Enhancing diffusion of scientific contents: Open data in Open Archives

Daniela Luzi, Rosa Di Cesare*, Roberta Ruggieri#, Marta Ricci**
*National Research Council, Institute of Research on Population and Social Policies
Via Palestro, 32, Rome, Italy
#Senate of the Republic, Piazza Madama, Rome, Italy

1. Introduction

The free availability of data gathered during research activities is becoming one of the new challenges facing the Open Access Movement. New scientific instruments and technologies used in highly collaborative fields such as molecular biology, astronomy and environmental sciences, make it possible to collect a great amount of data in different formats. Moreover, data are often associated with tools that can aggregate them as well as with direct references to the publications – conventional or non-conventional – that report the results of their analysis. The benefits of the availability of these data are evident, and include assessment of research results, along with the reproduction and re-utilisation of data, potentially to draw new insight for future research.

According to the National Science Foundation: “digital data are the currency of the data collection universe, which, like currency in the financial realm, comes in many different forms”. They are different in nature, generally depending on the very specific field of study; they are produced for different purposes using varying methods and/or instruments; they have their own lifecycle before they are “translated” into scientific results and diffused in scientific publications. Understanding all these aspects makes it possible to determine whether to preserve them and how, who is responsible for their curation and/or diffusion, what type of archive, or better infrastructure, should be developed. This in turn implies issues related with data ownership, as well as funding resources, types of institutions and services to be involved.

Several policy papers (NSF, 2005, OECD, 2007, US National Research Council, 1995; 1999) are advocating free access of datasets and are outlining recommendations to coordinate efforts for the development of successful data repositories and infrastructures. What is clear is that “*one-size-fits-all* approach to policy development is inadequate” (NSF, 2005).
That is why debate on data ranges from the analyses on issues related to data sharing (Gold, 2010, Piwowar et. al., 2010, Piwowar, 2011) to studies in specific scientific fields (NIH, 2003, 2007, Karasti, et al., 2006, Baker et al., 2009, Waaijers et al., 2011) including surveys on usage patterns (Brown, 2003, Piwowar et al., 2007,) and researchers’ attitude to make them available (Savage CJ, Vickers AJ (2009).

A few studies deal with the analysis of the existing dataset archives and compare their different characteristics (Marcial, 2010). Our paper intends to follow this type of survey, but with a different approach. In fact, the decision to use data archives listed in OpenDOAR enabled us to select a random sample given by the providers that had registered their archives in OpenDOAR. This approach throws light on an emerging reality such as IRs, that theoretically at least, have started to include datasets along with other digital objects. Insight can also be gained into archives of large scale and well-established datasets. Clearly, the adoption of a random sample affected our survey. In contrast to Marcial’s empirical survey mentioned above, that found a cluster of elements common to different archives, our study revealed elements of dataset archives listed in OpenDOAR, in order to bring them into line with traditional archive classification (Armbruster & Romary, 2010). This enables the tracking of possible trends in dataset archive expansion policy.

In this paper we present the result of an exploratory analysis of a dataset archive in OpenDOAR. After the dataset definition given in paragraph 2., we describe the method used to select the sample and their main variables. In the fourth paragraph the results of our survey are reported.

2. Dataset definition

Research data are complex objects (Borgman, 2010) and that explains why there is no common agreed definition. They are very generally described and definitions, especially those reported in policy documents, include a very broad variety of digital objects (see Box 1). This is evident if we consider the definition given by the U.S. National Research Council in 1995, where research data are exclusively associated with numerical quantities. Following definitions encompass a wider range of digital objects (for instance images, sounds, etc.), thus representing research outputs in all scientific fields.

A more general agreement is reached when it comes to the definitions of dataset, considered as a meaningful and systematic representation of the subject being investigated. What is importantly stressed here is the importance of its re-use for validation and future investigations.

In this paper the term dataset is used to denote the digital collections managed in data archives.

Box 1. Data definitions

Research data definitions

- National Research Council (1995): *Data are numerical quantities or other factual attributes derived from observation, experiment or calculation*
  http://www.nap.edu/catalog.php?record_id=4871)
- National Research Council (1999): Data *are facts, numbers, letters, and symbols that describe an object, idea, condition, situation, or other factors.*
  (http://www.nap.edu/openbook.php?record_id=9692&page=15)
- National Science Foundation (2005): *The term 'data' is used in this report to refer to any information that can be stored in digital form, including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc. Such data may be generated by various means including observation, computation, or experiment.*
  (http://www.nsf.gov/pubs/2005/nsb0540/start.jsp)
- OECD (2007): *Research data are defined as factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings.* (http://www.oecd.org/dataoecd/9/61/38500813.pdf)
- PARSE.INSIGHT (2009): *Digital research data is used for all output in research. In practical terms, raw data, processed data and publications are all covered by the same term. A distinction between these sorts of research data is only made when necessary (for example when policies for publications are compared with other data).*
  (http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf)
- HLWIKI Canada (2011) *Research data is often defined as the information (e.g. data sets, microarray, numerical data, clinical trial information, textual records, images, sound, etc.) generated or used as quantitative evidence in primary biomedical research. This research data is distinguished by the fact that it is accepted by the research community as a means to validate research findings, observations and hypotheses.*
  (http://hlwiki.slais.ubc.ca/index.php/Data_curation)

