The Landscape of Research Data Repositories in 2015

A re3data Analysis

Maxi Kindling maxi.kindling@hu-berlin.de [1], Heinz Pampel heinz.pampel@gfz-potsdam.de [2], Stephanie van de Sandt stephanie.van.de.sandt@cms.hu-berlin.de [1], Jessika Rücknagel ruecknagel@sub.uni-goettingen.de [1], Paul Vierkant paul.vierkant@gfz-potsdam.de [2], Gabriele Kloska gabriele.kloska@kit.edu [3], Michael Witt mwitt@purdue.edu [4], Peter Schirmbacher schirmbacher@hu-berlin.de [1], Roland Bertelmann roland.bertelmann@gfz-potsdam.de [2], and Frank Scholze frank.scholze@kit.edu [3]

[1]*Humboldt-Universität zu Berlin,*
Berlin School of Library and Information Science (BSLIS), Germany
[2]*GFZ German Research Centre for Geosciences,*
Section 7.4 Library and Information Services (LIS)
[3]*Karlsruhe Institute of Technology (KIT), KIT Library, Germany*
[4]*Purdue University Libraries, West Lafayette, Indiana, USA*

Abstract

This article provides a comprehensive descriptive and statistical analysis of metadata information on 1,381 research data repositories worldwide and across all research disciplines. The analyzed metadata is derived from the re3data database, enabling search and browse functionalities for the global registry of research data repositories. The analysis focuses mainly on institutions that operate research data repositories, types and subjects of research data repositories (RDR), access conditions as well as services provided by the research data repositories. RDR differ in terms of the service levels they offer, languages they support or standards they comply with. These statements are commonly acknowledged by saying the RDR landscape is heterogeneous. As expected, we found a heterogeneous RDR landscape that is mostly influenced by the repositories' disciplinary background for which they offer services.

Keywords: Research Data Repositories, RDR, Statistical Analysis, Metadata, re3data, Open Science, Open Access, Research Data, Persistent Identifier, Digital Object Identifier, Licenses

1 Introduction

1.1 Research data repositories for open research data

The idea of Open Science is becoming increasingly important (Nielsen 2011; Bartling and Friesike 2014; OECD 2015). An essential part of Open Science is Open Access to research data (Pampel and Dallmeier-Tiessen 2014). By sharing research data and other research materials, third parties can assess scholarly knowledge based on these data. The availability of research data and research materials fosters transparency and trustworthiness in research processes as well as enabling the reuse of research data. Accurately generated and curated datasets can be reanalyzed to validate research findings or reused and repurposed to answer different research questions. If datasets are easily accessible, new discoveries are facilitated and duplicate work can be reduced (Simons and Richardson 2013). This may lead to the high economic benefit that open research data may have for research and public societies. Several studies expose this economic impact, which the curation and sharing of valuable research data produces (Beagrie and Houghton 2014; Houghton and Gruen 2014). Considering these benefits, research data have to be regarded as a core element of the scholarly record. It is for this reason that funding organizations established data policies to influence the data management practices of researchers receiving public funding (European Commission. Directorate-General for Research & Innovation 2016).

The term "research data" can mean many things. As Borgman stated correctly, "data is a difficult concept to define, as data takes many forms, both physical and digital" (Borgman et al. 2012). In Pampel, Vierkant, et al. (2013) we define digital research data as "a (descriptive) part or the result of a research process", covering all stages of research. "Digital research data occur in different data types, levels of aggregation and data formats, informed by the research disciplines and their methods." (Pampel, Vierkant, et al. 2013)

To share or reuse research data, researchers and other interested parties need to be able to find information about datasets and the respective source making research datasets available. In this respect research data repositories (RDR) greatly contribute to realize the sharing and reuse of research data. Policies and recommendations for proper research data management stipulate that research data should be made available in an appropriate RDR to comply with the principles of research integrity (Pampel, Vierkant, et al. 2013; The Royal Society 2012). A RDR is a technical and organizational information system that helps researchers to manage, store and provide their own datasets, and to easily find and access datasets from other sources. Different approaches exist for operating a RDR. It can be described as "a subtype of a sustainable information infrastructure which provides long-term storage and access to research data" (Rücknagel et al. 2015). RDRs are an essential part of the research infrastructure of different "facilities, resources and related services used by the scientific community to conduct top-level research in their respective fields." (European Commission 2016)

At the same time, it is not easy to find an appropriate RDR for storing or reusing datasets. Marcial and Hemminger (2010) conducted a websample study of 100 RDR, analyzing information which is presented on the respective webpages. As they stated "there are many differences in the size, types, and organizations" of RDR (Marcial and Hemminger 2010), in part with differing scopes of services. It is ambiguous how many RDR are operated, but a tendency of growth can be assumed. Thus, a heterogeneous (and constantly changing) landscape of RDR can be described (Pampel, Vierkant, et al. 2013).

1.2 re3data — Registry of Research Data Repositories

re3data was a research project funded by the German Research Foundation (DFG) from 2012 until 2015 to create the Registry of Research Data Repositories called re3data. Project partners were the Library and Information Services (LIS) of the GFZ German Research Centre for Geosciences, the Library of the Karlsruhe Institute of Technology (KIT) and the Berlin School of Library and Information Science (BSLIS) at Humboldt-Universität zu Berlin. The project team developed the re3data metadata schema and provided a web interface to facilitate RDR search and browse functionalities. In 2014, the two major international registries for RDR — DataBib and re3data — joined forces and merged to one service, making the Library of the Purdue University a project partner. Since January 2016, re3data has been a service of DataCite (Brase et al. 2015) to ensure the registry's sustainable development.

The objective of our service re3data is to index and describe RDR so as to present detailed information about existing services. Being a starting point for researchers, funders, publishers, and other stakeholders to find and evaluate suitable services that support the management, storage, access and usage of research data such as research data repositories, portals and other service providers, re3data also helps different target groups to decide which RDR is appropriate for different purposes. The registry currently indexes over 1,821 RDR (as of 26 February 2017) with an extensive metadata description. Our approach for re3data is to provide extensive and quality-approved descriptions of RDR. For this reason, the "Metadata Schema for the Description of Research Data Repositories" is a comprehensive set of 42 properties (Rücknagel et al. 2015). The re3data description of a RDR, also referred to as a re3data metadata entry, provides the following information:

- General information about the RDR, such as the repository name, URL, disciplinary scope and a descriptive paragraph.
- Information concerning the responsible institutions of a RDR, such as the institution's name, type, location, and the type of responsibility.
- Legal issues including access and upload regulations as well as the availability of policies.

- Technical aspects such as information concerning supported persistent identifier systems, application programming interfaces or software in use if determinable.

All research data repositories listed in re3data are indexed and reviewed by our re3data editorial team. Up to the end of the funding period in December 2015, the editorial team consisted of project members from the partner institutions. During the funding period, we primarily checked other registries listing research data and other repositories to include them in re3data. The editorial team analyzes the website of a research data repository thoroughly using an internal handbook that provides practical information on how to obtain the metadata properties of the re3data schema. All of the information gathered was then reviewed by a second editor to improve the quality of the metadata entries provided on the website re3data.org. We are currently planning to enlarge the editorial team led by DataCite, to enable the indexing of repositories which do not use English as the interface language.

Every re3data record is persistently accessible and citable via a Digital Object Identifier (DOI). The registry offers a suggestion form to add RDR to re3data. By using the form, users and repository managers can provide detailed information about repositories that have not yet been indexed. Furthermore, re3data enables machine access to the registry via an Application Programming Interface (API).

Funding agencies such as the European Commission (Tarazona Rua et al. 2015) and the National Science Foundation (National Science Foundation. Directorate for Biological Sciences 2015) include references to re3data in their guidelines and policies related to data management and sharing. Additionally, several publishers and journals such as Copernicus Publications, PeerJ, PLOS ONE and Nature's Scientific Data recommend re3data in their editorial policies as a tool for authors to deposit data that support research findings published in their journals.

2 Methods

By analyzing the re3data metadata entries, we want to reveal the state of the art within the field of RDRs by the end of 2015. Furthermore, the findings presented in this analysis are meant to provide initial information on the RDR landscape to help identify areas requiring improvement. We explore the landscape of RDR focusing on the following research questions:

- Which types of RDR can be identified?
- What is the institutional background of the RDR covered?
- What are the disciplines and content types for which RDRs offer services?
- How do RDRs differ in terms of technical standards and access regulations?

