IST AUSTRIA
Institute of Science and Technology

Document Name Initiatives_and_projects_related_to_RD.pdf

Author Jana Porsche
Version 1.0
Modified 20.03.2013
Location http://repository.ist.ac.at/id/eprint/113

Table of Contents

Purpose

This document is created as a part of the project "Repository for Research Data at IST Austria". It summarises the actual initiatives, projects and standards related to the project. It supports the preparation of standards and specifications for the project, which should be considered and followed to ensure interoperability and visibility of the uploaded data.

Related benchmarks are listed to provide an overview of best practice examples and so enable eventual cooperation. The intention is to summarize associated initiatives and projects, which don´t need to be crucial for the project at the moment because of their very early stage, but shouldn´t be missed in the future.

The list of possible software in the conclusion of this document should offer a clear overview of potential software choice with Pros and Cons.

All lists in the document are sorted in ascending order.

List of related and inspiring initiatives and projects

ANDS Research Data Australia[1]

Internet-based discovery service designed to provide connections between data, projects, researchers and institutions and promote visibility of Australian research data collections in search engines.

CKAN[2]

Data management system making data accessible by providing tools to streamline publishing, sharing, finding and using data. CKAN is aimed at data publishers wanting to make their data open and available.

Databib[3]

Searchable catalogue, registry and bibliography of research data repositories.

Data Citation Index[4]

Introduced by Thomson Reuters, planning to be indexing preselected data repositories, providing connection to peer-reviewed literature under Web of Knowledge.

DataCite[5]

Global not-profit-organisation supporting citation, finding and reusing of data by assigning persistent identifiers to datasets and requisition of metadata minimum standard.

DFG, German Research Foundation[6]

Germany's largest self-governing funding organisation for science and research in Germany, serves all branches of science and the humanities. Funding also projects supporting research data infrastructure and management in Germany.

Funded related projects (selection):

Radieschen[7]

Roadmap showing recommendations for a multi-disciplinary research data infrastructure in Germany.

Re3data[8]

Registry of research data repositories.

---

[1] http://researchdata.ands.org.au/
[2] http://ckan.org/
[3] http://databib.org/
[4] http://wokinfo.com/products_tools/multidisciplinary/dci/
[5] http://datacite.org
[6] http://www.dfg.de
[7] http://www.forschungsdaten.org/
[8] http://www.re3data.org/

EUDAT, European Collaborative Data Infrastructure[9]

Developing initiative with vision to create a European collaborative data infrastructure.

Figshare[10]

Established cloud based research data repository for all data formats with wide visualising possibilities, offers badges for presentation of uploaded data on own homepage. OA journal publisher PLOS concluded partnership with Figshare for all published articles.

JISC Managing Research Data Programme

Focused on digital technology for education and research in UK, funding broad variety of projects, programs and services also related to research data management.

ODE (Opportunities for Data Exchange) Project[11]

ODE´s report on 10 tales of Drivers and Barriers in Data Sharing is representing best practices in sharing, re-using, preserving and citing data.

Research Data Alliance[12]

Organisation founded by 3 international research funding organisations aiming to accelerate international data-driven innovation and discovery through development and adoption of joint infrastructure, policy, practice and standards sharing.

Useful initiatives

This is a selection of initiatives, which could be used in our project to ensure high visibility of the data stored in the repository and also interoperability of the related metadata.

DataCite (worldwide)

DataCite is a worldwide initiative focused on assignment of unique Digital Object Identifier and standardization of related metadata.
Could be used for: minimal metadata standard, ensuring of citation possibility by assignment of DOI and increasing visibility of data through aggregation.

DCC[13] **(UK)**

DDC (Digital Curation Centre) is a centre of expertise in digital information curation (based in UK). It has collected disciplinary metadata standards.
Could be used for: broad range of How-to Guides and recommendation on policies related to publishing data and providing case studies, useful collection of disciplinary metadata standards.[14]

[9] http://www.eudat.eu
[10] http://figshare.com
[11] http://www.alliancepermanentaccess.org/index.php/community/current-projects/ode/
[12] http://rd-alliance.org/
[13] http://www.dcc.ac.uk

DDI (worldwide)
Data Citation Index will be a new citation database for usage of research data in publications based on the Web of Knowledge Platform from Thomson Reuters (similar to Web of Science for publications).
Could be used for: in future ensuring visibility of the repository.

