DOI: http://dx.doi.org/10.20396/rdbci.v14i3.8646333en

ARTICLE

THE ROLE OF LIBRARIANS IN SCIENTIFIC DATA MANAGEMENT

O PAPEL DOS BIBLIOTECÁRIOS NA GESTÃO DE DADOS CIENTÍFICOS

EL PAPEL DE LOS BIBLIOTECARIOS EN LA GESTIÓN DE DATOS DE INVESTIGACIÓN

Fabiano Couto Corrêa da Silva[1]

ABSTRACT

We present an analysis of the possibilities of the librarian profession with researchers for efficient management of scientific data. The result obtained through a survey of tools and currently available management techniques, demonstrates an analytical framework of support actions that librarians can provide for the development of a project on the life cycle of scientific data. It is concluded that the scientific data management requires planning solutions that include specific knowledge about the repository choice and storage techniques for the conservation and the permanent use of the data as a key to the success of a research project.

KEYWORDS: Scientific data. Data life cycle. Digital preservation. University Libraries. Scientific Repositories. E-science. Data curation. Free access. Academic librarians.

RESUMO

Apresentamos uma análise das possibilidades de atuação do bibliotecário em conjunto com pesquisadores para um gerenciamento eficiente de dados científicos. O resultado obtido, por meio de um levantamento de ferramentas e técnicas de gestão atualmente disponíveis, demonstra um quadro analítico de ações de apoio que os bibliotecários podem fornecer para a elaboração de um projeto para o ciclo de vida dos dados científicos. Conclui-se que a gestão de dados científicos exigem soluções de planejamento que incluem conhecimentos específicos sobre a escolha do repositório e técnicas de armazenamento para a conservação e o uso permanente dos dados como chave para o êxito de um projeto de pesquisa.

PALAVRAS-CHAVE: Dados científicos. Ciclo de vida dos dados. Preservação digital. Bibliotecas universitárias. Repositórios científicos. E-ciência. Curadoria de dados. Acesso livre. Bibliotecários acadêmicos.

RESUMEN

Se presenta un análisis de las posibilidades de acción del bibliotecario en conjunto con los investigadores para una gestión eficiente de los datos científicos. El resultado obtenido a través de un análisis de herramientas y técnicas de gestión disponibles en la actualidad, demuestra un marco analítico de acciones de apoyo que los bibliotecarios pueden ofrecer para el desarrollo de un proyecto para el ciclo de vida de los datos científicos. Se concluye que la gestión de datos científicos requieren soluciones de planificación que incluyen conocimientos específicos sobre la elección de las técnicas de depósito y almacenamiento para la conservación y el uso permanente de los datos como una clave para el éxito de un proyecto de investigación.

PALABRAS CLAVE: Datos científicos.Ciclo de vida de los datos. Preservación digital. Bibliotecas universitarias. Repositorios científicos. E-ciencia. Curaduría de datos. Acceso libre. bibliotecarios académicos.

[1] Doutorando em Información y Documentación en la Sociedad del Conocimiento (Universidad de Barcelona), Professor do Instituto de Ciências Humanas e da Informação (ICHI) da Universidade Federal do Rio Grande (FURG). Porto Alegre, RS. Email: fabianocc@gmail.com - ORCID: http://orcid.org/0000-0001-5014-8853.
Submitted: 03/08/2016 – **Accepted**: 30/08/2016.

1 INTRODUCTION

Scientific data preservation and exchange have become topics of interest on an international scale for managers, funding agencies and researchers in general. Currently, the demand for management of scientific data support requires that librarians understand and meet the new demands of the researchers, not only as consumers of information, but also as producers (TENOPIR, et. Al, 2015).

Given this new scenario in the scientific communication flow, librarians are aiding researchers at a broader level during the research process, rather than focusing solely on formal means of scientific communication.

From the access to research data point of view, libraries are developing support services during the phases of scientific data life cycle (CARLSON, 2012), ie, when researchers are generating and using the data in their planned work. Often, these services should be provided in close collaboration with researchers and may include the development of management plans to document and organize data through the use of tools or resources to safely store data.

One of the main scientific data management problems is associated with the lack of continuity of data recording due to the departure of any project member; so it is normal that difficulties would arise to reuse a set of data, since, without proper documentation, it is difficult to understand how, when and why the data were captured (CARLSON, 2012).

Efficient data management reduces the amount of work required for the interpretation and compilation of information obtained at the end of a research project when all is documented, thus the ongoing investigation does not need to be rebuilt at a later date. Therefore, the librarian assistance to researchers helps evaluate what they really need to understand, to understand how to organize a variety of data types and make the right decisions on access and preservation of data for their projects.

2 GOALS

a) To place a theoretical framework on the importance of scientific data in libraries;
b) To propose initiatives for the organization of scientific data

The research aims to offer recommendations for choosing scientific data management repositories. It comes to defining recommendations for the preservation and reuse of data from the raw survey data.

3 METHODOLOGY

To perform a qualitative analysis of the scientific data repositories typology and point recommendations for librarians, we adopted the model suggested by Higgins (2007), member of the DCC (Digital Curation Centre), and related the set of actions that the librarian can offer to the scientific community in its projects.

4 STAGES OF THE SCIENTIFIC DATA LIFE SCICLE

The planning of scientific data life cycle is an important tool to outline the necessary infrastructure for the preservation of data. Each stage of the cycle requires different levels of organization and planning is essential to describe the necessary procedures (HIGGINS, 2007). The following describes the fundamental steps for data preservation.

4.1 First stage

The first stage is data attainment. At this time, researchers must find an efficient way to store the collected data, chose a metadata strategy and ensure that all survey participants understand the metadata schema.

There is a range of support means in which the librarian can get involved, from the development of a management plan or metadata strategy, to choosing a suitable repository for the researchers' needs.

