Analysis of current RDM applications for the interdisciplinary publication of research data

André Langer, Ellen Bilz and Martin Gaedke

VSR.Informatik.TU-Chemnitz.de

SEMANTiCS Conference Karlsruhe 2019
1st International Workshop on Approaches for Making Data Interoperable (AMAR 2019)
Research Data
Research data

„any kind of digital artifact that is associated with scientific research”

[Sousa et. al, 2014]
Data as increasingly FAIR Digital Objects

Re-useless data (80%)

Findable

FAIR metadata

FAIR data-restricted access

FAIR data-Open Access

FAIR data-Open Access/Functionally Linked

[Mons et. al, 2017]
Research Publications

Research Data

Research Information
Institutions (optional)

Enter the name of the institution

Categories for this data

Enter the name of the category

Description of this data

3000 characters left

Steps to reproduce (optional)

3000 characters left
Specify subjects from a taxonomy or controlled vocabulary. Each term must be uniquely identified (e.g. a URL). For free form text, use the keywords field in basic information section.

<table>
<thead>
<tr>
<th>Term</th>
<th>Identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term</td>
<td>Identifier</td>
</tr>
</tbody>
</table>

*Add another subject*
Research Question
Which **relevant platforms** for research data publishing are currently used?
To which extent do they support Linked Data based meta data annotations for interdisciplinary reuse?
Assumptions

There is an interest in publishing existing research data

Appropriate domain-specific ontologies already exist

Benefits from a structured, unambiguous data description
Approach
Interdisciplinary Access to Research Data

Associated Concepts and Related Resources

Appropriate Ontologies

General-purpose description

Domain-specific description

Research Object

PID
In order to identify relevant systems for research data publishing, we ran a **systematic mapping** on scientific publications of the last ten years.
Inclusion criteria

+ dealing with research data management and data publishing or data sharing
+ papers published between 2008 and 2018
+ in English
Exclusion criteria

- domain-specific solutions
- different topic focus
- deprecated solutions
Conduction

Conduct Search
- Google Scholar: 21,100 (1,000)
- ISI Web of Science: 326
- ACM Digital Library: 286
- IEEE Xplore Digital Library: 115
- Springer Link: 21,031 (1,000)
- Mendeley: 73
- Science Direct: 187
- Emerald Insight: 44
- SAGE Journals: 10

Scan article titles (and abstracts) based on criteria
- Google Scholar: 151
- ISI Web of Science: 228
- ACM Digital Library: 76
- IEEE Xplore Digital Library: 69
- Springer Link: 304
- Mendeley: 35
- Science Direct: 85
- Emerald Insight: 25
- SAGE Journals: 9

Import to Mendeley
Remove Duplicates
Reassess based on criteria
- 865 after removing duplicates
- 487 after reassessing criteria

Clustering
- general papers publishing and sharing research data
  - on research data systems in academic environments (16)
  - on publishing and sharing research data in general (73)
- possible types of systems
  - Journals (8)
  - Repositories (70)
  - Systems and Software (213)
  - Platforms (107)
## Search Results

<table>
<thead>
<tr>
<th>System</th>
<th>Count</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>GitHub</td>
<td>101</td>
<td><a href="https://github.com/">https://github.com/</a></td>
</tr>
<tr>
<td>Dryad</td>
<td>67</td>
<td><a href="https://datadryad.org/">https://datadryad.org/</a></td>
</tr>
<tr>
<td>e!DAL</td>
<td>59</td>
<td><a href="http://edal.ipk-gatersleben.de/">http://edal.ipk-gatersleben.de/</a></td>
</tr>
<tr>
<td>DSpace</td>
<td>51</td>
<td><a href="https://duraspace.org/dspace/">https://duraspace.org/dspace/</a></td>
</tr>
<tr>
<td>Figshare</td>
<td>44</td>
<td><a href="https://figshare.com/">https://figshare.com/</a></td>
</tr>
<tr>
<td>Fedora</td>
<td>37</td>
<td><a href="https://duraspace.org/fedora/">https://duraspace.org/fedora/</a></td>
</tr>
<tr>
<td>Eprints</td>
<td>35</td>
<td><a href="https://www.eprints.org/">https://www.eprints.org/</a></td>
</tr>
<tr>
<td>Dropbox</td>
<td>32</td>
<td><a href="https://www.dropbox.com/">https://www.dropbox.com/</a></td>
</tr>
<tr>
<td>CKAN</td>
<td>27</td>
<td><a href="https://ckan.org/">https://ckan.org/</a></td>
</tr>
<tr>
<td>Dataverse</td>
<td>27</td>
<td><a href="https://dataverse.org/">https://dataverse.org/</a></td>
</tr>
<tr>
<td>Zenodo</td>
<td>25</td>
<td><a href="https://www.zenodo.org/">https://www.zenodo.org/</a></td>
</tr>
<tr>
<td>myExperiment</td>
<td>23</td>
<td><a href="https://www.myexperiment.org">https://www.myexperiment.org</a></td>
</tr>
<tr>
<td>Globus</td>
<td>22</td>
<td><a href="https://www.globus.org/">https://www.globus.org/</a></td>
</tr>
<tr>
<td>Virtuoso</td>
<td>22</td>
<td><a href="https://virtuoso.openlinksw.com/">https://virtuoso.openlinksw.com/</a></td>
</tr>
<tr>
<td>B2SHARE / EUDAT B2 Services</td>
<td>21</td>
<td><a href="https://b2share.eudat.eu/">https://b2share.eudat.eu/</a></td>
</tr>
<tr>
<td>Drupal</td>
<td>20</td>
<td><a href="https://www.drupal.org/">https://www.drupal.org/</a></td>
</tr>
<tr>
<td>XSEDE</td>
<td>14</td>
<td><a href="https://www.xsede.org/">https://www.xsede.org/</a></td>
</tr>
<tr>
<td>Dendro</td>
<td>13</td>
<td><a href="http://dendro.fe.up.pt/">http://dendro.fe.up.pt/</a></td>
</tr>
<tr>
<td>D2R</td>
<td>11</td>
<td><a href="http://d2rq.org/d2r-server">http://d2rq.org/d2r-server</a></td>
</tr>
<tr>
<td>HUBzero</td>
<td>10</td>
<td><a href="https://hubzero.org/">https://hubzero.org/</a></td>
</tr>
<tr>
<td>Google Drive</td>
<td>9</td>
<td><a href="https://drive.google.com/">https://drive.google.com/</a></td>
</tr>
</tbody>
</table>
Comparison
In order to compare the identified research data publishing systems, we ran a **comparative study** based on 15 criteria derived from the FAIR principles for data sharing.
## Findable (LD)
- C1 Is a particular research data set in a current version accessible via a unique PID?
- C2 Is the research data information through that platform indexed in data catalogs, registries and search engines?
- C3 Is a search interface available with filter possibilities for structured Linked Data?