To answer these research questions, we conducted a quantitative data analysis of a total of 1,381 RDR metadata entries in the re3data database at the end of the project funding phase, as of 3 December 2015. All metadata entries included in this database dump were indexed according to the "re3data.org Schema for the Description of Research Data Repositories" Version 2.2 (Vierkant et al. 2014) or below. re3data does not necessarily contain a representative sample of all RDR worldwide at that time. It is important to note that the schema was adjusted during the indexing process. Further the majority of the RDRs are indexed by our editorial team by analyzing RDRs' websites. As a consequence of the re3data indexing process, we cannot assess whether a respective metadata entry contains all information applicable to the RDR. The analysis reflects the RDR landscape in the indexing period in re3data up to the end of 2015.

The re3data metadata entries support drawing conclusions at the repository level. For instance, we can make statements on the technical capabilities of RDR to provide research data of a special content type such as text documents, audio and video material as often stated on the RDRs' website. Accordingly this paper does not provide an investigation on the distribution of individual content types in research data repositories on the level of data sets and collections.

The re3data database snapshot was provided as an SQL (Structured Query Language) dump and imported into a PostgreSQL database. A number of SQL queries based on the research questions were used to extract the required information from the database into a csv format (comma separated values).

We used the unique internal database identifier of repositories ("re3data.repository_id") that served as foreign key within the relational database for the statistical analysis of the metadata entries. Most properties of a repository are modeled as 1:n relation in the re3data metadata schema, allowing the selection of more than one value (such as more than one "content type"). Thus multiple tuples of each identifier and the appropriate property could occur. In order to obtain a useful dichotomous system for the statistical analysis with SPSS (mainly used SPSS Version 22.0.0 Mac OS X on Mac OS X 10.7.5) we modeled the existence of repository characteristics with a binary representation, transposing the property values into new variables for each research data repository identifier (ID) with the help of a Python script. This form allowed us to perform analytical tests for correlations with SPSS Statistics.

The script matched rows that are dependent on the internal database identifier. If a value exists for an ID, it will receive a 1 as new value for this variable. For example if a RDR has "plain text" and "raw data" as values in the variable "content type", it receives a 1 for the variable "contenttype:plain text" and "contenttype:raw data", but a 0 as a value for all other "content type" variables. Moreover, data cleansing included to consolidate multiple equal values for one RDR to a single value (e.g. some IDs have several "other" certificates etc.). Hence, the parent population of data differs in some cases from the original database population. In the following text, we always refer to this cleansed dataset. We needed the raw quantities only in one case, which is explained at that point.

As the given data are dichotomous nominal values, our range of statistical methods was limited to nonparametric procedures. Apart from the presentation of numbers and relative occurrences of variables, we also identified correlations between nominal variables in the data, e.g. dependencies between "content type" and "subject". We chose contingency tables and chi-square distribution (χ^2) tests combined with Cramer's V value for this.

By using contingency tables we compare expected values in case of statistical independence calculated by SPSS with the observed values to reject the null hypothesis in order to find significant correlations. Mostly those correlations substantiate preliminary assumptions on the disciplinary influence on RDR. A number of correlation effects that passed the test on at least a 95% significance level are presented in this article. Unless otherwise stated, the presented correlations are significant on the 95 % level. The strength of dependency was tested by Cramer's V, whereby a value lower than 0.3 is considered to be a low effect size. Values equal to 0.3 and lower than 0.5 are considered to be medium effect (Cohen 1988).

The research data set consisting of five data tables, a matrix of all correlations and a research data documentation is publicly available (Sandt et al. 2017). The correlation matrix can be used as an overview on all results of our analysis, whereas this article only presents results that are of main importance. The research data documentation provides detailed information on the methods and data transformation and aims to help reproduce the data analysis.

3 Results

3.1 Institutional background

re3data provides information on the institutions that are responsible for developing, maintaining, funding, etc. RDR listed in re3data. The metadata schema differentiates between subcategories further describing the type of responsibility, whether an institution is profit or non-profit and the institution's country of origin. Institutions can have more than one type of responsibility for a respective RDR, whereas only the head quarter's country is indexed. The categories are explained within the metadata schema on page 20 (Vierkant et al. 2014).

3.1.1 Responsibility types of institutions

We list a total of 4,311 institutions that are responsible for the 1,381 registered RDR. One institution may occur multiple times and with differing names or organizational units. Each RDR may also have multiple responsible institutions (1:n relation). The re3data metadata schema distinguishes between "general", "funding", "technical" and "sponsoring" responsibility of institutions. Nearly half of the listed institutions have a general responsibility for the associated RDR (43.0 %). That means the particular institution is responsible for the content as well as for the management of the RDR. One third are funding institutions (33.4 %). Additionally, 21.7 % of the institutions are technical hosts while only 1.9 % of all institutions are RDR sponsoring institutions, meaning that funds are granted to a research data repository in exchange for advertising. About one third of all repositories (29.3 %) are related to institutions with more than one area of responsibility.

The distribution of responsibility types of institutions according to the total amount of 1,381 RDR indexed in re3data is shown in a Venn diagram (cf. Figure 1). It is important to note that occurrences of multiple responsibility types are reduced to one for each type. Most RDR (49.4 %) are operated by institutions that are responsible in general, technical and funding terms; 20.5 % are operated by technically and generally responsible institutions. Only 2.3 % of all RDR are operated by institutions of all types of responsibility.

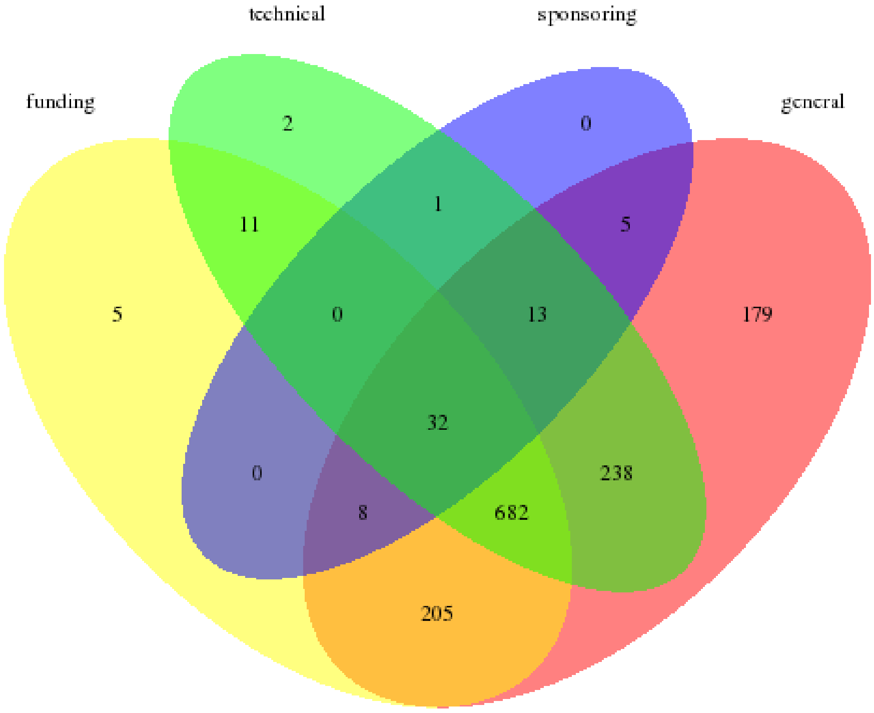

Figure 1: Responsibility types of institutions operating research data repositories indexed in re3data (n = 1,381)

3.1.2 Profit or non-profit institutions

Nearly all (96.6 %) of the above mentioned institution entries in re3data (n = 4,311) are non-profit organizations. Most of the repositories are funded solely by non-profit organizations, but hybrid forms also exist (4.9 % of all RDR have more than one institution type) as well as those that are profit institutions (2.1 %).

3.1.3 Breakdown by country

For the analysis of the institutions' countries, we focused on the absolute number of institutions for each RDR regardless of multiple values (a RDR might be related to several institutions from identical countries. For example a RDR can be supported by five institutions, all of them from the USA). We therefore refer to the uncleansed dataset.

Most institutions (n = 1,936) are from the USA. Germany (n = 521), Great Britain (n = 378) and Canada (n = 216) are also prevalent origins of institutions. Fifty-seven countries are represented below a 5 % level. International cooperation seems to be widely spread depending on the repository context, whereby the number of different countries per RDR varies widely between 1 and 22 (cf. Figure 2). Important to consider is that these findings depict only RDR indexed in re3data requiring an English graphical user interface (GUI) according to the re3data registration policy until the end of 2015.

