B2FIND – Searching for Research Data across Disciplines

Claudia Martens[1] and Markus Demleitner[2,3]

[1]Deutsches Klimarechenzentrum, Abteilung Datenmanagement
[2]Universität Heidelberg, Zentrum für Astronomie
[3]German Astrophysical Virtual Observatory GAVO

B2FIND is a cross-disciplinary data search engine developed within the pan-European Collaborative Data Infrastructure EUDAT CDI. It has been the central indexing tool for EOSC-hub and it plays a major role in the Data Infrastructure Capacity for EOSC (DICE). Technically, it is a comprehensive joint metadata catalogue that includes metadata records for data stored in various data centres and using different schemas on widely differing granularity. Hence, the major challenge B2FIND faces is bridging the semantic and structural gaps between different disciplines' practices. This contribution discusses how these bridges were built using the example of the Virtual Observatory, a global data infrastructure in Astronomy which itself looks back on 20 years of evolution in metadata management.

1 Introduction

The first letter in the FAIR acronym stands for "Findable", and indeed providing interfaces and APIs that allow for various modes of research data discovery is a very basic requirement for any sort of data infrastructure. Since data structures differ a lot between disciplines, so do the metadata schemas of disciplinary data infrastructures.

On the other hand, data discovery spanning disciplines promises rich rewards in terms of cross-fertilising research. A federation of the various tailored discovery models in the disciplines allows them to retain those while still enabling the cross-disciplinary discovery.

B2FIND has been providing data discovery federated across many disciplines since 2014. This contribution shows how the federation is effected, both centrally at B2FIND (section 2) and on the level of the individual disciplines, which can, to some extent, already accomodate some of the requirements of cross-disciplinary data discovery (section 3). We conclude with a brief look at the bigger picture: What parts of the picture of cross-disciplinary data discovery are still not quite in place?

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021.
Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13729

2 B2FIND – an interdisciplinary discovery portal for research data

2.1 Description

B2FIND is a discovery service for research data distributed within the European Open Science Cloud (EOSC) and beyond. It was built in 2014 as part of the European Union Research and Innovation programme Horizon2020 for the European Data e-Infrastructure Initiative (EUDAT), which is now the pan-European Collaborative Data Infrastructure (EUDAT-CDI), currently consisting of 28 partners, including most major European data centres and research organisations[1]. B2FIND has been the central indexing tool for EOSC-hub[2] and is a basic service of the Data Infrastructure Capacity for the EOSC[3] (DICE).

To fulfill these roles, a comprehensive metadata catalogue was built that includes metadata records for data that is stored in various data centres, using different data formats and metadata schemas on widely divergent granularity levels – where resources are sometimes single files, sometimes complex hierarchies involving millions of files –, representing all kinds of scientific output: from huge netCDF files produced by climate modeling to small audio records of Swahili syllables and phonemes; from the immigrant panel data in the Netherlands to a paleoenvironment reconstruction of the Mozambique Channel and from an image of the “Maison du Chirurgien” in ancient Roman Pompeii to an Excel file giving concentrations of calcium, magnesium, potassium and sodium in throughfall, litterflow and soil in an Oriental beech forest.

In order to enable this interdisciplinary perspective, different metadata formats, schemas and standards are homogenized on the B2FIND metadata schema[4], which is based on the DataCite schema[5] extended with the additional elements `<Discipline>` and `<Instrument>`, allowing users to search and find research data across scientific disciplines and research areas as well as searching for certain measurement tools, e.g., data produced by specific beamlines or measurement stations.
Additionally, the B2FIND schema includes a`<TemporalCoverage>`that allows users to search for a certain time range research data is related to.

Good metadata management is guided by FAIR principles, including the establishment of common standards and guidelines for data providers[6].

In this effort, close cooperation and coordination with scientific communities, research infrastructures and other initiatives dealing with metadata standardisation (OpenAIRE Advance, RDA interest and working groups, and several EOSC related projects) are es-

[1] https://www.eudat.eu/eudat-cdi
[2] https://www.eosc-hub.eu
[3] https://www.dice-eosc.eu
[4] http://b2find.eudat.eu/guidelines/mapping.html
[5] https://schema.datacite.org/
[6] http://b2find.eudat.eu/guidelines

sential in order to establish standards that are both reasonable for community-specific needs and usable for enhanced interoperability.