Dataset definitions

- ODLIS (Online dictionary for library and information science): *A logically meaningful collection or grouping of similar or related data, usually assembled as a matter of record or for research, Also spelled* dataset.
  (http://www.abc-clio.com/ODLIS/odlis_A.aspx)
- OECD (2007): *A research data set constitutes a systematic, partial representation of the subject being investigated*
  (http://www.oecd.org/dataoecd/9/61/38500813.pdf*)*
- DOE (Department of Energy): *No-text scientific and technical information*
  (http://www.osti.gov/data/index.shtml)
- University of Edinburgh: *A set of files containing both research data and documentation sufficient to make data re-use.*
  (http://datashare.is.ed.ac.uk/)

3. Methods

The information source of our analysis was the directory OpenDOAR (The Directory of Open Access Repositories) that currently lists more than 2000 repositories worldwide providing a detailed description of each of them. OpenDOAR categorizes and provides access to Institutional and Subject-based repositories, but also includes open access archives developed by funding agencies, governmental institutions and digital libraries.

The inclusion of different types of archives allowed us to analyse:

- Types of archives that collect datasets;
- Types of providers;
- Relationship between dataset characteristics and types of archives.

Moreover, OpenDOAR archive description, built on the information submitted by their providers, are then categorized, allowing users to sort the listed archives according to different criteria. We used the option "dataset" reported in the OpenDOAR content type categories to identify our first sample of analysis.

The purpose of our analysis was to track current trends in the development of data archives in the general framework of open access repositories, using the random sample provided by OpenDOAR listed archives. For these reasons, the OpenDOAR archive classification needed to be supplemented with the additional categories: *Directory and Digital library.* This was necessary in order to group archives with features different from "traditional" IRs or Subject-based repositories. Moreover, the category Digital library was introduced, even if limited to a single case, to show trends in data archives provided by libraries that may make their collections available and re-usable in digital forms.

The second step of our analysis concerned the identification of datasets provided in each archive of the OpenDOAR sample, which was performed searching for dataset, if the archives had this search option, or analysing the archives' collections manually.

According to the NSF definitions for *data origin* and *digital data collections* the sampled archives were analysed in terms of:

- *Data origin* (experimental, observational, computational)
- *Types of Data collection* (Research data, Resource or community, Reference data collections)

Moreover, datasets were classified as follows:

- Dataset content (Numeric, Scientific image, Image of artifacts, Maps, text-image)
- Dataset format
- Contextual information associated with datasets ("traditional documents", project descriptions, etc.).

Archives with a limited number of datasets (i.e. > 4) were excluded, as they were considered not representative of a stable commitment to dataset collection. Moreover, archives containing video, audio or other multimedia, were not considered in our analysis, this can be the subject of further analysis.

The latest update of our analysis was completed in October 2011.

4. Results

4.1. The sample

In OpenDOAR there are 80 archives that claim to contain datasets in their content type. The analysis of each of the selected OpenDOAR archives showed that only 29 out of 80 actually contain datasets, while 13 archives were discarded for the limited number of datasets available. In 33 archives no datasets at all were found, whereas the remaining 7 archives were not accessible (fig. 1).

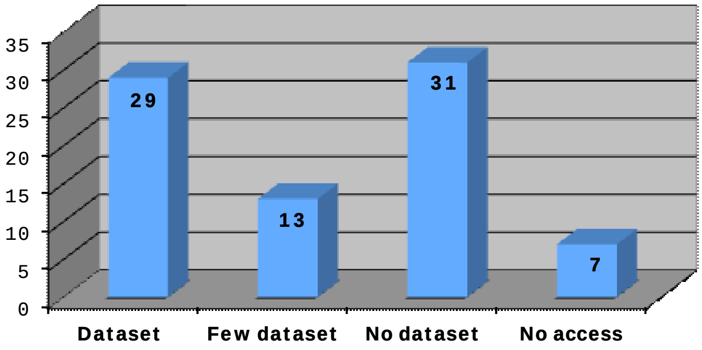

Fig. 1. - Dataset archives listed in OpenDOAR

Our sample of analysis consequently numbers 29 archives. Given the variety of archives listed in OpenDOAR, it should be noted that 59% of them (equal to 17 archives) exclusively contain datasets, while 41% (equal to 12 archives) contains both datasets and other digital objects, such as journal articles, reports, theses, etc. (fig. 2).

