e-IRG
e-Infrastructure Reflection Group

e-Infrastructure Reflection Group

"Long term preservation of Research Data (e-IRG document)"

***Dr. Juan Miguel González-Aranda** on behalf of Drs.*

Erik (Fledderus), Fotis (Karayannis), Françoise (Genova), Gabriele (von Voigt), Fernando (Aguilar), Jan (Wiebelitz), Jesús (Marco de Lucas), Naomi (Messing), Rosette (Vandenbroucke), Sverker (Holmgren), Trøels (Tvedegaard Rasmussen)

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA, INDUSTRIA Y COMPETITIVIDAD

E-IRG "Workshop on Long-term Sustainability for e-Infrastructures"
Session on Long-term Sustainability: Service & Data
Qawra, St. Paul's Bay – MALTA, June 8-9, 2017

DOCUMENT Table-of-Contents-TOC
Current version of this "LIVING" document:
draft long term preservation V2.3 2017-06-05

- **FOREWORD**
- **Management Summary –**
- **I. What is Long-term preservation**
- **II. What is available**
- **III. What is needed**
- **IV. Cost of long term preservation**
- **V. Possible solutions**
- **VI. Recommendations**

FOREWORD

- ✓ **Research data needs to be accessible and re-useable for a long of time or for "always" to enable multidisciplinary research, to enable new research on old data and to look for new science results when applying new technical computing possibilities on larger chunks of data.**

- ✓ **The increasing amount of research data that is generated at increasing rates never known before, provide another challenging issue with regard to the long-term storage, curation and preservation of data. The loss of data can have "just" economic impact if the data can be recovered through a re-run of experiments but in other cases data are lost forever and thus scientific insights or discoveries are at least postponed.**

- ✓ **Data curation and preservation of research data is a challenging task that needs strong support on the policy level but also advice for research institutions and funders. However many technical (e.g. data formats, metadata, data carrier), managerial (data ownership, curation) and financial (storage costs, conversion costs, curation costs) arise. There is not yet a clear picture how to deal with these issues.**

- ➢ **To this purpose, the efforts in the establishment of the European Open Science Cloud by following the FAIR principles, Research Data Alliance-RDA, OpenAir, EUDAT…among others, are essential.**

Management Summary (I)

- ✓ **These guidelines are intended to show a set of technical recommendations, methodologies and standards, providing technical details on the recommended practical implementation. The document addresses five main "themes" consisting of "guiding principles" that should be applied to guarantee the preservation, accessibility, and usability of research data in the long term:**

  - *State-of-art, what is long-term preservation*
  - *What is available*
  - *What is needed*
  - *Costs associated to long term preservation*
  - *Possible solutions and further recommendations*

- ✓ **The importance of long-term preservation of data is fast becoming one of the main concerns of large research initiatives (including associated Infrastructures). It goes beyond the data, and extends to their (meta-)data preservation and curation, and therefore, including the quality of research (meta-)data, as data are often accessible via metadata, and thus ensuring metadata quality is a means to provide long term accessibility.**

Management Summary (II)

- ✓ **For that purpose, preservation of data for long-term use will require data management strategies that include curation and preservation planning and implementation. While data management and curatorial activities have been an integral part of some scientific domains for years (see for example, astrophysics and high energy particle physics), these are new concepts in other areas of science (including inter-disciplinary such as Climate Change ones-related). Concepts such as provenance, representation for re-use, and work-flow capture are rarely understood, let alone addressed.**

- ✓ **Therefore, preservation of research data for long-term use requires careful planning, and would benefit from some new approaches, which are presented in the document.**

What is Long-term preservation

- ✓ **Long-term is defined as a period of time long enough for there to be concern about the loss of integrity of digital information held in repositories, including deterioration of storage media, changing technologies, support for old and new media and data formats (including standards), and a changing user community.**

- ✓ **This concept is also related to how to provide the proper mechanisms in order to guarantee sustainable format, defined as the ability to access an electronic record throughout its lifecycle, in spite of the technology used when it was originally created. A sustainable format is one that increases the likelihood of a record being accessible in the future.**

- ✓ **<u>IMPORTANT !!!:</u> In our e-IRG document, the Long-term preservation concept is more oriented towards the Open Science paradigm-area of action.**

What is available (I)

- ✓ This section is intended to show the on-going efforts from European Union in regards to long-term research data preservation from two complementary-synergetic perspectives :

  - ➢ The (recent) efforts from existing panEuropean & others related initiatives: **EOSC, GO-FAIR, OpenAIR, RDA (including GEDE RDA-Europe Working Group), OECD, EUDAT, LIBER Europe, etc., among others.**

  - ➢ **Member & Associate States (from now on MS & AS respectively) by themselves…**

    …through an "almost" in-detail review performed for 30 Countries: Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Norway, the Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, Turkey & United Kingdom

❖ **MS make a clear distinction between policies aimed at preservation and policies aimed at dissemination.**

What is available (II)

On the one hand, some of the MS have lately experienced a rapid development of e-Infrastructures aimed at preservation, curation, long-term preservation and increase in computing supporting capabilities for research. Examples of this are **Czech Republic**, **Estonia**, **Hungary**, **Poland**, **Slovakia** and **Spain**. These countries have considerably invested in e-Infrastructures for research and general use, either from national budgets, EU funds (ERDF, EU Framework Programmes project funds or others) and/or even from private investment.

