DATASEA
DATOS ABIERTOS DE INVESTIGACIÓN - OPEN RESEARCH DATA

e-LiS [2003-2018] e-prints in library & information science 15th Anniversary

Research data for E-LIS repository

Research data in practice

Antonella De Robbio
E-LIS Admin Board

DATASEA FINAL

Valencia, 22/06/2018

Datos generados por científicos: el futuro de la investigación

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

Summary

1. Introduction to Open Science

2.Data definition for Humanities Science: PARTHENOS

3. E-LIS the international repository for Library and Information Science in OS framework

4. Data: Government Data Research data and metadata

5. Big Data or Small data?

6. Some small data about E-LIS: statistic data

7. Metata and data in E-LIS structure

8. What are research data: types and life cycle

9. A world of data: Open, Shared, Reused, Published, Restricted Data

10. Why it is important to manage research data

11. European projects and research data management: the FAIR principles

12. Legal framework

13. Basic aspects of data curation activities : back up, storage, preservation…

14. File formats and transformation for privacy and sensible data

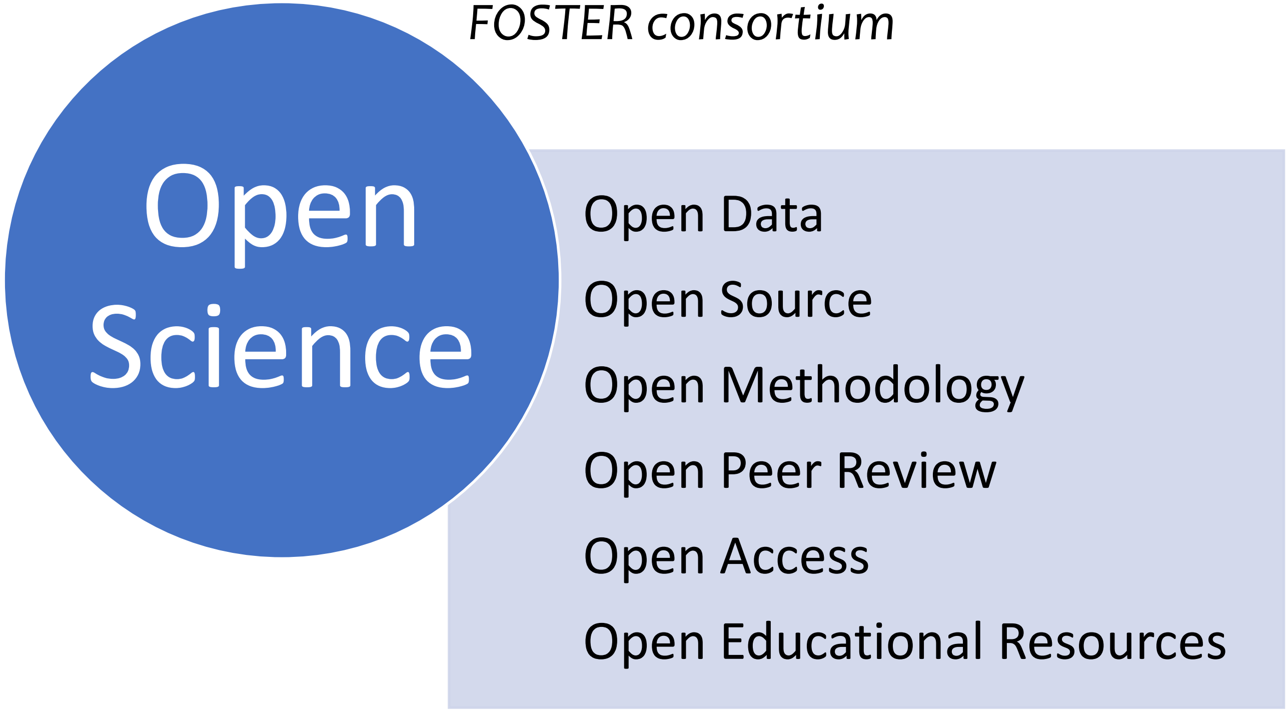
15. Data Citation and Schema.org project

16. Reliability of Data Repositories: which repository for my data? Re3data.org

Introduction to Open Science

“Open science is the movement to make scientific research, data and dissemination accessible to all levels of an inquiring society”

FOSTER consortium



Open Science

- Open Data
- Open Source
- Open Methodology
- Open Peer Review
- Open Access
- Open Educational Resources

Andreas E. Neuhold – Opera propria – CC BY 3.0

WHAT IS DATA?

PARTHENOS

Pooling Activities, Resources and Tools
for Heritage E-research Networking,
Optimization and Synergies

In this section...

- PARTHENOS training
  - Introduction to Research Infrastructures
    - What IS Infrastructure?
    - Interoperability
      - What is Data?
      - What is Metadata?
      - What are Standards?
      - What Are Knowledge Representation Systems and 'Ontologies'?

There are many definitions of what constitutes 'data', and often it depends on what your area of study is. On a conceptual level, data can been seen as the basic starting point for research investigation, the 'raw material' from which a researcher begins to construct his or her understanding of a particular field or question. These materials are often called 'raw data,' although that is a highly contextual term, given that in many cases they have already been created or collected by another person or institution. As the work of finding and collection continues, this will gradually become what is known as 'research data,' that is, the collected material from which the researcher will construct their final theories and arguments.

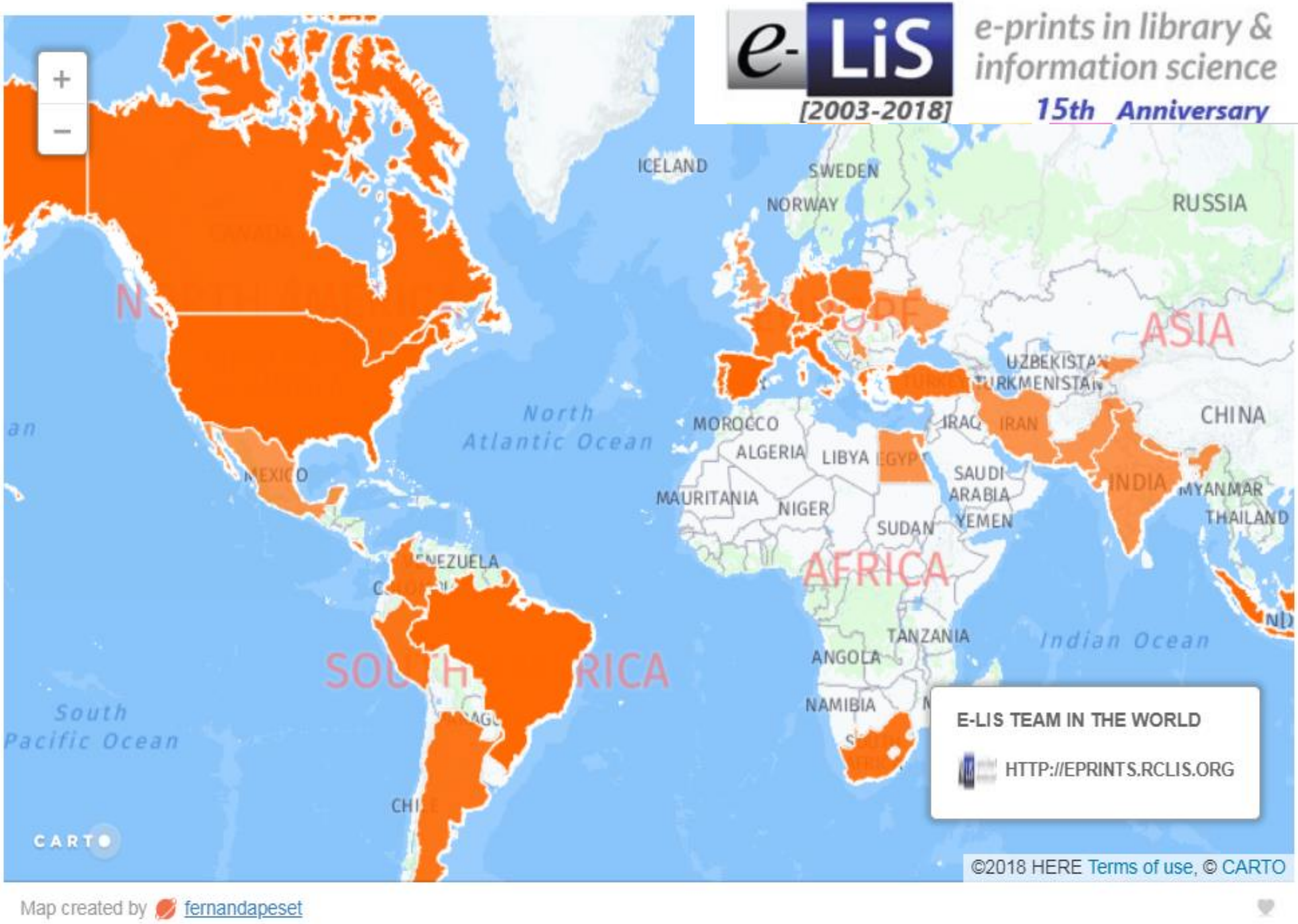
At a very simple level, 'data' is a collection of observations, facts, objects, texts or statistics that can be analysed, sometimes also referred to as 'sources' or 'evidence. Other definitions include "*citations, software code, algorithms, digital tools, documentation, databases, geospatial coordinates (for example, from archaeological digs), reports, and articles.*" (NEH, 2015) But even this long list can be expanded, as humanists also study audio and video recordings, collections of images, and other hybrid media.