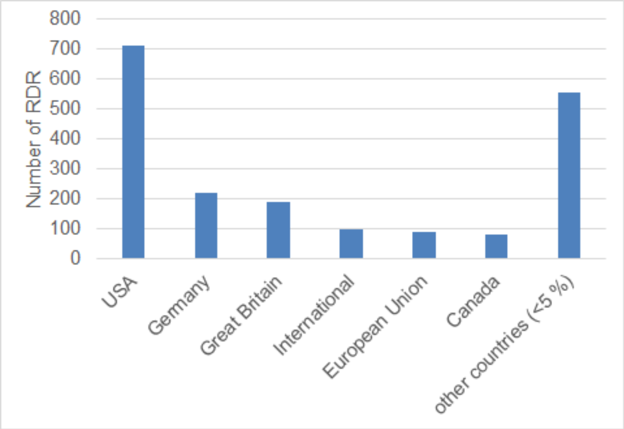

Figure 2: Countries of the responsible institutions operating research data repositories indexed in re3data (n = 1,381, multiple values possible)

3.2 Repository types

To describe the registered RDR in more detail, re3data collects some general information about the repository. This includes a description of the repository type. The re3data metadata schema distinguishes between "disciplinary", "institutional" and "other" repository types, defined in the metadata schema on pages 18 and 19 (Rücknagel et al. 2015).

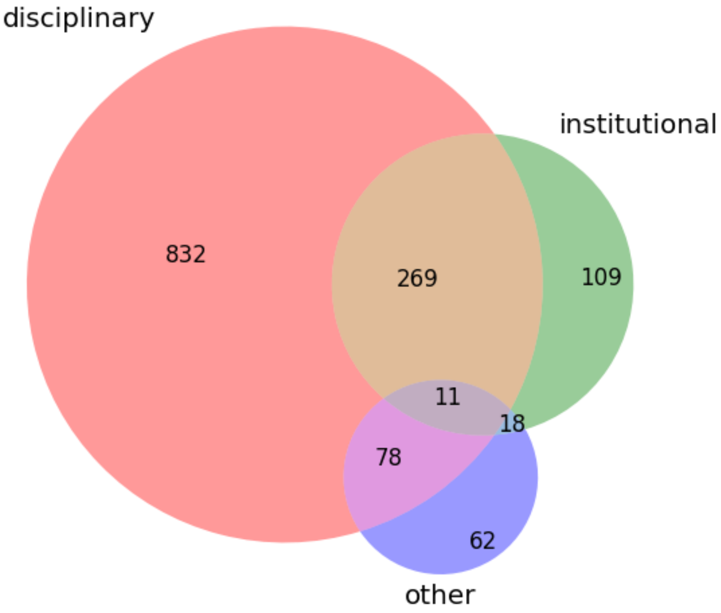

Figure 3: Types of research data repositories indexed in re3data (n = 1,379, 2 RDR with missing values, multiple values possible)

As shown in the Venn diagram above the majority (86.2 %) of the registered RDR are "disciplinary". Also, 29.5 % of the repositories are "institutional" and 12.2 % fulfill the "other" criteria. This "other" category includes, among others, portals or commercial data storage services. Further, 0.8 % of the repositories fulfill all categories, like the "Australian Ocean Data Network Portal" (*re3data.org: Australian Ocean Data Network Portal* 2014) (cf. Figure 3).

With 19.5 % there is a high number of RDR that are described as being institutional as well as disciplinary. Of all RDR, 72.9% can be clearly categorized into one type: 109 institutional RDR, 832 disciplinary RDR, and 62 other RDR.

3.3 Subjects and content

3.3.1 Subjects

re3data implemented the classification of subject areas of the German Research Foundation (Rücknagel et al. 2015). The following Venn diagram (cf. Figure 4) shows the distribution of registered RDR according to the four main categories "Humanities and Social Sciences" (SSH), "Life Sciences", "Natural Sciences" and "Engineering Sciences". A RDR can offer research data and services for more than one subject.

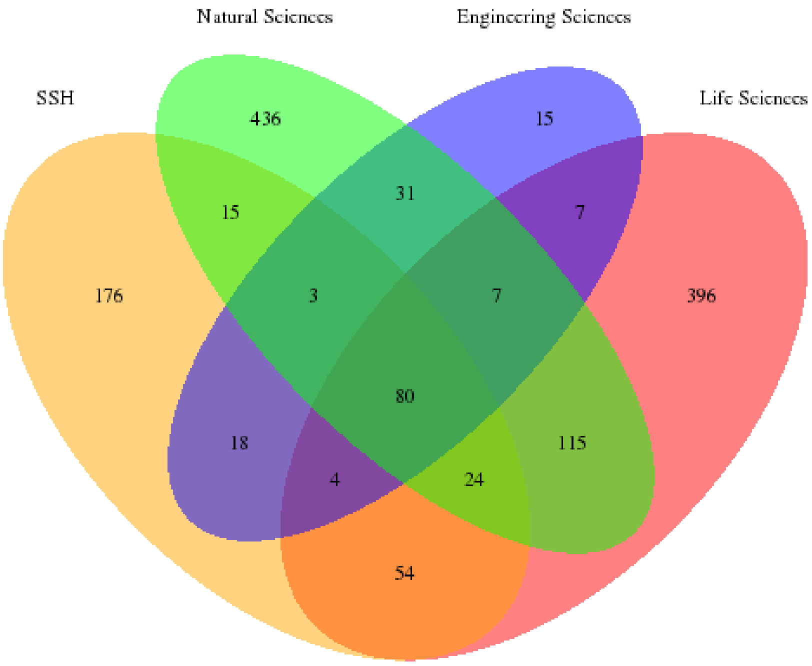

Figure 4: The four main subject categories of research data repositories indexed in re3data (n = 1,381, multiple values possible)

Only 5.8 % of 1,381 RDR cover all four subject areas, Humanities and Social Sciences, Natural Sciences, Life Sciences and Engineering Sciences, and can be described as being multidisciplinary RDR. Most RDR have a disciplinary focus on one subject area (mean = 1.4, median = 1.0, standard deviation = 0.8). RDR covering Natural Sciences (51.5 %) and Life Sciences (49.8 %) are the most frequent, while RDR covering Humanities and Social Sciences are represented with 27.1 % and RDR covering Engineering Sciences with 12.0 % (cf. Figure 4).

3.3.2 Content types

re3data differentiates between 15 content types according to the PARSE.insight survey (PARSE.insight team 2016) that are represented in the following table (cf. Table 1). All 15 content types are covered by "Research Data Australia" (*re3data.org: Research Data Australia* 2015), "Europeana" (*re3data.org: Europeana* 2015) covers 14 categories.

re3data content type	Number (n = 6,340)	Number (n = 1,381)
	Count (n)	Percentage (%)
Scientific and statistical data formats	881	63.8
Standard office documents	786	56.9
Plain text	690	50.0
Images	686	49.7
Raw data	586	42.4
Structured graphics	541	39.2
Structured text	490	35.5
Other	446	2.3
Archived data	339	25.5
Software applications	258	18.7
Audiovisual data	253	18.3
Databases	204	14.8
Networkbased data	104	7.5
Source code	49	3.5
Configuration data	27	2.0

Table 1: Content types of research data repositories indexed in re3data (n = 1,381, multiple values possible)

3.3.3 Content types according to the main subject categories

In the following we differentiate RDR content types according to the four main subject categories. The mean value of the content type variety shown according to the four main subject categories is 4.59 different types of content per RDR (median = 4.00, standard deviation = 2.13). Research data in Humanities and Social Sciences' RDR contain more "Standard office documents", "Plain text" or "Images" than expected. The number of occurrences is only significantly higher (n = 218) than the expected value (n = 186.9) for "Plain text". The test for independence showed a significant dependency between the variables "contentType" and "subject" with a low effect size (Cramer's V = 0.101). The existence of "Images" and the subject Humanities and Social Sciences correlate with a low effect size (Cramer's V = 0.156). The observed value (n = 185.8) is much higher than was expected (n = 138) on a 99.9 % level. "Standard office documents" are clearly overrepresented in this discipline (observed n = 276; expected n = 212.9), but also with a low effect size (Cramer's V = 0.208). For all cases mentioned in this paragraph, the test indicated a dependency on 99.9% level.

