ESFRI WORKING GROUP ABOUT DIGITAL REPOSITORIES

ESFRI POSITION PAPER

Introduction

In February 2007, the European Commission has issued a communication to the European Parliament, the Council and the European Economic and Social Committee entitled “On scientific information in the digital age: access, dissemination and preservation” referred to as COM(2007) 56 final. In that communication, the Commission aims to signal the importance of and launch a policy process on (a) access to and dissemination of scientific information and (b) strategies for the preservation of scientific information across the Union. In this respect, the Commission invites the Member States to explore common strategies and to discuss the relevant issues and challenges – organizational, legal, technical and financial – highlighted in the communication. Fora such as CREST and ESFRI could contribute to shaping the discussion. This paper represents a contribution of ESFRI on the topic, with suggestions for policy and recommendations.

This contribution is on a similar line to the OECD "Recommendation" on "Access to Research Data from Public Funds". However, because of the wide extension of the subject, it was necessary to focus the ESFRI statement on issues specifically related to Research Infrastructures, i.e. their contribution to progress of science and technology.

This paper is structured around 5 topics:

1. Availability
2. Permanency
3. Quality
4. Right of use
5. Interoperability

For each topic, we make recommendations regarding:

- the main policy requirements, and
- some implementation issues related to these policies.

1. Availability of data

Research Infrastructures are key ingredients to build and sustain the European Research Area (ERA). As a consequence, Research Infrastructures' production, in terms of scientific publication, technology development, human skills and knowledge must also be considered as part of the ERA and must be based on open access principles. Research Infrastructures should guarantee that raw research data are made available through portals and databases. Thus, the Research Infrastructure component of the ERA would not be considered just as a collection of large scale facilities with a whole set of scientific communities (covering all scientific disciplines), but also as a world of knowledge produced by all Research Infrastructure stakeholders and readily accessible for public research from all inter-operable digital repositories.

Recommendation #1a (Policy): *Any ESFRI-labelled Research Infrastructure must have a policy on availability of data and metadata, agreed by its user community, and an implementation of that policy.*

In order to guarantee the overall availability, quality, origin, etc of data and its context, it is advised that data repositories remain as close as possible to the data sources. From raw data

to grey literature, the Research Infrastructures must play an active role to maintain the availability of data, including the metadata which are required to interpret and re-use them effectively and an appropriate search interface which allows data to be found. Beyond the grey literature, scientific publications (white literature) are also relevant for being part of the Research Infrastructure digital repository content. However, this aspect of digital repositories should not be considered as a stand-alone issue but rather complement the normal activity of the scientific user community.

Recommendation #1b (Implementation): *Research Infrastructures must be responsible for implementing and managing the availability of data, from raw data to the grey literature, associated with corresponding research activities.*

2. Permanency

Permanency/preservation of data is a difficult problem to face today because of the size and complexity of the data, the rapid changes in technology, and the diversity of experiments. Even though most experimental data are analyzed when they become available, their permanency is often equally important. This is true for many scientific disciplines, where data are not just the recording of an event, but a contribution to the big book of knowledge: Biology and Medicine, Earth and Environmental Sciences, Astronomy and Astrophysics, Engineering, Social Sciences and Humanities, are among the many sciences which require long term conservation of data. Data must remain available for later use, re-interpretation or confrontation with more recent methodologies or theories. Permanency is not just continuity of storage conditions, it also means maintenance (media migration) and curation (availability for use) of data.

Recommendation #2a (Policy): *Any ESFRI-labelled Research Infrastructure must have a policy covering preservation, maintenance and curation of data, agreed with their user community, and an implementation of that policy.*

The permanency of data requires that they remain usable over time, independently of evolving formats, supports and standards. Research Infrastructure managed digital repositories must therefore be responsible for the preservation, maintenance and curation of data. The permanency of data is a real added value from digital repositories which must be tackled properly by all Research Infrastructures. Together with "availabaility", "permanency" is a prerequisit for the "interoperability", "quality" and "right of use", described below, which provide a sustained and usable "world of scientific data".

Recommendation #2b (Implementation): *Permanency of the scientific data is a requirement necessary for the usability of the repositories, which must be part of the normal implementation and operation of digital repositories by Research Infrastructures.*

3. Quality

Information on data, which is accessible (open access) with ad-hoc rights of use is not complete and reliable unless its quality is properly recorded. This qualification can come in several forms such as description of laboratory procedures for data or peer-review for publication. It does not make sense to release data unless their intrinsic quality is stated. The scientific community should not be provided with unqualified data by Research Infrastructures. Research data should have sufficient associated information to allow a user to judge its intrinsic quality (correctness) and extrinsic quality (usefulness for purpose).