E-LIS the international repository for Library and Information Science in the Open Science framework

Editorial Team
65 editors

e-LiS e-prints in library & information science

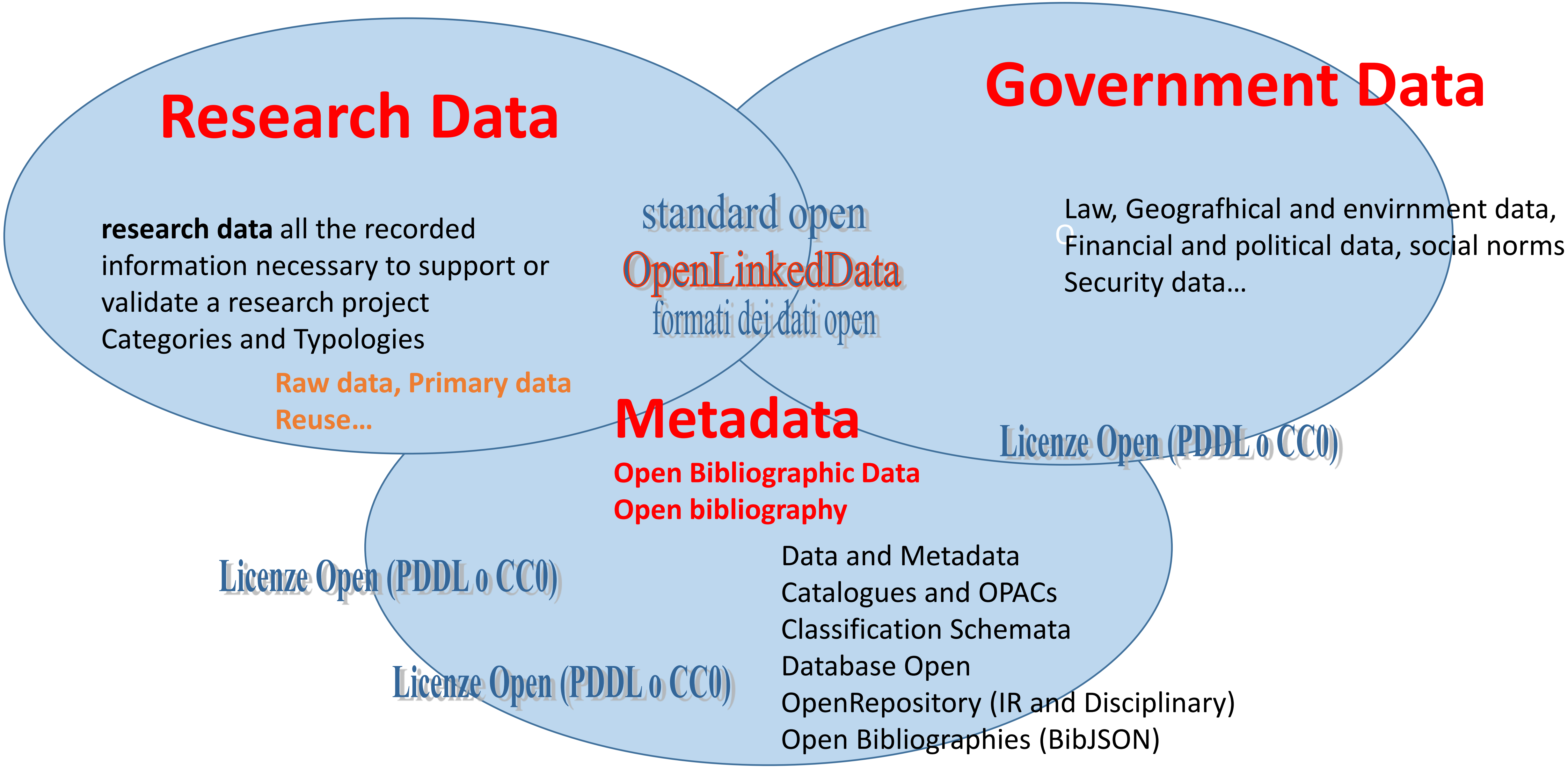
[2003-2018] 15th Anniversary



©2018 HERE Terms of use, © CARTO

Map created by fernandapeset

Three Open Data Layers



Research Data

research data all the recorded information necessary to support or validate a research project
Categories and Typologies

Raw data, Primary data
Reuse…

standard open
OpenLinkedData
formati dei dati open

Government Data

Law, Geografhical and envirnment data,
Financial and political data, social norms
Security data…

Licenze Open (PDDL o CC0)

Metadata

Open Bibliographic Data
Open bibliography

Licenze Open (PDDL o CC0)

Licenze Open (PDDL o CC0)

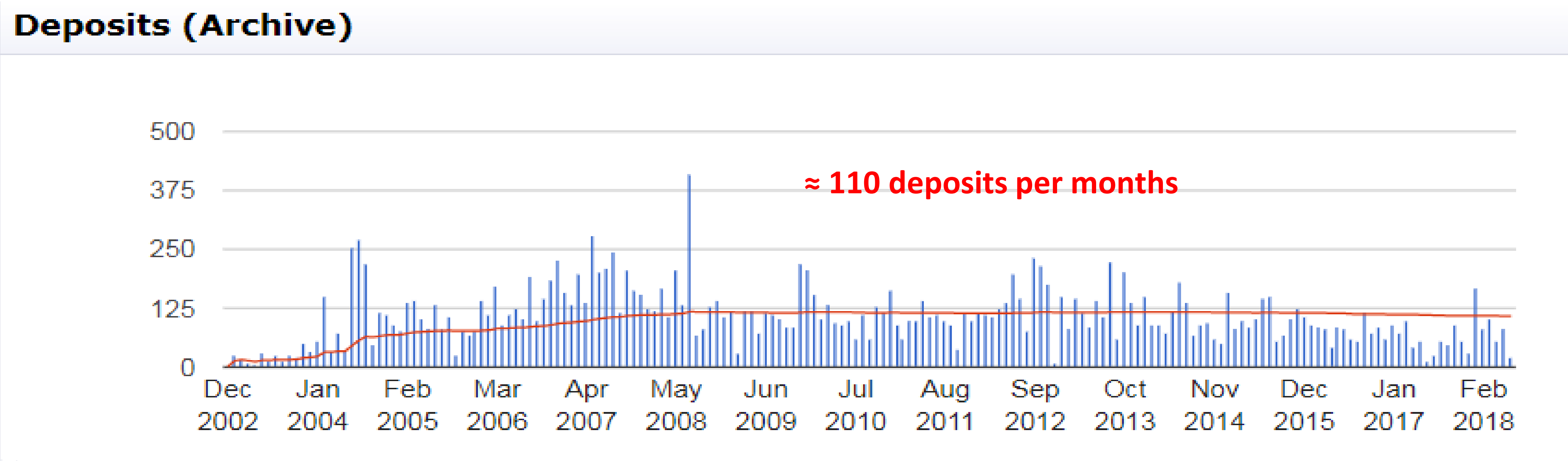
Data and Metadata
Catalogues and OPACs
Classification Schemata
Database Open
OpenRepository (IR and Disciplinary)
Open Bibliographies (BibJSON)

Big data or Small data?

Big data is the topic **in the world of marketing**
Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information
Include analysis, capture, data curation, search, sharing, storage, transfer, visualization.

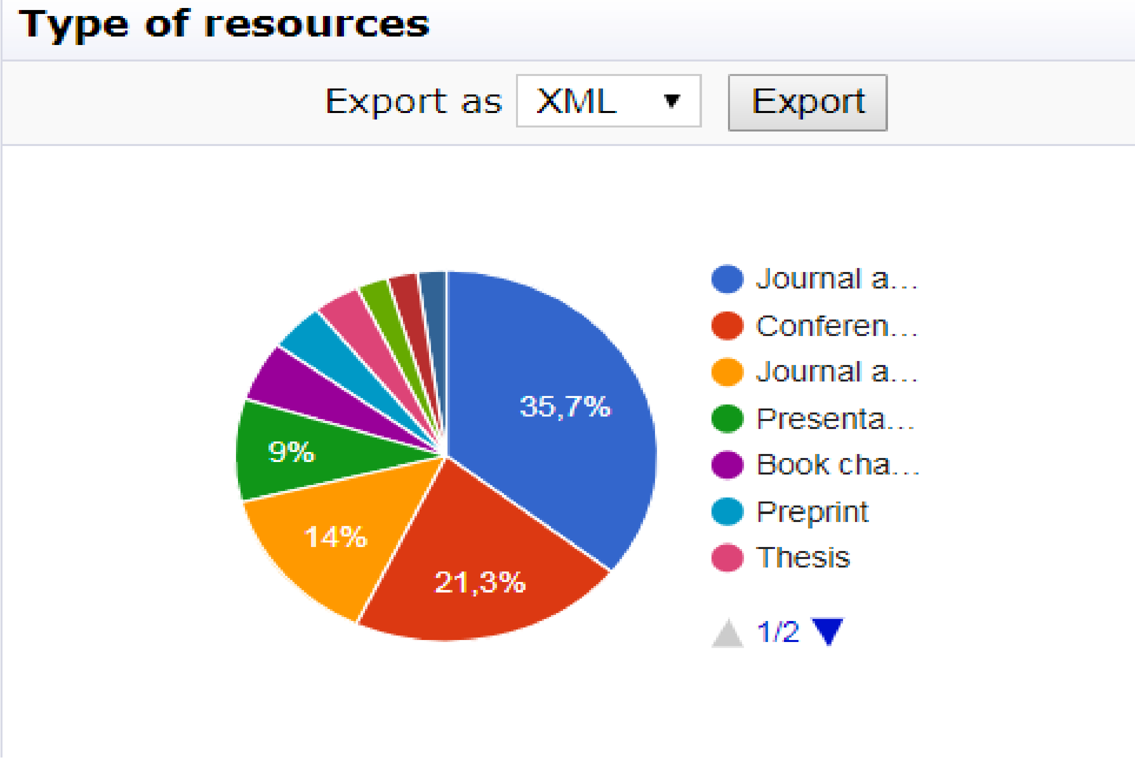
Importance of small data: Small data is data that is 'small' enough for human comprehension.
It is data in a volume and format that makes it accessible, informative and actionable.
The small set of specific attributes produced by the **Internet of Things**.

