LIBER Case Study:

Raw Data at the Spanish National Research Council and Related Services by the Institutional Repository DIGITAL.CSIC

AUTHOR: Isabel Bernal, DIGITAL.CSIC Technical Office, URICI, CSIC, isabel.bernal@bib.csic.es

KEYWORDS: generic, institutional

1 What was the starting point?

The institutional repository of the Spanish National Research Council, DIGITAL.CSIC, was launched in 2008 as a project under the Unit of Information Resources for Research (previously called CSIC Libraries Coordination Unit), with a primary goal of collecting, describing and enabling open access and preserving scientific outputs (mostly peer-reviewed articles) by CSIC centres and institutes across 8 broad scientific areas (agricultural sciences, natural resources, physics, chemistry, materials sciences, food sciences and technology, humanities and social sciences). However, as time has passed by, we have broadened the scope of content accepted for submission due to interest raised by some researchers, and datasets have been included in the list of typologies since March 2010 (For more information about this, please refer to the DIGITAL.CSIC policies by visiting https://digital.csic.es/politicas/ (Spanish only). Data policies were published in October 2013 only. The first document in which we talked about research data in DIGITAL.CSIC involved an interview with the authors of this first dataset in March 2010 where they explained their motivation for doing so (CSIC Abierto, No. 1, pp. 2-4).

Since 2010, the technical office of the repository has been raising awareness about the possibility of housing certain types of datasets in DIGITAL.CSIC and about the very concept of disseminating raw data in an open access context along with the various benefits, management issues and other considerations and copyright issues. Such efforts to improve awareness of this topic have been targeted at both the CSIC scientific and the librarian communities. Most content in the repository is uploaded via the Mediated Archiving Service, a service offered by CSIC libraries and the repository technical office to authors, which is why we consider it essential for librarians to become familiar with the management of raw data as much as the researchers themselves. The DIGITAL.CSIC technical office has taken the lead in unfolding this new service through a number of initiatives that are explained below. Since last year, greater emphasis has been placed on this aspect of the repository.

2 What kind of research data is targeted?

CSIC is a multidisciplinary research-performing organisation, meaning that there are research groups in many different disciplines ranging from agricultural sciences to zoology that generate and manage raw data. So far, datasets available through DIGITAL.CSIC mostly deal with historical research (https://digital.csic.es/handle/10261/28394), climatology and geography (https://digital.csic.es/handle/10261/23139, available in 3 different formats: plain text, raw binary, I.NetCDF), chemistry (https://digital.csic.es/handle/10261/38910), oceanography (https://digital.csic.es/handle/10261/65674), soil (https://digital.csic.es/handle/10261/81504), and agriculture (https://digital.csic.es/handle/10261/83834). We recently uploaded our first archaeology dataset collection comprising 900 micrographies, http://digital.csic.es/handle/10261/85731. This is the first of 3 dataset collections to be made available through the repository as a result of the Proyecto Au

June 2014

which analysed golden artifacts produced in the Iberian peninsula from Prehistory until the end of Ancient Times. In an interview with the director of the project, she explains the need to embark on open data (CSIC Abierto, No. 10, pp. 7-8). The formats of these items vary, but there are both Access and Excel formats available along with others that are more discipline specific such as I.NetCDF. The structure also varies as there are datasets that comprise one single item with many files (https://digital.csic.es/handle/10261/23139) while others are distributed amongst separate items under a single collection (such as https://digital.csic.es/handle/10261/81055 which is under development and gathers descriptions and images of fruit types across Spain; this other collection https://digital.csic.es/handle/10261/52002 gathers soil descriptions throughout southern Spain). Researchers' needs and preferences as regards formats and structure have been largely respected.

When it comes to reusing these contents, most of them carry Creative Commons or Open Database licences, at the recommendation of DIGITAL.CSIC technical office. When selecting open licences, many research groups have chosen licences that accept commercial uses.

There is however a growing number of researchers who have shown interest in uploading their datasets to the institutional repository, while many research groups manage and, in some cases, disseminate their raw data through other devices, most notably PCs, USB sticks, DVDs, other institutional servers, disciplinary data repositories, etc. In some respects, there is still a lack of information on the part of researchers with regard to what DIGITAL.CSIC offers them for their raw data. Although the technical office of the repository has relatively high knowledge of research groups/institutes that create raw data and provide them with open access, there is still much to discover. This is the reason why the technical office sent a survey to the CSIC scientific community at the end of September 2013 in order to learn about their habits as regards raw data and gather their opinions as to what repository services would be more valued in this field.

The survey results highlight that many different types of raw data are being generated within CSIC, including those gathered with sensors and other tools, images and scanned material and X-rays, Excel files, photos, laboratory notes, SPSS, Access and Word files, data from field work, automatically generated data, etc. A great deal of this is experimental data, with observational and simulation data to follow in volume. Respondents have indicated that their raw data ranges between less than 1 GB and 50 GB most of the time. Those generating and managing raw data over 500 GB account for around 12% of responses. As regards reproducibility, most respondents also mentioned that they use open formats.