Life Sciences' repositories cover more "other" types of content than other disciplines (observed n = 252 instead of expected n = 221.9). The dependency is significant around 99 %, but with a very low effect size (Cramer's V = 0.093). We did not find any statistically significant dependency on other content types, though we observed more "Structured text", "Software", "Scientific and statistical data formats" and "Structured graphics" than was statistically to be expected.

"Images" as a content type largely depends on the discipline. There are significantly more "Images" in Natural Sciences' RDR than expected (observed n = 429; expected n = 353.2). The test is significant on 99.9 % level, though the effect size is low (Cramer's V = 0.220). The dependency between Natural Sciences and "Raw data" (observed n = 357, expected n = 301.7) reflects the same probability, even though the effect size is just as low (Cramer's V = 0.162). The value of 479 "Scientific and statistical data formats" was higher than expected (n = 453.6). The dependency is on a 96 % level with a minor effect (Cramer's V = 0.077).

In Engineering Sciences more research data repositories cover "Audiovisual data" (observed n = 67) than expected (n = 30.2; Cramer's V = 0.212). There is also significantly more "other" content than expected (observed n = 74 instead of expected n = 53.39 on 99.9 % level and Cramer's V = 0.099).

3.4 Policies

Of the 1,381 RDR, 85.4 % provide at least one policy document of some kind. A policy document is a document that expresses a guiding framework for a RDR regulating different aspects of the implementation or operation of the repository (Martín and

Ballard 2010). Thus, different thematic issues and priorities can be depicted. The property "policyType" did not exist in version 2.2 of the re3data metadata schema, on which this metadata analysis is based. It was integrated in the metadata schema version 3.0. Before this, only the names of the respective policies were collected.

Since numerous names are used for policy documents, we cannot provide a quantitative analysis revealing the most popular policy types. Based on the observations of the policies' content, and with regard to the team's knowledge of the repository landscape, we gathered the following policy types "Access policy", "Collection policy", "Data policy", "Metadata policy", "Preservation policy", "Submission policy", "Terms of use", "Usage policy" and "Quality policy". In order to provide information concerning conditions of use we strongly encourage RDR to have at least one policy in place. It should cover general issues including the aspects in table 2 (cf. Table 2, following Australian National Data Service 2010). The policy types mentioned above are defined in the re3data metadata schema 3.0 on page 21 (Rücknagel et al. 2015).

Policy aspect	Details
Name of the policy	E.g. "Policy of the RDR"
Purpose of the document & key principles	E.g. commitment to long term preservation
Application of the policy	E.g. user of the website or person uploading research data; responsibilities
Licensing and copyright matters	E.g. obligation or recommendation of standard licences
Access & usage regulations	Embargo period, restrictions, privacy issues
Retention period	Removal of datasets
Formal	Dates (commencement, review, versioning), contact information, links to other relevant documents, glossary

Table 2: Policy aspects of research data repositories

3.5 Access and Licenses

re3data distinguishes between "open", "restricted" and "closed" as access categories with respect to the following levels: to the RDR ("database"), to the research data itself ("data") and to the data submission services ("upload") (Rücknagel et al. 2015). Each level can be "open", "restricted", or "closed". If the repository, research data and submission services can be accessed without financial and technical barriers, it means that the value for the respective property is "open". Access barriers to the RDR and its services that a user can overcome are "restricted", e.g. a user account needs to be created or an agreement has to be signed by the user to obtain access to the RDR, the data sets or to submit research data. "Closed" access means that an external user cannot overcome access barriers, e.g. if a service is solely for a respective community that he/she cannot become a member of. The access type "embargoed" is also used for research data. "Embargoed" research data cannot be accessed by third persons until the data sets have been released for "open" or "restricted" access (Rücknagel et al. 2015). Each RDR can have several access values for each level since, e.g., parts of data collections can be accessed openly and other parts may be restricted.

3.5.1 Access to the RDR and database licenses

The vast majority of all RDR (n = 1,381) are openly accessible (95.5 %). A few RDR have access restrictions to the database (4.3 %) while only 2 RDR (0.1 %) are closed (cf. Figure 5).

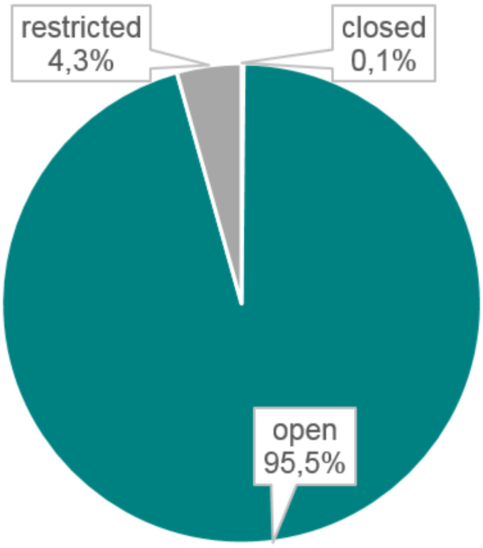

Figure 5: Access to the database of research data repositories indexed in re3data (n = 1,381)

Three hundred forty-two database licenses are mentioned that belong to 336 RDR. Six RDR make use of more than one license. Despite the fact, that we are thus unable to make any statements on the use of license information for most RDR, we do know that copyright information is the most common, followed by "other" license information and the use of one license type from the Creative Commons license family. All other license information does not seem to be widespread at all. The distribution of all license information is shown in the table below (cf. Table 3).

Database license	Number (n = 342)	Number (n = 1,381)
	Count (n)	Percentage (%)
Copyrights	139	10.1
Other (e.g. a policy to clarify legal aspects)	114	8.3
CC (Creative Commons license family)	51	3.7
Apache License 2.0	19	1.4
ODC (Open Data Commons)	8	0.6
BSD (Berkeley Software Distribution)	6	0.4
Public Domain	4	0.3
CC0 (Creative Commons Public Domain Dedication)	1	0.1

Table 3: Database licenses found in research data repositories indexed in re3data (n = 342, multiple values possible)

3.5.2 Access to the research data

As shown in the Venn diagram below, most RDR offer "open" access to the research data at least partly (86.2 %). However, 45.8 % of the RDR provide "restricted" or at least partly "restricted" access, 7.5 % "closed" access and 0.8 % "embargoed" access. RDR provide research data in all four access categories (cf. Figure 6). Overlapping categories in the Venn diagram such as RDR that are "open" and "restricted" at the same time mean that a RDR can be accessed openly in parts (e.g. "Australian Data Archive") (*re3data.org: Australian Data Archive* 2015).

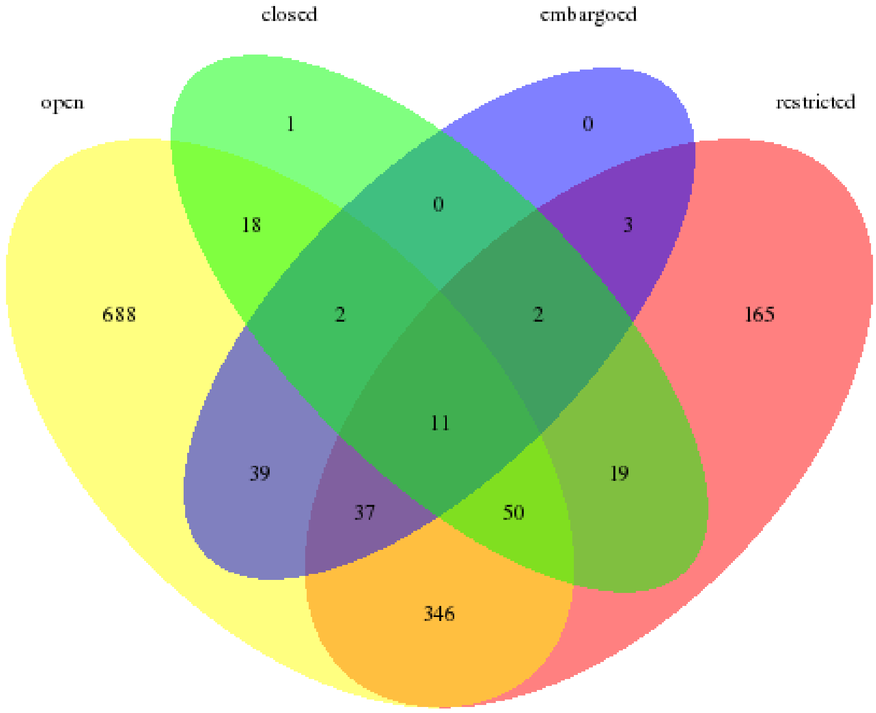

Figure 6: Access to the research data of repositories indexed in re3data (n = 1,381, multiple values possible)

A correlation of access to the research data and the aforementioned four subject categories shows a low effect size on "closed" access (Cramer's V = 0.156) for all RDR covering Humanities and Social Sciences. More RDR (n = 53) provide research data in "closed" access than was expected (n = 27.9). This also becomes obvious when comparing the relative numbers of "Humanities and Social Sciences" with all other disciplines. One effect of this is that there are significantly fewer "open" data access repositories (observed n = 302) than statistically expected (n = 322.5). The effect size is 0.097 (Cramer's V). In addition, more RDR offer "restricted" data access (observed n = 242 instead of expected n = 171.49). The effect size is a little stronger here (Cramer's V = 0.231).