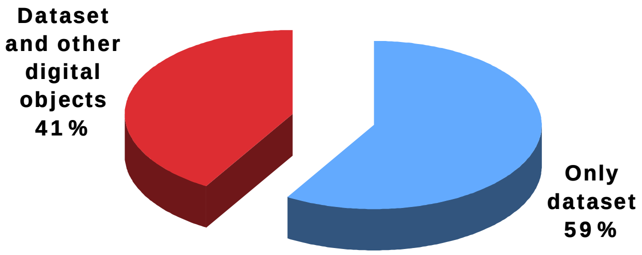

Fig. 2 Sampled archives by type of digital objects (n= 29)

4.2. The data archives' providers

In the analysis of the type of providers that insert datasets in their archives we also wanted to verify whether they are single institutions or have built consortia. Our hypothesis is that consortia may have developed internal rules, specific metadata and or format to describe and exchange data to be shared within a specific scientific community. Results are reported in table 1.

Table 1. Dataset providers by type of organisation

Research Institution	
Single (15)	**Consortium (5)**
• *Chiba University*, JP • *Spanish National Research Council, ES* • *Cambridge University Library and Computing Service, UK* • *Data Archiving and Networked Services (DANS)*, NL • *University of Southampton (Soton), UK* • *Data Library, University of Edinburgh, UK* • *Inter America Institute for Global Change Research (IAI)*, BR • *International Food Policy Research Institute (IFPRI)*,US • *University of Minnesota* ,US • *Monash University Library* - Australia • *Scripps Institution of Oceanography (SIO)*,US • *University of Delaware Library*, US • *University of Hull* ,UK • *Centre de Données astronomiques de Strasbourg (CDS)*, FR • *Marine Biological Laboratory & Woods Hole Oceanographic Institution (MBL & WHOI) Library*, US	• *Mineralogical Society of America, Mineralogical Association of Canada, University of Arizona, Schweizerbart Science Publisher*, INT • *COD Consortium*, INT • *Center for Research Libraries (CRL)*, US • *Alfred Wegener Institute for Polar and Marine Research (AWI), Center for Marine Environmental Sciences (MARUM), University of Bremen*, DE • *Department of Geosciences, University of Arizona University of Arizona (UA), CALTECH (California Institute of Technology)*, US
Indexing abstracting service	
Single (3)	**Consortium (1)**
• *Archaeology Data Service*, UK • *National Center for Biotechnology Information (CBI)*, US • *National Library of Medicine (NLM)*, US	• *Ontario Council of University Libraries*, CA
Publisher	
Single (1)	**Consortium (1)**
• *FigShare*, UK	• *Dryad*, INT
Government	
Single (3)	**Consortium (0)**
• *Coordenação de Biblioteca / CGDI / SAA / SE, Minisétrio da Saúde*, BR • *U.S. Department of Energy (DOE)*, US • *Deutschen Zentrum für Luft- und Raumfahrt (EDINA)*, DE	---

The majority of the providers of dataset archives are research institutions (20 out of 29), among which 8 universities and 7 research institutes.

Datasets are also available in archives developed by Indexing/abstracting services, governmental institutions and publishers. Such providers reflect the growing interest in datasets to be diffused for different purposes. At governmental level, for instance, the request for open data has been met by different countries that are progressively diffusing data collected within their institutions. In our sample we found the Brazilian Health Ministry, a German governmental agency for transport, and the U.S. Department of Energy that has a long tradition in the diffusion of technical information. Further, the presence of publishers represents the tendency to request datasets together with journal articles. In our sample a consortium of scientific journal publishers has developed Dryad that allows authors to submit their data and connect them with peer-reviewed articles. Similar features are provided by the publisher FigShare that provides citations of the datasets downloaded by authors.

4.3. Types of archives

In the analysis of type of archives we have adopted the traditional distinction between IR and subject based repositories. This classification is influenced by the information source we have chosen for our analysis, that has the advantage of exploring small dataset collections and verifying whether IRs are also beginning to consider datasets in their research results.

We introduced the category *Directory* to group heterogeneous types of archives, websites of governmental institutions, large databases that provide access to different data sources.

