International Journal on Digital Libraries (2021) 22:251–255
https://doi.org/10.1007/s00799-021-00309-8

FAIR Data and Cultural Heritage Special Issue Editorial Note

Sorin Hermon[1] · Franco Niccolucci[2]

Published online: 15 October 2021

1 Introduction

Topics from the papers gathered in this special issue were presented at the CIDOC CRM conference held in Heraklion, Crete, 2018, during a special session entitled "Heritage data-centric research: are FAIR data fair enough?", chaired by Franco Niccolucci. The session focused on the steps needed to be taken in order to align archaeology, and in more general terms, Cultural Heritage, to modern e-Science requirements and transform the discipline into a, collaborative, computationally intensive data-driven one. A number of initiatives are addressing how to manage and use data produced by heritage research, most notably the ARIADNE[1] one in the archaeological domain, presently involving the most important research centres from all European countries in creating a comprehensive and integrated archaeological data infrastructure that so far has already registered little less than 2.000.000 archaeological datasets. Such infrastructure, implemented by ARIADNE, is bringing archaeology out of the "long tail of science", i.e. those disciplines that make little use of data-centric research. It is revolutionising the concept of Big Data: not relatively few datasets, each with terabytes of numbers, as in nuclear physics; but millions of small datasets, all potentially relevant to a specific research question but including a large (and unknown) majority probably irrelevant at all.

E-Science relies on the well-known FAIR principles,[2] stating that data should be Findable, Accessible, Interoperable and Re-usable. Now, if "F", "A" and "I" mainly depend on the technical way in which data and metadata are generated, stored, managed and curated, the "R" has less technical (but not less important) implications. It involves theoretical, methodological and epistemological aspects that have not received enough attention in the current debate. It has been argued that e-science discovery could be modelled as a deterministic discovery process; nevertheless, even in this perspective, simply modelling the provenance of data is not sufficient, but the provenance of the hypotheses and results generated from analyzing the data need to be modelled as well. Thus, to reuse data in cultural heritage it is necessary to expand the "R" facet of the FAIR principles at least into R3: Re-usable, Relevant and Reliable. Judging relevance and reliability may appear obvious to a human eye, but it is not to machine processing. Data reliability depends on a chain of trust that needs to be adequately supported by documentation, and on this regard the CIDOC CRM may play a key role. If in the past reference to previous discoveries published in journals and books was based on the academic practice of peer-review and on the authoritativeness of the author and of the publication, re-using data created by others is still lacking a similar good practice.

The session discussed such aspects and proposed ways to address the issue. Contributions came from purely cultural heritage practice ("What would you need to rely on somebody else's data?") to semantics ("What would you suggest to document, in order to support reliability?"). Both aspects will be analysed in light of the CRM: does it already provide a sufficiently rich toolbox, or additions are required? If so, which ones? Following is a short description of the session's presented talks which were published elsewhere, while those published in this volume will be described further below in a separate section of this editorial.

2 Presented talks

Panos Constantopoulos, from the Athens School of Economics and Business, Greecce, presented the Scholarly Ontology, in a talk entitled "Ontology-based research process documentation as a reusability enabler". The Scholarly Ontology (SO) is an ontology for modelling research processes derived from CIDOC CRM. It has evolved as a

✉ Sorin Hermon
sorin.hermon@gmail.com

Franco Niccolucci
franco.niccolucci@gmail.com

[1] STARC, The Cyprus Institute, Nicosia, Cyprus

[2] VAST-LAB, PIN, Prato, Italy

[1] https://www.ariadne-infrastructure.eu/.

[2] https://www.force11.org/fairprinciples/.

generalization of the NeDiMAH Methods Ontology (NeMO) and enjoys extensive empirical grounding. Due to its cross-disciplinary character, the SO enables documenting and analyzing research processes unfolding in one or more domains, and, correspondingly, associating data from disparate, domain-specific sources. The research process is addressed from four complementary perspectives: activity, procedure, resource and agency. We view the contextualized, structured, process-oriented documentation of scholarly work using SO as an enabler of the reusability of cultural heritage data.

Martin Doerr, from ICS-FORTH, Greece presented a new proposal for an extension of CIDOC-CRM, aimed at describing inferences, in a talked entitled "CRMInf: Supporting Facts by Arguments". The presentation is based on the observation that in the current practice of documenting cultural heritage, maintainers of databases mostly present facts as their best knowledge, adding citations but without analyzing the reasons why a particular fact is believed or not. Archaeological records may contain more detailed justifications, but only in limited cases related to individual facts. On the other side, computer scientists have developed advanced argumentation systems, but more to support an expert dialogue than to justify and maintain the validity of facts in documentation systems. CRMInf is a CIDOC CRM—compatible extension designed for the latter. It contains a basic model of ways to acquire new knowledge, and it is being specialized for supporting more directly the discourse with historical sources and with scientific observations.