3.5.3 Access restriction to the research data

For 839 RDR, re3data provides metadata information concerning access restriction to the research data. "Restricted" data access includes restrictions such as "registration" (33.9 % of 1381), "other" (18.3 %, e.g. request data sets via email) , "institutional membership" (1.2 %), and "feeRequired" (7.4 %) (cf. Figure 7).

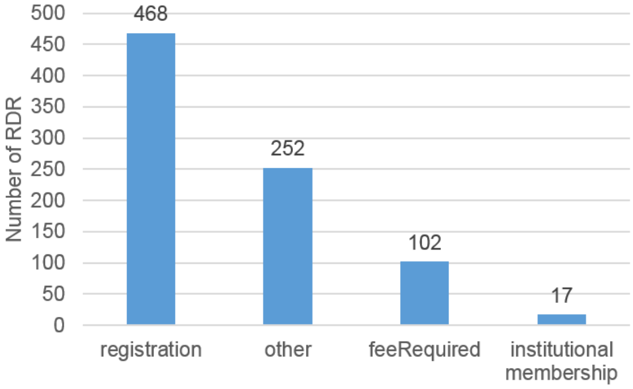

Figure 7: Access restrictions to the research data of repositories indexed in re3data (n = 839)

For RDR covering Life Sciences, a significance was proven for "embargoed" data access with an effect size of Cramer's V = 0.105. More repositories offering "embargoed" data access (n = 65) than expected (n = 46.8) may indicate that scientists in these fields require more time to analyze their data before allowing third parties to access the research data. Another reason may be the need to protect research data for patent submissions. Slightly more "open" data access repositories (observed n = 616, expected n = 592.5) are indicated in Life Sciences' fields that already have an established open access and data sharing culture to facilitate scientific progress. The effect is low though (Cramer's V = 0.099). There are fewer "restricted" RDR in Life Sciences (n = 284, expected n = 314.9) on nearly the same effect size (Cramer's V = 0.090).

Engineering Sciences has an effect (Cramer's V = 0.113) on "embargoed" data access RDR (n = 24, expected n = 11.2). No dependencies above the 95 % significance level were identified for Natural Sciences.

3.5.4 Research data licenses

There are a total of 2,122 metadata entries on research data licenses for all unique 1,381 RDR. As some RDR have more than one "other" license, the cleansed data only counts 1895 licenses. The table below shows the distribution of different kinds of legal use and license information RDR provide to access and use research data. The largest group of data licenses within re3data relate to individual license information prepared by the RDR on regulations for use. These licenses are recorded within the "Other" category (57.2 %). Copyright information is provided by 38.6 % of the 1,381 RDR. The Creative Commons license family is the most frequently used (21.8 %). Research data is declared to be "public domain" less often (14.2 %), while "ODC" (2.0 %), CC0 (1.9 %), OGL (0.9 %) were almost negligible (cf. Table 4).

Research data license	Number (n = 1895)	Number (n = 1381)
	Count (n)	Percentage (%)
Other	790	57.2
Copyrights	553	38.6
CC (Creative Commons license family)	301	21.8
Public Domain	196	14.2
ODC (Open Data Commons)	27	2.0
CC0 (Creative Commons Public Domain Dedication)	26	1.9
OGL (Open Government License)	12	0.9
BSD (Berkeley Software Distribution)	6	0.4
Apache License 2.0	2	0.1
OGL C (Open Government License)	1	0.1
RL (Restrictive License)	1	0.1

Table 4: Research data licenses found in research data repositories indexed in re3data (n = 1,895, multiple values possible)

The effect size on using a "Creative Commons" license in RDR covering Humanities and Social Sciences is high (Cramer's V = 0.132). We observed 115 RDR using "Creative Commons" licenses (expected n = 81.5). Sixteen RDR (instead of the 7 expected) use "CC0" for their Humanities and Social Sciences research data, though there is only a low correlation (Cramer's V = 0.107). There is also a low effect size on the "BSD" data license (Cramer's V = 0.084) for the Humanities and Social Sciences because more RDR (n = 5) use these than statistically expected (n = 1.6).

In Life Sciences, there is again a tendency towards an established culture of data sharing. This is depicted by a significantly (Cramer's V = 0.152) higher usage of "Creative Commons" licenses (observed n = 193, expected n = 149.7) and slightly more research data repositories make use of "Public Domain" (observed n = 121) than expected (n = 97.5), with a low effect size (Cramer's V = 0.098).

Significantly more "Creative Commons" licenses (observed n = 67, expected n = 36, Cramer's V = 0.168), "Apache" licenses (observed n = 2, expected n = 0.2, Cramer's V = 0.103) and "BSD" licenses (observed n = 5, expected n = 0.7, Cramer's V = 0.145) were found for research data repositories in Engineering Sciences.

3.5.5 Access and restriction to research data upload

The majority of RDR restricts data upload (57 %). Slightly fewer RDR are "closed" in terms of data upload (39.5 %). This means that no external data submissions are accepted for inclusion in the data collection of the RDR. "Closed" data uploads are found for RDR that collect project-specific research data or institutional output in particular. Only 3.9 % offer an open data upload, meaning that there is no registration or contact information required to submit research data to the RDR (cf. Figure 8). There are rare overlapping values indicating, e.g., that access to upload is partly open or that there are several options with varying restrictions for submitting data to a repository (e.g. "NCBI Genome repository") (*re3data.org: NCBI Genome repository* 2015).

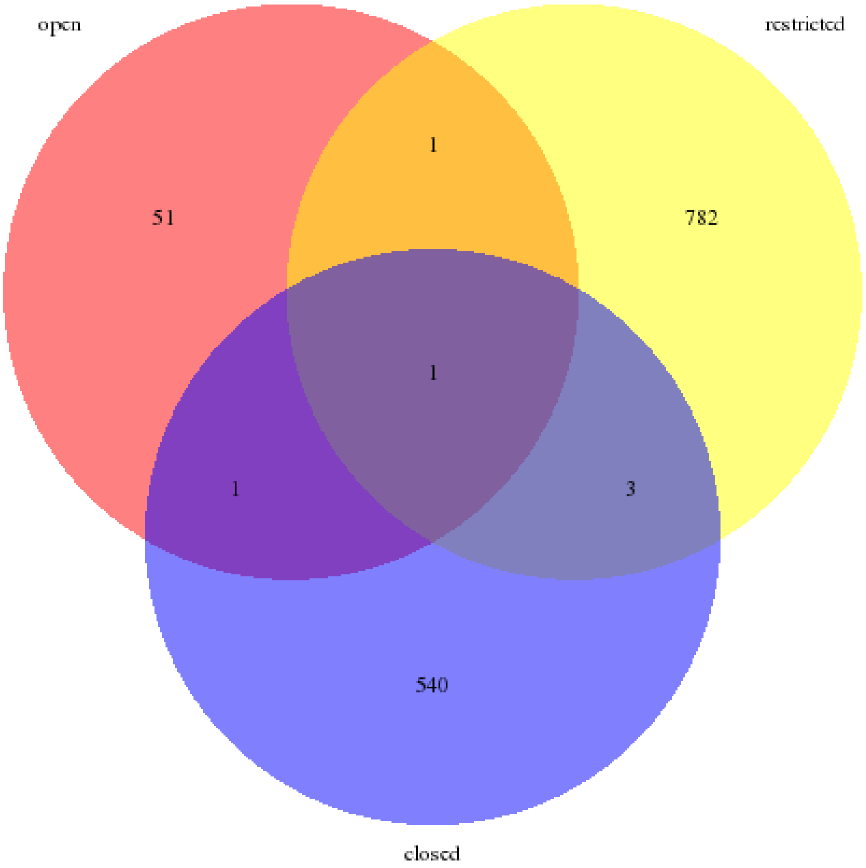

Figure 8: Research data upload access to the repositories indexed in re3data (n = 1,379, 2 RDR with missing values, multiple values possible)

The data upload is likely to happen by "registration" (37.6 %) or "other" mechanisms, such as sending data media by postal mailing (16.0 %). Only 4.1 % of the 1,381 RDR provide the upload services solely to "institutional members", and only 0.7 % require "submission fees" (cf. Table 5). The number of RDR in re3data missing information on research data upload restriction in general or on the specific type of upload restriction is 576. Specifying the type of upload restriction was not required in the indexing process.