In OpenDOAR we also found an archive in the form of a Digital library, which we included in our analysis because we consider it a good example of providing a re-usable dataset from digitalised documents. In fact the South Asian Digital library not only digitalised an old text containing statistical data from the colonial period, but also provided an excel file that reported the datasets of the document. In our opinion this is a good example of making datasets re-usable, even if they are not digitally born.

Table 2. Archives by type

Subject-based Repository (15)
American Mineralogist Crystal Structure Database – http://rruff.geo.arizona.edu/AMS/amcsd.php
Archaeology Data Service - http://archaeologydataservice.ac.uk/
Crystallography Open Database (COD) - http://www.crystallography.net/
EDNA-the e-depot for Dutch archaeology - http://www.dans.knaw.nl/en/content/categorieen/projecten/edna-e-depot-dutch-archeology
eCrystals - Southampton) - http://ecrystals.chem.soton.ac.uk/
IAI Search - http://mercury.ornl.gov/iai/
Metropolitan Travel Survey Archive - http://www.surveyarchive.org/
PubChem - http://pubchem.ncbi.nlm.nih.gov/
PANGAEA® (Publishing Network for Geoscientific and Environmental Data) – http://www.pangaea.de/
RRUFF Project - http://rruff.info/
ShareGeo Open - http://www.sharegeo.ac.uk/
SIOExplorer Digital Library Project (SIOExplorer) - http://siox.sdsc.edu/
Verkehrsmodelle – http://modelle.clearingstelle-verkehr.de/
VizieR Catalogue Service - http://vizier.u-strasbg.fr/
Woods Hole Open Access Server (WHOAS) - https://darchive.mblwhoilibrary.org/
Institutional repository (7)
Chiba University's Repository for Access To Outcomes from Research (CURATOR)- http://mitizane.ll.chiba-u.jp/curator/
Digital.CSIC - http://digital.csic.es/
DSpace @ Cambridge - http://www.dspace.cam.ac.uk/
Edinburgh DataShare - http://datashare.is.ed.ac.uk/
Monash University ARROW Repository - http://arrow.monash.edu.au/vital/access/manager/Index
University of Delaware Library Institutional Repository - http://dspace.udel.edu:8080/dspace/
University of Hull Institutional Repository - https://hydra.hull.ac.uk/
Directory (6)
Biblioteca Virtual em Saúde - http://bvsms.saude.gov.br/php/index.php
Dryad - http://www.datadryad.org/
FigShare - http://figshare.com/
IFPRI Publications (International Food Policy Research Institute Publications) - http://www.ifpri.org/publications
OSTI (Office of Scientific & Technical Information) - http://www.osti.gov/
OZone (OZone provided by Ontario Scholars Portal) - https://ospace.scholarsportal.info/
Digital Library (1)
DSAL (Digital South Asia Library) - http://dsal.uchicago.edu/

4.4. Which science area?

The majority of dataset archives in our sample cover hard science (52%), but there is also a meaningful percentage of archives that provide datasets in Humanities and social sciences (fig. 3).

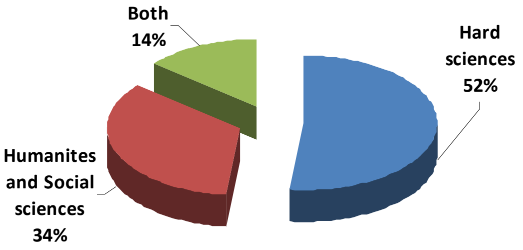

Fig. 3 Distribution of archives by science area (n=29)

If we group them in broad disciplinary fields, the most prevalent are Environment (21 %) and Demography (21%) (fig. 4).

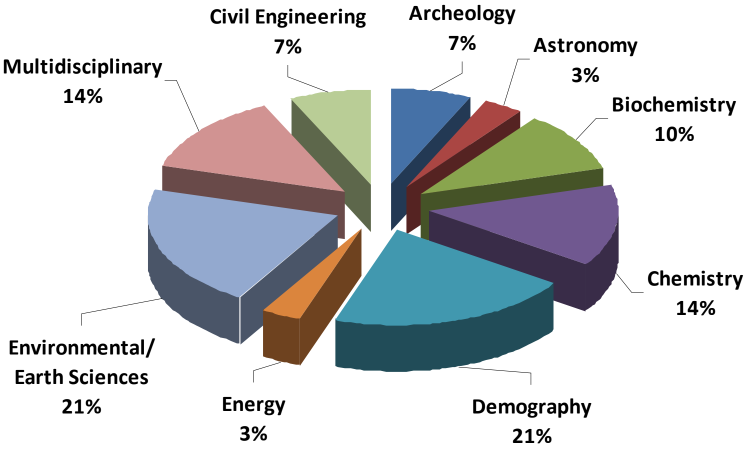

Fig. 4. - Distribution of archives by disciplinary fields (n=29)

Important criteria for the analysis of datasets depend on their origin, that is whether they are produced measuring specific phenomena at a given time, or are generated by experiments, or developing computational models or simulations to predict certain phenomena. Further, these variables are important when deciding whether it is important to preserve the data, considering that some of them cannot be so easily reproduced and/or collected. Figure 5 shows this variable linked with the scientific area.