The main question still is how to find a balance between community-specific metadata that serves their communities' needs on the one side and a metadata schema that is sufficiently generic to capture interdisciplinary research data, but at the same time is specific enough to enable targeted queries yielding results useful to the querier's research. This balance is not a static point, but rather an ongoing process depending on new technical developments as well as on political decisions. Even within the European Open Science Cloud consensus on a "Core Minimum Set" has not been reached yet, and what techniques for metadata exposure and exchange will prevail has yet to be seen.

2.2 Workflow

B2FIND's workflow for metadata ingestion basically consists of three steps: harvesting metadata, mapping them and uploading the final JSON records to a database for indexing and search. These steps are briefly described here in order to provide additional perspective on why the close community involvement described in sect. 3 is helping B2FIND.

2.2.1 Harvesting

Preferably, B2FIND uses the Open Archives Initiative Protocol for Metadata Harvesting OAI-PMH [8] to harvest metadata from data providers. OAI-PMH offers several options that makes it a suitable protocol for harvesting:

(a) a facility to define diverse metadata prefixes (minimal requirement is Dublin Core, further prefixes are optional),

(b) a facility to create subsets for harvesting (useful for large amounts of records or divergent records, e.g., from different projects or sites or measurement stations), and

(c) a facility to configure incremental harvesting.

Nonetheless, B2FIND supports other harvesting methods as well, e.g., the Open Geospatial Consortium Catalogue Service for the Web (OGC-CSW) or various REST APIs. Harvesting triples from SPARQL endpoints is implemented only in a beta version.

2.2.2 Mapping

The mapping process is twofold as it includes a format conversion as well as a semantic mapping based on standardized vocabularies (e.g., the field "Language" is mapped onto

ISO 639 codes, and “Discipline” is mapped on a standardized closed vocabulary[7]). First, entries from XML (or JSON) records are being parsed to assign them to the keys specified in the B2FIND schema. The resulting key-value pairs are stored in JSON dictionaries and checked/validated before being uploaded to the B2FIND repository. Formerly depending on XPATH rules, the current version of the B2FIND ingestion software[8] includes a very flexible mapping procedure: several harvesting endpoints using different metadata standards may be integrated within one “Community”. For each endpoint distinct mapping “issues” may be implemented (e.g., specifically defined methods for certain values and attributes, perhaps for representing different sorts of `<identifier>`, for using the header datestamp as `<PublicationYear>` or assigning additional `<keyword>`s)[9].

Currently B2FIND supports generic metadata schemas such as DataCite and Dublin Core. Community-specific metadata schemas are supported as well, e.g., ISO19115/19139 (which is the basis for Inspire) and FGDC (which is a DublinCore crosswalk for ISO 19135) for Environmental Research Communities, or FF for Nordic Archaeologists. In principle, our ingestion software allows us to integrate any metadata schema and we support the implementation of new schemas. The integration of DDI for Social Sciences is currently developed within the frame of a FAIRsFAIR project that aims to improve interdisciplinary research data discovery using DDI-CDI (an “enhanced” version of DDI attempting to make metadata interoperable across disciplines) and DCATv2 (an RDF vocabulary designed to facilitate interoperability between data catalogues published on the Web, also a W3C Recommendation).Integration of further RDF vocabularies in our metadata ingestion source code is planned but depends on the option to harvest triples, which again requires resources for software development.

2.2.3 Upload and indexing

B2FIND’s search portal and GUI is based on the open source portal software CKAN[10], which comes with an Apache Lucene SOLR servlet. We use it to index the mapped JSON records and offer performant faceted search functionalities. CKAN has a very limited internal metadata schema which has been enhanced for B2FIND by creating additional metadata elements as CKAN “extra” fields.

B2FIND offers a full text search, in addition to which results may be narrowed down using 12 (as of mid-2021) facets, including spatial and temporal search (using a map and an extension for “Timeline Search”) and Communities, Keywords, Creator, Publication Year, Discipline, Language, Publisher, Contributor, Resource Type, and Information on licensing and/or access restrictions.

[7] As there is no useful generic classification for scientific disciplines (except perhaps simple taxonomies issued by funding bodies), the B2FIND closed vocabulary is based on re3data’s subject schema. A revision is currently ongoing as part of a collaboration project Classification for Research Areas, clara.science.

