resources produced by DH practices, e.g. digital editions.

At the core of the information retrieval problem outlined above, we find the lack of a comprehensive data model that allows one to describe the peculiarities of the DH research products in the first place. While several data models exist and are shared in the broader scholarly community, they describe research outputs in general terms, without considering the diversity and specificity of DH outputs. Notable examples include the OpenAIRE Graph[9], which provides a Scholarly Knowledge Graph [17] collecting metadata on the following core entities: Research products, Data sources, Organisations, Projects, and Communities. Research products include "Publication", "Data", "Software", and "Other research product". RO-Crate[10] (Research Object Crate) offers another approach for packaging research data along with its metadata and associated component files [18]. RO-Crates are based on the concept of Research Object (RO), defined as a semantically rich aggregation of resources [19], and serve data according to Schema.org[11] in JSON-LD format. The current data model (v1.1.3) distinguishes between *Data entities* (e.g. directories, files) and *Contextual entities* (person, organisations, equipment) [20]. Within this framework, an RO-crate resource is treated as a root data entity with type `schema:Dataset`. The SKG-IF[12] (Scientific Knowledge Graph Interoperability Framework) Working Group has recently developed a metadata model targeting interoperability among Scientific Knowledge Graphs and their usability [21]. The model (v1.1) is structured around six core entities: Research product, Agent, Grant, Venue, Topic, and Data source. Research products are described via the FaBiO Ontology (FRBR-aligned Bibliographic Ontology) [22]; namely, `fabio:Dataset` (research data), `fabio:ScholarlyWork` (literature), and `fabio:Software` (software). Lastly, the KNOT project[13] aims to showcase the Digital Cultural Heritage of Italian universities [23]. The ontology[14] (v1.2) leverages entities from DCAT, PROV-O, and CIDOC-CRM, and the KNOT knowledge graph mainly focuses on Research Projects, Digital Objects (e.g. Datasets, Knowledge Graphs, Ontologies), and Web Services (e.g. Digital Editions, Digital Libraries, Endpoints). However, the model does not focus on identifying Cultural Heritage artefacts, using the generic `dcterms:subject` property to broadly indicate related disciplines and Wikidata keywords. In addition, no information is retrieved directly from available sources (e.g.: datasets, TEI encodings).

[5] These include catalogues of Digital Scholarly Editions [10, 11], heterogeneous projects gathered by national associations (AIUCD), research centres (/DH.arc, VeDPH, DH@FBK), international associations (EADH), disciplinary surveys [12, 13].
[6] https://zenodo.org/.
[7] https://openaire.eu/.
[8] https://dh-ch.openaire.eu/.
[9] https://graph.openaire.eu/.
[10] https://researchobject.org/ro-crate/.
[11] https://schema.org/.
[12] https://skg-if.github.io/.
[13] https://projects.dharc.unibo.it/knot/records.
[14] http://purl.org/knot/ontology.

In conclusion, despite such remarkable achievements, the models fall short of addressing all complexities set by the current DH landscape. Even advanced schemas, such as OpenAIRE Graph and SKG-IF, which introduce higher levels of granularity, fail to capture the heterogeneity of research outputs in the Digital Cultural Heritage domain. In fact, diverse projects can result in a variety of outcome types –such as text collections, Digital Scholarly Editions, Linked Open Data datasets, RDF vocabularies, and software–, each of which deserves to be described accordingly. Firstly, specialised terminologies are needed to identify the different products, particularly those peculiar to Digital Cultural Heritage, such as digital textual archives and Digital Scholarly Editions. Secondly, the existing models lack semantic attributes and controlled resources designed to adequately describe the methodological aspects of DH research. Crucial issues, such as textual typologies and edition criteria, which are critical for a comprehensive representation of peculiar outcomes and research practices, remain insufficiently addressed. Lastly, existing models do not provide adequate solutions for linking research activities to their corresponding Cultural Heritage objects, despite the potential offered by Linked Open Data. This results in two main consequences, namely: (1) it limits users and researchers in discovering products and perspectives on Digital Cultural Heritage resources, and (2) hinders Cultural Heritage resources retrieval and valorisation.

Further limitations derive from services and websites that do not include such information when providing access to research products metadata. These shortcomings affect both the data collection processes, due to the lack of suitable tools for extracting meaningful entities from available resources, and the dissemination stage, where the absence of dedicated systems for data visualisation hampers discovery. To address the challenges, the CLEF application is actively working on developing novel solutions, including data entry and exploration services such as Intermediate Templates, Advanced Knowledge Extraction, and Data Visualisation tools.

While hindering findability, current limitations prevent serendipitous discoveries and limit the effective reuse of research outputs in Humanities research. Bridging this gap requires the development of a semantic model that accommodates the diversity of DH outputs while facilitating the integration of Cultural Heritage metadata into services. To this extent, existing software solutions for cataloguing scholarly data lack the means to (1) leverage complex data models, and (2) automatically extract information from data sources (e.g. extracting the Cultural Heritage resources mentioned in a research product). Moreover, (3) they lack web-based solutions for performing data analysis without requiring users' advanced technical skills [24, 25].