After the data has been created or collected, the researchers start to process it. This envolves the transition, digitalization, validation, cleaning and storing of data during the registration process. The librarians can offer storage solutions or tools to help researchers generate the investigation metadata. During the analysis stage, researchers interpret the data and develop the investigation. It is also possible to offer training workshops for the researchers on the data management plan that can be projected to integrate an investigation project from the beginning.

4.2 Second stage

The next step is the preservation of data, which involves migration to suitable formats for preservation, including creating backups and generating additional metadata. At this stage, librarians can help researchers who do not know the requirements for data preservation, providing information about preservation formats, or migrating data directly. It is important to consider that many repositories have restrictions for data access. For example, when the data is stored in a repository it is also possible to determine when the datasets will be publicly available. This is an useful option for when a researcher wants the data to be preserved, but is not prepared to make them available for public consultation. Furthermore, most of the data repositories are under development and it may be difficult to distinguish the

most suitable for the specific needs of each investigator, so the use and verification of possibilities for each should be a continuous process.

The first question to ask about a repository is who is in charge of it; for example, the research investor or the university. For a researcher or research group that develops long term, it will be important that the repository is well managed over time.

The second question is for how long the data should be stored. This question is not always explicitly answered, but if the deposit is based on LOCKSS (Lots of Copies Keep Stuff Safe), CLOCKSS (Controlled LOCKSS) Portico2 is a good sign. These services ensure that data remains available even if the repository is not. Another issue is in what way the repository will index the data sets, ie, the file formats and metadata standardization (CORREA, 2016).

After choosing the repository, the researcher must provide information on how the data will be used and the access requirements, such as restrictions and information to assess the quality of data. It is very important that the implementation of a data repository be a project where the data life cycle involves joint decisions between researcher and librarian, so that the needs and technical possibilities can be addressed in the management plan. All these factors help ensure that the datasets will follow a usable format, and have controls that facilitate their search in other repositories. Finally, it never hurts to ask other researchers and librarians what they recommend for data preservation in their field.

If a researcher wants to share data from a published article, the first place to check is the magazine that published the article. A growing number of journals require the exchange of data and recommends specifically where data must be deposited to facilitate the peer review process. For example, the Scientific Data magazine has a list of recommendations repositories for authors (Scientific Data, 2015). Some magazines facilitate the inclusion of their data in a repository of their own, such as the magazines integrated with the Dryad repository (DRYAD, 2016). Understandably, not all magazines have these recommendations repository, but it is advisable to follow these guidelines, when available.

A second aspect to consider is where researchers in a particular research area share their data. Selecting a repository recognized by the scientific community in a particular area makes it more prone to be discovered by researchers in related fields and more likely to be cited.

It is also advisable to seek local options for data exchange through an institutional repository. One should take into account that some universities take part in data repositories consortiums, such as 4TU Centre for Reserch Data (2016), with the support of the Eindhoven University of Technology, Delft University of Technology and the University of Twente. If

[2] http://www.portico.org/digital-preservation/

an institutional repository is available, it is worth considering it as a good place to deposit the data, even if only to improve local service. As data management and exchange are becoming increasingly prominent, many institutions now offer support for the process of scientific data preservation.

4.3 Data management planning

The scientific data management planning is gradually becoming a requirement of investors and institutions that pay for data collection, as it shows concern for the data availability in the future. As can be seen in the calls, for example the Horizon 2020 program, in which the European Commission launched a pilot project called "Open research data pilot" to promote and optimize the management and reuse of research data generated by the projects it finances.

The collection, organization and preservation of information has been entrusted to librarians from different sectors who work with the information flow; the articulation of data management plans is simply a modern manifestation of these functions. One way to support many researchers at a time is offering workshops on the development of data management plans. Moreover, it is possible to offer individual consultations for the elaboration of data management plans, usually conducting interviews about the content they need to develop. Often, there are preservation and documentation problems that the researchers hadn't considered and the librarian, through these interviews, can alert them.

In general, researchers are not sure if they properly follow rules for filing and organization of scientific data, and are also not knowledgeable about intellectual property and have questions about the requirements for the development of data management plans. Researchers should ensure that institutional standards are being met and that all decisions are adjusted in accordance with these rules or with the specific project needs.

In conclusion, the data management plans can improve through a collaboration between librarians, professionals, researchers and repository, because they all present different experiences and knowledge to help create a data management plan. To make a simplified work method planning, either as a team or individually, you can use the Data Management Planning Tool (DMPTool) as an assistance to prepare data management plans through a drafting process. The application provides guiding questions to develop responses in accordance with the requirements set by financiers and can be used by anyone interested in the creation of data management plans. The goal is to document how data is produced or collected during an investigation and, after its completion, to define the way it will be described, shared and preserved. Thus, the whole process is formalized in a single document, much like a set of elements and information previously scattered between different people and documents, so that all necessary information is provided for monitoring of the project and its results.

The Digital Curation Centre (DCC) also provides a verification tool for data management plan that gathers information about the funding requirements and best practices in data management planning: the DMPonline. It is a program that asks a few questions in order to determine the appropriate model for every need and then provides a guide to help interpret and answer questions from the financiers.

4.3.1 Interviews with the researchers

Although preservation holds a very specific meaning for librarians, researchers have very different concepts on the archiving process. Librarians should begin the discussion on preservation making sure that both parties have a common understanding of terms.

The interview is the right moment to identify the requirements of the data management plan. Researchers at Purdue University (Carlson, 2012) conducted a study that indicates that librarians have specific skills to conduct interviews that may be useful for the development of data management plans, especially the ability to negotiate and manage user expectations. They point out that librarians need to know the resources available and the types of research being conducted in the researchers' institutions before making the first interview, as well as the types of resources available on campus departments and more frequent research practices. This does not mean that they need to have expert knowledge, because in many cases the librarian will learn more during the joint work with the researchers. The most important skills are knowledge about what questions should be asked and where to find the answers.

