#12 - Facilitating Data Citation:

An app to import metadata in a reference management software

Christina Bornatici (FORS), Stefan Buerli (FORS),
Yassine Gafaiti (FORS), Roberto Genesi (FORS)

This idea comes from the fact that research data are often not correctly cited in scientific publications. Researchers mention and describe the data they use in the method section of their publication, but omit to cite the data in the bibliographic section. This recurring observation was made both in publications using primary and secondary data, despite the fact that, for archived data, repositories suggest a citation and attribute a Persistent Identifier (e.g., a DOI) to each dataset. This incorrect citation causes several problems. First, data authors are not properly credited for their work, which could lead in reduced incentive to share new datasets. Also, this limits datasets' usage tracking and the development of Open Research Data metrics, like reusability scores as hacked by group #1.

The goal is to facilitate and improve data citation practice by making the data citation process identical to the citation process of other sources (e.g., research articles, book chapters). The idea is to build an application to import in a reference management software (e.g., Zotero, EndNote, Mendley) all the metadata needed to build the data citation in different bibliographic standards. Repositories could then include this application as a plugin for their catalogues.

The first objective was to define a minimum viable product. We analysed which file formats could be used to import the metadata in reference management software and which metadata where relevant to cite datasets. We decided to use the RIS format (Research Information System Document) which is compatible with commonly used reference management software that are Zotero, Mendley and Endnote (Wikipedia 2021). Also, we selected 8 mandatory and 6 extra metadata fields to be imported, based on a prior analysis conducted by the FORS data archive team of data citation recommendations at different institutions (e.g., data archives and repositories, university libraries, scientific journals and specialised organisations) in Switzerland and abroad. The next step was to find the codes used in the RIS format for each selected metadata field (see Table 1). The extra metadata fields would be included to the application if there is enough time.

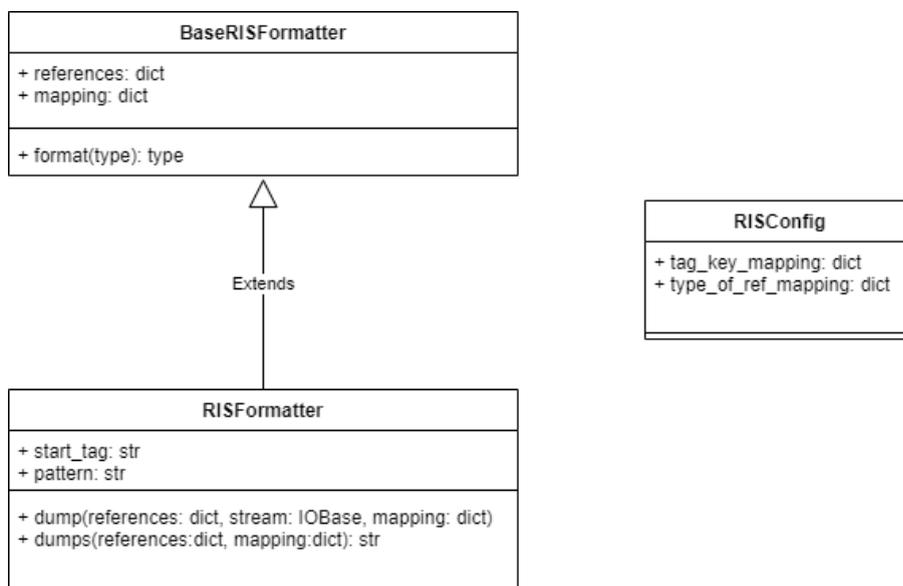Table 1: Mandatory and extra metadata fields and their relative RIS codes

Mandatory fields	RIS codes & examples	Extra fields	RIS codes & examples
Reference/Item type	TY - Dataset*	Dataset number	NV - Number
Authors/Creators	AU - Surname, Name	URL	UR - URL
Publication year	PY - Year	Place published	CY - Location
Title	TI - Title	Access date	Y2 - Date
Data type	C4 - Dataset*	Abstract	AB - Abstract
Publisher/Distributor	PB - Publisher	Language	LA - Language
Version	ET - Version		
DOI	DO - DOI number		

* The information is always "Dataset" for these fields.

We decided to use FORS data archive's catalogue FORSbase (https://forsbase.unil.ch/) as our test case. Given that the whole team knows well FORSbase, having developed the platform and/or working with it, the analyses of the available metadata was rapidly done.

Technically, we developed a model as shown in Figure 1. First, the BaseRISFormatter class contains the RIS format which has a a {tag} - {value} pattern. In our specific case this starts with a reference type, and is extended with other tags as illustrated in the RISFormatter class. In order to match the metadata extracted from the existing repository, in our case FORSbase, with the tags injected into the RIS file, we defined a configuration class (RISConfig), in which all the constants necessary for the mapping are defined. This package is included as a module. To use it, we retrieve all needed metadata from FORSbase. Then we pass it to the dump function in the RISFormatter class and we create the RIS file (export) to be downloaded by the final user.

