FAIR Digital Humanities scholarly metadata. The ATLAS project

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

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

Purpose

Several platforms play a pivotal role in the scholarly scenario, ensuring the persistent identification, preservation, and enhanced accessibility of research data. Examples of key initiatives include Zenodo, OpenAIRE, and Research Infrastructures (RIs) such as CLARIN and DARIAH.

However, existing ontologies and models fail to adequately capture the complexities of the contemporary Digital Humanities (DH) landscape. DH projects generate a wide range of outputs, each necessitating tailored descriptive strategies. Key factors, such as textual typologies and editorial criteria, are not sufficiently addressed. Moreover, current models lack effective mechanisms for linking research activities to their related Cultural Heritage objects, despite the potential provided by Linked Open Data (Daquino et al., 2024b). This results in two main consequences, namely: (1) it limits users and researchers in discovering products and perspectives on Digital Cultural Heritage resources, and (2) hinders Cultural Heritage resources retrieval and valorisation.

The ATLAS project aims to overcome these limitations and create a semantic framework capable of representing the varied outputs of DH research, by introducing the ATLAS Ontology and a knowledge graph of DH research related to Italian Digital Cultural Heritage (IDCH). ATLAS tackles the challenges of describing and interlinking scholarly data ensuring enriched and accessible metadata to enhance both discoverability and reusability of these cultural assets.

Methods

Considering the current landscape of research product catalogues, we assessed strategies for the optimisation of cataloguing practices for DH projects based on the IDCH. Our research focused on three key questions: What types of research products exist? How can we represent different types of research products in a way that highlights their distinctive features? Which metadata should we employ to ensure long-term preservation and improve the findability and reusability of research products?

In the first phase, we identified pilot research products related to Italian cultural heritage to determine the most suitable metadata for the catalogue. These pilots were selected as key references in the Italian DH research field and span five categories, namely: Text collections, Digital Scholarly Editions, Linked Open Data, Ontologies, Software tools (Daquino et al., 2024b).

Our analysis of the pilot research products yielded several key findings. First, we identified both common and category-specific metadata for use in the catalogue. Second, we uncovered critical issues affecting data usability and long-term preservation of selected pilots. Common issues across research products included (i) lack of data storage in trustworthy repositories, (ii) unclear dataset access points and methods, (iii) missing information about dataset status (e.g., completed, under development), (iv) unavailable data models and references to existing standards, and insufficient documentation about usage, methodologies, and technologies.

Based on these identified issues, we developed recommendations to ensure research products' FAIRness (Wilkinson et al., 2016), including best practices specific to each product type (Bardi et al., 2024).

We then refined the identified metadata fields through mapping (Daquino et al., 2024a) with major existing models for describing research products, specifically: RO-Crate, KNOT, OpenAIRE Graph, OpenAIRE Application Profile, SKG-IF, IRIS. To translate the metadata into RDF properties, we primarily used Schema.org.

Results

We produced a data model formalized as an OWL 2 DL ontology, the ATLAS ontology (Tomasi et al., 2024), and the first version of the knowledge graph (Daquino et al., 2024a), accessible through an extended version of CLEF (Daquino et al., 2023), a LOD native software for crowdsourcing.

In the data model, research products are modeled as schema:Dataset. Different types of research products are implemented as subclasses of schema:Dataset and aligned with subclasses of frbr:Expression from the FaBiO ontology.

Each research product can be linked to a research project, represented by the class schema:ResearchProject, along with representations of people, organizations, websites, and computer programs.

Some metadata selected for describing research products are common across most existing models, such as: title, description, creator, publisher, release date, landing page, access rights, and license. New properties introduced in our model highlight specific information that is typically hard to find or absent in the products' documentation, and namely are:

- Research activities, which describes the activities enabled by the research product.
- Status which captures the research product's current lifecycle state.
- Documentation URL.
- Metadata standards, which indicates models and standards used for metadata modeling alongside format for data modeling.
- Access point which complements the "landing page" concept.
- Academic field which indicates the disciplinary areas the research product pertains to.

- Methodology alongside software reuse which describes development processes, including specific activities and tools used.