### Accessible (LD)

<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
<th>C7</th>
<th>C8</th>
<th>C9</th>
<th>C10</th>
<th>C11</th>
<th>C12</th>
<th>C13</th>
<th>C14</th>
<th>C15</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDM Systems</td>
<td>CKAN</td>
<td>+</td>
<td>+</td>
<td>o</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>o</td>
<td>+</td>
<td>o</td>
<td>-</td>
<td>(-)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>eDLAL</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>o</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>?</td>
<td>o</td>
<td>+</td>
<td>+</td>
<td>?</td>
<td>+</td>
<td>o</td>
</tr>
<tr>
<td></td>
<td>ePrints</td>
<td>+</td>
<td>+</td>
<td>o</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>o</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>o</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Dataverse</td>
<td>+</td>
<td>+</td>
<td>o</td>
<td>+</td>
<td>+</td>
<td>o</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>o</td>
<td>+</td>
<td>+</td>
<td>o</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>DSpace</td>
<td>+</td>
<td>+</td>
<td>o</td>
<td>o</td>
<td>(o)</td>
<td>+</td>
<td>+</td>
<td>(-)</td>
<td>(+)</td>
<td>(-)</td>
<td>(-)</td>
<td>o</td>
<td>+</td>
<td>o</td>
<td>-</td>
<td>(o)</td>
</tr>
<tr>
<td></td>
<td>HUBzero</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>%</td>
<td>o</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Invenio</td>
<td>+</td>
<td>+</td>
<td>o</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>o</td>
<td>-</td>
<td>-</td>
<td>o</td>
<td>-</td>
<td>o</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>RDM Applications</td>
<td>Dryad</td>
<td>+</td>
<td>+</td>
<td>o</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>EUDAT B2</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>o</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Figshare</td>
<td>+</td>
<td>+</td>
<td>o</td>
<td>+</td>
<td>o</td>
<td>+</td>
<td>o</td>
<td>-</td>
<td>-</td>
<td>o</td>
<td>o</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>o</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Globus</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>-</td>
<td>o</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>o</td>
<td>-</td>
<td>o</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>myExperiment</td>
<td>+</td>
<td>o</td>
<td>o</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>o</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>XSEDE</td>
<td>o</td>
<td>-</td>
<td>o</td>
<td>o</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Zenodo</td>
<td>+</td>
<td>+</td>
<td>o</td>
<td>o</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>o</td>
<td>-</td>
<td>o</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
</tbody>
</table>

### Interoperable (LD)
- C8 Is the metadata provided in a way such as:
- C9 Can participate in registries
- C10 Can domain-specific metadata
- C11 Can each contribute a URI?

### Reusable (LD)
- C12 Can a data license be obtained?
- C13 Can the data be reused?
- C14 Are data sets in multiple community?
- C15 Is the provision of the data sustainable?

### RDM Tools
- D2R
- Dendro
- Fedora
- Virtuoso

+ was assigned, if the criterion was entirely fulfilled
o was assigned, if the criterion was partially fulfilled
- was assigned, if the criterion was not fulfilled
% was assigned, if the criterion was not applicable
? was assigned, if it was not possible to assess the mentioned criterion
() was assigned, if the feature is limited in the native version but might be there with plugins.
Conclusion

• Comparative study with 18 identified web-based general-purpose research data publishing solutions

• **Support for Linked Data** input and exposure differs among all examined platforms

• Input capabilities for LD meta data often limited to basic discovery meta information
Inspired and Interested?

Andre.Langer@Informatik.TU-Chemnitz.de

VSR.Informatik.TU-Chemnitz.de

@myVSR    /myVSR
Metrics Selection

Either upload a DaQAR requirement profile

Drop files here to upload

Or select the criteria that you are interested in:

- Check all possible criteria listed below

Accessibility dimensions

Availability

- http://res.semquire.net/concepts/DereferencedForwardLinksMetric
- http://res.semquire.net/concepts/DumpDownloadAvailableMetric
- http://res.semquire.net/concepts/NoMisreportedContentTypeMetric
- http://res.semquire.net/concepts/SPARQLAccessibilityMetric
Keywords

Please select a fitting mapping for Semantic Web

Semantic Web
extension of the Web to facilitate data exchange

Semantic Web: interoperability, usability, applicability
journal published by IOS Press

(None of the above)
None of the above options matches to my keyword


