Towards a Galician data commons

Lourdes Pérez González.

Biblioteca Universitaria, Universidade de Santiago de Compostela , Galicia, Spain

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

SAA ARCHIVES 360°
AUGUST 22-27, 2011
CHICAGO

Introduction

Development in information and communications technologies is transforming the natures and scale of research and education in the way that they have become increasingly data-intensive. The ease with witch digital data can be store, disseminated and accessible to secondary user via science infrastructures means that institutions embrace in the benefits of sharing of research data to increase its impact and visibility.

The research community of Galicia needs to establish a policy that ensures that data and research derived from publicly-funded research should be available and accessible for public use. Treating to provide a framework for developing good practice to all research community from Galicia (universities, research institutions, libraries, researchers and research founders) we prepare some principles reinforced with a number of programs to disseminate it, taking account the need for sensitivity to issues related to access, integrity, transparency, professional responsibility, interoperability, protection of intellectual property and ultimately understand that long term preservation of digital material is the central problem for e-science.

In this work we move on several fronts: our acceptance of the principles of access and sharing of research data established by the Organization for Economic Cooperation and Development (OECD), the commitment to long-term preservation thought the use and implementation of the Standard Reference Model for an Open Archival Information System (OAIS), we insist in a good data management policy according with the data lifecycle, working close to this stage where we define roles and responsibilities of the different stakeholders in the process.

Since our community is at different levels of work at local, national and international it claims cost models that balance the high cost resources digital preservation consume. Based on the different levels of collection established by the National Science Foundation (NSF) we work with the recommendations and best practices of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, and the ongoing models. Keeping Research Data Safe (I,II).

Metodology

Our study try to offer analysis of trends, best practices and key issues that could help us to construct robust infrastructures, content and services in the new research knowledge and information.

Our first steep was taking account of international requirements and developments. There have been significant developments that indicate policy, frameworks and proposal for a national and international data service.

It recognized that different type of data created and managed across research discipline require specific approach to data managing and sharing. In this way we gathered information on research attitudes and practices in the different research community in Galicia. Researchers create and collect different kinds of data; with different purposes and thought different process. We must determine which data should be public available.

At organizational level, we have to break walls between libraries, data centers and researchers, knowing that no single institution can be responsible. We assume that we need to work with key agents and stakeholders at national and international level in order to maximize the value of digital data for current and future researchers in Galicia. A strongest future for research data management and share is one in which multi-institutional collaborations evolvable cyberinfarestructures and services

We work from the lifecycle management of the digital data to ensure its re-use and continuity in a foreseeable future.

Our project adopt two principles: I[st] science system is based on openness and free exchange of data, information and knowledge, ideas derived from publicly funded research should be made available and accessible for the public use: availability, access and usability should de assured in public institution; 2nd digital research data of value arising from the actual data intensive e-science must be preserved and remain accessible for current and future generation.

Strategyc Approach

3 Levels

Principles
and policy

Management

Data lifecycle
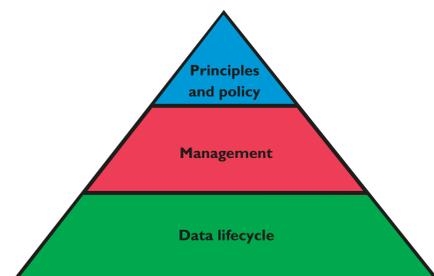
Principles and policy: International, European and National policy (governments and agencies across the world encourage effective dissemination and sharing of research outputs and promote the preservation of them).

Principles: Public-funded research data are public god, produced in public interest; public-funded research data should be openly available to the maximum extent possible.

Policy statement:

Open access to all research outputs, wholly or partially funded by the public found in Galicia;

Research generated by public funded must be well-managed by the different stakeholders during the research process, in this way research data must be accompanied by high-quality metadata enable further re-use and in case long-term preservation. .

Management: Project managers and program managers in multi-institutional collaborative approach.

It should be desirable a mandatory data management and sharing plan in the different discipline according the international standards where it exists.

Any data management and sharing plan should include any cost for its implementation.