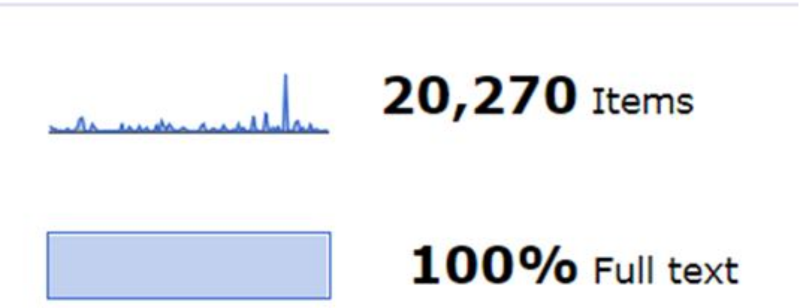
Deposits (Archive)



≈ 110 deposits per months

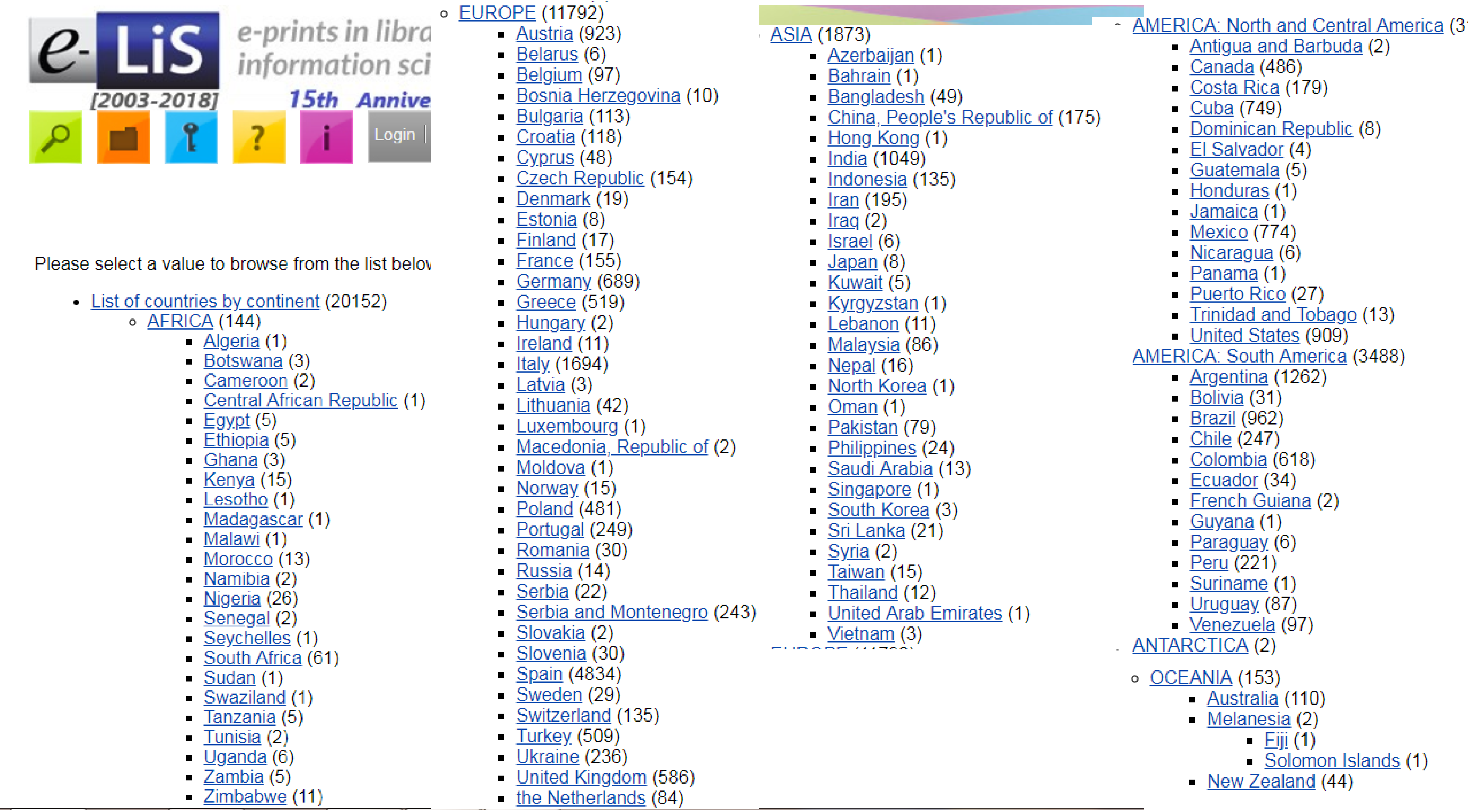
Type of resources





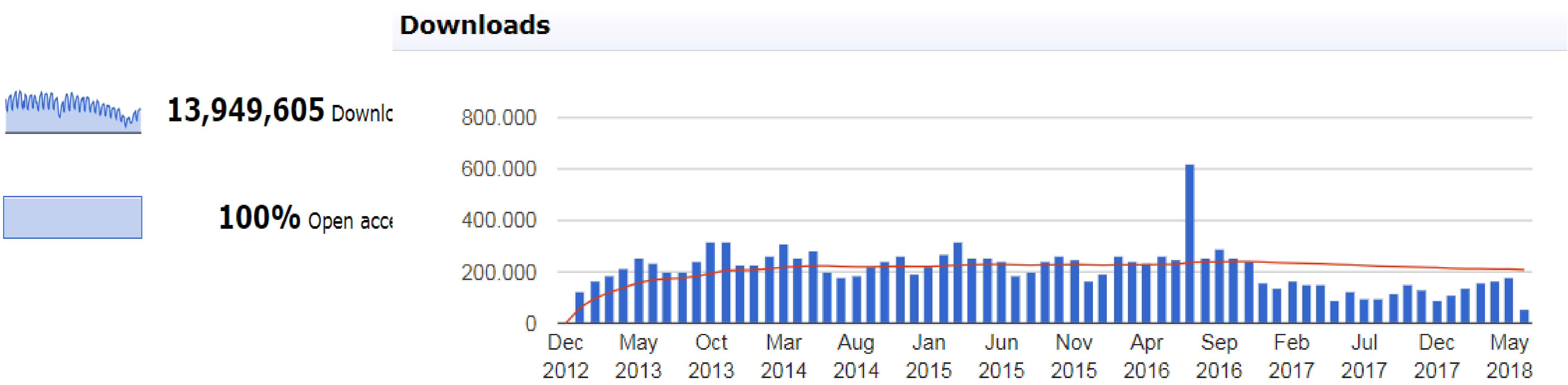
20,270 Items

100% Full text



Please select a value to browse from the list belov

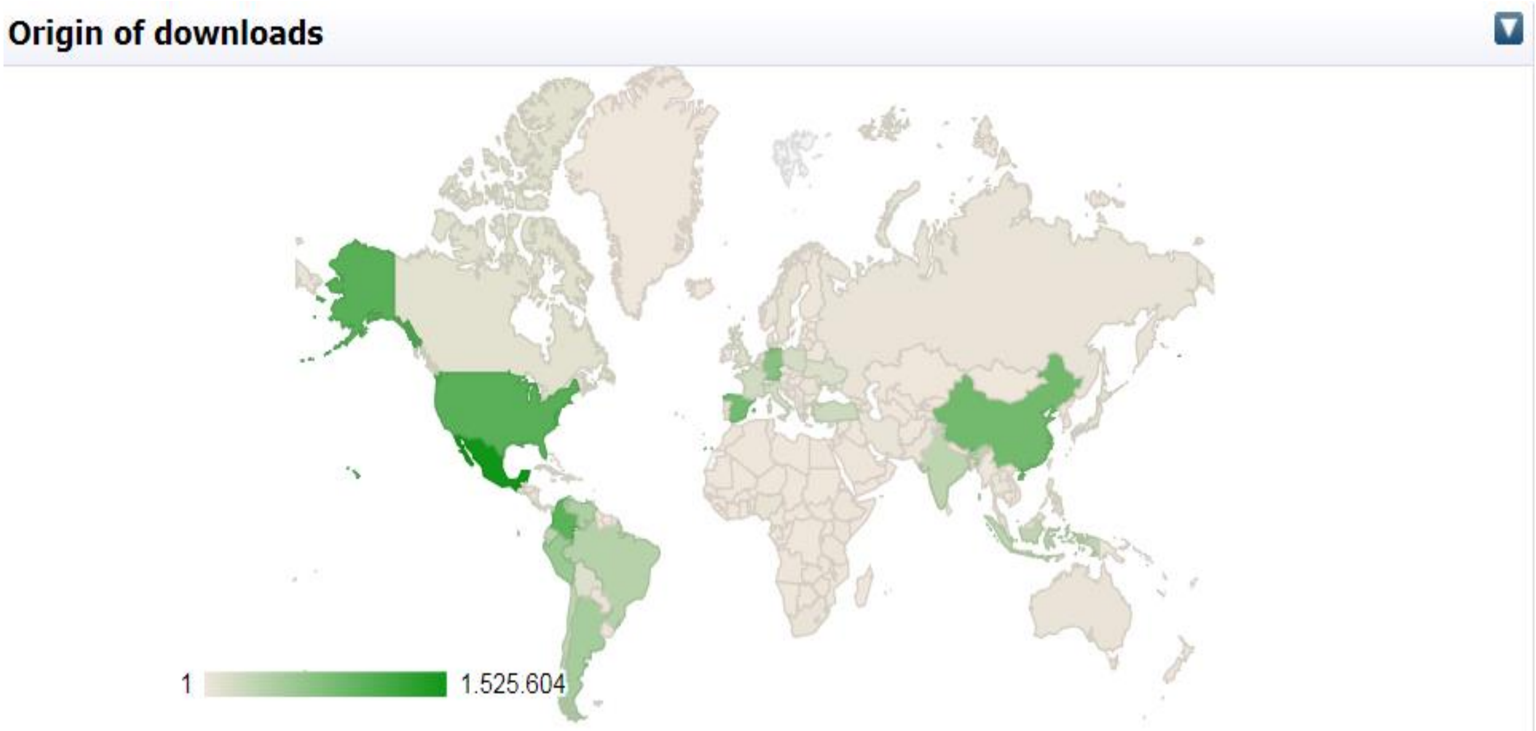
- List of countries by continent (20152)
  - AFRICA (144)
    - Algeria (1)
    - Botswana (3)
    - Cameroon (2)
    - Central African Republic (1)
    - Egypt (5)
    - Ethiopia (5)
    - Ghana (3)
    - Kenya (15)
    - Lesotho (1)
    - Madagascar (1)
    - Malawi (1)
    - Morocco (13)
    - Namibia (2)
    - Nigeria (26)
    - Senegal (2)
    - Seychelles (1)
    - South Africa (61)
    - Sudan (1)
    - Swaziland (1)
    - Tanzania (5)
    - Tunisia (2)
    - Uganda (6)
    - Zambia (5)
    - Zimbabwe (11)
  - EUROPE (11792)
    - Austria (923)
    - Belarus (6)
    - Belgium (97)
    - Bosnia Herzegovina (10)
    - Bulgaria (113)
    - Croatia (118)
    - Cyprus (48)
    - Czech Republic (154)
    - Denmark (19)
    - Estonia (8)
    - Finland (17)
    - France (155)
    - Germany (689)
    - Greece (519)
    - Hungary (2)
    - Ireland (11)
    - Italy (1694)
    - Latvia (3)
    - Lithuania (42)
    - Luxembourg (1)
    - Macedonia, Republic of (2)
    - Moldova (1)
    - Norway (15)
    - Poland (481)
    - Portugal (249)
    - Romania (30)
    - Russia (14)
    - Serbia (22)
    - Serbia and Montenegro (243)
    - Slovakia (2)
    - Slovenia (30)
    - Spain (4834)
    - Sweden (29)
    - Switzerland (135)
    - Turkey (509)
    - Ukraine (236)
    - United Kingdom (586)
    - the Netherlands (84)
  - ASIA (1873)
    - Azerbaijan (1)
    - Bahrain (1)
    - Bangladesh (49)
    - China, People's Republic of (175)
    - Hong Kong (1)
    - India (1049)
    - Indonesia (135)
    - Iran (195)
    - Iraq (2)
    - Israel (6)
    - Japan (8)
    - Kuwait (5)
    - Kyrgyzstan (1)
    - Lebanon (11)
    - Malaysia (86)
    - Nepal (16)
    - North Korea (1)
    - Oman (1)
    - Pakistan (79)
    - Philippines (24)