[8] `https://eudat.eu/news/eudat-unveils-new-improved-b2find-30`

[9] All B2FIND software is openly accessible on Github: `https://github.com/EUDAT-B2FIND`

[10] `https://ckan.org/`

"Community" here is the data provider or research infrastructure that B2FIND harvests from. Using a new OAI-PMH extension for CKAN, all metadata records within our database are also harvested by OpenAIRE, widening the scope of discoverability for research objects. B2FIND in this way acts as a metadata aggregator, exposing metadata that are (in some cases) "hidden" in data repositories which are not crawled by big players like Google[11].

2.3 When is a standard a standard?

The terms metadata "schema" and metadata "standard" are often used interchangeably, and both refer to "the formal specification of the attributes (characteristics) employed for representing information resources" [1]. Thus a metadata schema could be seen as a set of elements with a precise semantic definition and optionally rules how and what values can be assigned to these elements [5]; a metadata standard then is a schema which is developed and maintained by an institution that is a standard-setting one. That means that "a standard is a standard insofar as there is an institutional or organizational standardization unit developing and maintaining a standard - whereas all parties and persons involved agree this institution to be trustworthy and reliable" [9].

Metadata standards evolve in different ways. One way is to enforce their adoption through policies such as EU directive 2007/2/EC[12], which requires institutions and organisations publishing spatial datasets and services to implement Inspire as their metadata schema. Another way is to develop a metadata schema that is useful and thus widely adapted. DataCite may serve as an example here. Nonetheless, as new methods and techniques come into use standards need to continually be developed and adjusted, which is hard work (still) done by humans. As there is no one-and-only standard, interoperability is key.

Reliability is crucial for all aspects of FAIR data principles, escpecially for persistent identification of digital resources. However, while DataCite allows only one type of value for `<identifier>` (its DOI), B2FIND has found it necessary to admit several other types, too (in the present case, the IVOA identifiers discussed below). The internal ranking is: if a DOI is offered it will be displayed; if another PID is offered this will be displayed; if neither DOI nor other PID are offered, B2FIND will display whatever URI is given as `<source>`.

A metadata schema may also be defined as a "logical plan showing the relationships between metadata elements, normally through establishing rules for the use and management of metadata specifically as regards the semantics, the syntax and the optionality" [6], where "syntax" describes the structure of a schema and "semantics" describes the

[11] There are many arguments why a proprietary index like Google dataset search should be viewed with suspicion, many of which have been pointed out elsewhere (e.g., [7]). We mention in passing that Google's revenue from advertisment for less then a second corresponds to the funding B2FIND gets for a year.

[12] `https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX%3A32008R1205&from=EN`

meaning of its elements, properties or attributes. We will follow this definition here and try to describe both low-hanging fruit and concrete obstacles in syntax and semantics mapping.

3 Case Study: Mapping VOResource for B2FIND

In practice, communities with discipline-specific metadata schemas are encouraged to help B2FIND by building their half of the bridge to the other communities themselves – in particular because that is usually much simpler for domain experts than for a generic infrastructure. A discipline-specific service will then emit already processed metadata in, for instance, the DataCite schema or possibly even directly in B2FIND's schema; this will in general be a lossy transformation.

```
1  <ri:Resource created=”2020–11–17T11:00:00Z”
2      updated=”2021–01–29T13:52:42Z”
3      status=”active” xsi:type=”vs:CatalogResource”>
4    <title>Gaia eDR3 source catalogue ”light”</title>
5    <identifier>ivo://org.gavo.dc/gaia/q3/edr3lite</identifier>
6    <curation>
7      <publisher>The GAVO DC team</publisher>...</curation>
8    <content>
9      <subject>astrometry</subject>
10     <description>This is a ”light” version...</description>
11     <referenceURL>http://dc.g–vo.org/tableinfo/gaia.edr3lite</referenceURL>
12     <relationship>
13       <relationshipType>served–by</relationshipType>
14       <relatedResource ivo–id=”ivo://org.gavo.dc/tap”>
15         GAVO Data Center TAP service</relatedResource>
16     </relationship>...</content>
17   <capability standardID=”ivo://ivoa.net/std/TAP#aux”>
18     <interface role=”std” version=”1.1” xsi:type=”vs:ParamHTTP”>
19       <accessURL use=”base”>http://dc.g–vo.org/tap</accessURL>
20     </interface>
21   </capability>
22   <facility>Gaia</facility>
23 <ri:Resource>
```

Figure 1: A sketch of a VOResource metadata record. VOResource includes Dublin Core (e.g., lines 3, 7, 9, 10), and extends it towards items present in DataCite at least in compatible form (e.g., lines 12ff). The core operational metadata (here, the capability in lines 17ff) does not map to anything outside of the VO.