All the process requires appropriate funding for data management and different institutions must provide their policy.

Libraries and Research institutes archives as data providers are responsible for providing clear guiadence to researchers and end users of their data management and sharing plans.

Different stakeholders in process of data lifecycle: Data creation, data analysis, research outputs, data curation, services

We must incur in two strategically points: data creation and end of research, in this sense we try to implement a new research culture that ensure that all funded projects develop and implement a data management plan to ensure that data are well managed through the duration of a research project. The goal is to create or develop a research resource for the benefit of the research community.

Steps in data lifecycle

Data creation
- Its a priority to promote and support good data management.
- Data literacy, promoting appropiate skills.
- Start in the scientific process, it sould bacome part of standard research practice.
- The importance of standarsds on data collection.
- Data entry or digitalization: file format, data documentation.

Data Curation
- On the base of the research outputs
- Submission Information Package, SIPs are sent to the archive archive by producers.
- SIP should be negotiated between the research community and the Archive. Specifying criteria like file formats, subject matter, ingest schedule, access restrictions, and verification protocols

Service onset research data:
- Offer online ressource through data catalogues
- Licensing agrraements to acknoklege data rigts.
- Promoting the re-use of data
- Monitoring the secondary usage of data

Data creation

Data management is an essential precondition, it should be an indivisible part of any research project. Comprehensive and standardised documentation start when the research is in the onset of a project, and continued through the research process. Data undergo various stages of transformation in the course of research process; all of them must be included in data documentation.

Data documentation includes: the data management plan of the project, the type of data and instrument, its predictable use, context of collection, data collection methods, data protocol, dataset structure of data files; expected difficulties in data sharing,; explicit mention of consent, confidentiality, anonimysation and other ethical considerations; copyright and intellectual property ownership of the data going to open access and data sharing. A good standard is the Data documentation Initiative (DDI)

Norms and practices must be wrote, research libraries and archives must be responsible for providing guidance to help researchers and users in preparing data management and sharing plan for effective implementation through all the research life-cycle.

Good practice: Data Seal of Approval from Data Archiving and Networked services, DANS.

Data curation

Data long-term preservation and data management are central issues surrounding data access, data security and ethical use. Data management occurs in the lifecycle of the research proccess, data curation and digital preservation it should be carried by data archive and specialsed staff.

The value of data depends on the quality of data. Data creator and data managers in Galicia must pay attention to ensure compliance with quality standards in different international institutions and research institutions.

Data curation starts at the end of research on the base of research outputs whit the Submission Information Package, SIPs are sent to the archive archive by producers and it should be negotiated between the research community and the Archive. Specifying criteria like file formats, subject matter, ingest schedule, access restrictions, verification protocols.

The goal of data curation is the log-term data sharing. The value of data collections to be perserved dependes on the data policy and data management during the reaserch process.

Metadata are the core: Should provides standardised structured information explaining the origin, purpose, creator, access conditions.

Considering re-use of data in long term, we work with the most appropriate and usable software and data format to use. Guarantee long term data access and usability imply use open standard formats. Most of times when data collection is at end of research, the research team should convert data to preferred data preservation format.

Data quality control, stakeholder must assign clear roles and responsibilities for quality assurance at all stage of research.

Value must be added to data through annotation, addition of additional datasets by researchers and by curation aggregation and enhancement.

Research outputs: According to level of collections of NSF, there is different approach on depositing and shared data: centralised and specialist data centres are useful in expertise and resource in data curation, but that centers do not accept all. We promote distributed, local data storage in research collections but in Galicia we must work our expertise to data management at local level.