3. Methodology and Approach

The ATLAS project has investigated some pilots, representative of Italian DH projects and resources [14] to classify them into five main groups, namely:

- Text collections: ALIM (Archive of the Italian Latinity of the Middle Ages); Biblioteca Italiana; BUP - Digital Humanities; Musisque Deoque
- Digital Scholarly Editions: VaSto (VArchi STOria fiorentina); Codice Pelavicino Digitale; Leges Langobardorum; Digital Edition of Aldo Moro's works
- Software: EVT (Edition Visualisation Technology); Voyant Tools
- Linked Open Data: Zeri & LODE; DanteSources; LiLa - Linking Latin; Biflow - Toscana Bilingue Catalogue
- Ontologies: CIDOC-CRM; SPAR; HiCO

Pilots served two main purposes, namely (1) identifying essential metadata for building the ATLAS catalogue and its semantic model, and (2) validating and populating the ontology with scholarly data resulting in a knowledge graph. Additionally, this analysis also aimed to produce a set of guidelines to help improve data management practices in the Digital Humanities projects.

The results of the pilot analysis offered an initial base for evaluating existing standards for the description of research products. Metadata from pilot projects were systematically collected, assigning

a label and corresponding values to each piece of information. Labels provided a starting point for a preliminary mapping of existing data models and frameworks, enabling a semantic alignment and arrangement of identified metadata. Detailed mapping tables are provided in the supplementary materials of the ATLAS Ontology and include the following vocabularies and frameworks: RO-Crate[15], KNOT[16], OpenAIRE Graph[17], OpenAIRE Application Profile[18], SKG-IF[19], IRIS[20].

The preliminary analysis revealed the need for a novel data model capable of addressing the current issues highlighted in the state of the art, ensuring a nuanced representation of research outputs, enhancing metadata completeness, and improving accessibility. The resulting ATLAS Ontology[21] imports several models. The backbone is based on classes and properties from Schema.org (v28.0)[22], a vocabulary that has already proved to be suitable for describing and aggregating Cultural Heritage objects metadata [26]. However, the complexity of the Digital Cultural Heritage research domain required integrating other models, particularly those offering granularity concerning the DH domain. Among these, particular attention was paid to FaBiO, the FRBR-aligned Bibliographic Ontology [22], and DC Terms[23], both suggesting the importance of working on multiple levels of cultural objects [27]. To test and validate the newly created model, metadata collected from the preliminary analysis of pilot resources were reused to develop a first Knowledge Graph populating the novel ontology. In this stage, the CLEF web application [24] provides users with a system to verify the adequacy of the semantic schema and to streamline data entry activities. CLEF supports the collaborative creation of Linked Open Data collections through customisable "Templates" corresponding to ontological classes and rendered as user-friendly Web Forms. The platform's key features, including automatic Entity Reconciliation and Knowledge Extraction features, enable the development of a Knowledge Graph of interlinked Records, managed by the Blazegraph[24] triplestore and simultaneously serialised in Turtle format for milestones data publication and versioning purposes.

To meet the granularity requirements of the ATLAS Ontology and make proper use of the content in available resources (e.g. datasets, TEI documents), ATLAS worked on extending CLEF functionalities. This effort focused on three key areas, namely: innovative solutions for representing complex data models in data entry, streamlining data entry processes, and providing data processing tools to enhance user experience and catalogue exploration and visualisation.

4. Results

4.1. ATLAS Ontology v1.0

The ATLAS Ontology is an OWL 2 DL ontology[25] [28] designed to effectively represent scholarly research projects on the Italian Cultural Heritage and their outcomes. Its primary goal is to describe features of DH research products, highlighting their unique attributes to the broader landscape of scholarly artefacts. As aforementioned, the ATLAS Ontology leverages terms from different existing models to facilitate the alignment between the ATLAS catalogue and existing data sources. Schema.org (prefix `schema`, https://schema.org) serves as the backbone of the vocabulary, and it is enriched with terms from DCTerms, and FaBiO (prefix `fabio`, http://purl.org/spar/fabio/) [22]. To enhance granularity and be representative of the terminology used by practitioners in the DH, ATLAS has also introduced new Classes and Properties (prefix `atlas`, https://w3id.org/dh-atlas/), aligned to existing models. In

[15] https://w3id.org/ro/crate/1.1.
[16] http://purl.org/knot/ontology.
[17] https://graph.openaire.eu/docs/.
[18] https://openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/v4.0.0/.
[19] https://w3id.org/skg-if/context/docs/skg-if.json.
[20] https://wiki.u-gov.it/confluence/display/public/UGOVHELP/IRIS+-+Institutional+Research+Information+System.
[21] https://w3id.org/dh-atlas/.
[22] https://github.com/schemaorg/schemaorg/tree/main/data/releases/28.0/.
[23] http://purl.org/dc/terms/.
[24] https://blazegraph.com/.
[25] https://w3id.org/dh-atlas/.

Figure 1 we show an overview of classes and properties.