    - Saudi Arabia (13)
    - Singapore (1)
    - South Korea (3)
    - Sri Lanka (21)
    - Syria (2)
    - Taiwan (15)
    - Thailand (12)
    - United Arab Emirates (1)
    - Vietnam (3)
  - AMERICA: North and Central America (3
    - Antigua and Barbuda (2)
    - Canada (486)
    - Costa Rica (179)
    - Cuba (749)
    - Dominican Republic (8)
    - El Salvador (4)
    - Guatemala (5)
    - Honduras (1)
    - Jamaica (1)
    - Mexico (774)
    - Nicaragua (6)
    - Panama (1)
    - Puerto Rico (27)
    - Trinidad and Tobago (13)
    - United States (909)
  - AMERICA: South America (3488)
    - Argentina (1262)
    - Bolivia (31)
    - Brazil (962)
    - Chile (247)
    - Colombia (618)
    - Ecuador (34)
    - French Guiana (2)
    - Guyana (1)
    - Paraguay (6)
    - Peru (221)
    - Suriname (1)
    - Uruguay (87)
    - Venezuela (97)
  - ANTARCTICA (2)
  - OCEANIA (153)
    - Australia (110)
    - Melanesia (2)
      - Fiji (1)
      - Solomon Islands (1)
    - New Zealand (44)



Downloads

13,949,605 Downlc

100% Open acce



Origin of downloads

Mexico	1,525,604
USA	1,029,585
SA over	3,000,000
CN	848,042
ES	824,705
DE	684,788

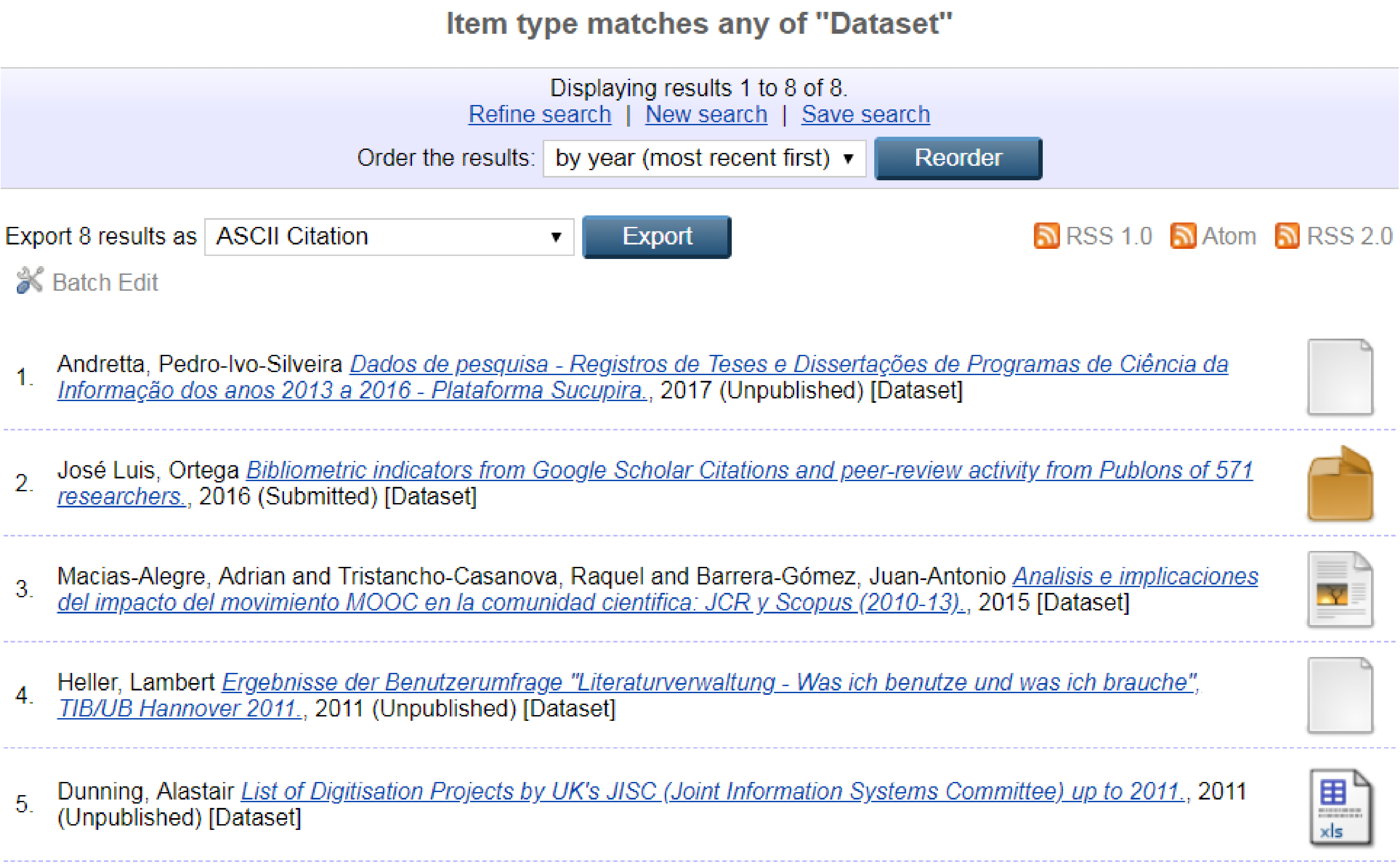
Metadata

E-LIS puts a great attention on metadata quality

Cultural and memory institutions have a long tradition of setting up, publishing, and sharing vast amounts of metadata, such as library catalogues and archival finding, providing inventories of books and documents with detailed descriptions of individual items using many different formats and approaches (i.e.: bibliographic approach vs historical approach). There are various categories of metadata, used to support different use cases in the digital domain.

A set of Metadata should at least specify:

- an identifier (or handlr),
- the name of the main researchers,
- the title of the data set,
- the name of the institution that holds the dataset,
- a **publication date** and the **type of resource** you are describing.



Item type matches any of "Dataset"

Displaying results 1 to 8 of 8.

