Navigating the long tail

Towards practical guidance for researchers on how to select a repository for long tail data

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA Research Data Management Helpdesk

Roman Gerlach, Markus Baaske, Birgitta König-Ries

MichaelStifelCenterJena for Data-Driven and Simulation Science

http://www.researchdata.uni-jena.de/

Motivation

With over 2000 entries in the Registry of Research Data Repositories (re3data.org, January 2018) researchers are confronted with a plethora of repositories to deposit research data. Given the diversity of these services, we have noticed that researchers find it challenging to make an informed decision, especially when they are dealing with data from the so-called "long tail" (small, diverse, individual, less standardized data). Although, re3data.org provides a very comprehensive list of criteria (i.e. filters) to narrow down the number of choices, there is still advice needed, for example, on evaluating the importance of a criterion (e.g. type of repository) or the impact of a certain choice (e.g. which PID?).

Objectives

From the perspective of a research data management helpdesk we investigated how we could address this selection challenge. The aim is to first compare five generic repositories, evaluate the elements available on re3data.org to describe them, and later develop a practical guide for researchers.

Methodology

A) Comparison of generic data repositories based on **criteria available at re3data.org.**

Repository selection criteria in this study:

- Generic repository without domain specific focus
- "common" researchers selection to choose from, based on our experience
- Globally well-known and well established (Dryad, figshare, Zenodo) vs. a national (RADAR) and an institutional repository (Digital Library Thuringia, DBT) with a similar aim
- Registered with re3data.org (DBT was registered during this study)

Descriptions and properties of each repository were gathered using the re3data.org API, e.g. https://www.re3data-.org/api/v1/repository/[r3d100000044,r3d100010066,r3d100010468,r3d100012330](accessed on 04/01/2018). Data for the DBT was collected by the authors and approved by DBT staff following the same re3data.org schema (version 2.2). We also explored the individual websites of the repository providers using the URLs provided in re3data.

B) Comparison based on **criteria frequently mentioned by researchers** and evaluation whether these criteria match with the information available at re3data.org. For this study, we deliberately have chosen three criteria (visibility, data curation, cost) that seem to be important to researchers and are hard to assess on re3data.org. Challenges arise because information is:

- spread over multiple filters (e.g. visibility)
- "hidden" in rather technical terms (e.g. visibility)
- lacking detail and precision (e.g. review & curation)
- only available at individual repository website (e.g. cost)