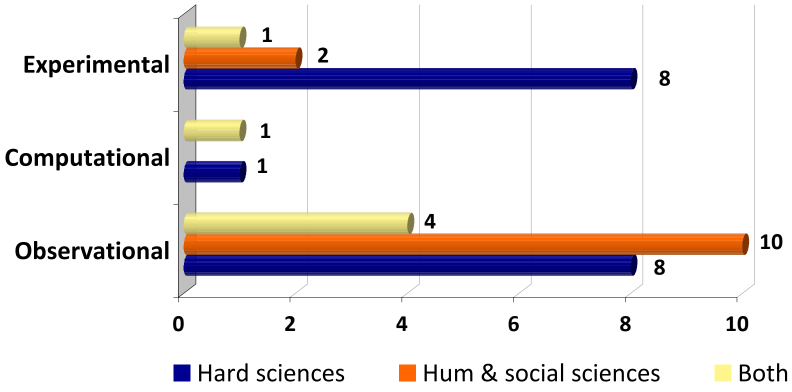

Fig. 5 Archives by data origin

The dataset archives in our sample are predominantly observational, and this is true in all science areas. Experimental data are collected mainly in hard sciences.

4.5. Functional categories of digital data collections

The National Science Foundation introduced three functional categories to analyse data collections referring to databases, infrastructures and organisations and individuals essential to managing this collection (NSF, 2005). This classification aims to distinguish between research data collected within a project of a certain size and budget as well as with different types of funds and funding sources. This distinction is also made to evaluate efforts necessary to preserve and diffuse datasets. Of course, a Research data collection can progressively become a Resource or Reference data collection, this was the case for instance of the well-known Protein data bank.
We applied these categories to the archives listed in OpenDOAR and compared them with the type of archives (fig.6).

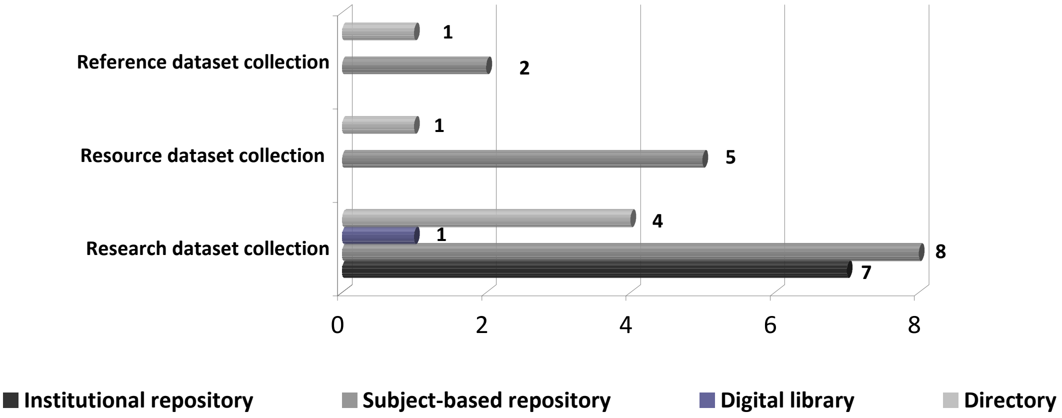

Fig. 6. – Archives by digital data collection

IRs exclusively contain datasets that fall into the category of Research data collections. Subject-based repositories contain datasets in all 3 categories, with a prevalence of Research data collections, while directories contain 1 Reference data collection.

4.6 Dataset content

For each archive in our sample we examined datasets with a view to analysing their content.
Figure 7 shows that the majority of archives contain numeric data, followed by scientific images, maps, text-images (i.e. digitized text) and images of artefacts.

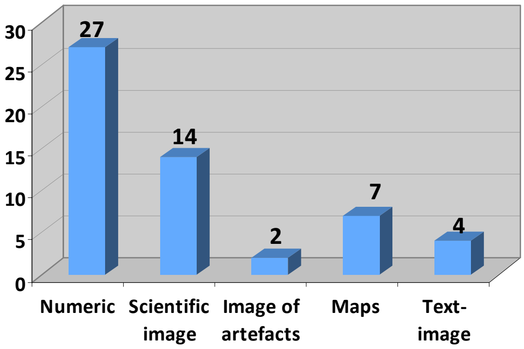

Fig. 7 Dataset content in our sample

Considering that the results of scientific observations, experiments or computational models can be expressed using different representations, not limited to numeric values, we associated the numeric content with other types of representation. Figure 8 shows the number of archives that only contain numeric datasets and/or images and those that associate numeric data with other digital objects.

