



THE PRESERVATION
AND REUSE OF
SCIENTIFIC DATA IN
SPAIN. REPORT OF THE
GOOD PRACTICES
WORKING GROUP.



GOBIERNO
DE ESPAÑA

MINISTERIO
DE ECONOMÍA
Y COMPETITIVIDAD



FECYT

FUNDACIÓN ESPAÑOLA
PARA LA CIENCIA
Y LA TECNOLOGÍA

Edition, Design and Layout

Spanish Foundation for Science and Technology, FECYT, 2012

Conclusions

Spanish Foundation for Science and Technology, FECYT.

Authors

Working Group on “The depositing and management of data in Open Access” as part of the RECOLECTA project.

Coordination

Cristina González Copeiro (FECYT)
Jordi Serrano-Muñoz (UPC)

Participants

Alicia García-García (UCV)
Antonia Ferrer-Sapena (UPV)
Fernanda Peset (UPV)
Isabel Bernal (CSIC)
Izaskun Lacunza (FECYT)
Javier Gómez (UA)
Luís Martínez-Urbe (Juan March Foundation)
Manuela Palafox (UCM)
Mercedes de Miguel Estévez (FECYT)
Paz Fernández (Juan March Foundation)
Pilar Rico Castro (FECYT)
Ricard de la Vega (CESCA)
Victoria Rasero (UC3M)

Collaborators

Agnes Ponsati (CSIC)
Florencia Dieci (UPV)

Date of publication

December 2012

How to cite this document

Working Group on “The depositing and management of data in Open Access” as part of the RECOLECTA project. *The preservation and reuse of scientific data in Spain. Report of the good practices working group* [online] Madrid: Spanish Foundation for Science and Technology, FECYT (2012) [Date of access 14/01/2013]. Available at WWW.FECYT.ES



This report is under a [Creative Commons License:](http://creativecommons.org/licenses/by-nc-nd/3.0/)
[Attribution-Noncommercial-NoDerivatives 3.0 Unported](http://creativecommons.org/licenses/by-nc-nd/3.0/)

CONTENTS

<i>Introduction</i>	4
1. Research data.....	5
2. Actors involved in scientific data management.....	8
3. What is research data?.....	10
3.1 Definition	10
3.2 Types of data	10
3.3 The management of data.....	11
4. Infrastructure and sustainability.....	13
5. Good practices for research data management	15
5.1 Developing a data management plan	15
5.2 Formats.....	17
5.3 Metadata.....	17
5.4 Digital identifier of data.....	19
5.5 Legal framework related to the management and dissemination of research data	20
5.6 Preservation	23
6. Examples of good practices by disciplines and actors.....	24
6.1 Guides for data management:	24
6.2 Data by disciplines:.....	24
7. Case studies in Spain.....	26
7.1 The evolution of Spanish contributions. Scientific data management	27
7.1.1 Bibliographic review of academic and professional literature	27
7.1.2 Meetings and conferences concerning research data management	30
7.1.3 Projects related to data management and contact with professionals from the sector.....	32
8. Case study: ODiSEA	37
8.1 Background	37
8.2 Aim	37
8.3 Team.....	38
8.4 Methodology.....	38
8.5 The product: “ODiSEA: International Registry on Research Data”	39
8.6 Lessons learned.....	39
9. Good Practices	40
10. Regarding case studies in Spain.....	41
11. Conclusions	43
12. Bibliography	47
Regarding the participating institutions	54

Introduction

This report comes as a response to the challenge facing the open access movement about how to include research data with scientific publications in repositories. It therefore serves to reinforce the application of Law 14/2011, of 1 June, on Science, Technology and Innovation, regarding article 37 of open access dissemination. Its aim is to assist in the standardisation of data management in repositories in order to facilitate its preservation, access and distribution. It reflects on all the important aspects involved in the management of data, from its definition, to types of data, actors involved, good practices for management and a general overview of the situation in Spain.

The Spanish Foundation for Science and Technology (FECYT), in collaboration with the Network of University Libraries (REBIUN) of the Conference of Rectors of Spanish Universities (CRUE), manages and coordinates RECOLECTA, a project for the creation of a network of interoperable institutional repositories which may be considered as the first initiative in Spain towards the creation of an infrastructure to facilitate open science. The aim is to give greater visibility and services to research results and Spanish scientific production.

Within the framework of this project, 2012 saw the launch of a working group set up to study the general situation of the management of scientific data from research and its use in repositories.

We would like to thank the following institutions for their participation in the working group: the Polytechnic University of Catalonia (UPC), the Carlos III University of Madrid (UC3M), the Complutense University of Madrid (UCM), the Spanish High Council for Scientific Research (CSIC), the University of Alicante (UA), the Centre for Scientific and Academic Services of Catalonia (CESCA), the Juan March Institute and the Polytechnic University of Valencia (UPV).

We trust that this study will be of help and interest to those involved in the management of research data.

1. RESEARCH DATA

In recent years, the movement for Open Access to scientific information has generated a debate around new trends in access, use and business models for information produced with public funds. This movement has an important presence in open access to publications in scientific journals. Hence, many funding agencies and institutions currently involved in research have policies guaranteeing open access to publicly-funded scientific publications.

The movement towards open access and the creation of e-infrastructures to support the use of scientific information by the scientific community has also started a debate about the importance of research data. This research data is becoming recognised as a source of knowledge in itself and independently of the publications which may be used in the validation of research results published in articles, to generate new knowledge and be used by humans and machines on an interdisciplinary level.

To ensure that this data can be used it needs to be available and accessible on the internet in the same way as publications. However, the nature of research data is much more variable and depends on the discipline and its particular life cycle. Furthermore, the technical and legal requirements for ensuring access are more complex than those covering publications. There are already disciplines in science with a tradition of depositing and reusing the data available in thematic repositories, but there are also many others which have not adopted this practice in their research routines. The adequate management of data also requires investment, personnel specialising in data management, its usage and subsequent preservation, coordination to ensure the interoperability of infrastructure nodes, change of culture in research personnel, etc.¹.

There is currently an international agreement which contemplates the creation of a transnational and multidisciplinary infrastructure to guarantee access to research data which will contribute to improving the quality of science and multiply its results and avoid duplication²³. Great progress has been made in this area, particularly by the funding agencies, to stimulate an “*open science*” culture to include research data as part of an e-infrastructure to support science in the 21st century.

Without attempting to be exhaustive, and to show international tendencies on this point, this introduction highlights some European documents and communications which are setting trends in the redefinition of

¹A surfboard for riding the wave: Towards a four country action programme on research data; Knowledge Exchange, 2011; <http://www.knowledge-exchange.info/Default.aspx?ID=469> [Date of access 6/12/2012]

²High level expert group on scientific data: Riding the Wave: How Europe can gain from the rising tide of scientific data; European Union, 2010; <http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf> [Date of access 6/12/2012]

³OECD Principles and Guidelines for Access to Research Data from Public Funding, OECD, 2007; <http://www.oecd.org/dataoecd/9/61/38500813.pdf> [Date of access 6/12/2012]

access to scientific information, conceived as an e-infrastructure available to the research community and general public and available in open access when the knowledge comes from publicly-funded projects.

In 2007, the European Commission published a communication on scientific information in the digital age, highlighting the first measures foreseen by the Commission to coordinate the transition from the age of scientific information on paper to the digital environment⁴. These recommendations focused on facilitating access to scientific publications, co-funding research infrastructures (repositories), and stimulating the debate for future policies in relation to the debate between the different players.

This Communication was followed by the Council's conclusions on scientific information⁵ which made rapid access to publications and research data a crucial element in the development of the European Research Area.

As a result of these conclusions, the European Commission launched a pilot Seventh Framework Programme, which stimulated the beneficiaries of seven areas of the programme to deposit their scientific research articles in thematic or institutional repositories, respecting an embargo period of between 6 and 12 months⁶. To support this pilot, the OpenAire project also received funding to provide technological infrastructure and technical support for compliance of the pilot⁷.

Also in 2007, the Organisation for Economic Cooperation and Development (OECD) published a guide for access to publicly-funded scientific research data laying down general recommendations for those responsible for scientific policy and funding agencies of member states to stimulate access to research data⁸.

In 2010, the European Commission asked the "High level group on research data" to prepare a report on its vision about access, use, reuse and quality of scientific research data in 2030⁹. This report has become the European roadmap towards an e-infrastructure to maximise the benefits of access to scientific information.

In response to this report, "*Knowledge Exchange*", an association with members of institutions dedicated to the creation of e-infrastructures for research and higher-education in four European countries has drawn

⁴ Communication on scientific information in the digital age: access, dissemination and preservation (Com 2007 56 Final); http://ec.europa.eu/research/science-society/document_library/pdf_06/communication-022007_en.pdf [Date of access 6/12/2012]

⁵ Council Conclusions on scientific information in the digital age: access, dissemination and preservation, European Union, 2007; http://www.consilium.europa.eu/ueDocs/cms_Data/docs/pressData/en/intm/97236.pdf [Date of access 6/12/2012]

⁶ Commission Decision on the adoption and a modification of special clauses applicable to the model grant agreement of FP7 C(2008) 4408 final http://ec.europa.eu/research/press/2008/pdf/decision_grant_agreement.pdf [Date of access 6/12/2012]

⁷ OpenAire FP7 project <http://www.openaire.eu/> [Date of access 6/12/2012]

⁸ OECD Principles and Guidelines for Access to Research Data from Public Funding, OECD, 2007; <http://www.oecd.org/dataoecd/9/61/38500813.pdf> [Date of access 6/12/2012]

⁹ High level expert group on scientific data: Riding the Wave: How Europe can gain from the rising tide of scientific data; European Union, 2010; <http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf> [Date of access 6/12/2012]

up a proposal for the creation of an action plan for the United Kingdom, Denmark, the Netherlands and Germany concerning research data¹⁰.

The European Commission was preparing new recommendations for open access and preservation of scientific information for the end of 2012, and these are expected to go further with the stimulation of open scientific content (for publications and data), open and interoperable infrastructures and “*open culture*” (for researchers and the general public).

In Spain, the recently approved “Science, Technology and Innovation Law”¹¹ added stimulus to the creation of infrastructures to support scientific information, with a section of the Law dedicated specifically to the deposit of scientific articles funded by the General State Budgets in institutional or thematic repositories.

This report stems from the Recolecta project and emphasises certain important considerations which need to be taken into account in the design and implementation of a research data-management policy, with particular emphasis on the situation in Spain compared to other countries. In this report, we define the variety of types of research data together with those involved in its management (institutional and thematic repositories, funding agencies, existing data centres, researchers, librarians and data-management experts, etc.). We also reflect on financial issues stemming from the creation of an interoperable, data-management infrastructure. Finally, this report aims to contribute to future initiatives which will be required for the management of research data under the scope of the new Science, Technology and Innovation Law.

¹⁰ A surfboard for riding the wave: Towards a four country action programme on research data; Knowledge Exchange, 2011; <http://www.knowledge-exchange.info/Default.aspx?ID=469> [Date of access 6/12/2012]

¹¹ Ley 14/2011 de la Ciencia, la Tecnología y la Innovación
<http://www.boe.es/boe/dias/2011/06/02/pdfs/BOE-A-2011-9617.pdf> [Date of access 6/12/2012]

2. ACTORS INVOLVED IN SCIENTIFIC DATA MANAGEMENT

E-science has changed research practices in all areas of science. The increase in computational capacity allows researchers to process and share large quantities of information. To facilitate the reuse of scientific data we need to adopt the standards used by the research-data community, develop and promote guides of good practices to help researchers manage their research data adequately, stimulate training programmes to provide the scientific community with the necessary skills, protect the intellectual property of data producers and establish the mechanisms required to ensure quality. To do this, it is vital for there to be a high level of coordination between those involved in data management.

In this section, we describe the role played by those involved in the management of scientific data and their associated responsibilities¹².

- **Researchers/data producers**

They provide the evidence and scientific validation of research. Although this category consists mainly of researchers, in some cases there are sets of data which already exist and scientists use them to validate their hypotheses. The research community can be thought of as producers, authors and users of research data.

- **Universities and Research Centres**

Their main responsibility is to lay down internal policy for scientific data management. They establish the standards for the different types of data and the guide of good practices. Institutions must assume promotional responsibility to ensure that the research results of their researchers are deposited in the institutional repositories for their short-term custody and preservation, providing suitable training for this. Inside universities and research centres, some of the most important data management services are provided by the IT services, libraries and research services. Each has complementary roles (IT services in data storage; libraries in metadata, support for publication and rights; and research services in institutional policies, management plans and ethical issues) and they need to coordinate between them to provide a complete institutional service.

- **Institutional repositories**

They play a basic role in short-term data storage, as opposed to the role of the long-term data preservation centres. The use of standards is fundamental to facilitate interoperability between repositories and data centres. The reliability and robustness of the links and the establishment of mechanisms for the migration of data between repositories is highly important, as is the maintenance of the versions of data if they are in different storage areas. The sustainability of the archiving of research data represents one of the key challenges and problems.

¹² Lyon, Liz (2007) Dealing with Data: Roles, Rights, Responsibilities and Relationships. Consultancy Report. UKOLN http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.doc [Date of access 6/12/2012]

- **Data centres**

They establish guides of good practices and the selection of data which must be preserved in the long-term, facilitating their dissemination. They protect the property rights of those who have produced the data and provide tools for its reuse. They develop data recovery plans in the event of disasters.

- **Data managers**

The professional profile of the data manager requires IT skills, knowledge of the discipline, of research practices and workflows, understanding of specific technical standards, metadata schema and common vocabularies.

They must also know which are the national and international research data centres for that discipline and have a broad understanding of the data publication requirements for the leading academic journals¹³. It is the data managers' responsibility to manage and promote the use of data once it is created to ensure its use and availability to be located and reused¹⁴.

- **Users who reuse data**

They must comply with the license conditions and permissions of use, recognising the intellectual property rights of the researchers who have produced the data.

- **Funding agencies**

The funding agencies implement data policies with the actors involved, they set preservation dates, resolve problems of confidentiality, data protection and use of licenses. Since 2000, the funding agencies in some countries (*National Institutes of Health, Wellcome Trust*, etc.) have started to ask for the liberation of data in different degrees and with different levels of compliance, in order to maximise the return on research funding. Since 2010, the *National Science Foundation* requires funding proposals to be accompanied by a Data Management Plan¹⁵.

- **Scientific publications**

In the same way as the funding agencies, the editors of scientific publications have started linking journal articles with their research data, in order to share this data with readers and researchers.

¹³ Lyon, Liz (2012) The Informatics Transform: Re-Engineering Libraries for the Data Decade. The International Journal of Digital Curation. Volume 7, Issue 1, 2012

<http://www.ijdc.net/index.php/ijdc/article/view/210/279> [Date of access 6/12/2012]

¹⁴ Martínez-Urbe, Luis, Macdonald, Stuart (2008). Un nuevo cometido para los bibliotecarios académicos: data curation. El profesional de la información, v.17, n. 3, mayo-junio 2008

¹⁵ Borgman, C.L. (2011). The conundrum of sharing research data. Journal of the American Society for Information Science and Technology
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=186915 [Date of access 6/12/2012]

3. WHAT IS RESEARCH DATA?

3.1 Definition

Defining research data is not a simple task: the data produced by researchers forms an extremely heterogeneous and complex group of materials, created for different purposes and via different processes. Data is the “soul” of research, and it rarely consists of simple objects which can be shared easily, instead it embodies the epistemological perspectives of its creators¹⁶.