EUDAT[15] (EU)
EUDAT (European Collaborative Data Infrastructure) is an international initiative for creation of strategy of sharing research data at pan-European level.
Could be used for: in future to see the development within whole Europe and be so able to take eventually part in further initiatives.

JISC[16] (UK)
JISC (Joint Information Systems Committee) is supporting UK education and research for more than 15 years with the goal to champion the use of digital technology in the UK on world-class level in research, teaching and learning.
Could be used for: cooperation with institutions receiving funds for developing research data management strategies and related repositories and do already have more experience.

Re3data (GE)
Re3data should be a registry of research data repositories, which goal is to create a global registry covering research data repositories from different scientific fields. Thanks to agreed cooperation with DataCite, there will be a further development of complementary user services and dialog on joined development of metadata schemas, data models and web service interfaces.
Vocabulary for describing of research data repositories is available in the version 2.0[17]
Could be used for: increasing visibility of the institutional repository.

Standards

To ensure possible high interoperability of the uploaded metadata, it is important to follow standards established in the research data-sharing environment. This will increase possibility of finding and hereby validating and reusing of the data. To enable harvesting of the metadata and aggregation of the data, OAI-PMH should be supported. To build the links between publications and the data OAI-ORE has to be considered.
The data has to be described by good quality metadata and should fulfil a minimal metadata standard. For a proper reusing and hereby citing of the data, a unique identifier of the data is essential. DataCite offer assignment of DOI´s for data. For DOI registration metadata is required for each object and so a metadata schema for clients was developed to support citation and discovery of datasets.

---

[14] http://www.dcc.ac.uk/resources/metadata-standards
[15] http://www.eudat.eu
[16] http://www.jisc.ac.uk/
[17] http://dx.doi.org/10.2312/re3.002

Another standards for describing data are DCAT and related DCIP used e.g. by CKAN[18].

DataCite Metadata Schema
DOI (digital object identifier)
- actual version 2.2[19]
- list of core metadata properties chosen for accurate and consistent identification of data for citation and retrieval purposes
- 3 types of metadata (with naming conventions, allowed values, controlled lists and other constrains)
    -5 mandatory properties (must be supplied for initial metadata submission)
    - 12 optional properties (can be added to offer better description of the data)
    - 2 administrative metadata properties (assigned by DataCite automatically)

DataCite Metadata Schema can be mapped to Dublin Core Simple elements and Qualified terms, to the International DOI Foundation (IDF) Metadata Kernel, to OECD Dataset metadata elements and to DDI 3.1 elements.
DataCite offers also an OAI - PMH Data Provider[20], a service, which exposes metadata stored in the DataCite Metadata Store.

OAI–PMH[21]
The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a low-barrier mechanism for repository interoperability; it is repository and metadata centric.
Data Providers are repositories that expose structured metadata via OAI-PMH. Service Providers then make OAI-PMH service requests to harvest that metadata. OAI-PMH is a set of six verbs or services that are invoked within HTTP.

OAI–ORE[22]
Open Archives Initiative Object Reuse and Exchange (OAI-ORE) define standards for the description and exchange of aggregations of web resources; it is web and resource centric.
The aggregations, sometimes called compound digital objects, may combine distributed resources with multiple media types including text, images, data and video. It very strongly supports semantic web and linked data.
OAI-PMH and OAI-ORE are complimentary, one can be used without the other or they can be used together.

[18] Introduction of CKAN see chapter „Open Source Solutions“.
[19] http://schema.datacite.org/meta/kernel-2.2/index.html
[20] http://oai.datacite.org/
[21] http://www.openarchives.org/pmh/
[22] http://www.openarchives.org/ore/

Data Catalogue Vocabulary (DCAT)[23]
DCAT is an RDF vocabulary designed to facilitate interoperability between online data catalogues. It is mainly used to describe governmental data catalogues, but can be extended by using additional RDF properties anywhere.