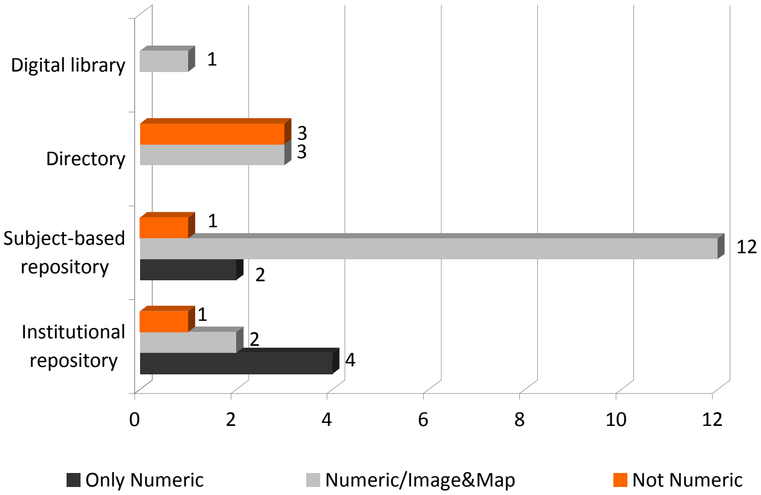

Fig. 8 Dataset content by type of archive

If we relate the content of a dataset with the type of archive, we see that there is a tendency to represent research results through numeric data associated with images. This is evident particularly in the case of Subject-based repositories (12 archives), while Institutional repositories tend to collect only numeric datasets (4 out of 7). Of course depending on the subject, some datasets are represented only by images (i.e. Not numeric) and this is present in all the kinds of archives in our sample.

The research results in the Subject-based repositories of our sample seem to provide a richer representation of datasets as a whole. For instance in the case of crystallography, the crystal structure described in the CIF format (see below) is combined with the graphical representation of its chemical structure, adding value for both crystallographers and chemists. (Cragin et al., 2010)

4.7 Dataset format

On the one hand file formats give evidence of the content of dataset (formats used to view images, texts and/or to store structured data already recognisable from their format extension). On the other, they also show how easily datasets can be exchanged. For instance the use of flat files, that is files that transform a record of a database into text, can be easily exchanged because they are not connected with proprietary systems. The disadvantage of using this format is that one needs to have additional information to interpret the data. In the archives listed in our sample we found different formats and sometimes the same archive provides the dataset in different formats so that users can easily access the data in the format he/she prefers. It follows that data format can also be considered an indicator of sharing and re-use. The formats more commonly used in our sample are reported in table 3.

Table 3. Dataset formats in our sample

File type category	File type/extension
Flat files	.txt, .ascii, .csv
Word processor	.doc, .pdf
Image	.tiff, .jpeg, .gif, .jmol
Spreadsheet	.xls
Statistical analysis	SPSS

As sharing and re-use are crucial for the dataset environment, we also looked for other file formats that facilitate their exchange. We found that some datasets were associated with the so-called readMefile, that contain important information, such as copyright, or how to install the database. We found readMefile especially in IRs (57%) and in directories related to Research data collections (67%).

It is of course the development of a specific standard format to exchange datasets that assures the highest degree of exchange and re-use. Their application indicates that a certain scientific community has a tradition in data sharing and has already agreed upon an exchange format that has a specific structure and meaning. An example of this standard is the CIF format used in crystallography to describe crystal structures or the standard used in astronomy to describe latitude, longitude and size of astronomical objects. It comes as no surprise that such exchange formats were present in Subject-based repositories (53%) and both in Resource and Reference data collections.

4.8 Datasets and "traditional documents"

Usually datasets are not self-describing, we need to know the context in which they are produced, how, and in which period, etc. Moreover their analysis can be described in other "traditional" documents, such as journal articles, reports, and theses (fig. 9).

In our sample we found that in the majority (72.4%) of archives, datasets are linked with traditional documents, and this is true for all types of archives.

Some archives also connected the dataset with the description of the project in which datasets were collected: this we found especially in large Subject-based repositories. Other archives also described the entire collection and this was the case especially in IRs.

