ARTchives: a Linked Open Data Native Catalogue of Art Historians' Archives*

Marilena Daquino[1][0000−0002−1113−7550], Lucia Giagnolini[1][0000−0002−4876−2691], and Francesca Tomasi[1][0000−0002−6631−8607]

Digital Humanities Advanced Research Centre (/DH.arc), Department of Classical Philology and Italian Studies, University of Bologna, Italy
{marilena.daquino2, francesca.tomasi}@unibo.it
lucia.giagnolini@studio.unibo.it

Abstract. Art historians' personal archives include a variety of sources documenting creators' work, opinions, and methodologies. Such a wealth of information is fundamental to trace the trajectories of art history through the lenses of historiographical research. However, the potential of such collections is still unveiled, and performing cross-collection research is not possible via online catalogues. The ARTchives project aims at crowdsourcing curated information on notable art historians' archives and providing scholars with a centralised access point to this heritage. In this paper we present the agile cataloguing process developed to support ARTchives contributors. ARTchives is based on a Linked Open Data native cataloguing system that leverages Semantic Web technologies and Natural Language Processing to facilitate data entry, editorial process, and data quality.[1]

Keywords: Linked Open Data · Art History · Archives.

1 Introduction

Art historians' personal archives include a variety of sources (papers, expertises, correspondances, photographs etc.) documenting creators' work, opinions, primary sources, and scientific methodologies. Such a wealth of information is fundamental to trace the trajectories of art history through the lenses of historiographical research. However, such a vast heritage is only partially available online and the extent and scope of such collections is still unveiled.

The objective of ARTchives[2] is to create a knowledge graph of art historians' archives for historiographical research purposes. Scholars can identify and retrieve archival fonds relevant to their studies, gather bibliographic sources, and can answer research questions related to historiographical topics with quantitative analysis methods - such as historians' network analysis, topic analysis of debates, collections interlinking.

[1] M. Daquino is responsible for Section 2 and 3; L. Giagnolini is responsible for Section 4. All authors are responsible for Section 1 and 5.

[2] http://artchives.fondazionezeri.unibo.it/

Nonetheless, crowdsourcing curated information is a hard task. Several issues may affect data quality, such as data duplication, incompleteness, and vagueness. In order to efficiently support curators contributing to ARTchives, we developed an agile cataloguing process that leverages Semantic Web technologies and Natural Language Processing techniques to allow archivists to save time and provide high-quality contents at the same time.

The remainder of the paper is as follows. In section 2 we give a brief overview of archival and cataloguing systems leveraging Semantic Web technologies. In section 3 we describe the cataloguing system developed for ARTchives. In section 4 we briefly address benefits arisen by the usage of such technologies in pursuing quantitative art historical analysis, and in section 5 we conclude and present future works.

2 Related Work

Galleries, Libraries, Archives, and Museums (GLAM) have been leveraging Semantic Web technologies data for over a decade. Consortia of museums and archives [11, 12, 8] foster the adoption of LOD as a *lingua franca* to develop aggregators and serve high-quality data to scholars and developers.

Nevertheless, only few pioneers abandoned legacy cataloguing and archiving systems to fully embrace the Linked Open Data (LOD) paradigm and manage their catalogues through LOD native management systems [14]. Institutions seem to prefer to maintain legacy systems for managing data life-cycle (addressing aspects such as data entry, review, validation, and publication), and to provide dedicated services to access their 5 stars data, whether these represent complete collections [10], subsets [7], or project-related data [6].

Along with official releases of cultural heritage data, crowdsourcing campaigns have been launched by institutions to enrich their data with experts' knowledge [9]. Likewise, scholarly projects leverage cultural heritage Linked Data to collaboratively develop new resources and data aggregators ([5] for an updated overview of projects). To the best of our knowledge, among the latter only the Listening Experience Database (LED) [1] adopts Semantic Web technologies to support data management, from data collection to publication. Currently, LED relies on an application developed to serve project-related goals and its reusability is not immediate in new projects.

In recent years, a few content management systems have been introduced to facilitate LOD publication via reusable platforms. Omeka S[3] is a popular platform for collaborative data collection and creation of virtual exhibitions. Data are served as JSON-LD via API, but these cannot be accessed in other syntaxes or queried via a SPARQL endpoint. Moreover, while user groups (roles) can be defined, editors do not have any means to supervise changes in the records. Another popular tool is Semantic MediaWiki[4], used in well-known projects like Wikidata. The system allows a fine-grained editorial control and serves data

[3] https://omeka.org/s/.

[4] https://www.semantic-mediawiki.org/wiki/Semantic_MediaWiki.

as LOD. However, integration and reuse of external data sources is a time-consuming activity that can only be performed manually.