Melbourne University in Australia gives the following definition in its institutional data policy:

Data are facts, observations or experiences on which an argument, theory or test is based. Data may be numerical, descriptive or visual. Data may be raw or analysed, experimental or observational. Data includes: laboratory notebooks; field notebooks; primary research data (including research data in hardcopy or in computer readable form); questionnaires; audiotapes; videotapes; models; photographs; films; test responses. Research collections may include slides; artefacts; specimens; samples. Provenance information about the data might also be included: the how, when, where it was collected and with what (for example, instrument). The software code used to generate, annotate or analyse the data may also be included.

3.2 Types of data

The National Science Foundation (2007) proposes the following categorization of research data based on its origin which gives a clearer picture of the variety of types and their different management requirements:

- **Observational data.** These are historical records, which can only be obtained in one place and at one moment in time. This characteristic is particularly important when it comes to preserving the data because, in the event of it being lost, it could not be reproduced. Examples: the opinion polls of the Centre for Sociological Research (CIS is its Spanish acronym) are surveys on different subjects which concern people in Spain The National Bank of Climatological Data would be another example of this type as it has information on the precipitation recorded in Spain over the last 150 years.
- **Experimental data.** This is the data which accompanies experiments from the planning and preparation stage until results are obtained. In many cases experiments can be repeated to obtain the same data; however, sometimes the cost of repeating the experiment makes this unfeasible. Examples: the CERN particle accelerator in Geneva produces a vast quantity of experimental data capable of

¹⁶ Borgman, CL (2012) On Local or Global? Making Sense of the Data Sharing Imperative. Talk at University of Southern Carolina on 9th April 2012

filling 100,000 DVDs each year. In research laboratories, whether these are chemical, biological or from other disciplines, a great quantity of data is produced with specialised instruments.

- **Computational data.** This is data which accompanies simulations which tend to include input data, certain programmes and results. For this type of data, in most cases the results are not needed because with the input data, programmes and the computer which generates the information, it should be possible to reproduce them. Examples: This can be data produced by advanced computation centres which simulate the working of organs in the human body, the movement of stars or which predict the weather.

This way each scientific discipline will base its research on these typologies and on those in which it can be subdivided. Whether they are qualitative, quantitative, geographical, spatial, or any other, they will belong to one or several of the mentioned axes.

3.3 The management of data

The correct management of research data is a fundamental part of the research process. This management consists of decision-making and actions from before the creation of the data, during their creation and use and throughout their life-cycle. Some of the stages which should be included in correct data-management are:

- A data-management plan as part of the funding proposal to anticipate management challenges and propose solutions to them.
- Deal with the appropriate ethical and legal questions referring to sensitive personal data, copyright and access licenses and data usage.
- The organization and documentation of data in compliance with disciplinary and international standards which make it possible to know what the data is and how it was created so that it may be reused.
- Appropriate storage mechanisms, back-up and information security to ensure confidentiality, integrity and availability of information.
- Share data in a way that means it is cited in a standard manner, thereby giving credit to its creators.
- Filing of a final copy of the data in specialised data centres which take the necessary measures for the preservation and dissemination of the data.

To make it possible for data to be managed in this way, policies are needed at a funding agency and institutional level to define and clarify the roles and responsibilities of the different actors. The responsibility for this management throughout the life-cycle should lie with a variety of institutions such as

funding agencies, universities, libraries, IT centres and the researchers themselves. However, the starting point has to be the researchers themselves and their needs.

The Ligue des Bibliothèques Européennes de Recherche - Association of European Research Libraries (LIBER) created in 2010 a Working Group on e-Science, which issued a final report¹⁷ with ten recommendations for libraries starting out in the management of research data. In its conclusion, it emphasised that libraries can and should assist researchers in the management and planning of data.

¹⁷ Christensen-Dalsgaard, Birte et al (2012) Ten recommendations for libraries to get started with research data management: Final report of the LIBER working group on E-Science / Research Data Management.
http://www.libereurope.eu/sites/default/files/WGSC_20120801.pdf [Date of access 9/12/2012]

4. INFRASTRUCTURE AND SUSTAINABILITY

Data must be managed by a reliable and stable infrastructure which guarantees its trustworthiness and integrity. The white paper “Strategy for a European Data Infrastructure”¹⁸ includes the main infrastructure requirements for various discipline-oriented initiatives and research communities at a European level. In short, they are:

- Long-term data preservation including authenticity and other checks that guarantee data quality.
- Data access (data life-cycle), data curation services and infrastructure computational capacity (data mining, data processing...).
- Distribution of data and federations, not just for preservation reasons but also for the optimization and increase in access performance.

In addition to these requirements, data must be duplicated to achieve high availability, which is a common requirement for this type of system.

Three aspects must be taken into account to provide a solution to these requirements:

- Software systems capable of managing data life-cycles.
- Mass data storage systems. Several technologies are available for this purpose, such as NAS (Network Attached Storage) architecture for horizontal growth, which allows rapid scaling via commodity nodes depending on demand. In relation to the data life-cycle, there may be many factors depending on their nature or discipline; however, bit flows which are stored in a physical medium can be treated homogeneously.
- High capacity networks for the transmission of data between different nodes. In Spain, the academic and research network (RedIRIS) provides these advanced communications services to the national scientific and university community.

These infrastructures have to be taken into account when analysing the viability of data management initiatives, because their costs, of both purchase and maintenance, are high. It is estimated that the maintenance costs of scientific data repositories are on a higher order of magnitude than the traditional repositories for publications¹⁹.

¹⁸ Strategy for a European Data Infrastructure
<http://www.csc.fi/english/pages/parade> [Date of access 6/12/2012]

¹⁹ Beagrie N, Chruszcz J and Lavoie B (2008). Keeping Research Data Safe 1. JISC
<http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf> [Date of access 12/12/12]

There are two basic principles for a better capitalization of these costs:

- Data selection processes. Not all data has to be or can be curated or preserved. A good selection integrated within the life-cycle of the data and performed from the point of view of the specific knowledge of the data and thinking not just about its main use, but also about how this data may be reused at a later stage is essential.
- Use of economies of scale in relation to infrastructures. What is required is a data layer which can group infrastructures transversally, as is done by Geant, RedIRIS or the Anella Científica in the connectivity layer, or which like the Driver projects make different research repositories interoperable. Not only could costs be shared, but also greater synergies could be achieved between different research groups and even between different disciplines.

We mentioned earlier that data can be very heterogeneous, and depending on this the costs associated to the infrastructure may vary substantially. On the extreme of high infrastructure costs we would find projects with mass datasets such as those for the data produced by the Large Hadron Collider or the European Bioinformatics Institute, while on the other extreme, for example, we could find the Worldwide Protein Data Bank Archive, a repository with more than 80,000 molecular structures in 3D, but which barely requires 150GB for storage. In this last case, the infrastructure costs are insignificant compared to the 69 FTE staff working on the project²⁰.

Even if we just manage the data which is “useful” or vital, and do so in infrastructures which take advantage of economies of scale and whatever the size of the infrastructure required, long-term infrastructure funding policies are required for the management of scientific data, because data is accumulative and it is typically preserved beyond technological cycles.

As we mentioned earlier, details of a data-management plan, including its financial viability, should be given in project funding proposals.

²⁰ The Royal Society (2012). Science as an open enterprise <http://royalsociety.org/policy/projects/science-public-enterprise/report>
[Date of access 12/12/12]

5. GOOD PRACTICES FOR RESEARCH DATA MANAGEMENT

Research data constitutes one of the main assets in the scientific research process. Optimum management of this data encourages innovation and the development of innovation, because it enables the exploitation of high quality data (share – reuse).

In the global framework of e-Science, the specific object of the control, organisation, description and preservation of scientific data is the ‘dataset’, which is defined as a collection of data brought together during the execution of a research project. Datasets are compound and heterogeneous digital objects. In other words, they may consist of different elements or types of data: text documents, spreadsheets, files with mathematical operations, graphs, images, etc. The dataset constitutes the basis of a research project and it is associated to a scientific publication as a result of this research. The dataset acquires added value if it is integrated with the related publication (‘linking data’: citation and link), regardless of its location.

Datasets are stored and managed in interoperable repositories in integrated networks in a global research structure, developed in compliance with international standards.

Higher education institutions and research funding agencies from several countries are involved in initiatives to create data management infrastructures to enable the reuse of datasets, through the adoption of policies promoting open access and data sharing, and guaranteeing the sustainability and accessibility of data in the long-term.

The Open Data movement, within the Open Access framework, defines open data as that which can be used, reused and redistributed without any restriction other than the requirement to attribute and/or share-alike²¹.

5.1 Developing a data management plan

The responsibility for data management lies firstly with researchers, but institutions ought to provide their research community with technical and organisational assistance. At an organisational level, in a research data management service, it is vital for there to be collaboration between researchers, data producers and data librarians within the institution.

The researchers are the experts who should provide the contextual information required to determine the origin and life-cycle of the data. The librarians are the experts in the management of information and they

²¹<http://opendefinition.org/okd/> [Date of access 6/12/2012]

have to provide specialised and customised support to the researchers, and use the technical resources required for the data to be understood and interpreted by other researchers.

Given the diversity of scientific data, which is by nature heterogeneous and affected by the specific culture of each scientific community, the institution needs to provide researchers with a data-management plan to save time and effort in the research process²². Planning brings with it a series of advantages:

- Data can be found and understood when it needs to be used.
- Project continuity is guaranteed regardless of the participation of the researchers.
- Duplication and unnecessary tasks are avoided.
- The upkeep of the generated dataset allows results to be validated.
- Data can be shared allowing a high level of collaboration and progress in research.
- If the data is open it will have high visibility.
- Other researchers using the data may cite it and the research will gain prestige.

The minimum description of data should cover the following aspects:

- Context, description of the project and purpose of the research, methodology used;
- Nature of the data, data history, content and structure, terminology, software, date of creation and modification, versions, responsible personnel and participants;
- File formats, structure and file nomenclature, storage system, procedure for back-up copies;
- Legal aspects, access and security policies.

The technological paradigm of a scientific data management system includes the following requirements:

- The logical data model (relational) and its management system (database) have to allow its description, representation and recovery;
- The management system must enable optimum organisation of data, documenting it, preserving it and making it accessible;
- A software capable of analysing a large quantity of data, processing, treating and obtaining different secondary products ('Data Mining').

²² There are tools for the elaboration of plans of this type, such as, for example DMPTool (<https://dmp.cdlib.org/>) [Date of access 6/12/2012]

5.2 Formats

The format in which the data is archived is an essential factor for ensuring its preservation and accessibility. The evolution of technologies explains why both hardware and software become obsolete. Researchers use the format and software which best suits their needs; but to guarantee access and preservation in the long-term, the following considerations need to be taken into account:

- Whenever possible, open, non-proprietary formats should be used.
- The format used has to allow the indexing of content to make it recoverable.
- A data compression format uses less storage space.
- The format chosen should be standard (IANA mime types), or de facto standard for the research community.

Files and folders must be well-organised in an ordered structure. The nomenclature system is important for identifying contents.

File version control is required to make it possible to locate successive versions to identify changes from one to the next.

5.3 Metadata

Metadata is a set of structured information which has to include the origin, purpose, time reference, geographical location, creator, access conditions and terms of use of a dataset. Metadata fulfils different interrelated functions: the management and administration, preservation, description, dissemination and recovery of data. The documentation and description of data facilitates its location, understanding and use.

Dataset documentation provided by the researcher will be included in the metadata register. Metadata should include at least the following information:

- *Title*: Name of the project of the dataset or research which produced it.
- *Names of the creators* and the *addresses* of the organisation or people who have created the data.
- *Data identification code*, even if it is a reference for internal use.
- Words or phrases which describe the *subject or content of the data*.
- *Sponsors*: The organisations or agencies which funded the research.
- *Rights*: Any type of intellectual property rights of the data.
- *Access to the information*: Where and how is data accessible for other researchers?
- *Language of the content*.
- *Key dates associated to the data*, including: start of the project and completion date, launch date, period of time covered by the data, and other data related to the useful life of the data, for example, the maintenance cycle, updating of the programme.

- *Place which the data refers to* (e.g. physical location, spatial coverage, etc).
- *Methodology*: How was the data generated, including the equipment or software used, experimental protocol, etc?
- *Data processing*: all the information about how the data has been altered or processed.
- *Sources*: Citations to materials for the data from other sources, including details of source data.
- *List of file names from the list of all the data files associated to the project*, with their names and file extensions (e.g. 'stone.mov').
- *Data file formats*, for example, FITS, SPSS, HTML, JPEG, RIF-CS and the software required to read the data.
- *File organisation*: structure of the data file(s) and the disposition of the variables, when applicable.
- *List of variables* in the data files.
- *Explanation of the codes or abbreviations used* in any of the names of files or variables in the data files.
- *Versions of date / date and time for each file*, and use a different ID for each version (see organisation of its files).
- Verification operations to check whether files have changed over time. (Checksum algorithm to protect data integrity).

Metadata is structured in registers in accordance with standardised schema. The criteria for the adoption of one schema or another will depend on the data-management objectives established by the organisation. Standardisation is a priority if interoperability is to be achieved with other data-management systems. To comply with all the previously mentioned functions, different metadata schemas are normally combined through the declaration of the space for the names corresponding to each schema.

There are several metadata standards, although here we shall cite those used most widely:

- ***Dublin Core Metadata Terms***²³. This is a very simple universal schema, which can be applied to resources of any type or origin.
- ***Data Documentation Initiative (DDI)***²⁴. This is a schema designed specifically for the description of social and economic datasets. It allows the full life-cycle of the data to be documented.
- ***General International Standard Archival Description (ISAD(G))***²⁵. This is a set of elements to describe files with several aggregation levels. The descriptive processes can be simultaneous to the production of the documents and continue throughout the whole of its life-cycle.

²³<http://dublincore.org/documents/dcmi-terms/> [Date of access 6/12/2012]

²⁴<http://www.ddialliance.org/what> [Date of access 6/12/2012]

²⁵[http://www.icacds.org.uk/eng/ISAD\(G\)es.pdf](http://www.icacds.org.uk/eng/ISAD(G)es.pdf) [Date of access 6/12/2012]

- *Metadata Encoding and Transmission Standard (METS)*²⁶. This is a standard for the coding and grouping of metadata which are administrative, technical, descriptive and for preservation, offering a highly exhaustive representation of complex digital objects. It also allows the relations between the parts of a digital object to be expressed, together with the relations between different objects.
- *ISO 19115 for geographic information*²⁷. Schema used for the description of geographical information and services. It is applicable to geographical datasets.

The metadata registers are grouped in information search and recovery systems, and can be collected using the OAI-PMH protocol.