Each research product type provides additional specific metadata beyond the properties common across different types. For example, it is possible to specify imported models and RDF ontologies used in the modeling of ontologies and linked open data, while input and output formats can be included in the description of software tools, to facilitate workflow creation across tools. A novel aspect of the ATLAS modeling approach is that it describes digital scholarly editions and text collections primarily as datasets, to emphasize features and methodologies specific to the “digital paradigm” (Sahle, 2016). For text collections and digital scholarly editions our model also includes traditional cataloguing metadata to describe content: work, author, and genre. Beyond the “work” (Riva et al., 2020) level properties, we have added properties describing documents and witnesses used by editors (reference to the edited text and bibliographic reference of edited text). This enables future catalogue users to filter search results to view different editions of the same textual resource. These properties also allow users to assess the scientific quality of digital editions and text collections, alongside the type of edition, which briefly describes how the text was edited using terms of the Parvum Lexicon Stemmatologicum (Roelli & Macé, 2015) as values.

Value

The ATLAS ontology leverages and builds upon established models for describing digital cultural heritage, providing a comprehensive framework with carefully selected terminology and granular detail levels. This approach enables precise descriptions of the diverse and unique characteristics found across different types of research outputs within the Digital Humanities field. Additionally, the ontology serves as guidelines for producing FAIR DH research data and facilitates detailed analysis of the methodologies employed in creating these outputs, offering valuable insights into the research process itself.

Acknowledgments

PRIN 2022 project funded by the European Union - Next Generation EU, Mission 4 Component 1 CUP J53D23013000006.

References

Bardi, A., et al. (2024). DH ATLAS: Whitebook v1.0. Zenodo https://doi.org/10.5281/zenodo.14169357.

Carriero, V. A., et al. (2019). ArCo: The Italian Cultural Heritage Knowledge Graph. In The Semantic Web – ISWC 2019 (pp. 36–52). Springer International Publishing.

https://doi.org/10.1007/978-3-030-30796-7_3.

Daquino, M., et al. (2024a). DH ATLAS (Version v1.0.1) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.13993057

Daquino, M., et al. (2024b). The ATLAS: A Knowledge Graph of Digital Scholarly Research on Italian Cultural Heritage. In A. Di Silvestro & D. Spampinato (Eds.), Me.Te. Digitali. Mediterraneo in rete tra testi e contesti, Proceedings del XIII Convegno Annuale AIUCD2024 (pp. 588–592). https://doi.org/10.6092/unibo/amsacta/7927.

Daquino, M., et al. (2023). CLEF. A Linked Open Data Native System for Crowdsourcing. J. Comput. Cult. Herit., 16(3), 41:1-41:17. https://doi.org/10.1145/3594721

Di Giorgio, S. (2015). DATI.CULTURAITALIA.IT. Un progetto pilota dedicato ai dati aperti e ai Linked Open Data. Archeologia e Calcolatori, Supplemento 7, 103-106 http://www.archcalc.cnr.it/indice/Suppl_7/13_Di%20Giorgio.pdf.

Frosini, L., et al. (2018). An Aggregation Framework for Digital Humanities Infrastructures: The PARTHENOS Experience. SCIRES-IT - SCIentific RESearch and Information Technology, 8, 33–45.

https://doi.org/10.2423/i22394303v8n1p33.

Riva, P., et al. (2020). IFLA Library Reference Model: Un modello concettuale per le informazioni bibliografiche. https://repository.ifla.org/handle/123456789/44.

Roelli, P., & Macé, C. (2015, November 13). Parvum Lexicon Stemmatologicum. https://wiki.helsinki.fi/xwiki/bin/view/stemmatology/Parvum%20lexicon%20stemmatologicum/.

Sahle, P. (2016). What is a Scholarly Digital Edition? In M. J. Driscoll & E. Pierazzo (Eds.), Digital Scholarly Editing (1st ed., Vol. 4, pp. 19–40). Open Book Publishers. https://www.jstor.org/stable/j.ctt1fzhh6v.6.

Tomasi, F., et al. (2024). The ATLAS Ontology (Version 1.0) [Dataset]. http://www.w3id.org/dh-atlas/.

Wilkinson, M. D., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1), Article 1. https://doi.org/10.1038/sdata.2016.18."