According to the study from Purdue University, a good way to start thinking on how to work with the researchers is to define a data curators' profile. Researchers at Purdue University Libraries and the Graduate School of Library and Information Science at the University of Illinois have developed a set of data curators' profiles describing how researchers create and manage primary data. They then organized a workshop and created a set of tools to help librarians understand what they needed to know before doing an interview. The product of this work, the Data Curation Profiles Toolkit, includes a worksheet for teachers and an interview manual for librarians. The manual's authors established a series of simple and easily understandable questions, searching for a common language between them. The questions allow librarians and researchers analyze the data life cycle, how information will be shared during the project and how the access should be provided both during the investigation and after its completion. The spreadsheet can also be useful for the researcher to describe all the problems related to data and view them in one place.

A similar study was conducted at the University of Colorado. Lage and Maness (2011) conducted extensive interviews with researchers and developed eight profiles that represent teachers and post-graduate students who were interviewed. These profiles show that the needs, practices and understanding of the research and the data vary widely between the different subjects. And although each institution, department and researcher exhibit unique

qualities, these profiles can be useful to understand the problems that arise when working with investigators. The profiles designed by the Colorado University team describe the level of interest and the support that researchers feel they receive, storage problems they face and the privacy that the data requires. One factor that the team observed is that there is not always a positive return between the level of support that researchers need and the librarian participation. Some interviews correlated positively the receptivity of a researcher for the librarians participation in data management. It also highlighted the lack of support for storage and data preservation, a positive bias toward the open access movement and lack of support for data management during the investigation process. The Colorado team also noted that the researchers in earth sciences seemed more open to the library and to provide their data. Those who work in highly competitive fields such as the sciences, proved to be less receptive to the participation of the library in the data management process.

Although Lage (2011) and his colleagues have discovered a wide range of attitudes towards data sharing and curation, there were some commonalities amongst many of those interviewed researchers. Most of them did not identify the data from his research as public data, which does not necessarily mean they were not open to sharing, since researchers often seek to maintain a certain level of control over the data they share.

The study by the University of Colorado also shows that researchers agree with the procedures or data storage departmental services, that most researchers have some research data subsets that are not being maintained or preserved with a developing plan and that they perceive the data management tasks as distractions from their research projects. These attitudes reveal that it is important for data management to be less complicated for researchers, and that planning should be considered before starting data collection.

After identifying the researchers receptive to working with the library, whom would request help with data management planning, it is time to conduct interviews particularly on their research projects. A draft of a management plan or a model of the data life cycle can guide the interview.

The librarian can start by asking if the researcher is aware of the tools that can be used to preserve data, and help identify file formats suitable for storage and storage plans.

After these questions, it is recommended to continue questions on security. It is useful in this case to be aware of the options for backing up data available for researchers, as is the knowledge of the advantages and disadvantages of different storage options. Data can be stored on internal or external local hard drives, cloud-based systems or servers. For more information about storage options, it is advisable to consult the information technology unit whenever possible. Garritano and Carlson (2009) suggest defining the workflow including processing and analytical steps performed after data collection. Carrying out a good data management is not about just selecting a good storage option, but also policies, best practices and support for backup and storage.

Researchers often keeps the collected data after the project is finished in the same place where they were stored during the process of investigation, regardless of how the procedure adopted may affect the long-term accessibility usability. Archiving data implies the active preservation of data, as well as the adoption of measures to increase the ability to find and access it. It is, among other things, recording unique identifier codes for the data and for the realization of common controls for its replication.

A data repository works by indexing processes that add value to the content available rather than simply saving it. It is important that researchers understand the difference between simple storage and correct filing. If researchers store digital data on servers or hard drives without regularly performing the necessary preserve actions, over time the data will become unusable. It is also necessary to know who owns and controls the data and if there are privacy issues. In many cases, researchers have no easy answers to these questions, so librarians can help them to go beyond the immediate answers. Librarians possess unique skills to organize disparate information in order to develop a research project and to find reference sources such as manuals, which may help researchers better understand the tools that can be useful, the context of the research and the required format index data. It is also advisable to talk with the project laboratory director or responsible for the data collection instruments: it can help understand how the data was generated or collected. If there is no main investigator of the project, it may be useful to discuss this workflow with the people who are gathering data directly. The librarian can also help researchers make connections with others in the same field, which can provide support with tools and all necessary information on the topic in question. This is the point where librarian skills and their ability to find information is useful because despite not knowing all repositories or metadata standards, the librarian can find the answers, thanks to their universal knowledge of information sources. Once the researchers are familiar with data management, from planning and managing metadata to long-term preservation, they will be more aware of the components they could leave out or include in the planning process to ensure that the preservation is effective.

In addition, an interview could be a conversation or a series of consultations to teach the resources they need through the life of the research data cycle. It is likely that the data store is already part of the work of a researcher, but the methods must be explicitly discussed and agreed upon at the beginning of a project. The workflow flows properly if each employee of a research project uses the same data storage methods and is familiar with them.

4.3.2 Repository selection

In the process of repositories selection, librarians can help understand the fundamental characteristics of each. For instance, thematic repositories provide visibility within the correspondent community, to the extent that the option for the institutional repository can be more efficient to give visibility to the development of researches with the closest pairs (for example, from the same institute in a university).

Publishers may accept datasets associated with articles; the librarian can also recommend the deposit on a third-party repository. Libraries can offer support to the publication on institutional repositories or instructions on how to deposit a dataset of a specific area.

There are several available options to deposit scientific data, such as re3data[3], which features an extensive directory of repositories in all subjects. The appropriate selection must be made at the beginning of a research project; It may also be appropriate to store data in more than one repository. For example, all project data can be stored in an institutional repository, as subsets of data can also be stored in specific domain repositories to increase the possibilities of being located. The librarian can help researchers in this decision process, researching available options and guiding them into making a meaningful data evaluation. The selection of the repository should take into account issues such as the existence of a specific repository for an appropriate discipline or a theme repository for a dataset access, the repository access and preservation policies.