5.4 Digital identifier of data

The stored dataset must be associated to a unique and persistent digital identifier which facilitates data verification, reuse, dissemination, impact and access in the long-term. In conformity with the semantic web, identifiers must have a URI (Uniform Resource Identifier) form. The URI is a chain of characters which condenses the URL (Uniform Resource Location) address and the URN (Uniform Resource Name) of the resource²⁸.

There are many different systems, such as for example:

- *PURL Uniform Resource Locator*. Functionally, a PURL is a URL. However, instead of pointing straight at the location of an internet resource, some PURL points point to an intermediate resolution service. The PURL resolution service associates the PURL with the real URL address and returns the URL to the client.
- *DOI Digital Object Identifier*. This is a name for an entity in digital networks. It provides a permanent and viable identification system and enables interoperable exchange of the information handled in digital networks.
- *ACCESSION*²⁹– Numbers used by the National Center for Biotechnology Information (NCBI) are unique and citable.
- *InChI*³⁰ The IUPAC International Chemical Identifier (InChI™) is a non-proprietary identifier of chemical substances which can be used in the sources of printed and electronic data, thereby enabling an easier link of the compilations of different data.

²⁶<http://www.loc.gov/standards/mets> [Date of access 6/12/2012]

²⁷http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020 [Date of access 6/12/2012]

²⁸<http://www.w3.org/TR/uri-clarification/> [Date of access 6/12/2012]

²⁹<http://www.ncbi.nlm.nih.gov/> [Date of access 6/12/2012]

³⁰<http://www.iupac.org/home/publications/e-resources/inchi.html> [Date of access 6/12/2012]

5.5 Legal framework related to the management and dissemination of research data

The production, management and dissemination of data should fit into a legal framework with rights and agreements which must be respected. The key questions here are:

- What legal rights exist for data and datasets?
- Who do these rights belong to?
- What legal restrictions have to be applied for the dissemination of data and datasets?
- What contracts, permissions and licenses must be used to comply with current legislation?

The following rights have to be taken into consideration:

- Intellectual property rights
- The confidentiality, privacy and protection of data

Access and data: Taking into account legal restrictions, it is necessary to identify which data is accessible, identify who can access it and for what purpose. Depending on the nature of the data we must address the following categories:

- Public data: this may be made available without restrictions for any user in open access.
- Restricted data: may only be consulted by certain users.
- Private data: may not be made public. It is confidential.

Privacy and confidentiality: Any research containing personal data has to comply with the precepts of data protection legislation. In Spain, these aspects are regulated by *Organic Law 15/1999, of 13 December, on the Protection of Personal Data*, the object of which is to “guarantee and protect in that concerning the treatment of personal data, the public liberties and the fundamental rights of individuals, in particular personal and family honour and privacy”. The law applies to personal data registered in any physical medium. The treatment of data covers the activities of collection, registration, storage, recovery, consultation, use and dissemination. To guarantee the right to data protection, the persons affected must be informed and their consent must be requested for the treatment of their data.

Intellectual property and data: In Spain the main legislation regulating intellectual property rights is the Law on Intellectual Property (*Royal Legislative Decree 1/1996 of 12 April which approved the revised text of the Law on Intellectual Property*) which has undergone several modifications, including Law 23/2006 of 7 July which adapted Spanish legislation to the new circumstances created by the information society.

Data collections and databases are protected as intellectual property, according to Art. 12 of the above-mentioned law on intellectual property through the so-called *sui generis* right, in that they constitute

intellectual creations. “Protection refers solely to their structure as a form of expression of the selection or disposition of contents”, not to the data itself. Royalties belong to the authors, as long as they are original works.

Moral rights are personal rights which belong exclusively to authors and they are inalienable. By virtue of these rights, it is fundamentally down to the authors to decide whether their work has to be disseminated and in what way, and demand recognition of authorship.

Exploitation rights or copyright are transferable. Title holders of these rights possess the sole rights to them and they may not be exercised without their authorisation, apart from the limits established by law. Exploitation rights constitute a series of acts such as reproduction, distribution, public communication and transformation.

There are exceptions to the exercise of exploitation acts, such as reproduction for exclusively private use, uses for the benefit of disabled people, use for citation or illustration for educational purposes.





Works in the public domain, when the protection period for rights has expired, may be used freely and free of charge³¹.

Data deposit: The deposit of datasets in a repository involves the exercise of exploitation rights, so that the explicit permission is required from the title holder of these rights through a non-exclusive cession agreement of the necessary rights.

In accordance with the open access movement, data produced by publicly-funded projects represents a good of public interest, so that it must be available in a repository in open access without prejudice of legal or ethical precepts.

Alternative licenses to the copyright: As we mentioned beforehand, the title holder of the exploitation rights has the power to determine who may access data and under what conditions. There are standard and free licenses which authors may apply to their research data to provide the terms for sharing and reusing this data on the internet. One example of such licenses is Creative Commons, which consists of six licenses which allow the copying, distribution, downloading and transformation of the digital documents:

³¹ The law on intellectual property establishes a period of 70 years after dissemination in which a work is covered by rights, and 70 years from its creation if it has not been disseminated.

	ATTRIBUTION: In any exploitation of the work authorised by the license, authorship must be recognised.
	NON COMMERCIAL: The exploitation of the work is limited to non-commercial uses.
	NO DERIVATIVE WORKS: Authorisation to exploit the work does not include the transformation to create a derivative work.
	SHARE ALIKE: Authorised exploitation includes the creation of derivative works as long as they maintain the same license when disseminated.

Through a combination of these four precepts six licenses are obtained:

- *Attribution alone (CC BY)*
- *Attribution + Share alike (CC BY-SA)*
- *Attribution + Without derivative work (CC BY-ND)*
- *Attribution + Non-commercial (CC BY-NC)*
- *Attribution + Non-commercial + share in identical conditions (CC BY-NC-SA)*
- *Attribution + Non-commercial + no derivative work (CC BY-NC-ND)*

The licenses of CC version 4.0 tackle the specific characteristics of data.

Science Commons is an initiative within Creative Commons which, amongst other things, aims to pull down barriers and develop tools to encourage the reuse of data from research projects. On these lines, Science Commons Open Access Data Protocol³² includes a methodology and good practices for the creation of tools to enable integration between scientific databases and put them in the public domain.

Following the Creative Commons model, the Open Knowledge Foundation has created some specific licenses for collections of data: *"The Open Data Commons License"*³³. It is important to distinguish between the license for data included in the database and the license regime for the actual database. Of the Open Data Commons' licenses, the Database Contents License stands out: it refers to the contents of the database, and the most radical of all is the Public Domain Database License, in which the title holders of the rights renounce these for the benefit of all.

³²<http://sciencecommons.org/projects/publishing/open-access-data-protocol/> [Date of access 12/12/2012]

³³<http://opendatacommons.org/licenses/> [Date of access 12/12/2012]

5.6 Preservation

Data must be preserved and remain accessible and usable for future research. Data management must include a preservation plan in conformity with international standards.

The questions to be asked are: Which data has to be saved? How should it be saved?

Make regular back-up copies which can be used to restore original files. File integrity needs to be verified by checking the MD5 code checksum value, the size of the file and the date.

The data storage strategy must contemplate the obsolescence of hardware and software. We recommend that data be copied in different types of physical medium, for example one digital copy and another on a hard drive. Preservation factors, such as changes in temperature, relative humidity, light, etc. should be taken into account.

6. EXAMPLES OF GOOD PRACTICES BY DISCIPLINES AND ACTORS

6.1 Guides for data management:

- Australian National Data Service: [HTTP://ANDS.ORG.AU/RESEARCHERS/MANAGE-DATA.HTML](http://ANDS.ORG.AU/RESEARCHERS/MANAGE-DATA.HTML) [Date of access 8/12/2012]
- Australian National University. Data Management: Information from courses and a manual on data management: [HTTP://ILP.ANU.EDU.AU/DM/](http://ILP.ANU.EDU.AU/DM/) [Date of access 8/12/2012]
- CIESIN: Geospatial Electronic Records- Resources on managing and preserving geospatial data and related electronic records: [HTTP://WWW.CIESIN.COLUMBIA.EDU/GER](http://WWW.CIESIN.COLUMBIA.EDU/GER) [Date of access 8/12/2012]
- Data Management for Researchers: [HTTP://ANDS.ORG.AU/RESEARCHERS/MANAGE-DATA.HTML](http://ANDS.ORG.AU/RESEARCHERS/MANAGE-DATA.HTML) [Date of access 8/12/2012]
- Data management in the Humanities: [HTTP://ERCIM-NEWS.ERCIM.EU/EN89/SPECIAL/DATA-MANAGEMENT-IN-THE-HUMANITIES](http://ERCIM-NEWS.ERCIM.EU/EN89/SPECIAL/DATA-MANAGEMENT-IN-THE-HUMANITIES) [Date of access 8/12/2012]
- ICPSR Guide to Social Science Data Preparation and Archiving: Outlines best practices throughout the research process, including applying for a research grant, collecting data, and preparing data for deposit in a public archive. [HTTP://WWW.ICPSR.UMICH.EDU/FILES/ICPSR/ACCESS/DATAPREP.PDF](http://WWW.ICPSR.UMICH.EDU/FILES/ICPSR/ACCESS/DATAPREP.PDF) [Date of access 8/12/2012]
- Oak Ridge National Laboratory. Best Practices for Preparing Environmental Data Sets to Share and Archive. Describes the practices to make data sets ready to share with others: [HTTP://DAAC.ORNL.GOV/PI/BESTPRACTICES-2010.PDF](http://DAAC.ORNL.GOV/PI/BESTPRACTICES-2010.PDF) [Date of access 8/12/2012]
- UK Data Archive: Create & Manage Data: Provides best practice strategies and methods for creating, preparing and storing shareable datasets. [HTTP://WWW.DATA-ARCHIVE.AC.UK/CREATE-MANAGE](http://WWW.DATA-ARCHIVE.AC.UK/CREATE-MANAGE) [Date of access 8/12/2012]
- UK Data Archive: Managing and Sharing Data: a Best Practice Guide for Researchers 3rd. ed. [HTTP://WWW.DATA-ARCHIVE.AC.UK/MEDIA/2894/MANAGINGSHARING.PDF](http://WWW.DATA-ARCHIVE.AC.UK/MEDIA/2894/MANAGINGSHARING.PDF) [Date of access 8/12/2012]

6.2 Data by disciplines:

- Annotation and description of biomedical databases (Harvard University): [HTTP://ESCHOLARSHIP.UMASSMED.EDU/CGI/VIEWCONTENT.CGI?ARTICLE=1000&CONTEXT=JESLIB](http://ESCHOLARSHIP.UMASSMED.EDU/CGI/VIEWCONTENT.CGI?ARTICLE=1000&CONTEXT=JESLIB) [Date of access 8/12/2012]
- Archaeology: [HTTP://ARCHAEOLOGYDATASERVICE.AC.UK/](http://ARCHAEOLOGYDATASERVICE.AC.UK/) [Date of access 8/12/2012]
- Astronomy: [HTTP://ADSWWW.HARVARD.EDU/](http://ADSWWW.HARVARD.EDU/) [Date of access 8/12/2012]
- Bioinformatics: [HTTP://WWW.EBI.AC.UK/INFORMATION/DATABASES_SITEMAP.HTML](http://WWW.EBI.AC.UK/INFORMATION/DATABASES_SITEMAP.HTML) [Date of access 8/12/2012]
- Marine sciences: [HTTP://WWW.MARINE-GEO.ORG/CONTRIBUTE.PHP](http://WWW.MARINE-GEO.ORG/CONTRIBUTE.PHP) [Date of access 8/12/2012]

- Chemical sciences: [HTTP://WWW.CHEMSPIDER.COM/](http://www.chemspider.com/) [Date of access 8/12/2012]
- Geospatial data: [HTTP://EDINA.AC.UK/PROJECTS/SHAREGEO/](http://edina.ac.uk/projects/sharegeo/) [Date of access 8/12/2012]
- Energy: [HTTP://EN.OPENEL.ORG/WIKI/MAIN_PAGE](http://en.openel.org/wiki/Main_Page) [Date of access 8/12/2012]
- ISATOOLS Biomedical data: [HTTP://ISATAB.SOURCEFORGE.NET/](http://isatab.sourceforge.net/) [Date of access 8/12/2012]
- Linguistics: [HTTP://WWW.LANGUAGE-ARCHIVES.ORG](http://www.language-archives.org) [Date of access 8/12/2012]
- List of data repositories: [HTTP://DATACITE.ORG/REPOLIST](http://datacite.org/repolist) [Date of access 8/12/2012]
- Medicine: [HTTP://WWW.NCBI.NLM.NIH.GOV/GENBANK/](http://www.ncbi.nlm.nih.gov/genbank/) [Date of access 8/12/2012]
- Music Brainz: [HTTP://MUSICBRAINZ.ORG/](http://musicbrainz.org/) [Date of access 8/12/2012]

7. CASE STUDIES IN SPAIN

This section of the report reflects on all the initiatives we are aware of concerning scientific data management in which Spanish agents are involved. We adopted the following methodology to compile this section: Bibliographical review, monitoring of meetings and conferences, contacts with known groups, review of projects and, finally, identification of Spanish repositories and datasets in data banks registered in the international ODiSEA inventory. We also take a detailed look at this project as a case study in Spain.

First of all we detected **authors and subjects of interest** through academic and professional literature, details of which are given in the following section. The first author to publish a paper on scientific data management in the area of information was Martínez Uribe in 2008 who had been working in Great Britain. The following years began to witness publications and communications on scientific data management and sharing in public forums: the group from Granada, Torres Salinas, Robinson-García and Cabezas-Clavijo, (*Anuario ThinkEPI and El profesional de la información*), Pérez González from Galicia (*Jornada SEDIC and 75th Annual Meeting of the Society of American Archivists*) and in *Blok de BiD: reseñas de biblioteconomía y documentación* Melero and Peset from Valencia and Keefer and Borrego from Barcelona. We also found individual contributions in other fields, such as psychology (Botella Ausina and Ortego Maté) or earth sciences (Bermúdez, Barragán and Alonso).

Secondly, three initial **meetings** appeared whose contributions are outlined in the following section. The first was promoted by FECYT-RECOLECTA and sponsored by the UNED in November 2011 (*Almacenamiento, la conservación y la gestión de los datos de investigación* [Storage, preservation and management of scientific data]), and was oriented towards the role of repositories in relation to data. Active Spanish agents only included FECYT and the data library of the Centre for Advanced Studies in the Social Sciences of the Juan March Foundation –presented by Fernández y Martínez Uribe-, and López Medina as coordinator. The other three were organised by GrandIR, a *spin off* led by de Castro: the first in August 2011 (*STM research data Management*), the second in May 2012 (*Advances in research data management in Spain*) and the third in November 2012 (*EuroCRIS autumn membership meeting*). At *STM research data Management*, researchers from different disciplines explain how they manage data and their requirements; a path which has been consolidated in *Advances in research data management in Spain*, together with the working group of data repositories and information managers of GrandIR, UPC and UOC.