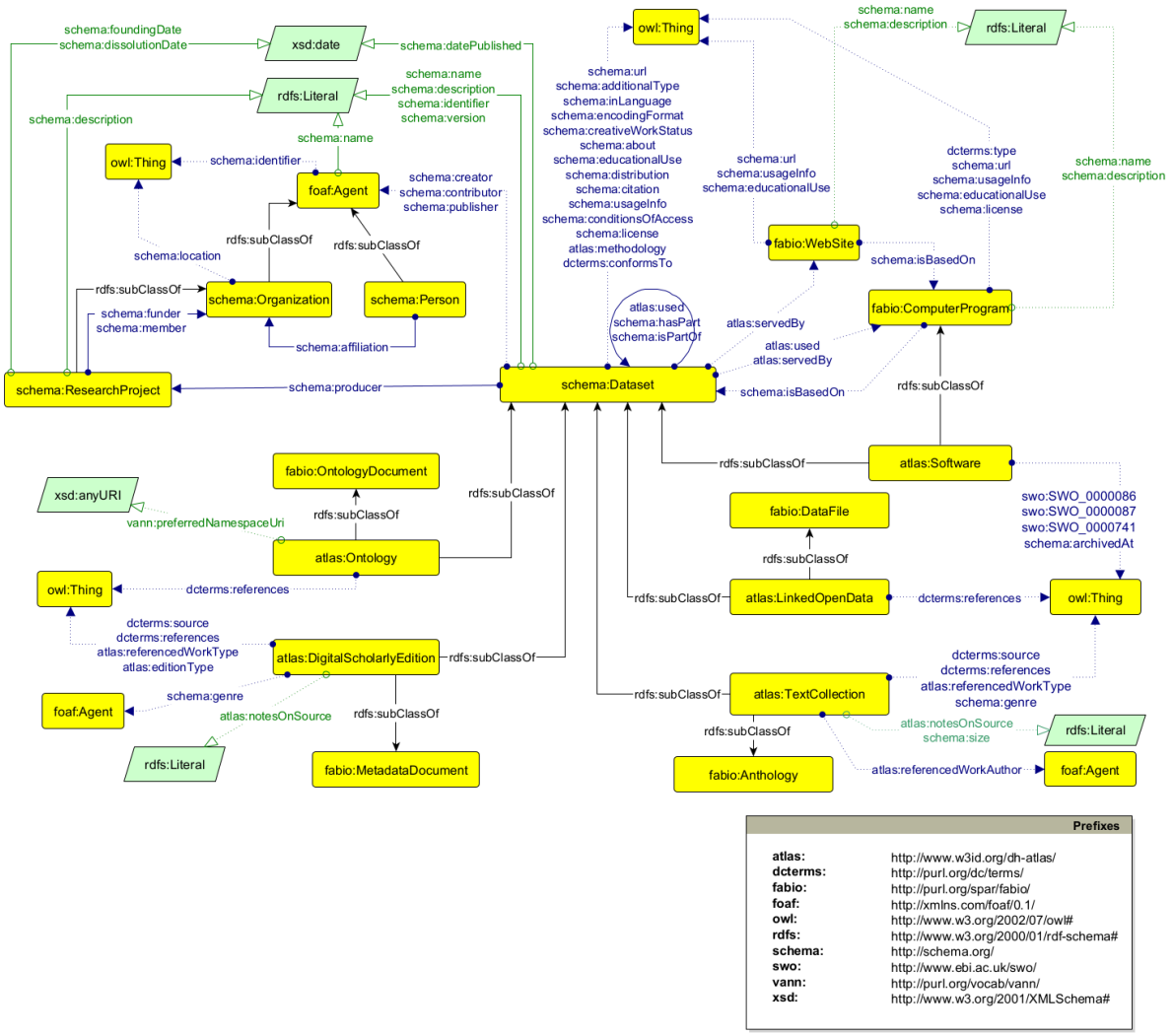

Figure 1: A visual diagram of the ATLAS Ontology: classes and properties.

Research Product The results of research activities are first-class entities in many reviewed models. The ATLAS Ontology follows this common and makes research products the core of the new vocabulary, represented by the class `schema:Dataset`, a subclass of `schema:CreativeWork`. While Schema.org broadly defines `schema:Dataset` as any "*body of structured information describing some topic(s) of interest*", additional specifications clarify its intended applications [29]. Usage examples include collections of packaged data, such as those "*published in scientific, scholarly or governmental open data repositories*", as well as "*data that is stored in collections of spreadsheet files, or as digital images, or in dedicated scientific, geospatial and engineering file formats*".

To better frame the nature of scholarly outcomes in the DH, the property `schema:additionalType` allows us to associate instances of `schema:Dataset` with subclasses of the class `frbr:Expression`, namely: `atlas:TextCollection`, `atlas:DigitalScholarlyEdition`, `atlas:LinkedOpenData`, `atlas:Ontology`, and `atlas:Software`.

Depending on the associated class, additional properties can be used to describe scholarly products. In ATLAS we distinguish artefact-dependant properties from general properties. General properties include information such as the title (`schema:name`), a description (`schema:description`), the release date (`schema:datePublished`), the current version (`schema:version`), the current work

status (schema:creativeWorkStatus), external identifiers (schema:identifier), the resource link (schema:url), and links to distributions (schema:distribution).

Further details focus on the technical content of the resource, such as the subject matter (schema:about), used languages (schema:inLanguage), the encoding format (schema:encodingFormat), bibliographic references (schema:citation), adopted standards (dcterms:conformsTo), and documentation web pages (schema:usageInfo).

To refine the description of DH artefacts and allow a more practical use of ATLAS cataloguing data, two properties describe the research activities afforded by the research product (schema:educationalUse) and those performed during the production of the outcome at hand (atlas:methodology): in both cases, values are expected to be taken from the TaDiRAH[26] taxonomy. The properties schema:license and schema:conditionsOfAccess are expected to provide information on the license and access rights respectively.

Relations between artefacts and people/organisations, i.e., instances of the class foaf:Agent, include authors (schema:creator), contributors (schema:contributor), publishers (schema:publisher), and the Research Project the object is a result of (schema:producer). Relations between Research Products can be expressed through schema:hasPart, schema:isPartOf, and atlas:used, the latter specifying external resources reused to generate the product although not being part of it. At the same time, the atlas:isServedBy property introduces those services and tools that make available the content of the Research Products (e.g. Visualisation Software, SPARQL endpoints). In Table 1, we summarise properties associated with the five classes defined in the ATLAS Ontology.

