Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
Consiglio Nazionale delle Ricerche

ISTI Technical Reports

DH ATLAS: White book v1.2

Alessia Bardi, ISTI-CNR, Pisa, Italy
Marina Buzzoni, Università di Venezia, Venezia, Italy
Marilena Daquino, Università di Bologna, Bologna, Italy
Riccardo Del Gratta, ILC-CNR, Pisa, Italy
Angelo Mario Del Grosso, ILC-CNR, Pisa, Italy
Roberto Rosselli Del Turco, Università di Torino, Torino, Italy
Franz Fischer, Università di Venezia, Venezia, Italy
Sebastiano Giacomini, Università di Bologna, Bologna, Italy
Chiara Martignano, Università di Venezia, Venezia, Italy
Giorgia Rubin, ILC-CNR, Pisa, Italy
Francesca Tomasi, Università di Bologna, Bologna, Italy

ISTI-TR-2025/019

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
Consiglio Nazionale delle Ricerche

DH ATLAS: White book v1.2

Alessia Bardi, Marina Buzzoni, Marilena Daquino, Riccardo Del Gratta, Angelo Mario Del Grosso, Roberto Rosselli Del Turco, Franz Fischer, Sebastiano Giacomini, Chiara Martignano, Giorgia Rubin, Francesca Tomasi

ISTI-TR-2025/019

The DH ATLAS whitebook presents the ATLAS catalogue and its underlying data model and provides guidelines and best practices for producing "FAIR" (Findable, Accessible, Interoperable, and Reusable) scholarly outputs in Digital Humanities and enhancing Italian digital cultural heritage.

Keywords: Digital Humanities Data FAIRness.

Citation
Bardi A. et al… DH ATLAS: White book v1.2.
ISTI Technical Reports 2025/019. DOI: 10.32079/ISTI-TR-2025/019.

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
Area della Ricerca CNR di Pisa
Via G. Moruzzi 1
56124 Pisa Italy
http://www.isti.cnr.it

THE ATLAS OF ITALIAN DIGITAL CULTURAL HERITAGE

Whitebook

Alessia Bardi (alessia.bardi@isti.cnr.it)
Marina Buzzoni (mbuzzoni@unive.it)
Marilena Daquino (marilena.daquino2@unibo.it)
Riccardo Del Gratta (riccardo.delgratta@ilc.cnr.it)
Angelo Mario Del Grosso (angelo.delgrosso@ilc.cnr.it)
Roberto Rosselli Del Turco (roberto.rossellidelturco@unito.it)
Franz Fischer (franz.fischer@unive.it)
Sebastiano Giacomini (sebastiano.giacomin2@unibo.it)
Chiara Martignano (chiara.martignano@unive.it)
Giorgia Rubin (giorgia.rubin@ilc.cnr.it)
Francesca Tomasi (francesca.tomasi@unibo.it)

Abstract

In the last twenty years, numerous Web-based Digital Humanities projects have emerged to collect data in the field of Italian Cultural Heritage. However, the research outputs produced by these projects risk obsolescence if not well-documented and based on shared standards and best practices. Moreover, these research products are often difficult to discover, and currently, no dedicated comprehensive catalogue exists. The ATLAS project—a joint effort of the University of Bologna, the Ca' Foscari University of Venice, and the CNR Unit of Pisa (that includes the Institute for Computational Linguistics "A. Zampolli" - ILC - and the Institute of Information Science and Technologies "Alessandro Faedo" - ISTI)—aims to fill this gap. It seeks to create a catalogue that leverages Semantic Web technologies to interlink scholarly outputs and enhance the Italian Cultural Digital Heritage. The main expected outcomes of the ATLAS project are:

- a knowledge graph representative of Italian heritage, accessible online via dedicated services and preserved in CLARIN;
- a web application that allows users to explore the catalogued scholarly products and their contents, starting from the pool of selected research products used as pilots for the project;
- a whitebook for the curators of the catalogue that also analyses the state of the art and offers scholars a set of guidelines and best practices to produce FAIR (Findable, Accessible, Interoperable and Reusable) and high quality scholarly data.

The whitebook begins with a review of the state of the art and presents the ATLAS project's goals and methodologies. Chapter 1 evaluates the research products used as pilots for designing the ATLAS data model and provides guidelines for producing FAIR data. Chapter 2 presents the ATLAS data model, with particular attention to metadata for cataloguing research products. Chapter 3 introduces the ATLAS web application and provides instructions for data entry and access for future curators and users.

Disclaimer

This version of the whitebook is provisional and is focused on the project's initial outcomes. A refined version (planned for M23, November 2025) will present the complete ATLAS data model and all features for data entry and access through the ATLAS web application.

Table of contents

Introduction

In recent years, the growing integration of the World Wide Web and its technologies has profoundly reshaped scholarly research, particularly within the Digital Humanities (DH) domain. These advancements have unlocked new avenues for preserving, sharing, and reusing research outputs, fostering unprecedented collaboration and dissemination. Yet, as vast amounts of scholarly data proliferate, the need for standardised models and guidelines to manage, aggregate, and explore this information effectively has become more pressing.

Several platforms in the broader scholarly landscape play a crucial role in providing persistent identification, long-term preservation, and enhanced findability of research data. Key services include Zenodo[1] and OpenAIRE.[2] The OpenAIRE network integrates various services, including community web portals such as the Digital Humanities and Cultural Heritage gateway, which facilitate the discovery and sharing of research outcomes and Open Science practices.

Currently, multiple catalogues document DH research, including digital scholarly editions,[3] and project lists from national[4] and international associations,[5] and research centres. However, there remains a lack of comprehensive catalogues specifically focused on DH projects related to Italian Cultural Heritage, and no structured collections exist for DH research outputs using Semantic Web technologies.

In Italy, institutions have made significant progress in digitising and aggregating Cultural Heritage through Linked Open Data (LOD) collections. Notable initiatives include the *dati.culturaitalia*[6] platform by the Italian Ministry of Culture and the ArCO[7] project, which has created a Knowledge Graph based on the General Catalog of Italian Cultural Heritage. These efforts align with European digitisation projects like ARIADNE[8] and Europeana.[9] While these initiatives provide interoperable LOD, a gap persists in the research framework for supporting best practices and improving the findability and reusability of Italian heritage-related DH data.

Current models inadequately address the complexities of today's DH landscape. DH projects produce very diverse outputs—including text collections, digital scholarly editions, Linked Open Data datasets, RDF vocabularies, and software—each requiring specific descriptive approaches. Critical elements like textual typologies and edition criteria remain insufficiently addressed. Furthermore, current models lack effective solutions for connecting research activities to their corresponding Cultural Heritage objects, despite the opportunities offered by Linked Open Data.

---

[1] https://zenodo.org/.
[2] https://www.openaire.eu/.
[3] Franzini, G. (2012-) Catalogue of Digital Editions, https://doi.org/10.5281/zenodo.1161425.
Sahle, Patrick et al., a catalog of Digital Scholarly Editions, v.4.112 2020ff, last change 2024-06-06.
[4] For example AIUCD's list of DH projects: https://www.aiucd.it/progetti/.
[5] For example EADH's list of DH projects: https://eadh.org/projects.
[6] https://dati.cultura.gov.it/.
[7] https://dati.beniculturali.it/arco/index.php.
[8] https://ariadne-infrastructure.eu/.
[9] https://www.europeana.eu/.

The DH field also needs clear guidelines for creating high-quality scholarly data. While best practices and guidelines (e.g., RIDE[10] reviewing guidelines) already exist along with FAIR principles,[11] these resources need expansion to cover the full range of DH resources.

The ATLAS[12] project aims to address these challenges by creating a comprehensive knowledge graph of DH research related to Italian Digital Cultural Heritage. By developing the ATLAS Ontology and its associated knowledge graph, the project seeks to establish a semantic framework that captures the diverse outputs of DH research, including digital editions, text collections, and datasets. ATLAS addresses the complexities of describing and linking scholarly data, ensuring that metadata is enriched and accessible, thus promoting the findability and reusability of these valuable cultural resources. Through detailed analysis and mapping of existing models and vocabularies, the project offers a new approach to integrating Italian DH resources into the global knowledge landscape, fostering greater collaboration and discovery across disciplines and institutions.

The ATLAS project will contribute to the Italian DH research landscape through the following key outcomes:

1. the evaluation of pilot projects, identifying strategies for handling the mapping of knowledge, data manipulation, and ensuring access and persistence for different types of digital research outputs;
2. the ATLAS ontology, mapping excerpts of schemas and ontologies reused by the pilot projects, further enhancing the interoperability of data and tools;
3. a knowledge graph on DH projects and scholarly data related to Italian Cultural Heritage, accessible via the ATLAS web application and preserved in a trustworthy repository;
4. a search portal built on the OpenAIRE CONNECT Gateway,[13] focused on scholarly literature and data relevant to the pilots and beyond;
5. a whitebook outlining good practices for FAIR (Findable, Accessible, Interoperable, Reusable) scholarly data, and ensuring high-quality content. The whitebook also contains the pilots analysis results, describes the ATLAS data model, and presents the ATLAS web application with guidelines for catalogue curators.

References

Brogan, Martha L. 2003. 'Survey of Digital Library Aggregation Services', Digital Library Federation, Washington, District Columbia, USA [online] http://old.diglib.org/pubs/dlf101/dlf101.htm (accessed 19 December 2024).

Carriero, Valentina Anita, Aldo Gangemi, Maria Letizia Mancinelli, Ludovica Marinucci, Andrea Giovanni Nuzzolese, Valentina Presutti, and Chiara Veninata. 2019. 'ArCo: The Italian Cultural Heritage Knowledge Graph'. In *The Semantic Web – ISWC 2019*, edited by

[10] RIDE Criteria for Reviewing Digital Text Collections, version 1.0: https://www.i-d-e.de/publikationen/weitereschriften/criteria-text-collections-version-1-0/.
[11] https://www.go-fair.org/fair-principles/.
[12] ATLAS is a project funded by the Next Generation program of the European Commission for 24 months (October 2023 - October 2025).
[13] https://connect.openaire.eu/.

Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtěch Svátek, Isabel Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon, 36–52. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-30796-7_3.

Franzini, Greta, Simon Mahony, and Melissa Terras. 2016. 'A Catalogue of Digital Editions'. In *Digital Scholarly Editing: Theories and Practices*, a cura di E. Pierazzo e M. Driscoll, 4:161–82. Cambridge, UK: Open Book Publishers. https://doi.org/10.11647/OBP.0095.

Manghi, Paolo, Alessia Bardi, Claudio Atzori, Miriam Baglioni, Natalia Manola, Pedro Principe, Michele Artini, et al. 2019. 'The OpenAIRE Research Graph Data Model'. Zenodo. https://doi.org/10.5281/zenodo.2643199.

Peroni, Silvio, Francesca Tomasi, and Fabio Vitali. 2013. 'The Aggregation of Heterogeneous Metadata in Web-Based Cultural Heritage Collections: A Case Study'. *International Journal of Web Engineering and Technology* 8 (4): 412–32. https://doi.org/10.1504/IJWET.2013.059107.

Rettberg, Najla, and Birgit Schmidt. 2012. 'OpenAIRE - Building a Collaborative Open Access Infrastructure for European Researchers'. *LIBER Quarterly: The Journal of the Association of European Research Libraries* 22 (3): 160–75. https://doi.org/10.18352/lq.8110.

Tomasi, Francesca, Fabio Ciotti, Maurizio Lana, Fabio Vitali, Silvio Peroni, e Diego Magro. 2013. «Dialogue and Linking between TEI and Other Semantic Models». In *The Linked TEI: Text Encoding in the Web*, 145–58. Roma: DIGILAB Sapienza University & TEI Consortium. https://hdl.handle.net/11585/185113.

Tomasi, Francesca. 2022. *Organizzare la conoscenza: Digital Humanities e Web semantico. Un percorso tra archivi, biblioteche e musei*. Milano: Editrice Bibliografica. https://cris.unibo.it/handle/11585/848605.

1. Pilot Analysis

Abstract

The ATLAS project aims to evaluate a pool of research products in order to identify descriptive metadata characterising research objects and projects via a bottom-up approach, as well as defining good practices for accessing and manipulating Cultural Heritage data. Metadata of selected scholarly products (also called pilots) were incorporated into the initial ATLAS knowledge graph (ATLAS-KG). The pilots fall into five main categories:

1. collections of text documents (ALIM, the Archive of the Italian Latinity of the Middle Ages; Biblioteca italiana; Musisque Deoque; BUP - Digital Humanities);
2. digital scholarly editions (VaSto, VArchi STOria fiorentina; Codice Pelavicino Digitale; Digital Edition of Aldo Moro's works);
3. software tools for text processing and visualisation (EVT, Edition Visualization Technology; Voyant tools);
4. linked open data (Zeri & LODE; DanteSources; LiLa - Linking Latin; Biflow - Toscana Bilingue Catalogue);
5. ontologies for the Cultural Heritage (CIDOC-CRM; HiCO; SPAR).

The pilots listed above were selected based on two main criteria. First, they represent advanced Digital Humanities projects in the Italian Cultural Heritage domain, covering Italian and Latin sources and spanning from classical to contemporary periods. Second, for each pilot, at least one member of the ATLAS project team is directly involved in its development, providing access and in-depth knowledge to related resources.

The pilots were evaluated according to existing standards and best practices established in the field of Digital Humanities, as well as the FAIR principles. The results of this evaluation are presented in this first chapter, which is divided into five sections. Each section is dedicated to a category of digital research products, as listed above. We summarise trends and practices that emerged from the pilots analysis and we suggest how to best implement that kind of research product. The analysis and recommendations focus on the FAIRness of the scholarly data, while we retain considerations on the content and the effectiveness from a theoretical standpoint. For instance, we do not evaluate the pertinence of the philological approach applied to a digital scholarly edition. The recommendations outlined in this chapter aim to fill the current gap of a comprehensive set of guidelines for producing FAIR scholarly data in the field of Digital Humanities.

Methodological notes

The initial phase of the ATLAS project focused on the analysis of a curated selection of research products. The analysis served two purposes: identifying cataloguing metadata to describe research projects and products, and exploring methods to automatically extract data from various types of research products.

Selected pilots primarily consist of research products related to Italian historical and literary fields, mainly developed by Italian researchers. Notable exceptions include the Voyant Tools software and the CIDOC-CRM ontology, both widely adopted in the Italian scholarly environment, but developed abroad. The pilots span five categories: text collections, digital editions, ontologies, linked open data, and software tools. Pilots were chosen to represent ongoing Digital Humanities research on the Italian Cultural Heritage, covering periods from classical to 20th-century and encompassing both vernacular Italian and Latin. We prioritised research products considered valid examples in Italian Digital Humanities, wherein at least one ATLAS project team member is directly involved, ensuring high information quality and direct access to resources. Nonetheless, we include references to other notable research products to present a broader landscape of Digital Humanities research.

We evaluated the pilots using standards and guidelines for producing FAIR research products that are shared within the broader scholarly community, although we focused on the humanities' specificities.

The analysis addresses four distinct yet interrelated dimensions:

- metadata, examining accessibility and adherence to standards;
- data implementation, evaluating licence clarity, use of standard formats, availability of persistent URLs or identifiers, and links to semantic web datasets;
- data access and consultation methods, assessing user interface and/or API accessibility, data download availability, and adoption of long-term preservation policies;
- documentation, reviewing website information clarity and completeness, availability of high-level and technical documentation, citation guidelines, and public detailed changelog and roadmap.

It is important to note that our analysis does not assess the scientific merit or theoretical soundness of the research products. For instance, when examining a digital scholarly edition, we do not evaluate the effectiveness of the chosen editorial approach. We assume that any research product included in the ATLAS catalogue adheres to the theories and practices of its respective discipline. Instead, our analysis focuses on how the data is created and managed from a FAIR perspective.

Standards and guidelines

The pilots' analysis was mainly based on the following standards and guidelines:

- The FAIR Guiding Principles for scientific data management and stewardship.
- Dublin Core metadata specification.
- PARTHENOS Guidelines to FAIRify data management and make data reusable.
- OPERAS Common Standards White Paper, June 2021.
- Research Data Alliance standards.
- GoTriple Content Providers Handbook.
- MDR-MAA standards for web publication.
- Open Archives Initiative Protocol for Metadata Harvesting.

Additional standards and guidelines specific to each category of research product are detailed in the related section.

Our analysis identified both strengths and areas for improvement in the pilots. Evaluation results are supported by the study of the state of the art, which revealed a number of best practices and strategies for creating FAIR research products that were differently applied according to the category of research object. The analysis' final outcome is a set of both generic and specific recommendations for implementing FAIR research products in the Cultural Heritage domain.

General recommendations

This section presents recommendations for all types of research products, organised into two lists. The first list summarises existing standards and guidelines—particularly the FAIR principles and PARTHENOS' 20 guidelines for making data management FAIR and reusable—focusing on recommendations for scholars creating research products. The second list presents additional recommendations developed from critical insights during our analysis. These focus on effectively presenting research products and improving their usability.

Recommendations to ensure data and metadata FAIRness

- Model and structure data according to domain-relevant community standards.
- Use a formal, accessible and shared language for knowledge representation.
- Use standard and non-proprietary file formats to ensure long-term preservation and interoperability. Use current popular file formats next to archival formats to increase reuse (e.g., Excel and CSV).
- Describe data with rich metadata, using a plurality of accurate and relevant attributes. Choose appropriate metadata schemas. Use open well-defined metadata vocabularies that follow FAIR principles (e.g., Dublin Core).
- Clearly state access rights and choose Open Access when possible. Use a data embargo when needed.
- Release the research product under a clear and accessible data usage licence.
- Assign data and metadata globally unique and persistent identifiers (PIDs), to enhance findability and citability of the research product. In metadata clearly and explicitly include the PID of the data they describe. Examples of PIDs forms are: Handle,[14] DOI,[15] PURL,[16] and URN.[17]

[14] https://www.handle.net/.
[15] https://www.doi.org/.
[16] https://purl.archive.org/.
[17] https://datatracker.ietf.org/doc/html/rfc3986.

- Use persistent author identifiers, to create links between research products and allow recognition. Examples are VIAF,[18] ISNI,[19] and ORCID.[20]
- Make metadata and data retrievable on the web by their identifier using a standardised communications protocol (e.g., HTTP(S), SMTP,[21] FTP[22]).
- Deposit the research product in a certified repository, to guarantee long-term accessibility. Examples of certification standards are CoreTrustSeal,[23] nestor seal,[24] and ISO 16363 certification.[25] Repositories automatically assign permanent identifiers to research products and maintain version tracking.
- Register and index the research product in a searchable resource (e.g., OpenAIRE,[26] CLARIN,[27] DARIAH[28]).
- Provide comprehensive data and metadata documentation that clearly references all standards and models used.
- Maintain data integrity. Preserve and ensure long-term access to all official versions of the research product. Document changes and updates, ideally formatted as a changelog.
- Ensure data quality, providing references to other (meta)data and detailed provenance. Create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data.

Recommendations to enhance research products' presentation and usability

- Choose a clear, descriptive title for your research product, including relevant keywords. If using an acronym, provide the full name for clarity.
- Create a website or landing page for the research product, where the following information is prominently displayed:
    - access points to the data;
    - licence;
    - version number;
    - status (e.g., completed, under development).
- Provide complete credits, including names, affiliations and persistent identifiers of creators -when available, collaborators, and institutional partners. Include contact information for users to report issues or suggest collaborations.
- Enhance citability by providing citation guidelines.
- For products developed within a research project, describe the project or provide a link to its landing page, explaining how the product contributes to the project's goals.

---

[18] https://viaf.org/.
[19] https://isni.org/.
[20] https://orcid.org/.
[21] https://datatracker.ietf.org/doc/html/rfc5321.
[22] https://www.ietf.org/rfc/rfc959.txt.
[23] https://www.coretrustseal.org/.
[24] https://www.langzeitarchivierung.de/Webs/nestor/EN/Zertifizierung/nestor_Siegel/siegel.html.
[25] https://www.iso.org/standard/56510.html.
[26] OpenAIRE's knowledge graph search interface: https://explore.openaire.eu/.
[27] List of depositing services provided by certified CLARIN centres and other european institutions: https://www.clarin.eu/content/depositing-services.
[28] DARIAH's tools and services catalogue: https://www.dariah.eu/tools-services/tools-and-services/.

For time-limited projects, specify the end date and outline future plans for preservation and maintenance.
- In the documentation describe the research product's design and development process, including used tools and applied methodologies.
- Provide a user guide and examples demonstrating how to utilise the research product.
- To ensure efficient documentation management, maintain a single source document in a repository and generate different format versions from it as needed.
- Offer metadata and documentation in English (in addition to other languages) to reach a broader audience.
- If possible, make data freely downloadable to encourage reuse.

Text collections

A text collection is a type of digital archive that focuses on textual heritage, preserving texts digitally and providing web access. Beyond the texts, a text collection typically offers bibliographic metadata and editorial annotations. Texts may be available in various formats, including digital facsimiles of documents (e.g., the corpus "Incunaboli in volgare" within Biblioteca Italiana) or critical editions (e.g., Musisque Deoque).

For a text collection to be considered scientifically valid, its texts must be scholarly edited following shared methodologies, and its documentation must clearly explain the editorial criteria and workflow. For this reason, there is often a fine line between digital scholarly editions and digital text collections:

"Editions widen their content. When they aim at including ever more documents and finally at completeness, and when the first level of representation may be just a digital facsimile with some metadata, then the edition looks more and more like an archive. [...] On the other hand, digital archives are already critical on the bibliographic level and imply the possibility to incrementally add further critical information, other forms of representation (such as transcription) and may finally even present an edited text. [...] If we take the critical engagement and the application of scholarly knowledge as the defining characteristics of an edition, then we can say that from a certain point on, an archive starts to be an edition." (Sahle 2016)

The OAIS (Open Archival Information System) reference model[29] serves as the conceptual foundation for most long-term digital preservation projects today. Among the pilot text collections, only Biblioteca Italiana explicitly mentions using the OAIS reference model in its implementation. The OAIS reference model outlines key mechanisms for long-term information preservation and access. Its functional model defines six core services: Ingest, Archival Storage, Data Management, Preservation Planning, Access, and Administration. The model centres on information packages containing both the preserved object and its metadata. These come in three forms: Submission (SIP), Archival (AIP), and Dissemination (DIP) Information Packages. Originally developed in 2003 and approved as ISO standard

[29] See references in the text collections' standards and guidelines section.

14721, OAIS has become the definitive reference for long-term digital preservation at both national and international levels.

Creating and maintaining a textual archive demands considerable time, funding, and resources. It is no coincidence that three of the analysed pilots boast a thirty-year history, resulting from collaborations among numerous universities and support from various institutions. Simultaneously, we can perceive the diverse levels of completeness and technological sophistication in the databases published within the "Digital Humanities" collection, which primarily stem from doctoral projects.

Over time, ALIM, MQDQ, and Biblioteca Italiana have undergone extensive restructuring to adapt to technological advancements. All three databases have adopted the XML/TEI standard for text preparation. ALIM and Biblioteca Italiana are based on Muruca, a framework for creating and publishing digital libraries. Using existing software is preferable to developing new software to ensure easier maintenance of the text collection. However, long-term access to resources still depends on the maintenance of the publishing platform. MQDQ's solution of publishing its texts on an external repository (ILC4CLARIN)[30] appears to be the only zero-cost option currently available to guarantee long-term data access. MQDQ also sets a commendable example by implementing RESTful APIs to query the database, thereby promoting interoperability and data reuse.

All pilots lack clear indications of database status and detailed documentation of changes and updates. These shortcomings stem from two main challenges in the field of research and Cultural Heritage. First, it is difficult to predict funding availability, which affects planning for database implementation and maintenance. Second, tracking editorial work is complex when it is carried out collaboratively by large groups. The latter issue can be addressed by adopting tools that facilitate collaborative editing on shared resources and by publishing resources on external repositories (e.g., Perseus Digital Library's GitHub repositories).[31] Planning regular publications—ideally annually or biannually—or updates (e.g., Corpus Corporum's[32] "What's new?" section) after substantial modifications makes it easier to document editorial work and indicate progress status.

Today, most text collections are encoded in XML/TEI, but research projects often lack the time and resources to produce detailed encodings that capture all philological and material phenomena in texts. Various projects, including ALIM and Biblioteca Italiana, have implemented a multi-level encoding system comprising: a "light" or "base" level mainly focused on structural aspects, such as page divisions, and one or two more advanced levels. The basic level is applied to all texts, while the other levels are progressively implemented in the collection. To encourage the adoption of the XML/TEI standard, despite limited resources, we recommend considering starting the editorial workflow with the automatic transcription of reference texts and then using automatic tools for conversion into the XML/TEI format.

The text collections we selected as pilots are:

[30] https://ilc4clarin.ilc.cnr.it/.
[31] https://github.com/PerseusDL.
[32] https://mlat.uzh.ch/home.

- **ALIM** (Archive of the Italian Latinity of the Middle Ages),[33] an ongoing project whose main goal is to collect and publish all Latin texts produced in Italy during the Middle Ages. ALIM offers reliable TEI-based editions of both literary and documentary texts, making ALIM an invaluable resource for philologists, historians and literary scholars.
- **Biblioteca italiana**,[34] a digital library of more than 3000 texts representative of Italian heritage spanning between the Middle Ages and the 20th century. All texts are associated with detailed bibliographic metadata, while texts within the BibIt corpus are also encoded in XML/TEI.
- **Musisque Deoque** (MQDQ),[35] a digital archive of scholarly edited poetic texts in Latin. The project's main corpus includes 642 works, spanning between the origins and the 7th century AD and amounting to a total of 343.709 verses and about 2.300.000 tokens. The "Poeti d'Italia in lingua latina" corpus comprises Latin poetic texts written in Italy during the Middle Ages and is composed of 3.200.000 tokens. The scholarly editions of the texts are encoded in XML/TEI and are characterised by rich critical apparatuses and detailed metrical analysis.
- **BUP - Digital Humanities**,[36] an editorial collection of digital scholarly editions and databases. All resources are described with detailed bibliographic metadata. The digital scholarly editions are encoded in XML/TEI with a very rich markup and are published via the open source software EVT.

ALIM

The ALIM database (Archivio della Latinità Italiana del Medioevo, website: https://alim.unisi.it/) aims to collect all texts written in Latin in Italy during the Middle Ages. Initiated in the mid-1990s[37] and continuing to this day, ALIM is the product of a collaborative effort among six Italian universities: Verona (coordinated by Prof. Antonio De Prisco), Suor Orsola Benincasa Naples (Prof. Edoardo D'Angelo, who is also the national coordinator of the entire project), Palermo (Prof. Giorgio Di Maria), Ca' Foscari Venice (Prof. Marina Buzzoni), Siena-Arezzo (Prof. Francesco Stella), and Basilicata (Prof. Fulvio Delle Donne).

The primary objective of the database is to provide scholarly reliable texts, mainly targeting academics such as philologists, literary historians, and historians specialising in medieval culture, science, and institutions. Additionally, the text collection serves as a valuable teaching resource.

In addition to its primary objective, ALIM contributes to the European Dictionary of Medieval Latin, sponsored by the Union Académique Internationale in Brussels. To support this effort, ALIM's digital library includes both public and private documentary sources from the 8th to the 15th centuries. The ALIM website also features "Lexicon," a tool for lexical

[33] http://en.alim.unisi.it/.
[34] http://www.bibliotecaitaliana.it/.
[35] http://mqdq.it.
[36] https://bup.unibas.it/library/DH.
[37] The ALIM project has been sustained by PRIN funds (granted by the Italian government) from 1996 to 2017, with additional support from the National Research Council (CNR), the National Academic Union (UAN), and the participating universities.

analysis that generates indexes, frequency diagrams, and comparisons of linguistic forms across texts.

The current version of ALIM's digital library (ALIM 2.0) debuted in 2016, following a comprehensive overhaul of the original version's architecture and editorial workflow.[38] ALIM 2.0 employs XML/TEI (P5) encoding for both documentary and literary sources and is exclusively accessible via its website. This website is constructed using Muruca,[39] a modular framework designed for creating and publishing digital libraries. All texts are open access, allowing users to freely download them in HTML, XML, TXT, or PDF formats.

The website's homepage highlights two key components: diverse access points to the texts and various collections. The access points include the authors' index, works' index, and searchable lists for literary and documentary sources. The collections function as thematic sections within the ALIM database, showcasing unique and rare texts. However, the homepage lacks crucial information such as citation guidelines and the date of the last update.

ALIM contains 774 literary sources and 6,654 documentary sources. While these numbers are not explicitly stated in the database's presentation, they are evident from the respective lists. Both literary and documentary source lists offer full-text and proximity search functions. Users can filter the literary texts by author, work, type (prose or verse), historical period, genre, and collection. The documentary sources can be filtered by collection, historical period, place, and "corpus" (referring to the codex or other document where the sources are preserved).

Each text is accompanied by a set of "bio-bibliographical" metadata. These include: work title, author name, historical period, text type (e.g., letter), record entry date (often empty), style (prose or verse), genre (e.g., comedy, lyric poetry), source type (documentary or literary), text dimensions (total words and characters), encoding information, notes (editorial and similar), and source. The bibliographic reference of the text source is provided in full. While the source type (e.g., "critical edition") is not explicitly listed among the metadata, the database's general presentation states that authoritative and scholarly editions were used as text sources.[40] For documentary sources, the style and genre are not specified.

[38] ALIM 1.0 remains accessible at http://www.alim.dfll.univr.it/.
[39] https://www.muruca.org/.
[40] Metadata attached to documentary sources are slightly different. However, it was not possible to analyse them thoroughly, as the consultation of documentary sources is currently malfunctioning.

Dante Alighieri's *De vulgari eloquentia* as presented in the ALIM database.

The database presentation is concise yet effectively illustrates its contents and guides users through the website's features. The documentation provides detailed information on the website's architecture and, most importantly, the data model for TEI text encoding. This includes the schema in DTD and RNG formats, a teiHeader template, and a comprehensive handbook. ALIM's TEI encoding procedure comprises three distinct levels:

- a "base-level" that marks up the text structure;
- a "medium-level" that encodes semantic features such as quotations, names, places, and works;
- an "advanced-level" that adds an "editorial" layer, addressing abbreviations, corrections, and critical notes.

The base-level encoding has been applied to all texts, while the medium-level has been implemented for only some. The advanced-level remains experimental, having been applied to just a limited selection of texts.

The documentation fails to specify whether an existing model or standard was used to define the metadata. Additionally, it lacks a clear indication of the ALIM database's completion level and future plans for long-term preservation.

Moreover, when switching to the English version of the website, some content remains in Italian, such as the metadata labels.

FAIRness evaluation summary

Strengths

- The database is open access, enhancing accessibility for researchers.
- Texts are encoded in the standard XML/TEI format, promoting interoperability.

- Users can download all texts in various standard formats.
- The documentation concisely yet effectively presents the text collection and its contents.
- The technical documentation details the encoding format and schemas used for implementing the digital library.
- The texts are accessible via intuitive and user-friendly search forms alongside indexes.
- The XML/TEI files of the literary sources are deposited in ILC4CLARIN ensuring their long-term preservation.