There have been other fragmented contributions in *Primeres Jornades sobre Gestió de la Informació Científica-JGIC* [First Open-days on Scientific Information Management-JGIC], in April 2012, and the *5º Os-Repositorios* (May 2012) where we identified several contributions and people involved in data management. In October 2012, *EUDAT European Data Infrastructure 1st Conference* was held without much Spanish participation (RedIRIS and Barcelona Supercomputing Center as partners).

We detected two clear groups of interest in Barcelona and Madrid for **related projects** and **personal contacts** with professionals from the information management sector. Of the many projects related to data in general, only a few would fall under the definition of “research data” used in this report: Wf4Ever, agINFRA and SeaDataNet. The rest, seeming very close, take source data not considered as “research data”, which has been verified by contacting with their directors (Baeza-Yates or Larriba). Finally, we also detected some recommendations from the national reference centres. Carlos III Health Institute (biobanks register) or the National Centre for Polar Data and Polar Archive (SCOR Report). This is all reviewed extensively in the section “Evolution of Spanish contributions”.

Finally, as mentioned earlier, we performed searches in the **data banks registered in ODiSEA**. These were performed in general fields, by author or geography, if they had them, using the terms Spain or Spanish. From this work we can conclude the following: i) of the 183 items reviewed from ODiSEA (1-15 October) we only identified two Spanish registers: CEACS –the data library of the Juan March Foundation- and Digital.CSIC –the institutional repository of the CSIC, developed in Dspace-; ii) hardly any of the banks let one limit by country; iii) the searches offer ambiguous and weak results: only 20 approximately offered Spanish datasets. Although the figure is very low, the number of deposited *datasets* may be greater. This analysis has not been performed because the only purpose at this point was to gauge activity by Spanish researchers in terms of data; iv) we found that the only subjects represented were Mineralogy, Social Sciences, Earth Sciences and primarily Biomedicine in its array of branches. Other experts have also mentioned Spanish presence in Economic History, Chemistry and Climatology.

7.1 The evolution of Spanish contributions. Scientific data management

7.1.1 Bibliographic review of academic and professional literature

- Martínez Uribe and Macdonald (2008) discussed the revolution involved in open movements in scientific communication, the new forms of scientific work (e-science), and the new roles which must be assumed by information units as entities of preservation. They coined the term ***data curation*** as the management of data throughout its life-cycle, from its creation to the addition of value for its reuse. They cited the most important initiatives, especially ones in the UK: both lines of study and specific results.
- Torres Salinas (2009) analysed the **benefits and ways of sharing data**, as well as the **DAF initiative** in two ThinkEPI articles. In the first article the author reviewed bibliographically the opinions of the scientists themselves and government funding agencies –NIH mandate-, and ended by highlighting that this was a new opportunity for academic libraries, just as the open access movement was for publications. The second article described the *Data Audit Framework* (DAF) as a possible roadmap for any institution contemplating data management. The aim of DAF was to get a

real picture of the situation in order to propose improvements. The results of the audits in *Edinburgh, Bath, Glasgow, King's College, Southampton* and *Oxford* revealed a home-grown management system which required institutional policies to preserve data, and guides to help the researchers themselves.

- Melero (2010) reviews *Riding the Wave*, a key report commissioned by the European Commission to a group of experts to identify the **benefits and costs** of the implementation of a **reliable and stable global data infrastructure**. This report considered the following elements as vital: maintain flexibility; create incentives for sharing without losing privacy; preserve data enriching them with their context and origin; and funding models. Amongst other measures looking towards 2030, it proposed: creating a collaborative international data infrastructure with additional funds; measuring and rewarding the value of this data; or creating an international and inter-ministerial committee to direct this infrastructure. It identified certain sectors which badly need globalisation of data: climate change and the environment, and subjects such as energy, epidemiology, etc. It identifies the main concern as how to integrate the different parts of the process and stimulate participation.
- Keefer (2011) analysed the report by Itaca S&R on the role of funding agencies on data sustainability. Based on a survey of 25 European and North-American agencies, it detected a lack of procedural uniformity. However, it highlighted the role which these agencies could play if a data management plan were required to guarantee data preservation.
- In 2012 Torres Salinas, Robinson and Cabezas brought together the most noteworthy aspects of **data sharing** with a hybrid approach, linking our profession with the trends emerging in the world of research which is after all where data sharing commences. It provided a detailed discussion of the actual concept of research data, of ways of sharing and banks available, and the policies of funding agencies and journals.

Borrego (2012a) reviewed a report from the project ODE-Opportunities Data Exchange on the **integration of primary data and publications** to maintain the relation between them. The report provided examples illustrating the desire to reuse other researchers' data but a certain reticence to share their own data, giving legal issues as a reason. The preferred channels for storage are repositories and editorial platforms, although in reality there is little activity in either area. Relating them to publications has additional advantages: it helps to interpret data and provides value both to the researchers who share it and the publications themselves. Validation and preservation are the problems detected if it stays in the hands of publishers. Finally, it points out how in the research process, it appears that data centres –in charge of collecting and processing data- and libraries – where publications are stored- are going to adopt complementary roles as a result of data management.

- Peset (2012) reviewed the results of a survey on scientific information in the digital age. Unlike access to publications, access to scientific data is perceived as being more problematic due to the

lack of infrastructures, incentives and national policies. It is worth pointing out that for data producers the greatest problem are incentives, while for managers the problem is infrastructure.

- Finally, Borrego (2012) discusses the results of four reports: the first three were European (from the KE and ODE initiatives) and the fourth was North-American, from the Council of Library and Information Resources-CLIR.
 - The first studied a possible European level infrastructure which was mentioned by Castro. It focussed on the analysis of the incentives for the creators of data, training initiatives and finally the characteristics and funding requirements for the technological infrastructure using four examples.
 - The second report, based on a survey of librarians, tackled the implications of a correct citation of datasets, which, amongst other questions, would make it possible to quantify its use. In any event, it highlights the lack of demand in libraries for services using this data published as additional material to research articles.
 - The third report, which also came from ODE work, was based on interviews with experts who identified key issues: the role of editors, funding models, training requirements and specific standards, etc.
 - Finally, the fourth came from CLIR and was based on a qualitative study on researchers in the social sciences and on the analysis of training offered for data management. Among scientists it is noteworthy the lack of interest which it produces and why it is not stimulated and the difficulty in managing data due to the complexity of its life-cycle. While the last part of the work identifies a small number of centres which offer this training and in any case, always an advanced level of studies.
- From another field, psychology, Botella Ausina and Ortego Maté (2010) proposed **courses of action** given the particular reticence to data-sharing which they detected in scientists in their field. They characterised the methods and epistemological nature of psychology which has an effect on the rare custom of sharing. They summarise the benefits in their field of sharing because “it is an ethical principle, helping to prevent fraud and protecting against some threats to reliability” (p.266). They recommend the creation of tools to allow sharing and establishing strategic alliances with leading individuals and institutions in their discipline.
- From the *sector of Antarctic research*, Bermúdez, Barragán and Alonso (2011) presented the **PRADDA policy and protocol**, <http://hielo.igme.es> of the National Centre for Polar Data, from which they share information with the signatories of the Antarctic Treaty. This is a pioneering sector due to its internationalisation and the creation in 1989 of the *Committee in the Coordination of the Antarctic Data-CCAD*. Since 2005 the projects funded by the National R+D+I Plan are obliged to send a copy of their data to the National Centre of Polar Data upon which it is made freely available for approximately four years.

7.1.2 Meetings and conferences concerning research data management

- During the RECOLECTA **webinars** (2011) FECYT was presented as the Spanish partner of the OpenAire plus project (until 1 June 2014) whose objectives are to include data, link it to publications and generate additional services for the research community. On the other hand, Fernández and Martínez Uribe drew attention to an interesting case study: the digital collection of data and books of codes gathered by the CEACS library. Since 1987 it has been collecting microdata, aggregate and geographical data, as a result of purchase, strategic alliances (ICPSR) or produced by researchers on opinion, political systems, elections, social surveys, geographical and Spanish data. They pointed to three nuclei of specialist services for the researcher:
 - Collaborates in the selection and purchase of data in different formats, describes and offers a reference service, and participates in the establishment of access licenses or in communication with other centres;
 - Assists researchers in the use of applications for statistical analysis, visualisation, and utilities for extracting data (data scraping);
 - Assesses researchers from the moment the data is created until it is deposited for preservation (selection of formats, elaboration of codebooks, storage in the Foundation and in Dataverse...).

The use of this last open code application from Harvard University has several benefits, the most important of which are the following: it works with persistent identifiers and standardised metadata schema for social sciences (DDI) and automatically generates citations in standardised formats. It highlights the close relation with researchers and their contribution to creating a data-handling culture between them and their students.

- At the **STM meeting**, Estrada and Echenique (2011) outlined the state of databases in quantum chemistry and the need for the Quixote project given the lack of a standardised data model in the field of calculation. In 2010 they proposed the management of data resulting from the calculation of the size of small and medium sized molecules. They developed a technological infrastructure to standardise and semantically enrich data formats which can be implemented at an individual or project level, etc. At this first meeting the participants agreed about the need to:
 - Set up an interdisciplinary working group.
 - Analyse how STM researchers store their data internationally.
 - Encourage people to share data.
 - Promote the initiatives of eEspacio UNED and Digital CSIC.
 - Study whether they are being deposited on editorial platforms.
 - Create a protocol in research groups.
 - Provide incentives for data sharing (for example, recognising this as a scientific contribution)

- Finally, they detected great differences between disciplines and highlighted the potential of libraries and the Spanish Network of e-Science of the MICINN as a support for these activities.
- During the XIII SEDIC meeting, Pérez González (2010) presented **three cost models** from English-speaking countries for planning a digital information preservation policy in general, and data in particular, given that it is compulsory to deposit data in Great Britain. The first stemmed from the *Blue Ribbon Task Force on Sustainable Digital Preservation and Access*, a foundation with the participation of public and private organisms from the USA and Great Britain. Its approach is general, although among the digital information to be preserved it mentions academic discourse and research data. It emphasises that the problem has to be approached from a perspective of coordination between multiple agents. The second model, *Keeping Research Data Safe*, comes from the JISC. It studies three cases, Cambridge, King's College and York University, to establish its recommendations based on OAIS. It highlights the exhaustive detail of the benefits on several levels. The last of the models, LIFE, is based on the life-cycle proposed by the *British Library*, with OAIS once again. Depending on the size and purpose of the file it implements models on Excel templates to provide a summary of the costs for the institution.
- At the 75th Annual Meeting of the Society of American Archivists, Pérez González (2011) presented the **essentials for a management project** for data from the autonomous region of Galicia, **based on international good practices** amongst which he expressly mentioned the Netherlands, USA and Great Britain. The basis of his approach lies in the life-cycle of data, from its creation to preservation and services.
- At JGIC, Peset mentioned the work of **research data** distinguishing between data produced by researchers –small sets of data which can be stored in editorial platforms- and the data of large, generally governmental, producers, which are sometimes semantically compatible.

The remaining contributions dealing with data at this meeting were oriented towards the **measurement of scientific systems and the evaluation of science**, something which does not fall within the scope of this report: description of statistical systems which include data about science (UNEIX and Institut català d'estadística) or the bibliometric analysis of data (Borrego).
- The second meeting on **Research Data Management** (May 2012) brought together researchers and information managers.

Sorribas presented the work of the Marine Technology Unit (UTM) of CSIC in support of marine and polar research. They work on several, national and international data-management projects: RedICTS, EuroFleets, ICOS, ESONET, CMIMA and OGC. In 2008 the working group SCOR Spain (Comité Científico sobre Investigación Oceánica [Scientific Committee on Oceanic Research]) conducted a DAFO analysis on "*Reflections on the management and custody of oceanographic data in Spain. Existing resources and recommendations for the future*". It highlighted their international obligations and that marine data is complex due to its scale, subject matter, instruments..., but that they have a data model with specialized vocabularies, etc.

In turn, the phonetician at the UPM, Lahoz detected that there was no raw data repository which would admit among its metadata the representation of any singularity in this field.

Vallverdú presented the case of the Department of Signal Theory and Communications (UPC) where they are now storing and processing signal data.

There was quite a large representation of information managers on this occasion.

Castro mentioned the recent increase in awareness about data-handling and several national projects integrated in the Knowledge Exchange-KE initiative: British (JISC-MRD), German (DFG), Dutch (SURF) and Danish (DK). He emphasised the need to:

- Recognise data sharing through incentives
- Encourage training programmes after identifying interest groups among editorials, researchers, project assessors, information managers...
- Study the necessary infrastructure, taking into account previous successful cases and distinguishing the challenges stemming from the work with data from those resulting from technical implementation.
- Find cost models suited to the three main areas: physical sciences, biomedical sciences and social sciences and humanities. The presentation included some success cases of Dspace: the DataShare institutional repository in Edinburgh, LAGO in Colombia and Dryad.

Amongst other aspects, it concluded that “It is not vital to have a data repository before starting to plan data treatment strategies”.

In turn, Serrano presented the progress made by the working group FECYT/RECOLECTA for data repositories.

Zúñiga presented the initiatives of the UOC for data management, linked to the CRIS and the repository, but taking into account the researcher’s perspective.

Finally, FECYT gave details of OpenAirePlus, with the following objectives: linking publications, datasets and sources of funding to obtain “Enhanced Publications” implementing OAI-ORE. The aim was to offer more contextual information of very different classes (from copyright to the community to which it is directed) for two pilot disciplines: social sciences and humanities and life sciences.

- At **OS-Repositories** Castro, García and Rodríguez Miranda presented an international overview of the progress made in data-management and the ADDI Laboratory project of geographical documentation of heritage, UPV/EHU; and García García and Rodríguez Gairín presented a preliminary report on the results of ODISEA, which are described in greater detail below.

7.1.3 Projects related to data management and contact with professionals from the sector

The third method for identifying other agents meant conducting a search in Spain for related projects and personal contacts with professionals in the information management sector. We detected the following **groups of interest** in Barcelona and Madrid.

- i-VIU: Grup d'estudis mètrics sobre el valor i ús d'informació at the Universitat de Barcelona (Borrego & Urbano), where work consisted of metric and statistical studies of information in the digital environment.
- At CESCO (Ricard de la Vega) they have been working on infrastructures and search and dissemination platforms for scientific activity –for example data transmitted between the CERN and the Anella Científica-. The Anella Científica [Scientific Ring] is a high velocity communications network which allows scientists to re-access the data accumulated from their experiments for analysis.
- At the Carlos III University of Madrid, the Tecnodoc group (Méndez, Gómez & Hernández) has been studying activities related to scientific communication and measurement of research and last year they had Jane Greenberg (Dryad) as the Chair of Excellence.

On the other hand, there are many **projects related with data** in general. Some of these should be taken into consideration in this report because they work on research data.