Data Catalogue Interoperability Protocol (DCIP)[24]
DCIP is based on DCAT with some recommendations about usage and includes a mapping to JSON.
It is defined by:

- a JSON and RDF representation for key data catalogue entities such as Dataset (DatasetRecord) and Resource (Distribution) based on the DCAT vocabulary
- a read only REST based protocol for achieving basic catalogue interoperability

Lists of already recognized subject repositories

The institutional repository at IST Austria should be a service provided towards our own scientists and doesn´t intend to substitute the broad range of good established subject repositories. Because some journals are already requesting upload of related data during the publication process within preselected subject repositories, we want to encourage the scientists to upload the data only once. The data uploaded externally should have registered metadata in the institutional repository to ensure complete information about publication and related data.

https://www.lib.umn.edu/datamanagement/datacenters
Selection of publicly accessible data repositories categorized by subject.

http://oad.simmons.edu/oadwiki/Data_repositories
Assorted list of OA subject repositories, part of Open Access Directory.

http://datacite.org/repolist
List offering a view into Databib.[25]
The catalogue doesn´t include only subject repositories, but also institutional, governmental and commercial research data repositories. Users and bibliographers create and curate records that describe data repositories that users can search. The list is a working document only for information purposes. It is a collaborative, annotated bibliography of primary research data repositories developed with support from the Institute of Museum and Library Services.

[23] http://www.w3.org/TR/vocab-dcat/
[24] http://spec.datacatalogs.org/
[25] http://databib.org/

Subject repositories used by researchers at IST Austria

Subject repositories already used by researchers at IST Austria according the requirements from journals, have to be implement into the list of established subject repositories within the institutional repository.

Name	URL	Description
Dryad	http://datadryad.org	repository of data underlying scientific and medical publications
GenBank	http://www.ncbi.nlm.nih.gov/genbank	NIH **genetic sequence database**, an annotated collection of all publicly available DNA sequences
mldata.org	http://mldata.org	repository of **machine learning data**
mloss.org	http://mloss.org	repository of **machine learning** open source **software**
NCEAS Data Repository	http://knb.ecoinformatics.org/knb/style/skins/nceas/	research data sets collected and collated as part of NCEAS' funded activities (**ecological data**)
TreeBASE	http://treebase.org	repository of **phylogenetic information**
UCI	http://archive.ics.uci.edu/ml	repository for **machine learning datasets**

Examples of already established institutional repositories

There is worldwide a broad range of institutional data repositories in beta version, which aren´t used for real datasets at the moment, but are planned to be released soon.
Following institutional repositories are examples of already well-established and used datacentres.

3TU.Datacentrum[26]
3TU.Datacentrum is a joint cooperation of three technical university libraries, operating under the umbrella of the 3TU.Federation (established in February 2007): TU Delft Library, TU Eindhoven Information Expertise Centre and University of Twente Library & Archive. 3TU.Datacentrum currently hosts about 5000 datasets.

eCrystals[27]
eCrystals - Southampton is the archive for Crystal Structures generated by the Southampton Chemical Crystallography Group and the EPSRC UK National Crystallography Service. The repository is based on EPrints and includes today almost 800 records.

[26] http://datacentrum.3tu.nl/en
[27] http://ecrystals.chem.soton.ac.uk/

Institutions already cooperating

At the moment IST Austria is mainly cooperating with the University of Essex, which prepared an institutional repository[28] based on EPrints for research data within a JISC Managing Research Data Programme 2011-2013 currently available at the beta version.

The cooperation itself consists sharing of experience done and the possible sharing of the developed metadata schema in the form of a plug-in in case we would decide to build the own repository for research data on software EPrints.

Possible Software

To offer a reliable repository, it is important to choose the most sufficient software according to internal needs and also possible resources. At the moment, the offer of open source solutions is wide and also well supported by communities, which is very welcomed because of short resources on our side.