Improvements

- Register and index the text collection in a searchable resource (e.g., OpenAIRE, CLARIN, DARIAH).
- Ideally, all XML/TEI files should be published on a non-proprietary, open-access platform (such as Zenodo)[41] to ensure their long-term preservation.
- The website's homepage lacks essential information about the database:
    - citation guidelines for the database;
    - date of the last update;
    - clear indication of the database's status, completion level, and number of collected texts;
    - link to the downloadable version of the text collection.
- Change the texts' licence terms from CC BY NC to a more FAIR-compliant licence such as CC BY SA.
- The documentation lacks information about future plans for long-term preservation.
- Support permalinks for texts.
- Assign a persistent identifier (e.g., DOI, Handle) to the database.
- Maintain a proper changelog to document all updates and changes.
- A RESTful API would be very valuable for retrieving text metadata and performing textual queries.

References

Boschetti, Federico, Riccardo Del Gratta, Monica Monachini, Marina Buzzoni, Paolo Monella, and Roberto Rosselli Del Turco. 2020. '"Tea for Two": The Archive of the Italian Latinity of the Middle Ages Meets the CLARIN Infrastructure'. *CLARIN Annual Conference*, 37–46. https://doi.org/10.3384/ecp1805.

Ferrarini, Edoardo. 2017. 'ALIM ieri e oggi.' *Umanistica Digitale*, no. 1 (October). https://doi.org/10.6092/issn.2532-8816/7193.

Biblioteca Italiana

Biblioteca Italiana (website: http://www.bibliotecaitaliana.it/) is a digital library that collects representative texts of Italian tradition and literature from the Middle Ages to the 20th century. Originating in 1996 from the work of a consortium of sixteen Italian universities, the Centro Interuniversitario Biblioteca Italiana Telematica (CIBIT), the digital library is now

[41] https://zenodo.org/.

directed by Beatrice Alfonzetti and Stefano Asperti, with Amedeo Quondam as its founder and president. The current website, launched in 2016, is hosted by the Department of Literature and Modern Cultures at La Sapienza University of Rome. This version of the digital library was developed in collaboration with MiBAC - Direzione Generale Biblioteche e Istituti Culturali,[42] ICCU - Istituto Centrale per il Catalogo Unico,[43] and BEIC - Biblioteca Europea di Informazione e Cultura.[44]

The digital library offers three distinct sections or collections. "BibIt", the core component of Biblioteca Italiana, contains 1,632 complete works in text format. These are based on authoritative scholarly editions, encoded in XML/TEI, and are all freely accessible, downloadable, and searchable. BibIt provides tools for contextual searches, proximity searches, full-text searches, and the creation of dynamic concordances in KWIC (Keywords in Context) format.

"Scrittori d'Italia" is the digital reproduction of the eponymous book series, founded in 1910 by the Laterza publishing house. It comprises 179 works (in 287 volumes), totalling 125,171 text-images.

Lastly, "Incunaboli in volgare" features 1,604 Italian incunables, amounting to more than 200,000 images. These incunables are freely available for online consultation and are accompanied by technical and management metadata.

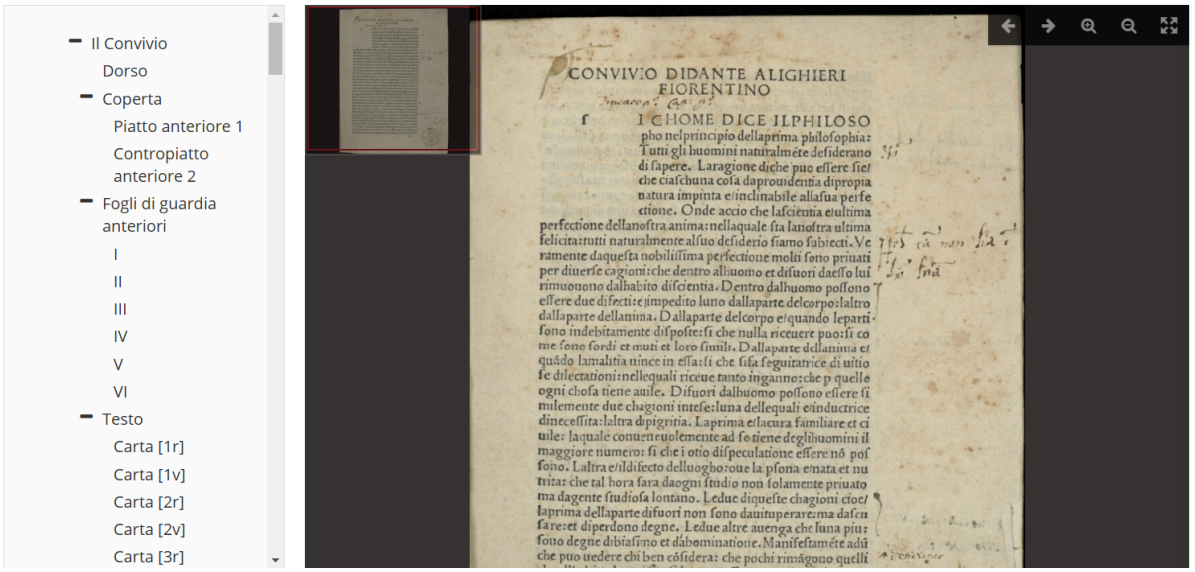
Biblioteca Italiana's text collections are accessible exclusively through their website, which is based on Muruca. Users can access all three sections via a unified list and search form. The list offers filtering options by collection, genre, historical period, author, and title. While metadata searches (e.g., title, author, publisher, etc.) are available across all collections, full-text search functionality is limited to the "BibIt" collection.

While primarily targeting scholars, Biblioteca Italiana's text collection also serves as a valuable teaching resource.

As noted earlier, only the works in the "BibIt" collection are available as XML/TEI files. The website offers a handbook for the XML/TEI (P4) encoding, along with the corresponding DTD schema and its extensions. The incunables and volumes from "Scrittori d'Italia" are accessible solely as images, which users can navigate using interactive indexes, as shown in the illustration below.

[42] http://librari.beniculturali.it/it/.
[43] https://www.iccu.sbn.it/it/.
[44] https://www.beic.it/it.

Alighieri, Dante
Il Convivio

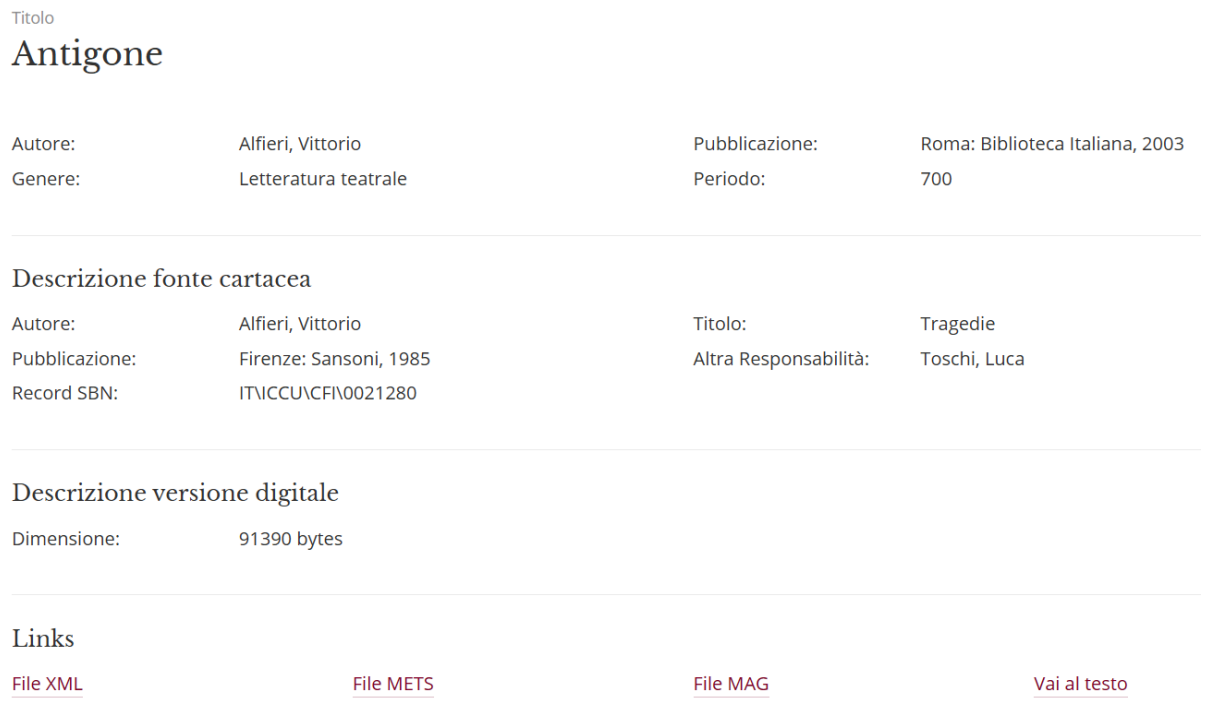


Screenshot of the incunable *Convivio* by Dante Alighieri published by Biblioteca Italiana.

The images from both "Incunaboli in volgare" and "Scrittori d'Italia" collections do not implement the IIIF[45] framework, and their licensing terms and reuse permissions are not clearly specified.

Each text is associated with a concise yet comprehensive set of metadata. Some metadata fields are common across all three collections (such as author, title, genre, and historical period), while others are specific to the collection to which the text belongs. For the complete description of the print sources of the published texts, Biblioteca Italiana relies on integration with SBN for the Bibit and Scrittori d'Italia sections; for the Incunaboli section, it refers to the descriptive record of the Incunabula Short Title Catalogue (ISTC) of the British Library and to the General Index of Incunabula in Italian Libraries (IGI). The print sources used in the BibIt collection are typically critical editions, though this is not explicitly stated in the metadata. Biblioteca Italiana's editorial work is limited to XML/TEI-encoding of the published texts from these print sources, without additional editing.

[45] See the text collections' standards and guidelines section.



Screenshot of the metadata associated with the text of Vittorio Alfieri's "Antigone" published in the BibIt collection.

The architecture of Biblioteca Italiana is based on the logical model OAIS (Open Archival Information System).[46] The metadata management system is based on the METS[47] framework, integrated with a series of auxiliary sub-schemas, including the MODS[48] schema for the bibliographic description of digital documents and their original sources, and the MIX[49] schema for the description of technical metadata for digital images. From the metadata stored in the METS records, metadata in TEI Header format are generated and inserted into the XML/TEI files of the texts, as well as metadata in MAG[50] format—a schema developed at the Central Institute for the Union Catalogue of Italian Libraries and Bibliographic Information (ICCU)—for interoperability with other national projects. The metadata for each text is available for download in various formats. Notably, the text and its metadata are presented on separate web pages. While the metadata page includes a link to the text page, the text page lacks a direct link back to the metadata page.

Each text appears to have a unique identifier for internal use. This identifier is present in the URL of the webpage displaying the text. However, the URL is not provided as a permalink in the metadata associated with the text.

FAIRness evaluation summary

Strengths

- The database is open access, enhancing accessibility for researchers.

[46] See references in the text collections' standards and guidelines section.
[47] http://www.loc.gov/standards/mets/.
[48] http://www.loc.gov/standards/mods/.
[49] http://www.loc.gov/standards/mix/.
[50] http://www.iccu.sbn.it/opencms/opencms/it/.

- Metadata are encoded in standard formats, promoting interoperability with existing national and international catalogues.
- Texts are encoded in the standard XML/TEI format, promoting interoperability.
- Users can download the metadata of all texts and the TEI-encoded texts from the BibIt collection.
- The documentation concisely yet effectively presents the three collections and their contents.
- The technical documentation details the encoding formats, standards, and schemas used for implementing the digital library.
- The texts are accessible via an intuitive and user-friendly search form.

Improvements

- Register and index the text collection in a searchable resource (e.g., OpenAIRE, CLARIN, DARIAH).
- Ideally, the XML/TEI files should be published on a non-proprietary, open-access platform (such as Zenodo)[51] to ensure their long-term preservation.
- To enhance interoperability and long-term durability, update the texts' TEI-encodings from P4 to P5.
- The landing page (bibliotecaitaliana.it) lacks essential information about the database:
    - citation guidelines for the database;
    - licence terms for the available texts;
    - clear indication of the database's status, completion level, and future plans.
- The "Incunaboli in volgare" and the "Scrittori d'Italia" collections would benefit from implementation via the IIIF framework, allowing for image reuse—currently subject to copyright restrictions.
- A RESTful API would be very valuable for retrieving text metadata and performing textual queries.
- Support permalinks for texts.
- Assign a persistent identifier (e.g., DOI, Handle) to the database.
- Maintain a proper changelog to document all updates and changes.

References

Quondam, Amedeo. 2021. 'Memorie per una storia dell'italianistica digitale: «Biblioteca italiana»'. *Griseldaonline* 20 (2): 137–47. https://doi.org/10.6092/issn.1721-4777/12360.

Musisque Deoque

"Musisque Deoque: Un archivio digitale di poesia latina, dalle origini al Rinascimento italiano" (MQDQ, website: https://www.mqdq.it/) is a comprehensive database of Latin poetry. Established in 2005,[52] this digital archive was designed to enable researchers to

[51] https://zenodo.org/.
[52] Paolo Mastandrea, a professor at the University Ca' Foscari of Venice, initiated the project with four PRIN grants from the Italian Government (1999, 2001, 2005, and 2007).

explore texts not only in their authoritative versions but also to examine textual variations found in critical apparatuses, as stated on the database's homepage:

> "At present, main collections of classical texts have been transferred onto digital device while resources, mostly online, allow quick lexical searches. In most cases, however, search engine inquiry only provides results key-words inside a fix and 'authoritarian' text. *Musisque Deoque* set out to overcome this limitation, making it possible to find not only the forms chosen and reported by the reference edition, but also the variants presented in the apparatus and selected under the responsibility of a 'digital editor'."

Over the years, the initial database expanded with new textual archives: "Carmina Latina Epigraphica" for Latin inscriptions, "Poeti d'Italia in lingua latina" for Italian poets who composed in Latin, and "Hellenica" for ancient Greek poetry.

In 2018, the Venice Centre for Digital and Public Humanities relaunched the project as "MQDQ Galaxy" to ensure long-term sustainability and open access to the MQDQ archives. This recent funding enabled API access to the MQDQ database. The project now involves a large group of researchers coordinated by Italian universities: Ca' Foscari University of Venice (Prof. Paolo Mastandrea), University of Calabria (Prof. Raffaele Perrelli), University of Parma (Prof. Gilberto Biondi), University of Perugia (Prof. Loriano Zurli), and University of Naples Federico II (Prof. Valeria Viparelli).

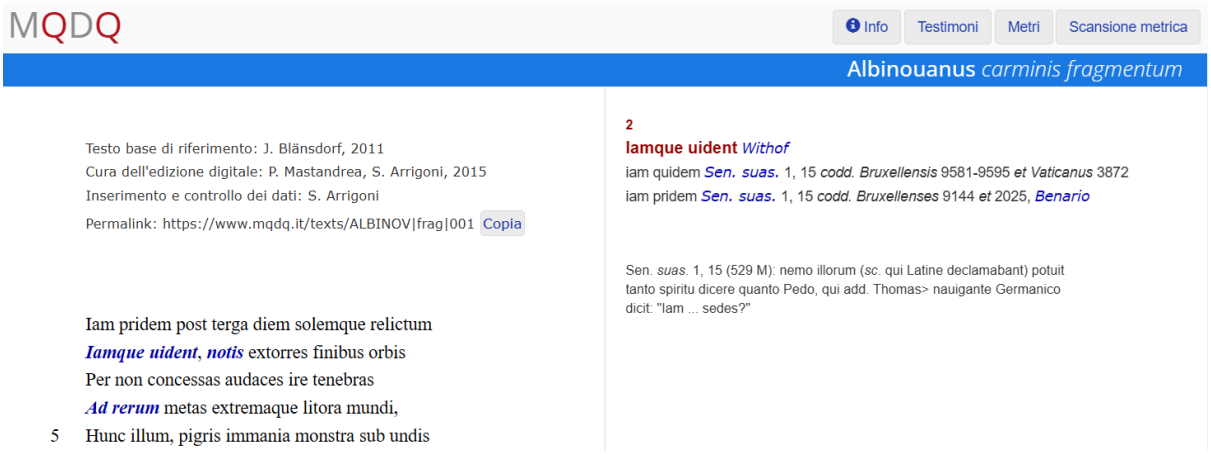
The MQDQ website is available in Italian and English. However, its landing page lacks some crucial information: citation guidelines for the digital archives, a standard licence for the texts, and an external identifier like a DOI code. The footer provides an official email address and a brief update date. A more detailed changelog is available on the Ca' Foscari University website,[53] though it's not formatted as a proper changelog.

MQDQ's primary aim is to provide an extensive corpus of Latin poetic texts for advanced lexical searches, including both edited texts and critical apparatuses. While primarily targeting scholars, MQDQ's authoritative and comprehensive text collection also serves as a valuable teaching resource.

Users can access the digital archive through various means: a general search interface focusing on lexical and metrical features, an alphabetic index, a chronological index, and specialised search features for lexical and metrical co-occurrences.

From a technical perspective, MQDQ consists of digital editions encoded in XML/TEI. However, the standard and format used for the texts' metadata are not clearly specified. The digital editions are based on previous printed editions. MQDQ editors revise, transcribe, and mark up these printed editions in XML/TEI. Each text includes information about the base text, the digital edition's editor(s), the data curator(s), and a permalink. Furthermore, each text is accompanied by a critical apparatus, a list of witnesses, the meter, and a metrical scan. Within the text body, portions subject to variation in the tradition are highlighted. When users click on these highlighted portions, the corresponding entry from the critical apparatus appears on the side.

[53] https://pric.unive.it/projects/mqdq-galaxy/home#c11784.



Screenshot of the edition of the "carminis fragmentum" by Albinouanus Pedo as presented on the MQDQ website (https://www.mqdq.it/texts/ALBINOV|frag|001).

The bibliographic references for the base texts and witnesses are neither provided in full nor linked to external web resources. Additionally, the website lacks a comprehensive bibliography.

The data modelling process is not documented, and the information structure within the database is left implicit.

The metrical scans of the texts were developed as part of the Pedecerto[54] project, which created a tool for automatic verse meter analysis and a search function that queries the entire MQDQ database. These metrical scans are available for download from the MQDQ website as XML/TEI files.[55]
As mentioned earlier, the MQDQ database comprises three distinct archives: Carmina Latina Epigraphica, Poeti d'Italia in lingua latina, and Hellenica. These archives vary in terms of completeness and functionality. Hellenica offers a limited range of works, accessible through the MQDQ portal via a dedicated alphabetic index[56] and search feature.[57] Poeti d'Italia in lingua latina presents a broader range of works, explorable through chronological and alphabetical indexes, a search feature, and an index organised by metrical scheme.[58] Carmina Latina Epigraphica is a more focused collection, accessible through its own index and the platform's shared indexes and search feature. This sub-collection is uniquely presented in detail from a theoretical perspective on a dedicated web page.[59]

The overall scope of the MQDQ text collection is not clearly defined. The homepage does not indicate the number of collected texts or tokens. However, the Pedecerto website reports that the MQDQ database contains 345,996 lines.[60] It is difficult to assess the database's completeness relative to the research project's goals. While one might assume the project is ongoing and that the database owners intend to preserve and maintain it long-term, the website does not explicitly state these intentions or future plans.

[54] https://www.pedecerto.eu/public/.
[55] https://www.pedecerto.eu/public/pagine/autori.
[56] https://mizar.unive.it/hellenica/public/indici/autori/idautori/1.
[57] https://mizar.unive.it/hellenica/public/ricerca/avanzata.
[58] https://www.poetiditalia.it/public/indici/metri.
[59] https://www.mqdq.it/public/ce/presentazione.
[60] Information taken from https://www.pedecerto.eu/public/pagine/arte.

The complete MQDQ database can be downloaded as a zipped folder containing XML/TEI files from the CLARIN website.[61] These files are available under the CC BY-SA 4.0 licence. The downloadable version on the CLARIN platform dates from 2021, and there is no indication of a more recent version available for download. However, the website states that the database was last updated on October 21, 2024.

The MQDQ database can be queried through RESTful APIs developed in recent years as part of the "MQDQ Galaxy" project. Access to these APIs is restricted to authenticated users. The RESTful API provides access to data such as the total number of authors, verses by meter, and multiword occurrences.

FAIRness evaluation summary

Strengths

- The database is open access, enhancing accessibility for researchers.
- Texts are encoded in the standard XML/TEI format, promoting interoperability.
- Users can download the entire text collection and metrical scans.
- The database offers multiple access points: a graphical user interface (MQDQ website) and RESTful APIs.
- Both simple and sophisticated search queries are supported.
- The website features comprehensive indexes for easy navigation.
- Texts are scholarly edited and include essential scientific information (author, title, editors, base text reference, witnesses, previous editions, and critical apparatus).
- Each text has a unique permalink, facilitating precise referencing.
- The text collection is indexed in OpenAIRE.

Improvements

- The landing page lacks essential information about the database:
    - citation guidelines for the database;
    - licence terms for the available texts;
    - clear indication of the database's status, completion level, and number of collected texts;
    - link to the downloadable version of the text collection;
    - link to the API.
- Assign a persistent identifier (e.g., DOI) to the database.
- Maintain a proper changelog to record all updates and changes.
- The scientific documentation needs expansion, detailing the editorial criteria, data model, and metadata standards used. It should also include a bibliography of the base texts and witnesses.
- Regular updates to the downloadable version of the database should be implemented, ideally every six months to a year.

References

[61] https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/OPEN-555.

Mastandrea, Paolo. 2007. 'Musisque Deoque. Un archivio digitale di poesia latina, dalle origini al Rinascimento italiano.' https://iris.unive.it/handle/10278/30263.

Venuti, Martina, Angelo Mario Del Grosso, Federico Boschetti, Luigi Tessarolo, Alessia Prontera, Dylan Bovet, Gianmario Cattaneo, and Valeria Melis. 2023. 'La "Galassia MQDQ:" un concetto di filologia tradizionale, digitale, sostenibile.' *Magazèn* 4 (1): 50. https://dx.doi.org/10.30687/mag/2724-3923/2023/07/003.

BUP - Digital Humanities

The editorial collection "Digital Humanities," published by BUP (Basilicata University Press), focuses on digital scholarly editions and databases. Professor Fulvio Delle Donne directs this collection.

The publications in the "Digital Humanities" collection are developed within the Department of Human Sciences, primarily by PhD candidates in the doctoral program "Culture e Saperi dell'Europa mediterranea dall'Antichità all'Età contemporanea." An editorial board guides this process, coordinated by Prof. Alessandro Di Muro and comprising Dr. Cristiano Amendola, Prof. Teofilo De Angelis, and Dr. Martina Pavoni.

Although not a traditional text collection, "Digital Humanities" and its set of digital scholarly editions serve as an interesting pilot for the ATLAS project for several reasons. Firstly, it exemplifies how to enhance and promote digital products by treating them with the same care given to printed editions. For instance, each digital scholarly edition is assigned an ISBN code. When an edition is available in multiple formats, such as XML/TEI and PDF, each format receives its own ISBN code. Additionally, the editorial collection itself has an ISSN code. Secondly, the collection has refined a publishing framework and editorial workflow based on XML/TEI encoding and the EVT visualisation software (see EVT), enabling the production of high-quality digital scholarly editions at a low cost.

The collection's website presents each digital scholarly edition with comprehensive metadata, including: the author and title of the edited work; the editor's name, ORCID, and brief biography; a short biography of the author; keywords; an abstract; page count; languages used; ISBN codes; links to both the PDF file and the XML/TEI edition (visualised with EVT); and the publication date.

The "Digital Humanities" collection also includes eight databases:

- DiLiBas-MA: Digital Libraries of Basilicata - Modern Age;
- DiLiBas-MOL: Digital Libraries of Basilicata - Modern Literature;
- DiLiBas-MELL: Digital Libraries of Basilicata - Medieval Latin;
- BiDiVi: Biblioteca Digitale Vichiana;
- ReDiAr: Reti Diplomatiche Aragonesi. Inventario digitale della Congiura dei Baroni (1485-1487);
- Human_It: Collecting, editing, analysing Italian humanist's letters (1400-1499);
- BiBas: Ricostruzione dei fondi librari delle Biblioteche monastiche e conventuali della Basilicata napoleonica, data-base digitale.

Although it is not stated explicitly, most of these databases are still under development or incomplete, while others are temporarily inaccessible. For this reason, they will not be analysed individually. Each database is presented on the editorial collection's website with metadata similar to that of digital scholarly editions. Some databases collect texts encoded in XML/TEI format. Additionally, the databases seem to have been implemented using a custom, ready-to-use publishing framework developed by the University of Basilicata's technical support team.

References

Amendola, Cristiano. 2021. 'Editoria universitaria, open access e nuove frontiere del lavoro umanistico: la Basilicata University Press e la collana "Digital Humanities."' *DigItalia: Rivista del Digitale nei Beni Culturali* 16 (2). https://doi.org/10.36181/digitalia-00042.

FAIRness evaluation summary

Strengths

- All publications are open access, enhancing accessibility for researchers.
- Digital scholarly editions are encoded in the standard XML/TEI format, promoting interoperability, and can be downloaded freely in PDF and XML/TEI formats.
- Digital scholarly editions are assigned one or multiple ISBN codes.
- All publications are introduced by a well-structured set of comprehensive metadata.

Improvements

- A crucial piece of information is missing: the status and level of completeness, especially for the databases.
- Deposit all editorial products in a repository for long-term preservation.

Standards and guidelines

- Standard encoding formats for various types of texts: TEI; MEI - music notation; Epidoc - inscriptions, and CEI - charters.
- Standard for the texts' metadata: Dublin Core.
- Standard for administrative metadata: METS (Metadata Encoding & Transmission Standard).
- Publication and sharing of facsimiles' digital images: IIIF. IIIF-compliant viewers: OpenSeadragon; Universal Viewer; Mirador.
- General recommendations to produce high-quality scholarly digital text collections: RIDE Criteria for Reviewing Digital Text Collections, version 1.0.
- National Information Standards Organization (NISO) Framework of Guidance for Building Good Digital Collections ((3rd edn, 2007).
- OAIS Introductory Guide (2nd Edition) and Reference model for an Open Archival Information System (OAIS).

Recommendations

- Organise texts in sub-collections to guide users in consulting texts (e.g., ALIM, Perseus Digital Library)[62] or to add value to the products of different research projects that are published within the collection (e.g., Mirabile).[63] In this second case, provide users with crucial information about each research project, including the project's objectives, contents, methodologies, editorial work, and progress status.
- Encode texts in XML/TEI or other standard non-proprietary formats established in digital humanities. When detailed encoding is impractical due to time or resource constraints, consider implementing a "light" encoding (e.g., ALIM, Biblioteca Italiana).
- Always cite the source text used for the text preparation, providing a complete bibliographic reference and/or a link to the descriptive web resource and specifying the nature of the text (e.g., manuscript, typescript, previous printed edition, previous digital edition, etc.).
- Use existing standards (e.g., Dublin Core) when defining metadata for texts.
- In the documentation, specify collection criteria and editorial criteria, stating the philological methodologies applied and the edition type (e.g., diplomatic transcription, critical edition).
- On the landing page, include citation guidelines for the database, the number of available texts, indicating the collection's completeness relative to its scientific objectives and, if applicable, a roadmap about the evolution of the text collection and the tools for its exploration.
- Facilitate text exploration through search functionalities and indexes.
- Link authors' and works' records to corresponding authority records if available, e.g., VIAF, Wikidata[64] (e.g., The Perseus Catalog).[65]
- Define a data management plan that includes actions to ensure long-term preservation of the text collections.

References

Fenlon, Katrina, Jacob Jett, and Carole L. Palmer. 2017. 'Digital Collections and Aggregations.' *DH Curation Guide: a community resource guide to data curation in the digital humanities*. https://guide.dhcuration.org/contents/digital-collections-and-aggregations/.

Johnston, Pete, and Bridget Robinson. 2002. 'Collections and Collection Description.' Collection Description Focus Briefing Paper, 1. Bath: UKOLN.

Lavoie, Brian. 2000. 'Meeting the Challenges of Digital Preservation: The OAIS Reference Model.' OCLC Newsletter, no. 243 (January/February): 26-30. (archived October 2023). https://cdm15003.contentdm.oclc.org/digital/collection/p267701coll28/id/527/rec/27.

Digital scholarly editions

[62] https://www.perseus.tufts.edu/hopper/.
[63] https://mirabileweb.it/content/info.
[64] https://www.wikidata.org/.
[65] https://catalog.perseus.org/.

To introduce digital scholarly editions we rely on Sahle's (2016) well-known definition:

"Scholarly digital editions are scholarly editions that are guided by a digital paradigm in their theory, method and practice".

According to Sahle, one of the major implications of the above mentioned "digital paradigm" is that a digital edition "as a publication is a process rather than a product. It grows incrementally not only before its final release, but also during its availability to the public" (*ibid.*).

The digital paradigm has not changed the editorial workflow completely (Mancinelli and Pierazzo, 2020), but new activities are necessary in order to produce a digital edition alongside the ones (e.g., transcription, collation) usually applied in the printed-oriented workflow.[66] These new activities, namely data modelling, digitisation and encoding, are part of the "source-output conceptual and technological model" (*ibid.*), where the source is an annotated text, usually encoded in XML/TEI, from which various outputs in multiple formats are derived.

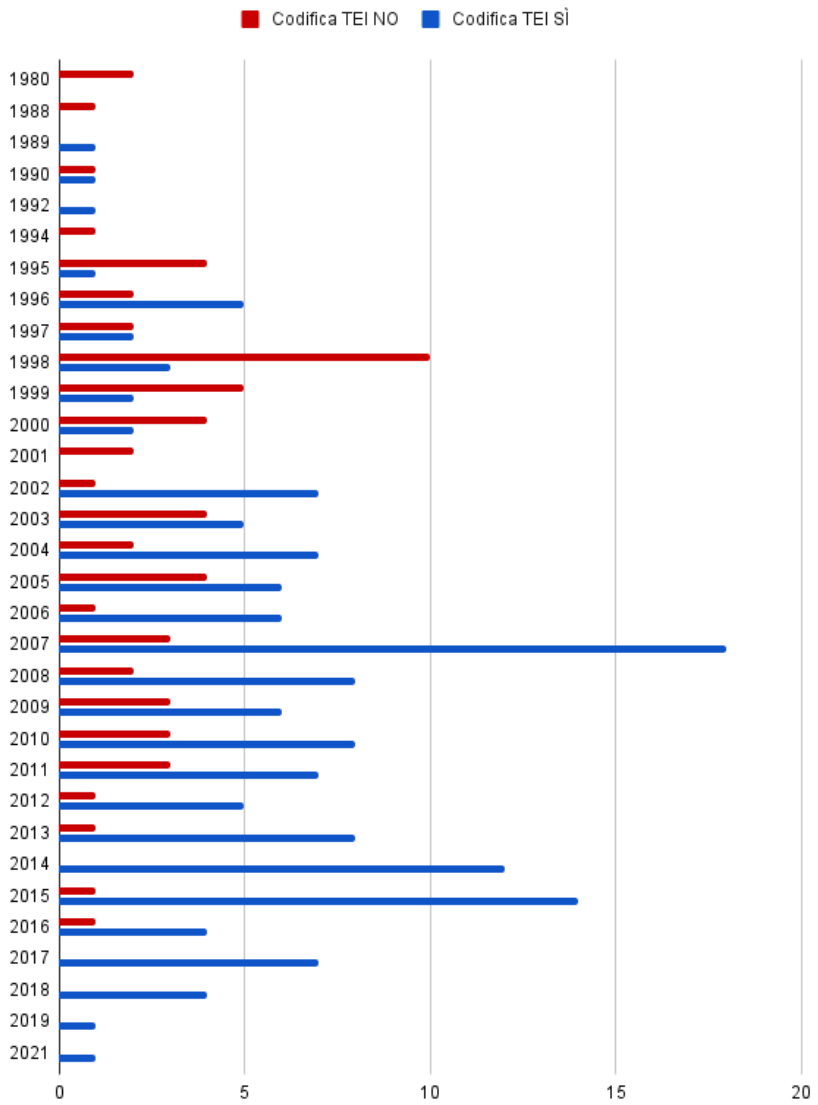
Each edition can be considered as a model or a picture of the relationship between the text and its tradition. Text editing often unveils extensive data not only about the textual tradition, but also about the work, author, content (named entities, lexicon, etc.), sources, and the text's reception. While it is challenging for a single project to exhaustively explore all these aspects, editors typically focus on specific areas aligned with their research goals. Adhering to international standard models and thoroughly documenting how the data was modelled is crucial to produce FAIR digital editions.

