Research Data Documentation: The Landscape of Research Data Repositories in 2015. A re3data Analysis

Version 1.1 - 07 March, 2017

Stephanie van de Sandt[b], Maxi Kindling[b], Heinz Pampel[a], Jessika Rücknagel[b], Paul Vierkant[a], Gabriele Kloska[c], Michael Witt[d], Peter Schirmbacher[b], Roland Bertelmann[a], Frank Scholze[c]

a) GFZ German Research Centre for Geosciences, Library and Information Services (LIS), Potsdam, Germany

b) Humboldt-Universität zu Berlin, Berlin School of Library and Information Science (BSLIS), Germany

c) Karlsruhe Institute of Technology (KIT), KIT Library, Germany

d) Purdue University Libraries, West Lafayette, USA

Abstract

This documentation describes a data set aggregated from a database snapshot of re3data dated 3 December 2015. As a registry of research data repositories, re3data offered metadata about 1381 repositories worldwide at that time. The data set is supplement to a data analysis paper. The data set contains four data tables, which are individually described in this documentation, as well as a matrix table including statistical correlations.

Citation

van de Sandt, S.; Kindling, M.; Pampel, H.; Rücknagel, J.; Vierkant, P.; Kloska, G.; Witt, M.; Schirmbacher, P; Bertelmann, R.; Scholze, F. Research Data Documentation: The Landscape of Research Data in 2015. A re3data Analysis. 2016. Version 1.0: DOI:http://doi.org/10.5281/zenodo.49709

Licenses

Description

This research data documentation is supplement to a data analysis paper (cf. Kindling et al.,submitted). The research data consists of aggregated metadata information on research data repositories from re3data, the global registry of research data repositories. All metadata entries are based on the 2.2 version of the schema for the description of research data repositories (Vierkant et al. 2014). The database snapshot of re3data, dated 3 December 2015, was provided as SQL dump and imported into a PostgreSQL database to perfom SQL queries according to several research questions that are described in the data analysis paper. The query outputs were extracted as .csv files, cleansed and analyzed by using Python, R, SPSS and Microsoft Excel. The documented research data set consists of four data tables in csv. format and a matrix table in xlsx. format (coloured).

Data table 1: re3data cleansed repository id.csv

The data table "re3data cleansed repository id.csv" is based on raw quantities with multiple options for an ID and blank cells since most properties of a repository are modeled as 1:n relation in the re3data metadata schema, e.g. an ID is usually linked to more than one "contenttype" (images, raw data etc.). Thus there were several options represented by a string per ID for one variable though

they differ and cannot be treated as duplicates. All property values thus were transposed into new variables and a binary representation with the help of a Python script (e.g. if a repository contains images as content type, this information will be represented by value "1" for the new variable "contenttype:image"). This form enabled us to perform the statistical analysis of the given data. The unique internal re3data database identifier "repository id" is used as foreign key in our PostgreSQL relational database. It must not be mixed up with the actual re3data identifier used on the re3data website.

Variables of data table 1

The main data table 1 "re3data cleansed repository id.csv" contains the following variable IDs. Please note that the data set only includes values for the four main subject classes according to the DFG classification (cf. Kindling et al. 2016).

- subject (ID 13)
- contentType (ID 15)
- institution (ID 18) including:
  - responsibilityType (ID 18.4)
- databaseAccess (ID 20) including:
  - databaseAccessType (ID 20.1)
- databaseLicense (ID 21) including:
  - databaseLicenseName (ID 21.1)
- dataAccess (ID 22) including:
  - dataAccessType (ID 22.1)
- dataLicense (ID 23) including:
  - dataLicenseName (ID 23.1)
- software (ID 26) including:
  - softwareName (ID 26.1)
- api (ID 28) including:
  - apiType (ID 28.1)
- pidSystem (ID 29)
- certificate (ID 34)

Data table 2: re3data countries language policies.csv

The data table "re3data countries language policies.csv" consists of uncleansed raw quantities of the variables "countries", "language" and "policy". It contains information on the language of the repository's GUI, the policy statements and on the countries of origins of repository institutions. Each ID can have multiple values (e.g. a repository may be linked to several institutions from the same country).