Table 1
Classes and properties for describing Research Products in ATLAS

ATLAS Type	RDF Property	Property Description
atlas:TextCollection, atlas:DigitalScholarlyEdition	dcterms:source	The cataloguing record of the main edited work(s)
atlas:TextCollection, atlas:DigitalScholarlyEdition	dcterms:references	The URL of a web resource that presents the main edited source(s)
atlas:TextCollection, atlas:DigitalScholarlyEdition	atlas:notesOnSource	Additional information on the edited text(s)
atlas:TextCollection, atlas:DigitalScholarlyEdition	atlas:referencedAuthor	The main author(s) of the edited text(s)
atlas:TextCollection, atlas:DigitalScholarlyEdition	atlas:referencedWorkType	The type of the edited text(s)
atlas:TextCollection, atlas:DigitalScholarlyEdition	schema:genre	The genre of the edited text(s)
atlas:DigitalScholarlyEdition	atlas:editionType	The type of edition
atlas:TextCollection	schema:size	The number of collected items
atlas:LinkedOpenData, atlas:Ontology	dcterms:references	Imported ontologies or vocabularies
atlas:Ontology	vann:preferredNamespaceUri	The preferred namespace URI to use terms from this vocabulary
atlas:Software	schema:archivedAt	The URL of the software's repository
atlas:Software	swo:0000086	The format of input data
atlas:Software	swo:0000087	The format of output data
atlas:Software	swo:0000741	Used programming language(s)
atlas:Software	schema:isBasedOn	Reused or extended software component(s)

[26] https://vocabs.dariah.eu/tadirah/en/.

People & Organisations Identifying communities and scholars involved in scholarly outcomes represents one of the desiderata of the ATLAS Ontology. ATLAS distinguishes between `schema:Person` and `schema:Organisation`, allows users to record their current or most recent affiliation (`schema:affiliation`) and differentiates contribution roles to research outputs (see Research Product above). Common attributes of agents include their name (`schema:name`), external identifiers (`schema:identifier`), such as ORCID[27], and links to authority records (`schema:sameAs`), e.g. Wikidata entities. For Organisations, additional details, such as their landing page (`schema:url`) and location (`schema:location`), are also captured.

Research Project All reviewed models provide information on research activities supporting the production of an outcome. However, the focus is usually set on specific aspects, such as funding agencies, grants, and open-access mandates. ATLAS attempts to combine all such aspects and identify the main actors. To represent Research Projects, the class `schema:ResearchProject` is used. Following the hierarchical arrangement by Schema.org, this is a subtype of `schema:Organisation`, thus it inherits all its properties. In ATLAS we are interested in the following attributes: description (`schema:description`), start date (`schema:foundingDate`), end date (`schema:dissolutionDate`), organisations part of the project (`schema:member`), and funding entities (`schema:funder`).

Website & Computer program Websites and tools that expose access points to research data play a pivotal role in enhancing the findability and reusability of scholarly outcomes. To provide an effective representation of these services, the ATLAS Ontology introduces two types: Websites (`fabio:WebSite`) and Computer Programs (`fabio:ComputerProgram`).

Computer Programs were previously mentioned in the context of Research Product subtypes. Specifically, `fabio:ComputerProgram` is one of the two parent types for `atlas:Software`. The description of a Computer Program includes the type of provided service (`dcterms:type`), the title (`schema:name`), a description (`schema:description`), the access URL (`schema:url`), a URL for a documentation page (`schema:usageInfo`), afforded research activities (`schema:educationalUse`), the license (`schema:license`), and links to other software components that the described program extends or reuses (`schema:isBasedOn`). A similar set of attributes is also available for Websites, except for `dcterms:type` and `schema:license`. In this context, the `schema:isBasedOn` expresses connections to domain-relevant tools (i.e., Computer Programs), such as deployed Visualisation software to present Digital Scholarly Editions.

The review of the current landscape of controlled vocabularies for scholarly data highlighted the lack of taxonomies to describe a few aspects relevant to DH resources. The ATLAS Ontology introduces several terms (named individuals) to address such an issue. For instance, we collected a preliminary list of different types of Digital Scholarly Editions (e.g. `atlas:BestManuscriptEdition`, `atlas:DiplomaticEdition`, `atlas:DocumentaryEdition`), created from the Parvum lexicon stemmatologicum [30], and categories of textual resources (e.g. `atlas:CollectedWorks`, `atlas:Paper`, `atlas:SingleManuscript`, etc...) from the Patrick Sahle Catalog of Digital Scholarly Editions [10].

4.2. ATLAS Knowledge Graph v1.0

The ATLAS Ontology has been populated with a preliminary Knowledge Graph (ATLAS-KG) [31] describing selected pilot projects and resources. The ATLAS-KG also served as a testing ground for validating the semantic model outlined in the previous paragraphs and testing the functionalities of the ATLAS platform. ATLAS-KG leverages SKOS Thesauri and Authority Records used in the

[27] https://orcid.org/.