Sorin Hermon, from The Cyprus Institute, Cyprus presented a talk with the subject "How FAIR are the FAIR principles for archaeological data?", focusing on the added value of making archaeological data FAIR, in particular primary data collected during fieldwork, such as 3D models of excavation units, analytical measurements and geodesic data. The main argument of the discussion was that without a formal representation of data provenance, such data can be FAIR but of little use in the archaeological research.

Joseph Padfield, from The National Gallery, London, UK presented "Putting theory into practice—Using a CIDOC based venue ontology to describe the movement of paintings within the National Gallery", where he detailed an innovative use of CIDOC CRM in the National Gallery with the Venue ontology, which allows considering how it is practically used, developing a simple, internal PID system and its incorporation within a practical tool for capturing and recording the movement of paintings, thus documenting their provenance and relationship with the parts of the gallery and the whole.

Christian-Emile Smith Ore, from the University of Oslo, Norway, presented "CIDOC-CRM as semantic glue for excavation data sets, site and monument registries and museum collections". In the talk, he described the situation in archeology, where a major issue of the last 10-15 years has been to rescue, preserve and give access to the data sets from archeological excavations. Following the implementation of the EU infrastructure project ARIADNE, a major driving force in this effort, a very large number of archaeological datasets are now accessible. One of the main challenges that remain open is how to apply the FAIR principles to these archeological datasets. Another open challenge is that often there are only weak links between the excavation and the data sets on the one hand and the museum collections (find repositories), site and monument registries and publications on the other. To strengthen the FAIR-ness of the datasets such links have to be strengthened. In Norway a new infrastructure project, ADED (Archaeological Digital Excavation Documentation) was launched in 2018 with the objective to create a repository for data sets and establish the aforementioned links. In this infrastructure, the CIDOC-CRM suite will be applied as semantic glue.

3 Articles in this special issue

The above presentations described either innovative applications of CIDOC-CRM within the Cultural Heritage domain or proposed new extensions to it. Moreover, they all highlighted the need for concrete steps for implementing the FAIR principles in the Cultural Heritage practice. Following is a description of further presentation which developed into papers published within this special issue.

Olivier Marlet, of University of Tours, France, presented "Logicist writing for reliability of data-centric research in archaeology", describing some of the activities conducted by the consortium MASA (Memory of Archaeologists and Archaeological Site) from the very large facility Huma-Hum, the Laboratoire Archéologie et Territoires (University of Tours/CNRS, France), in collaboration with the MRSH (University of Caen/CNRS, France). These aimed at setting up a logicist writing publication for the results of the excavation of the Rigny cemetery. Elisabeth Zadora-Rio, the archaeologist involved in the process, formalized her archaeological reasoning according to the precepts of Jean-Claude Gardin, proposing a clear structuring of the logic of inferences allowing to navigate from field observations to the most synthetic interpretations. The web application developed allows to read the publication in a synthetic way or to deepen the reading by going as far as excavation data, information directly linked to ArSol, the online database. The related paper, entitled "A way to express the reliability of archaeological data: data traceability at the Laboratoire Archéologie et Territoires (Tours, France) and co-authored by Olivier Marlet and Xavier Rodier details the good practices in archaeology disseminated by the MASA Consortium (Archaeologists and Archaeological Sites Memories) and the Laboratoire Archéologie et Territoires (Tours, France), with a focus on the evaluation of the

progress of ArSol database (Soil Archives) and its field data management database, with regard to the FAIR principles and the 5 Stars Linked Open Data. The work undertaken to achieve compliance with these precepts demonstrates that it is necessary to ensure the relevance and reliability of the published data as well. A pre-requisite for data to be reusable is to ensure its provenance. Various tools set up in the ArSol database make this possible, tracing data from the field recording, through its interpretation to the publication of excavation results and thus satisfactorily complying with the FAIR data principles requirements.