In ARTchives we rely on the experience of LED to develop an agile, efficient, reusable Linked Data native cataloguing system that tackles all aspects involved in collaborative data collection.

3 Linked Open Data native cataloguing in ARTchives

Data management system. ARTchives is an open catalogue of archival descriptions of notable art historians' personal archives. It is based on an open source data management system initially developed to answer ARTchives purposes and principles, namely:

- REUSE. Terms belonging to selected data sources are suggested while filling in the form for creating a new record. Reused sources include Wikidata, the Open Library (Internet Archive), the Getty ULAN, and the Getty Art and Architecture Thesaurus. Only terms missing in aforementioned sources are here given of a bespoke identifier.
- ENHANCEMENT. Long free-text descriptions entered by users are parsed to extract machine-readable data so as to avoid contributors repeating information (i.e. as both free-text and selected keywords).
- ACCURACY. Cataloguers can accept or reject aforementioned suggestions from the system and ensure contents comply with editorial standards.
- COLLABORATIVE. Contributors can access and modify all records, including the ones created by other institutions.
- CONSISTENCY. Records are peer-reviewed by the editorial board before publishing.
- CONTINUOUS PUBLISHING. Records can be published on a rolling basis and can be temporarily unpublished for review purposes.

The system leverages Linked Open Data since the creation of data and throughout all the curatorial/editorial management phases, therefore differentiating itself from systems described in section 2. Moreover, the original data management system[5] has been recently adapted to be customizable and reusable as-is in other crowdsourcing projects[6]. In detail:

- a configuration file allows adopters to select information relevant to their dataset, e.g. URI base, prefix, endpoint API;
- a JSON mapping document allows to specify data entry requirements, such as form field types (e.g. text box or dropdown), expected values, services to be called (e.g. autocomplete based on Wikidata), and the mapping between fields, ontology terms, and custom controlled vocabularies;

[5] ARTchives source code is available at: https://github.com/marilenadaquino/ARTchives

[6] Code available at: https://github.com/marilenadaquino/crowdsourcing under CC-BY license.

- HTML templates are available and can be easily customised to serve browsing and search interfaces over the catalogue;
- dereferencing mechanisms are up to the adopter, who can choose and set up redirection rules by means of their persistent URI provider (e.g. w3id).

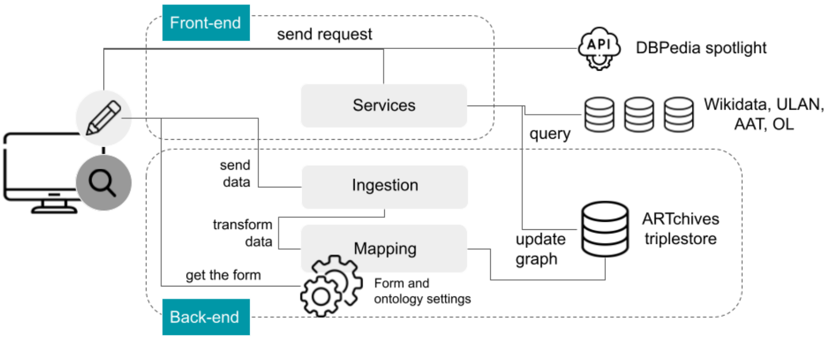

Fig. 1. ARTchives overview

Fig. 1 presents an overview of ARTchives data management system. The form for data entry is created according to settings specified by the user in a JSON document. While editing (creating, modifying, or reviewing) a record, both ARTchives triplestore (Blazegraph) and external services like DBPedia spotlight[7] and Wikidata APIs are called to provide suggestions. Every time a record is created/modified, data are sent to the ingestion module, developed as a Python framework (based on Webpy). The latter relies on the mapping module, which is in charge to transform data into RDF according to the ontology terms specified in the JSON mapping document, and to update the graph created for collecting data of the record.

The data management system is under continuous development to become a flexible tool for collaborative scholarly projects. A beta-version of the system has been tested with cataloguers of the six institutions sponsoring the project, namely: Federico Zeri Foundation (Bologna), Bibliotheca Hertziana (Rome), Getty Research Institute (Los Angeles), Kunsthistorisches Institut in Florenz (Florence), Scuola Normale Superiore (Pisa), and Università Roma Tre (Rome). Beyond ARTchives, other projects [4] actively provide new requirements to foster development and research.