- The Polytechnic University of Madrid-UPM, Astrophysics Institute of Andalusia and ISOCO, have joined forces to work on the Wf4Ever project which studies the evolution, exchange and collaboration in workflows. The main objective is to provide the means required to maximise participation and reuse of conserved research objects, while giving support to their evolution and versions and facilitating collaboration between scientists.
- Another international project agINFRA “Promoting data sharing and development of trust in agricultural sciences”, has the University of Alcalá de Henares as the Spanish member. This project aims to implement an open infrastructure to improve the transfer of scientific and technological results in agriculture, and also establish standards for exchange, including programmes and methodologies.
- The pan-European Network for Marine and Oceanic Data Management (SeaDataNet) has the participation of the Spanish Centre for Oceanographic Data of the Spanish Oceanographic Institute. The network is developing an interoperable system for the management of marine data and information, as we mentioned earlier. They also state that it is necessary to prepare legislation to regulate the rights and obligations of generators of data and information and the conditions of use of this information.
- At the Institute of Physics of Cantabria (IFCA) of the CSIC (Matorras) the research groups of Advanced Computation and e-Science³⁴ and that of Particle Physics take part in the Large Hadron Collider project of the CERN. In terms of pure data, the Worldwide LHC Computing Grid³⁵ forms part of this initiative of Particle Physics. This is an international project which connects more than 170 GRID computation centres in 36 countries which store, distribute and analyse some 25 million Gigabytes generated each year by the collider.

³⁴http://www.ifca.unican.es/computacion_avanzada_y_e-ciencia [Date of access 15/12/2012]

³⁵<http://wlcg.web.cern.ch/> [Date of access 15/12/2012]

- At the Pyrenean Institute of Ecology (IPE) and the Experimental School of Aula Dei (EEAD) of the CSIC, Vicente-Serrano and Begueria have developed a **SPEIbase** database gathers each month drought datasets at a global level based on a new drought index, the Standardised Precipitation-Evapotranspiration Index for the period 1901-2006. It can be consulted from its own search interface, although the storage and management of data is performed in the Digital repository.CSIC.
- Also a group of researchers from the Institute of History (IH) of the Centre of Human and Social Sciences (CCHS) of CSIC consisting of Crespo, Pérez, Maestre, and del Bosque have developed DynCoopNet -Dynamic Complexity of Cooperation-Based Self-Organizing Commercial Networks in the First Global Age-, a database with datasets on merchants, financiers, monopoly companies, trade routes, etc from 1400 to 1800, for Atlantic Europe, world Atlantic-American and Asia-Pacific. Available from Digital.CSIC.
- Other, seemingly very close projects take source data not considered as “research data”, which has been verified by examining their website or contacting with their directors: at the Pompeu Fabra University (Baeza-Yates), the Web Research Group is developing the project Modular and Extensible Platform for Data Mining on the Web, for all types of web data. The objectives proposed for this infrastructure include: the collection and extraction of Web data, the storage of objects (data, metadata, views and relations). The processes which include operations of recovery of information, of similitude, total or partial relevance (ranking), handling of objects, graphs and time sequences, and statistics which enable new views to be generated for data and the logical relations between them, and the presentation of views and relations using visualization techniques for structured data.
- At the Polytechnic University of Catalonia (Larriba), the Data Management group (DAMA-UPC) a research group for the management and analysis of large groups of data, has developed **Science-a**³⁶ [Date of access 9/12/2012] , a solution for the treatment and visualization of data, and support for the elaboration of projects. They take their data from Cordis, the EU Research and Development Information Service, which provides information on all the European projects which there have been to date. It manages all the project information, with abstracts, partners, quantity of money assigned, etc.

Many other examined projects were ruled out directly for this report, many of which were technological.

- **ADMIRE-Advanced Data Mining and Integration Research for Europe** (UPM) is driven by the difficulty involved in extracting significant information using data mining, combinations of multiple heterogeneous and distributed resources. ADMIRE proposes an architecture in DISPEL language to express data mining workflows and integration through user profiles.

³⁶<http://www.sciencea.com> [Date of access 15/12/2012]

- The **PlanetData** project, with participation of the UPM, establishes a sustainable European community of researchers to support organisations in the publication of their data in new and useful forms, making organisations more capable of interpreting the enormous quantities of data published online continuously. It includes structured and non-structured data, data flows, (micro) blog entries, digital files, e-Science resources, public sector datasets, and “cloud” linked data.
- Another project of the UPM, on a national level, is **MyBigData**, whose objective is to create a platform integrating new methods, techniques and tools to allow the integration of heterogeneous scientific data sources through the use of ontologies, incorporating new types of source data from researchers’ social networks, from the Web of Linked Data, and sensor networks.
- The objective of the **BabelData project** (UPM) is to develop techniques and algorithms for the construction of services capable of creating and using multilingual ontologies and data. To do this, the project deals with the following aspects:
 - Automatic localisation of ontologies
 - Multilingual mappings between ontologies in different languages
 - Models, methods, techniques and tools for the generation of multilingual linked data
 - Services for the integration of multilingual ontologies and multilingual Linked Data
- The **GeoBuddies project** (UPM) has developed an application which provides information and geo-spatial services for the pilgrims walking to Santiago de Compostela. The application supports mobile access and integration, is dynamic and context dependent with a set of resources and services, and uses data from the National Geographical Institute (IGN).

Finally, we have also detected some recommendations from national reference centres such as the Carlos III Health Institute. With the support of the Ministry of Economy and Competitiveness, it makes available to researchers an electronic platform for the register of biobanks and collections of samples, to facilitate public consultation and access to the materials they store.

The Spanish National Centre for Polar Data (CNDP) and the Geological and Mining Institute of Spain (IGME), deal with the generation of metadata and the storage, custody and management of raw data produced by research projects. Within the framework of the National Subprogram for Polar Research (SNIP) the CNDP drafted a data policy in 2007: “Proposal for a protocol of referral, storage and dissemination of Antarctic data in Spain”. PRADDA defines its area of application, the data types and formats, the procedures for sending data to the CNDP and the accessibility and availability of data through requests for controlled access.

Other recommendations have been given by national committees such as the Scientific Committee on Oceanic Research SCOR-Spain, in its report “Reflections on the management and custody of oceanographic data in Spain, existing resources and recommendations for the future”. To correct some of the problems

with oceanographic data management in Spain, it recommends that the oceanographic data management system should guarantee the following services:

- The collection, control of quality and storage of data to ensure they are available for the future. To do this, it is necessary to:
 - ✓ Rescue data and metadata which are currently inaccessible.
 - ✓ Data policy which contemplates:
 - Obligation to deposit data generated with public money in “Accredited Centres”.
 - Sources of funding to guarantee data management.
 - Regulations on restrictions and permissions of use.
 - ✓ A structure to enable the coordination and integration of the information of the centres working currently and which offers data services to the members of the scientific community who require it.
- Distribution of data to scientists, managers, industry and public: facilitate access to the data of meteorological and oceanographic stations operating in Spanish territorial waters. To do this, the existing databases need to be coordinated and integrated, to allow users to locate the necessary information through the Coordinating Centre.
- Establish protocols for data acquisition and processing: The appropriate protocols need to be selected and the relevant information sent to interested researchers.
- Development of data products geared to meet the evolution in demand.
- Adoption of a compulsory, national common data policy. In turn, data acquisition is work which needs to be recognised.

8. CASE STUDY: ODISEA³⁷

8.1 Background

As it has already been explained, the role of publishers in access to investigation data related to publications is a key aspect in its diffusion which has many advantages in the work of OpenAirePlus and Opportunities data Exchange (Fecyt 2012, Borrego 2012a).

Several researchers - Univ. Of Valencia, Univ. of Barcelona, Polytechnic Univ. of Valencia, Ramón Llull Univ., Univ. of Murcia and Catholic Univ. of Valencia- with different profiles (health, physical activity, documentation, and nanotechnology) have worked in identifying the research guidelines when depositing research data at the presentation in FECIES 2011 and 2012 (Peset and others, Ferrer-Sapena and others). This initial work was based on the analysis of the publisher's platforms which were used by the publications which had the highest impact in all the disciplines. A few other results in the social and humanities area were presented in 2012, in the 2nd Conference on the quality of social sciences and humanities publications, CRECS (García-García and others). This allowed the terminology used to be identified and the publishers to be classified in terms of the level of reuse they allowed.

Of the many processes which take place in the life cycle management of scientific data, the storage of said data is very important due to its implications in terms of preserving it for the future, citation, crediting authorship, using persistent links, and so on. For this reason, enquiries were made into the existence of a worldwide registry for storing research data, as there has been for open access repositories for many years: ROAR and OpenDoar. Nothing similar was found on a national or international scale except recently for Databib, who have been contacted in order to obtain more insight into their methodology and to coordinate projects.

The absence of a central system for gathering data has led to the need to establish a registry which gathers and classifies said data. The urgency of this action is due to the proliferation of specific deposits in various disciplines and to the decentralization of data storage deposits in the institutions' own repositories. In the first stages of data management in the investigation, basic tools are required such as ODISEA, which offer a global view of the situation, which in turn help to answer the various questions which may arise.

8.2 Aim

The research team proposed the creation of a tool which combined these deposits and classified them by subjects. The aim of this project is to facilitate the identification of research data storage sources in order to

³⁷ <http://odisea.ciepi.org/> [Date of access 15/12/2012]

allow professionals to at least be able to know, in an easy and trustworthy manner, where the researchers should log their data and if any disciplinary pools exist.

It is planned that ODISEA will also have the following tools available to the user and researchers: i) provide basic data on the registered banks, especially on their degree of openness; ii) be the starting point for the future meta-search tool for harvestable data, OPENDATASCIENCE; and iii) create awareness of research regarding management and production of research data.

8.3 Team

The team is made up of nine individuals; researchers from the following universities:

- Alicia García-García from the Catholic University of Valencia
- Antonia Ferrer-Sapena from the Polytechnic University of Valencia
- Fernanda Peset from the Polytechnic University of Valencia
- José Morales-Aznar from the Ramon Llull University
- Josep-Manuel Rodríguez-Gairín from the University of Barcelona
- Luís-Millán González from the University of Valencia
- Tomás Saorín from the University of Murcia
- Xavi García –Massó from the University of Valencia
- Florencia Atalia Dieci from the Polytechnic University of Valencia (Collaborator)

Currently the team is financed by the Ministry of Economy and Competitiveness's R&D&I National Plan: "OPENDATASCIENCE, resource centre for the preservation and management of open research data", CS02012-39632-C02-02, and will be part of the project's public website.

8.4 Methodology

The methodology used to gather existing data deposits and repositories has been based on several sources. Previous bibliographic studies have been checked, using the Web of Knowledge, Scopus, CSIC, and LISA databases, combining the keywords: "data sharing", "reuse", "data curation", "research data", and "data repositories". These works cited deposits such as DART (Treloar, 2006), ARROW (Payne y Treloar 2006), DRYAD (Greenberg, 2009), Protein Data Bank, and GenBank (Martínez-Urbe y Macdonald, 2009), and other mentioned combinations of the above (Torres-Salinas, 2012).

It has been observed that, in recent years, publishers have encouraged the delivery of this data, and most of them include guidelines regarding supplementary material in their author policy. The copyright policies of the most prominent scientific publishers have been examined regarding the articles' supplementary material, given that in areas such as Medicine or the Natural Sciences researches specify the public repositories where the data groups need to be deposited before the article can be published.

The Registry of Open Access Repositories (ROAR) has also been carried out, as well as OpenDoar (Directory of Open Access Repositories), and the digital files which contain the research data have been identified.

The classification of data banks is based on knowledge areas from Essential Science Indicators from the Web of Knowledge: Agricultural Science, Biology and Chemistry, Chemistry, Clinical Medicine, Computer Science, Economics and Business, Engineering, Environment Ecology, Geoscience, Immunology, Material Science, Mathematics, Microbiology, Molecular Biology and Genetics, Multidisciplinary, Neuroscience and Behaviour, Pharmacology.

The ODISEA website was developed from Drupal, within the CIEPI domain. The registry and search tool is administered by Rodríguez Gairín and is hosted separately.

8.5 The product: “ODiSEA: International Registry on Research Data”



Currently it has 183 deposits, among which are specialized banks, data libraries, repositories which accept data collections, and image banks.

This registry allows searches to be made in registered deposits by name, geographical area, scientific discipline, etc. Research is continuing to provide results as access to data is openly deposited and in the area of reuse policies.

The following information is collected from the data deposits, though it is not all available to the public. Name, Institution, Type of data stored, Format, Discipline, Geographical area, URL, Year, Quantity of data, OAI-PMH URL, Open access, Observations/notes: where software type and other information are included.

8.6 Lessons learned

During the development of ODISEA a series of common qualitative factors were perceived, which we have considered to be problematic for data management:

- Each of the deposits collects data in very different ways, linked more to each of the disciplinary sectors than to a common work framework.
- The information offered in each of the items on its page is not uniform and does not follow a standard pattern.
- In several instances the information which is returned in the form of results, the datasets, cannot be understood by anyone who is not an expert in the field.
- The websites do not offer enough information for users which are foreign to the system.
- It is not possible to easily know the number of datasets included for each item, or which have been registered in DOAR.

9. GOOD PRACTICES

Lastly, some good practices and important examples from the registered items are noted:

- Number of partners involved in the project: Dryad has 23 high-level partners, including scientific societies, publications, publishers, etc.
- Degree of openness: the American Association for the Advancement of Science, in its Science publication, explicitly allows any user to download, print, extract, reuse, archive, and distribute data related to the articles.
- Standardisation of the data: according to the Cyganiak graphic ³⁸ a given deposit also displays the data in the semantic web via linked open data such as UniProt
- Visibility of banks: certain items acquire "fame" among Spanish authors, which plays an important role in its diffusion: Dryad, Archer, DART, GenBank...
- Coordination on a national scale: agencies such as the National Institute of Health in the United States have led sector projects to promote data standards: NINS Common Data Elements³⁹
- Recommendations from large publishers: some of them have pointed out that deposits should be made in specific public banks depending on type and discipline. For example, the American Association for the Advancement of Science recommends molecular structure data be deposited in the Worldwide Protein Data Bank, protein sequences and DNA in GenBank, EMBL or DDBI and microarrays in Gene Expression Omnibus and Array Express.

³⁸<http://richard.cyganiak.de/2007/10/lod/> [Date of access 9/12/2012]

³⁹http://www.commondataelements.ninds.nih.gov/General.aspx#tab=Data_Standards [Date of access 9/12/2012]

10. REGARDING CASE STUDIES IN SPAIN

The time is right for a national strategy in our country, at the highest levels of authority and with the highest possible number of agents involved. A structure is required which allows for the coordination and integration of information in the centres which currently operate, and which offers data services to members of the scientific community.

This question should be discussed especially amongst **researches**, who have a vertical relationship with experts in their own discipline, on a national and international scale. It is vital to work on reference subject networks where there are mature models for data description, publishing technologies, and significant data groups. This expert knowledge by disciplines must be taken advantage of to encourage international good practice in our country. The protocols must come into force at the time the data is produced, that is, they should be aimed at being implemented at the research project level.

The sharing of data, recognized as part of the scientific work, should be encouraged in the **organisations** to which they belong, as well as in the evaluation agencies.

Funding entities should consider mandates which force data funded with public money to be deposited openly once the protocols and technological infrastructure are sufficiently developed.