The data submission process in the selected repository may be variable. Researchers must consult the repository requirements regarding the data preparation, for which most have a step by step guide with detailed instructions about the formats and other technical issues. Sometimes it takes the form of a delivery system via e-mail; thus, researchers can submit their data via email and the repository leaders perform the remaining work. Thus, librarians can assist researchers with data deposit or offering instructions with a step by step guide to common repositories with the help of library guides.

Eventually repositories are very specific and do not link to the publications and other datasets that contextualize them. Some may have inactive data packages due to lack of research continuity or other reasons. So it is advisable to check that the data type of the repository we are interested in is updated regularly.

To better understand the variety of existent repositories we stablished five categories which are described below, following exemples.

[3] Registry of Research Data Repositories. Available in: www.re3data.org

4.3.2.1 Institutional repositories

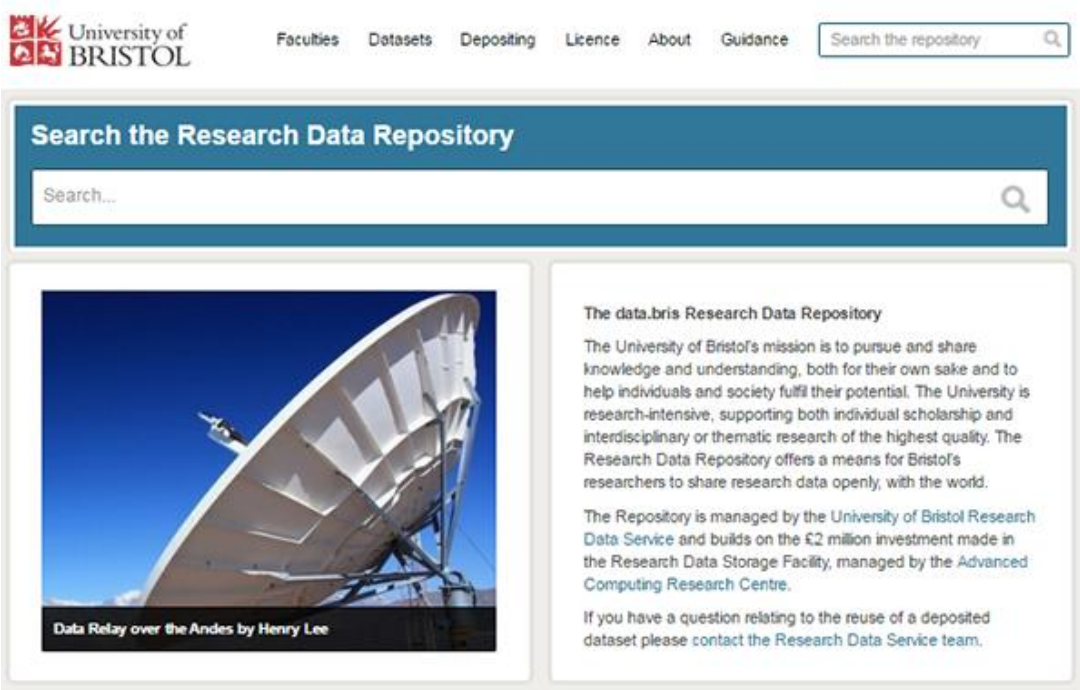

Figure 1. Institutional repository.
Source: University of Bristol http://data.bris.ac.uk/data/

Institutional repositories have gained notoriety in the 2000s with the emergence of software systems for their implementation such as Fedora and DSpace. They aim to collect, manage and maintain an academic or research institution's intellectual production and are platforms designed for the preservation and dissemination of scientific publications (articles, theses, administrative documents, etc) generated by the members of the institution. They facilitate researchers to automatically archive all publications, regardless of the original publication source.

Over time, repositories also started to allow the storage of data, facilitating the addition of basic and complex data descriptions, and generally emit identifiers that can be used to cite and retrieve data. Some institutional repositories offer also unlimited storage and, being supported by a university, they are usually administered by librarians.

Although institutional repositories provide confidence, they lack flexibility and control. Many have strict requirements to accept survey data files using very generic formats, missing Application Program Interfaces (APIs), making interoperability impossible with other systems and many use only a very general standard metadata like Dublin Core and do not support metadata fields domain of specific researches or specific data type and controlled vocabularies.

4.3.2.2 Theme repositories

The thematic repositories are those that include research data from a specific subject area. Some successful thematic repositories are arXiv, PubMed or Eprints.

