METADATA FOR RESEARCH DATA IN PRACTICE

by Barbara Petritsch

Abstract: *What data is needed about data? Describing the process to answer this question for the institutional data repository IST DataRep.*

Keywords: *Data repository; data publication; Eprints; DataCite; metadata*

METADATEN FÜR FORSCHUNGSDATEN IN DER PRAXIS

Zusammenfassung: *Welche Daten über Daten brauchen wir? Die Beschreibung des Prozesses diese Frage für das institutionelle Datenrepositourium IST DataRep zu beantworten.*

Schlüsselwörter: *Datenrepositorium; Datenpublikation; Eprints; DataCite; Metadaten*

Contents

The *Institute of Science and Technology Austria (IST Austria)* is a young international institution dedicated to basic research and graduate education in the fields of life sciences, formal and physical sciences, located in Klosterneuburg on the outskirts of Vienna. Still growing, IST Austria is committed to conducting world-class research. By 2026, up to 90 research groups will perform research in an international state-of-the-art environment.

1. Introduction

In order to be prepared for the growing need of data publication, IST Austria decided to build a data repository in 2012. The library of IST Austria engaged in the tasks of researching the field and implementing the repository, which is in operation since August 2015. In this article I describe IST DataRep in regard to its metadata and how this was shaped by our approach.

For us it was crucial that the users of this tool – the researchers – were involved throughout the whole process of implementation. At first, a survey on the actual state of research at IST Austria was carried out to get a better idea of the requirements for and needs of our scientists. Later in the process, when we had already decided on the open source software Eprints, they took part in extensive user tests. The scope of these tests was to adapt the depositing workflow to our researchers' needs and wishes. Implicit to this process was the design of the metadata set.

2. IST DataRep project

In the course of the IST DataRep project various aspects determined the final metadata set:

2.1. Data Publication

The top priority for the institutional data repository was to provide the opportunity for affiliated scientists of all disciplines to publish their data (underlying publications) especially if they lack a suitable subject repository. One key feature of academic publishing is that it renders scientific content citable. As advancement on URLs – which tend to be inconsistent and unreliable – the concept of persistent identifiers was introduced for digital publishing as a permanent addressing mechanism. The most popular and best-known persistent identifier in academic publishing is the Digital

Object Identifier (DOI). Therefore the obvious choice for IST DataRep was to register with DataCite, which provides DOIs especially for research data and university publications (theses, technical reports etc.).

2.2. Data Format & Data Types

Due to the interdisciplinary scope of research conducted at IST Austria, the institutional survey showed that a high variety of data types and formats are created.[1] With this precondition we were facing the usual role of an institutional repository: to take care of the "long tail" of research[2]. This implies universal values for data description and the consideration of all kinds of data types and formats. With this set of requirements we were equipped to evaluate the available proprietary and open source products, which resulted in us choosing the open source software Eprints.[3]

2.3. User Tests

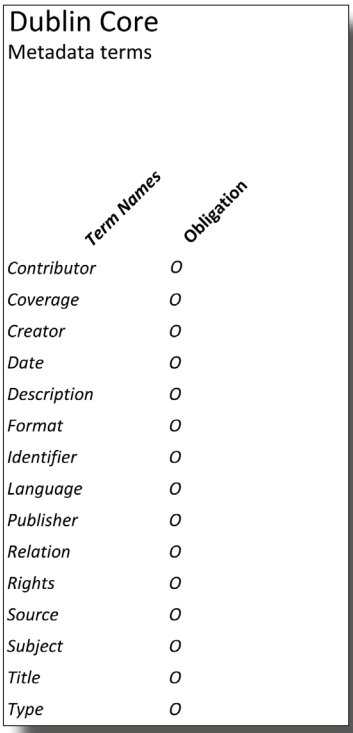
After implementing Eprints in combination with the plugin ReCollect[4] we ran extensive user tests with the barely surprising result that researchers wish to have an easy-to-use tool, which enables a quick upload. For the purpose of comparison they also tested Figshare. One distinctive feedback was that the test persons greatly appreciated metadata presets during upload. We therefore adopted this feature for IST DataRep.

2.4. Adaptation

Adapting Eprints was a process of considering the institutional concept on the one hand and complying with external factors on the other. One influential factor of metadata design was the metadata schema of DataCite[5] as a precondition for minting DOIs. It consists of mandatory, recommended and optional terms, which provides (us with) some orientation, flexibility and limits. It is based on Dublin Core[6] but with additional regulations and restrictions.