The second paper in this volume, with the title "FAIR data for prehistoric mining archaeology" and co-authored by Gerald Hiebel, Gert Goldenberg, Caroline Grutsch, Klaus Hanke and Markus Staudt, from the University of Innsbruck, Austria, presents an approach on how to create FAIR data for prehistoric mining archaeology, based on the CIDOC CRM ontology and semantic web standards. The interdisciplinary Research Centre HiMAT (History of mining activities in the Tyrol and adjacent areas, University of Innsbruck) investigates mining history from prehistoric to modern times with an interdisciplinary approach. One of its related activities is the multinational DACH project "Prehistoric copper production in the eastern and central Alps". Within this framework, data from a specific geographical region of the project was selected to transform it into an open and reusable data, according to the FAIR data principles, as part of an Open Research Data pilot project. Every archaeological investigation in Austria has to be documented according to the requirements of the Austrian Federal Monuments Office. This documentation is deposited in the CERN-based EU supported research data repository ZENODO. For each deposited file, metadata are created through the application of the conceptual metadata schema CIDOC CRM. Concepts specific to mining archaeology research are organized with the DARIAH Back Bone Thesaurus, a model for sustainable interoperable thesauri maintenance, developed in the European Union Digital Research Infrastructure for the Arts and Humanities. Metadata are created through the extraction of information from the documentation and the transformation to a knowledge graph using semantic web standards. To facilitate usage, graph data are exported to hierarchical and tabular formats representing sites and objects with their geographic locations, temporal and typological assignments and links to the research activities and documents. Metadata are deposited together with the documentation into the repository.

Achille Felicetti, from PIN, Prato, Italy, presented a talk entitled "Heritage Science and Cultural Heritage: a CIDOC CRM-enabled Model for Integration and Interoperability". The main goal of the presented model is to collect provenance data of scientific datasets resulting from Heritage Science research, and to document it in a standard and accessible way. The approach, inheriting and adapting common logics and concepts of existing models and taking inspiration from the semantic principles of CIDOC CRM, proposes a schema composed of reusable XML modules, intended to describe Heritage Science entities (including actors, devices, datasets, analysis and other events) in detail, and dynamically organised in a common framework by means of a set of internal links based on persistent identifiers. Such a structure implements a platform-independent meta-format able to express the essence of the data while remaining unbound to any specific system or software, and supports the necessary confidence in somebody else's data for re-use. The related paper, bearing the title "Heritage Science and Cultural Heritage: standards and tools for establishing cross-domain data interoperability" and co-authored by Lisa Castelli (INFN, Italy), Achille Felicetti & Fabio Proietti (INFN, Italy), describes the system for documenting scientific data produced in Heritage Sciences, presented at the aforementioned conference. The system is built around a general meta-model, flexible enough to provide descriptions, in a formal language, of the datasets produced by scientific research. Resulting metadata can be re-encoded and published in multiple formats. The underlying metadata schema is inspired by CIDOC CRM principles for data modelling and maintains a full compatibility with CIDOC CRM ontology to capture provenance and foster interoperability with Cultural Heritage information. The use of a wide set of thesauri and controlled vocabularies guarantees internal coherence at data and metadata level. A set of user interfaces has been designed to simplify and speed up the process of data gathering and metadata definition.

The following paper of this volume, entitled "A fuzzy approach to evaluate the attributions reliability in the archaeological sources" and authored by Marianna Figuera, from the University of Catania, Italy, presents a case study of data management and processing of archaeological information through a relational database. The unusual typology of the 'small finds' that were archaeologically analyzed and the specific history of the excavations at Phaistos and Ayia Triada (Crete, Greece) prompted our consideration of issues regarding data integrity. We sought to address the problem surrounding the relevance of archaeological sources by applying a reliability index to the subjective interpretations of archaeological data, which ultimately led to the implementation of a fuzzy method to determine the degree of uncertainty of attributions associated with function. The resulting database represents a 'container of memories' that allows the processing of all the typological and functional attributions from any source, without having to necessarily simplify or dilute the information in order to render it manageable. The concept of 'probability of belonging' and multi-assignment of source attributions seem to represent plausible methodological pathways to determining the reliability of archaeological data, thus warranting the research

presented herein. The paper was preceded by a presentation at the conference, bearing the same title as the paper. It presented the problem of the relevance of archaeological sources, addressed from a different perspective and considering the reliability concept linked to the subjectivity intrinsic to the archaeological data. The case study relates to small finds excavated at the archaeological sites of Phaistos and Ayia Triada (Crete). These unusual finds analyzed, and the specific history of excavations of the two sites led to the realization of a procedure in which a fuzzy approach has been used to preserve the degree of uncertainty of the functional attributions. The concept of "probability of belonging" and the management through multi-assignment of the sources' attributions could suggest a possible methodological approach to the validation of the relevance and reliability of the archaeological data.