3 What is the organisational framework?

Roles and responsibilities

As regards DIGITAL.CSIC, its technical office has guided the preparation, description, uploading and licence selection of most datasets available in the repository. In a few cases, researchers have done the whole job by themselves after having received approval from the technical office to go ahead. So far, no datasets have posed any challenges to the platform of the repository, with the exception of the first version of SPEIbase for which we needed to ask CSIC Central Informatics Services to enhance the capacity of the repository's hard disk. These central services are also responsible for daily back-ups of all the items in DIGITAL.CSIC.

As regards researchers who manage and disseminate their raw data without resorting to DIGITAL.CSIC, the results of our survey indicate that in most cases, research groups manage raw data and back-ups by themselves. The majority of groups make their data available to their research groups and other collaborating researchers, meaning that open access is not a common practice to date. The most-mentioned obstacles impeding open access were fear to wasting time by doing it, legal issues, misuse or misunderstanding of data, loss of authorship and fear of losing primacy in discoveries.

Policies

Given the momentum that raw data are gaining in DIGITAL.CSIC, another recent action by its technical office has been the publication of the repository policies as regards their management and open access. Through them, DIGITAL.CSIC seeks to encourage more researchers to make use of the institutional repository for the housing and dissemination of certain types of raw data, but also to raise more awareness within the community about the many options and the potential of sharing data openly. We also prepared a separate document containing a template with the basics for describing datasets in the repository, http://digital.csic.es/bitstream/10261/81323/3/Datasets_DC_plantilla.pdf. Last but not least, we published a divulging article on open access and research data in February 2014, mostly targeting the CSIC community: http://digital.csic.es/handle/10261/92393

The document also includes a set of recommendations for researchers who want to learn more about what to consider when managing their data. The policy is publicly available on the DIGITAL.CSIC website (https://digital.csic.es/politicas/politicaDatos.jsp) and formalises the practice that the repository has been informally promoting to those researchers that had contacted the technical office for advice. The policy intends to give a general view of the main issues at stake (phases in the generation and management of data, formats and versioning, structure, copyright issues and licences, confidentiality, preservation, citation, description, etc.), DIGITAL.CSIC specifications and preferred standards.

The survey results point to a majority of respondents without an explicit or implicit policy for managing raw data. Those with a policy in place include the complexity of data, funder requirements, the structure of the research group (many researchers generating data), and the need for third-party data access/analysing/annotation.

4 What kind of support services are provided to researchers?

Up until now, the DIGITAL.CSIC technical office has been providing advice on a one-to-one basis, responding to questions by single researchers and supporting them in making their data freely and openly available. Most researchers approach the DIGITAL.CSIC Technical Office because of their requirement to upload their datasets in a repository as a condition of publication or because of an open access data mandate. In a growing number of cases, CSIC researchers choose to upload their datasets to the repository because they already make use of the institutional Mediated Archiving Service for their publications, hence the repository is considered a good option to house, describe, disseminate and preserve their datasets effectively, effortlessly and at no cost. Promotion of this type of content in the repository is carried out through informative sessions about the institutional repository in CSIC centres and institutes.

Data-related training has started to be provided to librarian staff over the last year in the form of 3-day workshops and other training activities. Embedded librarians in research projects that include management of raw data are still a minority at CSIC, with most cases concentrated on humanities, social sciences and natural resources/agriculture.

5 What kind of infrastructure is provided?

The DIGITAL.CSIC policy on raw data seeks to set the circumstances under which the repository can house, disseminate and preserve raw data, and the standards to be used. Their policy was published in October 2013, coinciding with the celebrations of the Open Access Week. Our focus for the time being is on documentation, enhanced data use/e-research, data publication and re-use, and preservation.

In parallel, right now we are working on uploading new datasets and the above-mentioned survey will help us identify those CSIC researchers that would appreciate support in this respect. In fact, when surveyed about the services they would like to see in DIGITAL.CSIC to help them with organising, preserving and enabling open access through the institutional repository, training on how to manage raw data, related copyright/licensing issues and the preparation of guides and handbooks were most highly-valued.

6 What have you learned so far? What's next?

Organising and providing open access according to best practices and standards bears huge potential for a scientific organisation such as CSIC where most research groups generate and deal with raw data on a general basis. The data may be as interesting as resulting papers and, in many cases, hidden from the broad scientific community and readers at large.

At present, most raw data are locked in researchers' PCs or are scattered across multiple infrastructures, both institutional and external. In some fields like physics and geoscience, the practice of somehow managing data seems to be more developed than in other areas, but in general the CSIC researcher community thinks that CSIC should preserve and facilitate their access and would therefore value enhanced support from DIGITAL.CSIC.

Further information

Website: https://digital.csic.es/

Policy: https://digital.csic.es/politicas/politicaDatos.jsp

Contact: Isabel Bernal, DIGITAL.CSIC Manager, URICI, CSIC, isabel.bernal@bib.csic.es

References

Bernal, I.; Román-Molina, J. (2014). Prácticas en la gestión, difusión y preservación de datos de investigación en el CSIC. Report. http://digital.csic.es/handle/10261/92404

CSIC Abierto, http://digital.csic.es/revista-csic-abierto/