Editorial process. In ARTchives an archival record includes around 26 fields - compliant with archival content standards ISAD(G) and ISAAR - describing respectively the keeper of the archival collection, the creator of the collection, and

[7] https://www.dbpedia-spotlight.org/

the collection itself[8]. Every archival record is formally represented as a named graph [2]. Named graphs enable us to add RDF statements to describe those graphs, including for instance statements on their provenance (such as activities, dates, and agents involved in the creation and modification of a record). Provenance information is described by means of the well-known W3C-endorsed PROV Ontology [13]. Moreover, named graphs allow us to prevent inconsistency of competing descriptions for the same entities, for instance when different cataloguers describe the same creator of multiple collections.

The editorial process in ARTchives addresses four phases: record creation, record modification, review, and publication. Records can be created and modified by any accredited user (so far, these include mainly archivists and professionals of cultural heritage institutions). Members of the ARTchives editorial board peer-review contributions and decide when to publish the record. A published record can be searched and browsed from the website and can be retrieved as Linked Data from the SPARQL endpoint[9]. Every time a change is made to a record, both content data and provenance information are updated on the triplestore and on the file system.

Data collection support. When creating or modifying a record, contributors are supported in a few tasks, namely: (1) data reconciliation, (2) duplicate avoidance, (3) keyword extraction, (4) data integration.

In detail, when field values address real-world entities or concepts that are shared in the art history community, autocomplete suggestions are provided by live querying external selected sources and the knowledge base of ARTchives. Suggestions appear in the form of lists of terms, each term including a label, a short description (to disambiguate homonyms) and a link to the external record (e.g. Wikidata entity). If no matches are found, users can add a new entity that is added to the knowledge base of ARTchives.

When filling in specific fields (i.e. keepers and art historians' names), the system alerts the user in case they are entering information about an existing entity in ARTchives, preventing duplicates.

Several fields require contributors to enter long free-text descriptions (e.g. historians' biographies, scope and content of collections), which include a wealth of information that cannot be processed as machine readable data. To prevent such a loss, two concurrent Named Entity Recognition (NER) tools (i.e. DBpedia spotlight API and compromise.js) extract entities (e.g. people, places, subjects). The latter are reconciled to Wikidata and keywords are shown to users for approval/discard. Approved terms are included in the cataloguing data as subjects associated respectively to people and collections, avoiding user to input them again in the section of the record dedicated to subjects.

Whenever Wikidata terms are reused - either via autocomplete or via NER -, the system queries Wikidata SPARQL endpoint to retrieve relevant context information and store it in the ARTchives Knowledge base for analysis purposes.

[8] ARTchives documentation http://artchives.fondazionezeri.unibo.it/documentation

[9] http://artchives.fondazionezeri.unibo.it/sparql

For instance, subjects of collections like artists, artworks, and artistic periods are enriched with time spans; historians biographical information is enriched with birth and death places. Finally, it is worth noting that collections and keepers are geo-localised via OpenStreetMap APIs[10].

Data sustainability and data modelling choices. Long-term availability of scholarly projects is often hampered by time and resource constraints. Therefore, the wealth of data produced by noble initiatives becomes often unavailable in the mid/long-term. To prevent that, ARTchives reuses as much as possible Wikidata, both at schema level (using classes and properties) and at instance level (reusing individuals as suggested field values), with the idea to directly contribute to Wikidata in the near future with selected, curated metadata. Moreover, leveraging external ontologies only facilitates small-medium crowdsourcing projects, which do not have to develop and maintain bespoke ontologies. To pursue this objective, ARTchives data are realeased under a CC0 waiver. An analysis and estimate of ARTchives potential contribution to Wikidata is ongoing.

4 ARTchives Linked Open Data for quantitative art history

As aforementioned, one of the objectives of ARTchives is to adopt quantitative methods to answer art history and historiographical research questions. Being the crowdsourcing phase still in early stage, reliable large-scale analyses cannot be performed yet. However, a number of exploratory data analyses (EDA) performed over ARTchives actively contribute to refine project requirements in terms of data completeness, interlinking, and bias.[11] In particular, we investigated historians' networks and types of relations that are relevant in the Art History community. Through data visualization techniques we were able to show well-known geographical and relational patterns, such as as historians' communities based on provenance and places of activities, highlighting for instance Italian and German clusters. Less obvious patterns include institutional networks, highlighted by the correlation of their relevance in historians' biographies.