We customized our fields considering the institutional needs and our researchers' request for a convenient workflow. Therefore our initial effort was to limit the amount of fields required to fill in during deposit.[7]Because the researchers themselves, who most likely have no specific cataloging skills, carry out the data upload we focused on implementing a tool which is usable for them. This also implied to rethink field names to maximize clarity from the researcher's/user's perspective.

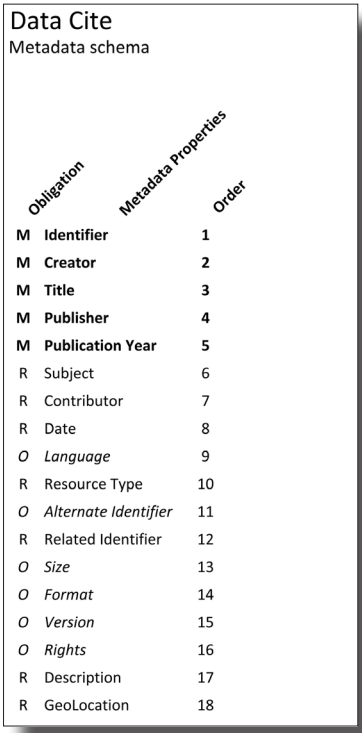
3. Dublin Core, DataCite Metadata Schema and IST DataRep metadata fields



Dublin Core
Metadata terms

Term Names	Obligation
Contributor	*O*
Coverage	*O*
Creator	*O*
Date	*O*
Description	*O*
Format	*O*
Identifier	*O*
Language	*O*
Publisher	*O*
Relation	*O*
Rights	*O*
Source	*O*
Subject	*O*
Title	*O*
Type	*O*

Fig. 1: Dublin Core Metadata terms



Data Cite
Metadata schema

Obligation	Metadata Properties	Order
M	**Identifier**	**1**
M	**Creator**	**2**
M	**Title**	**3**
M	**Publisher**	**4**
M	**Publication Year**	**5**
R	Subject	6
R	Contributor	7
R	Date	8
O	*Language*	9
R	Resource Type	10
O	*Alternate Identifier*	11
R	Related Identifier	12
O	*Size*	13
O	*Format*	14
O	*Version*	15
O	*Rights*	16
R	Description	17
R	GeoLocation	18

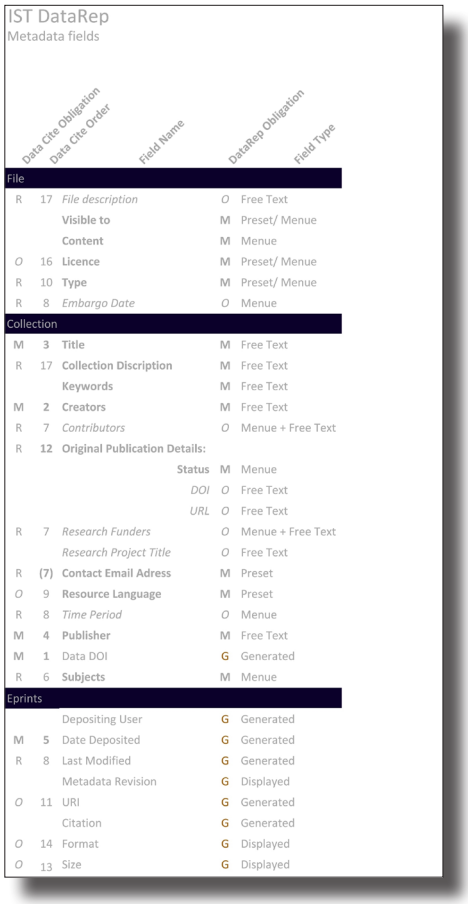
Fig. 2: DataCite Metadata schema

Tracing back the schema of the metadata fields used in IST DataRep, one ends up with Dublin Core. Dublin Core is a metadata schema developed by the DCMI (Dublin Core Metadata Initiative, founded in 1994) at the OCLC/NCSA Metadata Workshop in Dublin, Ohio in 1995. It is a simple set of 15 core elements for resource description intended to facilitate the discovery of digital objects. All of these elements are optional (*O*), repeatable, and they have no specific order (Fig. 1). It was designed specifically for non-catalogers.[8]