Research data upload restriction	Number (n = 805)	Number (n = 1,381)
	Count (n)	Percentage (%)
Registration	519	37.6
Other	221	16.0
Institutional Membership	56	4.1
Fee required	9	0.7

Table 5: Research data upload restrictions found in research data repositories indexed in re3data (n = 805, multiple values possible)

3.6 Services

3.6.1 Persistent Identifier systems

PID system	Number (n = 1,421)	Number (n = 1,381)
	Count (n)	Percentage (%)
None	924	66.9
DOI (Digital Object Identifier)	275	19.9
Handle	102	7.4
Other	77	5.6
PURL (Persistent Uniform Resource Locator)	16	1.6
URN (Uniform Resource Name)	16	1.6
ARK (Archival Resource Key)	11	0.8

Table 6: Persistent Identifier systems used by research data repositories indexed in re3data (n = 1,421, multiple values possible)

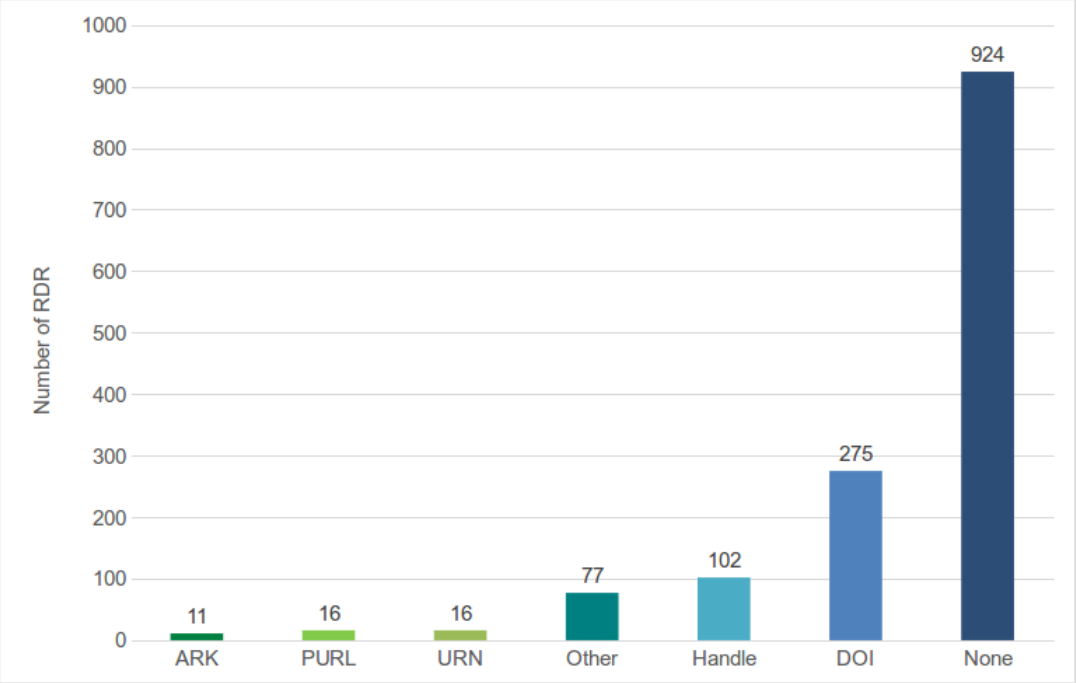

Figure 9: Persistent Identifier systems used by research data repositories indexed in re3data (n = 1421, multiple values possible)

Only 36 % of 1381 RDR use a Persistent Identifier (PID) system to ensure the persistent citability of the provided research data. For 46 RDR we found more than one PID system applied. The most common system is the "Digital Object Identifier (DOI)" system, with 19.9 % RDR in re3data using it. DOI is used widely in all four subject categories (as shown in the correlation matrix) (Sandt et al. 2017). Less than half as many RDR offer the "Handle" system (7.4 %). Least popular are the "URN" (1.6 %) and "ARK" systems (0.8 %) (cf. Table 6, Figure 9).

3.6.2 Repository software

Repository software	Number (n = 1,166)	Number (n = 1,381)
	Count (n)	Percentage (%)
Unknown	765	55.4
Other	271	19.6
DSpace	36	2.6
DataVerse	30	2.2
CKAN (Comprehensive Knowledge Archive Network)	16	1.2
Fedora	14	1.0
EPrints	11	0.8
MySQL	10	0.7
Nesstar	10	0.7
dLibra	2	0.1
eSciDoc	1	0.1

Table 7: Repository software used by research data repositories indexed in re3data (n = 1,166, multiple values possible)

We can make statements on the software used to run the RDR for 396 repositories of 1,381. The small number of registered software types within re3data is most certainly due to our manual indexing process. In most cases the repository software is not explicitly stated on the website and as a consequence not indexed. Furthermore, it is important to know that the "unknown" value was not part of the Metadata Scheme in the first version 1.0. Thus re3data includes RDR with unknown software that have no metadata information as well as RDR with the value "unknown" from the later indexing process. What we can say is that from the captured repository software, DSpace and DataVerse are the most prevalent followed by CKAN, Fedora and Eprints, MySQL and Nesstar. Combinations of different software solutions were rarely observed (0.4 %). The majority of RDR use another type of software (19.6 %) that is not explicitly listed in the controlled vocabulary of the metadata schema version 2.2 (cf. Table 7).

3.6.3 Application Programming Interfaces (API)

API	Number (n = 830)	Number (n = 1,381)
	Count (n)	Percentage (%)
FTP (File Transfer Protocol)	284	20.6
Other	178	12.9
REST (Representational State Transfer)	164	11.9
OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting)	85	6.2
SOAP (Simple Object Access Protocol)	52	3.8
SWORD (Simple Web-service Offering Repository Deposit)	21	1.5
NetCDF (Network Common Data Format)	20	1.4
OPeNDAP (Open Source Project for a Network Data Access Protocol)	16	1.2
SPARQL (SPARQL Protocol and RDF query language)	10	0.7

Table 8: APIs used by research data repositories indexed in re3data (n = 830, multiple values possible)

We found 608 (44 %) RDR that provide information concerning APIs for data and metadata exchange. The "Alaska Ocean Observing System" (*re3data.org: Alaska Ocean Observing System* 2015) repository offers a total of seven "APIs" as an extreme. The average number of APIs per RDR is 1.49 (median = 1.00, standard deviation = 0.84). Most RDR (20.6 %) offer data and metadata exchange via "FTP", 12.9 % via "other" APIs, followed by "REST", "OAI-PMH", and "SOAP" (cf. Table 8). Similar to the "software type" property, our editorial team experienced difficulties collecting information concerning API systems in use by RDR. Therefore we assume that the depicted information is biased by the fact that some interfaces are easier to identify by our editors, such as the FTP API.

Apart from this fact, we found subject dependencies for the distribution of the API types. We observed far fewer "FTP" APIs for RDR in Humanities and Social Sciences (n = 22) than expected (n = 76.9) on a medium low correlation level (Cramer's V = 0.221), but a lot more "OAI-PMH" APIs (observed n = 62 instead of expected n = 23.0). The correlation strength is slightly stronger here (Cramer's V = 0.264). "SOAP" is rarely used in this subject area (observed n = 3 instead of expected n = 14.1; Cramer's V = 0.095).

A "REST" API seems to be significantly popular among Life Sciences (observed n = 104 instead of expected n = 81.6; Cramer's V = 0.100). We observed the same popularity for a "NetCDF" API (observed n = 20 instead of expected n = 10.3; Cramer's V = 0.118) and "OpenDAP" (observed n = 16 instead of expected n = 8.2; Cramer's V = 0.105) for Natural Sciences RDR.