Recommendation #3a (Policy): *Any ESFRI-labelled Research Infrastructure must agree with their user community on a policy and implementation of data quality, covering also*

what additional information should be collected and preserved such that an end user can judge the quality both intrinsically (correctness) and extrinsically (for any intended use).

The quality of data must be evaluated and guaranteed, jointly by the Research Infrastructures and their user communities. Quality of data is understood differently from value of data. The latter being a scientific judgement under the control of the user community, whilst the former refers to the quality with which the data were produced, collected and stored. Research Infrastructures should maintain enough metadata regarding the data and its context to enable the assessment of the quality of the repository content. This metadata is also important to support decision processes around archival versus disposal.

Recommendation #3b (Implementation): *Any ESFRI-labelled Research Infrastructure should make available information regarding the quality of its data through the provision of sufficient metadata to assess it.*

4. Rights of use

"Open access to research data from public funding should be easy, timely, user-friendly and preferably Internet-based" (OECD Recommendation concerning Access to Research Data from Public Funding). When a publicly-funded Research Infrastructure is used to produce data and knowledge, the outcome of this research cannot be considered as closed (or restricted) information, unless specific prior arrangement is in place to do so. Raw data should certainly be considered as open access data. However, the concept of "public rights" may need some clarification. Raw scientific data are not covered by IPRs, nevertheless directive96/9/EC on the legal protection of databases protects efforts in organizing research data. Same concept holds naturally for any added-value work to the raw data, like publications and reports. At global scale, worldwide scientific collaborations yield new types of problems. However, it is difficult to imagine that raw data can be of any use without the proper handling tools. It may also be considered that raw data are not available immediately, for those researchers producing them to have time to analyze and publish. Therefore, beyond a reasonable time period to allow publications, right of use should be granted to the public for all raw data produced with public funding, except for : (1) prior publication reserved for the experimental team (2) where the value to the public is greater if restricted access allows generation of products or services.

Recommendation #4a (Policy): *Any ESFRI-labelled Research Infrastructure must have a policy agreed with their user community guaranteeing public accessibility of data, with restrictions if appropriate to their community, and an implementation to support that policy.*

Implementation of right of use relies on the proper definition of the scope of data and users communities. Through the general principle of granting the right of use of publicly funded research data, some specific contexts, like research cofunded with non-public bodies may eventually require the use of an authentication and authorization infrastructure to support the proper access and tracking of data use.

Recommendation #4b (Implementation): *Any publicly-funded Research Infrastructure will grant the right of use to the public, unless otherwise required for early publication constraint or pre-established contractual arrangement.*

5. Interoperability

Interoperability between repositories is also a requirement to develop interactions within and across disciplines. This interoperability is required at all levels. It may be enabled by data formatting conventions (mostly for raw data or standardized interfaces) or it may be provided by conversion built into the tools to browse and access the data bases which then must

guarantee standardized interfaces at the time of access or exploitation. From one repository, it should be possible to navigate across the world of data in other repositories. Furthermore, it is expected that cross-referencing between digital repositories will add value to the world of scientific data.

Recommendation #5a (Policy): *Any ESFRI-labelled Research Infrastructure must have a policy agreed with their user community covering interoperation based on open standards and an implementation supporting it.*

Interoperability of digital repositories relies primarily on the use of open standards to archive and access scientific data. A very basic requirement, which is linked also to the availability and the quality of data is the identification of data, for example based on the URI/URNs or DOIs. Such identification will support also the possibility of cross-referencing among digital repositories and, through taxonomies, Research Infrastructures from different scientific fields.

Recommendation #5b (Implementation): *Digital repository content must use some form of universal identification for data independently of the values. It is highly advisable that this interoperability framework is shared by all the ESFRI-labelled Research Infrastructures under a common European umbrella.*

WORKING GROUP COMPOSITION AND MODUS OPERANDI

Composition

- Reinhard Altenhoener (Germany)
- Sanzio Bassini (Italy)
- Juan Bicarregui (UK)
- Manuel Delfino (Spain)
- Ole Henrik Ellestad (Norway)
- Daniel Gomes (Portugal)
- Keith Jeffery (UK)
- Leif Laaksonen (eIRG, Finland)
- Carlos Morais-Pires (European Commission)
- Jean Moulin (Belgium)
- Louise Perbal (Netherlands)
- Lorenza Saracco (European Commission)
- Magnus Stenbeck (Sweden)
- Edda Lilja Sveinsdottir (Iceland)
- Francoise Thibault (France)
- Dany Vandromme (Chair, France)

Modus operandi

Most of the work was made via electronic tools (e-mail, wiki, etc.). One face to face meeting was organised in Bruxelles on August 31st, 2007 to confront views and build consensus about the recommendations. Then the final touch was made to the paper during the week following the meeting.