Service onset research data

Research libraries and archives as data service provider in Galicia should be responsible for ensuring long-term access to data that has been placed in their care; ensuring that these data are usable, reliable and available.
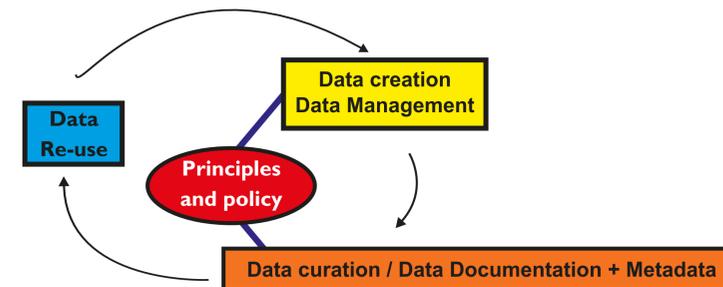
Should be responsible for guidance and advice for data creator on issues related to data management, confidentiality (specially in sensitive and confidential data), security (ensuring the protection of data from unauthorised use, change, disclosure or destruction in conformity with explicit security protocols) , copyright and data sharing.

Data service provider must promote the politics of sharing research data on the base taht they are valuable resource.

Promotting share data should be accompanied by: data catalogues, licenseing aggrements to acknowlge data ownership, monitoring of the secondary usage of data, safe kkepping of research data in a secure environement, management of access... etc. .

E-science data → Long Term data sharing

Data Re-use

Data creation Data Management

Principles and policy

Data curation / Data Documentation + Metadata

Conclusions

For data be shared in the research community and public in general, data must be properly curated and stored. It must be on time, preferably on the early phases of a research project increasing the value, visibility, impact and durability of data and decreasing the cost of storage, curation and preservation.

Archives and research community in Galicia must work closed adopting policies established by another organizations as the Data Seal of Approval by the Data Archiving and Networked Services (DANS) in The Neatherlands, the different levels of collection established by the National Science Foundation (NSF) and th especial recomendations of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, and the ongoing models. Keeping Research Data Safe (I,II).

Acknowledgments: We want to thank the intellectual guidance to the Digital Curation Centre, DCC UK and UK Data Archive.

Bibliography

Blue Ribbon Task Force on sustainable digital preservation and Access/ National Science Foundation; The Andrew W. Mellon Foundation; Library of Congress; Joint Information Systems Committee; Council on Library and Information Resources; National Archives and Records Administration (2008). Sustaining the digital investment : issues and changelles of economically sustainable digital preservation. Retrieved July 22, 2011 from http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

Blue Ribbon Task Force on sustainable digital preservation and Access/ National Science Foundation; The Andrew W. Mellon Foundation; Library of Congress; Joint Information Systems Committee; Council on Library and Information Resources; National Archives and Records Administration (2010). Sustainable economics for a digital planet : Ensuring long-tern Access for digital information. Retrieved July 22, 2011 http://brtf.sdsc.edu/biblio/BRTF_Report.pdf

Research Information Network(2008). To Share or not to share publication and quality Assurance of Research Data Outputs. Report to the Research Information Network / Swan, A.and Brown, S. Retrieved July 22, 2011 from http://eprints.ecs.soton.ac.uk/16742/

Identifying benefits arising from the curation and open sharing of research data produced by UK Higher Education and research institutes. (2008) / Jenny Fry, Suzanne Lockyer, Charles Oppenheim, John Houghton and Bruce Rasmussen. Centre for Strategic Economic Studies, Victoria University, Melbourne; Department of Information Science, Loughborough University. Retrieved July 22, 2011

Joint Information System Committee (2010). Keeping Research Data Safe. Retrieved July 22, 2011 from http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf

Joint Information System Committee (2010). Managing and sharing data : a best practice guide for researchers / Veerle Van den Eynden, Louise Corti, Matthew Woollard and Libby Bishop. Retrieved July 22, 2011

National Science Board, NSB (2005). Cyberinfrastructure Vision for 21st Century. Retrieved July 22, 2011 from http://www.arl.org/brm~doc/ci_vision_march07.pdf

National Science Board, NSB (2005). Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century. Retrieved July 22, 2011 from http://www.nsf.gov/pubs/2005/nsb0540/

Research Information Network(2008). To Share or not to share publication and quality Assurance of Research Data Outputs. Report to the Research Information Network / Swan, A.and Brown, S. Retrieved July 22, 2011 from http://eprints.ecs.soton.ac.uk/16742/