	DRYAD	figshare	zenodo	RADAR FIZ Karlsruhe	dbt Digitale Bibliothek Thüringen
repositoryName	DRYAD	figshare	Zenodo	RADAR	Digital Library Thuringia
description	DataDryad.org is a curated general-purpose repository that makes the data underlying scientific publications discoverable, freely reusable, and citable. Dryad is an international repository of data underlying peer-reviewed scientific and medical literature, particularly data ...	figshare allows researchers to publish all of their research outputs in an easily citable, sharable and discoverable manner. All file formats can be published, including videos and datasets. Optional peer review process. figshare uses creative commons licensing.	ZENODO builds and operates a simple and innovative service that enables researchers, scientists, EU projects and institutions to share and showcase multidisciplinary research results (data and publications) that are not part of the existing institutional or subject-based repositories ...	RADAR is an online service for the archival and publication of research data resulting from completed scientific studies and projects. RADAR is a generic, interdisciplinary service which offers two service levels: data archival and data publication (including archival). Data ...	DBT is the institutional repository of the FSU Jena, the TU Ilmenau and the University of Erfurt as well as members of the other Thuringian universities and colleges can publish scientific documents in the DBT. In individual cases, users (via the ThULB Jena) can also archive documents in the DBT.
type	other	other	other	other	institutional
size	...				
startDate	2008	2011	2013	2017	2000
repositoryLanguage	eng	eng	eng	deu, eng	deu, eng
subject	1 Humanities and Social Sciences, 2 Life Sciences,3 Natural Sciences, ...	1 Humanities and Social Sciences, 2 Life Sciences, 3 Natural Sciences, 4 Engineering Sciences	1 Humanities and Social Sciences, 2 Life Sciences, 3 Natural Sciences, 4 Engineering Sciences	1 Humanities and Social Sciences, 2 Life Sciences, 3 Natural Sciences, 4 Engineering Sciences	1 Humanities and Social Sciences, 2 Life Sciences, 3 Natural Sciences, 4 Engineering Sciences
missionStatementURL	...	...	...	...	...
contentType	Plain text, Scientific and statistical data formats, Software applications, Source code, Standard office documents, Structured text, other	Archived data, Audiovisual data, Images, Plain text, Raw data, Scientific and statistical data formats, Source code, Standard office documents, Structured graphics	Archived data, Audiovisual data, Images, Networkbased data, Plain text, Raw data, Scientific and statistical data formats, Source code, Standard office documents, Structured graphics, Structured text, other	Images, Standard office documents, other	Archived data, Audiovisual data, Configuration data, Databases, Images, Networkbased data, Plain text, Raw data, Scientific and statistical data formats, Standard office documents, Structured graphics, Structured text, other
providerType	dataProvider, serviceProvider	dataProvider, serviceProvider	dataProvider	dataProvider, serviceProvider	dataProvider
keyword	Biodiversity, FAIR, interdisciplinary, scientific and medical publications	data collection platform, multidisciplinary	FAIR, multidisciplinary	multidisciplinary	multidisciplinary
institutionName	Drexel University, College of Computing & Informatics, Metadata Research Center, Dryad, Institute of Museum and Library Services, JISC, National Evolutionary Synthesis Center, National Science Foundation, North Carolina State University	Digital Science,The Digital Preservation Network	European Commission, Horizon 2020,European Commission, Research & Innovation, Seventh Framework Programm - FP7,European Organization for Nuclear Research,OpenAIRE	FIZ Karlsruhe - Leibniz Institute for Information Infrastructure	Thüringer Universitäts- und Landesbibliothek Jena, Universitätsrechenzentrum Jena
institutionCountry	AAA , GBR, USA	GBR, USA	EEC	DEU	DEU
responsibilityType	funding, general, technical	general, sponsoring, technical	funding, general, technical	general, technical	general, technical
institutionType	non-profit	commercial, non-profit	non-profit	non-profit	non-profit
institutionURL	...	...	...	...	...
policyName	Joint Data Archiving Policy (JDAP), Membership Agreement, Membership Policy, Terms of Service	COPE principles of transparency and best practice in scholarly publishing, Figshare Support Portal, Privacy Policy	Policies, Terms of use	Customer contract, Upload policy	Leitlinien für den Betrieb des Publikationsserver
policyURL	...	...	...	...	...
databaseAccessType	open	open	open	open	open
databaseLicenseName	CC0	CC	CC0	CC0	
databaseLicenseURL	...	...	...	...	...
dataAccessType	embargoed, open	embargoed, open	closed, embargoed, open, restricted	closed, embargoed, open, restricted	closed, embargoed, open, restricted
dataAccessRestriction			registration	registration	other
dataLicenseName	CC0, other	Apache License 2.0, BSD, CC, CC0, other	CC, CC0, other	CC	CC,ODC-BY, ODbL, PDDL, other
dataUploadType	restricted	restricted	open	restricted	restricted
dataUploadLicenseName	Depositing data to Dryad	Terms and conditions	Policies	Creative Commons	Terms and conditions
dataUploadLicenseURL	...	...	...	...	...
softwareName	DSpace	other	other	eSciDoc	other
versioning	yes	yes			yes
pidSystem	DOI	DOI	DOI	DOI	DOI
citationGuidelineURL	...	...	...		
aidSystem	ORCID	ORCID	ORCID	ORCID	other
enhancedPublication	yes	yes	unknown	yes	unknown
qualityManagement	yes	yes	yes	no	yes
metadataStandardName	Dublin Core	DataCite, Dublin Core	DataCite, Dublin Core	DataCite	MODS v.3.6
metadataStandardURL	...	...	...	...	...
remarks	Dryad is covered by Thomson Reuters Data Citation Index. DRYAD is covered by SCOPUS. Dryad is covered by Elsevier DataSearch. Dryad is a nonprofit organization, governed by its member organizations, including journals, publishers, scientific societies, funding agencies, and other stakeholders, and an international repository of data underlying scientific and medical publications. For more information see http://datadryad.org/pages/membershipOverview	figshare is covered by Thomson Reuters Data Citation Index.c. figshare is partner of the Reproducibility Initiative. figshare uses Altmetric metrics.	Zenodo is covered by Thomson Reuters Data Citation Index. Zenodo uses Altmetric metrics. Zenodo uses invenio repository software. OpenAIRE Orphan Record Repository got a make-over and was re-branded as ZENODO. Zenodo uses Invenio repository software. ZENODO was launched within the OpenAIREplus project as part of a European-wide research infrastructure. Easy upload and semi-automatic metadata completion by communication with existing online services such as DropBox for upload, Mendeley/ORCID/CrossRef/OpenAIRE for upload and pre-filling metadata.	RADAR was established by an experienced consortium to develop an infrastructure that facilitates the preservation, publication and traceability of research data. The project was funded by the German Research Foundation (DFG) from 2013 to 2016 within the programme â€œScientific Library Services and Information Systems (LIS)â€ The technical infrastructure of RADAR was provided by the FIZ Karlsruhe â€“ Leibniz Institute for Information Infrastructure and the Steinbuch Centre for Computing (SCC). The Ludwig-Maximilians-UniversitÃ¤t Munich (LMU), Faculty for Chemistry and ...	The DBT uses the MyCoRe/MIR open source software. Authors are identified by GND or VIAF, ORCID will come in 2018.
entryDate	06/02/2013	22/08/2012	13/06/2013	17/03/2017	24/01/2018
lastUpdate	22/11/2017	18/04/2017	22/11/2017	02/08/2017	28/01/2018
apiType	OAI-PMH, other	OAI-PMH	OAI-PMH, REST	OAI-PMH, REST	OAI-PMH, REST, SWORDv2