Refine search | New search | Save search

Order the results: by year (most recent first) ▾

Export 8 results as ASCII Citation ▾

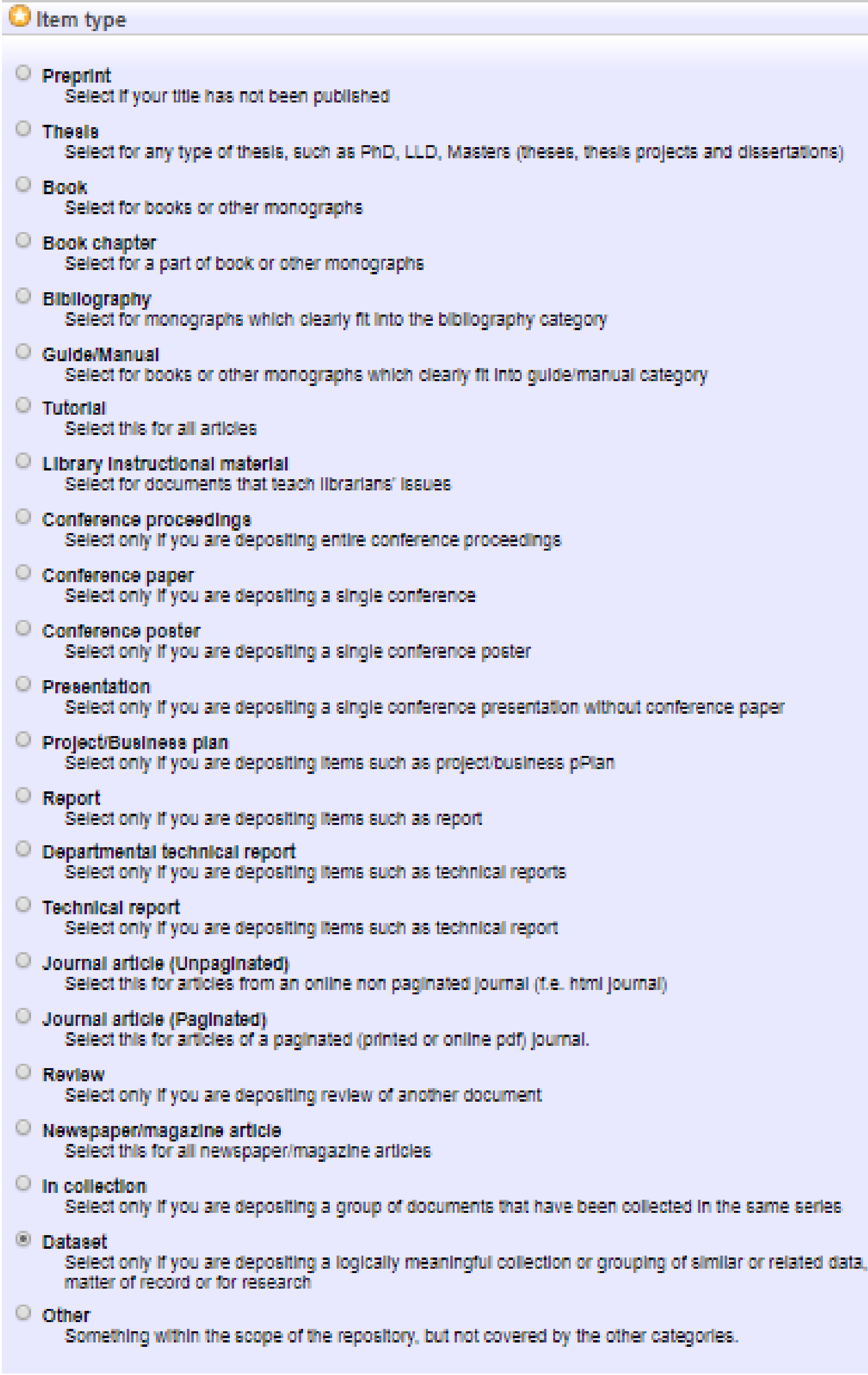
1. Andretta, Pedro-Ivo-Silveira *Dados de pesquisa - Registros de Teses e Dissertações de Programas de Ciência da Informação dos anos 2013 a 2016 - Plataforma Sucupira.*, 2017 (Unpublished) [Dataset]

2. José Luis, Ortega *Bibliometric indicators from Google Scholar Citations and peer-review activity from Publons of 571 researchers.*, 2016 (Submitted) [Dataset]

3. Macias-Alegre, Adrian and Tristancho-Casanova, Raquel and Barrera-Gómez, Juan-Antonio *Analisis e implicaciones del impacto del movimiento MOOC en la comunidad cientifica: JCR y Scopus (2010-13).*, 2015 [Dataset]

4. Heller, Lambert *Ergebnisse der Benutzerumfrage "Literaturverwaltung - Was ich benutze und was ich brauche", TIB/UB Hannover 2011.*, 2011 (Unpublished) [Dataset]

5. Dunning, Alastair *List of Digitisation Projects by UK's JISC (Joint Information Systems Committee) up to 2011.*, 2011 (Unpublished) [Dataset]



Item type

○ Preprint
Select if your title has not been published

○ Thesis
Select for any type of thesis, such as PhD, LLD, Masters (theses, thesis projects and dissertations)

○ Book
Select for books or other monographs

○ Book chapter
Select for a part of book or other monographs

○ Bibliography
Select for monographs which clearly fit into the bibliography category

○ Guide/Manual
Select for books or other monographs which clearly fit into guide/manual category

○ Tutorial
Select this for all articles

○ Library instructional material
Select for documents that teach librarians' issues

○ Conference proceedings
Select only if you are depositing entire conference proceedings

○ Conference paper
Select only if you are depositing a single conference

○ Conference poster
Select only if you are depositing a single conference poster

○ Presentation
Select only if you are depositing a single conference presentation without conference paper

○ Project/Business plan
Select only if you are depositing items such as project/business pPlan

○ Report
Select only if you are depositing items such as report

○ Departmental technical report
Select only if you are depositing items such as technical reports

○ Technical report
Select only if you are depositing items such as technical report

○ Journal article (Unpaginated)
Select this for articles from an online non paginated journal (i.e. html journal)

○ Journal article (Paginated)
Select this for articles of a paginated (printed or online pdf) journal.

○ Review
Select only if you are depositing review of another document

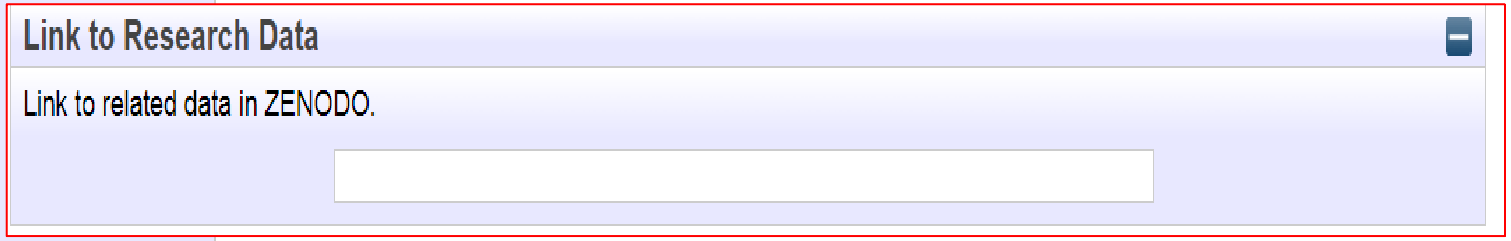
○ Newspaper/magazine article
Select this for all newspaper/magazine articles

○ In collection
Select only if you are depositing a group of documents that have been collected in the same series

● Dataset
Select only if you are depositing a logically meaningful collection or grouping of similar or related data, usually assembled as a matter of record or for research

○ Other
Something within the scope of the repository, but not covered by the other categories.

But in LIS research studies when we need to have kit of research data to prove validity the paper of our research? Which dataset for LIS argument? This is great question. Surveys? Spreadsheet with comparative data? Tutorial? Statistical data?

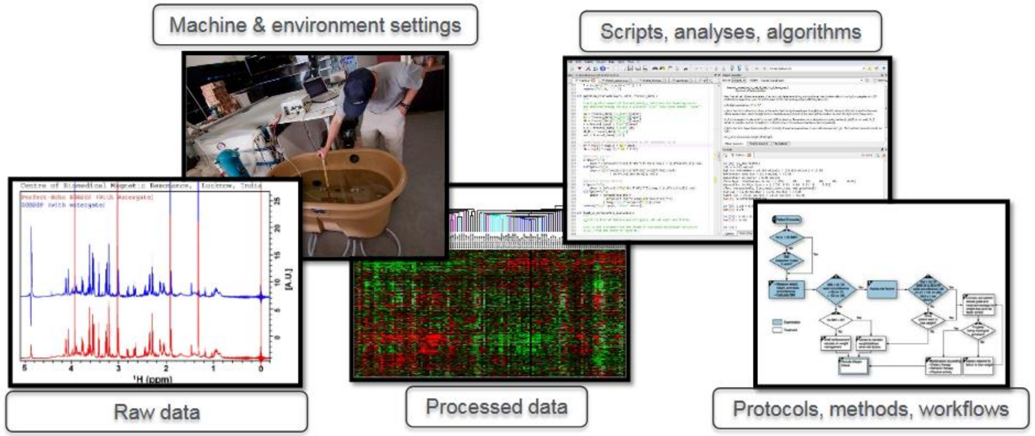


Link to Research Data

Link to related data in ZENODO.

- Status field.
- Refereed field.
- Public domain field.
- Authors field.
- Title field.
- Subjects field.
- Date field.
- English abstract field.
- Keywords field.
- Language field.
- Country field

What are research data: categories and types


Machine & environment settings
Scripts, analyses, algorithms
Raw data
Processed data
Protocols, methods, workflows

General categories of data:

- Observational (e.g. sensor readings, survey instruments)
- Experimental (e.g. lab equipment readings)
- Simulation (e.g. climate models)
- Derived or compiled (e.g. compiled databases, text or data mining)

Examples of research data:

•Digital texts or digital copies of text
•Spreadsheets
•Audio, video
•Computer Aided Design (CAD)
•Waveforms
•Statistics (SPSS, SAS)
•Databases
•Geographic Information Systems (GIS)
and spatial data
•Digital copies of images
•Matlab files
•Computer code
•Protein or genetic sequences
•Artistic products
•Web files

A world of data:

Open, Shared, Reused, Published, Restricted Data

Data is open if it can be freely accessed, used, modified, mined and shared by anyone for any purpose

Open data is defined by the Open Definition and requires that the data be:

Legally open = available under an open (data) license that permits anyone freely to access, reuse and redistribute (e.g. see Share-alike licenses)