DH community, such as TaDiRAH[28] and EU Vocabularies[29], but also national controlled vocabularies (Schema.gov)[30], COAR[31], Linked Open Vocabularies (LOV)[32], Wikidata[33], VIAF[34], Geonames[35], ORCID[36], ROR[37]. The Knowledge Graph is organised in a number of Named Graphs, each corresponding to the content of a record in the ATLAS platform, filled in using a template, which in turn corresponds to a class/concept described above, namely: Research Product, Research Project, Person, Organisation, Computer program, and Website. Created data are currently available in their Turtle serialisations, while the platform is soon to be published. To date, the graph accounts for 179 records, including 16 Research Products, 11 Research Projects, 76 instances of Person, 59 Organisations, and 17 Websites and Computer Programs. Figure 2 provides a graphical example of the description of a Research Product, i.e., the Zeri Photo Archive RDF Dataset [32], the primary research outcome of the Zeri & LODE project. For the sake of brevity, only a few core statements are presented here, while a complete serialisation is available in the graph repository. Pink circles represent instances of ATLAS classes, with their types represented in yellow boxes.

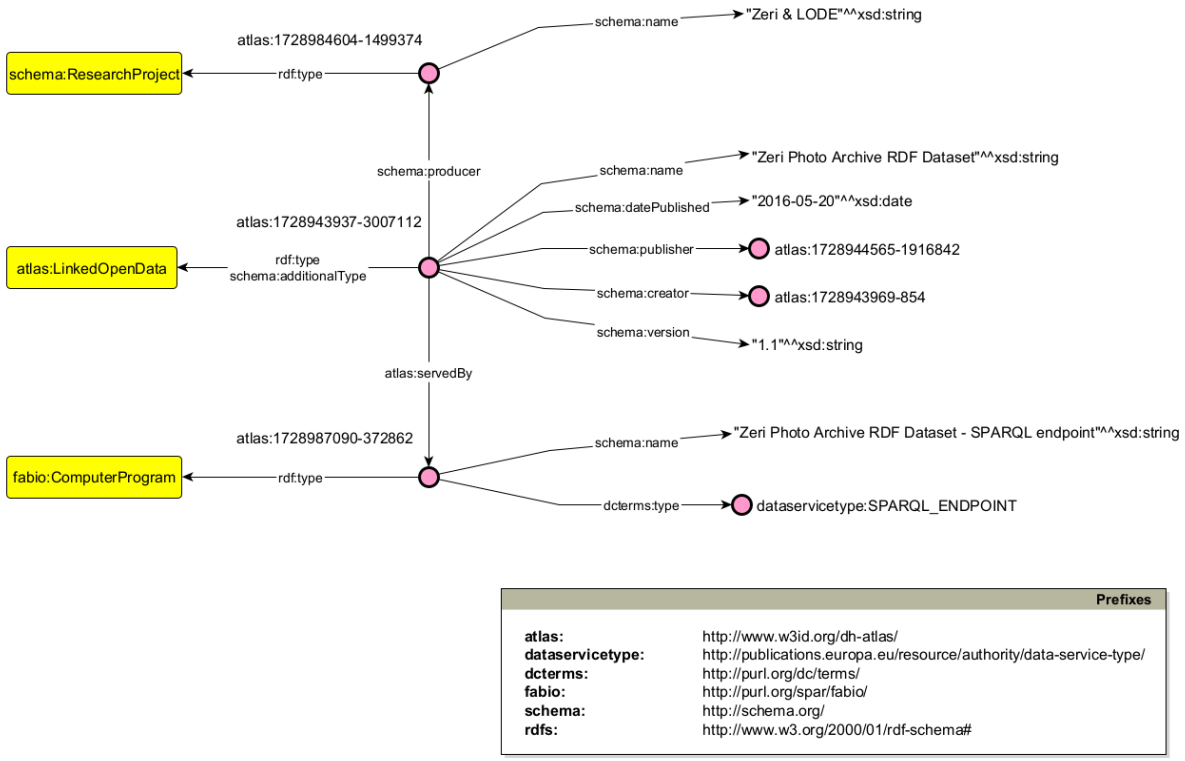

Figure 2: A visual diagram exemplifying the description of a Research Product and related entities.

Black arrows indicate predicates connecting entities to either entities or literal values. In the example, the Research Product named "Zeri Photo Archive RDF Dataset" (`atlas:1728943937-3007112`) is an instance of `atlas:LinkedOpenData` through the `rdf:type` and `schema:additionalType` properties. The relation with the Research Project responsible for its creation (`atlas:1728984604-1499374`), named "Zeri & LODE", is represented using the

[28] https://vocabs.dariah.eu/tadirah/en/.
[29] https://op.europa.eu/en/web/eu-vocabularies/controlled-vocabularies.
[30] https://schema.gov.it/.
[31] https://vocabularies.coar-repositories.org/.
[32] https://lov.linkeddata.es/dataset/lov.
[33] https://wikidata.org/.
[34] https://viaf.org/.
[35] https://geonames.org/.
[36] https://orcid.org/.
[37] https://ror.org/.

schema:producer property. Two object properties link the Research Product (schema:publisher and schema:creator) to the Agents (Person and Organisation) who contributed to its realisation. Lastly, atlas:servedBy connects the artefact to one of its access points, that is, an instance of fabio:ComputerProgram (atlas:1728987090-372862), labelled "Zeri Photo Archive RDF Dataset - SPARQL endpoint".

