

re3data.org - Making research data repositories visible and discoverable

Robert Ulrich, Karlsruhe Institute of Technology
Hans-Jürgen Goebelbecker, Karlsruhe Institute of Technology
Frank Scholze, Karlsruhe Institute of Technology
Michael Witt, Purdue University

May 15th, 2015

Abstract

This paper gives a brief introduction on re3data.org and some best practices and experiences gathered on encouraging the publication of research data. It is intended for the PSI-Workshop 2015 in Krems, Austria.

Introduction

Research data is valuable and needs to be available so that scientific results can be reviewed and verified. Visible data is a necessary condition that allows for the reuse of collected data, making it possible to build upon previous achievements and to create new knowledge.

Therefore, funding organizations, governmental institutions and scientific journals have begun to make the publication of data obligatory. Their policies offer incentives for researchers in order to encourage them to provide access to their data basis. In many cases, researchers are required to present possible strategies for the publication of their data.

For instance, the Organization for Economic Co-operation and Development (OECD) released their "Principles and Guidelines for Access to Research Data from Public Funding" with the intention "to promote data access and sharing among researchers"[1] .

But the information which is collected and generated in the various disciplines is not uniform. In contrast, it is heterogeneous and of great diversity. As a result, data has been published in many repositories and their number is still growing.

re3data.org is a registry of research data repositories (RDR) [2]. The project aims at providing a map of the RDR landscape and encourages a culture of sharing. It offers help for scientists to find storage places, to either

- 1) find research data
- 2) deposit research data

Their work gains visibility with parts of it stored in appropriate repositories, providing accessibility, stability and reliability. In addition to that, it supports the various stakeholders within the scholarly process.

re3data.org has its main focus on research and the stakeholders within the scientific field. It does not explicitly cover business models based on open data in science. Nevertheless accessible and visible data is

crucial to enable reuse within the economic system. Experiences and decisions made within the scholarly system will also affect commercial usage, relying on publicly funded data basis.







The re3data.org registry

The re3data.org register indexes research data repositories. [2] The website is available at <http://re3data.org>. It is the infrastructure used by the project to provide global access. The front-end of the online service documents the repositories and the service provides access to the records, utilizing an advanced search with filters. Users visiting the page are able to search and filter the datasets. The comprehensible result list presents the main properties of a repository and allows users to navigate and find the desired information easily.

The screenshot displays the re3data.org website interface. At the top, the logo reads "re3data.org" with "REGISTRY OF RESEARCH DATA REPOSITORIES" underneath. A dark navigation bar contains links: Home, Search, Browse, Suggest, FAQ, About, Schema, API, Contact, and Imprint. Below this, the page title is "Search for Repositories (1234 Reviewed Repositories)". A search bar with a magnifying glass icon and the word "Search" is present. Below the search bar are three filter sections: "Subject" with a dropdown menu labeled "Add subjects", "Content Type" with a dropdown menu labeled "Add content types", and "Country (of the responsible institutions)" with a dropdown menu labeled "Add countries". There are also three checkboxes: "Certificates" (checked), "Open Access" (unchecked), and "Persistent Identifier" (unchecked). A "Reset filter" button is located at the bottom right of the filter section. Below the filters, it shows "1234 results (1 - 25)" and a "Sort by" dropdown menu set to "weight". A pagination bar shows numbers from 1 to 27, with 1-25 highlighted. Below the pagination, the first search result is for "3TU.Datacentrum" with the URL "3TU.DC". To the right of the repository name are icons for various standards: ORCID, CC, DOI, and others. Below the repository name, a blue bar contains the text "Subjects: Agriculture, Forestry, Horticulture and Veterinary Medicine | Agriculture, Forestry, Horticulture and Veterinary Medicine".

Illustration 1: Browse and search for data repositories

A specifically developed icon set represents the main properties of a record and supports the user visually.

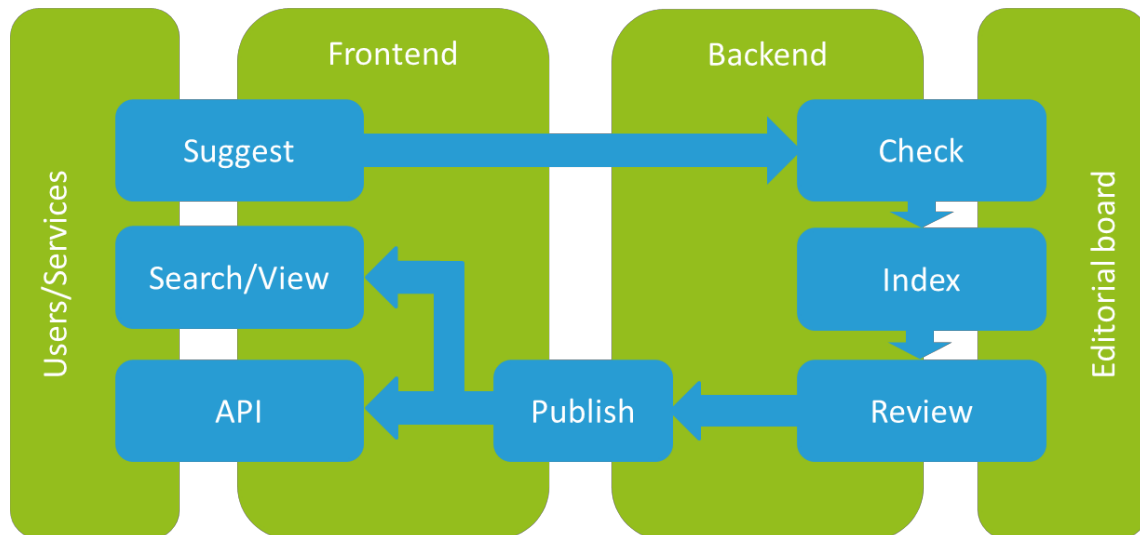
-  It provides additional information on its service
-  It provides information on the terms of access to its data, database and upload
-  It provides terms of use and licenses of the data
-  It uses a persistent identifier system to make its provided data persistent, unique and citable.
-  It is either certified or supports a repository standard.
-  It provides a policy