Figure 1: Unified Modeling Language (UML) diagram

The group succeeded in producing an application that works for FORSbase. This was developed within a test version. When researchers are on FORSbase catalogue and select a dataset, they would have the possibility to download a RIS file containing all the mandatory and the extra metadata fields clicking on the new button "Export citation" (Figure 2). This RIS file lands in the "Downloads" folder and can be automatically uploaded in both reference management software we tested, that are Zotero and EndNote. All the imported metadata lands in the correct Zotero and Endnote metadata fields. It should be noted that, contrary to EndNote, Zotero does not have a specific reference/item type for datasets. In our tests, the item type used by Zotero was "Journal article". We do not know why Zotero used this item type rather than another. When specific fields are not available, Zotero lists them in its "Extra" field. The fields related to the version, dataset number, place and publisher are thus inserted there (Figure 3).

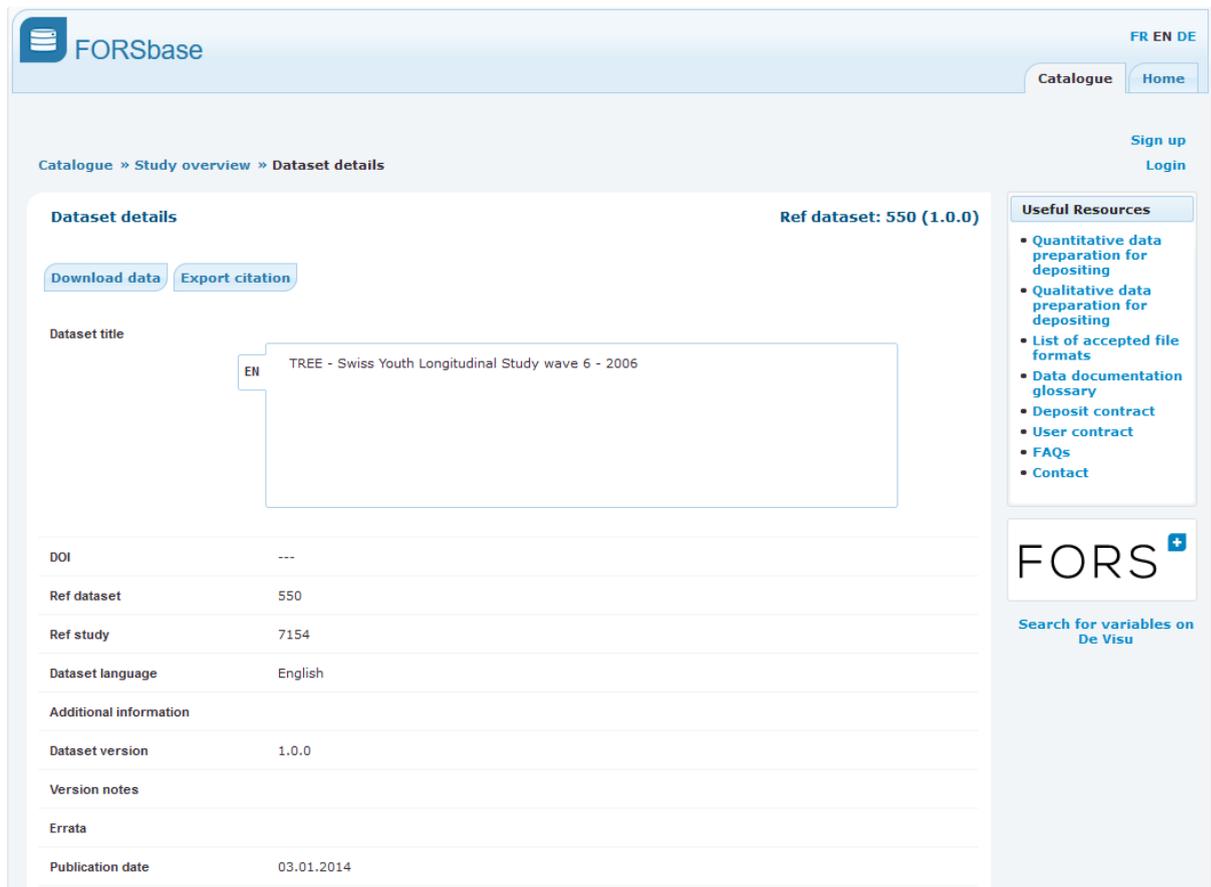Figure 2: New "Export citation" button in FORSbase (test version)

Figure 3: Printscreen of Zotero "Extra" field once the RIS file is imported

As the reference management software can be linked with a text processing software, the researchers can easily cite the dataset in their texts, as they would do for other sources. Depending on the bibliographic style in use, this creates an automatic reference in the bibliographic section. Bibliographic styles can be very easily and quickly changed for all the citations included this way. We were successful when testing our solution with Endnote and Word (no other tests have been done).

To sum up, the created application allows the export from an existing system – FORSbase in our case – the mandatory and extra metadata fields used for data citation in a RIS file format. This RIS file is then correctly imported in a reference management software. We can say that with this application, researchers can save metadata and cite data in their

texts as they would do for any other sources. Thus, the mission is accomplished concerning this minimum viable product. The resulting application could be adapted to be used in other data repositories. Next steps would be to extend the application to other formats than RIS and, as far as we are concerned, to reuse this module to SWISSUbase (https://info.swissubase.ch/fr/).

References

Wikipedia (2021). Comparison of reference management software: Import file formats. Accessed on 26.01.2021:
https://en.wikipedia.org/wiki/Comparison_of_reference_management_software.