DataCite was founded as a consortium by leading research libraries and information centers in 2009.[9] Its aim is to allow easy online access to research data and improve its citability. DataCite's Metadata Schema is based on Dublin Core with more regulations and restrictions, which clearly indicates a librarian's approach. It has 18 elements (terms) set

in a specific order with three different levels of obligation: mandatory (**M**), recommended (R) and optional (*O*) (Fig. 2). The element types are definable via subfields, e.g. the term contributor has to be refined as researcher, supervisor, advisor etc.[10]

The metadata form of IST DataRep consists of 28 fields in total with different qualities: free text fields to fill in manually (10), option menus to select (5), presets (5) which are still modifiable, and generated elements (8) (see exemplary fields in Fig. 4). Most fields can be mapped to DataCite's elements (Fig. 3). One option for the mapping process is automatically via the DataCite DOI plugin for Eprints. Subsequently, this information is transferred via an API to DataCite for DOI registration. The alternative is to prepare an XML file manually and register a DOI via DataCite's online metadata shop.


IST DataRep
Metadata fields

Data Cite Obligation	Data Cite Order	Field Name	DataRep Obligation	Field Type
File				
R	17	*File description*	*O*	Free Text
		Visible to	M	Preset/ Menue
		Content	M	Menue
O	16	**Licence**	M	Preset/ Menue
R	10	**Type**	M	Preset/ Menue
R	8	*Embargo Date*	*O*	Menue
Collection				
M	3	**Title**	M	Free Text
R	17	**Collection Discription**	M	Free Text
		Keywords	M	Free Text
M	2	**Creators**	M	Free Text
R	7	*Contributors*	*O*	Menue + Free Text
R	12	**Original Publication Details:**		
		Status	M	Menue
		DOI	*O*	Free Text
		URL	*O*	Free Text
R	7	*Research Funders*	*O*	Menue + Free Text
		Research Project Title	*O*	Free Text
R	(7)	**Contact Email Adress**	M	Preset
O	9	**Resource Language**	M	Preset
R	8	*Time Period*	*O*	Menue
M	4	**Publisher**	M	Free Text
M	1	Data DOI	G	Generated
R	6	**Subjects**	M	Menue
Eprints				
		Depositing User	G	Generated
M	5	Date Deposited	G	Generated
R	8	Last Modified	G	Generated
		Metadata Revision	G	Displayed
O	11	URI	G	Generated
		Citation	G	Generated
O	14	Format	G	Displayed
O	13	Size	G	Displayed

Fig. 3: IST DataRep Metadata fields

4. Conclusion

For the library of IST Austria it was crucial to create a tool that is usable, understandable, practical and tailored to specific needs. One main approach to the project was to focus on only one challenge at a time. Our first priority was data publication. After the repository was available for data deposit/publication we were aiming at a long-term storage solution. At the moment we are participating in the H2020 EUDAT project with the objective to establish off-site long-term storage for the deposited data, which is also crucial to the warranty of DOIs: to facilitate permanent access. In 2017 we will focus on two projects: one of which is very hands-on,