The ways in which the academic **information managers** should orient themselves are not unique, but should not be duplicated. On the one hand, the organisations where the data is produced should implement management models for the life cycles of the data linked to the CRIS, the repository...; but the option which at this moment in time is seen to offer the greatest advantages and lowest storage cost is storage together with the publication in publishing platforms. The hybridization of the repository with the data itself would appear to be an efficient solution, and is in fact favoured by the European Union. This option reduces many different costs, though we should be aware that in numerous occasions the data referred to in an article is only part of a larger database. The countries with the greatest record in data storage carry out storage inversely: the database is stored and the literature associated to said data is added...

In any case, there is a noticeable low level of demand for library services worldwide, which entails a significant effort to promote our capacities amongst researchers. The role of library services is important, as is the need for library personnel to be adequately trained in order to achieve a good level of data management (its life cycle as a whole) and a reciprocal trust relationship in services between librarians and researchers.

Lastly, research data should be directed at this time to the general tendency towards innovation brought about by the handling of large quantities of data (BigData) through specialized ontologies. We need to work towards making data more productive: "The age of data-driven science"

Data is cool, data is business, data is science

11. CONCLUSIONS

The **open access movement** has generated debate regarding access, use, and business models pertaining to information obtained using public money, including scientific **publications** and research **data**.

Research data is gaining a reputation as a source of knowledge in itself, independent of the publications which may be used in the validation of the research results published in the articles, in order to generate new knowledge and be used in an interdisciplinary manner.

In order to ensure that data is being used, they should be made available and be accessible online; however their **nature** is **highly diversified**, depending on their discipline and especially on their life cycle. As a result the technical and legal requirements for guaranteeing access are complicated. According to their origin, data can be divided into the following categories: observational, experimental, and computational.

An adequate **management of data** requires at least the following aspects:

- **Policies** which define the roles and responsibilities of the different parties at the level of funding and institutional agencies.
- **Financial resources** in the long term, given that data is cumulative and is preserved.
- Specialised **human resources** (for generating data, and for using and preserving it).
- Coordinated **infrastructure** to guarantee its interoperability. Amongst the infrastructural requirements the following can be highlighted: preservation, access, data curation, data processing, distribution.

In order to comply with these aspects adequate training (both for the creators and for those who are responsible for maintenance), equipment, high capacity data storage, and high capacity networks are required.

- **Cultural change** in the parties involves: researches/data creators; Universities and Research Centres; Institutional Repositories; Data Centres; Data Managers, Users who reuse the data; Funding agencies; Scientific publication editors.

There is already an international agreement to consider the creation of an international and interdisciplinary infrastructure which guarantees access to research data.

In order to encourage openness in data deposits, different international bodies (EU, OCDE...) have issued **recommendations**, in line with current trends, aimed at:

- Enable access to scientific publications and data.
- Co-fund research infrastructures (repositories).
- Stimulate debate for future policies in this area.
- Stimulate debate between the different parties involved in scientific data management.

In addition, within the scope of the 7th Framework Program (*VII Programa Marco*), several pilot **projects** are being carried out aimed at the creation of electronic infrastructure and the deposit of articles in repositories (e.g. *OpenAire*) and the deposit of data (e.g. *Open Aire Plus*).

The responsibility for managing data belongs to the researchers, libraries, information technology services, and institutions in general. The creation of data is the responsibility of the researchers; however the management of their life cycle belongs to the information managers, which are the specialised librarians. The institutions must provide technical and organizational support which enables a model for a **data management plan** which allows the following: find and understand the data when they need to be used, guarantee the continuity of the project, avoid duplicates, validate the results, share, increase visibility in the case of open deposits, and the reputation of the investigation when citing data.

It is worth highlighting that research data management should be carried out during the entire research process: before the creation of data, during its creation and use, and during its life cycle.

Any data management plan, which should be included in every funding proposal, should consider the following:

- **Organisation and documentation** of the data according to standards. Data **storage, back-up, security**, and **sharing** mechanisms.

The '**dataset**' is a collection of data gathered during the execution of a research project. It forms the base of a given research project and is associated to a scientific publication. Datasets are stored and managed in interoperable network repositories.

The **format** in which data is archived should be open. The indexation of content should be facilitated (for ease of access), as should be the compression of data (less storage space) in a standard format for the researcher community.

The naming system is important for identifying content. It is also necessary to monitor version updates.

The dataset documentation provided by the researcher is included in the **metadata** registry.

Normalisation is a priority for achieving interoperability with other data management systems. There are several metadata standards.

The dataset stored should be associated to a unique **digital signature** which should be a URI. The URI is a string of characters which condenses the URL address (Uniform Resource Location) and the URN name (Uniform Resource Name) of the resource. There are many different systems, for example: PURL Uniform Resource Locator, DOI Digital Object Identifier, Accession, InChI.

- Ethical and legal points: within the **legal framework** of the management the legal rights over data and datasets, intellectual property, confidentiality, privacy, and data protection (be they public, restricted, or private), agreements, permissions, and licences should all be taken into account.

Any research project which contains personal information must comply with data protection law.

Data collection and databases are protected by intellectual property, which comprises the following aspects:

- ✓ Author's rights: belong to the author as long as the work is original.
 - ✓ Moral rights: are personal in nature; they belong exclusively to the authors and are inalienable.
 - ✓ Exploitation rights or copyright: are transferable. There are other licences apart from copyright such as Creative Commons where the conditions for sharing and reusing data are specified.
- A data **preservation plan** according to international standards. A final copy of the data should be archived in specialized data centres with different types of support (accounting for hardware and software becoming obsolete).

In order to improve the profitability of costs two basic **sustainability** principles are suggested: selecting data based on those which can be enriched or preserved and the use of scale economies in infrastructures in order to achieve transversality.

The following is an account of **Spanish initiatives in scientific data management**. To this end literature, journals, conferences, and projects have been studied:

- In terms of the **academic and professional literature**, the coining of the term **data curation** is noted. The **benefits and ways of sharing data** are considered, as well as **pricing models** for planning a policy for preserving digital information and data, the **benefits and cost** of creating a **trustworthy and stable global data infrastructure**, the possible **routes of action** when faced with resistance to data sharing, and the foundation of an autonomous data **management project** based on international good practices.
- Amongst the **seminars and conferences** related to research data management, the Recolecta project **Webinars** can be highlighted, as well as the **STM seminar** on the status of quantum chemistry databases and the need for the Quixote project in given the lack of a standardized model in the calculation discipline, the more important points of data sharing, the **integration of primary data and publications**, the distinction between research data produced by researches and data sourced from large producers, the **Research Data Management** seminar, or the international overview of data management advances presented in **OS-Repositorios**.
- There are several **projects** related to data management. For example:
 - The *Wf4Ever* project, on the evolution, exchange, and collaboration in workflows.
 - The agINFRA, "Promoting data sharing and development of trust in agricultural sciences", international project, aimed at implementing an open infrastructure to improve the transfer of scientific and technological results in agriculture.

- *Red Paneuropea de Gestión de Datos Marinos y Oceánicos (SeaDataNet)*, aimed at developing an interoperable system for managing maritime data and information.
- The Science-a project, regarding the treatment, visualization of data, support in the development of projects.
- ADMIRE (Advanced Data Mining and Integration Research for Europe) aimed at extracting important information, mining data, and achieving integration through user profiles.
- The PlanetData project, aimed at establishing a sustainable European community of researchers who support publications by publishing their data in new and useful ways.
- The MyBigData project, aimed at creating a platform which integrates new methods, techniques, and tool which allow the integration of heterogeneous scientific data sources.
- The BabelData project, aimed at developing techniques and algorithms for building services which are capable of creating and using ontologies and data in multiple languages.
- The GeoBuddies project, aimed at the development of an application which provides geo-spatial information and services for the Camino de Santiago pilgrims.

The case study in our country is “**ODiSEA**: International Registry on Research Data” which aims at facilitating the identification of research data storage sources which allows researchers to know where to deposit their data, and to know whether there are any data pools specific to their discipline. In addition it has other purposes: i) to provide basic data regarding the banks registered; ii) to be the starting point for a future meta search engine for data groups, *opendatascience*; and iii) to create awareness of research in the area of management and production of research data.

ODiSEA has several deposits, including specialized banks, data libraries, repositories, and image banks.

It allows searches to be made between registered deposits, according to different criteria. Research is on-going to provide results as the data is deposited openly.

This report has been developed within the **Recolecta** framework, a project which is managed and coordinated by FECYT for creating a network of interoperable, institutional, scientific repositories, aimed at facilitating open science, in accordance with *article 37 of Law 14/2011 of June 1, on Science, Technology, and Innovation*. Recolecta also aims to better serve and give Spanish research results and scientific production higher visibility.

FECYT, together with a **group of experts**, has prepared this report to support the management of research data. Several experts have participated, from institutions such as the Carlos III (UC3M) and Complutense of Madrid (UCM) Universities, the Spanish High Council for Scientific Research (CSIC), University of Alicante (UA), the Centre for Scientific and Academic Services of Catalonia (CESCA), the Juan March Institute, and the Polytechnic University of Catalonia (UPC), the latter of which has played the role of coordinator. Later, the Polytechnic University of Valencia (UPV) also joined this group.

12. BIBLIOGRAPHY

- Australian Government. Department of Education, Science and Training (2007). *A proposal for an Australian National Data Service*.
[HTTP://WWW.PFC.ORG.AU/PUB/MAIN/DATA/TOWARDSTHEAUSTRALIANDATACOMMONS.PDF](http://www.pfc.org.au/pub/main/data/towardstheaustraliandatacommons.pdf) [Date of access 8/12/2012]
- Barragán, Antonio; Bermúdez, Óscar (2007). *Propuesta del Protocolo de remisión, almacenamiento y difusión de datos antárticos en España*. Centro Nacional de Datos Polares y Archivo Polar.
[HTTP://WWW.UIB.ES/DEPART/DFS/APL/AAC/AA/ANTARTIDA/PGCDCAE/08_CNDP/PROTOCOLODATOS.PDF](http://www.uib.es/depart/dfs/apl/aac/aa/antartida/pgcdcae/08_CNDP/PROTOCOLODATOS.PDF) [Date of access 9/12/2012]
- Bailey, Charles W., Jr. (2013). Research Data Curation Bibliography. [HTTP://DIGITAL-SCHOLARSHIP.ORG/RDCB/RDCB.HTM](http://digital-scholarship.org/RDCB/RDCB.HTM) [Date of access 15/01/2013]
- Bailey, Charles W., Jr. (2013). Digital Curation Bibliography: Preservation and Stewardship of Scholarly Works. [HTTP://DIGITAL-SCHOLARSHIP.ORG/RDCB/RDCB.HTM](http://digital-scholarship.org/RDCB/RDCB.HTM) [Date of access 15/01/2013]
- Beguería S., Vicente-Serrano S.M., Angulo M. A multi-scalar global drought data set: the SPEIbase. *Bulletin of the American Meteorological Society*, DOI: 10.1175/2010BAMS2988.1.
- Bermúdez Molina, Oscar; Barragán Sanabria, Antonio; Alonso Gallego, Francisco (2007). Evaluación de la producción científica de la investigación española en la Antártida. *10as Jornadas Españolas de Documentación: FESABID 2007* Santiago de Compostela, 9-11 of May.
- Bermúdez, Óscar; Barragán, Antonio; Alonso, Francisco (2011). La gestión de los datos polares en España: una aproximación a la contribución de las ciencias de la vida. *Ecosistemas*, v. 20, n.1, pp. 94-103. [HTTP://REVISTAECOSISTEMAS.NET/PDFS/675.PDF](http://revistaecosistemas.net/pdfs/675.pdf) [Date of access 9/12/2012]
- Borgman, C.L. (2011). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*. [HTTP://PAPERS.SSRN.COM/SOL3/PAPERS.CFM?ABSTRACT_ID=186915](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=186915) [Date of access 8/12/2012]
- Borgman, C.L. (2012). *On Local or Global? Making Sense of the Data Sharing Imperative*. Talk at University of Southern Carolina on 9th April 2012
- Borrego, Angel (2012). Enriquecer las publicaciones con datos empíricos. *Reseñas de Biblioteconomía y Documentación*. ISSN: 2014-0894,
[HTTP://WWW.UB.EDU/BLOKDEBID/ES/CONTENT/ENRIQUECER-LAS-PUBLICACIONES-CON-DATOS-EMPÍRICOS-0](http://www.ub.edu/blokdebid/es/content/enriquecer-las-publicaciones-con-datos-empiricos-0) [Date of access 9/12/2012]
- Borrego, Angel (2012). Los retos de la gestión de datos de investigación. *Reseñas de Biblioteconomía y Documentación*, ISSN: 2014-0894, [HTTP://WWW.UB.EDU/BLOKDEBID/ES/CONTENT/LOS-RETOS-DE-LA-GESTIÓN-DE-DATOS-DE-INVESTIGACIÓN](http://www.ub.edu/blokdebid/es/content/los-retos-de-la-gestion-de-datos-de-investigacion) [Date of access 9/12/2012]
- Botella Ausina, Juan; Ortego Maté, María del Carmen (2010). Compartir datos: hacia una investigación más sostenible. *Psicothema*, Vol. 22, n 2. Pages. 263-269 Available at:
[HTTP://WWW.PSICOTHEMA.COM/PDF/3725.PDF](http://www.psicothema.com/pdf/3725.pdf) [Date of access 9/12/2012]