The next paper of this volume, entitled "Towards an ontological cross-disciplinary solution for multidisciplinary data: VI-SEEM data management and the FAIR principles" is co-authored by Valentina Vassallo (The Cyprus Institute, Cyprus) and Achille Felicetti (PIN, Prato, Italy). It starts from the observation that different scientific communities produce different kinds of datasets that rely on different data descriptions, approaches, and logical organisations. In such an environment, it is essential to establish a knowledge communication framework that can guarantee some fundamentals, such as an inclusive description and documentation of the interdisciplinary digital resources, their long-term preservation, access, use, and reuse. The establishment of semantic knowledge integration aims at overcoming such inhomogeneity between data produced by different research communities. Specifically, we refer to those communities aggregated within the e-Infrastructure developed by the European project VI-SEEM: Life Science, Climate Science, and Digital Cultural Heritage. The current research proposes a framework based on CIDOC CRM and its extensions, in particular the CRMsci and CRMdig, and tested on examples identified as interdisciplinary respect to the different and various research areas of the project. Moreover, the semantic solution aims at fulfilling the FAIR principles.

The final paper in this volume, entitled "The R4 to Identify Born and Digitized Cultural Heritage: Re-usable, Relevant, Reliable and Resistant" and authored by Nicola Barbuti, Bari University, Italy, proceeds a presentation with the same title the author delivered at the special session of the 2018 CIDOC-CRM conference. The talk focused on the need to correctly identify what, and how much of the digital resources produced up to day can be identified as "born digital and digitized cultural heritage". According to the author, this process needs clear and homogeneous identity criteria, according to which one can distinguish digital cultural entities from the daily magmatic production of data. As the FAIR Principles alone do not seem to be sufficient for this purpose, the author proposes that the FAIR R should be quadrupled in R4: Re-usable, Relevant, Reliable and Resistant. These requirements will give the digital data the value of Cultural Heritage, as they are perfectly specular to the definition we can give of what we commonly consider tangible and intangible cultural heritage. The consequent paper, signed by the same author and bearing the title "Thinking digital libraries for preservation as digital cultural heritage: by R to R4 facet of FAIR principles" advises to rethink digital and digitization as social and cultural expressions of the contemporary age, in light of Art. 2 of the UE Council conclusions of 21 May 2014 on cultural heritage as a strategic resource for a sustainable Europe (2014/C 183/08), which states: "Cultural heritage consists of the resources inherited from the past in all forms and aspects–tangible, intangible and digital (born digital and digitized), including monuments, sites, landscapes, skills, practices, knowledge and expressions of human creativity, as well as collections conserved and managed by public and private bodies such as museums, libraries and archives". Consequently, the author suggests to rethink digital libraries produced by digitization as cultural entities and no longer as mere dataset for enhancing fruition of cultural heritage, by defining clear and homogeneous criteria to validate and certify them as memory and sources of knowledge for future generations. By expanding R: Re-usable of the FAIR Guiding Principles for scientific data management and stewardship into R4: Re-usable, Relevant, Reliable and Resilient, the author proposes a more reflective approach to creation of descriptive metadata for managing digital resource of cultural heritage, which can guarantee their long-term preservation.

4 Conclusion

To sum up, the conference session and the following papers published in this special issues focused on how to digitally transform cultural heritage into a data-driven discipline and fully implement the FAIR data principles in its practice. The AriadnePlus project, one of the major EU funded initiatives, is currently implementing its repository based on these principles. A need to further develop the R component of the FAIR principles, through theoretical, methodological and epistemological research was proposed by (Niccolucci. Consequently, formally expressing data quality, through modelling its provenance and modelling the hypotheses and results generated from analyzing the data was identified as a pre-requisite of this process of data FAIR-ification (Hermon), while Figuera suggests to adopt a fuzzy sets theory approach to express ambiguity and uncertainty of data. In order to model the research reasoning process in Cultural Heritage, Constantopoulos suggests to use Scholarly Ontology, Doerr proposes to define the CRMinf model, an

extension of CIDOC-CRM aimed at modelling archaeological inferences, while Marlet recommends to adopt a logicist approach (following the works of Jean Claude Gardin) to enhance the reliability of the published archaeological data. Finally, Barbuti suggests to extend the R data principle in order to more correctly capture the essence of Cultural Heritage and its digital components.

Several case-studies were presented as well, where CIDOC-CRM was applied in order to monitor movement of paintings within the various departments of the national gallery in London, UK (Padfield), integration of excavation data sets with monuments registries and museum collections (Ore) and the process of FAIR-ification of archaeological data (Hiebel et al). Finally, Felicetti et al. present a model and related system for representing Heritage Science data using the CRM, in order to achieve the needed interoperability of such data, while Vassallo and Felicetti propose to apply ontological solutions for integration of cross-disciplines data, while taking into consideration the FAIR principles as well.