As an example, in this section we discuss the integration of the records in the Virtual Observatory's Registry into B2FIND. As discussed in [2], the native metadata schema in the Virtual Observatory (VO) is built on the discipline-specific standard VOResource

[10] and its extensions. A (severely abridged) sketch of a record is shown in Fig. 1[13]. For illustration, we give the result of the mapping of this information to the oai_datacite-like XML consumed by B2FIND in Fig. 2.

When converting metadata from one schema to the other, there are five major categories of items:

- Elements with exactly matchable semantics
- Elements having simply mappable semantics
- Elements having having similar semantics but requiring more complex mapping procedures
- "Important" elements missing in the source schema
- "Important" elements missing in the target schema

We will discuss examples for each of these in turn.

3.1 Exactly Matching Semantics

When source and target schema have elements with exactly matching semantics, the conversion is a very simple syntactic operation. Even in our example, where both VOResource and DataCite are siblings with respect to their common parent Dublin Core, that is a rare exception. Only the `publisher` matches down to structural details (Fig. 1, line 7 vs. Fig. 2, line 3), and even there some adjustment is necessary, because the element directly sits in the record root in DataCite, where in VOResource it is part of a `<curation>` child element.

Usually, the different requirements in the domain show on the structural level. For instance, in VOResource's all-English world of professional astronomy, a single title is enough (Fig. 1, line 4). The more general DataCite admits multiple titles (Fig. line 2, line 2). As long as the target format admits a superset of the source format's features, this would only turn into a problem if the conversion would ever need to be reversed (e.g., turning DataCite to VOResource); in our case, this does not seem a likely requirement.

Sometimes, a single element in the source schema maps to a pair of element and attribute name in the target schema, as when adorning VOResource's `<description>` (Fig. 1, line 10) with an `descriptionType="abstract"` attribute (Fig. 2, line 16ff). Again, these are operations not trivially invertible.

None of these present technical challenges; the VO's B2FIND interface simply re-uses pre-existing XSLT[14] originally written to facilitate DOI minting from VO resource records.

[13] See `http://dc.zah.uni-heidelberg.de/getRR/gaia/q3/edr3lite` for a full version.

[14] `https://github.com/msdemlei/datalink-xslt`

```
1  <d:resource>
2    <d:titles><d:title>Gaia eDR3 source catalogue ”light”...</d:title></d:titles>
3    <d:publisher>The GAVO DC team</d:publisher>
4    <d:publicationYear>2020</d:publicationYear>
5    <d:subjects>
6      <d:subject>observational-astronomy</d:subject></d:subjects>
7    <d:resourceType resourceTypeGeneral=”Other”/>
8    <d:alternateIdentifiers>
9      <d:alternateIdentifier alternateIdentifierType=”ivoid”>
10       ivo://org.gavo.dc/gaia/q3/edr3lite</d:alternateIdentifier>
11     <d:alternateIdentifier alternateIdentifierType=”reference␣URL”>
12       http://dc.g-vo.org/tableinfo/gaia.edr3lite</d:alternateIdentifier>
13   </d:alternateIdentifiers>
14   <d:relatedIdentifiers/>
15   <d:formats/>
16   <d:descriptions><d:description descriptionType=”Abstract”>
17         This is a ”light” version...
18   </d:description></d:descriptions>
19 </d:resource>
```

Figure 2: A sketch of the DataCite-like metadata record generated from the VOResource shown in Fig. 1. The various classes of changes between this and the original record are discussed in the text.

3.2 Simply Mappable Semantics

VOResource has the notion of a reference URL (Fig. 1, line 11), which should resolve to a web page with human-readable information on the resource. Also, its identifier (an “ivoid” or IVOA identifier; Fig. 1, line 5) does not directly map to anything in DataCite. For the B2FIND mapping, we decided to map them into alternate identifiers (with types of “reference URL” and “ivoid”; Fig. 2, line 8ff). While for the ivoid, this obviously is the right thing to do – what software there is that knows how to resolve ivoids can easily be taught to pick up the particular alternate identifier type –, the reference URL only passes as an identifer in a very abstract sense; in reality, it would fit a “documentation”-like link a lot better.