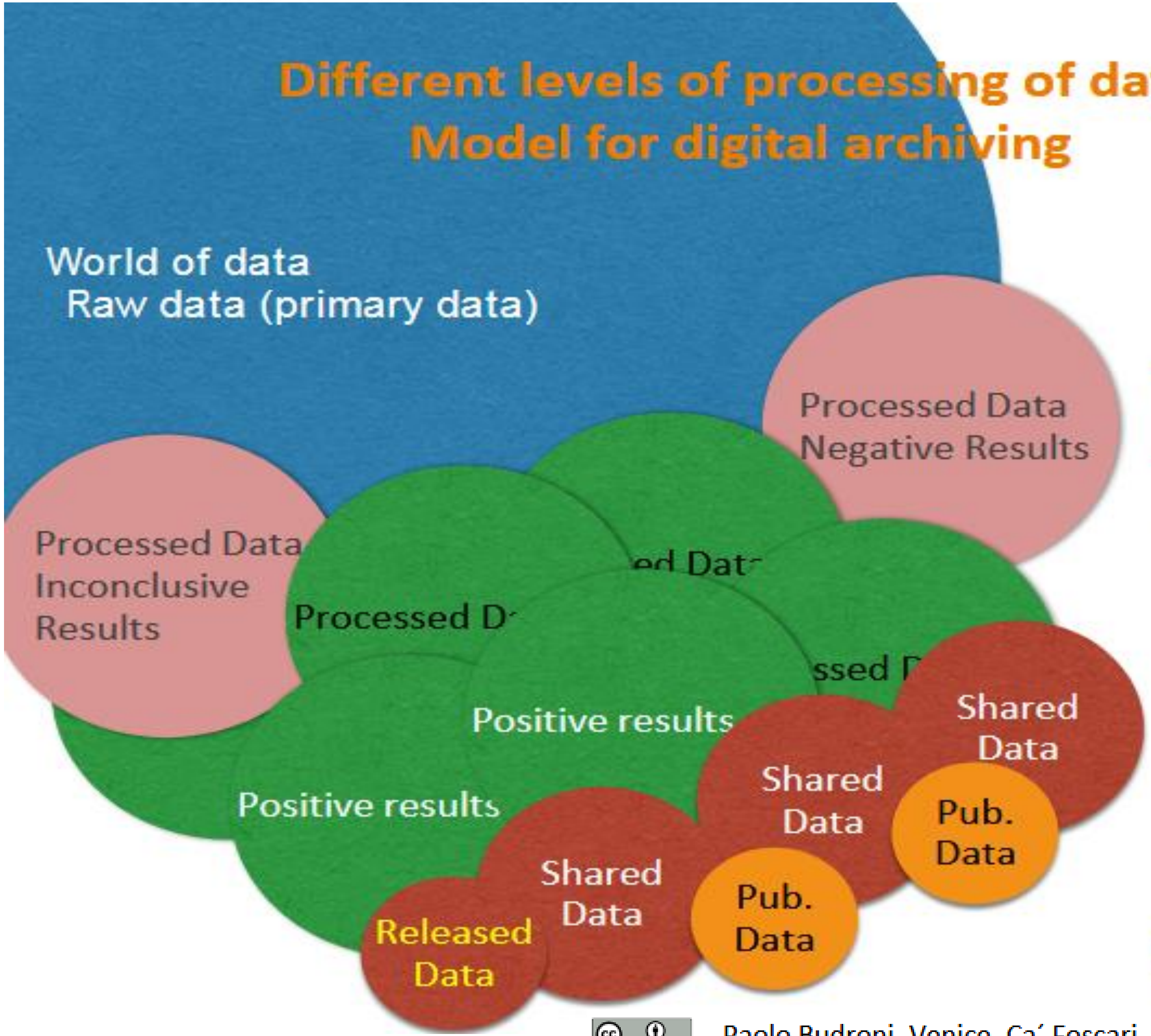
Technically open = available for no more than the cost of reproduction and in machine-readable and bulk form.

Here is a useful Checklist

http://www.codata.org/uploads/Legal%20Interoperability%20Principles%20and%20Implementation%20Guidelines_Final2.pdf

Open Knowledge International - https://okfn.org/

Why it is important to manage research data



Different levels of processing of data
Model for digital archiving

→ Strata of research data

→ Restricted Data

→ Open data

→ Published data

→ Open access published data

OA

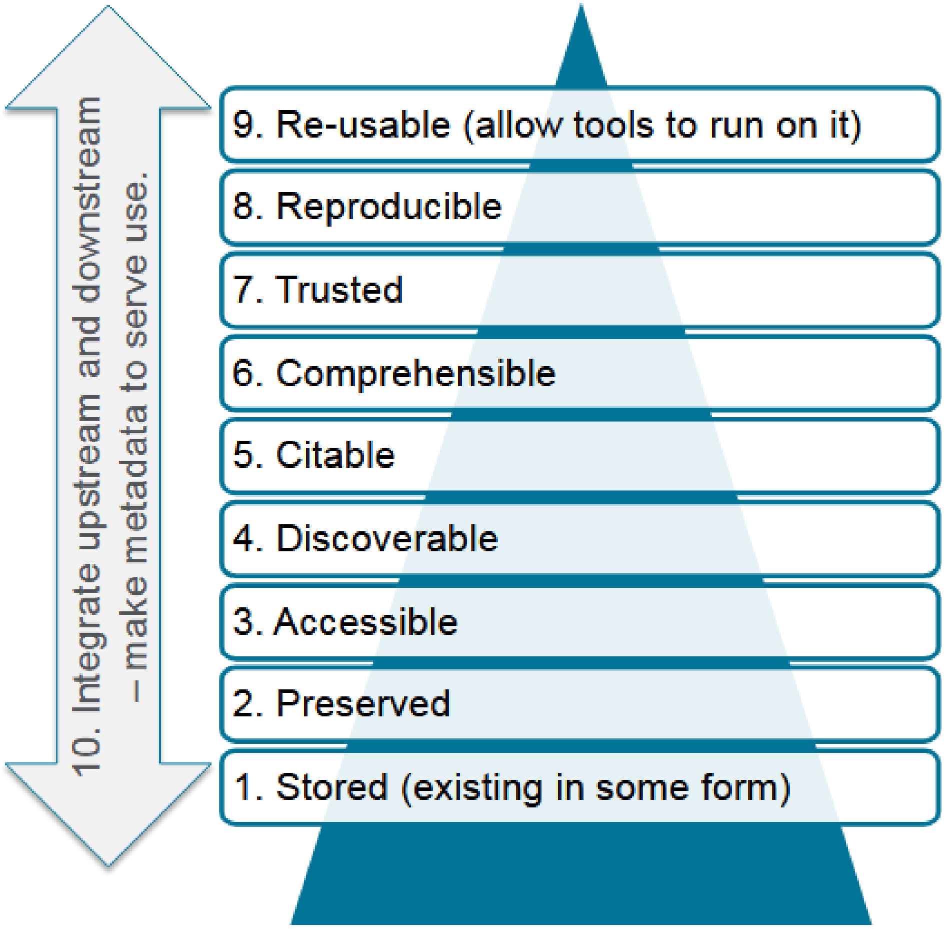
European projects and research data management: the FAIR principles

FAIR Principles

Find
Access
Interoperate
Re-use
Data

Data FAIRport

Find, Access, Interoperate & Re-use Data



9. Re-usable (allow tools to run on it)
8. Reproducible
7. Trusted
6. Comprehensible
5. Citable
4. Discoverable
3. Accessible
2. Preserved
1. Stored (existing in some form)

10. Integrate upstream and downstream – make metadata to serve use.

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier [...] Barend Mons

H2020 Programme

Guidelines on

FAIR Data Management in Horizon 2020

Findable

Accessible

Interoperable

Reusable

Legal framework

Intellectual property rights

Sensitive data

PSI Directive

Open Access and Open Data

Licensing frameworks

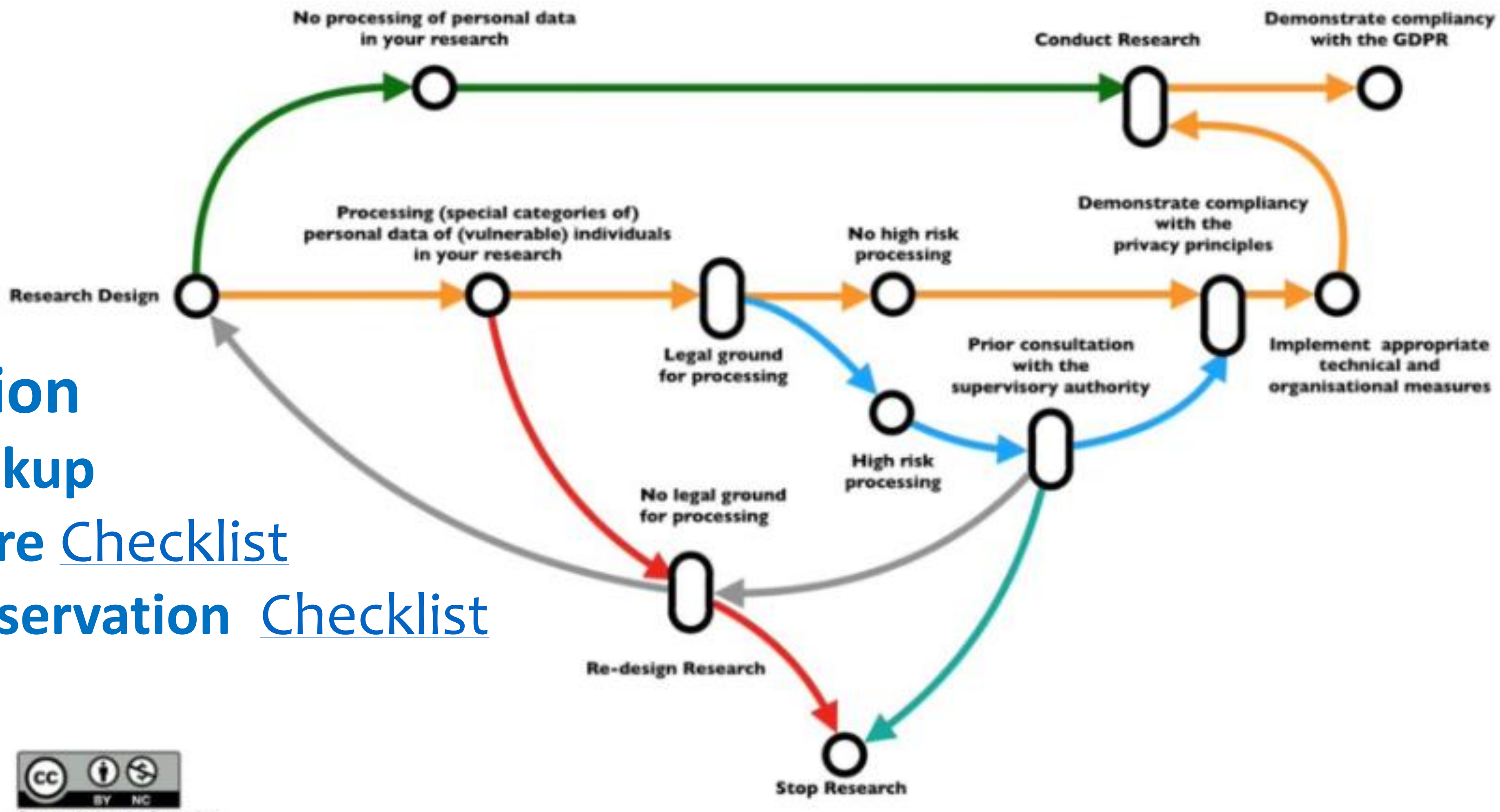
Rights Statements (RightsStatements.org)

