An approach to extend the metadata schema of Zenodo for Cultural Heritage datasets

Alberto Bucciero[‡], Emanuel Demetrescu[§], Bruno Fanini[§], Alessandra Chirivi[‡], Francesco Taurino[‡]

‡ CNR, Lecce, Italy
§ CNR, Rome, Italy

Corresponding author: Alberto Bucciero (alberto.bucciero@cnr.it)

Academic editor: Editorial Secretary

Abstract

In this article, we present an approach designed to extend the metadata schema of the Zenodo data management platform to strengthen the FAIRness of the published dataset. We focus on a bottom-up approach starting from a series of datasets ranging from the 3D digitalisation of monuments and sites to the creation of reconstructive records (including the scientific documentation they are based on), to the implementation of digital storytelling and to the development of open source-based web-apps. We propose the simplest possible set of metadata to be included in the Zenodo platform with the possibility, for the community, to adopt and further develop/modify them. This article will describe in detail the formalisation and the digital formats adopted providing the related metadata templates developed within the projects.

Keywords

metadata, data FAIRness, FAIR, cultural heritage, data management, heritage science

Overview and background

The ability to publish, share and ensure long-lasting research results is one of the main goals of the policy known as "Data FAIRness"(Wilkinson et al. 2016). At the same time, however, not much emphasis is placed on the data originally acquired, transformed and post-processed in the various stages of the work that led to the research result (Jacobsen et al. 2020). Often, these data are stored in the most disparate devices —almost always in a way that does not guarantee their security— and they are not commented on and lack meaningful structuring.

This situation is particularly evident in the field of Heritage Science (HS). Databases regarding the same cultural context are "stratified" —not integrated— over the years and

formalised in heterogeneous ways, relying on non-standardised data ingestion and preservation strategies, based mainly on customised and non-generalisable data modelling solutions (metadata schema redundancies). Only in recent years has an increased focus on interoperability and the adoption of semantic knowledge networks laid the groundwork for standardisation that respects the multifaceted reality of the cultural record.

Objectives

OpenAire's effort has led to the establishment of Zenodo as a valuable platform for the management and long-lasting preservation of research data, including scientific datasets. Despite its ambition, the extreme descriptive poverty of the meta-information that can be instantiated does not allow for an optimal description of either such datasets. On the other hand, the scientific research process requires numerous steps, each characterised by specific methodologies, tools and data that are processed and that produce other output data in each single step of the sub-process. To keep track of the entire scientific process and, ultimately, to make this process FAIR, it is, therefore, essential to be able to fully describe the datasets used.

These considerations led us to identify a technical-scientific method that allows us to use Zenodo to extend its meta-descriptive capabilities, to store an additional set of information in the HS domain that can more efficiently identify and make searchable search results and all related datasets.

Implementation

One of the main limitations recognised in the Zenodo platform is the lack of ability to decorate data with appropriate metadata qualifying the specific application domain. Essentially, Zenodo almost exclusively supports a basic metadata scheme derived from Dublin Core (Weibel et al. 1999). Although this scheme is generic enough to adapt to multiple cases, it is insufficient when you need to describe datasets in more detail to ensure more effective searchability, accessibility, interoperability and reuse.

In addition, the lack of support for extension towards customised metadata schemes is one of the main critical issues of Zenodo, especially in specific fields of applications, such as Cultural Heritage.

The first draft of a possible solution has been released on Zenodo (Bucciero and Demetrescu 2022) that consists of two different parts:

- an operational methodology that guides the user step by step in publishing datasets, specifying how to add a set of domain metadata compared to the basic ones provided by default by the platform;
- a set of metadata specific to the Heritage Science domain that can fully qualify the dataset by declaring some salient features.

This contribution will add more theoretical and practical context describing new approaches to the publication of semantic-based datasets and to the publication of structured data, as well as the software and algorithms used to modify and transform the data flow from the sources to the final scientific dataset.

Finally, we will present the design of an architectural model (Fig. 1) that, by using a specialised crawler, will be able to extend the searching capabilities of the Zenodo platform, allowing the findability of the datasets, based on the newly-added metadata.

We propose to enrich the basic descriptive capability of Zenodo (based just on Dublin Core) by adding a "sister file" containing all the metadata needed to fully qualify the dataset. In this way, a search query can be executed by:

- searching into the native metadata provided directly by Zenodo (blue route in Fig. 1);
- searching into the extended metadata description by calling the Zenodo API and obtaining the "sister file" corresponding to that specific dataset (red route in Fig. 1).

Conflicts of interest

References

- Bucciero A, Demetrescu E (2022) Manuale operativo di metadatazione dei dataset per Zenodo nei Beni Culturali. https://doi.org/10.5281/zenodo.6138586
- Jacobsen A, de Miranda Azevedo R, Juty N, Batista D, Coles S, Cornet R, Courtot M, Crosas M, Dumontier M, Evelo C, Goble C, Guizzardi G, Hansen KK, Hasnain A, Hettne K, Heringa J, Hooft RW, Imming M, Jeffery K, Kaliyaperumal R, Kersloot M, Kirkpatrick C, Kuhn T, Labastida I, Magagna B, McQuilton P, Meyers N, Montesanti A, van Reisen M, Rocca-Serra P, Pergl R, Sansone S, da Silva Santos LOB, Schneider J, Strawn G, Thompson M, Waagmeester A, Weigel T, Wilkinson M, Willighagen E, Wittenburg P, Roos M, Mons B, Schultes E (2020) FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence 2: 10-29. https://doi.org/10.1162/dint_r_00024
- Weibel S, Godby J, Miller E, Daniel R (1999) Dublin Core Metadata Standard. http://www.oclc.org:5046/conferences/metadata/dublin_core_report.html
- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, 't Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18

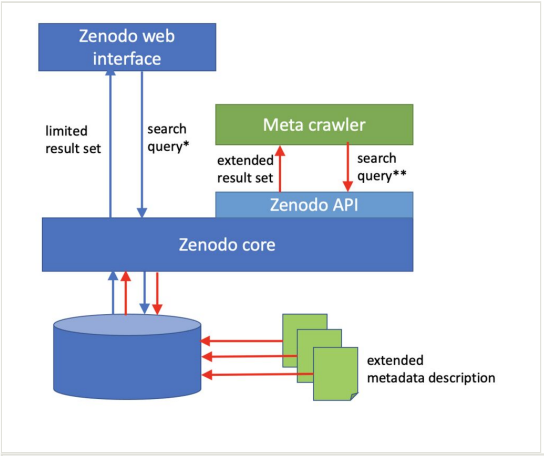

Figure 1.

Architectural model extending Zenodo's search capabilities.