Documents' digitisation is a necessary step when they are not already available in electronic form. Digitised documents serve both as a starting point of the editorial workflow to (automatically) transcribe the texts and as part of the final product, in order to provide readers with the possibility of verifying first-hand editors' readings. Thanks to IIIF and its compliant viewers it is now easier to integrate the attestations' facsimiles in a digital edition.

The encoding step aims at producing one or multiple files containing the data of the edition, i.e. the text(s) and the annotations that document editorial interventions and allow readers to interpret the text(s). The choice of using XML/TEI as the encoding format is not always straightforward. Editors often desire to use TEI but lack the technical skills for implementation. When funds and resources are available, this issue is frequently addressed by either entrusting the encoding to researchers with digital philology expertise or by developing bespoke WYSIWYG[67] editing tools that hide the TEI encoding from editors (see Pierazzo 2019 for the definition of "haute couture editing"). However, the TEI guidelines have gradually emerged as the "de facto standard" (Mancinelli and Pierazzo, 2020). The following graph[68] compares the number of digital scholarly editions, catalogued in Franzini's

[66] See also: Leonardi, Lino. 2021. "Filologia digitale del Medioevo italiano." Griseldaonline 20 (2): 77–89. https://doi.org/10.6092/issn.1721-4777/12817.
[67] WYSIWYG (What You See Is What You Get) refers to an interface that allows users to view the finished result.
[68] The graph is taken from: Martignano, Chiara. 2023. "Un modello concettuale per le edizioni critiche digitali." Tesi di Dottorato, Siena: Università degli Studi di Siena. http://dx.doi.org/10.25434/chiara-martignano_phd2023.

Catalogue of Digital Editions (updated to March 2022), that are based on XML/TEI against the number of those that are not TEI-compliant divided by their initiation year.[69]



Considering the trend depicted in the graph and the fact that the TEI guidelines have been continuously supported and developed by a large international scholarly community, it is logical to strongly recommend XML/TEI as the standard text encoding format. In case of peculiar types of texts, other TEI-compatible and XML-based models may be used, namely: MEI (Music Encoding Initiative) for music notation, Epidoc (Epigraphic documents in TEI XML) for inscriptions, and CEI (Charters Encoding Initiative) for medieval charters.[70]

[69] In the first years shown in the chart, corresponding to the period before and immediately after TEI's inception, the numbers are understandably low. Growth begins in 2002, with notable peaks in 2007 and again between 2014 and 2015. The graph's trend likely reflects biases in the catalogue's compilation timing and methods. Nevertheless, where specified in the catalog, XML/TEI has progressively emerged as the predominant encoding format.

[70] For more information refer to the digital scholarly editions' standards and guidelines section.

Questions may arise about the actual FAIRness of a digital edition using XML/TEI, particularly regarding interoperability and reuse. The TEI guidelines offer flexibility in element interpretation and usage, potentially leading to diverse encodings of identical textual phenomena across editions. However, this issue can be readily addressed by providing users with the applied schema and clearly describing editorial criteria in the documentation.

Addressing the initial concern of editors struggling with TEI, two strategies are now easily adoptable. First, use editing software that automatically exports data in XML/TEI, favouring generic tools over custom-developed ones. Numerous free tools are available for various editorial tasks like collation, transcription, and lemmatisation.[71] The second approach involves converting data from other formats to TEI. TEI has long offered OxGarage, recently relaunched as TEIGarage,[72] for this purpose.[73]

Finally, the last step of the "source-output" model is the presentation (or data visualisation). Various outputs can be derived from the edition's source file(s), e.g., PDF, e-book formats. However, most of the time the output is HTML pages generated by software tools such as EVT and TEIPublisher.[74] To foster a collaborative and cumulative research environment, a digital edition should provide users with downloadable outputs in reusable formats and also strive to connect with and build upon existing scholarly work. However, to the authors' notion, there is yet no successful example of data reuse in scholarly editing.

The digital editions we selected as pilots are:

- **VaSto, VArchi STOria fiorentina**,[75] an international project that aims at producing the annotated digital edition of the *Storia fiorentina* by Benedetto Varchi (1503-1565). The edition is TEI compliant and visualised through the open source software EVT, which has been customised to provide support for genetic editions.
- **Codice Pelavicino Digitale**[76] is a project providing a digital scholarly edition of a historical document of crucial importance for the Italian Cultural Heritage with regard to historical research of the XII-XIII century. The edition is TEI compliant and visualised through the open source software EVT.
- **Digital Edition of Aldo Moro's works**,[77] a critical edition of Aldo Moro's published and unpublished texts, with their historical introduction. The edition reused well-known ontologies in the CH domain and is available in RDFa and TEI/XML format.[78]

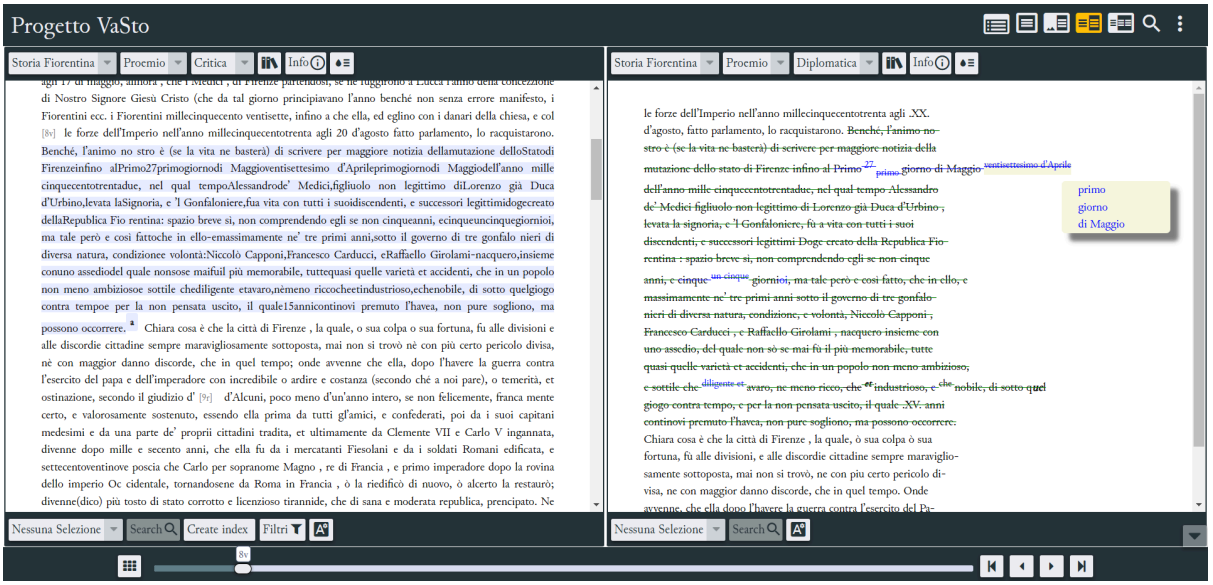
---

[71] A list of software for digital scholarly editing is available here.
[72] https://teigarage.tei-c.org/.
[73] While this method may not perfectly translate all textual phenomena (e.g., strikethroughs becoming `<hi rend="strikethrough">` instead of `<del>`), it is still a significant step towards enhancing the edition's FAIRness.
[74] Created in 2015 by the company eXist Solutions in the context of the Swiss e-codices association, TEIPublisher (https://teipublisher.com/index.html) allows you to publish XML files on the web (mainly according to the TEI standard), generating a website, or convert documents to other formats, such as PDF and ePUB.
[75] https://dharc-org.github.io/progetto-vasto/.
[76] https://pelavicino.labcd.unipi.it/.
[77] https://aldomorodigitale.unibo.it/.
[78] Leges Langobardorum, the diplomatic-interpretative edition of two witnesses of the Edict of Rothari, a renowned collection of Lombard laws, will also be used as pilot for the ATLAS project. Currently in preparation, the edition will be described in the next version of the whitebook. Roberto

VaSto - VArchi, STOria fiorentina. Edizione digitale

The VaSto project (VArchi, STOria fiorentina. Edizione digitale, website: https://dharc-org.github.io/progetto-vasto/) aims to publish the uncensored version of Benedetto Varchi's *Storia fiorentina*, as attested in the manuscript Corsiniano 1532 (Biblioteca Nazionale dei Lincei e Corsiniana, Rome). VaSto is a collaborative effort between /DH.arc[79] (University of Bologna) and Concordia University, with support from the CarisBo foundation, led by professors Dario Brancato and Paola Italia.

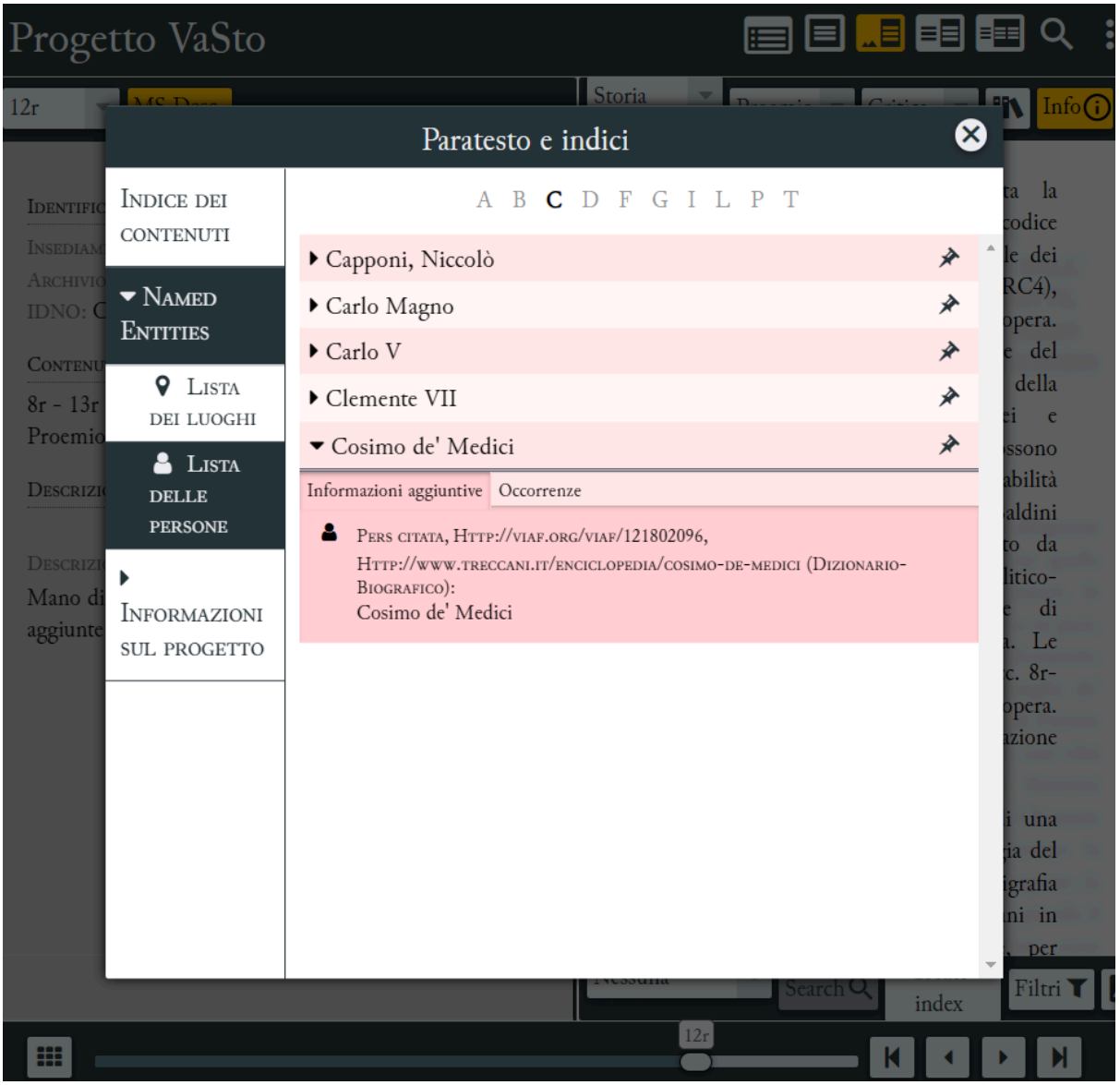
The project's primary output is an XML/TEI-encoded digital edition of the *Storia fiorentina*. To date, only the work's preface has been published as a pilot version. The initial release in May 2020 used the first beta version of EVT 2.0 software, while the current version, released a month later, employs the second beta. The edition is both diplomatic—presenting the text as it appears in the Corsiniano 1532 manuscript alongside digital images—and critical-genetic, incorporating editorial interventions that restore the author's intended meaning and enhance readability. EVT's interface allows users to toggle between these two edition levels or compare them side-by-side, as shown below.



The edition is also "annotated", as stated in the documentation, because names of people and places mentioned in the edited work are marked up and linked to their respective VIAF entries and/or descriptions on external web sources when available.

Rosselli Del Turco (University of Turin) and Marina Buzzoni (Ca' Foscari University of Venice) are curating the Codex Vercellensis (Vercelli, Biblioteca Capitolare Eusebiana, CLXXX)VIII) and the Codex Eporedianus (Ivrea, Biblioteca Capitolare, XXXIV.5), respectively. The edition will feature digital images of the manuscripts and will be published within ALIM using the visualisation software EVT. Prepared in XML/TEI format, the edition will adhere to ALIM's shared criteria for advanced text encoding, including corrections and critical notes (see ALIM). Currently, the edition's page in ALIM's website (http://alim.unisi.it/editto-di-rotari/) offers illustrative images of the expected outcome, along with a brief description and publication references.

[79] https://centri.unibo.it/dharc/en.



VaSto is also a collaborative edition: it was implemented with the participation of students from the *Scholarly Editions* lab held at the University of Bologna in 2020. The edition's XML/TEI document is freely accessible and downloadable from the project's GitHub repository.[80] This single document contains the diplomatic edition, the critical edition, and the annotated named entities.

The editors aim to complete the edition of the *Storia fiorentina*, presenting both the manuscript version of the author's last will and the censored version that has become the work's *vulgata*. However, the documentation lacks information about the full edition's publication timeline and the current status of the editorial work.

In addition to the digital edition, the VaSto project has produced three secondary outcomes:

80 https://github.com/ValentinaPasqual/ProgettoVasto/blob/master/evt2beta2/data/text/pilot_proemio.xml.

- a timeline developed with TimelineJS,[81] depicting the writing process of the *Storia fiorentina* and key historical events described in the work;
- a map created with Leaflet,[82] illustrating the places mentioned in the *Storia fiorentina*;
- VastoCollection, which showcases digital images of the RC4 manuscript and portraits of people mentioned in the *Storia fiorentina*. This collection is built using Omeka[83] and IIIF, with catalographic records structured according to the Dublin Core standard.

The website provides comprehensive documentation from both philological and technical perspectives. The philological section[84] offers a description of the edited text and its history, the author's biography, a bibliography, and a list of witnesses—though these are not linked to external web sources. The scientific section[85] includes a detailed guide for interacting with the edited text. That section also explains how the EVT 2.0 software was customised to meet the project's requirements, which is particularly valuable for EVT's developers as it fosters synergy with the user community. While the documentation does not provide the encoding schema, it lists the TEI elements used to represent editorial (corrections, critical notes, etc.) and material (page and line breaks, deletions, etc.) phenomena. The website's copy requires revision, as typos are present—for instance, "IIF" instead of "IIIF". The credits page is well-structured and comprehensive, presenting not only the editorial team and collaborators but also the official releases, attribution, and licence (Creative Commons 3.0) under which the edition is available.

References

Brancato, Dario, Milena Corbellini, Paola Italia, Valentina Pasqual, and Roberta Priore. 2021. 'VaSto: un'edizione digitale interdisciplinare'. *Magazén* 2 (1): 139–69. https://dx.doi.org/10.30687/mag/2724-3923/2021/03/006.

FAIRness evaluation summary

Strengths

- The edition is open-access and open-source, promoting accessibility and transparency.
- The use of XML/TEI encoding, the standard for digital editions, enhances interoperability and data reuse.
- People mentioned in the edited text are linked to VIAF, leveraging semantic web technologies for improved contextualisation.
- Comprehensive documentation includes attribution, licensing information, and version history, ensuring clarity and proper credit.

Improvements

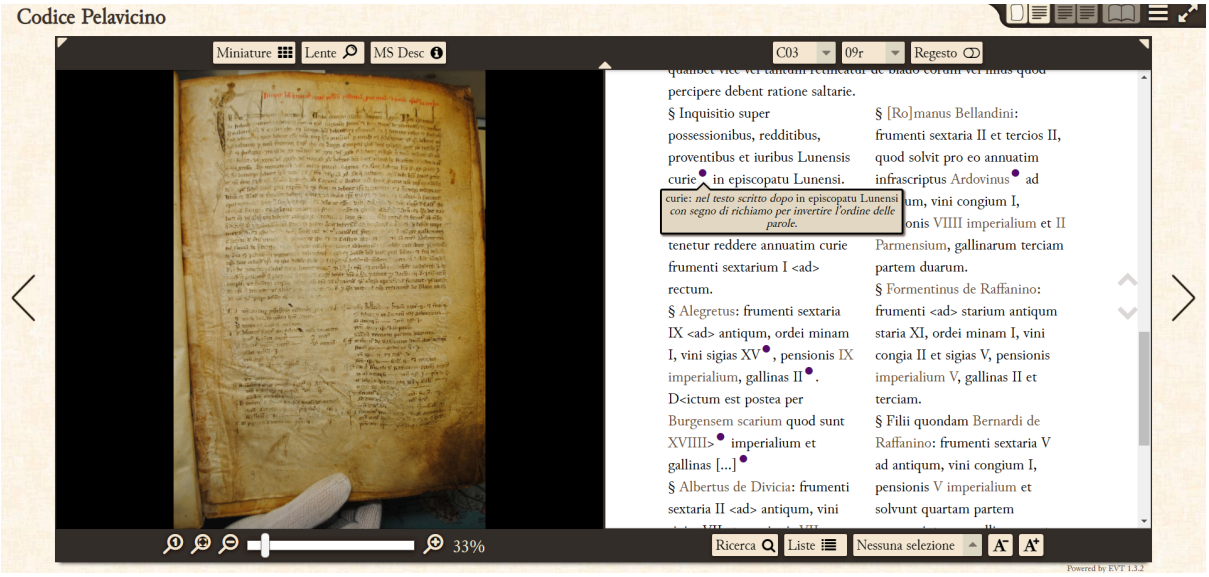
[81] https://timeline.knightlab.com/.
[82] https://leafletjs.com/.
[83] https://www.omeka.net/.
[84] https://dharc-org.github.io/progetto-vasto/Progetto.html.
[85] https://dharc-org.github.io/progetto-vasto/Funzionalit%C3%A0.html.

- Provide a clear indication of the edition project's status and timeline.
- Publish the edition's schema as a downloadable document.
- Release the pilot officially as a data dump on an external repository to enhance data versioning, and citability, and ensure long-term preservation.
- Register the digital edition in a searchable resource such as OpenAIRE to increase its findability.
- Polish the website's copy for clarity.

Codice Pelavicino Digitale

Codice Pelavicino Digitale (website: https://pelavicino.labcd.unipi.it/) is the digital documentary edition of the eponymous codex preserved at the Archivio Capitolare Lunense of Sarzana. Professor Enrica Salvatori (University of Pisa) initiated the project in 2014. Over the years, numerous collaborators have joined the extensive editing team, contributing to the codex transcription, encoding, and development of consultation tools. The edition is now complete, as clearly stated on the website's homepage, featuring 529 documents from the codex and a comprehensive set of accompanying materials. Currently, the edition undergoes periodic corrections, revisions, and progressive feature improvements.

The project's primary outcome is an open-access, TEI-encoded digital edition, visualised using EVT 1.0 software (version 1.3.2). The graphical user interface allows users to read diplomatic transcriptions of the documents alongside digital images of the respective folios. Each document is prefaced by a detailed, toggleable description. Within the transcribed text, users can interact with editorial notes (indicated by blue dots) and named entities.



Named entities mentioned in the text—specifically places, people, families, and organisations—can be explored via lists. Each entry in these lists includes a description and links to occurrences in the documents. In addition to the named entities' lists, users can access a glossary and a chronological index presenting the historical events mentioned in the documents, ordered by date or document. A full-text search function is also available.

To facilitate data reuse, the edition's XML/TEI files are available for download under the Creative Commons CC BY 4.0 licence from the ILC4CLARIN repository.[86] The named entities' lists can also be downloaded from the same platform.[87] However, the manuscript's digital images are not available due to copyright restrictions. Implementing the manuscript's facsimiles with IIIF and the named entities as linked open data would significantly enhance this already commendable work.

The edition's website offers a rich set of materials:

- interactive map of places mentioned in the codex;
- named entities' search form;
- list of all charters contained in the codex;
- notaries' list;
- comparative table of current and previous document numbering.
- timeline, developed with the TimelineJS tool, illustrating historical events mentioned in the codex.
- A TEI-based glossary of terms that were more challenging to interpret.

The documentation is comprehensive from both philological and technical perspectives, including attribution, licences, bibliography, and publications. The philological section clearly outlines the editorial criteria and standards applied to implement the digital edition, namely the Guidelines for Editors of Scholarly Editions[88] by the Modern Language Association and the Criteria for Reviewing Scholarly Digital Editions, version 1.1[89] published by the Institut für Dokumentologie und Editorik. While the technical section does not provide a downloadable TEI schema, it offers a detailed explanation of the elements used, illustrating how peculiar aspects of the texts, such as coins, professions, and notaries' signs, were encoded. This information can be invaluable for other edition projects of similar documents. Lastly, the documentation also illustrates how the visualisation software tool, EVT 1.0, was customised to meet the edition's requirements.

References

Salvatori, Enrica, Roberto Rosselli Del Turco, Chiara Alzetta, Chiara Di Pietro, Chiara Mannari, and Alessio Miaschi. 2017. 'Il Codice Pelavicino tra edizione digitale e Public History.' *Umanistica Digitale*, no. 1 (October). https://doi.org/10.6092/issn.2532-8816/7232.

FAIRness evaluation summary

Strengths

- The edition is open-access and open-source, promoting accessibility and transparency.

[86] http://hdl.handle.net/20.500.11752/OPEN-1012.
[87] https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/OPEN-1011.
[88] http://www.mla.org/cse_guidelines.
[89] http://www.i-d-e.de/aktivitaeten/reviews/criteria-for-reviewing-scholarly-digital-editions-version-1-1

- The use of XML/TEI encoding, the standard for digital editions, enhances interoperability and data reuse.
- Publication of the edition's XML/TEI files on an external repository, facilitating data reuse and ensuring long-term preservation.
- The digital edition is indexed in OpenAIRE.
- Active contribution to the development of the visualisation tool.
- Clear indication of the edition's status, licences, and attribution.
- Provision of indexes, search features, and various visualisation tools that enhance the discoverability of the contents.

Improvements

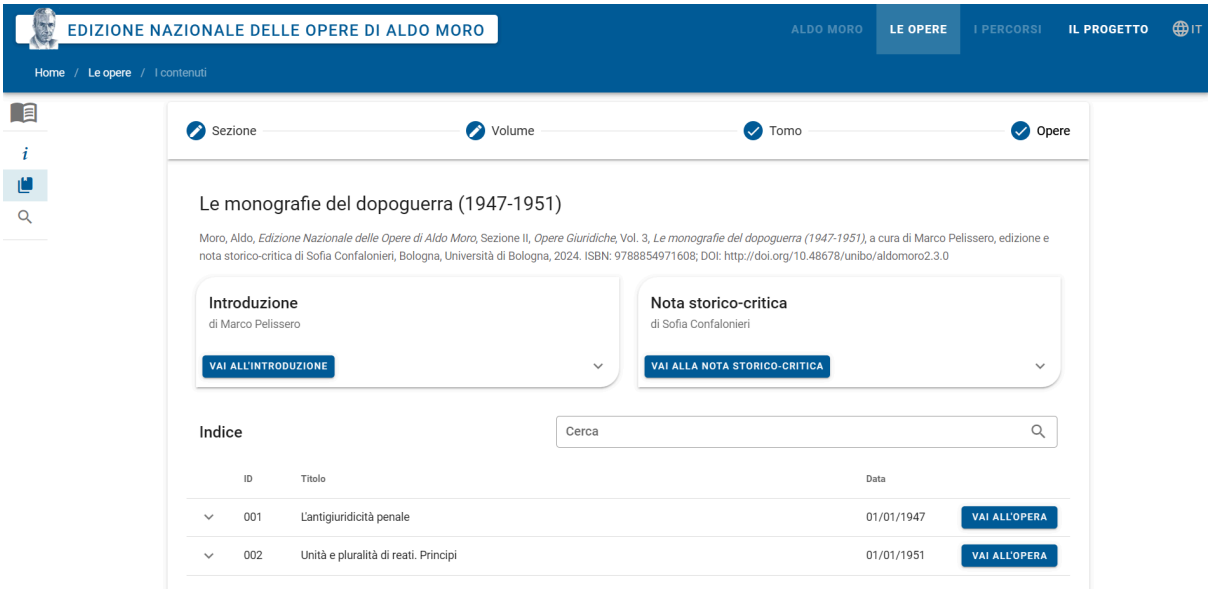
- Publish the manuscript's digital images using IIIF to enhance accessibility and interoperability.
- Implement named entities as linked open data to enhance the already high-quality scientific work.

Digital Edition of Aldo Moro's works

The National Edition of Aldo Moro's Works (website: https://aldomorodigitale.unibo.it/) is a collaborative effort involving numerous Italian universities and partner institutions. While scholars from various Italian universities curate the edition, a team of researchers from the University of Bologna manages its technical-scientific implementation. Initiated in 2021 and still ongoing, the project aims to publish all of Aldo Moro's works, both published and unpublished. To date, 806 works have been released across eight volumes, organised into two main sections: the first covering Moro's religious, journalistic, and political writings, speeches, and interviews; the second dedicated to his academic works. Both sections follow a chronological arrangement.

The published works are freely accessible on the edition's website. A banner displayed on the landing page informs users that the edition is under development, with new content and features forthcoming. Notably, the edition's bibliographic reference, complete with ISBN code and DOI, is prominently displayed in the page footer.

Users can access works through a search function, analytical indexes for people, places, organisations, and bibliographic references mentioned by Moro, or a step-by-step interface guiding them through the edition's volumes. Navigation proceeds from section to volume, tome, and finally to individual works. Each tome is preceded by its bibliographic reference and links to an introduction and historical-critical note, which are displayed in different web pages, each with its own DOI.



Alongside the title and the edited text, each work's edition comprises:

- an introductory section with an abstract, the full bibliographic reference of the edited work and of the witnesses;
- interactive lists of people, places, organisations, citations, and bibliographic references mentioned in the edited text, along with their occurrence counts;
- a metadata sheet detailing themes, data curator, author's roles, typologies, document status, date, event location, identifier, licence, and additional notes.

The texts are available for download in XML/TEI, HTML-RDFa, and PDF formats. The interface facilitates sharing via email and social media, and provides a copy function for the work's URL. Users can search works by title and keywords, and filter them based on document type, themes, Moro's roles, date, and publication status.

Beyond the corpus edition, the project has produced an RDF dataset of "structural, intertextual, and contextual data related to Aldo Moro's works". This knowledge base was used to design the edition's consultation interface and its search functionalities. The Turtle-encoded dataset is freely accessible and downloadable from the Zenodo repository.[90] Controlled vocabularies for roles, themes, and document types are also available on Zenodo in Turtle format. This data is integrated into the edition's documents using RDFa (Resource Description Framework in Attributes),[91] which embeds structured data within web pages. The added information includes:

- licence, persistent identifier (DOI), and bibliographic citation of the document;
- references to people, places, organisations, works cited by Moro or researchers, and quotations;
- mentioned entities, their attested forms, authority control based on Wikidata records, and controlled forms of personal names;
- notes comprising commentary by researchers and/or Moro himself.

[90] http://doi.org/10.5281/zenodo.5592157.
[91] Resource Description Framework in Attributes. https://www.w3.org/TR/rdfa-primer/.

The data modelling draws on several existing ontologies, including: Bibliographic Reference Ontology (BiRO),[92] Discourse Elements Ontology (DEO),[93] Dublin Core Metadata Terms (DCTerms),[94] FRBR-aligned Bibliographic Ontology (FaBiO),[95] Friend Of A Friend vocabulary (FOAF),[96] and Expression of Core FRBR Concepts in RDF (FRBRcore).[97]

The project's workflow is noteworthy for addressing a common challenge in digital edition projects. Works were initially transcribed in Word, then digitised, indexed, enriched with information, and published through a complex process involving multiple contributors. A custom software tool, KWICKWOCKWAC,[98] facilitated this process by converting Word documents to Web pages, enabling text markup, metadata insertion, and linking entities to existing Web authority records. This approach bypasses the frequent issue of editors lacking XML/TEI skills, which often leads to not adopting TEI for text encoding or, as in this case, developing custom editing tools.

The project's website offers comprehensive documentation on the edition's preparation, detailing the workflow, editorial criteria, data modelling, website development, and scientific outcomes. These outcomes encompass the corpus of edited works, the software code, the RDF dataset, controlled vocabularies, technical documentation on data modelling, and other materials produced during the edition's creation. To ensure long-term access and citability, all scientific outputs except the corpus are published on Zenodo under Creative Commons 4.0 licences. The website's "terms and conditions" page provides citation guidelines for the various outputs.

References

Barzaghi, Sebastian, Francesco Paolucci, Francesca Tomasi, and Fabio Vitali. 2024. 'KwicKwocKwac, a Tool for Rapidly Generating Concordances and Marking up a Literary Text.' arXiv. https://doi.org/10.48550/arXiv.2410.06043.

Barzaghi, Sebastian. 'National Edition of Aldo Moro's works' (RDF Dataset) (1.0.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5592157.

FAIRness evaluation summary

Strengths

- The edition is open access, allowing users to easily share and download the edited works in various standard formats.
- The edition uses standard and non-proprietary encoding formats.
- Indexes, a search function, and a user-friendly interface enhance the corpus' discoverability.