Creative Commons

Licensing framework in PARTHENOS Community

Authentication and authorization infrastructure

The London Metro Map Approach to a Privacy Impact Assessment (PIA) for Academic Research


No processing of personal data in your research
Conduct Research
Demonstrate compliancy with the GDPR
Processing (special categories of) personal data of (vulnerable) individuals in your research
No high risk processing
Demonstrate compliancy with the privacy principles
Research Design
Legal ground for processing
Prior consultation with the supervisory authority
Implement appropriate technical and organisational measures
High risk processing
No legal ground for processing
Re-design Research
Stop Research

Curation

- Backup
- Store Checklist
- Preservation Checklist

1.	PDF	14,052
2.	Text	6,590
3.	HTML	365
4.	Microsoft PowerPoint	306
5.	Slideshow	252
6.	Microsoft Word	167
7.	Binary	105
8.	Image	81
9.	Audio (RealAudio)	68
10.	Image (JPEG)	65
11.	Archive	41
12.	Text (Rich Text)	28
13.	Other	23
15.	Video	9
16.	Text (OpenDocument)	6
17.	XML Word Processing Document (DOCX)	6
18.	Image (PNG)	5
20.	Audio	4
21.	Postscript	4
22.	Microsoft Excel	4
23.	Image (GIF)	4
24.	Plain Text	4

Over 35 file formats

When preparing to collect research data, you should chose **open**, **well-documented** and **non-proprietary formats** wherever possible.

The choice of format will vary depending on how you plan to analyze, store and share your data.

Useful guides on formats

Data Citation

Data citation refers to the practice of providing a reference to data in the same way as researchers routinely provide a bibliographic reference to outputs such as journal articles, reports and conference papers.

Main information required:

DataCite

- Who produced the dataset (creator or author);
- The title of the dataset;
- The unique identifier of the dataset, preferably a Digital Object Identifier (DOI) or minimally a link to the dataset if it is online;
- The date the dataset was published and its version number, if it has one;
- The date and time the dataset was accessed;
- The distributor of the dataset.

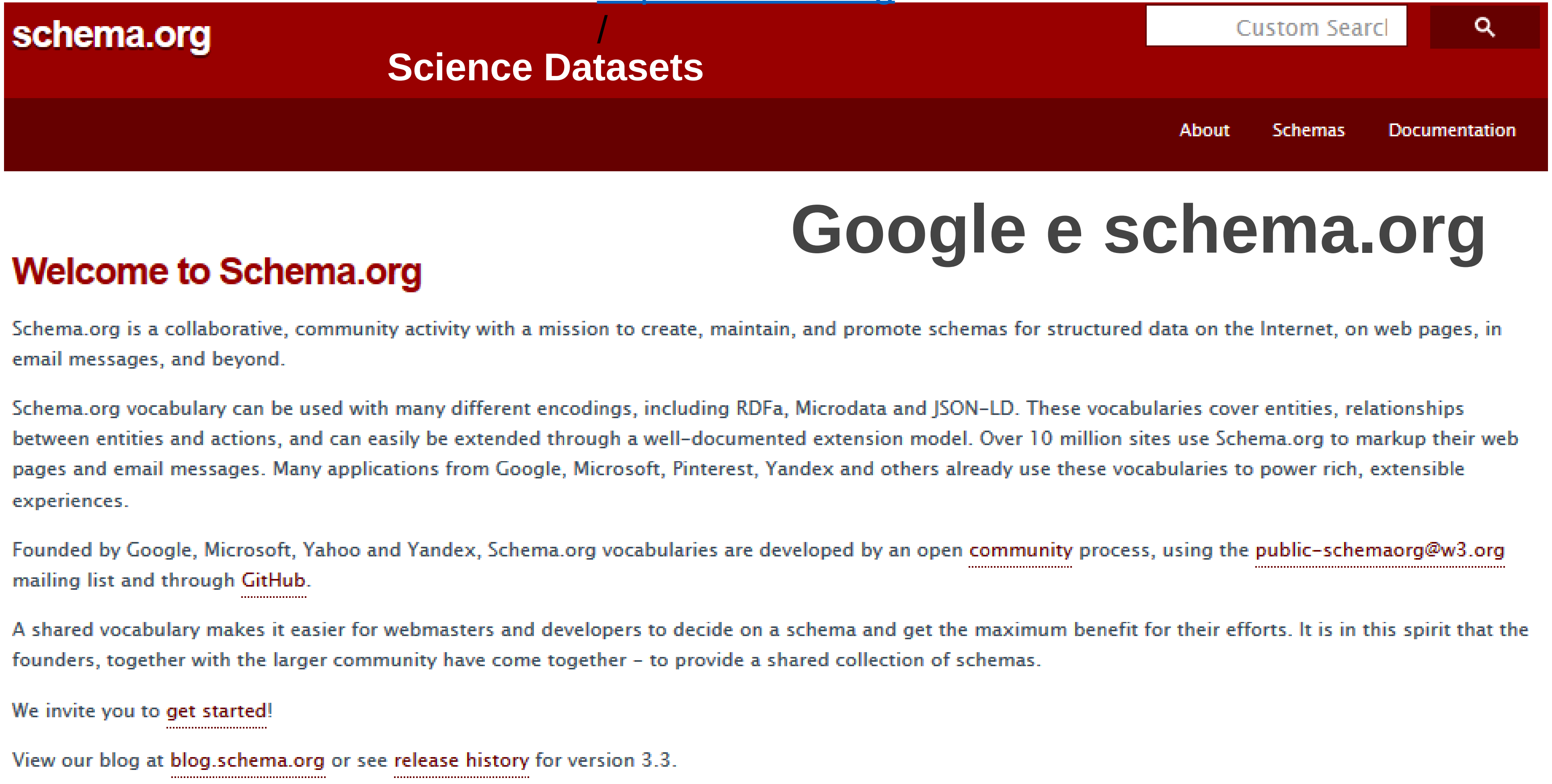
Important elements in citing data, regardless of citation style, publisher or repository guidelines, can be found in this short overview by Purdue University.

https://www.ands.org.au/working-with-data/citation-and-identifiers/data-citation

Data discovery

http://schema.org/
Science Datasets

Google e schema.org



Welcome to Schema.org

Schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond.

Schema.org vocabulary can be used with many different encodings, including RDFa, Microdata and JSON-LD. These vocabularies cover entities, relationships between entities and actions, and can easily be extended through a well-documented extension model. Over 10 million sites use Schema.org to markup their web pages and email messages. Many applications from Google, Microsoft, Pinterest, Yandex and others already use these vocabularies to power rich, extensible experiences.

Founded by Google, Microsoft, Yahoo and Yandex, Schema.org vocabularies are developed by an open community process, using the public-schemaorg@w3.org mailing list and through GitHub.

A shared vocabulary makes it easier for webmasters and developers to decide on a schema and get the maximum benefit for their efforts. It is in this spirit that the founders, together with the larger community have come together - to provide a shared collection of schemas.

We invite you to get started!

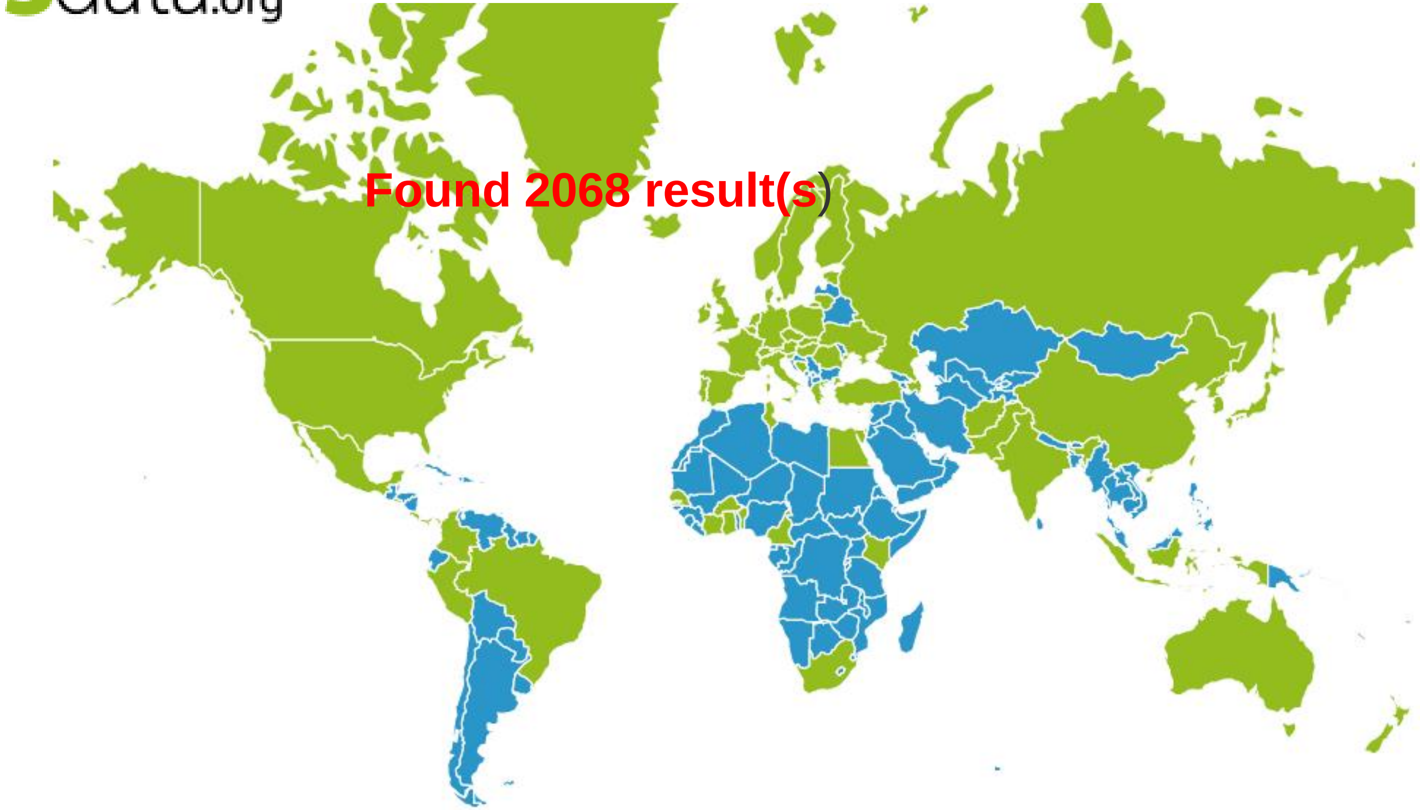
View our blog at blog.schema.org or see release history for version 3.3.