Open Source Solutions[29]

CKAN[30]

Pros: highly customisable, easy-to-use web interface, workflow support, data visualization and analytics for tabular data, on-going development to suit needs of RDM hereby of institutional research data repositories

Cons: hosted solution, mainly used for governmental data, individual metadata schema.

Examples: DataDotGC, Data.Gov.UK, Europe´s Public Data, Helsinki Data Portal, Data.Lincoln

DSpace[31]

Pros: supports communities, highly configurable and includes a flexible workflow for applying metadata to assets that will suit complex metadata, used in other projects (DRYAD, Edinburgh DataShare), developed by HP/MIT.

Cons: complex implementation (in comparation with EPrints more demanding setup).

Examples: QMRO aka Queen Mary Research Online , DRYAD, Edinburgh DataShare , http://www.dspace.org/whos-using-dspace

[28] http://researchdata.essex.ac.uk/

[29] http://rdm.c4dm.eecs.qmul.ac.uk/blog/digital-assets-management-software-comparison-pros-and-cons

[30] http://ckan.org

[31] http://www.dspace.org/

Fedora Commons[32]

Pros: store all types of content and its metadata, customer driven front-ends (e.g. Drupal), provide RDF search (SPARQL), access data via Web APIs (REST/SOAP), many storage options (database and file systems), versioning support.

Cons: very complex implementation and deployment process when compared with other alternatives.

Packages built on Fedora commons:
"eSciDoc": http://www.escidoc.org/JSPWiki/en/Overview
- e-research environment, can be used also for data, example DARIAH
"Islandora": http://islandora.ca/ (Drupal+Fedora)
- used for IslandArchives.ca[33], customized Islandora sites used by researchers at UPEI[34] and elsewhere to steward research data.

EPrints[35]

Pros: IST Austria already does have experience with using it for the institutional repository for publications, apparently easy to set up and manage, integrated front-end interface, extensible (Perl), used by more than 200 organisations, long development (since 2001), developed by University of Southampton, plug-in for metadata schema will be available from University of Essex.

Cons: focus on scientific publications, although it can be used for research data.

Examples: Essex RDR
list of further examples http://www.eprints.org/exemplar.php

Dataverse[36]

Pros: made specifically for managing datasets, linking them to publications, and sharing them; supports conversion to open formats, versioning, persistent URI, different levels of access to data, can link to other dataverse networks, developed by Harvard University.

Cons: all formats are accepted but only tabular files have full data support, designed mainly for tabular and statistical data, focus on social science at the moment, but developing also initiatives for other disciplines.

Example: IQSS Harvard University

[32] http://fedora-commons.org/
[33] http://islandarchives.ca/
[34] http://home.upei.ca/
[35] http://www.eprints.org/
[36] http://thedata.org/

Commercial solutions

The availability of a commercial solution was also investigated to ensure complex overview about accessible possibilities, but it is obvious, that the actual offer is at the moment very limited. We can assume this is because of the missing pressure on the publication of data so the commercial subjects don´t invest resources in field, which doesn´t show sufficient demand.

Open Repository[37]
It is a hosted solution for building and maintaining customized DSpace repositories.

CONTENTdm[38]
Will be actually used only for publications.

Digital Commons[39]
Will be actually used only for publications.

Digitool[40]
Will be actually used only for publications.

Equella[41]
Will be actually used only for publications.

Vital[42]
Will be mainly used for publication at the moment.

Zentity[43]
Besides it is a commercial product, the licence is free. Zentity is developed by Microsoft research division to be a semantically enabled repository platform. It strongly supports the relations between data elements.

[37] http://www.openrepository.com
[38] http://www.contentdm.org/
[39] http://digitalcommons.bepress.com/
[40] http://www.exlibrisgroup.com/category/DigiToolOverview
[41] http://www.equella.com/
[42] http://www.vtls.com/products/vital
[43] http://research.microsoft.com/en-us/projects/zentity/