4.3. CLEF v3.0

The first version of the Knowledge Graph we have briefly introduced was created by leveraging the new functionalities provided by the latest release of the CLEF web application. Although the contribution here presented does not aim to address all potential technical requirements underlying a catalogue of scholarly data, it provides a number of features that current solutions have so far overlooked [24, 25], namely: (1) the usage of intermediate templates to prevent users from delving into the complexities of an ontology while entering data, (2) the possibility to fill in the record by semi-automatically extracting data from online data sources, and (3) provide customisable data visualisations based on the data created.

Intermediate Templates CLEF supports Linked Open Data crowdsourcing by streamlining data entry processes. Users can create LOD by filling a user-friendly web form, wherein fields correspond to RDF properties and the record is an entity of a class. Each record complies with a template, i.e. a set of mandatory and optional fields/properties to be filled with appropriate values.

However, implementing complex data models could result in intricate templates and describing a single resource often requires creating and linking several records. For instance, in ATLAS, when creating the record of a Research Product, users must also define (1) Organisation and Person instances for related creators, contributors, and publishers, (2) the corresponding Research Project, and (3) available Computer Programs and Websites serving as access points. While in existing systems this would require users to create preliminary records for such secondary entities, and only then recall these entities in the main record, CLEF allows users to create multiple records at the same time using a mechanism of subtemplates, which graphically include fields for describing the secondary, ancillary entity along with the main one. Notably, the mechanism underlying this functionality is ontology-independent, and can be reused in any new template.

While this solution facilitates the implementation of complex data models on a practical level, other updates have focused on knowledge engineering improvements. These include allowing the association of multiple OWL classes with the same Template as well as the integration of Subclasses.

Enhanced Knowledge Extraction The 2.0 version of CLEF introduced a working area for Knowledge Extraction, allowing users to retrieve named entities or Linked Open Data from various types of sources, including SPARQL endpoints, API services, and Static Files (.csv and .json formats) [33]. To query Static Files, CLEF 2.0 relies on SPARQL Anything[38], a reengineering tool that facilitates SPARQL interrogations on diverse data formats and returns RDF data regardless of the input format.

ATLAS seeks to (gradually) make Knowledge Extraction accessible to users with more or less technical background, therefore overcoming the barrier posed by query languages. To achieve this goal, a Manual Extraction option has been introduced. This feature enables contributors to provide the URL of a document (i.e., a .json, .csv, or .xml file), which is automatically parsed to identify JSON keys, CSV columns, or XML tags. Users can then select desired elements through a suggestion dropdown to extract corresponding values. Additionally, filtering options can be specified, such as a minimum number of occurrences and regular expressions. In the end, provided parameters are automatically converted into a SPARQL Anything query.

To complete the Extraction process and return LOD, template creators can now configure fields by associating them with an automatic Entity Reconciliation system. So doing, extracted terms are matched to the most relevant URI in selected sources like Wikidata and VIAF.

[38] https://sparql-anything.cc/.

Data visualisation CLEF integrates new explorative tools for improving user interaction with cataloguing data. Specifically, the updated platform introduces a new Charts Template section, designed to support the editorial board in creating customised data visualisation interfaces. This feature allows one to combine and arrange several presentations, enriched with textual description. For greater customisation, contributors can use HTML tags and attributes can be used to modify captions, ensuring design flexibility. Available visualisations rely on SPARQL queries to extract data from the catalogue and showcase it by leveraging the amCharts js library[39]. Key options include a) Counters, displaying some key metrics as standalone numerical values associated with customisable labels, b) Charts, visualising trends and data distributions through a variety of chart types, including bar graphs, pie charts, and doughnut charts, c) Maps, providing geographic representations of data by plotting resource distribution on interactive maps (Figure 3).

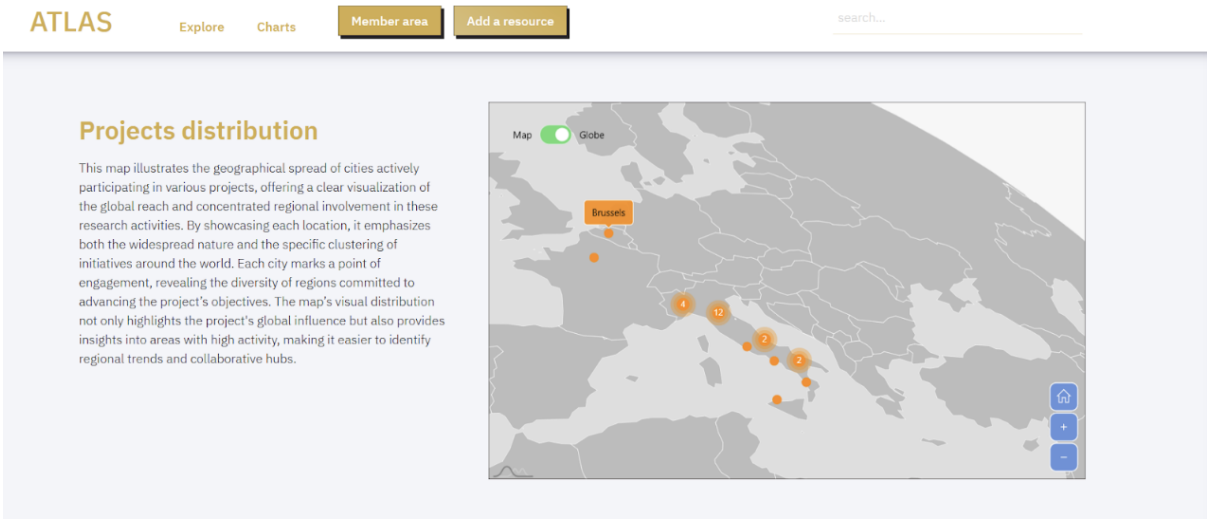

Figure 3: An example of an interactive map based on the ATLAS Knowledge Graph.