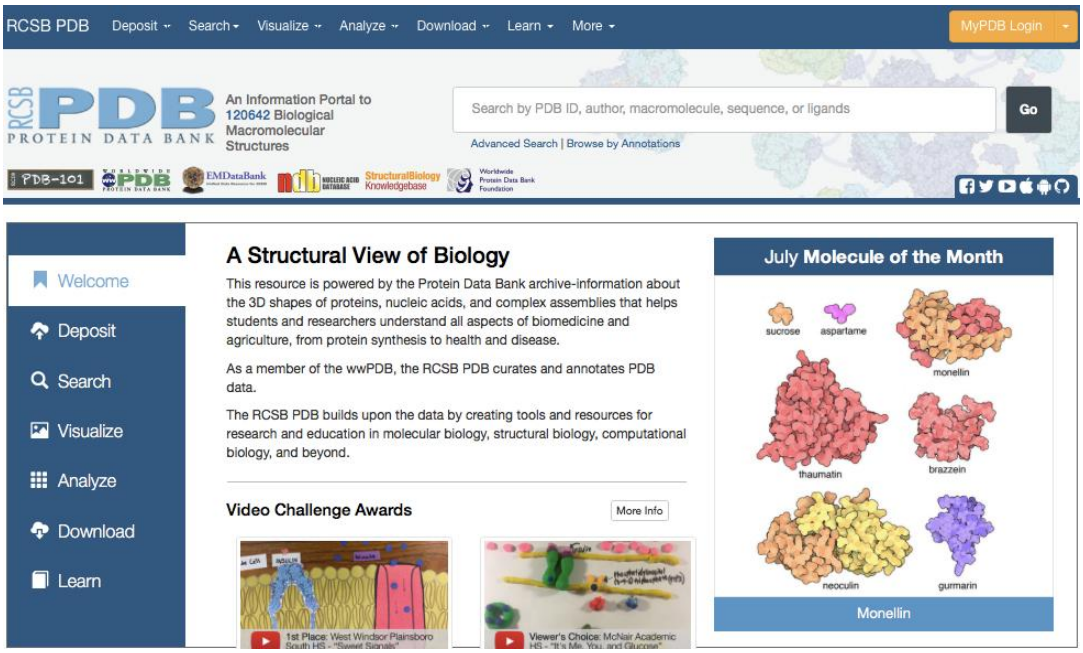

Figure 2. Protein Data Bank repository.
Source: Protein Data Bank www.wwpdb.org

Many subjects have repositories specifically designed for the data types in their domain. Some examples include Protein Data Bank of the Research Collaboratory for Structural Bioinformatics (RCSB, 2016), for 3D shapes of proteins, nucleic acids and complex sets; GenBank (2016), for DNA sequences; EMDataBank (2016), with 3D maps of electron microscopy density, atomic models and associated metadata; eCrystals to the crystallographic data of X-rays and the National Oceanographic Data Center (NODC, 2016) for oceanographic data.

Often times these thematic repositories have analytical and discovery tools available along with the data to promote its reuse. Some experts suggest that the data should be stored only in thematic repositories, because, they say, allows for specialized use of the metadata and further review and validation by field experts. However, not all subjects have data repositories and the specific nature and peculiarity of many data explains the difficulties in finding storage in existing repositories.

There are several databases that are maintained and administered by universities, research centers and government agencies, allowing free access to the raw data and processed information. OBIS (Ocean Biogeographic Information System) and MGDA (Marine Geophysical Data Access) serve as the most widespread examples in spatial databases at the

international level. Brazilian initiatives of the same type can also be cited, such as the National Bank of Oceanographic Data (BNDO), the Database for the Oil Industry (Banpetro), the Environmental Information System for the Biota/FAPESP (SinBiota), among others. The adoption of these can be an effective tool for data recovery, which is at times unavailable. However, most of these programs do not provide integration with different environmental databases, avoiding for example a correlation with geological, biological and hydrodynamic databases with the goal of planning and managing marine protected areas, especially in local and regional scales.

4.3.2.3 Editorial repositories

Editorial repositories offer similar characteristics to institutional, but with special features for specific communities.

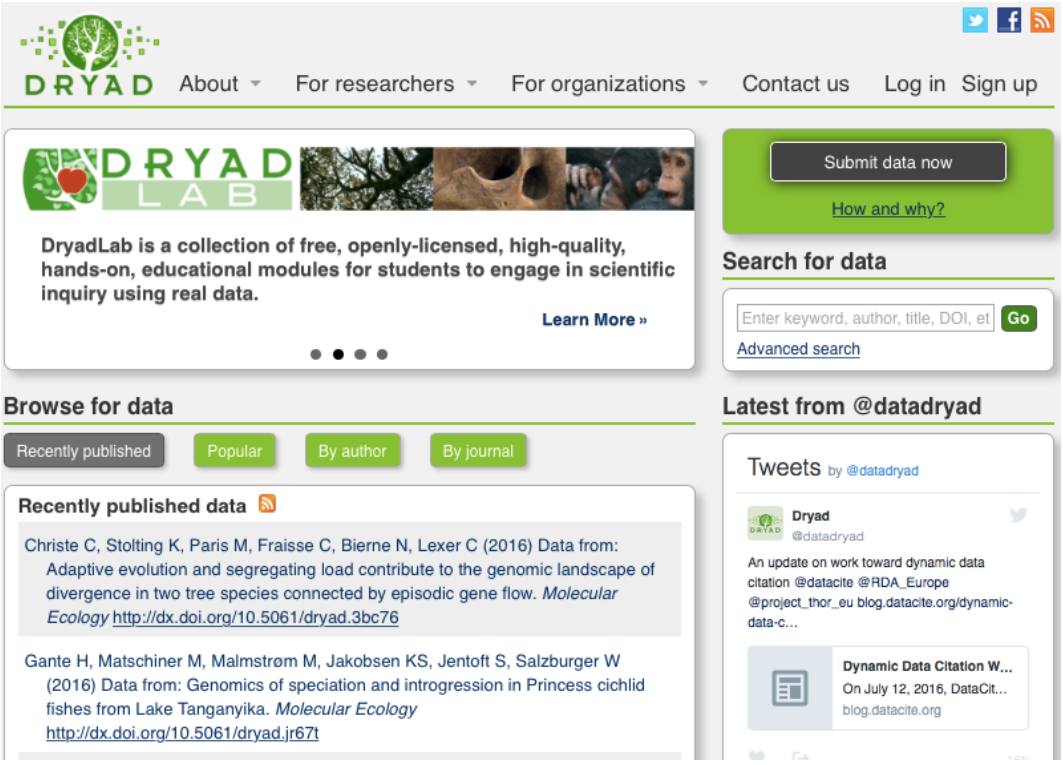

Figure 3. Dryad repository (registry example)
Source: Dryad datadryad.org

Dryad is an international scientific and medical research data repository, coming from scientific peer-reviewed journals. It works as a repository of various subjects and also facilitates Digital Object Identifier System (DOI) code, awarded by the EZID service of the Digital Library of California and registered by DataCite (DOI, 2016). Has both open access and commercial journals. One of its main features is the ability to accommodate any type of data (Correa, 2016). Moreover, Dryad minimizes the load of articles submissions, meaning

the repository makes an automatic reading the metadata provided by partner journals which provide bibliographic information for each item prior to publication.