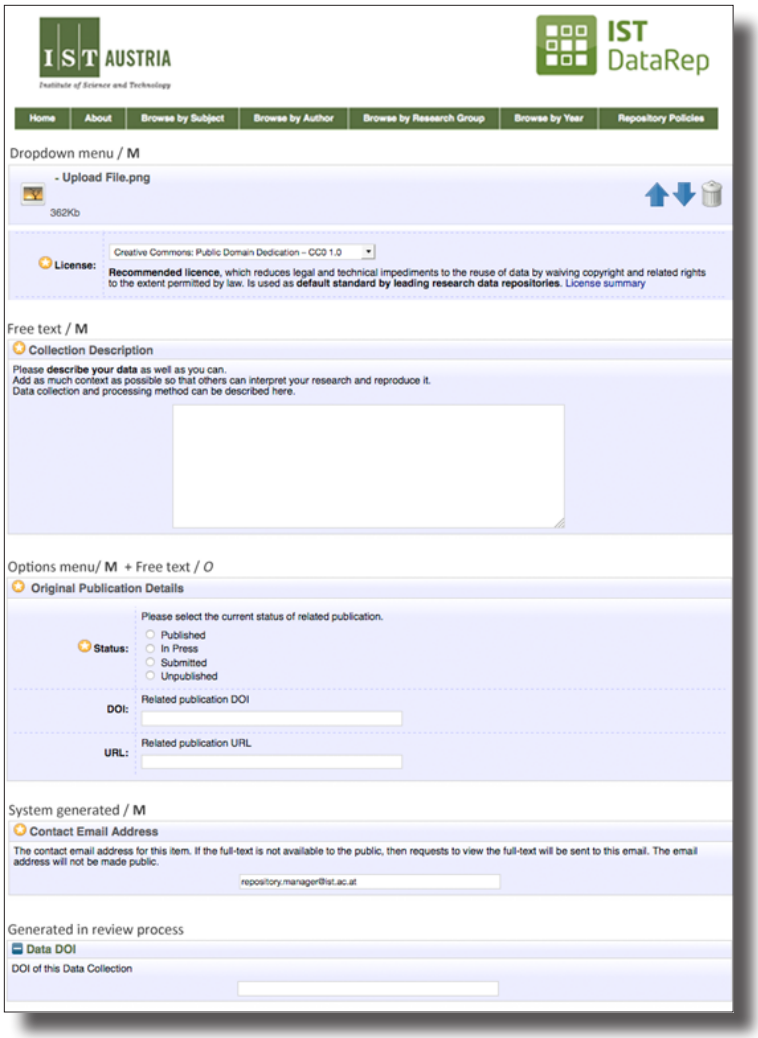


Fig. 4: Upload form

focusing on seamless workflows for data publication and another, which is debating the various aspects of research data on an institutional basis to achieve a consent in definition but also an understanding of the distinctive perspectives on research data.

After an initial phase of little awareness of the RDM (Research Data Management) services at IST Austria, researchers are now approaching the library for advice and cooperation. The increasing interest in data manage-

ment, publication and preservation is certainly due to the recent uptake of these issues in national and international funding policies as well as publishers' data policies. For us this proves that the implementation of the data repository happened at the right time and that the library of IST Austria is well prepared for future tasks of data management.

Barbara Petritsch
ORCID: http://orcid.org/0000-0003-2724-4614
Institute of Science and Technology Austria (IST Austria), Library
E-Mail: barbara.petritsch@ist.ac.at

1 Porsche, Jana: *Actual state of research data @ ISTAustria*. Project Report, IST Austria 2012. Online: https://repository.ist.ac.at/103/.
2 Tail (small, medium) data is, in contrast to head (big) data, described as heterogeneous, hand generated and created and processed under individual procedures. Therefore the curation is not likely to be organised centrally and/or automatically (i.e. subject repository). Furthermore, accessibility is a major issue due to legal uncertainties and lack of effective platforms (i.e. institutional repositories). See also: Heidon, Bryan P: *Shedding Light on the Dark Data in the Long Tail of Science*. Library Trends 2008, 57, 2, 280–299. Online: http://hdl.handle.net/2142/10672.
3 For further information on this process see: Porsche, Jana. *Technical requirements and features*. Project Report. IST Austria 2013. Online: https://repository.ist.ac.at/135/.
4 The ReCollect plugin transforms an EPrints install into a research data repository with an expanded metadata profile for describing research data (based on DataCite, INSPIRE and DDI standards) and a redesigned data catalogue for presenting complex collections. Developed by the UK Data Archive and the University of Essex, as part of the JISC MRD Research Data @Essex project. See also: http://bazaar.eprints.org/367/ (28.3.2017).
5 https://schema.datacite.org/.
6 http://dublincore.org/documents/dces/.
7 Recollect has 46 fields by default. We dropped the fields: alternative title, corporate creators, collection date, Bounding Box North, Bounding Box East, Bounding Box West, Bounding Box South, Data Collection Method, Legal and ethical issues, Lineage, Additional Information, Original Data Publisher, Restrictions, Copyright Holder, Note. See also: http://www.data-archive.ac.uk/media/375386/rde_eprints_metadataprofile.pdf (28.3.2017).

8 Weibel, Stuart:*The Dublin Core: A Simple Content Description Model for Electronic Resources*. Bul. Am. Soc. Info. Sci. Tech. 1997, 24, 9–11. DOI: https://dx.doi.org/10.1002/bult.70.

9 Brase, Jan: *European Initiative to Facilitate Access to Research Data*. D-Lib Magazine 2009, 15, 5/6. DOI: https://dx.doi.org/10.1045/may2009-inbrief.

10 Depending on the specific field subfields are mandatory or optional.