5. Discussion and Conclusions

The ATLAS Ontology seeks to enhance the description of Digital Cultural Heritage projects and their related outcomes by leveraging the potential of Linked Open Data. To this end, it integrates properties and entities from some of the most relevant semantic models within the DH domain and Schema.org, and provides terms to address the description of peculiarities relevant to scholars in the Humanities.

To evaluate the developed model, we extended the functionalities of CLEF, through which we created the ATLAS Knowledge Graph, including metadata of selected pilot projects. The newly implemented features, including intermediate templates, advanced knowledge extraction, and data visualisation tools, provided us with the instruments for populating and validating the ontology through the creation of a Knowledge Graph.

The level of granularity introduced by the ATLAS Ontology shows great potential for performing detailed data analyses on the Italian Cultural Heritage and its relation with Digital Humanities outcomes. In particular, its terminology has proven to effectively capture and describe different types of Research Products among selected resources, covering peculiar aspects such as DH methodologies. However, while the ontology provides a solid base for addressing a shared terminology, we will perform a user test to prove the goodness of our solutions and improve the terminology with user-contributed terms, so as to allow diversity and richness in the way scholars describe their results. Future developments will indeed expand ATLAS vocabularies, enabling better handling of this crucial gap and increasing the coverage of underrepresented concepts.

[39] https://amcharts.com/.

The extension of CLEF functionalities with scalable methods for Knowledge Extraction effectively simplifies this descriptive process by leveraging the Linked Open Data potential. Nonetheless, the road to facilitate LOD generation via user-friendly interfaces still poses a number of challenges, due to the variety of technical skills of scholars that would provide descriptions of their data. For this reason, the next stages of the ATLAS project will focus on extending the current Knowledge Graph through the analysis of new research initiatives. The insights and issues emerging from this process will inform the efforts to consolidate the developed model, while further usability tests will contribute to delivering a refined crowdsourcing platform. In so doing, ATLAS aims to offer an increasingly comprehensive tool, capable of advancing research in the DH domain and fostering the full valorisation of Italian Cultural Heritage.

Acknowledgments

Funded by the European Union- Next Generation EU, Mission 4 Component 1 CUP J53D23013000006.

References

[1] F. Tomasi, Organizzare la conoscenza: Digital Humanities e Web semantico, Editrice Bibliografica, 2022. doi:10.53134/9788893573573.

[2] N. Brügger, N. O. Finnemann, The Web and Digital Humanities: Theoretical and Methodological Concerns, Journal of Broadcasting & Electronic Media 57 (2013) 66–80. doi:10.1080/08838151.2012.761699.

[3] L. Frosini, A. Bardi, P. Manghi, P. Pagano, An Aggregation Framework for Digital Humanities Infrastructures: The PARTHENOS Experience, SCIRES-IT - SCIentific RESearch and Information Technology 8 (2018) 33–45. doi:10.2423/i22394303v8n1p33.

[4] K. Fenlon, J. Jett, C. L. Palmer, Digital Collections and Aggregations, 2017. URL: https://archive.mith.umd.edu/dhcuration-guide/guide.dhcuration.org/index.html%3Fp=77.html.

[5] F. Tomasi, Vespasiano Da Bisticci Letters, 2013. URL: http://vespasianodabisticciletters.unibo.it. doi:10.6092/unibo/vespasianodabisticciletters.

[6] S. Peroni, F. Tomasi, F. Vitali, The aggregation of heterogeneous metadata in web-based cultural heritage collections: a case study, International Journal of Web Engineering and Technology 8 (2013) 412–432. doi:10.1504/ijwet.2013.059107.

[7] M. L. Brogan, Survey of Digital Library Aggregation Services, Digital Library Federation, 2003. URL: http://old.diglib.org/pubs/dlf101/dlf101.htm.

[8] S. D. Giorgio, Gli archivi del MiBACT. L'integrazione dei dati archeologici digital, in: P. Ronzino (Ed.), Proceedings del Workshop L'integrazione dei dati archeologici digitali - Esperienze e prospettive in Italia (Lecce, Italia, 1-2 ottobre, 2015), 2016, pp. 47–55. URL: https://ceur-ws.org/Vol-1634/paper6.pdf.

[9] V. A. Carriero, A. Gangemi, M. L. Mancinelli, L. Marinucci, A. G. Nuzzolese, V. Presutti, C. Veninata, ArCo: The Italian Cultural Heritage Knowledge Graph, in: The Semantic Web – ISWC 2019, Springer International Publishing, 2019, pp. 36–52. doi:10.1007/978-3-030-30796-7_3.

[10] P. S. et al., Digitale edition v.4.112, Online, 2020. URL: https://www.digitale-edition.de.

[11] G. Franzini, M. Terras, S. Mahony, A Catalogue of Digital Editions, Open Book Publishers, 2016, pp. 161–182. doi:10.11647/obp.0095.09.

[12] Griseldaonline, Gli strumenti dell'Italianistica digitale, 2020. URL: https://site.unibo.it/griseldaonline/it/strumenti/strumenti-italianistica-digitale.

[13] C. Hall, Digital Humanities and Italian Studies: Intersections and Oppositions, Italian Culture 37 (2019) 97–115. doi:10.1080/01614622.2019.1717754.