[92] http://purl.org/spar/biro.
[93] http://purl.org/spar/deo.
[94] http://purl.org/dc/terms/.
[95] http://purl.org/spar/fabio.
[96] http://xmlns.com/foaf/0.1/.
[97] http://purl.org/vocab/frbr/core#.
[98] https://aldomorodigitale.unibo.it/markup/.

- All scientific outputs are highly citable, thanks to publication on the Zenodo repository and ISBN assignment. Full bibliographic references are readily available.
- The edition's data is also accessible as an RDF dataset, employing semantic web technologies to facilitate data reuse and further studies.
- The digital edition is indexed in OpenAIRE.

Improvements

- If possible, consider publishing the whole set of the edition files in various formats (XML/TEI, PDF and RDF-a) on an external repository in order to facilitate the reuse of the textual corpus as a whole.

Standards and guidelines

- Standard encoding formats for various types of texts: TEI; MEI - music notation; Epidoc - inscriptions, and CEI - charters.
- Standard for the texts' metadata: Dublin Core.
- Publication and sharing of facsimiles' digital images: IIIF. IIIF-compliant viewers: OpenSeadragon; Universal Viewer; Mirador.
- General recommendations to produce high-quality digital scholarly editions: RIDE Criteria for Reviewing Scholarly Digital Editions, version 1.1
- Existing digital scholarly editions as examples or references for further studies:
    - Franzini, G. (2012-) *Catalogue of Digital Editions*, https://doi.org/10.5281/zenodo.1161425.
    - Sahle, Patrick et al., a catalog of Digital Scholarly Editions, v.4.112 2020ff, last change 2024-06-06.

Recommendations

- Use XML/TEI as the encoding format to promote interoperability. To facilitate encoding, consider these strategies:
    - Employ an editing tool that exports data as XML/TEI documents. Several tools are available for various stages of editorial work, including transcription of primary sources (e.g., eScriptorium,[99] FairCopy,[100] Transkribus).[101]
    - Convert from other data formats to XML/TEI using TEIGarage.
- Provide users with the edition schema and data model to facilitate data reuse. Create schemas easily with the TEI Roma tool.[102]
- Include photographic reproductions of the attestations when available, preferably through IIIF. This allows readers to verify the editor's readings.
- The documentation should include the editorial criteria and a detailed description of the textual tradition, accompanied by links to relevant web resources.
- For visualisation, prioritise existing and non-proprietary tools. When a new custom feature is necessary, collaborate on an existing project rather than starting from scratch.

[99] https://escriptorium.openiti.org/.
[100] https://www.faircopyeditor.com/.
[101] https://www.transkribus.org/.
[102] https://roma.tei-c.org/.

- For extensive and rich texts, provide indexes and a search function to improve discoverability. For texts rich in references to people, places, and entities, consider creating a semantic edition to highlight these aspects.
- Enable users to download the edition, including a print-ready PDF version when appropriate.

References

Barabucci, Gioele, Elena Spadini, and Magdalena Turska. 2017. 'Data vs. Presentation. What Is the Core of a Scholarly Digital Edition?' In *Advances in Digital Scholarly Editing*, 37–46. Sidestone Press. https://serval.unil.ch/notice/serval:BIB_09C6C598108A.

Ciula, Arianna, Øyvind Eide, Cristina Marras, and Patrick Sahle. 2023. *Modelling Between Digital and Humanities: Thinking in Practice*. Open Book Publishers. https://doi.org/10.11647/OBP.0369.

Driscoll, Matthew James, and Elena Pierazzo, eds. 2017. *Digital Scholarly Editing : Theories and Practices*. *Digital Scholarly Editing : Theories and Practices*. Digital Humanities Series. Cambridge: Open Book Publishers. http://books.openedition.org/obp/3381.

Eide, Øyvind. 2014. 'Ontologies, Data Modeling, and TEI.' *Journal of the Text Encoding Initiative*, no. Issue 8 (December). https://doi.org/10.4000/jtei.1191.

Franzini, G., S. Mahony, and M. Terras. 2016. 'A Catalogue of Digital Editions.' In *In: Pierazzo, E and Driscoll, M, (Eds.) Digital Scholarly Editing: Theories and Practices. (Pp. 161-182). Open Book Publishers: Cambridge, UK. (2016)*, edited by E. Pierazzo and M. Driscoll, 4:161–82. Cambridge, UK: Open Book Publishers. https://doi.org/10.11647/OBP.0095.

Mancinelli, Tiziana, and Elena Pierazzo. 2020. *Che cos'è un'edizione scientifica digitale*. Carocci.

Pierazzo, Elena. 2019. 'What Future for Digital Scholarly Editions? From Haute Couture to Prêt-à-Porter.' *International Journal of Digital Humanities*, May. https://doi.org/10.1007/s42803-019-00019-3.

Pierazzo, Elena. 2014. *Digital Scholarly Editing: Theories, Models and Methods*. http://hal.univ-grenoble-alpes.fr/hal-01182162.

Sahle, Patrick. 2016. 'What Is a Scholarly Digital Edition?' In *Digital Scholarly Editing*, edited by Matthew James Driscoll and Elena Pierazzo, 1st ed., 4:19–40. Theories and Practices. Open Book Publishers. https://www.jstor.org/stable/j.ctt1fzhh6v.6.

Tomasi, Francesca, Fabio Ciotti, Maurizio Lana, Fabio Vitali, Silvio Peroni, and Diego Magro. 2013. 'Dialogue and Linking between TEI and Other Semantic Models.' In *The Linked TEI: Text Encoding in the Web*, 145–58. Roma: DIGILAB Sapienza University & TEI Consortium. https://hdl.handle.net/11585/185113.

Software tools

In the field of digital humanities, software tools are often developed by individual researchers or project teams relying on time-limited funds. This leads to fragmented and discontinuous development, making it challenging to maintain tools in the long-term and update them to new standards and technologies. For example, the software EVT has been rebooted twice to accommodate technological changes. Other tools, like Collate[103] (relaunched as CollateX)[104], required a complete rewrite to remain accessible and usable.

Two key factors that can render a software tool inaccessible are the lack of documentation and the absence of technical support. If a tool's development is unstable, it is unlikely to provide users with adequate technical support. Paid applications typically offer both documentation and technical support, but may be inaccessible due to costs. Some applications offer trial periods or free plans suitable for small to medium-sized edition projects. For instance, Transkribus provides free transcription for 3,000 printed pages and 500 handwritten pages. For open-source software, inaccessibility often stems from a lack of technical documentation and poor source code readability.

Free software tools are more accessible and likely to be adopted by a larger audience, but only if they are user-friendly, customisable, and meet users' needs. Researchers often prefer developing new tools from scratch rather than adopting existing ones (Pierazzo 2019), to maintain full control over the tool's functionality and user interface. To encourage reusability, tools should be highly customisable and open-source. EVT is a successful example of an open-source tool that has been reused and further developed through collaboration with various research projects.

A key design approach for producing effective and reusable software is called "Domain Driven Design (DDD)" (Evans 2004). According to DDD, developers should involve domain experts since the early stages of the software's design process, in order to effectively capture the software's application domain. A tool may be more usable and effective if it is designed with an understanding of how researchers work and the characteristics of the objects they study (texts, art collections, photo archives, etc.).

Another crucial factor in developing robust and reliable software is applying best practices and methodologies from software engineering. However, in the field of Digital Humanities, not all software developers have a strong background in software engineering, which can lead to poorly implemented tools:

> "The development of applications in the field of Digital Humanities (DH) does not adequately take into account domain modelling, software design principles and software engineering methodologies. In fact, many systems developed in the context of DH-related projects have not been conceived to be modular, extensible, and scalable: they only tend to solve specific problems such as data-driven and project-oriented tools. In addition, most projects focus on the requirements of humanists (as end users), but leave out the needs of software developers" (Del Grosso et al. 2016).

[103] https://digitalmedievalist.org/2012/04/01/collate-text-editing-software/.
[104] https://collatex.net/.

Finally, software tools should leverage current technologies and adhere to digital humanities standards (such as XML/TEI). This approach ensures interoperability among various tools, fostering a more cohesive and efficient digital humanities ecosystem.

The software we selected as pilots are:

- **EVT (Edition Visualization Technology)**, http://evt.labcd.unipi.it/, is an open source software to visualise digital scholarly editions on the basis of TEI/XML-encoded documents. It is easy to configure and deploy on the Web (using HTML, CSS and Javascript for long-term support), it is fully customisable, and it includes several useful research tools.
- **Voyant Tools**, https://voyant-tools.org/, is an open-source, web-based text reading and computer-assisted analysis environment for scholars. It allows scholars to explore and query texts in several linguistic tasks and analyses such as word frequency, keyword analysis, and topic modelling.

EVT

EVT (Edition Visualization Technology, http://evt.labcd.unipi.it/) is an open-source, client-only software for visualising digital scholarly editions in XML/TEI format. Conceived by Roberto Rosselli Del Turco and developed by Digital Humanities students at the University of Pisa, the tool was initially created for the digital edition of the Vercelli Book.[105] It later evolved to accommodate various editions through collaboration with other projects, notably the Digital Codex Pelavicino (see Codice Pelavicino Digitale), and is now developed by a mix of DH graduate students and professional software developers.

EVT 1.0, released in 2013, supports diplomatic-interpretative editions and "digital documentary editions", using XSLT transformations to generate HTML from XML/TEI. EVT 2.0, launched in 2016, was built with the JavaScript framework AngularJS[106] to support critical editions. It later incorporated all functionalities from the first version. A notable example of EVT 2.0 usage is the digital edition of Benedetto Varchi's *Storia fiorentina* (see VaSto). In October 2024, EVT 3.0 beta was released, essentially rebooting EVT 2.0 with Angular,[107] a more modern JavaScript framework. All EVT versions are free and open-source. This pilot analysis focuses on version 2.0, given its stability and widespread adoption.

EVT 2.0[108] is available under the AGPL-3.0 licence. It uses standard web technologies (JavaScript, CSS, HTML) and integrates existing tools, particularly OpenSeadragon[109] for image viewing and VisColl[110] for visualising manuscript structure.

To use EVT, users insert their materials (XML/TEI files, images, and VisColl files) into the software folder and specify paths in the configuration file. This file also allows interface customisation, such as selecting edition levels and visualisation modes. The EVT2-Config-Generator[111] tool simplifies this process. Users can customise TEI element

[105] http://vbd.humnet.unipi.it/beta2/.
[106] https://angularjs.org/.
[107] https://angular.dev/.
[108] EVT 2.0's source code GitHub repository: https://github.com/evt-project/evt-viewer.
[109] https://openseadragon.github.io/.
[110] https://viscoll.org/.
[111] http://evt.labcd.unipi.it/evt2-config/.

presentation via a dedicated CSS stylesheet. Comprehensive user information is available in the repository's README file[112] and in the manual provided in each release archive.

EVT 2.0's modular architecture facilitates collaborative development. The repository's README provides developer instructions, and the code is well-commented to encourage third-party contributions. However, EVT 2.0 lacks a detailed, structured presentation of data extraction and modelling from source materials. EVT 3.0 addresses this by modelling data through TypeScript's class system and providing software contribution guidelines in the repository wiki.[113]

A complete history of EVT 2.0 releases is not available on GitHub, which only shows recent releases. Previously, the EVT 2.0 GitHub repository was for private development use, with releases published on SourceForge.

FAIRness evaluation summary

Strengths

- The software is freely available and open-source, promoting accessibility and collaboration.
- Its modular and well-documented source code encourages reuse and customisation.
- It leverages standard web technologies and seamlessly integrates existing tools, enhancing its versatility and compatibility.
- The client-only model, combined with the use of standard web technologies, makes EVT-based editions accessible in the long term since they require little or no maintenance.
- It is registered in an interoperable Open Access institutional repository (Archivio Istituzionale AperTO - University of Torino) and therefore included in scientific knowledge graphs such as OpenAIRE.[114]

Improvements

- Include a detailed description of data extraction and modelling in the documentation.
- Provide external developers with a complete release and change history.
- Provide software contribution guidelines (addressed in EVT 3.0).

References

Rosselli Del Turco, Roberto. 2019. 'Designing an Advanced Software Tool for Digital Scholarly Editions: The Inception and Development of EVT (Edition Visualisation Technology).' *Textual Cultures* 12 (2): 91–111.

Rosselli Del Turco, Roberto, Chiara Di Pietro, and Chiara Martignano. 2019. 'Progettazione e implementazione di nuove funzionalità per EVT 2: lo stato attuale dello sviluppo.' *Umanistica Digitale* 3 (7). https://doi.org/10.6092/issn.2532-8816/9322.

[112] https://github.com/evt-project/evt-viewer/blob/master/USER_README_EN.md).
[113] https://github.com/evt-project/evt-viewer-angular/wiki.
[114] EVT 2.0 entry in OpenAIRE. https://explore.openaire.eu/search/software?pid=2318%2F1759002. Last accessed 20 November 2024.

Voyant Tools

Voyant Tools (https://voyant-tools.org/) is a web-based environment for text reading and analysis. First released in 2016, the software was designed and developed by Stéfan Sinclair, building on existing text analysis tools (HyperPo[115] and Taporware).[116] Following Sinclair's passing in 2020, Geoffrey Rockwell of the University of Alberta now leads the project, supported by Andrew MacDonald as principal programmer and Cecily Raynor at McGill University.

Voyant Tools aims to facilitate reading and interpretive practices for both the general public and digital humanities scholars. The software's interface is available in English and thirteen other languages. Its open-source code is available on GitHub[117] under a GPL3 licence, with the latest version (2.6.17) released in September 2024.

To use Voyant Tools, users can paste or upload texts or collections in various formats (plain text, HTML, XML, PDF, RTF, MS Word, and Pages) on the software's landing page. Multiple URLs to textual resources can also be inserted.

After tokenising and analysing the textual materials, the application displays results in a graphical interface with a "default skin" — a set of default tools including word distribution graphs, a word cloud visualising the most frequent words, and a summary providing information such as word count and distinctive words. Users can choose from 29 different tools to display in the GUI. These tools interact: clicking an element in one tool updates the information in others. Users can search the corpus using various tools, and export the tools and displayed data as HTML snippets for embedding in external web pages.

Uploaded texts and collections are cached on the software's servers for about a month. This allows users to bookmark and share URLs referring to a text collection, enabling multiple users to work on the same texts across different sessions. Voyant provides basic access management features to control who can access a given text collection. For those preferring local use, VoyantServer — a standalone version — allows running Voyant Tools without storing documents on the software's server.

Voyant Tools provides demonstrative corpora for users to explore its features. The software's comprehensive documentation details each tool individually and offers tutorials. It also includes citation guidelines. While the documentation doesn't specify the programming languages used (Java and XSLT), this information can be inferred from the GitHub repository. A notable section of the documentation outlines the design principles guiding the software's development, offering valuable insights into its conceptual framework.

- **"modularity**: tools should be able to fit together in various configurations
- **generalization**: tools should be designed to address a variety of types of text and uses
- **domain sensitivity**: tools need to be sensitive to the ways in which textual scholars think of and interact with digital texts

[115] https://web.archive.org/web/20121110191405/http://hyperpo.org/.
[116] https://edutechwiki.unige.ch/en/Taporware.
[117] Voyant Tools' source code GitHub repository: https://github.com/voyanttools.

- **flexibility**: tools should be able to work with local or network sources in different formats
- **internationalization**: tools should allow users to work in different languages
- **performance**: tools should be reasonably responsive in order to function in a web-based context
- **separation of concerns**: it may be best to separate back-end analytic procedures from front-end interface concerns
- **extensibility**: it should be easy to create new tools and adapt existing ones, especially for the purposes of experimentation
- **interoperability**: tools should provide public APIs so that they can interact with other tools on the web
- **skinnability**: tools should be able to present themselves differently for different user needs and preferences
- **scalability**: tools should provide functionality both for a small corpus (like a book) or a large corpus (like many books)
- **simplicity**: at least one view of the tools should be maximally simple in its interface
- **ubiquity**: tools should lend themselves to being embedded in content elsewhere on the web
- **referenceability**: tools and their results should lend themselves to being referenced and cited as academic resources"

FAIRness evaluation summary

Strengths

- The software is free and open-source, promoting accessibility and collaboration.
- The tools are modular and well-documented, facilitating ease of use and customisation.
- The interface is highly customisable, allowing users to tailor the environment to their specific needs.

Improvements

- Deposit the software source code in a certified repository, to guarantee long-term accessibility and preservation.
- Voyant Tools appears to be registered in OpenAIRE, however with a limited set of metadata that need to be revised.

References

Alhudithi, Ella. 2021. 'Review of Voyant Tools: See through Your Text.' *Language Learning & Technology* 25 (3): 43–50.

Welsh, Megan E. 2014. 'Review of Voyant Tools.' *Collaborative Librarianship* 6 (2): 96–98.

Standards and guidelines

- W3C standards for web development.
- FAIR Principles for Research Software (FAIR4RS Principles):
    - "F: Software, and its associated metadata, is easy for both humans and machines to find.
    - A: Software, and its metadata, is retrievable via standardised protocols.
    - I: Software interoperates with other software by exchanging data and/or metadata, and/or through interaction via application programming interfaces (APIs), described through standards.
    - R: Software is both usable (can be executed) and reusable (can be understood, modified, built upon, or incorporated into other software)."
- The Research Software MetaData Guidelines.
- Object-oriented design and SOLID principles.
- Semantic Versioning Specification (SemVer).
- Standard digital humanities encoding formats for data input and output, such as XML/TEI and IIIF (see more in the standards presented in the other sections).
- Community Development of Java Technology Specifications.

Recommendations

- Before creating new software from scratch, investigate existing similar solutions and explore opportunities to further develop or adapt them, promoting the reuse and enhancement of existing resources (e.g., EVT forks).
- Involve domain experts in software design to apply software engineering methodologies and best practices, including:
    - Utilise documented and shared design patterns (Gamma et al. 1994).
    - When applying the object-oriented programming paradigm, follow the SOLID principles (Silén 2024).
    - For complex software, implement the "domain-driven design" approach (Evans 2004).
    - Organise code into modules to facilitate the reuse of individual components.
    - Adopt DevOps practices to streamline development and deployment processes (plan, code, build and test, release, deploy, operate, and monitor) (Silén 2024).
    - Ensure that all software dependencies, whether libraries, frameworks, or operating system components, are clearly documented and managed. This also includes defining the operational requirements, such as minimum and optimal hardware resources (e.g., CPU, RAM, disk space) needed to ensure that the software works properly.
    - Define and implement software integration strategies with the goal of achieving a cohesive, scalable and maintainable software ecosystem, minimising the risks of incompatibility and the efforts required for adaptation:
        - Define integration approaches: whether these will be API-based or exchange files, for example, and prepare standard protocols to facilitate communication.
        - Ensure interoperability and compatibility between different systems by considering standard data formats and structured schemas.

        - Plan strategies for handling errors and malfunctions.
        - Ensure scalability and the ability to handle increased load without compromising overall performance.
    - Integrate a structured testing phase as part of the software development process, establishing clear metrics and goals to determine testing success.
    - In systems with numerous dependencies, adopt standard versioning strategies such as "Semantic Versioning". This approach uses version numbers to convey meaningful information about the code and the changes made in different versions.
    - With each released version, always attach a changelog document that provides a clear and organised chronology of updates, improvements, bug fixes, and other changes.
- Employ standard and non-proprietary programming languages and technologies to develop tools, ensuring greater longevity and easier maintainability.
- Choose a programming language with mature libraries that can ease the development and maintenance of your software (e.g., use Python for NLP software development to easily integrate available tools).
- Develop in *open source* and foster collaborative development by:
    - Writing clear, comprehensive code comments.
    - Providing guidelines for contributing to software development.
    - Utilising repositories such as GitHub that foster collaboration among developers.
    - Following shared methodologies and strategies for versioning and branching (e.g., GitFlow workflow).
- Release software officially through freely accessible channels (e.g., GitHub), providing detailed and user-friendly documentation.
- Publish your released research software in a trusted scholarly repository (e.g. Zenodo, CLARIN, institutional or thematic repository) with rich metadata to ensure citability and credit to the development team.

References

Barker, Michelle, Neil P. Chue Hong, Daniel S. Katz, Anna-Lena Lamprecht, Carlos Martinez-Ortiz, Fotis Psomopoulos, Jennifer Harrow, et al. 2022. 'Introducing the FAIR Principles for Research Software'. *Scientific Data* 9 (1): 622. https://doi.org/10.1038/s41597-022-01710-x.

Cohen, Jeremy, Daniel S. Katz, Michelle Barker, Neil P. Chue Hong, Robert Haines, and Caroline Jay. 2021. 'The Four Pillars of Research Software Engineering'. *IEEE Software* 38 (1): 97–105. https://doi.org/10.1109/MS.2020.2973362.

Del Grosso, Angelo Mario, Emiliano Giovannetti, and Simone Marchi. 2017. 'The Importance of Being… Object-Oriented: Old Means for New Perspectives in Digital Textual Scholarship.' In *Advances in Digital Scholarly Editing*, 269–74. Leiden: Sidestone Press.

Del Grosso, Angelo, D. Albanesi, Emiliano Giovannetti, and Simone Marchi. 2016. 'Defining the Core Entities of an Environment for Textual Processing in Literary Computing'. In

Digital Humanities 2016: Conference Abstracts, 771–75. Jagiellonian University & Pedagogical University, Kraków. https://dh2016.adho.org/abstracts/425.

Evans, Eric. 2004. *Domain-Driven Design : Tackling Complexity in the Heart of Software*. Addison-Wesley.

Gamma, Erich, Richard Helm, Ralph Johnson, and John Vlissides. 1994. *Design Patterns: Elements of Reusable Object-Oriented Software*. Pearson Education.

Pierazzo, Elena. 2019. 'What Future for Digital Scholarly Editions? From Haute Couture to Prêt-à-Porter.' *International Journal of Digital Humanities*, May. https://doi.org/10.1007/s42803-019-00019-3.

Zenzaro, Simone. 2024. 'Models for Digital Humanities Tools: Coping with Technological Changes and Obsolescence.' *International Journal of Information Science and Technology* 8 (2): 1–10. https://doi.org/10.57675/IMIST.PRSM/ijist-v8i2.283.

Silén, Petri. 2024. *Clean Code Principles And Patterns, 2nd Edition. A Software Practitioner's Handbook*. Leanpub.

Linked open data

According to (Bizer, Heath, Berners Lee 2009), the expression Linked Data can be used to define "a set of best practices for publishing and connecting structured data on the Web". Tim Berners-Lee's guidelines for five-star linked open data have become widely recognised. These guidelines outline a progressive approach to creating linked data:

1. Available on the web (whatever format) but with an open licence, to be Open Data.
2. Available as machine-readable structured data (e.g. Excel instead of an image scan of a table).
3. As (2) plus non-proprietary format (e.g. CSV instead of Excel).
4. All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff.
5. All the above, plus: Link your data to other people's data to provide context." (Berners Lee 2009)

Over the last years, many GLAM institutions have been leveraging the potential of Linked Data to open previously siloed collections and regain their centuries-old recognition as intermediaries between users and Cultural Heritage through high-quality data publication (Marden et al., 2013; Daquino 2021). Notable examples include the Library of Congress Linked Data Service, whose first dataset dates back to 2009. Since then, many cultural institutions have embraced the same challenge, converting their holdings into collections of RDF triples and making them available through dedicated query services and APIs.

The construction of linked datasets is made possible by some fundamental Web technologies, namely URIs (Uniform Resource Identifiers) and HTTP (HyperText Transfer Protocol), while on a conceptual level, their development is ensured by ontologies, which provide domain-specific vocabularies and a define the meaning of shared terms. However, the

production of LOD collections in a native way has received little intention so far and many Cultural Heritage datasets are still dependent on some intermediate technologies which store their content in a traditional format. Overcoming such a limitation would make data consistency easier to maintain while maximising knowledge reusability and entity reconciliation (Garcia, Kernerman, Bosque-Gil 2017).

To make linked open datasets truly FAIR, additional measures are necessary. Essential information such as licence, version, and version history must be included. Data provenance management requires careful attention, as it is paramount also for project management purposes, e.g., to monitor the editorial process and to keep track of data versions. Models and technologies used should be thoroughly documented. Adopting shared models, vocabularies, standards, and protocols at national and international levels is vital, as exemplified by Zeri&LODE's use of Italian government standards for photos and artworks and LiLa's integration of Latin into LLOD.

The FAIRness of data also hinges on usability. While open and free access is fundamental, accessibility depends on the ease of use of services like LodView and the quality of their documentation. Long-term access must be ensured, preferably through external open access repositories like Zenodo, CLARIN repository or other robust, actively maintained tools. URIs require careful consideration—they should be uniform across different data access modes and point to web resources. Lastly, for a dataset to be truly FAIR, it must be comprehensively documented, especially regarding its data model, used resources (standards, vocabularies, ontologies), workflow, and data preparation tools.

Zeri & LODE

Zeri&LODE (website: https://data.fondazionezeri.unibo.it/) is a University of Bologna project aimed at enhancing Federico Zeri's repository catalogue. This repository includes an art library (46,000 volumes, 37,000 auction catalogues, 60 periodicals) and a photo archive (290,000 photographs of monuments and artworks). The project's goal is to create an RDF dataset of Zeri's photo archive, building on the cataloguing work conducted by the Federico Zeri Foundation and the University of Bologna. This initiative is part of a larger endeavour led by PHAROS, an International Consortium of Photo Archives that aims to create an open and freely accessible digital research platform allowing for comprehensive consolidated access to photo archive images and their associated scholarly documentation.
The catalogue implementation employed two Italian metadata content standards issued by the ICCD (Istituto Centrale per il Catalogo e la Documentazione) of the Italian Ministry of Cultural Heritage: Scheda F[118] for photographs (Scheda di fotografia) and Scheda OA[119] for artworks (Scheda Opera d'Arte). These standards served as the foundation for the dataset modelling.
The initial release of the Zeri Photo Archive RDF dataset (April 2016) represents a significant subset of data already available on the Zeri Catalog website[120] and discoverable through the Europeana Portal.[121] The dataset primarily covers Modern Art (15th-16th centuries),

[118] http://www.iccd.beniculturali.it/index.php?it/473/standard-catalografici/Standard/10.
[119] http://www.iccd.beniculturali.it/index.php?it/473/standard-catalografici/Standard/29.
[120] http://catalogo.fondazionezeri.unibo.it/.
[121] http://www.europeana.eu/portal/it/search?q=PROVIDER%3A%22Federico+Zeri+Foundation%22.

describing about 19,000 artworks and over 30,000 photographs through approximately 11 million RDF statements.

Developed using W3C standard technologies—RDF and SPARQL—the linked open dataset is accessible via a dedicated SPARQL endpoint, a web user interface, and LodView, a linked open data browser. The documentation provides quick-access links to facilitate navigation in this browser. A downloadable version of the dataset is available in an open access repository at the University of Bologna.

The concise documentation includes licence information, attribution guidelines, references to models used in building the data model, and the URI creation schema to facilitate data reuse. However, some aspects of the dataset creation and workflow remain unclear, such as the use of external tools and development duration. While seemingly secondary, this information could benefit researchers undertaking similar projects. Additionally, the latest version date appears only in the attribution, and the dataset's status is ambiguous, as it represents only a subset of the materials in the Zeri Archive.

The data model incorporates existing ontologies: the CIDOC Conceptual Reference Model (CIDOC-CRM), the SPAR Ontologies, and the HiCO Ontology. To enhance interoperability, the Zeri&LODE project developed two custom ontologies: FEntry Ontology[122] and OAEntry Ontology,[123] which represent the aforementioned Scheda OA and Scheda F standards. The documentation also provides links to the RDF mappings of these standards. Technical terms for artwork and photograph descriptions were sourced from the AAT Getty Thesaurus. GeoNames[124] was used for place names, while VIAF, Getty ULAN,[125] DBpedia,[126] and Wikidata were used for people's names.

FAIRness evaluation summary

Strengths

- Uses standard W3C technologies.
- Provides multiple access points to the dataset.
- Allows users to download the dataset.
- Offers comprehensive guidelines for data reuse.
- Builds the data model by leveraging existing models and standards.
- Publishes native ontologies in open access, making them freely available.
- The complete dataset is available as a data dump on an external open access repository (AMSActa) that ensures its long-term preservation.
- Has persistent identifiers (DOIs, handle) assigned by the institutional repositories AMSActa and IRIS - University of Bologna

Improvements

[122] http://www.essepuntato.it/2014/03/fentry.
[123] http://purl.org/emmedi/oaentry.
[124] https://www.geonames.org/.
[125] http://www.getty.edu/research/tools/vocabularies/ulan/.
[126] http://wiki.dbpedia.org/services-resources/datasets/dbpedia-datasets.

- Clearly display the dataset's version number, as it is currently only mentioned in the attribution.
- Include a brief description of the dataset's creation process and any tools used in the documentation.
- Register the dataset in a searchable resource such as OpenAIRE.

References

Gonano, Ciro Mattia, Francesca Tomasi, Francesca Mambelli, Fabio Vitali, and Silvio Peroni. 2014. 'Zeri e LODE. Extracting the Zeri Photo Archive to Linked Open Data: Formalizing the Conceptual Model'. In *IEEE/ACM Joint Conference on Digital Libraries*, 289–98. https://doi.org/10.1109/JCDL.2014.6970182.

Daquino, Marilena, Francesca Mambelli, Silvio Peroni, Francesca Tomasi, and Fabio Vitali. 2017. 'Enhancing Semantic Expressivity in the Cultural Heritage Domain: Exposing the Zeri Photo Archive as Linked Open Data'. *Journal on Computing and Cultural Heritage (JOCCH)* 10 (4): 21:1-21:21. https://doi.org/10.1145/3051487.

DanteSources

DanteSources (https://dantesources.dantenetwork.it/) is a web tool for retrieving and visualising data about the sources of Dante Alighieri's works. Developed between 2013 and 2016, it is a joint effort of the Institute of Information Science and Technologies "Alessandro Faedo" (ISTI) of the Italian National Research Council and the Department of Philology, Literature, and Linguistics at the University of Pisa.

Built on semantic web technologies, DanteSources' RDF dataset encompasses 714 cited works, 273 cited authors, and 45 distinct thematic areas. Users can search the data by Dante's work, primary source, cited author, thematic area, and type of reference in both Dante's works and primary sources. The tool presents the distribution of primary sources, authors, and thematic areas cited by Dante through tables and charts, with the underlying data available for download as CSV files. The referenced publications are often not openly accessible or readily available. The dataset is also accessible via a SPARQL endpoint.[127] The URIs extracted from the SPARQL endpoint do not identify an existing resource.

While the documentation effectively guides data exploration, it falls short in detailing the data model and dataset. The website provides the project's RDF schema (which references the CIDOC-CRM and FOAF ontologies) and full bibliographic references for the analysed editions of Dante's works. However, it omits crucial information such as licensing and citation guidelines. The dataset's status remains ambiguous, though it appears comprehensive, covering all of Dante Alighieri's works except the Comedy.

FAIRness evaluation summary

Strengths

[127] https://dantesources.dantenetwork.it/sparql.

- The dataset is open access and utilises standard technologies (RDF).
- It offers users an intuitive graphical interface for exploring data.
- The documentation incorporates the RDF schema.
- The dataset is indexed in OpenAIRE.

Improvements

- Include licence and citation guidelines on the home page.
- Provide a detailed description of the data model, including the models and standards used. Additionally, include demo queries.
- URIs extracted from the SPARQL endpoint should identify existing resources.
- Describe in the documentation the website's development process, including tools and technologies employed. Additionally, provide users with a guide.
- Make the complete dataset available as a data dump on an external open access repository.