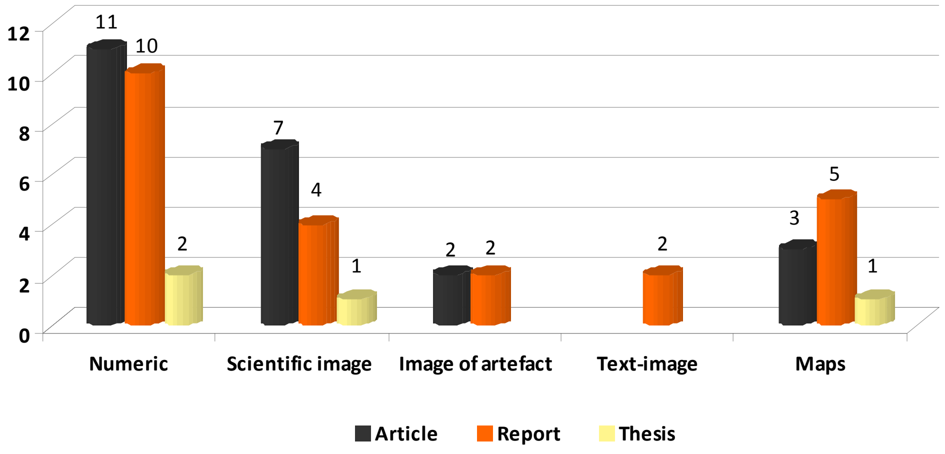

Fig. 9. – Dataset content by document type

5. Conclusions and discussion

Our sample enabled us to determine some common features, but also some characteristics that while not so widespread, may indicate possible trends in the development of dataset archives.

Considering the providers, our sample shows that along with research organisations and data services, governmental institutions and publishers too are developing archives making datasets available to the public. This is in line with the policy on open data announced in some countries as well as with the tendency of publishers to require datasets together with articles they are going to publish. Consortia are also frequently involved in building dataset archives, confirming the importance of collaboration in this field. Further analysis on the types of consortia (based on scientific collaboration, funding resources, and/or organisational models) should be carried out.

Datasets are collected in IRs along with other digital objects, while the majority of Subject-based repositories of our sample contain exclusively datasets. The introduced category Directory represents another way of organising data archives, combining different databases and linking various information sources. In our sample we also had an example of a Digital library that made datasets re-usable and an IR that contained only datasets.

Archives specifically focused on datasets and on specialized sub-disciplinary fields provide a richer environment in terms of data representations, of development of specific formats that facilitate data exchange, and of re-use and links to other digital objects and/or documents. This was evident in the Subject-based repositories and in Directories in our sample. This does not exclude that IRs cannot contribute to the collection and diffusion of datasets. Certainly, given the variety of datasets and their close relationship with the sub-disciplinary field in which they are collected, this poses different issues, such as self-archiving procedures and attitudes, ownership and copyright of data as well as their updating and maintenance. In this respect, the data collection categories proposed by the NSF provide a useful interpretative key and also suggest procedures to adequately construct and store data collections according to the type of archive and the mission of the archives' provider. In our sample, we had a prevalence of Research data collection in IRs, representing the outcomes of specific scientific projects with a limited user community and budget. Dataset availability in IRs along with other scientific results provide a more complete description of the research activities carried out in scientific institutions, while efforts concerning their visibility and usability should be further improved.

References

Anderson W.L. (2004). Some Challenges and Issues in Managing, and Preserving Access to, Long-live Collections of digital Scientific and Technical Data. *Data Science Journal*, 3.

Arms William Y., Larsen Ronald L. (2007). The Future of Scholarly Communication: Building the Infrastructure for Cyberscholarship.
URL: http://www.sis.pitt.edu/~repwkshop/NSF-JISC-report.pdf

Armbruster Chris, Romary Laurent (2010). Comparing Repository Types: Challenges and Barriers for Subject-based Repositories, Research Repositories, National Repository Systems and Institutional Repositories in Serving Scholarly Communication.
International Journal of Digital Library Systems, 1 (4).
URL: http://arxiv.org/abs/1005.0839

Baker Karen S., Yarmey Lynn (2009). Data Stewardship: Environmental Data Curation and a Web-of-Repositories. *The International Journal of Data Curation*, 4 (2).

Borgman C.L., Wallis J.C., & Enyedy N. (2007). Little Science Confronts the Data Deluge: Habitat Ecology, Embedded Sensor Networks, and Digital Libraries. *International Journal on Digital Libraries*, 7 (1-2).

Borgman L. Christine (2010). Research Data: Who Will Share What, With Whom, When, and Why?
URL: http://works.bepress.com/cgi/viewcontent.cgi?article=1237&context=borgman>

Brown C.M. & Abbas J.M. (2010). Institutional Digital Repositories for Science & Technology Information: A View from the Laboratory. *Journal of Library Administration Special Issue: Emerging Practices in Science and Technology Librarianship*, 50:181–215.