4.3.2.4 General use repositories

It comes to repositories that any researcher can use, regardless of their institutional affiliation, to preserve any kind of academic production. The best known examples are Figshare (2016) and Zenodo (2016).

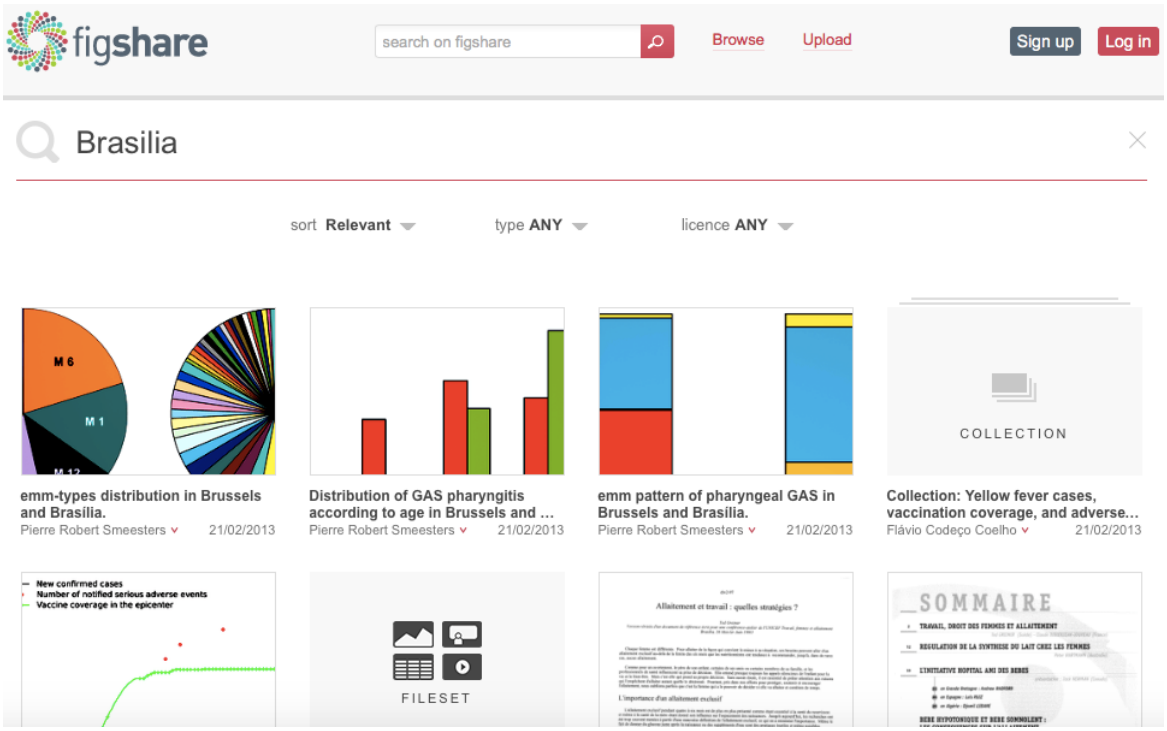

Figure 4. Figshare (results page).
Source: Figshare https://figshare.com.

Figshare is a platform created by Digital Science to share and display the results of multidisciplinary research and it is aimed at researchers, scientists, projects and institutions. It is currently associated with the F1000 Research (a repository of scientific articles recognized by the academic community for its high level of quality), collaborates with PLOS (the world's largest open access journal) and with Plum Analytics (a service that quantifies the impact of published research). All material published in Fisgshare is identified with a DOI to facilitate their location and the date. On the platform you can find: presentations, videos, posters, images, data, articles, etc. and the data preservation works with CLOCKSS technology, a nonprofit organization that promotes a partnership between publishing houses and academic libraries to sustainably archive all Web content produced in the scientific realm.

Users can integrate the data in a repository with other websites and blogs by copying and pasting a code. Readers can comment on the datasets and download files for their

reference managers for later use. The repository also offers the possibility to publish negative results or bad experiences so that other researchers can save efforts without having to go through already carried out tests as not to waste their time working in certain cases.