However, the B2FIND-internal mapping can easily be taught to pull the reference URL from the alternate identifier (presenting it as the main URI for the result). Again, the pre-existing XSLT mentioned above is sufficient for these transformations.

3.3 Complex Mapping

The case of <subject> (Fig. 1, line 9 vs. Fig. 2, line 6) is more complex; in the example, the keyword “astrometry” turns into “observational-astronomy”.

Both VOResource and DataCite admit multiple subject elements (albeit in different positions), but expecting a cross-disciplinary search engine to deal sensibly with the keyword schemes of all disciplines contributing is probably not realistic (although having all the respective vocabularies uniformly in RDF would probably help in such an effort).

Instead, B2FIND has a list of relatively coarse-grained subject keywords. For astronomy, it includes the top-level terms from the Unified Astronomy Thesaurus [4]. Hence, when preparing records for B2FIND, the Virtual Observatory side adds the top-level terms for all subject keywords found where these keywords can be located in the UAT. This is a relatively complex operation using external resources (the UAT vocabulary, database queries) that would hence be very hard indeed to implement in XSLT. Hence, this part is put directly into the code serving the oai_datacite metadata format[15].

3.4 Lacunae in the Source Schema

The most debatable part of the VO-B2FIND interaction is that so far the oai_datacite metadata schema is used with invalid records. That is because most of the resources in the VO do not have DOIs assigned at this point, and even those that have a DOI do not always declare it in their VOResource. However, oai_datacite requires an `<identifier>` element, and it must contain a DOI, so the VO's unique (though not persistent) identifier (Fig. 1, line 5) cannot be used.

The solution for now is to skip the identifier element, thus producing invalid DataCite records. As a bespoke practice between two parties, this works well enough, and it will confuse no generic clients with non-DOI material in identifier elements. As the VO moves to offering the B2FIND metadata schema, this can be cleaned up.

Much less serious is the lack of various optional metadata such as DataCite's `<formats>`. This particular element is not provided because in the VO, almost all resources come as services, which are primarily operated using protocols typically emitting VOTables (an XML-based format for tables with a focus on rich metadata). However, there is often underlying science data, which then presumably is of the primary interest to the researcher. But it is then hard to predict which formats are available, and VOResource records generally do not give the respective metadata. Within the VO, where data formats are relatively standardised (e.g., images will almost always come in FITS), this is not a problem; beyond the VO, it might be.

3.5 Lacunae in the Target Schema

In actual data discovery in the VO, the arguably most important items are the capabilities (Fig. 1, lines 17f), which define what standard protocols can be used to query the data within the resource described, and the tableset (not in the example), which defines the underlying table structure.

[15]The technical details are discussed in a blog post at `https://blog.g-vo.org/semantics-cross-discipline-discovery-and-down-to-earth-code/`.

Both items are not representable in DataCite, and for the capabilities, this is probably not even desirable, as they can, in general, only be used with specialised clients or libraries, which presumably are not available to users from outside astronomy. An exception is when VO resources declare capabilities that offer a web browser interface; in that case, the XSLT produces a URL-typed alternate identifier.

A perhaps less dramatic problem is what makes the relationship declarations in VOResource (Fig. 1, lines 12ff) disappear. What this declares is that the data collection can be queried through a TAP service (cf. [2]). Because DataCite focuses on data rather than services, there is no relationship type "there is a service for the present data" in the controlled vocabulary for relationType. Hence, this particular relationship cannot be declared in the DataCite record, even though the modelling of relationships in general is rather similar in VOResource and DataCite, and DataCite even allows record producers to define the related resources through generic URIs rather than DOIs; this latter property is used by the conversion XSLT to translate several other types of relationships.

Another important piece of metadata is the coverage in space, time, and spectrum, which for the example resource might look like this:

```
<coverage>
  <spatial>0/0-11</spatial>
  <temporal>57174 57174</temporal>
  <spectral>1.986e-19 4.966e-19</spectral>
</coverage>
```

Decoded, this means "data on the whole sky" (`0/0-11` is a representation of that in an astronomy-specific format called MOC [3] that allows operators to define very complex maps on the sky), for midnight 2015-05-31 (`57174` is a "Modified Julian Date" designating that point in time; to be very exact, one should add that this uses a time scale called TDB, and the clock sits in the solar system's center of mass, though for data discovery, this sort of precision probably does not matter), and that the data pertains to the (somewhat augmented) optical band between 400 and 1000 nm (it is the photon energy in Joule).