- Castro Martín, Pablo de; García Gómez, Consol; Rodríguez Miranda, Álvaro (2012). Gestión de datos de investigación en repositorios de acceso abierto: una visión panorámica y un caso práctico en la UPV/EHU. *Jornadas Os-Repositorios (5as Bilbao, May 23 to 25). Universidad del País Vasco*
- Castro, Pablo de (2012). Avances recientes a nivel internacional en la gestión de datos de investigación. *Advances in Research Data Management (Barcelona, May 10) GrandIR / Universitat Politècnica de Catalunya*. [HTTP://WWW.GRANDIR.COM/EN/TECNICAL-SESSION/ADVANCES-IN-RESEARCH-DATA-MANAGEMENT-IN-SPAIN/PROGRAMME](http://www.grandir.com/en/technical-session/advances-in-research-data-management-in-spain/programme) [Date of access 9/12/2012]
- Christensen-Dalsgaard, Birte et al (2012). *Ten recommendations for libraries to get started with research data management. Final report of the LIBER working group on E-Science / Research Data Management*. [HTTP://WWW.LIBEREUROPE.EU/SITES/DEFAULT/FILES/WGSC_20120801.PDF](http://www.libereurope.eu/sites/default/files/WGSC_20120801.pdf) [Date of access 9/12/2012]
- European Commission (2010). *Una agenda Digital para Europa*, [HTTP://EUR-LEX.EUROPA.EU/LEXURISERV/LEXURISERV.DO?URI=COM:2010:0245:FIN:ES:PDF](http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2010:0245:FIN:ES:PDF) [Date of access 9/12/2012]
- *Commission Decision on the adoption and a modification of special clauses applicable to the model grant agreement of FP7 C(2008) 4408 final* [HTTP://EC.EUROPA.EU/RESEARCH/PRESS/2008/PDF/DECISION GRANT AGREEMENT.PDF](http://ec.europa.eu/research/press/2008/pdf/decision_grant_agreement.pdf) [Date of access 8/12/2012]
- *Communication on scientific information in the digital age: access, dissemination and preservation (Com 2007)56*; [HTTP://EC.EUROPA.EU/RESEARCH/SCIENCE-SOCIETY/DOCUMENT LIBRARY/PDF_06/COMMUNICATION-022007_EN.PDF](http://ec.europa.eu/research/science-society/document_library/pdf_06/communication-022007_en.pdf) [Date of access 8/12/2012]
- National Science Foundation (2007). *Cyberinfrastructure Vision for 21st Century Discovery* [HTTP://WWW.NSF.GOV/PUBS/2007/NSF0728/INDEX.ISP](http://www.nsf.gov/pubs/2007/NSF0728/index.jsp) [Date of access 8/12/2012]
- *Digital Curation Center. All standards for any lifecycle action*. [HTTP://WWW.DCC.AC.UK/RESOURCES/STANDARDS/DIFFUSE/STANDARDS?FRAMEWORK_ID=0&LIFECYCLE_ID=0&SORT=TYPE](http://www.dcc.ac.uk/resources/standards/diffuse/standards?framework_id=0&lifecycle_id=0&sort=type) [Date of access 8/12/2012]
- Echenique, Pablo (2011). The Quixote Project: a pioneering work in managing Computational Chemistry research data. *STM Research Data Management*. Grandir ZCAM, (Zaragoza, August 25, 2011) [HTTP://DIGITAL.CSIC.ES/BITSTREAM/10261/39026/1/2011_08_QUIXOTE_MEETING.PDF](http://digital.csic.es/bitstream/10261/39026/1/2011_08_QUIXOTE_MEETING.PDF) [Date of access 9/12/2012]
- Estrada, Jorge; Echenique, Pablo (2011). From Databases in QC 2010, ZCAM, Sep 2010 onwards: a brief history of Quixote. *STM Research Data Management*. Grandir ZCAM, (Zaragoza, August 25, 2011) [HTTP://DIGITAL.CSIC.ES/BITSTREAM/10261/39038/1/2011_8_25_QUIXOTE_MEETING_V2.PDF](http://digital.csic.es/bitstream/10261/39038/1/2011_8_25_QUIXOTE_MEETING_V2.PDF) [Date of access 9/12/2012]
- Estrada, Marta; Álvarez, Enrique; Barragán, Antonio; Bermúdez, Óscar; García, M^a Jesús; Lavín, Alicia; Masqué, Pere; Pérez, Fiz F; Piera, Jaume (2011). *INFORME SCOR Comité Científico sobre Investigación Oceánica Representación española. Reflexiones sobre la gestión y la custodia de datos*

- oceanográficos en España. Recursos existentes y recomendaciones para el futuro.* [HTTP://WWW.SCOR-ES.ORG/DOCUMENTACION/REFLEXIONES_GESTION_DATOS.PDF](http://www.scor-es.org/documentacion/reflexiones_gestion_datos.pdf) [Date of access 9/12/2012]
- European Commission. (2010). *Global Research Data Infrastructures: The GRDI2020 Vision. GRDI2020 project.* [HTTP://WWW.GRDI2020.EU/REPOSITORY/FILESCARICATI/FC14B1F7-B8A3-41F8-9E1E-FD803D28BA76.PDF](http://www.grdi2020.eu/repository/filescaricati/FC14B1F7-B8A3-41F8-9E1E-FD803D28BA76.PDF) [Date of access 8/12/2012]
 - European Union. (2007). *Council Conclusions on scientific information in the digital age: access, dissemination and preservation;* [HTTP://WWW.CONSLIUM.EUROPA.EU/UEDOCS/CMS_DATA/DOCS/PRESSDATA/EN/INTM/97236.PDF](http://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/intm/97236.pdf) [Date of access 8/12/2012]
 - Ferrer-Sapena, A.; Villamón, M.; González-Moreno, L.M.; Peset, F.; Aleixandre, R.; García-García, A.; Morales-Aznar, J. Gestión de los datos de investigación como medida de calidad. M^a Teresa Ramiro, M^a Paz Bermúdez e Inmaculada Teva (Comps.) (2012). *Evaluación de la Calidad de la Investigación y de la Educación Superior (IX Foro)*, pp. 483. Granada: Asociación Española de Psicología Conductual (AEPC). ISBN: 978-84-695-3701-5.
 - García-García, A.; García-Massó, X.; Ferrer, A.; González-Moreno, L.M.; Peset, F.; Aleixandre, R. Mejores prácticas en reuso de conjuntos de datos publicados online como material adicional a los artículos. *2a Conferencia sobre calidad de revistas de ciencias sociales y humanidades (CRECS 2012)* [HTTP://WWW.THINKEPI.NET/CRECS2012](http://www.thinkepi.net/crecs2012) [Date of access 9/12/2012]
 - García-García, Alicia; García-Massó, Xavi; Ferrer-Sapena, Antonia; González, Luis-Millán; Peset, Fernanda; Rodríguez-Gairín, Josep-Manuel; Saorín, Tomás (2012). ODISEA: International Registry on Research Data. *5as Jornadas OS-Repositoryos "La motricidad de los repositorios de acceso abierto"* 23 to 25 of May 2012. Universidad de País Vasco.
 - Greenberg, Jane (2009). Theoretical Considerations of Lifecycle Modeling: An Analysis of the Dryad Repository Demonstrating Automatic Metadata Propagation, Inheritance, and Value System Adoption. *Cataloging & Classification Quarterly*, vol. 47, no. 3, p. 380-402.
 - *Is it Open Data?* [HTTP://ISITOPENDATA.ORG/](http://isitopendata.org/) [Date of access 8/12/2012]
 - Head of State. (2011). Law 14/2011, of June 1, on Science, Technology, and Innovation. *Official State Gazette*, vol. no. 131, no. 2nd of June 2011, pp. 54387 to 54455. [HTTP://WWW.BOE.ES/BOE/DIAS/2011/06/02/PDFS/BOE-A-2011-9617.PDF](http://www.boe.es/BOE/DIAS/2011/06/02/PDFS/BOE-A-2011-9617.PDF) [Date of access 9/12/2012]
 - Keefer, Alice (2011). La preservación de los datos de investigación y las agencias de financiación de la I+D. *Reseñas de Biblioteconomía y Documentación*. ISSN: 2014-0894, [HTTP://WWW.UB.EDU/BLOKDEBID/ES/NODE/130](http://www.ub.edu/blokdebid/es/node/130) [Date of access 5/1/2013]
 - Lacunza, Izaskun. OpenAIREplus: a European initiative as a driver for national RDM activity. *Advances in Research Data Management* (Barcelona, May 10, 2012) GrandIR / Universitat Politècnica de Catalunya [HTTP://WWW.GRANDIR.COM/EN/TECNICAL-SESSION/ADVANCES-IN-RESEARCH-DATA-MANAGEMENT-IN-SPAIN/PROGRAMME](http://www.grandir.com/en/technical-session/advances-in-research-data-management-in-spain/programme) [Date of access 9/12/2012]
 - Lahoz, José María (2012). Una perspectiva de la gestión de datos desde las Humanidades. *Advances in Research Data Management* (Barcelona, May 10) GrandIR / Universitat Politècnica de Catalunya.

[HTTP://WWW.GRANDIR.COM/EN/TECNICAL-SESSION/ADVANCES-IN-RESEARCH-DATA-MANAGEMENT-IN-SPAIN/PROGRAMME](http://www.grandir.com/en/technical-session/advances-in-research-data-management-in-spain/programme) [Date of access 9/12/2012]

- Lyon, Liz (2007). Dealing with Data: Roles, Rights, Responsibilities and Relationships. Consultancy Report. UKOLN.
[HTTP://WWW.UKOLN.AC.UK/UKOLN/STAFF/E.J.LYON/REPORTS/DEALING_WITH_DATA_REPORT-FINAL.DOC](http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.doc) [Date of access 8/12/2012]
- Lyon, Liz (2012). The Informatics Transform: Re-Engineering Libraries for the Data Decade. *The International Journal of Digital Curation*. Volume 7, Issue 1, 2012.
[HTTP://WWW.IJDC.NET/INDEX.PHP/IJDC/ARTICLE/VIEW/210/279](http://www.ijdc.net/index.php/ijdc/article/view/210/279) [Date of access 8/12/2012]
- Managing and sharing data. Best practice for researchers. *UK Data Archive*, 2011 (rev.).
[HTTP://WWW.DATA-ARCHIVE.AC.UK/MEDIA/2894/MANAGINGSHARING.PDF](http://www.data-archive.ac.uk/media/2894/managingsharing.pdf) [Date of access 8/12/2012]
- Marcos-Martín, Carlos; Soriano-Maldonado, Salvador-Luis (2011). Reutilización de la información del sector público y Open data en el contexto español y europeo. Proyecto Aporta. *El profesional de la información*, vol. 20, no. 3. p.291-297
[HTTP://ADMINISTRACIONELECTRONICA.GOB.ES/RECURSOS/PAE_020002228.PDF](http://administracionelectronica.gob.es/recursos/pae_020002228.pdf) [Date of access 9/12/2012]
- Martínez-Uribe, Luis, Macdonald, Stuart (2008). Un nuevo cometido para los bibliotecarios académicos: data curation. *El profesional de la información*, v.17, no. 3, May-June 2008
- Martinez-Uribe, Luis; Fernández, Paz (2011). La Biblioteca de Datos del Centro de Estudios Avanzados en Ciencias Sociales (CEACS) del Instituto Juan March como un servicio de apoyo a su comunidad científica. *Webinar FECYT/Recolecta sobre almacenamiento, conservación y gestión de los datos de investigación*. 7th of November to the 19th of December. Available at:
[HTTP://WWW.RECOLECTA.NET/BUSCADOR/WEBMINARS_PDF/CEACS_DATA_LIBRARY.PDF](http://www.recolecta.net/buscador/webminars_pdf/ceacs_data_library.pdf) [Date of access 9/12/2012]
- Martínez-Uribe, Luis; Macdonald, Stuart (2008). Un nuevo cometido para los bibliotecarios académicos: data curation. *El profesional de la información*, vol. 17, no. 3, p. 273-280.
[HTTP://WWW.ELPROFESIONALDELA INFORMACION.COM/CONTENIDOS/2008/MAYO/03.PDF](http://www.elprofesionaldelainformacion.com/contenidos/2008/mayo/03.pdf) [Date of access 9/12/2012]
- Martinez-Uribe, Luis; Macdonald, Stuart (2009). User Engagement in Research Data Curation. *Lecture Notes in Computer Science*, vol. 5714, p. 309-314.
- Matorras, Francisco (2009). The CMS Computing Model,
[HTTP://INDICO.CERN.CH/GETFILE.PY/ACCESS?CONTRIBID=2&RESID=0&MATERIALID=SLIDES&CONFID=68690](http://indico.cern.ch/getFile.py/access?contribId=2&resId=0&materialId=slides&confId=68690)
[Date of access 9/12/2012]
- Melero, Remedios (2010). Una pleamar de datos. *Reseñas de Biblioteconomía y Documentación*. ISSN: 2014-0894, [HTTP://WWW.UB.EDU/BLOKDEBID/ES/CONTENT/UNA-PLEAMAR-DE-DATOS](http://www.ub.edu/blokdebid/es/content/una-pleamar-de-datos) [Date of access 9/12/2012]
- Murillo, Angela; Greenberg, Jane (2012). Data-at-Risk, Metadata Registration for Data, and Dryad. *Advances in Research Data Management* (Barcelona, May 10, 2012) GrandIR / Universitat

Politécnica de Catalunya [HTTP://WWW.GRANDIR.COM/EN/TECNICAL-SESSION/ADVANCES-IN-RESEARCH-DATA-MANAGEMENT-IN-SPAIN/PROGRAMME](http://www.grandir.com/en/TECNICAL-SESSION/ADVANCES-IN-RESEARCH-DATA-MANAGEMENT-IN-SPAIN/PROGRAMME) [Date of access 9/12/2012]

- OECD (2007). *Principles and Guidelines for Access to Research Data from Public Funding*; [HTTP://WWW.OECD.ORG/DATAOECD/9/61/38500813.PDF](http://www.oecd.org/dataoecd/9/61/38500813.pdf) [Date of access 8/12/2012]
- European Parliament (2003). Directive 2003/98/ce of the European Parliament and Committee from the 17th of November 2003, regarding the reuse of public sector information. *Official Journal of the European Union*, vol. No. 345, no. 31st of December 2003 [HTTP://EU.VLEX.COM/SOURCE/DOUE-23/ISSUE/2003/12/31/1](http://eu.vlex.com/source/DOUE-23/ISSUE/2003/12/31/1) [Date of access 9/12/2012]
- Payne, Geoff; Treloar, Andrew (2006). The ARROW Project after two years: are we hitting our targets?. *Proceedings of VALA*, Melbourne. [HTTP://WWW.VALACONF.ORG.AU/VALA2006/PAPERS2006/57_TRELOAR_FINAL.PDF](http://www.valaconf.org.au/vala2006/papers2006/57_TRELOAR_FINAL.PDF) [Date of access 9/12/2012]
- Pérez González, Lourdes (2010). Modelo/s de coste para la preservación de los datos científicos en la e-ciencia. *XII Jornadas de Gestión de la Información. Valor económico de la información: mercados, servicios y rentabilidad*. SEDIC, 18-19th of November. Available at: [HTTP://EPRINTS.RCLIS.ORG/BITSTREAM/10760/8555/1/PEREZ.PDF](http://eprints.rclis.org/bitstream/10760/8555/1/PEREZ.PDF) [Date of access 9/12/2012]
- Pérez González, Lourdes (2011). E-ciencia y la información como bien público, algunas propuestas. *XIII Jornadas de Gestión de la Información*. BNE, Madrid 17th and 18th of November. Available at: [HTTP://WWW.SEDIC.ES/SERVICIOS-ETICA-DCHOS-HUMANOS.PDF](http://www.sedic.es/servicios-etica-dchos-humanos.pdf) [Date of access 9/12/2012]
- Pérez González, Lourdes (2011). Towards a Galician Data Commons. 75th Annual Meeting of the Society of American Archivists. [HTTP://DLC.DLIB.INDIANA.EDU/DLC/BITSTREAM/HANDLE/10535/7870/TOWARDS%20A%20GALICIAN%20DATA%20COMMONS.PDF?SEQUENCE=1](http://dlc.dlib.indiana.edu/dlc/bitstream/handle/10535/7870/TOWARDS%20A%20GALICIAN%20DATA%20COMMONS.PDF?SEQUENCE=1) [Date of access 9/12/2012]
- Pérez, Esther; Maestre, Roberto; Bosque, Isabel del; Crespo Solana, Ana; Sánchez-Crespo, Juan Manuel (2010). DynCoopNet-CSIC-Objetivos y Estado Actual del Proyecto [HTTP://HUMANIDADES.CCHS.CSIC.ES/CCHS/SIG/PDF/PDF/DYNCOOPNET/ESTHERPEREZ_DYNCOOPNET.PDF](http://humanidades.cchs.csic.es/cchs/sig/pdf/pdf/dyncoopnet/estherperez_dyncoopnet.pdf) [Date of access 9/12/2012]
- Pérez, Esther; Maestre, Roberto; Bosque, Isabel del; Crespo Solana, Ana; Sánchez-Crespo, Juan Manuel (2012). Modelling and Implementation of a spatio-temporal historic GIS. Self-organizing Networks and GIS Tools. Cases of Use for the Study of Trading Cooperation (1400-1800). *Journal of Knowledge Management, Economics and Information Technology*. [HTTP://DIGITAL.CSIC.ES/BITSTREAM/10261/59170/1/Modelling_and_Implementation_of_a_Spatiotemporal_Historic_GIS.pdf](http://digital.csic.es/bitstream/10261/59170/1/Modelling_and_Implementation_of_a_Spatiotemporal_Historic_GIS.pdf) [Date of access 9/12/2012]
- Pérez, Esther; Maestre, Roberto; Bosque, Isabel del; Crespo Solana, Ana; Sánchez-Crespo, Juan Manuel (2010). Integración de bases de datos históricas en una IDE. Comercio mundial y redes de cooperación en la primera Edad Global (1400-1800), *Jornadas Técnicas de la Infraestructura de Datos Espaciales de España*