Just as we had fewer "FTP" APIs in Humanities and Social Sciences, the same phenomena could be observed in Engineering Sciences RDR (observed n = 10 instead of expected n = 33.9; Cramer's V = 0.132). Both subject classes also show the same behavior for "OAI-PMH". Engineering Sciences offer 27 "OAI-PMH" APIs instead of the statistically expected 20.2 if both variables were independent (Cramer's V = 0.156).

3.6.4 Certificates

Certificates	Number (n = 261)	Number (n = 1,381)
	Count (n)	Percentage (%)
Other	109	7.9
WDS (World Data System)	72	5.2
DSA (Data Seal of Approval)	52	3.8
CLARIN certificate B (Common Language Resources and Technology Infrastructure)	14	1.0
RatSWD (German Data Forum)	6	0.4
TRAC (Trustworthy Repositories Audit & Certification)	3	0.2
DINI Certificate (Deutsche Initiative für Netzwerkinformation e.V. — DINI-Zertifikat)	2	0.2
DIN 31644 (Deutsches Institut für Normung German Institute for Standards)	1	0.1
ISO 16363 (International Standard Organisation)	1	0.1
Trusted Digital Repository	1	0.1

Table 9: Certificates used by research data repositories indexed in re3data (n = 261, multiple values possible)

Two hundred sixty-one of 1,381 RDR (18.9 %) have been awarded a certificate according to the re3data metadata. The metadata schema offers 9 options for known certificates or standards and the "other" category. Only one RDR has a maximum of three certificates (*re3data.org: UK.Data Archive* 2015). Most RDR only have one certificate, the mean value is 1.14 (median = 1.00, standard deviation = 0.36). The World Data System ("WDS") certificate is awarded most often (5.2 %, n = 72), followed by 52 RDR (3.8 %) that are compliant with the "Data Seal of Approval".

Only 109 RDR (7.9 %) offer information on any "other" kinds of certification and standards compliance (cf. Table 9).

With respect to the RDR's subjects, we can prove a low effect for Humanities and Social Sciences (Cramer's V = 0.166) with 14 observed "CLARIN certificate B" (expected n = 3.8). Since the "CLARIN certificate B" clearly focuses on Humanities and Social Sciences' disciplines, we consequently have no further RDR from another subject that is compliant. The effect on the Data Seal of Approval ("DSA") for Humanities and Social Sciences is average (0.307). Here we identified 50 certificates.

In contrast to Humanities and Social Sciences, Life Sciences have a certificate significantly less often (effect size is 0.167).

4 Key findings, interpretation and discussion

4.1 General observations

As expected, we found a heterogeneous landscape of RDR that is mostly influenced by the repositories' disciplinary background for which they offer services. Most repositories are disciplinary services (cf. Figure 3). We identified statistically significant dependencies in connection with the four main subject categories Humanities and Social Sciences, Natural Sciences, Life Sciences, and Engineering Sciences for a number of properties.

4.2 Access

Openness seems to be a common factor because the vast majority of the RDR grant Open Access to their databases and at least partly to dataset collections (cf. Figures 6 and 7). Research data repositories in the Life Sciences group appear to be most willing to share data sets openly. The re3data analysis provides a first overview of access with respect to information at the repository level. Before any assumptions can be made on the distribution of open research data datasets and collections itself, it's metadata needs to be investigated on the data level. The most frequent data license types according to re3data observed on the repository level are "other" and "copyright" (cf. Table 4). This means that self-written policy documents or copyright notices are mainly used to regulate access and usage conditions for the database and the content it provides. The dissemination of standard licenses, especially for research data and metadata, should be accorded even greater support. Although the Creative Commons License family is used by 21.8 % of the repositories (cf. Table 4), this number could be improved and needs to be investigated further e.g. in terms of license types in use. Creative Commons is an approved standard to clarify how content can be used, commonly in a digital form. Research data repository operators should consider adding appropriate standard licenses to their range of licenses a data submitter can choose from.

More than half of the RDR offer restricted access to upload research data through registration or other means, e.g. data upload via API or postal mailing (cf. Figure 8 and Table 5). The authentication of a data submitter seems to be a major concern for research data repository to restrict the upload processes. Hardly any repository listed in re3data is requiring a submission fee. There are two possible reasons for this: either very few RDR demand such a fee, or the information was not found during the indexing process. To clarify access and usage regulations and further determine the terms on which users can submit data sets to a repository's collection, it is of utmost importance to provide this information on the research data repository's website.

4.3 Services

Well-established standards and services in the field of text repositories are far from common within the RDR environment. Aspects such as the persistent citability of research data sets and collections are most important to realize an adequate reuse of research data sets. Consequently, the provision of PID systems needs to be expanded from this aspect, as do other issues. With respect to those repositories that use a PID system, the DOI and Handle system are the most widespread solutions to enable a persistent reference to research data. DOI and Handle together account for 27.3 % (cf. Table 6 and Figure 9). Over 50 % of all captured PID values are DOIs (55.3 %). This number indicates that DOI is a well-established standard. The foundation of

DataCite can be named as one possible reason for this development, as the well-known non-profit organization has been providing PIDs for data sets since 2009.

4.4 Repository software and API

As for the software solutions in use to run a research data repository, we realize that the software types used by the majority of repositories are hard to record since the information is not readily provided by all RDR. What we can indicate is that a large number of "other" software types were registered during the indexing process (cf. Table 7). This shows that several repositories have developed their own software solution. Presumably, some repositories have special needs that require a self-developed software tool or the existing (standard) software packages do not match the institutional research infrastructure development strategy.

Far more RDR than were recorded need to have (standardized) APIs to provide metadata for service providers and enable metadata search to improve findability of research data sets and collections. Enabling metadata aggregation is an important feature of appropriate repository software to improve discovery of research data.

4.5 Certificates

Around 19 % of all RDR have a certificate or comply with a standard. This number seems rather small despite the fact that re3data does not only list repositories with a long-term commitment to preserve data sets. Complex standards such as DIN and ISO in particular, which require a high implementation effort, can only be demonstrated for a few exceptions. The Data Seal of Approval is an approach that seems to be well on it's way to becoming a common "soft" standard for data curation (cf. Table 9). WDS is also widespread and ensures the sustainability of an information system, even though the certification process requires the membership of responsible institutions. Apart from these fundamental, quality-proven standards and audits that focus on the repository level, quality standards for the research data and metadata itself are crucial for the research community to guarantee that RDR are perceived as reliable information systems suitable for the indefinite storage of valuable research data sets and materials.

5 Conclusions

RDR differ in terms of the service levels they offer, languages they support or standards which they comply with. These statements are commonly acknowledged by saying that the RDR landscape is heterogeneous. By conducting an analysis of re3data metadata entries, we were able to further present differences between the 1,381 listed RDR on a statistical basis. Although this analysis is limited to the re3data metadata entries indexed by the end of 2015, we identified tendencies concerning the global landscape of RDR.

One outstanding fact is that compliance with important standards of the Information Infrastructure as well as the Library and Information Science (LIS) community are underused. This data analysis supports this conclusion especially regarding the provision of Persistent Identifiers for research data sets or the use of common APIs (cf. Tables 6 and 8). The API entries recorded within our sample do not allow extensive remarks on the possibilities of cross-institutional metadata exchange on the data level. The compatibility with community-approved standards nevertheless is essential for the implementation of any service that serves one or several research communities. This factor should be considered when planning, operating and improving a RDR.

The use, trust and reputation of repository services are of vital importance if they are to be widely recognized as a reliable information system. We also demonstrated that there are several RDR that are already recognized services since they are recommended in funder and publisher policies (Pampel, Vierkant, et al. 2013).

Based on the findings of this initial analysis further research on the landscape of RDR can differentiate and sharpen our exemplary picture of RDR. Within the re3data metadata schema, and consequently in the course of the analysis, we used a broad subject classification on the first level of main subject categories. A more in-depth analysis of the sub-categories may reveal significant differences between subjects that are summarized, for example, under the category "Humanities and Social Sciences". Similar applies for the content type category. More granular information on the RDR's content, such as data formats, might be a starting point for follow up research.

An active community of stakeholders (researchers, research organization, funding possibilities) is needed to either build or evaluate RDR that support the research community by managing research data in a sustainable way. Less information on sustainability and long-term preservation is provided with respect to the metadata schema of re3data. Nevertheless, this is of utmost importance for a sustainable landscape of Open Science services and infrastructure.