https://researchdata.jiscinvolve.org/wp/2016/11/04/google-role-research-data-discovery/

Reliability of Data Repositories: which repository for my data? Re3data.org

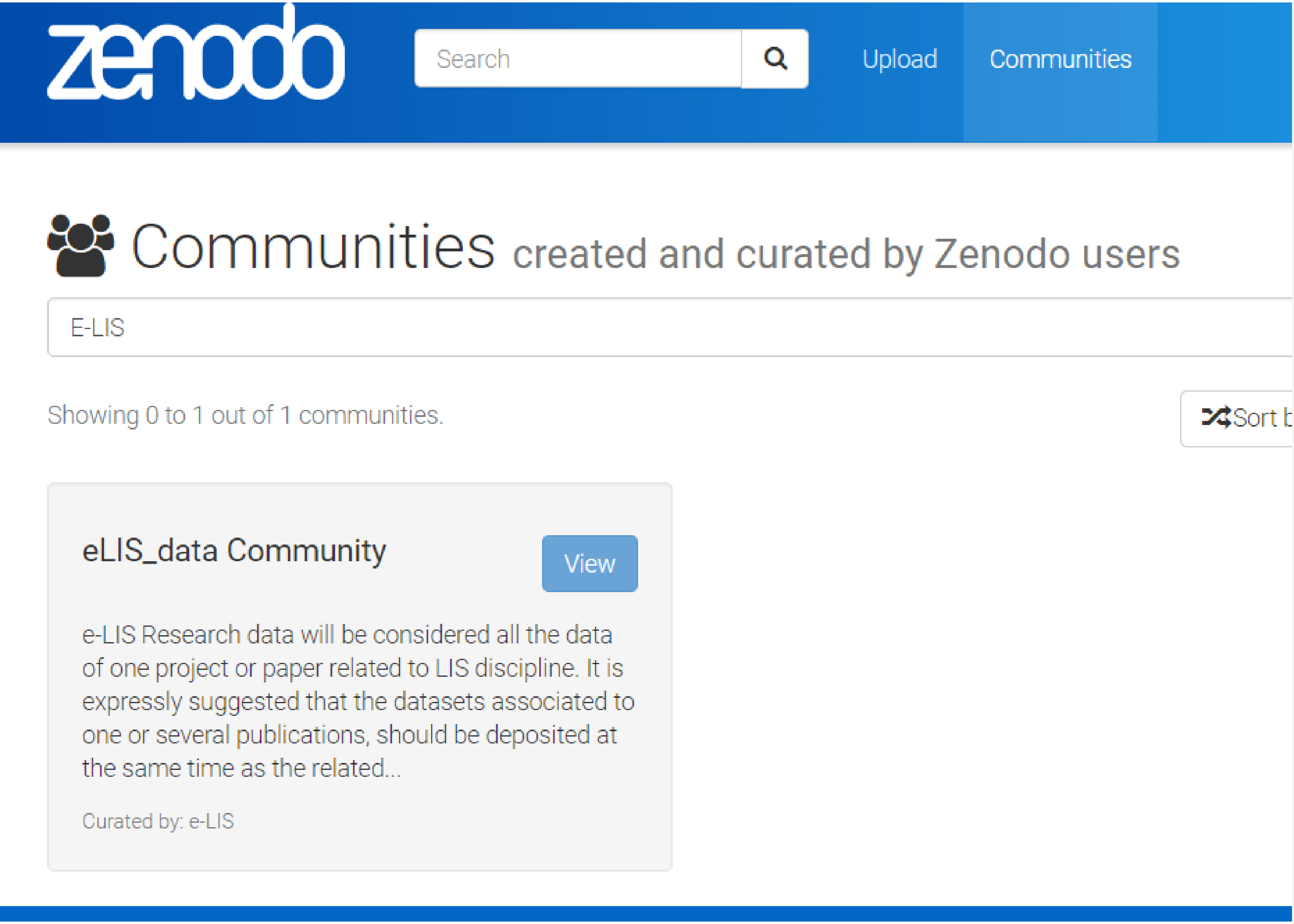
Data repository Directory

re3data.org



re3data.org Metadata Schema for the Description of Research Data Repositories. Version 3.0, December 2015. doi:10.2312/re3.008

... towards Zenodo



Communities created and curated by Zenodo users

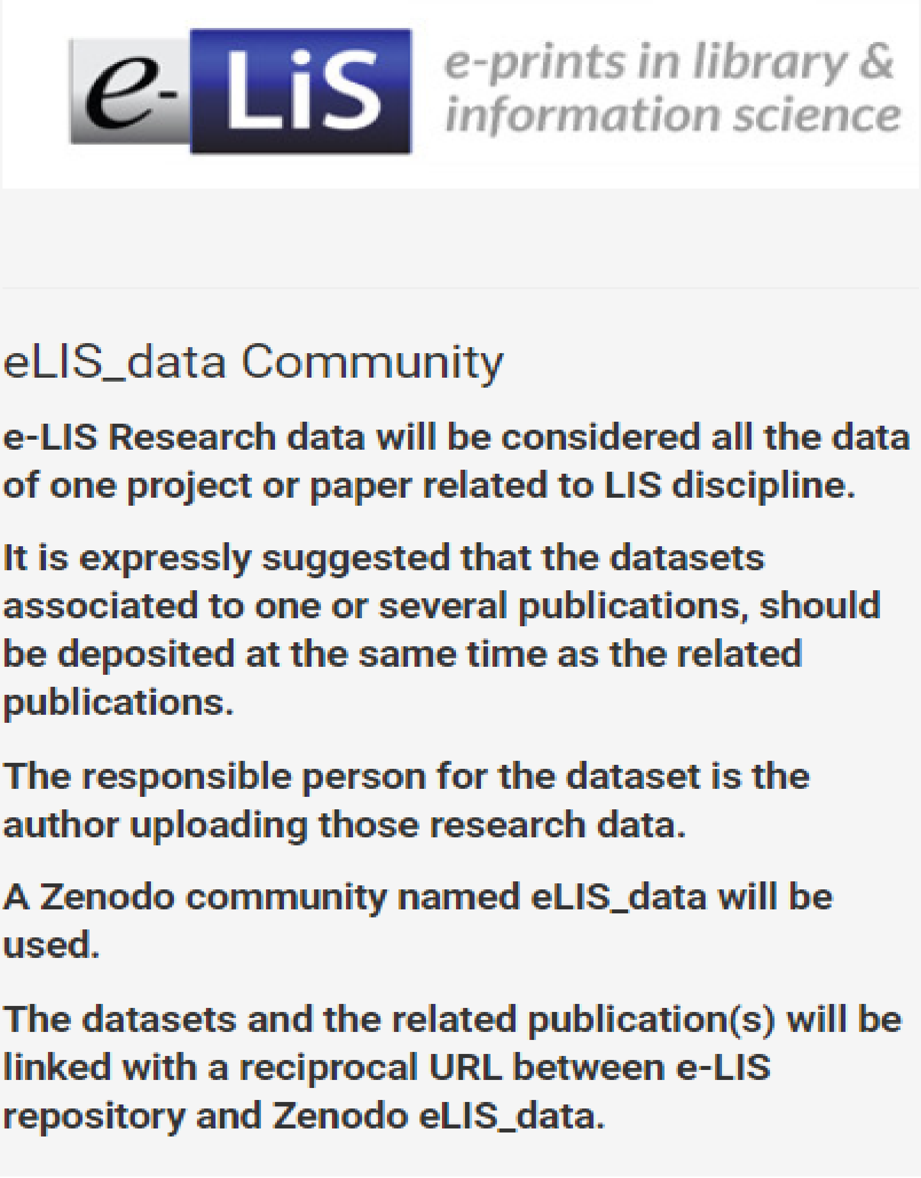
E-LIS

Showing 0 to 1 out of 1 communities.

eLIS_data Community

e-LIS Research data will be considered all the data of one project or paper related to LIS discipline. It is expressly suggested that the datasets associated to one or several publications, should be deposited at the same time as the related...

Curated by: e-LIS



eLIS_data Community

e-LIS Research data will be considered all the data of one project or paper related to LIS discipline.

It is expressly suggested that the datasets associated to one or several publications, should be deposited at the same time as the related publications.

The responsible person for the dataset is the author uploading those research data.

A Zenodo community named eLIS_data will be used.

The datasets and the related publication(s) will be linked with a reciprocal URL between e-LIS repository and Zenodo eLIS_data.