The details of a data repository are described using the re3data.org schema. It is a list of metadata properties covering a research repository regarding its general scope, content and infrastructure as well as its compliance with technical, metadata and quality standards. The schema includes required metadata properties and optional properties providing additional information [5]. The schema is designed to:

- 1) Recommend a standard for describing a research data repository
- 2) Provide the basis for interoperability between research data repositories and re3data.org
- 3) Be one of the first steps towards the goal of a certificate for research data repositories

Entries in the registry are gathered in a manual workflow backed by the experience of an editorial team. Any user or institute can suggest new repositories. The editorial team will check the repositories on the minimum requirements set by the re3data.org policy and add missing information. This may involve direct communication with the repository operator. The procedure includes quality assurance. The repository description is therefore revised by a second person, ensuring that only relevant records will be published by the service.

To support automation and further use within other works and services, the records are provided in a machine readable format and can be accessed and crawled via a RESTful-API.



Experiences and best practices

The previous two chapters gave a short overview on the motivation and functional description of the project's components. This chapter attempts to describe some best practices and experiences gathered by the project members. The professional context of the project members and their work on re3data.org provide the basis for some lessons learned:

Join forces

re3data.org has merged with DataBib [3], a similar project started around the same time. It is currently operated and driven by the Berlin School of Library and Information Science, GFZ German Research Centre for Geosciences, KIT Library and Purdue University Libraries. By the end of the year 2015 it will become a service under the auspices of DataCite [4].

Despite increased organizational overhead, required time to build up trust and align work flows, the combined efforts have turned out to be a good boost for accomplishing tasks.

- Prevents doing the same work twice, therefore it saves limited resources
- Results and experience gathered by one project member can be shared, resulting in a better overall service
- Both, the project and the repositories gain more visibility thanks to the international character of the collaboration

These benefits are not specific to research data and are likely to be achieved in other public projects, as well.

Technical infrastructure

Building a standalone service with human readability in mind is a common task. Nevertheless, defining a data structure is always a trade-off. re3data.org has chosen to define an own schema when the project launched, since no other schema available described research data repositories in such detail. The developed schema meets the current requirements well.

However, when it comes to sharing data with others in machine readable formats, both sides have to agree on it. The desired interfaces, formats and vocabularies are as heterogeneous as the research data itself. Some users, for instance, may be capable of crawling web resources and asking for RDF, but others might expect excel files for easy processing.

The general guideline to keep it small and simple certainly applies. In addition, the provision of multiple interfaces, which enable most users to access or deposit data, is found beneficial.

Raise awareness and clarify responsibilities

Raising awareness among researchers is considered the primary step in promoting the publication of research data. Publishing data affects many aspects ranging from the technical infrastructure over legal constraints to things like patents. The involved parties should therefore work together on data management plans. This will clarify rights and responsibilities from the beginning and prevent misunderstanding or implicit expectations, therefore it will help to

ensure publication. As good practice, scientists authorize the publication of the data and ensure its correctness, but service units at universities like libraries, computer centers or legal departments are responsible for the handling and processing within specific environments.

This good practice of data management and sharing needs to be accepted as part of good scientific practice. In addition, it needs to be communicated as such within the scientific community, as well as among students.

References

- [1] Organisation for Economic Co-operation and Development (OECD) (2007) Principles and Guidelines for Access to Research Data from Public Funding. Paris. Available: <http://www.oecd.org/dataoecd/9/61/38500813.pdf>
- [2] Pampel H, et al. (2013) Making Research Data Repositories Visible: The re3data.org Registry. PLoS ONE 8(11): e78080. doi:10.1371/journal.pone.0078080
- [3] Merger of Databib and re3data.org, first version of API available, Available <http://www.re3data.org/2015/03/merger-and-first-version-of-api/>
- [4] DataCite, re3data.org, and Databib Announce Collaboration, <https://www.datacite.org/news/datacite-re3dataorg-and-databib-announce-collaboration.html>
- [5] Vierkant, et al. (2014). Schema for the Description of Research Data Repositories. Version 2.2. doi: 10.2312/re3.006