References

Bartalesi, Valentina, Paola Andriani, Daniele Metilli, Carlo Meghini, and Mirko Tavoni. 2016. 'DanteSources : Una Biblioteca Digitale Delle Fonti Dantesche.' *Forme e La Storia : Rivista Di Filologia Moderna : IX, 1, 2016*, 63–82. https://doi.org/10.1400/256610.

Bartalesi, Valentina, Carlo Meghini, Daniele Metilli, Mirko Tavoni, and Paola Andriani. 2018. 'A Web Application for Exploring Primary Sources: The DanteSources Case Study.' *Digital Scholarship in the Humanities* 33 (4): 705–23. https://doi.org/10.1093/llc/fqy002.

LiLa

The LiLa (Linking Latin, website: https://lila-erc.eu/) project, led by Professor Marco Passarotti, is funded by the European Research Council and based at the Cattolica University's CIRCSE (Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione) research centre. Running from 2018 to 2023, LiLa aimed to build a Linked Data-based Knowledge Base of Linguistic Resources and Natural Language Processing (NLP) tools for Latin. Its primary goal was to connect and leverage existing resources, enhancing the study and analysis of Latin texts. More specifically:

"LiLa integrates existing and new linguistic data for Latin. It combines resources like corpora, lexica, and NLP tools from various providers. LiLa also generates new data by enhancing existing resources. This includes:

- Adding PoS-tagging and lemmatisation to Latin texts
- Standardising annotations across Latin treebanks
- Expanding *Latin WordNet* and *Latin-Vallex*
- Increasing coverage of the *Index Thomisticus Treebank*

Additionally, LiLa develops new models for PoS-tagging and lemmatisation, aiming to create an optimal NLP pipeline for Latin."

The LiLa Knowledge Base comprises a lemma bank, lexical resources (e.g., Word Formation Latin, Latin Vallex 2.0, Latin WordNet), and textual resources (e.g., Index Thomisticus Treebank (ITTB), UDante, Lucani Pharsalia). LiLa's webpage presents each data component with three links: a Zenodo publication for attribution, a GitHub repository for download, and a link to the LodView browser for data access. LiLa offers three exploration methods for its knowledge base: a triplestore, a user-friendly query interface, and an interactive search platform. The SPARQL query interface includes useful sample queries to familiarise users with the Knowledge Base's content and data model. Additionally, LiLa developed the "TextLinker" web service to populate its knowledge base by lemmatising and PoS-tagging raw Latin text and linking tokens to the LiLa Lemma Bank.

While the TextLinker and interactive search platform have intuitive interfaces, they lack user examples and instructions. The documentation also omits information on tool development and open-source status.

The project aimed to create a knowledge base adhering to FAIR principles through the following procedures:

- Assigning unique URIs to all (meta)data, lemmas, word types, and metadata tags, enabling precise identification and linking of linguistic elements across resources.
- Using HTTP for data retrieval and SPARQL for querying, facilitating data reuse and citation tracking.
- Archiving all resources on Zenodo for long-term preservation.
- Employing standard semantic web vocabularies and ontologies to describe object relations, though specific ontologies and vocabularies are not mentioned in the documentation.
- Providing open access to all resources.
- Using standardised URIs and offering detailed provenance information.
- Releasing new data under the CC BY-SA licence, software under GNU LGPL3, and Zenodo metadata under CC0.
- Integrating Latin into the multilingual Linguistic Linked Open Data (LLOD) cloud.[128]

The documentation lacks crucial information about the LiLa Knowledge Base's underlying data model, newly created ontologies, and the application of existing RDF vocabularies.

Having seemingly achieved all its goals, the project can be considered complete.

FAIRness evaluation summary

Strengths

- The dataset was designed and implemented with FAIR principles in mind from the outset.
- Data are published on external repositories, ensuring long-term access and preservation.
- The knowledge graph was built by reusing and leveraging existing resources.

[128] http://linguistic-lod.org/llod-cloud.

- The data can be accessed via user-friendly tools with intuitive interfaces that facilitate exploration.
- The datasets are indexed in a searchable resource (OpenAIRE).

Improvements

- Include examples and user guides for all tools.
- Describe the implementation process of the tools, considering publishing them as open-source to facilitate reuse.
- Provide a detailed description of the knowledge graph's data model and newly created ontologies, specifying how existing RDF vocabularies were utilised.
- In the documentation, specify the tools used for creating the knowledge base.

References

Mambrini, Francesco, and Marco Carlo Passarotti. 2023. 'The LiLa Lemma Bank: A Knowledge Base of Latin Canonical Forms'. *Journal of Open Humanities Data* 9 (1). https://doi.org/10.5334/johd.145.

Passarotti, Marco. 2023. 'La Knowledge Base LiLa. Interoperabilità tra risorse testuali e lessicali per il latino'. *Chimera* 10 (June):45–72. https://doi.org/10.5281/zenodo.8076839.

Biflow

BIFLOW (Bilingualism in Florentine and Tuscan Works, website: https://catalogobiflow.vedph.it/) is a research project funded by the European Research Council and hosted by the Ca' Foscari University of Venice and EHESS (École des hautes études en sciences sociales), Paris. Directed by Prof. Antonio Montefusco, the project investigates literary documents that circulated simultaneously in multiple languages in medieval Tuscany, particularly in Florence, from the late 13th to the early 15th century.

The project's primary outcome is an RDF-based catalogue, accessible through a SPARQL endpoint[129] implemented with the tool RDF store-js[130] (which was last updated in 2016, raising concerns about its long-term viability), or via a more user-friendly website interface. However, it is worth noting that the data provided through this service doesn't fully align with the catalogue's RDF serialisations, particularly regarding the base URI of entities and properties.

The catalogue is organised into entries collecting bilingual textual dossiers, which include the source text and its various translations. Each dossier is identified by a code derived from the author's name and the work's title. Every dossier provides a content summary, an essential bibliography, and an interactive graph showing relationships between versions—though currently, the graph functionality appears to be impaired.

[129] https://catalogobiflow.vedph.it/sparql/.
[130] https://github.com/antoniogarrote/rdfstore-js.

Within each dossier, descriptive sub-entries for each text version offer varying degrees of detail on textual history, manuscript tradition, and editorial history. Sub-entries are denoted by a letter added to the dossier code: uppercase for source texts, and lowercase for translations. All texts include a census and description of the codices comprising their textual traditions. The manuscripts are described with essential information and links to other cataloguing sites. However, links to authority records are missing for works and authors. Users can download the entire dossier—including entries for each version and the manuscript census—in PDF and RDF formats. The RDF documents lack information about data provenance.

The catalogue supports querying via full-text search or by author/translator, manuscripts, title, incipit/explicit, language, genre, and textual typology.

While the catalogue has an associated ISSN code and each dossier includes the editor's name and attribution, the website lacks permalinks to dossiers and other catalogue objects.

The catalogue's data model is formalised as an OWL ontology—the Biflow-Toscana Bilingue ontology—visualisable with the WebVOWL[131] tool and documented with Ontospy.[132] This ontology expands on existing models, particularly eFRBRoo, CIDOC-CRM, and the Biblissima ontology.[133]

The catalogue's completion status remains unclear. The dataset isn't available for wholesale download and appears accessible only through the Biflow website, potentially jeopardising its long-term preservation. The website requires refinement, notably in its publications section. Moreover, the English version is incomplete and only partially translated.

FAIRness evaluation summary

Strengths

- The dataset is built on standard technologies (RDF and OWL).
- The documentation includes the dataset's ontology.

Improvements

- Homogenise the base URIs.
- Link data, such as work titles, author names, and manuscripts, to external authority systems (e.g., GeoNames, VIAF).
- Add data provenance in the RDF documents.
- Use a more up-to-date and stable tool for the implementation of the SPARQL endpoint.
- Publish the dataset as a data dump on an external repository.
- Refine the website by adding missing content and completing translations.
- Clearly indicate the project's current status.
- Register the dataset in a searchable resource such as OpenAIRE.

[131] https://service.tib.eu/webvowl/.
[132] https://lambdamusic.github.io/Ontospy/.
[133] https://doc.biblissima.fr/ontologie/bibma/.

References

Mancinelli, Tiziana, and Antonio Montefusco. 2020. 'The BIFLOW-Toscana Bilingue Catalogue: A Digital Representation of the Socio-Cultural History of Translation in the Tuscan Middle Ages (1260–1430)'. *Textual Cultures* 13 (2): 82–110. https://doi.org/10.14434/textual.v13i2.31597.

Standards and guidelines

- Standard languages: RDF, RDF-a, SPARQL, SKOS, OWL.
- Domain-relevant models, ontologies and vocabularies, used on a national and international level: CIDOC-CRM, FOAF, Dublin Core, LRMOO (ex FRBR), DataCite, DCAT, schema.org.
    - Use LOV to explore existing models.
- Data reuse and entity reconciliation based on authority records and controlled vocabularies: Wikidata, VIAF, Open Library, EU vocabularies, WorldCat.
- 5-star open data.

Recommendations

- Reuse existing resources, clearly specifying which ontologies, vocabularies, tools, and models are used.
- Entity reconciliation: link your data extensively to external resources (e.g., authority records like VIAF, Wikidata, SKOS vocabulary like TaDiRAH, etc.) to harness the semantic web's potential.
- Offer a SPARQL endpoint with accompanying search examples.
- Implement a user-friendly GUI for data visualisation and navigation, catering to users unfamiliar with SPARQL.
- Facilitate data reuse by providing the data model and URI patterns. In case the data model comprises new classes and properties, that could be useful for other scholars, formalise and release the data model as an ontology.
- Ensure data provenance, storing provenance information along with content data in order to prevent inconsistencies when integrating sources, emphasise content responsibility, and foster data credibility.

References

Berners Lee, Tim. 2009 'Linked Data - Design Issues.' https://www.w3.org/DesignIssues/LinkedData.html.

Bizer, Christian, Tom Heath, and Tim Berners-Lee. 2009. 'Linked Data - The Story So Far'. *International Journal on Semantic Web and Information Systems (IJSWIS)* 5 (3): 1–22. https://doi.org/10.4018/jswis.2009081901.

Daquino, Marilena. 2021. 'Linked Open Data Native Cataloguing and Archival Description'. *JLIS.It* 12 (3): 91–104. https://doi.org/10.4403/jlis.it-12703.+

García, Jordi Gràcia, Ilan Kernerman, and Julia Bosque Gil. 2017. 'Toward Linked Data-Native Dictionaries'. In *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference*, 550–59. Lexical Computing.

Marden, Julia, Carolyn Li-Madeo, Noreen Whysel, and Jeffrey Edelstein. 2013. 'Linked Open Data for Cultural Heritage: Evolution of an Information Technology'. In *Proceedings of the 31st ACM International Conference on Design of Communication*, 107–12. SIGDOC '13. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2507065.2507103.

Ontologies

Ontologies are typically designed to provide "a vocabulary describing a domain of interest and a specification of the meaning of terms in that vocabulary" (Euzenat and Shvaiko, 2007). While ontologies often target specialists in specific fields, they also serve as valuable models for broader disciplinary areas, and practical applications. Through their properties and structure, they aim to provide a detailed vocabulary, explicitly defining some terms and the relationships existing between them (Guarino et al., 2009). CIDOC exemplifies this approach, having emerged from the need to standardise diverse models and practices in the field of Cultural Heritage. Its comprehensive scope enables various applications, encourages reuse, and allows for extensions into more specific models. To effectively reach a general audience, providing detailed documentation with numerous concrete examples is crucial.

Formalising an ontology requires specific markup languages. Among notable examples –such as RDF and RDF Schema–, the Web Ontology Language OWL has been standing as the most popular semantic technology for the definition of vocabularies since its release in 2004 (Matentzoglu et al., 2013). Its current version (OWL 2) has offered a Description Logic (DL) for various well-known vocabularies in the DH domain, including HiCO and the set of SPAR ontologies (Peroni and Shotton, 2018).

The effectiveness of a semantic vocabulary hinges on its modelling quality. Therefore, following shared guidelines and standardised best-practices is essential. For instance, the Simplified Agile Methodology for Ontology Development (SAMOD)[134] has been designed to guide ontology engineers through an iterative workflow. Additionally, involving the user community is paramount, as demonstrated by the SPAR ontologies. It is essential to test the efficacy of an ontology through real-world applications, as seen with HiCO's use in various research projects. A notable difference in the frequency of updates and maintenance often emerges when an ontology is developed by a large community rather than individual scholars. Finally, maintaining the functionality of IRIs is a critical aspect of ontology upkeep.

A successful ontology must adhere to FAIR principles. This can be supported by several tools, including catalogues (e.g. LOV, ODP),[135] documentation services (e.g. LODE,[136] WIDOCO,[137] WebVOWL),[138] ontology repositories.

[134] https://essepuntato.it/papers/samod-owled2016.html.
[135] ODP (Ontology Design Patterns): http://ontologydesignpatterns.org/wiki/Main_Page.
[136] Live OWL Documentation Environment: https://essepuntato.it/lode/.
[137] WIzard for DOCumenting Ontologies (WIDOCO): https://zenodo.org/badge/latestdoi/11427075.
[138] https://github.com/VisualDataWeb/WebVOWL.

CIDOC-CRM

The CIDOC Conceptual Reference Model (CRM, website: https://www.cidoc-crm.org/) is a tool for integrating Cultural Heritage information across diverse datasets. It provides a formal structure for describing concepts and relationships in Cultural Heritage documentation, enabling data integration from multiple sources.

CIDOC CRM aims to create a shared understanding of Cultural Heritage information through a common semantic framework. It serves as a language for domain experts and implementers to define information system requirements and guide conceptual modelling, acting as a "semantic glue" between different Cultural Heritage information sources.

The CRM consists of a base standard (CRMbase) and modular extensions. These extensions, developed with or by research communities, support specialised research questions while remaining compatible with the base ontology. This approach ensures high information integrity and integration. The extensions are:

- LRMOO — Library Reference Model;
- PRESSOO — Model for publishing periodicals;
- CRMact — Model for activity plan;
- CRMarchaeo — Excavation model;
- CRMba — Model for archaeological buildings;
- CRMdig — Model for provenance metadata;
- CRMgeo — Spatiotemporal model;
- CRMinf — Argumentation model;
- CRMsci — Scientific observation model;
- CRMsoc — Model for social phenomena;
- CRMtex — Model for the study of ancient texts.

CIDOC CRM is developed by a volunteer community, the CIDOC CRM Special Interest Group, under ICOM's International Council for Documentation. Members include institutions involved in researching and documenting human history.

The development of the CRM commenced in 1996. In 2006, CRM gained recognition as an official ISO standard.[139] While the latest version of the model is 7.3,[140] the most recent stable and official version—which also serves as the ISO standard—is 7.1.3,[141] released in February 2024. All previous versions are presented on the website in a chronologically organised table. Each stable version can be downloaded in various standard formats (XML, JSON-LD, RDF) and is accompanied by comprehensive documentation in PDF and DOCX formats. The website offers users learning materials, use cases, and best practices, along with mappings and information to facilitate the model's reuse.

CRMbase can be visualised online via an intuitive and friendly interface. Its documentation is available in English, German, Greek, French, Portuguese, Russian and Chinese. All classes and properties are assigned a URI.

[139] https://www.iso.org/standard/85100.html.
[140] https://www.cidoc-crm.org/Version/version-7.3.
[141] https://www.cidoc-crm.org/Version/version-7.1.3.

FAIRness evaluation summary

Strengths

- Comprehensive and extensive documentation, accompanied by learning materials, use cases, and examples to facilitate the model's adoption and reuse.
- The model is an official ISO standard. All stable versions are available for download in various standard formats.
- Detailed presentation of current and previous versions, clearly differentiated by stability status.
- The model is formalised as an ontology using standard and non-proprietary languages.
- The model supports interoperability with other domain-relevant standards.
- Users can report issues, fostering continuous improvement.

Improvements

- Consider publishing the stable versions of the ontology on an external repository in open access.
- Register the model and its compatible models in a searchable resource.

References

Chryssoula Bekiari, George Bruseker, Erin Canning, Martin Doerr, Philippe Michon, Christian-Emil Ore, Stephen Stead, Athanasios Velios. 2024. 'Volume A: Definition of the CIDOC Conceptual Reference Model' version 7.1.3. https://www.cidoc-crm.org/Version/version-7.1.3.

Doerr, Martin. 2005. 'The CIDOC CRM, an Ontological Approach to Schema Heterogeneity'. *Dagstuhl Seminar Proceedings, Volume 4391*, 1–5. https://doi.org/10.4230/DagSemProc.04391.22.

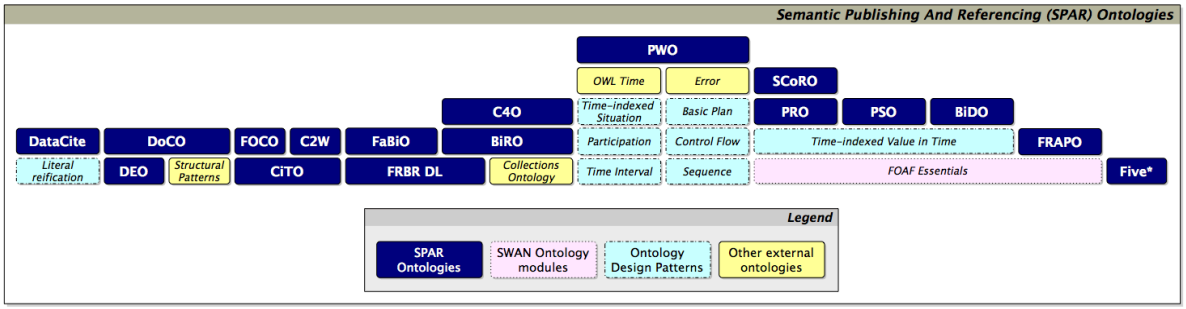
SPAR Ontologies

The Semantic Publishing and Referencing Ontologies (SPAR Ontologies, website: http://www.sparontologies.net/) are a suite of OWL 2 DL[142] ontology modules for creating machine-readable RDF metadata for semantic publishing and referencing. David Shotton created the first ontology, CiTO (Citation Typing Ontology), in 2009. Since 2010, Shotton and Silvio Peroni have led the development of the SPAR suite. With contributions from numerous collaborators, the SPAR ontologies have evolved into interoperable ontological modules that reuse existing vocabularies. The SPAR suite now encompasses sixteen distinct ontologies:

- FRBR-aligned Bibliographic Ontology (FaBiO): Describes publishable entities and bibliographic references.
- Citation Typing Ontology (CiTO): Characterises citation types and rhetoric.

[142] http://www.w3.org/TR/owl2-syntax/.

- Bibliographic Reference Ontology (BiRO): Defines bibliographic records, references, and collections.
- Citation Counting and Context Characterisation Ontology (C4O): Records in-text citations, contexts, and global citation counts.
- Document Components Ontology (DoCO): Provides vocabulary for document components.
- Publishing Status Ontology (PSO): Characterises publication status throughout the publishing process.
- Publishing Roles Ontology (PRO): Describes roles in the publication process.
- Publishing Workflow Ontology (PWO): Describes publication workflow steps.
- Essential FRBR in OWL2 DL Ontology (FRBR): Expresses IFLA's Functional Requirements for Bibliographic Records.
- Discourse Elements Ontology (DEO): Provides vocabulary for rhetorical document elements.
- Scholarly Contributions and Roles Ontology (SCoRO): Describes scholarly contributions and roles.
- Funding, Research Administration and Projects Ontology (FRAPO): Describes research project administrative information.
- DataCite Ontology (DataCite): Enables description of DataCite metadata properties
- Bibliometric Data Ontology (BiDO): Describes bibliometric data.
- Five Stars of Online Research Articles Ontology (FiveStars): Characterises online journal article attributes.
- FAIR* Reviews Ontology (FR): Describes reviews of scholarly resources.



The structure of the SPAR ontologies.

On the suite's website, each ontology has its own page featuring comprehensive documentation, usage examples, graphical representations, references, and a link to the GitHub repository housing the ontology's source code. Users can access and download the ontology document in various standard non-proprietary formats: RDF/XML, Turtle, N-Triples, and JSON-LD. Each ontology can be visualised online using the LODE tool, which provides a user-friendly interface for browsing classes and properties. All ontologies are available under the Creative Commons 4.0 licence and are assigned a DOI and IRIs, created using the PURL and W3id.org[143] systems. However, the PURL-generated links do not work properly.

The SPAR ontologies' website offers extensive and detailed documentation, including examples of ontology usage and references to external projects that have implemented the

[143] http://W3id.org.

suite. Additionally, the documentation provides contribution guidelines for those interested in proposing a new ontology for inclusion in the SPAR suite.

FAIRness evaluation summary

Strengths

- All ontologies in the suite are open source, open access, and available for download in various standard formats.
- The suite is built by reusing existing vocabularies and employs a modular structure, facilitating the reuse of individual or multiple ontologies and the integration of new ones.
- Each ontology is well-documented and assigned a unique DOI.

Improvements

- Fix the PURL IRIs.
- Publish all versions of the ontologies as different official releases on GitHub.
- Register all ontologies in a searchable resource.

References

Euzenat, Jérôme, and Pavel Shvaiko. 2007. *Ontology Matching*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-49612-0.

Matentzoglu, Nicolas, Samantha Bail, and Bijan Parsia. 2013. 'A Snapshot of the OWL Web'. In *The Semantic Web – ISWC 2013*, edited by Harith Alani, Lalana Kagal, Achille Fokoue, Paul Groth, Chris Biemann, Josiane Xavier Parreira, Lora Aroyo, Natasha Noy, Chris Welty, and Krzysztof Janowicz, 331–46. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-41335-3_21.

Peroni, Silvio, and David Shotton. 2018. 'The SPAR Ontologies'. In *The Semantic Web – ISWC 2018*, edited by Denny Vrandečić, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, and Elena Simperl, 11137:119–36. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-00668-6_8.

HiCO

The Historical Context Ontology (HiCO, website: https://marilenadaquino.github.io/hico/) is an OWL 2 DL ontology designed to represent the context of scholarly claims about Cultural Heritage objects. Developed by Marilena Daquino (University of Bologna), with contributions from Silvio Peroni and Francesca Tomasi, HiCO wasn't born from a research project but has found application in various research initiatives. It is used to describe art historians' attributions in the Zeri dataset (see above) and in digital editions of Paolo Bufalini's notebook[144] and Vespasiano da Bisticci's letters.[145]

[144] https://projects.dharc.unibo.it/bufalini-notebook/.
[145] https://vespasianodabisticciletters.unibo.it/.

HiCO extends PROV-O,[146] a W3C-recommended ontology for data provenance description, adding terms to describe aspects of hermeneutical activity. It also incorporates the CiTO Ontology (from the SPAR ontologies, see above) to link attributions to related sources. HiCO has been developed according to the SAMOD methodology. The ontology is accessible online via the LODE tool and downloadable as an OWL document under the Creative Commons 4.0 licence. The current version, 2.0, was released in 2020. HiCO is also hosted in a public GitHub repository,[147] although the version available appears to be older than the one published via LODE. The documentation provides detailed component descriptions, prefaced by the ontology's scope, a visual representation of its components, and a real-world usage example. The IRIs are built using PURL, but they currently do not function correctly. Consequently, properties used in Linked Open Data collections (e.g., Zeri Archive) are identified by URIs that do not lead to existing web pages.

FAIRness evaluation summary

Strengths

- The ontology is built using a standard language (OWL), follows a shared methodology, and reuses existing ontologies.
- The ontology is associated with a permanent URI (https://w3id.org/hico/).
- The documentation is comprehensive.
- HiCO's current version is assigned a specific IRI, while previous versions remain accessible online.

Improvements

- Fix the PURL IRIs to ensure proper resolution of ontology identifiers.
- Publish all versions of the ontology as distinct official releases on GitHub for better version control and accessibility.
- Publish each version of the ontology on a repository such as Zenodo or CLARIN to ensure long-term preservation.
- Register the ontology in a searchable resource.

References

Daquino, Marilena, and Francesca Tomasi. 2015. 'Historical Context Ontology (HiCO): A Conceptual Model for Describing Context Information of Cultural Heritage Objects'. In *Metadata and Semantics Research*, edited by Emmanouel Garoufallou, Richard J. Hartley, and Panorea Gaitanou, 424–36. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-24129-6_37.

Standards and guidelines

- W3C standard languages: SKOS, OWL, RDF Schema.

[146] https://www.w3.org/TR/prov-o/.
[147] https://github.com/marilenadaquino/hico.

- Domain-relevant models, ontologies and vocabularies, used on a national and international level: CIDOC-CRM, FOAF, Dublin Core, LRMOO (ex FRBR), DataCite, DCAT, schema.org.
    - Use LOV to explore existing models.
- Simplified Agile Methodology for Ontology Development (SAMOD).
- Shared tools and technologies for ontology development and publication:
    - Workbench for storing and searching: graphdb, Virtuoso.
    - Editing tool: Protégé.
    - Web environment for visualising ontology documentation: LODE.
    - Tool for ontologies visual representations: Graffoo.

Recommendations

- Encourage the reuse of existing ontologies to build upon established knowledge structures.
- Foster interoperability with existing ontologies and facilitate integration through comprehensive mapping.
- Adhere to shared methodologies for ontology design and implementation to ensure consistency and best practices.
- Engage the user community and domain experts throughout the design phase to create more robust and relevant ontologies.
- Provide users with detailed documentation, complete with practical usage examples and intuitive graphical representations of the ontology.
- Ensure the permanence of URIs, considering the use of W3id and PURL services for long-term stability.
- Conduct regular maintenance checks to ensure the continued functionality of links to your ontology.
- Utilise external services for ontology publication to guarantee long-term accessibility and preservation.

References

Guarino, Nicola, Daniel Oberle, and Steffen Staab. 2009. 'What Is an Ontology?' In *Handbook on Ontologies*, edited by Steffen Staab and Rudi Studer, 1–17. International Handbooks on Information Systems. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-92673-3_0.

Noy, Natalya F, and Deborah L McGuinness. n.d. "Ontology Development 101: A Guide to Creating Your First Ontology."

2. Data Model

Abstract

In this chapter we present the data model of the ATLAS knowledge graph, focusing on the outcomes from the project's first phase, namely the cataloguing metadata for describing various types of research products (identified through the pilots' analysis), and the framework serving as the foundation for the data model. The next version of the whitebook will also detail how we have modelled the contents of research products.

The framework was constructed by comparing and mapping existing models, particularly Schema.org,[148] RO-crate,[149] DCAT,[150] and OpenAIRE.[151] It is based on two main concepts: `Research Project` and `Research Product`. The latter represents all types of digital objects produced by scholarly research projects. The `Research Product` class offers a suite of properties that help identify the research product from both cataloguing (e.g., `title`, `description`, `release date`, `version`, `licence`) and technical (e.g., `format`, `access points`) perspectives.

To distinguish different types of research products, from the `Research Product` have been derived five sub-classes: `Text Collection`, `Digital Scholarly Edition`, `Ontology`, `Linked Open Data`, and `Software`. This modelling approach highlights the unique characteristics of each type of digital object produced in scholarly research. Each subclass expands on the `Research Product` class, allowing for the description of specific types of research products with additional properties. After some methodological notes, the second chapter presents these classes and their related properties, illustrated with examples from the pilots and accompanied by recommendations for data entry.

Methodological notes

Two key research questions guided our data modelling process:

1. How can we represent research products and the different types of research products?
2. How can we represent research projects and their relationship to the products they develop?

[148] http://Schema.org.
[149] Research Object Crate: https://www.researchobject.org/ro-crate/.
[150] Data Catalog Vocabulary: https://www.w3.org/TR/vocab-dcat-2/.
[151] Data model documentation of OpenAIRE's knowledge graph: https://graph.openaire.eu/docs/data-model/.

To address these questions, we began by comparing existing models with similar aims and scope. We focused initially on the Data Catalog Vocabulary (DCAT), Schema.org, DC Terms,[152] and the FRBR-aligned Bibliographic Ontology (FaBiO).[153]

DCAT is a W3C recommended RDF vocabulary designed to represent data catalogues published on the Web. It is based on seven main classes, namely:

- A `dcat:Catalog` is a collection of metadata about resources like datasets or services.
- A `dcat:Resource` can be a dataset, a data service, or any other type of resource described in a catalog. It is not used independently but serves as a base for more specific types like `dcat:Dataset`, `dcat:DataService`, and `dcat:Catalog`.
- A `dcat:Dataset` is a collection of data published or managed by a single person, group, or organisation. It can include various types of information like numbers, text, images, or sounds.
- A `dcat:Distribution` is a specific dataset form that people can access, such as a downloadable file.
- A `dcat:DataService` is a set of operations (like an API) that allows access to one or more datasets.
- A `dcat:DatasetSeries` is a group of related datasets that are published separately but share some common characteristics.
- A `dcat:CatalogRecord` contains information about a catalogue entry, such as who added it and when.

Some of DCAT's classes and properties, particularly "Dataset," were incorporated into Schema.org. Using Schema.org instead of DCAT allows access to a wider pool of classes and properties while still utilising the main classes defined in DCAT. Among these, the class "Dataset" seemed to be a valuable starting point for describing research products in our model.

Dublin Core is a crucial international standard for cataloguing metadata. Initiated in the mid-1990s, the DCMI Metadata Terms ontology now comprises fifteen main elements applicable to the metadata description of all resource types, including digital scholarly objects. These elements are contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title, and type. The ontology also provides a set of classes (the DCMI Type Vocabulary) to categorise the nature or genre of the resource. Some of these classes were viable options to represent different types of research products in our data model, namely: Collection, Dataset, InteractiveResource, Service, Software, and Text.

Lastly, FaBiO, one of the SPAR ontologies (see Chapter 1), describes publications and other publishable entities (e.g., journal articles, conference papers, books) from a bibliographic perspective. While this ontology does not offer classes that perfectly fit the different types of pilots we analysed, it may prove useful in connecting the concept of "research product" to IFLA's "WEMI" (Work, Expression, Manifestation, Item) framework. This connection may

[152] DCMI Metadata Terms: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.
[153] https://sparontologies.github.io/fabio/current/fabio.html.

facilitate the integration of our data model with relevant models in the Cultural Heritage field, such as CIDOC-CRM.

In the subsequent phase of the modelling process, we examined the data models of systems and knowledge graphs that like ATLAS aim at representing and cataloguing scholarly outputs, in particular, the models of: the OpenAIRE scholarly knowledge graph, SKG-IF (Scholarly Knowledge Graphs Interoperability Framework),[154] RO-Crate, IRIS (Institutional Research Information System)[155] and the KNOT catalogue.[156] The comparison between these models enabled us to identify common properties and select those most appropriate for our objectives. We formalised the comparison between these models and our own data model as a comprehensive mapping.[157]

The data model of the OpenAIRE scholarly knowledge graph comprises different classes to represent organisations, projects, communities, individual people, and, more importantly, research products and data sources, i.e., the sources from which the metadata of graph objects are collected. In this data model, research products may be further described as publications, data, software, or other products with additional properties.