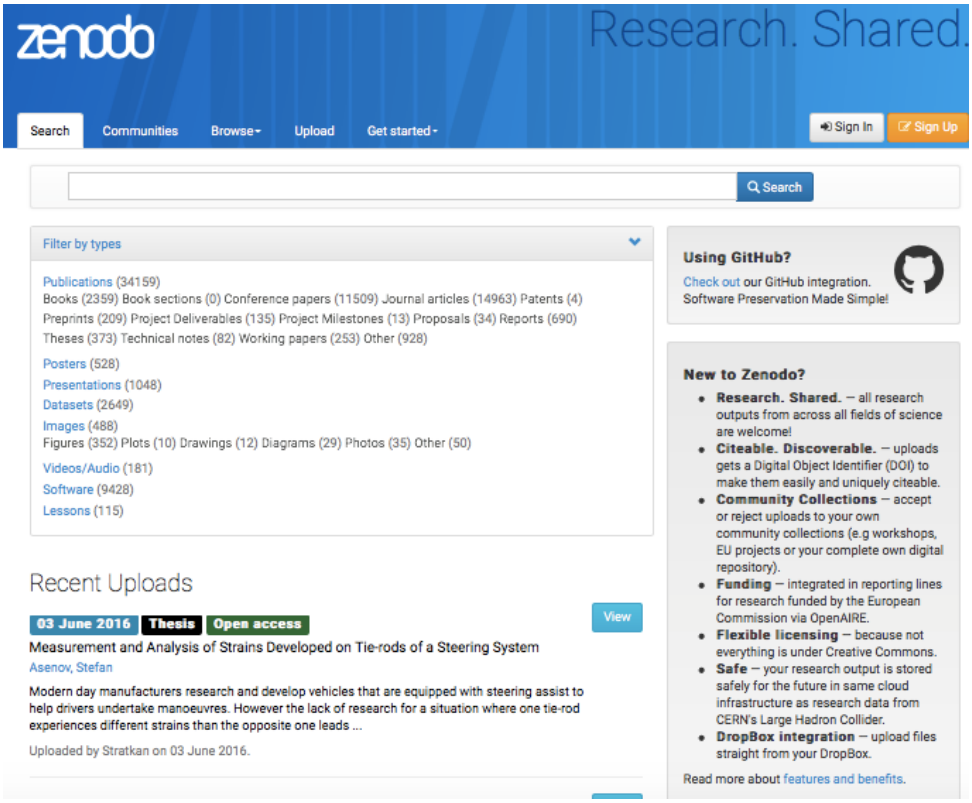

Figure 5: Zenodo repository (home page)
Source: Zenodo https://zenodo.org

Zenodo is an OpenAIRE portal initiative that provides adequate infrastructure to accommodate datasets and other European project's research results. It was developed with the Invenio platform by the European Organization for Nuclear Research (CERN), the center also responsible for managing the huge amount of data of the Large Hadron Collider data (LHC). As in the case of Figshare, access to the deposit is free, allocates DOI and allows datasets to be available in BibTeX, EndNote and other bibliographic formats. Users can add metadata to their files in a much more detailed way than in Figshare. All data are likely to be collected by other platforms through the OAI-PMH protocol (ZENODO, 2016). Zenodo drives data loading through communication with services such as Mendeley, DropBox, CrossRef and ORCID. It also includes long-term digital preservation strategies, allows to establish flexible licenses to manage rights and allows users to create their own collections in their own space using metadata under license dedicated to the public domain, without restrictions or authorization request (called CCO licenses), except for e-mail addresses. In addition, where permitted, other Zenodo users can review your files, and an interesting

feature is that it makes it easier to register with the ORCID identifier or GitHub account (Correa, 2016).

4.3.2.5 Repositorios próprios

Sometimes researchers archive their research data on a personal or projects designated cloud server. It is possible to find various technological options, including the free Dropbox, and other commercial ones, like Amazon Cloud Drive or Microsoft Azure. Keeping your own repository depends on a researcher's ability to accommodate the needs of their projects and to undertake appropriate actions for backup and replication. In this sense, it is one of the least recommended options due to its lack of guarantees in organization, maintenance and preservation.

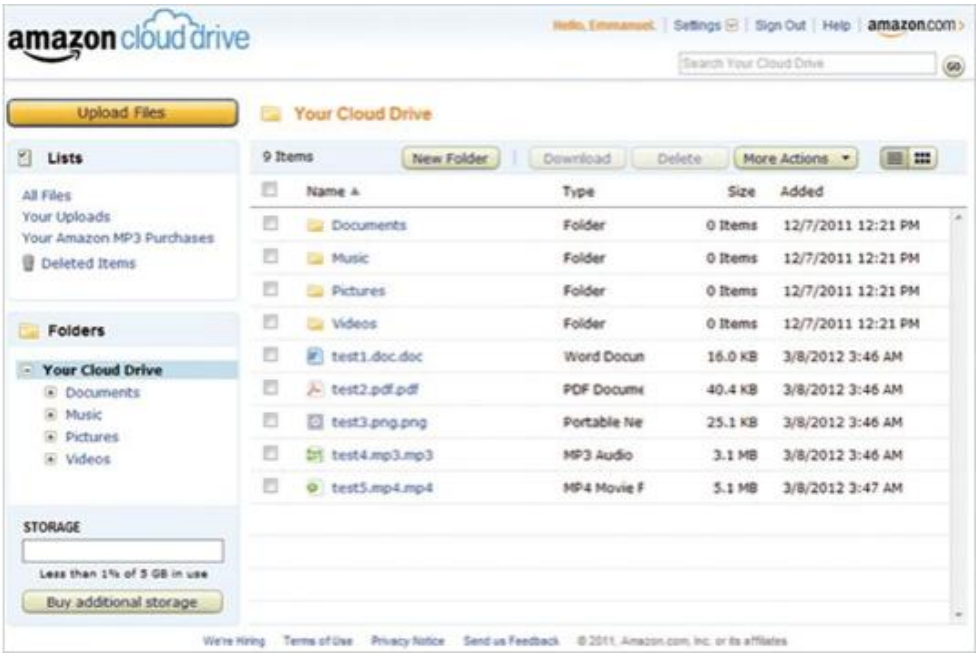

Figure 6: Amazon Cloud Drive repository (user page).
Source: Amazon Cloud Drive

4.3.3 The reusability of data

Although the requirements for the research data sharing are relatively new, management systems are still developing. So now the search for a particular dataset is not as easy as finding a published article. To find a dataset it is recommended to start by searching for articles on the topic of interest. Typically, the data will be disclosed as supplementary material, or an article link.

If the location of the data is not in the article, there are some alternatives. The first is to search for the author's curriculum or website to see if there are any references to the availability of data somewhere. If that does not work, it is possible to contact the author to request access to their data. The policies of some journals and funding agencies require a copy of the data, as long as the data are not confidential. None of these strategies is infallible since older data may become lost and email addresses are changed, but it can be a good strategy to gain access to data corresponding to an article.

If the search is directed towards a subject's general data and not the data of a specific article, the search strategy should be different. A good place to start searching for a topic is an index of specific themes of a specialty, when these are available. For example, the Integrated Ocean Observing System (IOOS) lists a wide range of marine resources on the web and has a search engine to help find specific data on oceans research. These indices do not necessarily collect data, but point to a number of resources on a particular topic together along with databases, which may also be available in libraries.

In the absence of a database or library, it is possible to consider searching in data repositories that are popular in particular fields and can be found in the re3data list. One should also take into account external data sources, such as governmental agencies, research foundations, special interest groups and other organizations, as they often make their data available. For example, the National Oceanic and Atmospheric Administration (NOAA) of the United States is an excellent resource for all that refers to climate-related data. As with any type of information, it is always recommended to evaluate the data source to ensure its credibility.