[HTTP://DIGITAL.CSIC.ES/BITSTREAM/10261/24908/1/IIDEE09_DYNCOOPNET_FINAL.PDF](http://digital.csic.es/bitstream/10261/24908/1/IIDEE09_DYNCOOPNET_FINAL.PDF) [Date of access 9/12/2012]

- Peset, F.; Aleixandre, R.; Villamón, M.; González-Moreno, L.M.; Ferrer, A. (2012). Open Data in the scientific world: OpenDataScience project. In: Lidia Cabello y M^a Paz Bermúdez (Comps.) (2011). *Evaluación de la Calidad de la Investigación y de la Educación Superior (VIII Foro FECIES)*, pp. 481-482. Granada: Asociación Española de Psicología Conductual (AEPC). ISBN: 978-84-694-3488-8.
- Peset, Fernanda (2012). Opiniones del sector científico sobre la preservación de la información. Blok de BiD. *Reseñas de Biblioteconomía y Documentación*. ISSN: 2014-0894, September.
[HTTP://WWW.UB.EDU/BLOKDEBID/ES/CONTENT/OPINIONES-DEL-SECTOR-CIENT%C3%ADFICO-SOBRE-LA-PRESERVACIÓN-DE-LA-INFORMACIÓN](http://www.ub.edu/blokdebid/es/content/opiniones-del-sector-cient%C3%ADfico-sobre-la-preservaci%C3%B3n-de-la-informaci%C3%B3n) [Date of access 9/12/2012]
- *Riding the Wave: How Europe can gain from the rising tide of scientific data* (2010).
[HTTP://CORDIS.EUROPA.EU/FP7/ICT/E-INFRASTRUCTURE/DOCS/HLG-SDI-REPORT.PDF](http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf) [Date of access 8/12/2012]
- Serrano-Muñoz, Jordi (2012). FECYT/Recolecta Working Group for Data Repositories. Advances in Research Data Management (Barcelona, 10 mayo) GrandIR / Universitat Politècnica de Catalunya.
[HTTP://WWW.GRANDIR.COM/EN/TECNICAL-SESSION/ADVANCES-IN-RESEARCH-DATA-MANAGEMENT-IN-SPAIN/PROGRAMME](http://www.grandir.com/en/technical-session/advances-in-research-data-management-in-spain/programme) [Date of access 9/12/2012]
- Sorribas Cervantes, Jordi (2012). La gestión de datos (marinos) desde la perspectiva de un centro de datos de investigación. *Advances in Research Data Management* (Barcelona, 10 mayo) GrandIR / Universitat Politècnica de Catalunya. [HTTP://WWW.GRANDIR.COM/EN/TECNICAL-SESSION/ADVANCES-IN-RESEARCH-DATA-MANAGEMENT-IN-SPAIN/PROGRAMME](http://www.grandir.com/en/technical-session/advances-in-research-data-management-in-spain/programme) [Date of access 9/12/2012]
- Special Online Collection: Dealing with Data (2011). *Science*. 11 February
[HTTP://WWW.SCIENCEMAG.ORG/SITE/SPECIAL/DATA/](http://www.sciencemag.org/site/special/data/) [Date of access 8/12/2012]
- Torres-Salinas, Daniel (2010). Compartir datos (data sharing) en ciencia: contexto de una oportunidad. *Anuario ThinkEPI*, vol. 4, p. 258-261. [HTTP://WWW.THINKEPI.NET/COMPARTIR-DATOS-DATA-SHARING-EN-CIENCIA-EL-CONTEXTO-DE-UNA-OPORTUNIDAD](http://www.thinkepi.net/compartir-datos-data-sharing-en-ciencia-el-contexto-de-una-oportunidad) [Date of access 9/12/2012]
- Torres-Salinas, Daniel (2010). Hacia la gestión de datos de investigación en las universidades: la Data asset framework. *Anuario ThinkEPI*, vol.4, p. 262-265. [HTTP://WWW.THINKEPI.NET/PRIMEROS-PASOS-HACIA-LA-GESTION-DE-DATOS-DE-INVESTIGACION-EN-LAS-UNIVERSIDADES-LA-INICIATIVA-DAF](http://www.thinkepi.net/primeros-pasos-hacia-la-gestion-de-datos-de-investigacion-en-las-universidades-la-iniciativa-daf) [Date of access 9/12/2012]
- Torres-Salinas, Daniel; Robinson-García, Nicolás; Cabezas-Clavijo, Álvaro (2012). Compartir los datos de investigación: introducción al data sharing. *El profesional de la información*, March-April, vol. 21, no. 2, p. 173-184. [HTTP://HDL.HANDLE.NET/10760/16786](http://hdl.handle.net/10760/16786) [Date of access 9/12/2012]
- Treloar, A (2006). The Dataset Acquisition, Accessibility, and Annotation e-Research Technologies (DART) Project: building the new collaborative e-research infrastructure. *Proceedings of AusWeb06, the Twelfth Australian World Wide Web Conference*, Southern Cross University Press, Southern Cross University, July. [HTTP://AUSWEB.SCU.EDU.AU/AW06/PAPERS/REFEREED/TRELOAR/PAPER.HTML](http://ausweb.scu.edu.au/aw06/papers/refereed/treloar/paper.html) [Date of access 9/12/2012]

- Treloar, A.; Groenewegen, D.; Harboe-Ree, C. (2007). The Data Curation Continuum. Managing Data Objects in Institutional Repositories. *D-Lib Magazine*, vol. 13, no. 9/10.
[HTTP://WWW.DLIB.ORG/DLIB/SEPTEMBER07/TRELOAR/09TRELOAR.HTML](http://www.dlib.org/dlib/september07/treloar/09treloar.html) [Date of access 9/12/2012]
- University of Melbourne Research Data Management Policy
[HTTP://RESEARCH.UNIMELB.EDU.AU/INTEGRITY/CONDUCT/DATA/REVIEW](http://research.unimelb.edu.au/integrity/conduct/data/review) [Date of access 8/12/2012]
- Vallverdú, Francesc (2012). Research Data Management: A Perspective from a University Department. *Advances in Research Data Management* (Barcelona, 10 mayo) GrandIR / Universitat Politècnica de Catalunya. [HTTP://WWW.GRANDIR.COM/EN/TECNICAL-SESSION/ADVANCES-IN-RESEARCH-DATA-MANAGEMENT-IN-SPAIN/PROGRAMME](http://www.grandir.com/en/technical-session/advances-in-research-data-management-in-spain/programme) [Date of access 9/12/2012]
- Van der Graaf, M. and Waaijers, L. (2011). *A Surfboard for Riding the Wave. Towards a four country action programme on research data*. [HTTP://WWW.KNOWLEDGE-EXCHANGE.INFO/SURFBOARD](http://www.knowledge-exchange.info/surfboard) [Date of access 8/12/2012]
- Vicente-Serrano S.M., Beguería S., López-Moreno J.I. (2010). A Multi-scalar drought index sensitive to global warming: The Standardized Precipitation Evapotranspiration Index – SPEI. *Journal of Climate* 23(7), 1696-1718, DOI: 10.1175/2009JCLI2909.1
- Vicente-Serrano S.M., Beguería S., López-Moreno J.I., Angulo M., El Kenawy A. A global 0.5° gridded dataset (1901-2006) of a multiscalar drought index considering the joint effects of precipitation and temperature. *Journal of Hydrometeorology* 11(4), 1033-1043, DOI: 10.1175/2010JHM1224.1.
- Wacowicz, Monica; Crespo Solana, Ana; Bernabé Poveda, Miguel Ángel (2010). Visualization and Space Time representation of Dynamis, non linear Spatial Data in DynCoopNet Project.
[HTTP://DIGITAL.CSIC.ES/BITSTREAM/10261/23414/1/SCIENTIFICREPORT.%20WACHOWICZ.CRESPO.BERNA%20BE.PDF](http://digital.csic.es/bitstream/10261/23414/1/SCIENTIFICREPORT.%20WACHOWICZ.CRESPO.BERNA%20BE.PDF) [Date of access 9/12/2012]
- Zúñiga, Anna (2012). Reptes i dificultats en la implementació d'estratègies institucionals per la gestió de dades. *Advances in Research Data Management* (Barcelona, 10 mayo) GrandIR / Universitat Politècnica de Catalunya. [HTTP://WWW.GRANDIR.COM/EN/TECNICAL-SESSION/ADVANCES-IN-RESEARCH-DATA-MANAGEMENT-IN-SPAIN/PROGRAMME](http://www.grandir.com/en/technical-session/advances-in-research-data-management-in-spain/programme) [Date of access 9/12/2012]

Regarding the participating institutions

The **Spanish Foundation for Science and Technology (FECYT)** is a public foundation dependent on the Spanish Ministry of Finance and Competitiveness. Under the principles of rationalisation, transparency and efficiency, it works to develop social participation instruments that favour science; to be the right tool for science dissemination and boosting science culture; to act as a communication channel for the Spanish scientific community abroad; and to become a metric reference for Spanish R&D&i. FECYT also supports scientific information and resource management structures.

Among the activities carried out by FECYT is the RECOLECTA⁴⁰ project, which coordinates the creation of an interoperable institutional repository network that could be considered the first national initiative to create an infrastructure that enables open science. The aim is also to better serve and give Spanish research results and scientific production higher visibility.

Carlos III University of Madrid (UC3M)⁴¹ was established in 1989 and in 2010 received accreditation as a Campus of International Excellence. The University has three centres: the Social and Juridical Sciences Faculty, the Humanities Faculty, and the Higher Polytechnic School, which are all located in three different campuses in Getafe, Leganés, and Colmenarejo.

Complutense University of Madrid (UCM)⁴² was established in Alcalá de Henares by Cardinal Cisneros in 1499. In 2009 it received accreditation as a Campus of International Excellence. The Complutense University has two campuses: one in Moncloa and one in Somosaguas. Its 78 degree programs cover a wide range of specialities, and are grouped into five knowledge branches: Humanities, Experimental Sciences, Health Sciences, Social and Juridical Sciences, and Technology.

The **Spanish High Council for Scientific Research (CSIC)**⁴³ is the largest public institution dedicated to research in Spain and the third largest in Europe. It is dependent on the Ministry of Economy and Competitiveness⁴⁴, through the Research, Development, and Innovation Secretary of State. Its main aim is to develop and encourage research in scientific and technological progress, and to this end it is open to collaboration with other Spanish and international bodies. The engine behind its research is made up by its centres and institutions, distributed through all the autonomous communities, including more than 15,000 workers, of which over 3,000 are staff researchers and a few others are doctors and scientists in training. Due to its multidisciplinary and diverse nature, CSIC works in all areas of knowledge. Its activity, which

⁴⁰ <http://www.recolecta.net> [Date of access 6/12/2012]

⁴¹ <http://www.uc3m.es> [Date of access 6/12/2012]

⁴² <http://www.ucm.es/> [Date of access 6/12/2012]

⁴³ <http://www.csic.es/> [Date of access 6/12/2012]

⁴⁴ <http://www.micinn.es/> [Date of access 6/12/2012]

comprises everything from basic research to technological development, is organized around eight scientific and technical areas. In addition, the CSIC manages a group of important infrastructures, the most extensive and complete network of specialized libraries, and has mixed investigation units. As a result of the signing of the Declaration of Berlin in 2006, the institutional repository DIGITAL.CSIC was born in 2008. In 2010, the raw data was included as content. The pioneering experience in this respect was *SPEIbase: a global 0.5° gridded SPEI data base*, featured in the first CSIC Open Bulletin⁴⁵.

University of Alicante (UA)⁴⁶ was established in 1979 on the structure of the Centre for University Studies (CEU) which had begun to operate in 1968. The University has fifty degree programs, over sixty University Departments and research units and groups in the area of Social and Juridical Sciences, Experimental Sciences, Technology, Humanities, Education, and Health, as well as fifteen University and Cross-university Institutes, and nine university headquarters.

Polytechnic University of Catalonia BarcelonaTech (UPC)⁴⁷ was established in 1971 and is specialized in engineering, architecture, and sciences. It is represented in eight Catalan cities. UPC is close to its environment and plays an active role in its economic, cultural, and social development. UPC is also represented in the rest of the world, always with the same willingness to serve and be involved.

Polytechnic University of Valencia (UPV)⁴⁸ has been a university since 1971 and is a dynamic and innovative public institution, dedicated to research and teaching. At the same time it also has strong ties with its social environment where it carries out its activities, and has elected to have a strong presence abroad. One of the pillars of Polytechnic University of Valencia's social status has been its capacity for research. Its departments, research centres and institutions carry out applied research projects together with national and international bodies and companies.

Centre for Scientific and Academic Services of Catalonia (CESCA)⁴⁹ manages infrastructures based on the information and communication technologies (e-infrastructures) in order to serve the university and research. The Centre has a vision to be leaders in the management and use of TIC for improving the quality and efficiency of the university and research system, taking advantage of scale economies through cross-university cooperation, professional good practice, and sharing resources.

The **Juan March Foundation**⁵⁰ was created in 1955. In 1987 the Juan March Institute for Research and Study was born, as well as its dependant, the Centre for Advanced Studies in Social Science (CEACS).

⁴⁵ <http://digital.csic.es/handle/10261/26261> [Date of access 6/12/2012]

⁴⁶ <http://www.ua.es/> [Date of access 6/12/2012]

⁴⁷ <http://www.upc.edu> [Date of access 6/12/2012]

⁴⁸ <http://www.upv.es> [Date of access 13/12/2012]

⁴⁹ <http://www.cesca.cat/> [Date of access 13/12/2012]

⁵⁰ <http://www.march.es/> [Date of access 13/12/2012]

Currently CEACS is a post-doctorate research centre which supports the research of the researches who work there and the scientific community of the Centre as a whole. From 1991 it has been purchasing quantitative databases from international bodies and data providers, and as of 2010 they have a Social Sciences scientific data Library and Repository hosted in Harvard University.



FECYT
FUNDACIÓN ESPAÑOLA
PARA LA CIENCIA
Y LA TECNOLOGÍA