The Data source and Research product classes, with the same categorisation for the latter, are also proposed in the SKG-IF, a model developed by an Interest Group of the RDA (Research Data Alliance) on Open Science Graphs for FAIR Data. In this model, the concept of "research product" is categorised as:

- "**Literature:** Intended for reading by humans (article, thesis, peer-review, blog posts, books, reports, patents, etc.)
- **Research data:** Self-contained, persistently identified digital assets intended for processing (e.g. files containing: tables, metadata collections, dumps; persistent dynamic queries to scientific databases)
- **Research software:** (definition from RDA WG) Research Software includes source code files, algorithms, scripts, computational workflows, and executables that were created during the research process or for a research purpose. [...]
- **Other products**: any digital asset, uniquely identified, whose nature does not fall in the first three types" (https://skg-if.readthedocs.io/en/v1.0/products.html#research-product)

RO-Crate is a community-driven initiative aimed at developing a lightweight approach for packaging research data along with their associated metadata. The research object (RO) is fundamentally a collection of data or a "crate" (e.g., papers, data files, software, references to other research). In order to make it easier to track, archive, and attribute, the crate is accompanied by a plain text file named the *RO-Crate Metadata Document,* which includes metadata for each item within the collection. The RO-crate specification is based mainly on

[154] https://skg-if.readthedocs.io/en/v1.0/.
[155] https://wiki.u-gov.it/confluence/display/public/UGOVHELP/IRIS+-+Institutional+Research+Information+System.
[156] https://projects.dharc.unibo.it/knot/.
[157] The mapping is deposited in Zenodo and can be consulted freely: https://doi.org/10.5281/zenodo.13993057.

Schema.org and other existing standards, such as the W3C Web Annotation Data Model,[158] W3C PROV, Dublin Core Terms, and ORCID.

IRIS[159] is a Java-based platform for managing and enhancing research outputs adopted by numerous Italian universities. While originally focussing on publications, IRIS's modular structure was progressively expanded to describe research and public engagement activities and projects as well. Two modules are particularly relevant to ATLAS:

- Activities & Projects (AP): Gathers information on research projects, contracts, and initiatives. Allows data entry to highlight scientific value and collaborations.
- Institutional Repository / Open Archive (IR/OA): Stores and enhances publication outputs. Provides an interoperable system for managing and disseminating publications, compliant with MIUR and OpenAIRE requirements. (see Bollini et al., 2016, p. 739)

IRIS's data model is mainly based on DC Terms.

The data model developed for the KNOT catalogue[160] proved particularly useful as a reference for our model's general framework. The KNOT data model (KNOT-DM)[161] is based on DCAT, CIDOC-CRM, and PROV-O, which are used to describe published data, Cultural Heritage information, and academic provenance, respectively. The KNOT data model (KNOT-DM) distinguishes between research projects and their products (termed "digital scholarly objects" in KNOT), representing products as `dcat:Catalog` and `prov:Entity`, and projects as `prov:Activity`. Following DCAT, it further differentiates between data (`dcat:Dataset`), data access services (`dcat:Dataservice`), and data publications (`dcat:Distribution`).

To differentiate types of research products, the KNOT-DM uses a `type` attribute and a native controlled vocabulary, the KNOT Taxonomy,[162] as values. These include corpus, database, dataset, digital archive, digital catalogue, digital edition, digital library, digital platform, digital repository, knowledge base, knowledge graph, ontology, and software.

Going back to the above mentioned research questions, we decided to maintain KNOT-DM's conceptual distinctions between research products and projects, and between research products and their access services and publications. However, we took a different approach to representing research product types. We modelled various research products, starting with those from our pilots, as subclasses of a generic "research product" concept. These subclasses ("ontology," "software," "linked open data," "digital edition," and "text collection")

---

[158] https://www.w3.org/TR/annotation-model/.

[159] IRIS is developed by Cineca—an Italian consortium comprising universities, research centers, and the Italian Ministry of Education (MIUR).

[160] The University of Bologna and the Central Institute for the Digitisation of Cultural Heritage (ICDP) developed the KNOT catalogue to explore ways of integrating digital Cultural Heritage from Italian universities, with a focus on Digital Humanities, into the ICDP Digital Library. The Digital Humanities Advanced Research Center (/DH.ARC) created the catalogue.

[161] KNOT data model documentation: https://icdp-digital-library.github.io/KNOT/website/ENG/data_model.html.

[162] KNOT Taxonomy: https://github.com/icdp-digital-library/KNOT/blob/main/data_model/controlled_vocabularies/1.2/ktx.ttl.

are more thorough and specific to the domain of the ATLAS catalogue than the ones offered in the models presented above, usually limited to “Publication”, “Software” and “Other”. For more details, please refer to the next section.

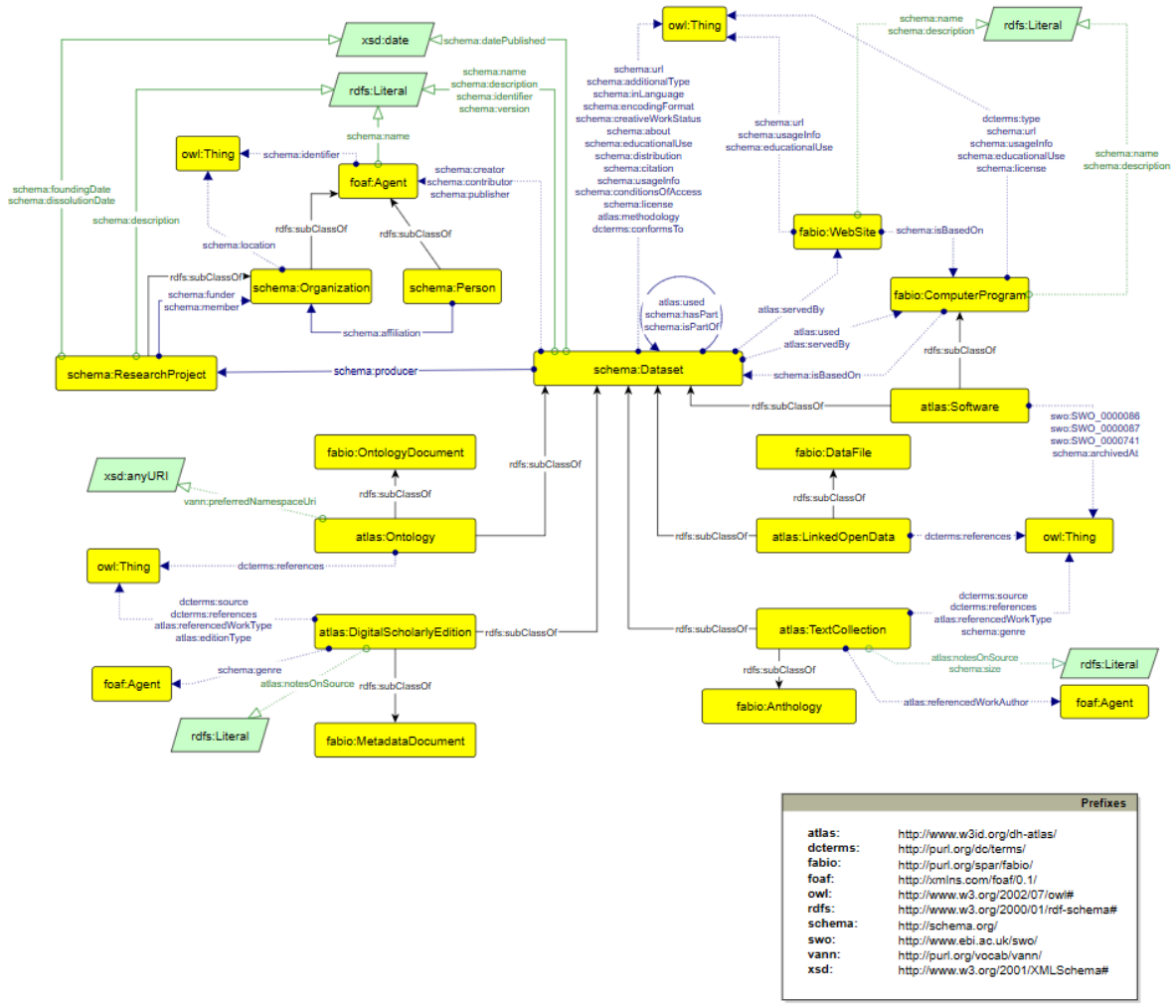
After addressing the research questions and establishing our general approach, we continued data modelling by analysing the pilots. The pilots’ analysis was crucial for identifying common properties across different types of research products and those specific to particular product types. This analysis, extensively described in the previous chapter, yielded valuable insights and led to the selection of domain-specific metadata.[163]

Finally, the data model has been formalised as an OWL 2 DL ontology. The ATLAS Ontology[164] integrates complementary entities from Schema.org, DC Terms, and FaBiO—and introduces new ATLAS Classes and Properties to enhance granularity and specificity, thereby facilitating coherent connections between Classes across different vocabularies.

Description

At the core of the ATLAS model is the `schema:Dataset` class, which defines Research Products as structured sets of information focused on specific topics of interest. This class branches into five specialised subclasses (`atlas:LinkedOpenData` ; `atlas:DigitalScholarlyEdition` ; `atlas:Ontology` ; `atlas:Software` ; `atlas:TextCollection`), each designed to provide a more detailed definition and categorisation of various types of Research Products. Consequently, each subclass is characterised by distinct properties, and for enhanced clarity, each is also aligned as a subclass of a corresponding class from the FaBiO ontology. This alignment also establishes a meaningful connection with the FRBR model, positioning ATLAS Research Products as Expressions within that framework.

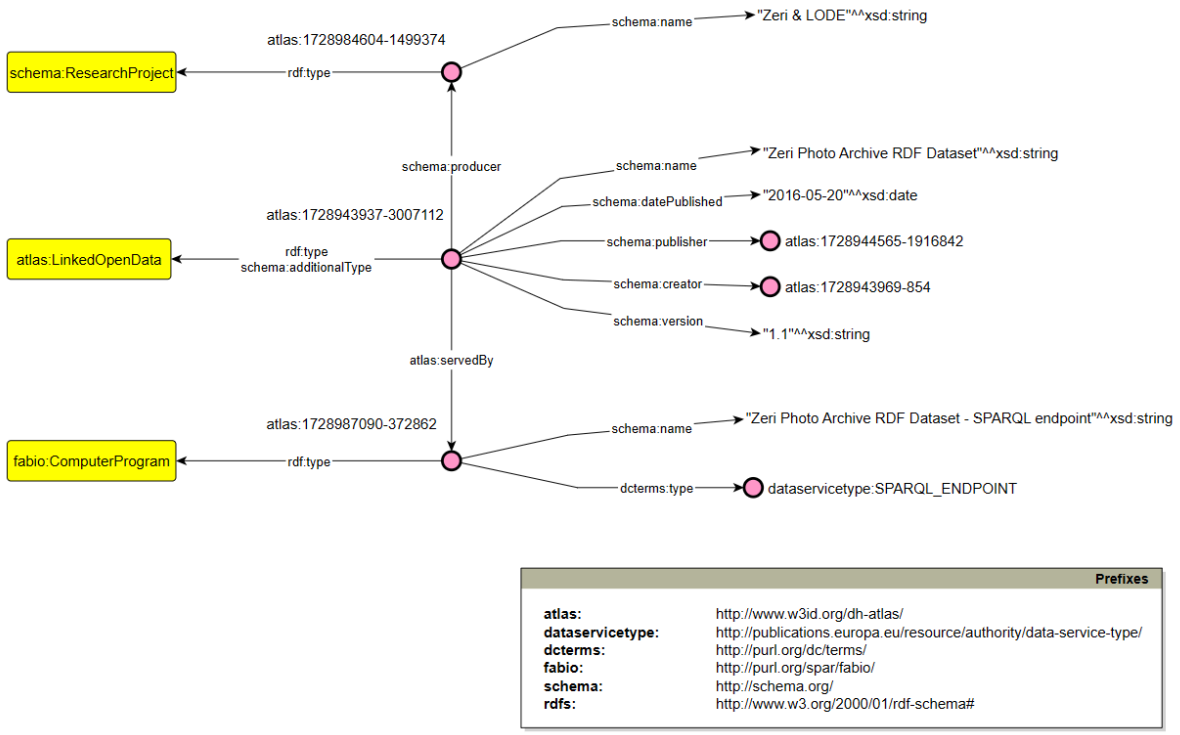
[163] A comprehensive list of metadata and examples illustrating how they were used to describe the pilots is freely available on Zenodo: https://doi.org/10.5281/zenodo.13993057.
[164] The ontology is available on the project’s GitHub repository and can be consulted here.



The relationship between a Research Product and the Research Project responsible for its creation is expressed through the `schema:producer` property, which links a `schema:Dataset` instance to a `schema:Organization`, more specifically to a `schema:ResearchProject`.

The ATLAS ontology also aims to provide detailed information on the access methods for Research Products. To support this, two key classes from the FaBiO ontology—`fabio:ComputerProgram` and `fabio:WebSite`—have been integrated. The former encompasses any type of computer program, including SPARQL endpoints or API services, which facilitates access to the content of a `schema:Dataset` by offering specific functionalities. In other cases, a Research Product may be made accessible through a dedicated website, as seen with digital edition visualisation platforms. These websites typically use advanced software to process the content of the Research Product and generate an interactive browsing interface.

The illustration below presents a simplified use case drawn from the pilots. A Research Product, specifically an `atlas:LinkedOpenData` (Zeri Photo Archive RDF Dataset) is connected to its Research Project (`schema:ResearchProject`), Zeri & LODE, and SPARQL Endpoint (`fabio:ComputerProgram`).



In the following sections, we provide a detailed presentation of the Research Product class and its subclasses. For information on other classes and properties included in the ATLAS ontology, please refer to the ATLAS ontology documentation.[165]

Research Product

In the ATLAS-DM research products are modelled as instances of `schema:Dataset`, which is defined as "a body of structured information describing some topic(s) of interest".

As illustrated in Schema.org's data and datasets overview,[166] `schema:Dataset` describes collections of packaged data, alongside `schema:DataCatalog` for the overall collection and `schema:DataDownload` for specific representations of a dataset. These three classes are designed for applications that publish or integrate different kinds of data, such as the ATLAS catalogue. In the ATLAS-DM, we implement this approach by associating each Research Product with a Research Project (where applicable) and providing a `downloads` property.

The Research Product class includes multiple properties, some mandatory to ensure essential identifying information. The recommended properties are designed to thoroughly describe research products following "FAIR" principles.

Prefixes used in RDF examples

```
@prefix dcterms: <http://purl.org/dc/terms/>
```

[165] https://dh-atlas.github.io/deliverables/ontology/index-en.html.
[166] https://schema.org/docs/data-and-datasets.html.

```
@prefix foaf: <http://xmlns.com/foaf/0.1/>
@prefix ns1: <http://schema.org/>
@prefix ns2: <https://w3id.org/dh-atlas/>
@prefix ns3: <http://dbpedia.org/ontology/>
@prefix ns4: <http://www.ebi.ac.uk/swo/> .
@prefix prov: <http://www.w3.org/ns/prov#>
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
@prefix schema: <https://schema.org/>
@prefix skos: <http://www.w3.org/2004/02/skos/core#>
@prefix vann: <http://purl.org/vocab/vann/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>
```

Title

Mandatory

Value type: ***String***	*Domain:* ***Dataset***, ***Research Project***, ***Web Site***
Cardinality: ***MANY***	
RDF property: ***schema:name***	

The title of the research product and its abbreviations or aliases. If multiple titles in different languages are available, it is possible to indicate them all.

```
<https://w3id.org/dh-atlas/1729030147-4807222> a schema:Dataset ,
        <http://purl.org/spar/fabio/Anthology> ,
        ns2:TextCollection ;
    ...
    ns1:name "Archivio della Latinità Italiana del Medioevo (ALIM)
Digital Library"@en ;
```

Description

Mandatory

Value type: ***String***	*Domain:* ***Dataset***, ***Research Project***, ***Web Site***
Cardinality: ***MANY***	*Same as:* ***schema:description***
RDF property: ***atlas:description***	*Subproperty:* ***atlas:notesOnSource***

Brief text that describes the research product, its content and main features.

The description should highlight the RP's innovative features.

```
<https://w3id.org/dh-atlas/1729156722-0141878> a schema:Dataset ,
        <http://purl.org/spar/fabio/Anthology> ,
        ns2:TextCollection ;
    rdfs:label "Biblioteca Italiana"@en ;
...
    ns1:description "Biblioteca Italiana is a digital library of texts
representative of the Italian cultural and literary tradition from the
Middle Ages to the 20th century, with more than 3,500 titles in its
catalog. It is divided into three main collections: BibIt, Incunaboli,
Scrittori d'Italia."@en ;
```

Creator

Mandatory

Value type: ***URI***	*Domain:* ***Dataset***
Cardinality: ***MANY***	*Range:* ***foaf:Agent***
RDF property: ***schema:creator***	

The research product's creator(s). It may be a single person, a group of people or an organisation.

We recommend using ORCID identifiers for people.

```
<https://w3id.org/dh-atlas/1729160920-6521974> a schema:Dataset ,
        <http://purl.org/spar/fabio/Anthology> ,
        ns2:TextCollection ;
    rdfs:label "Musisque Deoque (MQDQ)"@en ;
    ns1:creator <https://w3id.org/dh-atlas/1729161349-6181638>,
        <https://w3id.org/dh-atlas/1729169549-7938185>,
        <https://w3id.org/dh-atlas/1729169645-4338667>,
        <https://w3id.org/dh-atlas/1729169793-2088108>,
        <https://w3id.org/dh-atlas/1729169989-2397113>,
        <https://w3id.org/dh-atlas/1729170113-255162> ;
  ... > .
<https://w3id.org/dh-atlas/1729161349-6181638> rdfs:label "Luigi
Tessarolo"^^xsd:string .
<https://w3id.org/dh-atlas/1729169549-7938185> rdfs:label "Paolo
Mastandrea"^^xsd:string .
<https://w3id.org/dh-atlas/1729169645-4338667> rdfs:label "Raffaele
Perrelli"^^xsd:string .
<https://w3id.org/dh-atlas/1729169793-2088108> rdfs:label "Gilberto
```

```
Biondi"^^xsd:string .
<https://w3id.org/dh-atlas/1729169989-2397113> rdfs:label "Loriano
Zurli"^^xsd:string .
<https://w3id.org/dh-atlas/1729170113-255162> rdfs:label "Valeria
Viparelli"^^xsd:string .
```

Contributor

Value type: ***URI***	*Domain:* ***Dataset***
Cardinality: ***MANY***	*Range:* ***foaf:Agent***
RDF property: ***schema:contributor***	*Subproperty:* ***atlas:referencedAuthor***

A person, group, or organisation that has contributed to the creation or development of the research product.

We recommend using ORCID identifiers for people.

```
<https://w3id.org/dh-atlas/1729175944-0618887> a schema:Dataset ,
        <http://purl.org/spar/fabio/Anthology> ,
        ns2:TextCollection ;
    ns1:contributor <https://w3id.org/dh-atlas/1729174015-556427>,
        <https://w3id.org/dh-atlas/1729175360-5990732>,
        <https://w3id.org/dh-atlas/1729175473-5923057>,
        <https://w3id.org/dh-atlas/1729175545-28137> ;
... > .
<https://w3id.org/dh-atlas/1729174015-556427> rdfs:label "Alessandro Di
Muro"^^xsd:string .
<https://w3id.org/dh-atlas/1729175360-5990732> rdfs:label "Cristiano
Amendola"^^xsd:string .
<https://w3id.org/dh-atlas/1729175473-5923057> rdfs:label "Teofilo De
Angelis"^^xsd:string .
<https://w3id.org/dh-atlas/1729175545-28137> rdfs:label "Martina
Pavoni"^^xsd:string .
```

Publisher

Mandatory

Value type: ***URI***	*Domain:* ***Dataset***
Cardinality: ***MANY***	*Range:* ***foaf:Agent***
RDF property: ***schema:publisher***	

The publisher of the research product. It may also be a university or other institution hosting the research product.

We recommend using ORCID identifiers for people.

```
<https://w3id.org/dh-atlas/1729185052-855823> a schema:Dataset ,
        <http://purl.org/spar/fabio/MetadataDocument> ,
        ns2:DigitalScholarlyEdition ;
    rdfs:label "VaSto - Varchi, Storia fiorentina Digital Edition"@en ;
    ns1:publisher <https://w3id.org/dh-atlas/1729185609-482096> ;
... > .
<https://w3id.org/dh-atlas/1729185609-482096> rdfs:label "/DH.ARC -
Digital Humanities Advanced Research Centre"^^xsd:string .
```

Landing Page

Mandatory

Value type: ***URL***	*Domain:* ***Dataset***, ***Computer Program***, ***Web Site***, ***Organization***
Cardinality: ***ONE***	*Range:* ***xsd:anyURI***
RDF property: ***schema:url***	

The URL of the web page where the research product is presented, and is possible to access its distributions.

A landing page can be either a website built from scratch or a README file in a repository.

```
<https://w3id.org/dh-atlas/1729019160-8370216> a schema:Dataset ,
        <http://purl.org/spar/fabio/MetadataDocument> ,
        ns2:DigitalScholarlyEdition ;
    rdfs:label "Codice Pelavicino Digital Edition"@en ;
    ns1:url <https://pelavicino.labcd.unipi.it/evt/> ;
... > .
```

Identifier

Value type: ***String***	*Domain:* ***Dataset***, ***Agent***
Cardinality: ***MANY***	
RDF property: ***schema:identifier***	

A unique identifier of the research product.

The same research product can be assigned multiple identifiers of different kinds.

We recommend using international standard codes such as ISBN and ISSN, and persistent identifiers such as DOI, W3ID, PURL, and Handle.

```
<https://w3id.org/dh-atlas/1729190822-8668764> a schema:Dataset ,
        <http://purl.org/spar/fabio/MetadataDocument> ,
        ns2:DigitalScholarlyEdition ;
    rdfs:label "National Edition of Aldo Moro's Works"@en ;
    ns1:identifier "https://doi.org/10.6092/unibo/aldomoro"@en ;
... > .
```

Release Date

Mandatory

Value type: ***Date***	*Domain:* ***Dataset***
Cardinality: ***ONE***	*Range:* ***xsd:date***
RDF property: ***schema:datePublished***	

The date the research product was released.

If the full date can not be applied, indicate the year.

```
<https://w3id.org/dh-atlas/1729010538-925614> a schema:Dataset ,
        <http://purl.org/spar/fabio/ComputerProgram> ,
        ns2:Software ;
    rdfs:label "Edition Visualization Technology (EVT)"@en ;
    ns1:datePublished "2020-01-01"^^xsd:date ;
... > .
```

Current Version

Value type: ***String***	*Domain:* ***Dataset***
Cardinality: ***ONE***	
RDF property: ***schema:version***	

The number of the research product's current version.

We recommend using shared numbering schemes, such as:

- SemVer (Semantic Versioning),[167] e.g., 1.0.0
- CalVer (Calendar Versioning)[168]

```
<https://w3id.org/dh-atlas/1729187122-1272922> a schema:Dataset ,
        <http://purl.org/spar/fabio/ComputerProgram> ,
        ns2:Software ;
    rdfs:label "Voyant tools"@en ;
    ns1:version "2.06.14"@en ;
... > .
```

Access Rights

Mandatory

Value type: ***URI***	*Domain:* ***Dataset***
Cardinality: ***ONE***	
RDF property: ***schema:conditionsOfAccess***	

This field indicates how the research product may be accessed. If multiple access rights exist, use this field to specify access for the main part of the product. Use "Has Part" to detail access rights for any subsets with different permissions.

The expected values are defined in the COAR Access Rights vocabulary[169] and namely are:

- Embargoed access, when the publication of the research product in open access is delayed due to copyright constraints.
- Metadata only access.
- Open access.
- Restricted access, when the research product can be accessed only by authorised users.

```
<https://w3id.org/dh-atlas/1729073242-5875242> a schema:Dataset ,
        <http://purl.org/spar/fabio/DataFile> ,
        ns2:LinkedOpenData ;
    rdfs:label "Biflow-Toscana Bilingue RDF Dataset"@en ;
    ns1:conditionsOfAccess <http://purl.org/coar/access_right/c_abf2> ;
... > .
```

Access Point

[167] https://semver.org/.
[168] https://calver.org/.
[169] http://vocabularies.coar-repositories.org/documentation/access_rights/.

Value type: ***URI***	*Domain:* ***Dataset***
Cardinality: ***MANY***	*Range:* ***Computer Program***, ***Web Site***
RDF property: ***atlas:servedBy***	

URL of the website or data service (e.g., API, SPARQL endpoint) that grants access to the research product.

```
<https://w3id.org/dh-atlas/1728943937-3007112> a schema:Dataset ,
        <http://purl.org/spar/fabio/DataFile> ,
        ns2:LinkedOpenData ;
    rdfs:label "Zeri Photo Archive RDF Dataset"@en ;
   ...
    ns2:servedBy <https://w3id.org/dh-atlas/1728987090-372862>,
        <https://w3id.org/dh-atlas/1728988048-7197032>,
        <https://w3id.org/dh-atlas/1728993673-7262628> .
<https://w3id.org/dh-atlas/1728987090-372862> rdfs:label "Zeri Photo
Archive RDF Dataset - SPARQL endpoint"^^xsd:string .
<https://w3id.org/dh-atlas/1728988048-7197032> rdfs:label "Zeri Photo
Archive RDF Dataset - SPARQL query interface"^^xsd:string .
<https://w3id.org/dh-atlas/1728993673-7262628> rdfs:label "Zeri Photo
Archive RDF Dataset - LodView RDF Browser"^^xsd:string .
```

License

Mandatory

Value type: ***URI***	*Domain:* ***Dataset***
Cardinality: ***ONE***	
RDF property: ***schema:license***	

The licence under which the research product is available to the public.

Recommended values are the URIs to the terms defined in the Licences Vocabulary[170] by the Italian Government and the licences listed on the Open Source Initiative's website.[171]

```
<https://w3id.org/dh-atlas/1729034329-548886> a schema:Dataset ,
        <http://purl.org/spar/fabio/DataFile> ,
        ns2:LinkedOpenData ;
    rdfs:label "LiLa Knowledge Base"@en ;
    ns1:license
```

[170] https://schema.gov.it/lodview/controlled-vocabulary/licences.
[171] https://opensource.org/licenses.

```
<https://w3id.org/italia/controlled-vocabulary/licences/A31_CCBYSA40> ;
... > .
```

Downloads

Mandatory

Value type: ***URL***	*Domain:* ***Dataset***
Cardinality: ***MANY***	*Range:* ***xsd:anyURI***
RDF property: ***schema:distribution***	

The URL of the download or download page of the research product's files.

```
<https://w3id.org/dh-atlas/1728948776-872414> a schema:Dataset,
        <http://purl.org/spar/fabio/OntologyDocument> ,
        ns2:Ontology ;
    rdfs:label "Historical Context Ontology (HiCO)"@en ;
    ns1:distribution
<http://marilenadaquino.github.io/hico/current/hico.owl> ;
... > .
```

Status

Value type: ***URI***	*Domain:* ***Dataset***
Cardinality: ***ONE***	
RDF property: ***schema:creativeWorkStatus***	

The research product's status in its lifecycle.

The expected values are defined in the EU Dataset Status Vocabulary[172] and namely are:

- Completed.
- Deprecated, when it is recommended that the contents of this dataset be no longer used.
- Under development, the dataset may be in an incomplete or faulty state.
- Discontinued, if the dataset is no longer produced or updated.
- Withdrawn, when the dataset is no longer meant to be published.

```
<https://w3id.org/dh-atlas/1728995301-3097954> a schema:Dataset ,
```

[172] http://publications.europa.eu/resource/dataset/dataset-status.

```
        <http://purl.org/spar/fabio/DataFile> ,
        ns2:LinkedOpenData ;
    rdfs:label "DanteSources RDF Dataset"@en
    ns1:creativeWorkStatus
<http://publications.europa.eu/resource/authority/dataset-status/COMPLET
ED> ;
... > .
```

Format

Value type: ***URI***	*Domain:* ***Dataset***
Cardinality: ***MANY***	
RDF property: ***schema:encodingFormat***	

The file format(s) in which the research product's data are stored.

Expected values are URIs to the terms of the EU File Type Vocabulary[173] or of the IANA Media Types[174] list.

```
<https://w3id.org/dh-atlas/1728948776-872414> a schema:Dataset,
        <http://purl.org/spar/fabio/OntologyDocument> ,
        ns2:Ontology ;
    rdfs:label "Historical Context Ontology (HiCO)"@en ;
 ns1:encodingFormat
<http://publications.europa.eu/resource/authority/file-type/JSON_LD>,

<http://publications.europa.eu/resource/authority/file-type/OWL>,

<http://publications.europa.eu/resource/authority/file-type/RDF_XML> ;
... > .
```

Metadata Standards

Value type: ***URI***	*Domain:* ***Dataset***
Cardinality: ***MANY***	
RDF property: ***purl:conformsTo***	

Standards to which the research product's metadata conform.

Recommended values can be found in:

[173] http://publications.europa.eu/resource/dataset/file-type.
[174] http://www.iana.org/assignments/media-types/media-types.xhtml.

Riley, Jenn, 2018, “Seeing Standards: A Visualization of the Metadata Universe”, https://doi.org/10.5683/SP2/UOHPVH, Borealis, V3, UNF:6:gWl/jicj8wtJm4Grmph7TQ== [fileUNF]

```
<https://w3id.org/dh-atlas/1729010538-925614> a schema:Dataset ,
        <http://purl.org/spar/fabio/ComputerProgram> ,
        ns2:Software ;
    rdfs:label "Edition Visualization Technology (EVT)"@en ;
    dcterms:conformsTo <http://www.tei-c.org/> ;
... > .
```

Language

Value type: ***URI***	*Domain:* ***Dataset***
Cardinality: ***MANY***	
RDF property: ***schema:inLanguage***	

The language(s) in which the research product is expressed.

If multiple languages are used—for example, in labels of an ontology or various texts in a collection—we recommend indicating all of them.

Expected values are URIs to the terms of the EU Language Vocabulary.[175]

```
<https://w3id.org/dh-atlas/1728948052-580608> a schema:Dataset,
        <http://purl.org/spar/fabio/OntologyDocument> ,
        ns2:Ontology ;
    rdfs:label "Semantic Publishing and Referencing Ontologies (SPAR)
"@en ;
    ns1:inLanguage
<http://publications.europa.eu/resource/authority/language/ENG> ;
... > .
```

Type

Mandatory

Value type: ***URI***	*Domain:* ***Dataset***
Cardinality: ***ONE***	
RDF property: ***schema:additionalType***	

[175] http://publications.europa.eu/resource/authority/language.

The type of research product.

Expected values are: Digital Scholarly Edition; Text Collection; Linked Open Data; Ontology; Software.

Disclaimer: additional types of research products may be included in future versions of the ATLAS knowledge graph.

```
<https://w3id.org/dh-atlas/1728948736-9982028> a schema:Dataset,
        <http://purl.org/spar/fabio/OntologyDocument> ,
        ns2:Ontology ;
    rdfs:label "CIDOC Conceptual Reference Model (CRM)"@en ;
    ns1:additionalType <http://purl.org/spar/fabio/OntologyDocument> ;
... > .
```

Research Project

Value type: ***URI***	*Domain:* ***Dataset***
Cardinality: ***ONE***	*Range:* ***schema:ResearchProject***
RDF property: ***schema:producer***	

The research project within which the research product was created.

Not all research products originate from a research project.