[...] indicates that more information is available, but it has been truncated for this view

Visibility

relevant re3data element(s):

- **Database Access Type**
- **Enhanced Publication**
- **PID System**
- **Metadata standards**
- **API**
- **Remarks**

The visibility of a dataset is determined by the interoperability of the repository (i.e. selection of standards, protocols, interfaces). But the average researcher may not be able nor willing to evaluate that. So decisions are likely to be based on 'familiarity' ("*I know DOIs, never heard of ARKs*") and 'the more, the better' ("*two supported APIs must be better than one*").

Data Review & Curation

relevant re3data element(s):

- **Quality Management** (QM)
- **URL to Policies** at repository's website

Dryad	QM = yes	Dryad curation personnel will review and curate content prior to and following publication
figshare	QM = yes	No further information available
Zenodo	QM = yes	Data is deposited „as-is". No further information available
RADAR	QM = no	Data is deposited „as-is". No further review by the service provider
DBT	QM = yes	DBT curation personnel will review and curate content

Cost

relevant re3data element(s):

- **URL to Policies** at repository's website

Dryad	Data Pubishing Charge (DPC): US$ 120, unless: the associated journal, or another organization, has already contracted with Dryad to sponsor the DPC, or the submitter is based in a fee-waiver country; for more than 20GB, US$ 50 per 10GB
figshare	Free of charge for basic service (up to 5GB)
Zenodo	<50GB file size are "free of charge by those without ready access to an organized data centre"
RADAR	Publication (< 50TB): 0.46 € per GB/year + 595 € institution/year + 7.56 € per GB (once)
DBT	Free of charge for member organisations

eeFDM

Acknowledgement: This work was conducted as part of the eeFDM project funded by the Federal Ministry of Education and Research (Germany). We would like to thank Robert Ulrich and Gabriele Kloska from re3data.org for answering our questions and providing background information on re3data. Many thanks also to Annemarie Müller for preparing the data of the DBT as well as Thomas Scheffler and Michael Lörzer from the Thüringer Universitäts- und Landesbibliothek Jena for their support.

Bundesministerium für Bildung und Forschung