While temporal and spectral coverage could probably easily be made usable outside of astronomy, for the spatial coverage this seems a lot more problematic. DataCite's GeoLocation could perhaps be extended (somewhat oxymoronically) to sky coverages, but turning complex MOCs into polygons is only approximatively possible and probably unrealistic. Whether MOCs have sufficient utility outside of astronomy to make support for them desirable in a cross-discipline service is, on the other hand, rather questionable.

4 Concluding Remarks

This last question can be put into more general terms: "How do we find a balance between community-specific metadata that serves the needs of specific communities well but is un-understandable to non-experts on the one side, and a metadata schema that is sufficiently generic to enable meaningful, possibly even blind, interdisciplinary research data discovery?" We do not claim to have an answer, but we tentatively suggest that both the specialised and the generic schemas are needed and that one size will probably never fit all, which would then suggest that the present co-existence of discipline-specific and federating infrastructure is here to stay. Also, developing a generic metadata schema suitable for expressing metadata relevant for cross-disciplinary discovery is an evolving, ongoing process. What is already clear at this point is that creating machine-readable, mappable metadata is hard and requires experts assisting data publishers.

In closing, let us mention that we feel even the most fundamental question, "How do we enable reuse of research data across disciplines?", is not satifactorily answered at this point. Using data usually is quite a bit harder than reading a paper, which very typically already is hard for non-experts in some sub-discipline. Hence, as perhaps already discernible from our emerging collection of (mostly fictitious) user stories[16], cross-discipline data discovery quite likely will more often than not involve expert discovery.

Acknowledgements

Part of the work reported on here was funded by the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

Bibliography

[1] G. Alemu and B. Stevens. *An Emergent Theory of Digital Library Metadata - Enrich then Filter*. Amsterdam: Elsevier, 2015.

[2] M. Demleitner. Practical Interoperability in the Virtual Observatory. In *Proceedings of the 3rd Heidelberger e-Science-Tage (this volume)*, 2021.

[3] Pierre Fernique, Thomas Boch, Tom Donaldson, Daniel Durand , Wil O'Mullane, Martin Reinecke, and Mark Taylor. MOC - HEALPix Multi-Order Coverage map Version 1.1. IVOA Recommendation 07 October 2019, October 2019.

[4] Katie Frey and Alberto Accomazzi. The Unified Astronomy Thesaurus: Semantic Metadata for Astronomy and Astrophysics. *Astrophysical Journal Supplement*, 236(1):24, May 2018. `arXiv:1801.01021`, `https://doi.org/10.3847/1538-4365/aab760`

[5] B. Haslhofer and W. Klas. A survey of techniques for achieving metadata interoperability. *ACM Computing Surveys*, 42 (7), 2010.

[6] ISO TC46/SC 11. Managing metadata for records – part 3: Self-assessment method. ISO/TR 23081-3:2011, 2011. URL: `https://www.iso.org/standard/57121.html`.

[16] `https://github.com/msdemlei/cross-discipline-discovery`

[7] Peter Kraker. #dontleaveittogoogle - how open infrastructures enable continuous innovation in the research workflow. Presentation held at Open Science Conference, March 19th-20th 2021, Berlin., 2019. URL: https://www.open-science-conference.eu/wp-content/uploads/2019/03/Peter-Kraker.pdf.

[8] Carl Lagoze, Herbert Van de Sompel, Michael Nelson, and Simeon Warner. The open archives initiative protocol for metadata harvesting, version 2.0, 2002. URL: http://www.openarchives.org/OAI/openarchivesprotocol.html.

[9] Erwann Quimbert, Keith G. Jeffery, Claudia Martens, Paul Martin, and Zhiming Zhao. Data cataloguing. In Z. Zhao and M. Hellström, editors, *Towards Interoperable Research Infrastructures for Environmental and Earth Science*. Springer, 2020. https://doi.org/10.1007/978-3-030-52829-4_8

[10] Raymond Plante, Markus Demleitner, Kevin Benson, Matthew Graham, Gretchen Greene, Paul Harrison, Gerard Lemson, Tony Linde, and Guy Rixon. VOResource: an XML Encoding Schema for Resource Metadata Version 1.1. IVOA Recommendation 25 June 2018, June 2018. https://doi.org/10.5479/ADS/bib/2018ivoa.spec.0625P.