Indicating the research project provides context, helping to situate the research product within its timeline, development stage, and the broader research environment in which it was conceived and created.

```
ALIM
<https://w3id.org/dh-atlas/1729030147-4807222> a schema:Dataset ,
        <http://purl.org/spar/fabio/Anthology> ,
        ns2:TextCollection ;
    rdfs:label "Archivio della Latinità Italiana del Medioevo (ALIM)
Digital Library"@en ;
    ns1:producer <https://w3id.org/dh-atlas/1729017470-5332708> ;
... > .

ALIM's Research Project
<https://w3id.org/dh-atlas/1729017470-5332708> a foaf:Agent,
        schema:ResearchProject ;
    rdfs:label "Archivio della Latinità Italiana del Medioevo (ALIM)
Project"@en ;
    ns1:description "ALIM (Archivio della Latinità Italiana del Medioevo
/ Archive of the Italian Latinity of the Middle Ages) ..."@en ;
```

```
    ns1:foundingDate "1996-01-01"^^xsd:date ;
    ns1:funder <https://w3id.org/dh-atlas/1729001319-1277974> ;
    ns1:location <https://sws.geonames.org/2523920>,
        ... ;
    ns1:member <https://w3id.org/dh-atlas/1729012784-159052>,
        ... ;
    ns1:name "Archivio della Latinità Italiana del Medioevo (ALIM)
Project"@en ;
    ns1:url <https://alim.unisi.it/> .
```

Has Part

Value type: ***URI***	*Domain:* ***Dataset***
Cardinality: ***MANY***	*Range:* ***schema:Dataset***
RDF property: ***schema:hasPart***	*Is inverse of:* ***schema:isPartOf***

A research product that forms part of the catalogued one. This could be, for instance, a sub-collection of texts in a specific language within a larger text collection. An example is the "Hellenica" corpus within the Musisque Deoque collection.

We recommend filling out this field, to give value to the single components of the research product.

```
<https://w3id.org/dh-atlas/1729160920-6521974> a schema:Dataset ,
        <http://purl.org/spar/fabio/Anthology> ,
        ns2:TextCollection ;
    rdfs:label "Musisque Deoque (MQDQ)"@en ;
    ns1:hasPart <https://w3id.org/dh-atlas/1729012784-159053>;
... > .

<https://w3id.org/dh-atlas/1729012784-159053> a schema:Dataset ,
        <http://purl.org/spar/fabio/Anthology> ,
        ns2:TextCollection ;
    rdfs:label "Poeti d'Italia in lingua latina"@en ;
... > .
```

Is Part Of

Value type: ***URI***	*Domain:* ***Dataset***
Cardinality: ***ONE***	*Range:* ***schema:Dataset***
RDF property: ***schema:isPartOf***	*Is inverse of:* ***schema:hasPart***

Inverse of `Has Part` this field indicates the research product of which the catalogued one is a component.

```
<https://w3id.org/dh-atlas/1729160920-6521974> a schema:Dataset ,
        <http://purl.org/spar/fabio/Anthology> ,
        ns2:TextCollection ;
    rdfs:label "Musisque Deoque (MQDQ)"@en ;
... > .

<https://w3id.org/dh-atlas/1729012784-159053> a schema:Dataset ,
        <http://purl.org/spar/fabio/Anthology> ,
        ns2:TextCollection ;
    rdfs:label "Poeti d'Italia in lingua latina"@en ;
      ns1:isPartOf <https://w3id.org/dh-atlas/1729160920-6521974> ;
... > .
```

Documentation

Value type: ***URL***	*Domain:* ***Dataset***, ***Web Site***, ***Computer Program***
Cardinality: ***MANY***	
RDF property: ***schema:usageInfo***	

The URL of the documentation of the research product.

```
<https://w3id.org/dh-atlas/1728948052-580608> a schema:Dataset,
        <http://purl.org/spar/fabio/OntologyDocument> ,
        ns2:Ontology ;
    rdfs:label "Semantic Publishing and Referencing Ontologies (SPAR)
"@en ;
    ...
    ns1:usageInfo <http://www.sparontologies.net/examples> .
```

Research Activities

Mandatory

Value type: ***URI***	*Domain:* ***Dataset***, ***Web Site***, ***Computer Program***
Cardinality: ***MANY***	

RDF property: ***schema:educationalUse***	

Research activities enabled or supported by the research product.

Expected values are URIs to terms from the Taxonomy of Digital Research Activities in the Humanities (TaDiRAH).[176] Multiple activities can be specified, ranging from broad to specific levels of granularity.

```
<https://w3id.org/dh-atlas/1729187122-1272922> a schema:Dataset ,
        <http://purl.org/spar/fabio/ComputerProgram> ,
        ns2:Software ;
    rdfs:label "Voyant tools"@en ;
    ns1:educationalUse
<https://vocabs.dariah.eu/tadirah/contentAnalysis>,
        <https://vocabs.dariah.eu/tadirah/discovering>,
        <https://vocabs.dariah.eu/tadirah/structuralAnalysis>,
        <https://vocabs.dariah.eu/tadirah/visualAnalysis> ;
... > .
```

Academic Field

Value type: ***URI***	
Cardinality: ***MANY***	
RDF property: ***schema:about***	

The academic field(s) to which the research product pertains.

Expected values are the IDs and/or names from an official classification system. We recommend using the classification established by CUN (Consiglio Universitario Nazionale), which is also available as a controlled vocabulary.[177]

```
<https://w3id.org/dh-atlas/1728995301-3097954> a schema:Dataset ,
        <http://purl.org/spar/fabio/DataFile> ,
        ns2:LinkedOpenData ;
    rdfs:label "DanteSources RDF Dataset"@en ;
    ns1:about
<https://w3id.org/italia/controlled-vocabulary/classifications-for-unive
rsities/academic-disciplines/SSD-L-FIL-LET-08>,

<https://w3id.org/italia/controlled-vocabulary/classifications-for-unive
```

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

[177] https://schema.gov.it/lodview/controlled-vocabulary/classifications-for-universities/academic-disciplines.

```
rsities/academic-disciplines/SSD-L-FIL-LET-10>,

<https://w3id.org/italia/controlled-vocabulary/classifications-for-unive
rsities/academic-disciplines/SSD-M-STO-08> ;
... > .
```

Methodology

Value type: ***URI***	*Domain:* ***Dataset***
Cardinality: ***MANY***	
RDF property: ***ns2:methodology***	

Research activities are conducted to create or develop the research product.

We recommend filling out this field to enhance transparency in the research product's creation process and to share effective methodologies and workflows with fellow scholars.

Expected values are URIs to terms from the Taxonomy of Digital Research Activities in the Humanities (TaDiRAH). Multiple activities can be specified, ranging from broad to specific levels of granularity.

```
<https://w3id.org/dh-atlas/1729019160-8370216> a schema:Dataset ,
        <http://purl.org/spar/fabio/MetadataDocument> ,
        ns2:DigitalScholarlyEdition ;
    rdfs:label "Codice Pelavicino Digital Edition"@en ;
    ns2:methodology <https://vocabs.dariah.eu/tadirah/annotating>,
        <https://vocabs.dariah.eu/tadirah/capturing>,
        <https://vocabs.dariah.eu/tadirah/transcribing> ;
... > .
```

Reused Software

Value type: ***URI***	*Domain:* ***Dataset***
Cardinality: ***MANY***	*Range:* ***Dataset***, ***Computer Program***
RDF property: ***atlas:used***	

Software tools or other research products used to create or develop the research product.

We recommend filling out this field to enhance transparency in the research product's creation process and to share effective methodologies and workflows with fellow scholars.

Recommended values can be URIs of: software tools, software libraries, computer programs, digital editions, text collections, ontologies, and linked open datasets.

```
<https://w3id.org/dh-atlas/1729156722-0141878> a schema:Dataset ,
        <http://purl.org/spar/fabio/Anthology> ,
        ns2:TextCollection ;
    rdfs:label "Biblioteca Italiana"@en ;
    ...
    ns2:used <http://www.muruca.org/> .

<http://www.muruca.org/> rdfs:label "MURUCA"^^xsd:string .
```

```
<https://w3id.org/dh-atlas/1729019160-8370216> a schema:Dataset ,
        <http://purl.org/spar/fabio/MetadataDocument> ,
        ns2:DigitalScholarlyEdition ;
    rdfs:label "Codice Pelavicino Digital Edition"@en ;
...
ns2:used <https://w3id.org/dh-atlas/1729010538-925614> .

<https://w3id.org/dh-atlas/1729010538-925614> rdfs:label "Edition
Visualization Technology (EVT)"^^xsd:string .
```

Bibliographic Reference

Value type: ***URL***	*Domain:* ***Dataset***
Cardinality: ***MANY***	*Range:* ***xsd:anyUri***
RDF property: ***schema:citation***	

URL of a publication that describes or reviews the research product.

We recommend linking open access publications when available. Ideally, use the publication's persistent identifier, such as DOI or Handle.

```
<https://w3id.org/dh-atlas/1729185052-855823> a schema:Dataset ,
        <http://purl.org/spar/fabio/MetadataDocument> ,
        ns2:DigitalScholarlyEdition ;
    rdfs:label "VaSto - Varchi, Storia fiorentina Digital Edition"@en ;
    ns1:citation <http://doi.org/10.30687/mag/2724-3923/2021/03/006> ;
... > .
```

Text Collection

In the ATLAS-DM, a text collection is a subclass of `schema:Dataset` and `fabio:Anthology`. The latter is defined as "a collection of selected literary or scholastic works, for example, poems, short stories, plays or research papers."

`atlas:TextCollection` includes additional properties that describe the collected texts from both a bibliographical perspective and in terms of the corpus's dimensions and variety.

We recommend providing thorough descriptions of all texts in the collection. If this is not possible, prioritise describing the main texts or those that best demonstrate the collection's scientific objectives.

Edited work[178]

Mandatory

Value type: ***URI***	*Domain:* ***Text Collection*** (***Digital Scholarly Edition***)
Cardinality: ***MANY***	
RDF property: ***dcterms:source***[179]	

The URI of the cataloguing records of the edited works.

We recommend using records from the Open Library.[180] Another viable option is VIAF.[181]

If no record exists, enter the bibliographic reference formatted as follows: "Last name, name (year). Title".

```
<https://w3id.org/dh-atlas/1729156722-0141878> a schema:Dataset ,
        <http://purl.org/spar/fabio/Anthology> ,
        ns2:TextCollection ;
    rdfs:label "Musisque Deoque (MQDQ)"@en ;
    dcterms:source "https://openlibrary.org/works/OL16280210W/Georgicon"
... > .
```

Reference to the edited text[182]

[178] Disclaimer: this property is under revision and may be modified in future versions of the data model.
[179] We are currently evaluating alternative RDF properties to better represent the concept of "work".
[180] https://openlibrary.org/.
[181] We are currently testing the use of WorldCat (https://search.worldcat.org/it) for this specific field and the homonymous field associated with the Digital Scholarly Edition class. WorldCat may be available in future versions of ATLAS's application for data entry (see Chapter 3).
[182] Disclaimer: this property is under revision and will be modified in future versions of the data model in order to represent the URL of a web resource presenting the witnesses or documents of the edited text. For this reason, the label will probably change to "witness or document available at", and another RDF property may be used.

Mandatory

Value type: ***URI***	*Domain:* ***Text Collection*** *(****Digital Scholarly Edition****,* ***Linked Open Data****,* ***Ontology****)*
Cardinality: ***MANY***	
RDF property: ***dcterms:references***	

The URL of a web resource presenting the main edited sources (e.g., digital manuscript, cataloguing record in other sources than the Open Library).

Preferably use web resources that have permalinks or persistent identifiers.

If no existing web resource is available, provide the bibliographic or archival reference.

```
<https://w3id.org/dh-atlas/1729156722-0141878> a schema:Dataset ,
        <http://purl.org/spar/fabio/Anthology> ,
        ns2:TextCollection ;
    rdfs:label "Musisque Deoque (MQDQ)"@en ;
    dcterms:references
"https://archive.org/details/p.-vergili-maronis-opera-virgil-mario-geymo
nat-z-library"
... > .
```

Bibliographic reference of witness or document[183]

Value type: ***String***	*Domain:* ***Text Collection*** *(****Digital Scholarly Edition****,* ***Linked Open Data****,* ***Ontology****)*
Cardinality: ***MANY***	
RDF property: ***dcterms:references****/****rdfs:label***	

The bibliographic or archival reference of the edited text.

```
<https://w3id.org/dh-atlas/1729156722-0141878> a schema:Dataset ,
        <http://purl.org/spar/fabio/Anthology> ,
        ns2:TextCollection ;
    rdfs:label "Musisque Deoque (MQDQ)"@en ;
    dcterms:references
"https://archive.org/details/p.-vergili-maronis-opera-virgil-mario-geymo
```

[183] Disclaimer: this property is under revision and its label will probably be changed to "Bibliographic reference of witness or document" in future versions of the data model.

```
nat-z-library"
... > .

<https://archive.org/details/p.-vergili-maronis-opera-virgil-mario-geymo
nat-z-library> rdfs:label "GEYMONAT M. 1973 (2008) (ed.), P. Vergili
Maronis Opera, Augustae Taurinorum (ristampa anastatica con correzioni:
Roma 2008)." .
```

Type of edited text[184]

Value type: ***URI***	*Domain:* ***Text Collection*** (***Digital Scholarly Edition***)
Cardinality: ***ONE***	*Super property:* ***dcterms:type***
RDF property: ***atlas:referencedWorkType***	

The type of the edited text.

Recommended values were selected by the ATLAS team using as an example the categorisation provided in Patrick Sahle's catalogue of Digital Editions (a catalog of Digital Scholarly Editions, v.4.112 2020ff, edited by Patrick Sahle et al., last change 2024-06-06, https://www.digitale-edition.de/exist/apps/editions-browser/index.html.), and namely are:

- single work;
- collected works;
- collection of texts;
- single manuscript;
- papers;
- serial documents;
- letters;
- diaries;
- charters;
- inscriptions.

```
<https://w3id.org/dh-atlas/1729030147-4807222> a schema:Dataset ,
        <http://purl.org/spar/fabio/Anthology> ,
        ns2:TextCollection ;
```

[184] Disclaimer: this property is under revision and its label will probably be changed to "Bibliographic reference of witness or document" in future versions of the data model. Values will be physical objects, as defined in CIDOC-CRM (E19 Physical Object), and carrier types as defined in IFLA's Multilingual dictionary of cataloguing terms and concepts (MulDiCat): "A designation that reflects the format of the storage medium and housing of a carrier in combination with the type of intermediation device required to view, play, run, etc., the content of a resource. Carrier type reflects attributes of a manifestation" (https://www.iflastandards.info/muldicat/#CarrierType).

```
    rdfs:label "Archivio della Latinità Italiana del Medioevo (ALIM)
Digital Library"@en ;
    ns2:referencedWorkType ns2:charters, ns2:collectionOfTexts;
... > .
```

Specifications on the edited text

Value type: ***String***	*Domain:* ***Text Collection*** ***(Digital Scholarly Edition)***
Cardinality: ***MANY***	*Super property:* ***atlas:description***
RDF property: ***atlas:notesOnSource***	

Additional information on the edited texts.

This field may be used to specify peculiarities of the textual transmission of the edited texts, e.g., the indication of the folios preserving the edited text in a manuscript, or of the editorial criteria applied.

```
<https://w3id.org/dh-atlas/1729030147-4807222> a schema:Dataset ,
        <http://purl.org/spar/fabio/Anthology> ,
        ns2:TextCollection ;
    rdfs:label "Archivio della Latinità Italiana del Medioevo (ALIM)
Digital Library"@en ;
    ns2:notesOnSource "Testo codificato nel formato TEI XML, versione
P5, per essere incluso fra i testi che compongono il progetto ALIM -
Archivio della Latinità Italiana del Medioevo.";
... > .
```

Author of the edited text

Value type: ***URI***	*Domain:* ***Text Collection*** (***Digital Scholarly Edition***)
Cardinality: ***MANY***	*Range:* ***foaf:Agent***
RDF property: ***atlas:referencedAuthor***	*Super property:* ***schema:contributor***

URI of the cataloguing record(s) of the edited texts' author(s). We recommend using VIAF.

```
<https://w3id.org/dh-atlas/1729156722-0141878> a schema:Dataset ,
        <http://purl.org/spar/fabio/Anthology> ,
        ns2:TextCollection ;
    rdfs:label "Biblioteca Italiana"@en ;
```

```
    ns2:referencedAuthor <https://viaf.org/viaf/39389587/> ;
... > .
```

Genre of the items[185]

Value type: ***URI***	*Domain:* ***Text Collection (Digital Scholarly Edition)***
Cardinality: ***MANY***	
RDF property: ***schema:genre***	

The genres of the edited texts.

It is possible to specify both general categories such as “prose” or “poetry” as well as specific genres like “picaresque novel” and “sonnet”.

Recommended values are terms of the DYAS Humanities Thesaurus.[186]

```
<https://w3id.org/dh-atlas/1729160920-6521974> a schema:Dataset ,
        <http://purl.org/spar/fabio/Anthology> ,
        ns2:TextCollection ;
    rdfs:label "Musisque Deoque (MQDQ)"@en ;
    ns1:genre
<https://vocabs.dariah.eu/dyas/en/page/?uri=https%3A%2F%2Fhumanitiesthes
aurus.academyofathens.gr%2Fdyas-resource%2FConcept%2F2537> ;
... > .
```

Number of items

Value type: ***String***	*Domain:* ***Text Collection***
Cardinality: ***MANY***	
RDF property: ***schema:size***	

The quantity of individual texts included in the text collection.

It is possible to express the quantity in number of tokens or other textual items as well.

```
<https://w3id.org/dh-atlas/1729160920-6521974> a schema:Dataset ,
```

[185] Disclaimer: this property is under revision and will be modified in future versions of the data model. Values will be content types, as defined in IFLA’s Multilingual dictionary of cataloguing terms and concepts (MulDiCat): “A designation that reflects the fundamental form of communication in which the content is expressed and the human sense through which it is intended to be perceived. Content type reflects attributes of both work and expression” (https://www.iflastandards.info/muldicat/#ContentType).
[186] https://vocabs.dariah.eu/dyas/en/.

```
        <http://purl.org/spar/fabio/Anthology> ,
        ns2:TextCollection ;
    rdfs:label "Musisque Deoque (MQDQ)"@en ;
    ns1:size "2300000 tokens"@en ;
... > .
```

Digital Scholarly Edition

The class `atlas:DigitalScholarlyEdition` is a subclass of `schema:Dataset` and of `fabio:MetadataDocument`. We chose this alignment for two main reasons.

First, while the FaBiO ontology provides a class for critical editions (`fabio:CriticalEdition`), it lacks a specific class for digital scholarly editions. Using `fabio:MetadataDocument` makes sense since digital scholarly editions are primarily TEI-encoded documents where critical annotations serve as metadata. Second, ATLAS-DM models research products as types of expressions. Since `fabio:CriticalEdition` is a subclass of `frbr:Work` but `fabio:MetadataDocument` is a subclass of `frbr:Expression`, aligning `atlas:DigitalScholarlyEdition` with the latter maintains our desired hierarchy.

The class `atlas:DigitalScholarlyEdition` has been defined as follows: “An information resource which offers a critical representation of (normally) historical documents or texts following a methodology determined by a digital paradigm.”

`Edited work`[187]

Mandatory

Value type: ***URI***	*Domain:* ***`Digital Scholarly Edition (Text Collection)`***
Cardinality: ***MANY***	
RDF property: ***`dcterms:source`***	

The URI of the cataloguing records of the edited works.

We recommend using records from the Open Library. Another viable option is VIAF.

If no record exists, enter the bibliographic reference formatted as follows: “Last name, name (year). Title”.

[187] Disclaimer: all properties available for both text collections and digital scholarly editions are undergoing the same revision process and will be modified in future versions of the data model, in the ways illustrated in the previous footnotes.

```
<https://w3id.org/dh-atlas/1729185052-855823> a schema:Dataset ,
        <http://purl.org/spar/fabio/MetadataDocument> ,
        ns2:DigitalScholarlyEdition ;
    rdfs:label "VaSto - Varchi, Storia fiorentina Digital Edition"@en ;
    dcterms:source
"https://openlibrary.org/works/OL5780615W/Storia_Fiorentina" ;
... > .
```

Reference to the edited text

Mandatory

Value type: ***URI***	*Domain:* ***Digital Scholarly Edition (Text Collection, Linked Open Data, Ontology)***
Cardinality: ***MANY***	
RDF property: ***dcterms:references***	

The URL of a web resource presenting the main edited sources (e.g., digital manuscript, cataloguing record in other sources than the Open Library).

```
<https://w3id.org/dh-atlas/1729185052-855823> a schema:Dataset ,
        <http://purl.org/spar/fabio/MetadataDocument> ,
        ns2:DigitalScholarlyEdition ;
    rdfs:label "VaSto - Varchi, Storia fiorentina Digital Edition"@en ;
    dcterms:references
"https://www.mirabileweb.it/manuscript/roma-accademia-nazionale-dei-linc
ei-biblioteca-(bi-manuscript/19567" ;
... > .
```

Bibliographic reference of the edited text

Value type: ***String***	*Domain:* ***Digital Scholarly Edition, (Text Collection, Linked Open Data, Ontology)***
Cardinality: ***MANY***	
RDF property: ***dcterms:references/rdfs:label***	

The bibliographic or archival reference of the edited text.

```
<https://w3id.org/dh-atlas/1729185052-855823> a schema:Dataset ,
```

```
        <http://purl.org/spar/fabio/MetadataDocument> ,
        ns2:DigitalScholarlyEdition ;
    rdfs:label "VaSto - Varchi, Storia fiorentina Digital Edition"@en ;
    dcterms:references
"https://www.mirabileweb.it/manuscript/roma-accademia-nazionale-dei-linc
ei-biblioteca-(bi-manuscript/19567" rdfs:label "Roma, Accademia
Nazionale dei Lincei, Biblioteca (Biblioteca Corsiniana) 39.D.5 (Rossi
395; Cors. 1532)" ;
... > .
```

Type of edited text

Value type: ***URI***	*Domain:* ***Digital Scholarly Edition (Text Collection)***
Cardinality: ***ONE***	*Super property:* ***dcterms:type***
RDF property: ***atlas:referencedWorkType***	

The type of the edited text.

Recommended values were selected by the ATLAS team using as an example the categorisation provided in Patrick Sahle's catalogue of Digital Editions (a catalog of Digital Scholarly Editions, v.4.112 2020ff, edited by Patrick Sahle et al., last change 2024-06-06, https://www.digitale-edition.de/exist/apps/editions-browser/index.html.), and namely are:

- single work;
- collected works;
- collection of texts;
- single manuscript;
- papers;
- serial documents;
- letters;
- diaries;
- charters;
- inscriptions.

```
<https://w3id.org/dh-atlas/1729190822-8668764> a schema:Dataset ,
        <http://purl.org/spar/fabio/MetadataDocument> ,
        ns2:DigitalScholarlyEdition ;
    rdfs:label "National Edition of Aldo Moro's Works"@en ;
    ns1:referencedWorkType ns1:CollectedWorks ;
... > .
```

Specifications on the edited text

Value type: ***String***	*Domain:* ***Digital Scholarly Edition (Text Collection)***
Cardinality: ***MANY***	*Super property:* ***atlas:description***
RDF property: ***atlas:notesOnSource***	

Additional information on the edited text.

This field may be used to specify peculiarities of the textual transmission of the edited texts, e.g., the indication of the folios preserving the edited text in a manuscript, or of the editorial criteria applied.

```
<https://w3id.org/dh-atlas/1729190822-8668764> a schema:Dataset ,
        <http://purl.org/spar/fabio/MetadataDocument> ,
        ns2:DigitalScholarlyEdition ;
    rdfs:label "National Edition of Aldo Moro's Works"@en ;
    ns1:notesOnSource "The current edition includes 1,398 works, many of
which are transcriptions of speeches, making it difficult to provide
adequate references to the documents used and source works."@en ;
... > .
```

Author of the edited text

Value type: ***URI***	*Domain:* ***Digital Scholarly Edition (Text Collection)***
Cardinality: ***MANY***	*Range:* ***foaf:Agent***
RDF property: ***atlas:referencedAuthor***	*Super property:* ***atlas:contributor***

URI of the cataloguing record of the edited text's author. We recommend using VIAF.

```
<https://w3id.org/dh-atlas/1729190822-8668764> a schema:Dataset ,
        <http://purl.org/spar/fabio/MetadataDocument> ,
        ns2:DigitalScholarlyEdition ;
    rdfs:label "National Edition of Aldo Moro's Works"@en ;
    ns2:referencedAuthor <http://www.viaf.org/viaf/73876435> ;
... > .
<http://www.viaf.org/viaf/73876435> rdfs:label "Aldo Moro"^^xsd:string .
```

Type of edition

Mandatory

Value type: ***URI***	*Domain:* ***Digital Scholarly Edition***
Cardinality: ***MANY***	*Super property:* ***dcterms:type***
RDF property: ***atlas:editionType***	

The type(s) of edition. This indicates the editorial approach taken by the team and the main objectives of the edition.

Recommended values are:

- best-manuscript edition;
- critical edition;
- digital Edition;
- diplomatic edition;
- documentary edition;
- eclectic edition;
- monotypic edition;
- synoptic edition.

To describe how the digital edition was implemented from a technical and methodological point of view (e.g., "semantic edition", "LOD edition", "crowdsourced edition") use the Research Product's field "methodology" (e.g., https://vocabs.dariah.eu/tadirah/semantification, https://vocabs.dariah.eu/tadirah/linkedOpenData, https://vocabs.dariah.eu/tadirah/crowdsourcing).

Note

The recommended values include types applicable to both digital and printed editions, since scholarly goals are typically consistent across formats. These types were carefully selected to represent the main categories of digital editions, based on the following sources:

Roelli, Philipp, and Caroline Macé. 2015. "Parvum Lexicon Stemmatologicum." November 13, 2015. https://wiki.helsinki.fi/display/stemmatology/Parvum+lexicon+stemmatologicum.

For complete definitions, please consult the lexicon.

```
<https://w3id.org/dh-atlas/1729019160-8370216> a schema:Dataset ,
        <http://purl.org/spar/fabio/MetadataDocument> ,
        ns2:DigitalScholarlyEdition ;
    rdfs:label "Codice Pelavicino Digital Edition"@en ;
    ns2:editionType ns2:DocumentaryEdition ;
... > .
```

Genre

Value type: ***URI***	*Domain:* ***Digital Scholarly Edition (Text Collection)***
Cardinality: ***MANY***	
RDF property: ***schema:genre***	

The genre(s) of the edited text.

It is possible to specify both general categories such as “prose” or “poetry” as well as specific genres like “picaresque novel” and “sonnet”.

Recommended values are terms of the DYAS Humanities Thesaurus.

```
<https://w3id.org/dh-atlas/1729190822-8668764> a schema:Dataset ,
        <http://purl.org/spar/fabio/MetadataDocument> ,
        ns2:DigitalScholarlyEdition ;
    rdfs:label "National Edition of Aldo Moro's Works"@en ;
    ns1:genre
<https://vocabs.dariah.eu/dyas/en/page/?uri=https%3A%2F%2Fhumanitiesthes
aurus.academyofathens.gr%2Fdyas-resource%2FConcept%2F2536>,
<https://vocabs.dariah.eu/dyas/en/page/?uri=https%3A%2F%2Fhumanitiesthes
aurus.academyofathens.gr%2Fdyas-resource%2FConcept%2F1499> ;
... >
```

Software

In the ATLAS-DM a piece of software is a subclass of `schema:Dataset` and `fabio:ComputerProgram`. The latter is defined as “a unit of computer code in source or compiled form, employing one or more algorithms to be executed by a digital computer to undertake a particular task. Computer programs are collectively called ‘software’ to distinguish them from the equipment (‘hardware’) upon which they run”.

Programming Language

Mandatory

Value type: ***URI***	*Domain:* ***Computer Program***
Cardinality: ***MANY***	
RDF property: ***ns4:isEncodedIn***	

The programming language(s) employed in the software’s development.

Values

```
<https://w3id.org/dh-atlas/1729187122-1272922> a schema:Dataset ,
        <http://purl.org/spar/fabio/ComputerProgram> ,
        ns2:Software ;
    rdfs:label "Voyant tools"@en ;
    ns4:SWO_0000741 <http://www.wikidata.org/entity/Q2005>,
        <http://www.wikidata.org/entity/Q251>,
        <http://www.wikidata.org/entity/Q32110>,
        <http://www.wikidata.org/entity/Q46441> ;
... > .
<http://www.wikidata.org/entity/Q2005> rdfs:label
"JavaScript"^^xsd:string .
<http://www.wikidata.org/entity/Q251> rdfs:label "Java"^^xsd:string .
<http://www.wikidata.org/entity/Q32110> rdfs:label "XSLT"^^xsd:string .
<http://www.wikidata.org/entity/Q46441> rdfs:label "Cascading Style
Sheets"^^xsd:string .
```

Code Repository URL

Mandatory

Value type: ***URL***	*Domain:* ***Software***
Cardinality: ***ONE***	*Range:* ***xsd:anyUri***
RDF property: ***schema:archivedAt***	

The URL of the source code's repository (e.g., GitHub, Zenodo, FigShare).

```
<https://w3id.org/dh-atlas/1729010538-925614> a schema:Dataset ,
        <http://purl.org/spar/fabio/ComputerProgram> ,
        ns2:Software ;
    rdfs:label "Edition Visualization Technology (EVT)"@en ;
    ns1:archivedAt <https://github.com/evt-project/evt-viewer/> ;
... > .
```

Input Format

Value type: ***URI***	*Domain:* ***Software***
Cardinality: ***MANY***	
RDF property: ***ns4:hasSpecifiedDataInput***	

The file format(s) of data the software can process as input, if applicable.

Expected values are URIs to the terms of the EU File Type Vocabulary or of the IANA Media Types list.

```
<https://w3id.org/dh-atlas/1729187122-1272922> a schema:Dataset ,
        <http://purl.org/spar/fabio/ComputerProgram> ,
        ns2:Software ;
    rdfs:label "Voyant tools"@en ;
    ns4:SWO_0000086
<https://publications.europa.eu/resource/authority/file-type/HTML>,
<https://publications.europa.eu/resource/authority/file-type/TXT>,
<https://publications.europa.eu/resource/authority/file-type/XML> ;
... > .
```

Output Format

Value type: ***URI***	*Domain:* ***Software***
Cardinality: ***MANY***	
RDF property: ***ns4:hasSpecifiedDataOutput***	

The file format(s) of data the software can generate as output, if applicable.

Expected values are URIs to the terms of the EU File Type Vocabulary or of the IANA Media Types list.

```
<https://w3id.org/dh-atlas/1729187122-1272922> a schema:Dataset ,
        <http://purl.org/spar/fabio/ComputerProgram> ,
        ns2:Software ;
    rdfs:label "Voyant tools"@en ;
    ns4:SWO_0000087
<http://publications.europa.eu/resource/authority/file-type/HTML>,
<http://publications.europa.eu/resource/authority/file-type/JSON>,
<http://publications.europa.eu/resource/authority/file-type/RTF>,
<https://publications.europa.eu/resource/authority/file-type/PNG>,
<https://publications.europa.eu/resource/authority/file-type/TSV>,
<https://publications.europa.eu/resource/authority/file-type/TXT> ;
... > .
```

Based on

Value type: ***URI***	*Domain:* ***Computer Program***, ***Web Site***, ***Dataset***
Cardinality: ***MANY***	

RDF property: ***schema:isBasedOn***	

Software component(s) used or extended in the catalogued software.