Results of the analysis drew our attention on some recurrent patterns, such as the closeness of art historians' due to shared institutions and research topics, and the relevance of art historians' documents in other historians' archival collections based on the aforementioned closeness. While few obvious patterns are immediately recognizable, the lack of extensive data and the incompleteness of some records prevent us from identifying other known relations and, possibly, opening up new research paths. We believe this aspect should be further investigated, since the lack of knowledge may turn into an opportunity. In particular, we envisage the definition of inference rules based on heuristics (recurrent patterns) to associate similar collections and supervised classification methods to

[10] https://www.openstreetmap.org/

[11] https://mybinder.org/v2/gh/LuciaGiagnolini12/Tesi/main

predict relations between art historians, institutions and contents of the collections. In so doing we aim at unveiling patterns that can be generalised as peculiar of the Art History domain, improve ARTchives data completeness, and further develop methods to support experts in retrieving archival collections relevant to their studies.

Lastly, it is worth noting a few experiments leveraging both ARTchives and Wikidata have been performed by independent scholars to address biases in the scope of Art Historical Linked Open Data. A notable example is the project Martrioska[12] which highlights the gender bias in art history, how this affects the completeness of data aggregators, and how this gap can be filled with computational methods.

5 Conclusion

In this paper we presented the data management system of ARTchives, an ongoing crowdsourcing project to aggregate curated information on art historians' personal archives. Both specific and generic project requirements stimulated the development of a Linked Open Data native cataloguing system that could effectively support consistent, accurate cataloguing and editorial processes. Future works include the alignment of terms to RIC Ontology [3] to allow archives reusing ARTchives data seamlessly.

ARTchives data management system fully embraces the Linked Open Data paradigm, fostering data reuse and efficient cataloguing, and ensuring data quality and consistency across information systems. Future developments include extension of the code base to support small-medium projects in producing 5 stars data that leverage user-friendly repositories (e.g. github) instead of or along with a triplestore for data storage and update.

Lastly, preliminary results of the EDA require us to further investigate the well-known issue of incompleteness of crowdsourced data. Lack of complete data may turn into an opportunity to develop computational methods tailored on the domain at hand for data enrichment and recommendation. Future works will address heuristics for archival collections interlinking and recommendation.

6 Acknowledgements

This work is partially supported by a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004746 (Polifonia: a digital harmoniser for musical heritage knowledge, H2020-SC6-TRANSFORMATIONS).

[12] https://martrioska.github.io/

References

1. Adamou, A., Brown, S., Barlow, H., Allocca, C., d'Aquin, M.: Crowdsourcing linked data on listening experiences through reuse and enhancement of library data. International Journal on Digital Libraries **20**(1), 61–79 (2019)
2. Carroll, J.J., Bizer, C., Hayes, P., Stickler, P.: Named graphs. Journal of Web Semantics **3**(4), 247–267 (2005)
3. Clavaud, F., EGAD, I.: International council on archives records in contexts ontology (ica ric-o) version 0.1 (2019)
4. Daquino, M., Daga, E., d'Aquin, M., Gangemi, A., Holland, S., Laney, R., Penuela, A.M., Mulholland, P.: Characterizing the landscape of musical data on the web: State of the art and challenges (2017)
5. Davis, E., Heravi, B.: Linked data and cultural heritage: A systematic review of participation, collaboration, and motivation. Journal on Computing and Cultural Heritage (JOCCH) **14**(2), 1–18 (2021)
6. Davis, K.: Old metadata in a new world: Standardizing the getty provenance index for linked data. Art Libraries Journal **44**(4), 162–166 (2019)
7. Deliot, C.: Publishing the british national bibliography as linked open data. Catalogue & Index **174**, 13–18 (2014)
8. Delmas-Glass, E., Sanderson, R.: Fostering a community of pharos scholars through the adoption of open standards. Art Libraries Journal **45**(1), 19–23 (2020)
9. Dijkshoorn, C., De Boer, V., Aroyo, L., Schreiber, G.: Accurator: Nichesourcing for cultural heritage. arXiv preprint arXiv:1709.09249 (2017)
10. Dijkshoorn, C., Jongma, L., Aroyo, L., Van Ossenbruggen, J., Schreiber, G., Ter Weele, W., Wielemaker, J.: The rijksmuseum collection as linked data. Semantic Web **9**(2), 221–230 (2018)
11. Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C., Van de Sompel, H.: The europeana data model (edm). In: World Library and Information Congress: 76th IFLA general conference and assembly. vol. 10, p. 15 (2010)
12. Knoblock, C.A., Szekely, P., Fink, E., Degler, D., Newbury, D., Sanderson, R., Blanch, K., Snyder, S., Chheda, N., Jain, N., et al.: Lessons learned in building linked data for the american art collaborative. In: International Semantic Web Conference. pp. 263–279. Springer (2017)
13. Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., Zhao, J.: Prov-o: The prov ontology (2013)
14. Malmsten, M.: Exposing library data as linked data. IFLA satellite preconference sponsored by the Information Technology Section" Emerging trends in (2009)