Brown C.M. (2003). The Changing Face of Scientific Discourse: Analysis of Genomic and Proteomic Database Usage and Acceptance. *Journal of the American Society for Information Science & Technology,* 54(10): 926-938.

Cragin Melissa H., Palmer Carole L., Carlson Jacob R., & Witt Michael. (2010). Data Sharing, Small Science, and Institutional Repositories. *Philosophical Transactions of the Royal Society A*, 368(1926): 4023-4038.

Gold Anna (2010). Data Curation and Libraries: Short-term Developments, Long-term Prospects. *Data Curation and Libraries*, 4.

Graaf Maurits van der, Waaijers Leo (2011). KE Knowledge Exchange Primary Research Data Working Group. A Surfboard for Riding the Wave: Towards a Four Country Action Programme on Research Data.
URL: http://www.voced.edu.au/content/ngv48428>

Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council (2009). Harnessing the Power of Digital Data for Science and Society.
URL: http://www.nitrd.gov/About/Harnessing_Power_Web.pdf

Joint Information Systems Committee (JISC) Managing Research Data (MRD) Programme (2009).

URL: http://www.jisc.ac.uk/whatwedo/programmes/mrd.aspx

Karasti Helena, Baker Karen, Halkola Eija (2006). Enriching the Notion of Data Curation in E-science: Data Managing and Information Infrastructuring in the Long- term Ecological Research (LTER) Network. *Computer Supported Cooperative Work (CSCW)* 15: 321-358.

Marcial Laura Haak, Hemminger Bradley M. (2010). Scientific Data Repositories on the Web: an Initial Survey.
URL: http://onlinelibrary.wiley.com/doi/10.1002/asi.21339/pdf

National Institutes of Health (2006). Data Sharing Policy and Implementation Guidance.
URL: http://grants2.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

National Research Council (1995). Preserving Scientific Data on our Physical Universe: a New Strategy for Archiving the Nation's Scientific Information Resources. Washington D.C: National Academy Press.
URL: http://www.nap.edu/catalog.php?record_id=4871

National Research Council (1997). Bits of Power: Issues, in Global Access to Scientific Data. Washington, D.C.: National Academies Press.
http://www.nap.edu/catalog.php?record_id=5504

National Research Council (1999). A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases. Washington, DC: National Academy Press.
URL: http://www.nap.edu/openbook.php?record_id=9692&page=14

National Research Council (2003). Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences. Washington, D.C.: National Academy Press.
URL: http://selab.janelia.org/publications/Cech03/Cech03-reprint.pdf

National Science Foundation (2005). Long-lived Digital Data Collections: Enabling Research and Education in the 21st century.
URL: http://www.nsf.gov/pubs/2005/nsb0540/start.jsp

National Science Foundation (2011). Dissemination and Sharing of Research Results.
URL: http://www.nsf.gov/bfa/dias/policy/dmp.jsp

OECD (2007). OECD Principles and Guidelines for Access to Research Data from Public Funding.
URL: http://www.oecd.org/dataoecd/9/61/38500813.pdf

OpenDOAR (2011).
URL: http://opendoar.org/

PARSE.INSIGHT (2009). Insight into Digital Preservation of Research Output in Europe.
URL: http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

Pirowar Heather A., Day R.S., Fridsma D.B. (2007). Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLoS ONE* 2(3): e308. doi:10.1371/journal.pone.000030
URL: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.00003088

Pirowar Heather A., Chapman Wendy W. (2010). Public Sharing of Research Datasets: a Pilot Study of Associations. *Journal of Informetrics 4 (2): 148-156.* doi:10.1016/j.joi.2009.11.010.

Savage C.J, Vickers A.J (2009). Empirical Study of Data Sharing by Authors Publishing in PLoS Journals. *PLoS ONE* 4(9): e7078. doi:10.1371/journal.pone.0007078

Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E., Manoff M. Frame M. (2011). Data Sharing by Scientists: Practices and Perceptions; *PLoS ONE* (6)6
URL: http://dx.doi.org/doi:10.1371/journal.pone.0021101

UK data archive (2011). Managing and Sharing Data. [3rd ed].
URL: http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

Waaijers Leo, Graaf van der Maurits (2011). Quality of Research Data, an Operational Approach.
URL: http://www.dlib.org/dlib/january11/waaijers/01waaijers.html

Whitlock M.C. (2011). Data Archiving in Ecology and Evolution: Best Practices. *Trends in Ecology & Evolution* 26: 61-65. doi:10.1016/j.tree.2010.11.006.