While the software reuse property indicates a more general relationship of “reuse”, this property may be used from a more technical perspective to indicate a derivative relationship.

Expected values include catalogued research products or URLs of the software’s landing pages.

```
<https://w3id.org/dh-atlas/1729003408-5163774> a
<http://purl.org/spar/fabio/ComputerProgram> ;
    rdfs:label "DanteSources RDF Dataset - SPARQL Endpoint"@en ;
    ns1:isBasedOn <https://virtuoso.openlinksw.com/> ;
... > .
```

Linked Open Data

`atlas:LinkedOpenData` is a subclass of `schema:Dataset` and `fabio:DataFile`. The latter is defined as “A realisation of a fabio:Dataset (a frbr:Work) containing a defined collection of data with specific content and possibly with a specific version number, that can be embodied as a fabio:Digital Manifestation (a frbr:Manifestation with a specific format) and be represented by a specific fabio:ComputerFile (a frbr:Item) on someone’s hard drive.”.

RDF Vocabularies and Ontologies

Value type: ***URI***	*Domain:* ***Linked Open Data*** (***Digital Scholarly Edition***, ***Text Collection***, ***Ontology***)
Cardinality: ***MANY***	
RDF property: ***dcterms:references***	

URIs of the ontologies and vocabularies employed in data modelling.

Expected values can be URIs of catalogued research products or URIs drawn from LOV (Linked Open Vocabularies) service.

```
<https://w3id.org/dh-atlas/1728943937-3007112> a schema:Dataset ,
        <http://purl.org/spar/fabio/DataFile> ,
        ns2:LinkedOpenData ;
    rdfs:label "Zeri Photo Archive RDF Dataset"@en ;
        <https://www.w3.org/RDF/> ;
    dcterms:references <https://w3id.org/dh-atlas/1728948052-580608>,
... > .
```

```
<https://w3id.org/dh-atlas/1728948052-580608> rdfs:label "Semantic
Publishing and Referencing Ontologies (SPAR)"^^xsd:string .
```

Ontology

In the ATLAS-DM an ontology is a subclass of `schema:Dataset` and `fabio:OntologyDocument`.

Namespace

Mandatory

Value type: ***URI***	*Domain:* ***Ontology***
Cardinality: ***ONE***	
RDF property: ***vann:preferredNamespacePrefix***	

The namespace used by the terms of the ontology.

Values

```
<https://w3id.org/dh-atlas/1728948736-9982028> a schema:Dataset,
        <http://purl.org/spar/fabio/OntologyDocument> ,
        ns2:Ontology ;
    rdfs:label "CIDOC Conceptual Reference Model (CRM)"@en ;
    vann:preferredNamespaceUri <http://www.cidoc-crm.org/cidoc-crm/> ;
... > .
```

Imported or Referenced Models

Value type: ***URI***	*Domain:* ***Ontology*** (***Digital Scholarly Edition***, ***Text Collection***, ***Linked Open Data***)
Cardinality: ***MANY***	
RDF property: ***dcterms:references***	

Ontologies or vocabularies imported or partially reused by the catalogued ontology.

Expected values can be URIs of catalogued research products or URIs drawn from LOV (Linked Open Vocabularies) service.

```
<https://w3id.org/dh-atlas/1728948776-872414> a schema:Dataset,
```

```
        <http://purl.org/spar/fabio/OntologyDocument> ,
        ns2:Ontology ;
    rdfs:label "Historical Context Ontology (HiCO)"@en ;
    dcterms:references <http://purl.org/spar/cito>
... > .
```

References

Bollini, Andrea, Michele Mennielli, Susanna Mornati, and David T. Palmer. 2016. 'IRIS: Supporting & Managing the Research Life-Cycle'. *Universal Journal of Educational Research* 4 (4): 738–43. https://doi.org/10.13189/ujer.2016.040410.

Manghi, Paolo, Alessia Bardi, Claudio Atzori, Miriam Baglioni, Natalia Manola, Jochen Schirrwagen, and Pedro Principe. 2019. 'The OpenAIRE Research Graph Data Model'. Zenodo. https://doi.org/10.5281/zenodo.2643199.

Riva, Pat, Patrick Le Boeuf, and Maja Žumer. 2020. 'IFLA Library Reference Model: Un modello concettuale per le informazioni bibliografiche'. Edited by Istituto centrale per il catalogo unico delle biblioteche italiane e per le informazioni bibliografiche, November. https://repository.ifla.org/handle/123456789/44.

Soiland-Reyes, Stian, Peter Sefton, Mercè Crosas, Leyla Jael Castro, Frederik Coppens, José M. Fernández, Daniel Garijo, et al. 2022. 'Packaging Research Artefacts with RO-Crate'. *Data Science* 5 (2): 97–138. https://doi.org/10.3233/DS-210053.

3. ATLAS Catalogue

Abstract

The third and final chapter of the whitebook presents the ultimate outcome: the ATLAS knowledge graph and its associated services. This chapter serves as a guide for potential users and curators of the catalogue. It begins with a set of guidelines for data entry, including instructions on creating new records and using the support systems. Next, it details the data access services—namely, the GUI, API, SPARQL endpoint, and data dump. The chapter concludes with a comprehensive description of the catalogue.

In this version of the whitebook, the ATLAS platform and backend services are briefly introduced. The definitive version will provide a more detailed description of these components.

CLEF Overview

For the implementation of the ATLAS-KG, we used CLEF[188] (Crowdsourcing Linked Entities via web Form), a LOD-native crowdsourcing platform for collaborative data collection, peer review, and publication. Developed by an international team of researchers within the ERC-funded project Polifonia,[189] CLEF offers a highly configurable, web-ready solution for producing linked open data through a user interface.

At CLEF's core is the templating system. When setting up the application, administrators define templates for describing their resources. Each field in the template maps to an ontology predicate, ensuring consistent data entry and validation. These templates guide the peer-review process and enable data exploration through actionable filters. While administrators can specify custom ontology terms, CLEF encourages the reuse of established vocabularies by harmonising terms with labels from LOV Linked Open Vocabularies. Users then contribute data by completing the resulting web forms, aided by autocomplete suggestions from Wikidata, Geonames, and the existing catalogue.

CLEF supports both anonymous and authenticated contributions to the data catalogue. The editorial workflow consists of three steps: record creation, peer review, and publication. Each record is represented as an RDF named graph, with editing activities (such as dates and agents involved) automatically documented through RDF statements.

Data entered in CLEF becomes immediately accessible through the automatically generated SPARQL endpoint and "Explore" page, which offers filtered views of the collected data. The platform integrates with GitHub for user authentication, version control, and data backup.

[188] https://zenodo.org/badge/latestdoi/479251315. CLEF's GitHub repository: https://github.com/polifonia-project/clef.
[189] https://polifonia-project.eu/.

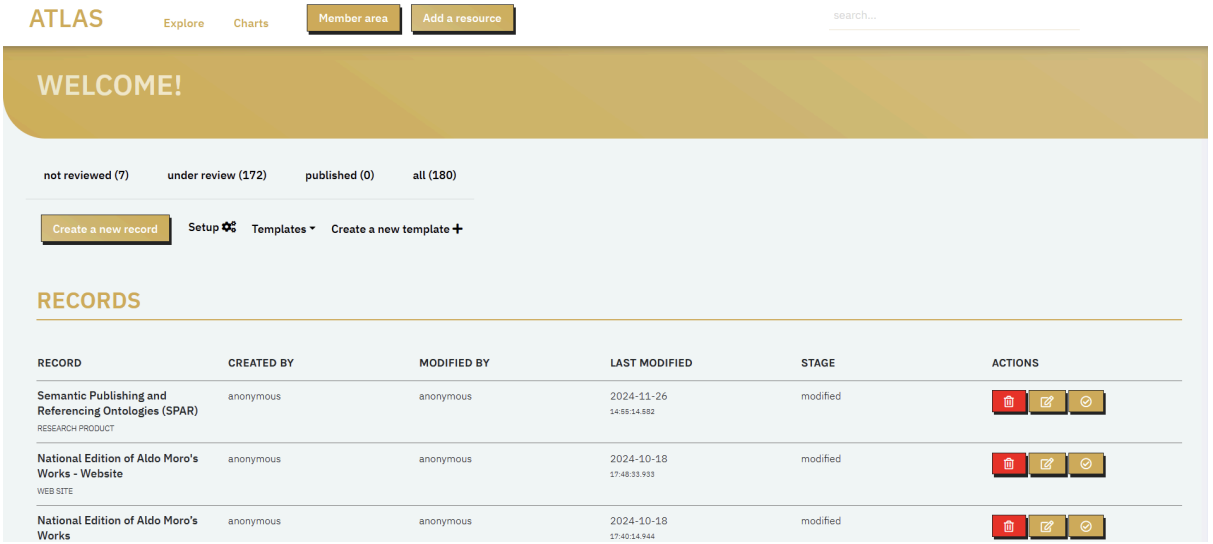
CLEF can run either locally or on a server, and its Python-based source code (built on Webpy) is available on both GitHub and Zenodo.

The ATLAS catalogue is built on CLEF v3.0, which will be officially released soon.[190]

Data entry

To contribute to the ATLAS catalogue users must authenticate via GitHub and access the application's member area. Thanks to the synchronisation between the ATLAS CLEF platform and its corresponding GitHub repository, all modifications to the catalogue are thoroughly tracked.

When first accessing the member area, users are prompted to a list of all records. The list can be sorted according to the record status: edited, reviewed and published.



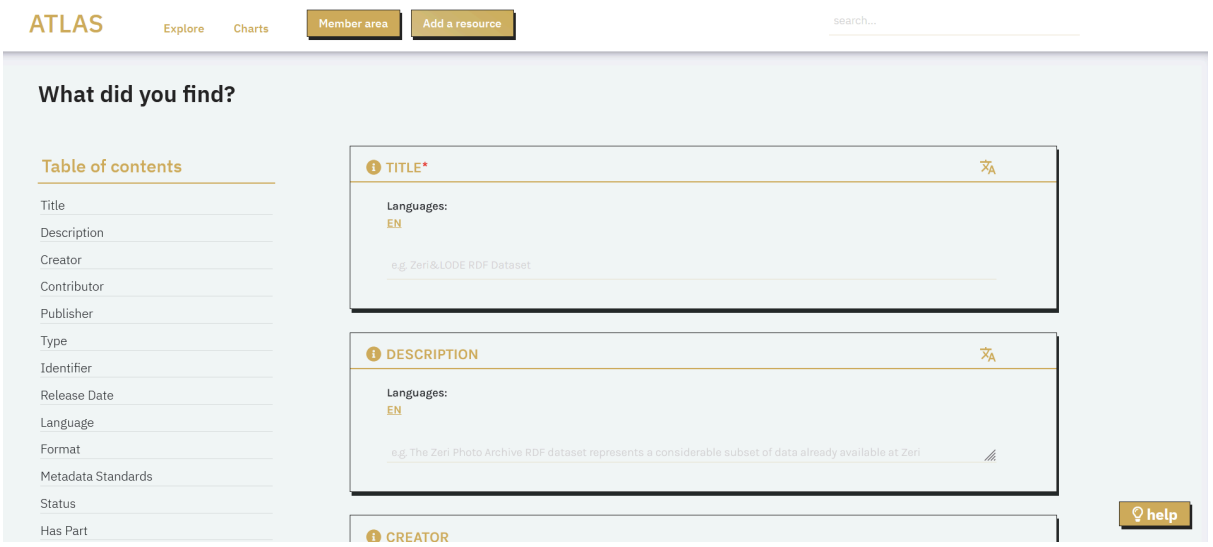
List of the ATLAS catalogue's records in the member area of the CLEF application.

Each record in CLEF corresponds to an RDF named graph. Contributors can create new records to describe research products, by filling out a user-friendly web form, wherein fields correspond to RDF properties and the record is an entity of a class. To add a new Research Product, users can click the "Add a new resource" button in the navigation bar and are prompted to the web form to create a new record.

The application's navigation bar.

The research product's web form comprises a set of fields, each corresponding to one of the properties of the Research Product class, as described in the data model. The interactive list on the left of the form helps users navigate between the properties. For more information about how the form is structured, please refer to the template's subsection.

[190] Full documentation of CLEF: https://polifonia-project.github.io/clef/.



Web form of the Research product's template.

After editing, users can save the record, which becomes automatically ready for revision. The peer-review system allows all authenticated users to review records. After the revision records can be published. It is possible to browse a published record from the Explore page, search it via text search, and retrieve it as Linked Open Data from the SPARQL endpoint via the REST API at <APP-URL>/sparql. It is not possible to unpublish a published record. This restriction prevents inconsistencies between records relying on other records' identifiers.

Templates

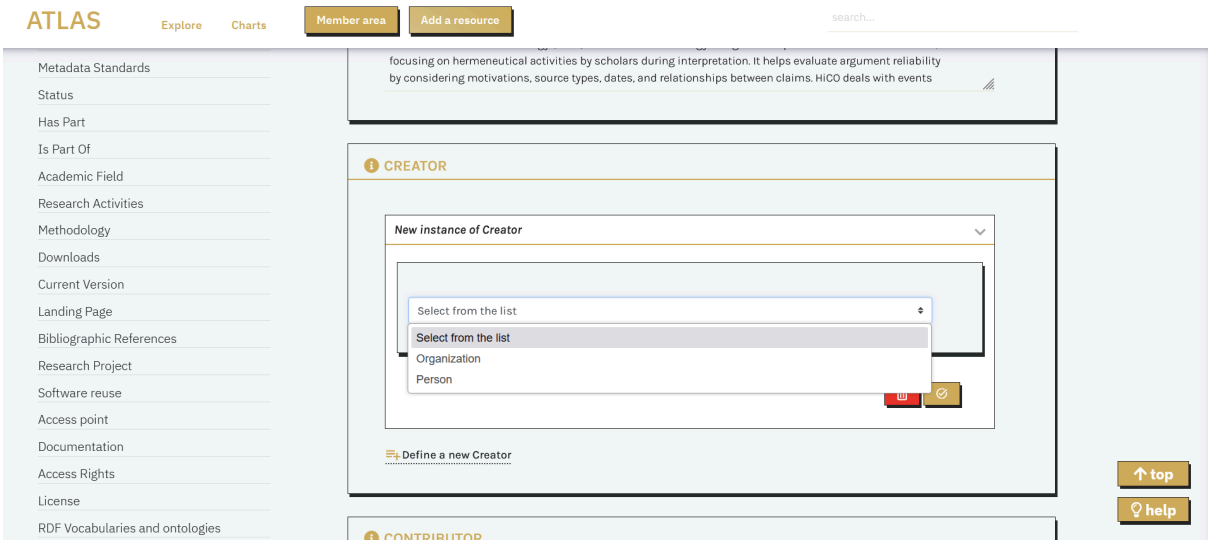
Each record in CLEF complies with a template, i.e., a set of mandatory and optional fields/properties to be filled with appropriate values. ATLAS's main template in CLEF describes research products. The template first displays fields that are common across all research product types. When a specific type is selected in the "type" field—Digital Scholarly Edition, Text Collection, Software, Linked Open Data, or Ontology—additional fields specific to that type appear. For a complete list of fields and their recommended values, please refer to Chapter 2.

The ATLAS-DM includes several ancillary classes—Person, Organization, Research Project, Computer Program, and Website—that represent creators, contributors, publishers, research projects, and access points of research products. CLEF simplifies data entry by allowing users to create multiple entities of these classes alongside the main research product entity. Through a system of subtemplates (or intermediate templates), the fields for each class appear within the Research Product template. This section presents each subtemplate and its fields.

Person and Organization

To represent individual scholars or organisations such as universities, research centres and cultural institutions involved in a research product as creators, collaborators or publishers, the "Person" and the "Organization" templates are used, respectively. For each person, data curators must specify the full name and may optionally indicate the person's affiliation,

ORCID identifier, and link to an authority record (e.g., Wikidata, VIAF) that unambiguously identifies the person. The “Organization” template requires curators to specify their name and location, optionally including the organisation’s website URL, persistent identifier, and a link to an authority record (e.g., Wikidata, VIAF).

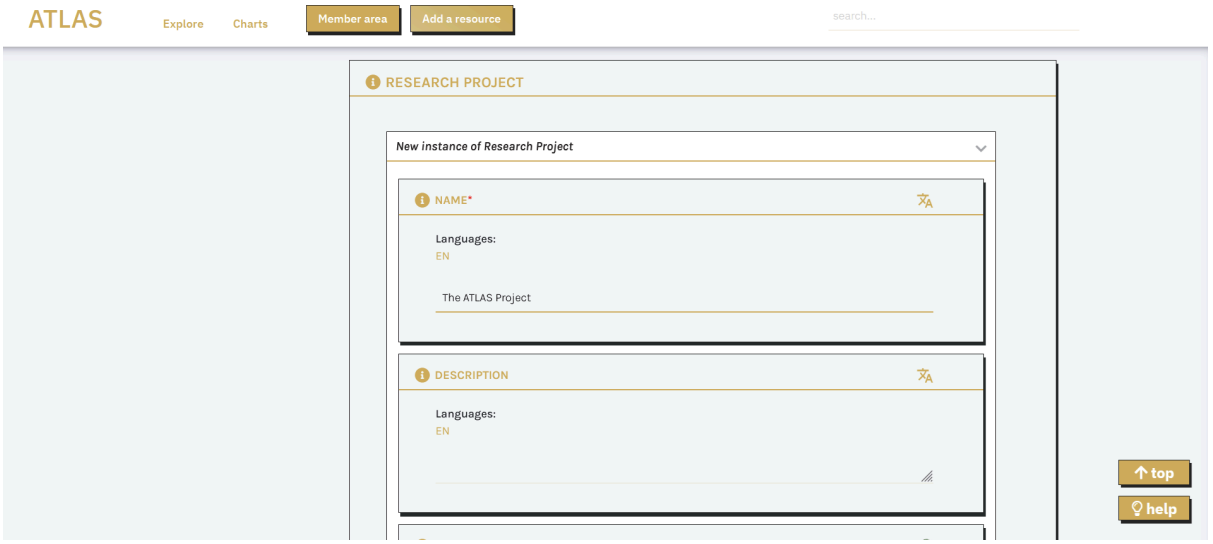


Selection of the Person or Organization intermediate template within the Creator field.

Research Project

Most research products are developed within research projects. To link a product to other project outcomes and better understand its context and purpose, users can describe the associated project using the “Research Project” template.

Data curators must provide the research project’s full title (including any abbreviations or aliases), a brief description, the project’s country location, funding agency, and website link. They may also include additional details such as the project’s start and end dates, member names, identifiers (e.g., Grant number), and a link to an authority record (e.g., Wikidata, VIAF) that uniquely identifies the research project.



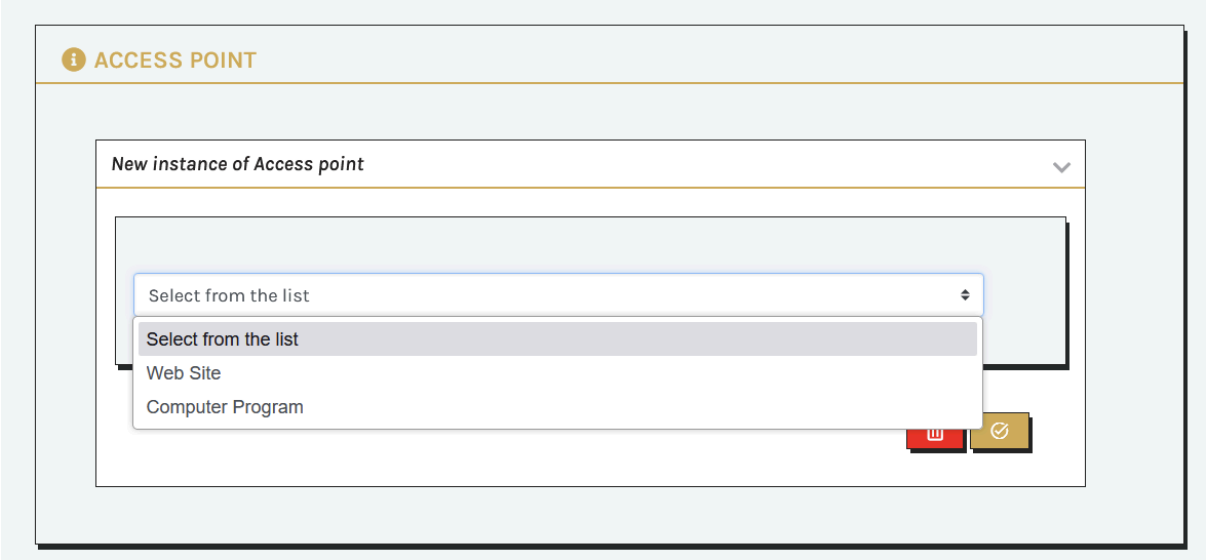
The Research Project template.

Computer Program and Website

The intermediate templates “Computer Program” and “Website” are used to describe software tools, data services, and websites that provide access to research products. Both templates require curators to specify the title and URL of the website or the access URL for the research product. In addition to these mandatory fields, curators may include:

- a brief description of the software tool’s or website’s functionality;
- URL to the research product’s documentation;
- research activities enabled by the service or software in relation to the dataset;
- software components that are reused or extended in the current tool or website, which can include research products or URLs to official software pages.

For computer programs, curators can also specify the licence terms and classify the type according to the EU Data service type Vocabulary, which includes: API, Download service, Human interaction service, and SPARQL endpoint.



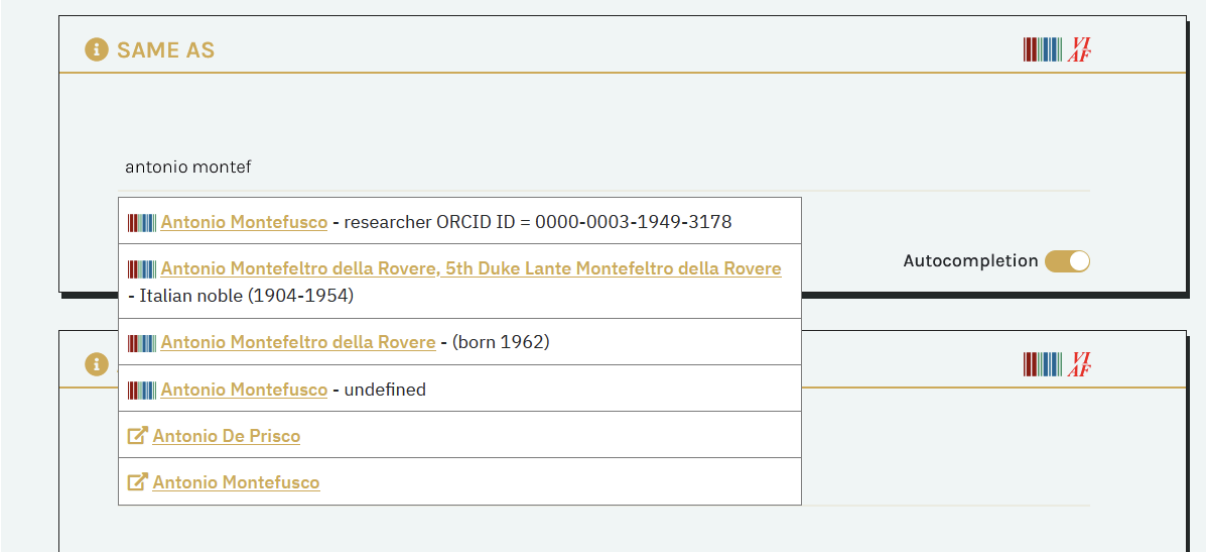
Selection of the Web Site or Computer Program intermediate template within the Access point field.

Data entry support

While editing a research product record, contributors are supported in tasks relevant to the data reusability, namely: data reconciliation, duplicate avoidance, and keyword extraction.

Data reconciliation and autocomplete suggestion

When field values refer to real-world entities or concepts that may appear in multiple records, the system provides autocomplete suggestions by querying both external sources (in real-time) and the ATLAS SPARQL endpoint. These suggestions appear as lists of terms, with each term showing a label, a short description (to help distinguish between similar items), and a link to the online record (such as a Wikidata page or an existing project record). If no matches are found, collaborators can create new entities that are added to the knowledge base and will appear in future suggestion lists.



Example of the autocomplete feature associated with the Same as field in the Person template.

The system retrieves suggestions primarily from Wikidata and VIAF for personal names, Geonames for locations, and specialised SKOS vocabularies like TaDiRAH. Users have the option to disable autocomplete functionality for any field where it is available.

Duplicate avoidance

When creating a new record, the application alerts users of potential duplicates already existing in the catalogue in order to prevent involuntary inconsistencies. Contributors may accept or ignore the recommendation. A possible scenario for this feature is when different contributors describe the same creator of multiple research products.

Keyword extraction

For specified fields, CLEF provides a knowledge extraction function that retrieves named entities and Linked Open Data from SPARQL endpoints, APIs, and static files (in CSV, JSON, and XML format). It uses SPARQL Anything[191] to query these static files and convert them to RDF format. This feature is currently not available for ATLAS, but will be implemented soon.

To make this feature more user-friendly, users can rely on manual extraction, simply providing a document URL, and the system automatically identifies its structure. Through a dropdown menu, users can select elements to extract and set basic filters like minimum occurrences. The system then handles the technical SPARQL query conversion. The extraction process includes automatic entity reconciliation: extracted entities are reconciled to URIs in sources like Wikidata and VIAF and are shown to users to approve/discard. Approved terms are included in the data as machine-readable keywords associated with the subject entity.

Data access

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

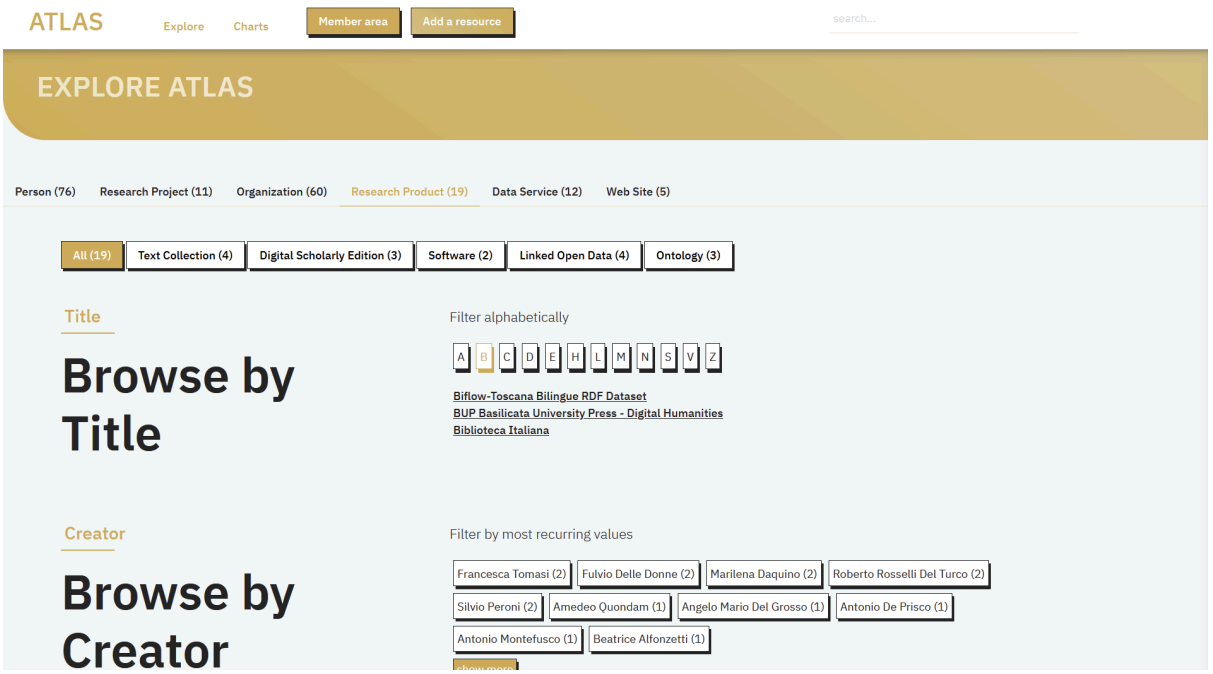
The ATLAS knowledge graph is accessible through CLEF's user-friendly "Explore" interface and its SPARQL endpoint. An API is currently under development and will provide additional access methods. The knowledge graph is also available as a data dump to support data reuse and integration.

CLEF's "Explore" page

CLEF's "Explore" page [LINK in nota] is accessible from CLEF's homepage through the corresponding navigation bar button.

Users can filter entities in the knowledge graph by type—including research product, person, organisation, research project, computer program, and website—by selecting the appropriate tab at the top of the "Explore" page. Research products can be further refined by sub-classes.

Once a type of entity is selected, the "Explore" page displays an alphabetical index and relevant filters. Research products, for instance, can be filtered by creator, type, language, and encoding format.



The "Explore" page.

The navigation bar on the right includes a search field where users can query the catalogue's entities by name or title. As users type their search terms, the field displays a list of relevant entity suggestions.

The home page and the search field in the navigation bar.

CLEF's SPARQL endpoint

CLEF's SPARQL endpoint [LINK in nota] is accessible from CLEF's homepage through the corresponding navigation bar button.

The SPARQL endpoint relies on Blazegraph triplestore.[192] The application provides users with a graphic interface, based on Yasgui,[193] for the editing of SPARQL queries and the visualisation of the search results.

For detailed indications on how to use ATLAS's SPARQL editor, please refer to Yasgui's documentation.

Data dump

Version 1.01 of the DH ATLAS knowledge graph is available as a set of Turtle files deposited in Zenodo: https://doi.org/10.5281/zenodo.14058144. It contains metadata from the pilot projects discussed in Chapter 1 and their related entities. The data structure adheres to the ATLAS ontology model (Chapter 2). The package also includes demo files that showcase the structure of the main entities within the model.

API

ATLAS's REST APIs are currently under development. They will allow users to access the RDF graph and the HTML page of published entities via their permaID.

Data Visualisation

[192] https://blazegraph.com/.
[193] https://yasgui.triply.cc/.

CLEF enables catalogers to display their data through charts, maps, and counters that retrieve data via the SPARQL endpoint. Each graphic component includes a brief comment explaining the displayed data. On CLEF's "Charts" page [aggiungere LINK in nota], we provide an overview of the catalogue's size—including the number of entities, research products, and research projects available—through a set of counters. We also present maps showing the geographical distribution of scholars and organisations involved in creating research products, as well as the distribution of research projects.

References

Daquino, Marilena, Mari Wigham, Enrico Daga, Lucia Giagnolini, and Francesca Tomasi. 2023. 'CLEF. A Linked Open Data Native System for Crowdsourcing'. *Journal on Computing and Cultural Heritage (JOCCH)* 16 (3): 41:1-41:17. https://doi.org/10.1145/3594721.

Giacomini, Sebastiano, Marilena Daquino, Francesca Tomasi and Laurent Fintoni. 2025. 'CLEF 2.0. Solutions for Native Linked Data Cataloguing of Italian Digital Cultural Heritage'. *JLIS.it*. In publication.

Daquino, Marilena. 2021. 'Linked Open Data Native Cataloguing and Archival Description'. *JLIS.It* 12 (3): 91–104. https://doi.org/10.4403/jlis.it-12703.