The metadata entries re3data provides do not claim to be complete or accurate. Our editorial board indexes and updates the RDR manually. Information might change over time or could not be found during the indexing process. The active cooperation of the community is needed to keep the re3data metadata entries as up-to-date as possible. These findings should also help the RDR community to improve their services or build new services if research institutions deem this necessary. For this reason, we have prepared the following recommendations based on our indexing experience as well as the analysis of the metadata collected.

6 Recommendations

The discussion on research data management has made big steps forward in the last years. The "FAIR Guiding Principles for scientific data management and stewardship" (findable, accessible, interoperable, reusable research data) from 2016 describe the current state of the art in a short and precise manner (Wilkinson et al. 2016). Quite a large number of the re3data indexed repositories began to operate long before that. We therefore recommend RDR managers to take a closer look at the following issues:

6.1 Visibility of a research data repository

1. Register RDR in re3data or similar registries to improve the repository's visibility. Services offered by the research data repository to the community are presented in an adequate and transparent way.

2. Review the metadata entry for the listed research data repository within re3data

6.2 Functionalities of a research data repository

3. Support PID systems to provide each dataset that is stored within the research data repository with a persistent identifier, thus allowing the persistent citation of the relevant dataset.

4. Use a data license to clarify access and usage conditions for the data sets provided. Existing machine-readable standard licenses can be supported to ensure that the license is legally effective and globally understandable.

5. Make metadata and related research data sets available to other services and research organizations through an API. This networking approach improves interoperability of the services, the visibility and findability of data sets within the research community and facilitates new services.

6. Ensure compliance with certificates and standards to ensure the reliability and trustworthiness of the RDR. It should be clearly communicated which certificates and standards are met, so that users can evaluate the service quality.

7. The institutional responsibility has to be clarified and communicated to the user, e.g. included in a data policy or a mission statement.

8. Create policies to describe the services offered, the terms under which the repository may be used and to clarify the principles that the RDR deems important.

9. When choosing or developing a repository software, ensure that the software supports technical standards that for example are described in the certificates mentioned above.

Taking into account the number of existing RDR as well as their different focuses in terms of subjects, content types and services, we strongly recommend to consider whether one or more of the listed RDR might be suitable for certain requirements. Before a single institution begins to build another service, probably with a similar orientation, re3data could be consulted to assess the appropriateness of available RDR.

Acknowledgements

We would like to thank all colleagues for their helpful input and comments on this article. The project re3data and writing of this paper was enabled by a grant from the German Research Foundation (DFG). We would like to thank the DFG for supporting our work. We would also like to thank the global repository community for its help and support in the establishment of re3data. re3data is operated as a service of DataCite.

References

Australian National Data Service (2010). *Outline of a Research Data Management Policy for Australian Universities / Institutions.* Version 1. URL: http://www.ands.org.au/__data/assets/pdf_file/0004/382072/datamanagementpolicyoutline.pdf.

Bartling, Sönke and Sascha Friesike, eds. (2014). *Opening Science. The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing.* Cham: Springer International Publishing. DOI: 10.1007/978-3-319-00026-8.

Beagrie, Neil and John Houghton (2014). *The Value and Impact of Data Sharing and Curation. A synthesis of three recent studies of UK research data centres.* Joint Information Systems Committee (JISC). URL: http://repository.jisc.ac.uk/5568/.

Borgman, Christine L., Jillian C. Wallis, and Matthew S. Mayernik (2012). "Who's Got the Data? Interdependencies in Science and Technology Collaborations". In: *Computer Supported Cooperative Work (CSCW)* 21.6, pp. 485–523. DOI: 10.1007/s10606-012-9169-z.

Brase, J., I. Sens, and M. Lautenschlager (2015). "The tenth anniversary of assigning DOI names to scientific data and a five year history of datacite". In: *D-Lib Magazine*, p. 1. DOI: 10.1045/january2015-brase.

Cohen, J (1988). *Statistical Power Analysis for the Behavioral Sciences.* 2nd ed. Hillsdale, N.J: Lawrence Erlbaum Associates Inc.

European Commission (2016). *Research & Innovation Infrastructures. What are RIs?* URL: http://ec.europa.eu/research/infrastructures/index.cfm?pg=about.

European Commission. Directorate-General for Research & Innovation (2016). *Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020.* Version 2.1.

Houghton, John and Nicholas Gruen (2014). *Open Research Data. Report to the Australian National Data Service (ANDS).* URL: http://ands.org.au/__data/assets/pdf_file/0019/393022/open-research-data-report.pdf.

Marcial, Laura Haak and Bradley M. Hemminger (2010). "Scientific data repositories on the Web: An initial survey". In: *Journal of the American Society for Information Science and Technology* 61.10, pp. 2029–2048. DOI: 10.1002/asi.21339.

Martín, E. and G Ballard (2010). *Data Management Best Practices and Standards for Biodiversity Data Applicable to Bird Monitoring Data.* U.S. North American Bird Conservation Initiative Monitoring Subcommittee. URL: http://www.prbo.org/refs/files/12058_Martin2010.pdf.

National Science Foundation. Directorate for Biological Sciences (2015). *Updated Information about the Data Management Plan Required for Full Proposals.* URL: https://www.nsf.gov/bio/pubs/BIODMP_Guidance.pdf.

Nielsen, Michael (2011). *Reinventing Discovery. The New Era of Networked Science.* Princeton University Press. URL: https://press.princeton.edu/titles/9517.html.

OECD (2015). "Making Open Science a Reality". In: OECD Science, Technology and Industry Policy Papers 25. DOI: 10.1787/5jrs2f963zs1-en.

Pampel, Heinz and Sünje Dallmeier-Tiessen (2014). "Open Research Data. From Vision to Practice". In: *Opening Science. The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing.* Ed. by Sönke Bartling and Sascha Friesike. Cham: Springer International Publishing, pp. 213–224. DOI: 10.1007/978-3-319-00026-8_14.

Pampel, Heinz, Paul Vierkant, et al. (2013). "Making Research Data Repositories Visible: The re3data.org Registry". In: *PLOS ONE* 8.11, pp. 1–10. DOI: 10.1371/journal.pone.0078080.

PARSE.insight team (2016). *PARSE.Insight Questionnaire*, p. 1129. URL: https://libereurope.eu/wp-content/uploads/2010/01/PARSE.Insight.-Deliverable-

D3.4-Survey-Report.-of-research-output-Europe-Title-of-Deliverable-Survey-Report.pdf.

re3data.org: Alaska Ocean Observing System (2015). DOI: 10.17616/R3JS7B.

re3data.org: Australian Data Archive (2015). DOI: 10.17616/R3DS3K.

re3data.org: Australian Ocean Data Network Portal (2014). DOI: 10.17616/r3cg82.

re3data.org: Europeana (2015). DOI: 10.17616/R3Q925.

re3data.org: NCBI Genome repository (2015). DOI: 10.17616/R3R89S.

re3data.org: Research Data Australia (2015). DOI: 10.17616/R33W46.

re3data.org: UK.Data Archive (2015). DOI: 10.17616/R3088K.

Rücknagel, J. et al. (2015). *Metadata Schema for the Description of Research Data Repositories. Version 3.0*, pp. 1–29. DOI: 10.2312/re3.008.

Sandt, Stephanie van de et al. (2017). *The Landscape of Research Data Repositories in 2015. A re3data Analysis.* DOI: 10.5281/zenodo.49709.

Simons, Natasha and Joanna Richardson (2013). *New Content in Digital Repositories. The Changing Research Landscape.* Chandos Information Professional Series. Chandos Publishing.

Tarazona Rua, Maria Monica, Daniel Spichtinger, Celina Ramjoue, and Jean-Francois Dechamp (2015). *Access to and Preservation of Scientific Information in Europe. Report on the implementation of Commission Recommendation C(2012) 4890 final.* European Commission. Directorate-General for Research and Innovation. DOI: 10.2777/975917.

The Royal Society (2012). *Science as an open enterprise.* The Royal Society Science Policy Centre report 02/12. URL: https://royalsociety.org/~/media/policy/projects/sape/2012-06-20-saoe.pdf.

Vierkant, Paul et al. (2014). *Metadata Schema for the Description of Research Data Repositories.* Version 2.2, pp. 1–27. DOI: 10.2312/re3.006.

Wilkinson, Mark D. et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3, p. 160018. DOI: 10.1038/sdata.2016.18.