Finally, we know that as public data archiving becomes more common, it will be easier to find data to be reused. The investigation process is transitioning into a data exchange system, which means that many of its trading and reuse systems are under development so that in the future it will be as easy to find an article's data as it currently is to find the article itself.

4.3.4 Data reutilization's rights

Once researchers find the data they need for their research, they should consider what it is allowed to do with them. For example, when data is copyrighted, most copyright laws allow reuse, with some exceptions. Therefore, when other researchers use the survey data, it is necessary to analyze the conditions of licensing and copyright. The most common type of data reuse is opened or public domain license data. Publicly shared datasets are often under a Creative Commons license or a license from Open Data Commons (CORREA, 2016). The advantage of using the data in these conditions is that the license must clearly state what can and cannot be done, making the reuse rights clear. Such open licenses, with occasional exceptions such as the commercial research limitation, for example, allow both the reuse and

republish of the original license data. Basically, open license data usage is the best option, because it has relatively extensive permissions to doing what is necessary with the data.

Data obtained through financing can also facilitate reuse, although often with more restrictions so that access can be opened with a license. As funding agencies usually have rules stipulating data usage restrictions, it is about following the provisions of the contracts. It is necessary to keep in mind that contracts can prevail over the rights granted by copyright law. For example, if the researcher obtains a dataset under a contract that prohibits open access publication, it will not be possible to share.

When the data does not have licenses that allow reuse, usage becomes more complicated. In this case, it is usually allowed to conduct research on the material, but permission to republish the original data is required. When data is composed of natural facts, it is usually allowed to publish citing the source.

Proper documentation is one of the most important requirements for datasets reuse. For example, it is difficult for researchers to use a dataset if they are unable to determine the meaning of the variable's names. Therefore, it is advisable to start by looking for the documentation for the reuse of a dataset. The best datasets are the ones that offer detailed documentation, including the README file, or an index of datasets. The best strategy to overcome insufficient documentation is to contact the data creator for more information.

Errors in a dataset is another problem that can be found when using data from another researcher. Such errors include inconsistencies, null, lost or incorrect values. Even when no errors are found in a superficial analysis of the data, it is worthwhile to do some simple tests to verify its quality. For example, making a simple data graph is an easy way to check for errors before using a dataset for more complex analysis. Another advantage of running these quality controls is that you gain a better understanding of the data (CORRÊA, 2016).

4.3.5 Data life cycle regeneration

Reusing scientific data is the last stage of its lifecycle, but also the beginning of a new one. The data can be used in new projects and be incorporated into the collection, analysis and dissemination of results phase. Data plays an important role in all processes, and its importance will grow as the sharing and reuse becomes more prominent. Therefore, the researcher should start to think about the data as a product that needs to be preserved properly, from a research project to completion. Only through efficient data management, we can realize the full potential of scientific data.

Now that we have come to the end of the data lifecycle and demonstrated the role of the librarian in this process, we can reflect on the path for research data management. We have contemplated from data management planning, documentation and analysis, to the

sharing and reuse of data, to different types of storage and security. Through the analysis we did, both researcher and librarian can find tools that will help properly manage scientific data.

As the researcher explores, investigates and incorporates data management strategies in the workflow, they will remember important points. Firstly, it is not necessary to do everything at once. Data management is a collection of a number of small steps that add up to good practices. It is advisable to work in one practice at a time until a correct management of scientific data becomes a constant habit.

Data management also becomes easier with time. This happens partly because of the need to adjust the research workflows, adapting new strategies. The trend is that good habits in relation to the data become part of the investigations routine. The goal is to reach a point where a good data management simply converts into an additional process of an investigation.

Finally, it is important to remember that data management is a living process. It is not a goal to be reached so that the researcher does not have to redo the data management. Correct data management requires effort, but this effort will be rewarded later when the researcher does not have to spend hours trying to find, understand or reuse the data. By making data management conscious and continuous, the researcher can significantly reduce the daily frustrations that come to those who work with survey data. Data should work for the researchers and not against them; scientific research is already in itself difficult enough.

5 FINAL THOUGHTS

If before researchers produced knowledge by seeking documents in different repositories, today the scientific data gained prominent role for the advancement of scientific production. As a result, this new importance given to scientific data acquired new demands to preserve and restore it.

Planning the scientific data lifecycle is critical to start outlining the infrastructure that the researcher wants to offer. Each lifecycle stage requires different levels of organization and this model proposes recommendations for efficient data management.

As shown, the librarian has the necessary profile to assist researchers in developing a data management project and to indicate the required tools for this procedure. Moreover, the proper selection of a repository according to the requirements of certain datasets is imperative, a process in which both librarians and researchers can work together.

BIBLIOGRAPHIC REFERENCES

CARLSON, Jake R. Demystifying the data interview: Developing a foundation for reference librarians to talk with researchers about their data. *Reference Services Review* v. 40, n. 1. 2012. p. 7-23. Disponível em: < http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1186&context=lib_research >. Acesso em: 03 ago. 2016.

European Commission. Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020. v. 1.0. Dec. 2013.

GARRITANO, Jeremy R.; CARLSON, Jake R. A Subject librarian's guide to collaborating on e-science projects. *Issues in Science and Technology Librarianship*, n. 57, 2009.

LAGE, Kathryn; LOSOFF, Barbara; MANESS, Jack. Receptivity to library involvement in scientific data curation: a case study at the University of Colorado Boulder. *Libraries and the Academy*, v. 11, n. 4. 2011. p. 915–937.

SCIENTIFIC DATA. Disponivel em: < http://nature.com/scientificdata >. Acesso em: 03 ago. 2016.

3TU. Datacentrum, 2015. 3TU. Datacentrum. Disponivel em: < http://datacentrum.3tu.nl/. en/home. Acesso em: 03 ago. 2016.

ACKNOWLEDGMENTS

This work was made with the support of CNPq (National Council for Scientific and Technological Development – Brazil).

ISSN 1678-765X 17>
9 771678 765003