[14] M. Daquino, A. Bardi, M. Buzzoni, R. Del Gratta, A. M. Del Grosso, F. Fischer, F. Tomasi, R. Rosselli Del Turco, The ATLAS: a knowledge graph of digital scholarly research on Italian Cultural Heritage, in: A. Di Silvestro, D. Spampinato (Eds.), Me.Te. Digitali. Mediterraneo in

rete tra testi e contesti, Proceedings del XIII Convegno Annuale AIUCD2024, 2024, pp. 588–592. doi:10.6092/unibo/amsacta/7927.

[15] M.-A. Sicilia, E. García-Barriocanal, S. Sánchez-Alonso, Community curation in open dataset repositories: Insights from zenodo, Procedia Computer Science 106 (2017) 54–60. doi:10.1016/j.procs.2017.03.009.

[16] N. Rettberg, B. Schmidt, Openaire - building a collaborative open access infrastructure for european researchers, LIBER Quarterly: The Journal of the Association of European Research Libraries 22 (2012) 160–175. doi:10.18352/lq.8110.

[17] S. Amodeo, A. Brunschweiger, F. Krauss, G. Malaguarnera, The OpenAIRE Graph: Empowering Research through Open Science, Zenodo, 2024. doi:10.5281/ZENODO.13885430.

[18] S. S.-R. et al., Packaging research artefacts with RO-Crate, Data Science 5 (2022) 97–138. doi:10.3233/ds-210053.

[19] S. B. et al., Why linked data is not enough for scientists, Future Generation Computer Systems 29 (2013) 599–611. doi:10.1016/j.future.2011.08.004.

[20] P. S. et al., Ro-crate metadata specification 1.1.3, 2023. doi:10.5281/ZENODO.7867028.

[21] S. Amodeo, T. Vergoulis, E. Papadopoulou, M. Buys, A. Mannocci, G. Malaguarnera, Eosc collaborative frontiers to achieve interoperability and enhance scholarly data, Zenodo, 2024. doi:10.5281/ZENODO.14055894.

[22] S. Peroni, D. Shotton, Fabio and cito: Ontologies for describing bibliographic resources and citations, Journal of Web Semantics 17 (2012) 33–43. doi:10.1016/j.websem.2012.08.001.

[23] L. Fintoni, Rethinking scholarly digital objects as cultural heritage: the KNOT project, in: A. Di Silvestro, D. Spampinato (Eds.), Me.Te. Digitali. Mediterraneo in rete tra testi e contesti, Proceedings del XIII Convegno Annuale AIUCD2024, 2024, pp. 582–587. doi:10.6092/unibo/amsacta/7927.

[24] M. Daquino, M. Wigham, E. Daga, L. Giagnolini, F. Tomasi, Clef. a linked open data native system for crowdsourcing, Journal on Computing and Cultural Heritage 16 (2023) 1–17. doi:10.1145/3594721.

[25] S. Giacomini, M. Daquino, F. Tomasi, L. Fintoni, CLEF 2.0. Solutions for Native Linked Data Cataloguing of Italian Digital Cultural Heritage, 2025. JLIS.it. In publication.

[26] N. Freire, V. Charles, A. Isaac, Evaluation of Schema.org for Aggregation of Cultural Heritage Metadata, Lecture Notes in Computer Science 10734 (2018) 225–239. doi:10.1007/978-3-319-93417-4_15.

[27] F. Tomasi, Digital humanities e organizzazione della conoscenza: una pratica di insegnamento nel LODLAM, AIB Studi 60 (2020). URL: https://aibstudi.aib.it/article/view/12068. doi:10.2426/aibstudi-12068.

[28] A. Bardi, M. Buzzoni, M. D. R. Del Gratta, A. M. Del Grosso, F. Fischer, F. Tomasi, R. Rosselli Del Turco, S. Giacomini, C. Martignano, G. Rubin, DH ATLAS: Ontology v1.0, Zenodo, 2024. doi:10.5281/ZENODO.14058232.

[29] Schema.org, Data and Datasets overview, 2024. URL: https://schema.org/docs/data-and-datasets.html.

[30] P. Roelli, C. Macé, Parvum lexicon stemmatologicum. A brief lexicon of stemmatology, Helsinki University Homepage, 2015. doi:10.5167/UZH-121539.

[31] A. Bardi, M. Buzzoni, M. D. R. Del Gratta, A. M. Del Grosso, F. Fischer, F. Tomasi, R. Rosselli Del Turco, S. Giacomini, C. Martignano, G. Rubin, DH ATLAS: Knowledge Graph v1.0, Zenodo, 2024. doi:10.5281/ZENODO.14058144.

[32] M. Daquino, F. Mambelli, S. Peroni, F. Tomasi, F. Vitali, Zeri Photo Archive RDF Dataset, Centro Risorse per la Ricerca (CRR-MM), Università di Bologna, 2016. doi:10.6092/UNIBO/AMSACTA/5497.

[33] M. Daquino, L. Fintoni, S. Giacomini, F. Tomasi, CLEF 2.0. Soluzioni per la catalogazione nativa Linked Data del patrimonio digitale culturale italiano, in: A. Di Silvestro, D. Spampinato (Eds.), Me.Te. Digitali. Mediterraneo in rete tra testi e contesti, Proceedings del XIII Convegno Annuale AIUCD2024, 2024, pp. 417–422. doi:10.6092/unibo/amsacta/7927.