Variables of data table 2

Data table 2 contains the following variable IDs:

- repositoryLanguage (ID 12)
- institution (ID 18) including:
  - institutionCountry (ID 18.3)
- policy (ID 19)

Data table 3: re3data database access id

The data table "re3data database access id" only contains the uncleansed raw quantities of the variable "databaseAccessRestriction" and is only related to the "database access identifier".

Variables of data table 3

Data table 3 contains the following variable IDs:

- databaseAccess (ID 20) including:
  - databaseAccessRestriction (ID 20.2)

Data table 4: re3data institution id

The data table "re3data institution id" only contains the uncleansed raw quantities of the variable "institutionResponsibilityType".

Variables of data table 4

Data table 4 contains the following variable IDs:

- institution (ID 18) including:
  - responsibilityType (ID 18.4)

Venn diagrams

We used Venn diagrams to represent logical relations in our variables in the above mentioned data analysis paper. For Venn diagrams with more than 3 variable sets we used the R library VennDiagram [1]. Circles in the Venn diagrams represent the amount of research data repositories that have a certain attribute. As the areas of the 3 circle Venn diagrams in R cannot be scaled proportional to the number of elements it contains, in one case [2] we used area-scaled figures created with the Python library matplotlib-venn[3].

Correlation matrix

The Microsoft Excel file "re3data correlation matrix" contains all correlating variables of the cleansed main data table 1. To obtain an overview and shorten the matrix we only selected the appropriate values of variables that were involved in significant correlations. A grey cell background indicates either no correlation was found or it was not tested. A blue cell background indicates that less values than expected were found and the Phi value was negative. A red cell background on the opposite indicates that more values than expected were found and the Phi value was positive. Accordingly such cells contain the Cramer's V value we based our analysis on. The level of significance is shown by the amount of asterisks. Two asterisks indicate a highly significant correlation on a 99% level. One asterisks indicates a significant correlation on at least a 95% level. The correlations we found are discussed in detail in the data analysis paper.

Correlation calculation

The correlations described in the data analysis paper were calculated using SPSS Statistics Version 22.0.0 Mac OS X on Mac OS X 10.7.5. For the purpose of reproducibility the calculation is demonstrated below using the example of the low correlation between research data repositories covering the subject "Humanities and Social Sciences" and "closed" access to the research data of the repositories (Cramer's V = 0.156).
First, χ^2 needs to be calculated:

$$\chi^2 = \sum[(f_o - f_e)^2/f_e]$$

f_o are the observed values (count); f_e are the statistically expected values.

The contingency table shows the following values:

			data access closed	
			0	1
SSH	0	count	957	50
		expected	931.9	75.1
	1	count	321	53
		expected	346.1	27.9

[1] `https://cran.r-project.org/web/packages/VennDiagram/VennDiagram.pdf`
[2] see Figure 3: Types of research data repositories
[3] `https://pypi.python.org/pypi/matplotlib-venn`

If we sum up all values, we get a χ^2 of 33.486.

The formula for Cramer's V looks this way:

$$V = \sqrt{\frac{\chi^2}{n * (R - 1)}}$$

n = 1381 (total amount of research data repositories in the snapshot)
R = 2 (the amount of rows is smaller than the amount of lines, so the smaller number was chosen)

with the following result: Cramer's V = 0.156.

References

Kindling, M., Pampel, H., van de Sandt, S., Rücknagel, J., Vierkant, P., Kloska, G., Witt, M., Schirmbacher, P., Bertelmann, R., Scholze, F. (2017) The Landscape of Research Data Repositories in 2015. A re3data Analysis (submitted data analysis paper)

Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., Goebelbecker, H.-J., Gundlach, J., Schirmbacher, P., Dierolf, U. (2013). Making Research Data Repositories Visible: The re3data Registry. PLOS ONE, 8(11), e78080. doi: http://doi.org/10.1371/journal.pone.0078080

Vierkant, P., Spier, S., Ruecknagel, J., Pampel, H., Fritze, F., Gundlach, J., Fichtmüller, D., Kindling, M., Kirchhoff, A., Göbelbecker, H.-J., Klump, J., Kloska, G., Reuter, E., Semrau, A., Schnepf, E., Skarupianski, M., Bertelmann, R., Schirmbacher, P., Scholze, F., Kramer, C., Witt, M., Fuchs, C., Ulrich, R. (2014). Schema for the Description of Research Data Repositories. Version 2.2. doi: http://doi.org/10.2312/re3.006