On the other one, there are a group of smaller (or with fewer financing capabilities) MS which adopted a Gold open access model for publications which does not necessarily require the use of institutional repositories. This is the case, for example, for **Croatia**, **Cyprus**, **Latvia** and **Romania**.

In any case, for the majority of MS, institutional repositories are very well developed and pursue the goal of curation and preservation of scientific data, information and (tacit or explicit) knowledge, **although some National reports stress that many of these institutional repositories are not certified to properly guarantee the long-term preservation of research information.**

What is available (III)

A whole set of **technology platforms, aggregators and portals have been devised in MS with a view to harvesting, linking and guaranteeing inter-operability by providing a single-access point of all information on scientific research.**

However, **portals offering scientific information are usually harvesters and aggregators of meta-data, rather than repositories hosting and providing access to the research results themselves, especially in the case of research data.** In spite of this, during the last years it can be observed a tendency among the latest wave of EU enlargement countries which are focusing efforts on developing **centralized national repositories for preservation to be connected to the existing national systems and to be inter-operable across the EU with, for example, OpenAIRE protocols** in order to provide easy-to-use single access platforms which might be used both by public authorities for monitoring RDI.

On the other hand, other countries with an earlier and more developed tradition in digitization policies & strategies have recently increased investment in digital e-Infrastructures for research and policies on the creation and use of research e-infrastructures. This would be the case, for example, in **Denmark**, **Finland**, **France**, **Germany**, **Ireland**, **Italy**, **the Netherlands**, **Norway**, **Spain** and the **United Kingdom**. In most of these cases, MS have put forward specific programmes, strategies and objectives covering ICT development in the mid- to long-term period, in order to prepare their RDI activities for rapid technological advances in automated analysis, large-scale computing and the exponential growth of data.

However, most of their repositories do not always provide access to full-text articles or data, except for theses and dissertations, which can usually be accessed freely online.

What is needed

This section is based on answering the following key questions which arise when considering how to preserve digital (research or not) data for a long-time period:

- ✓ **Where were the data stored?;**
- ✓ **Is there a backup of the data off-site?;**
- ✓ **How to ensure the integrity of the data over time?;**
- ✓ **What ICT security features do organizations require-are needed for storing and accessing the data?;**
- ✓ **What metadata standard should be used to document the data?;**
- ✓ **What sustainable file formats should be used for long-term storage?;**
- ✓ **...**

In addition, and in order to ensure that a preserved data set remains fully useable and understandable in the future, additional information - beyond the instrument data and the metadata - needs to be preserved as well. This 'associated knowledge' can include e.g. information on the structure and semantics of the data sets, on processors, or calibration. This must be complemented by providing a mission-specific list of information and tools to be preserved along with the instrument data to ensure long-term usability, from the point of view of what is needed for different user communities, and in general terms, which are the common needs.

Cost of Long term preservation

The elaboration of this section is being complicated & complex. It is initially based on an on-going review of existing support bibliography, e.g.:

- **"The Economics of Long-Term Digital Storage" (Rosenthal et al.)**
- **"The cost of long-term retention and access to research data" (Addis et al.)**
- There is also a very interesting initiative relevant to this effort: **The RDA Data Fabric Interest Group (DF IG)**

In addition, in 2013 it was presented a 2020 vision for **long-term data preservation in High-Energy Physics-HEP** to the International Committee for Future Accelerators (ICFA):

- In 2016, it was presented a paper at iPRES on the PRODUCTION services that CERN offers for the above

- In 2017, the importance of Open Data and Open Science was stated at Davos; Data Preservation for HEP is also a Science Demonstrator in the European Open Science Cloud pilot (based on fully generic services equivalent to those offered at CERN). The corresponding paper: **"A 2020 vision for long-term data preservation in HEP to the International Committee for Future Accelerators (ICFA)"**

https://indico.cern.ch/event/377026/attachments/1131045/1616570/DPHEP_BLUETOO-July22.pdf

In any case, it is very difficult to get a graph consisting of which are cash flows, funded to grants or how it is performed.

Possible solutions

This section is focused on the process of identifying **"Examples of long-term Research Data Preservation"**, as existing solutions can be considered as guidelines and best practices.

As an initial support bibliography it was suggested: **"Long-Term Preservation of Earth Observation Space Data. Preservation Guidelines" CEOS/WGISS/DSIG/EODPG v1.0 15 September 2015** and **"The ESA Earth Observation Long Term Data Preservation (LTDP) Programme" (Beruti et al.)**

In addition, concrete rules for long-term preservation are needed. In relation to this, an initial support bibliography was suggested: **"USGS_Guidelines_for_the_Preservation_of_Digital_Scientific_Data_Final" (USGS)** & **"Principes and good practice for Preserving Data" (ICPSR).**

Moreover, as an interesting remark, it is also consider here other thematic areas-disciplines as well, not only *"hard"* sciences but also Humanities, Economy, etc. (clear relation with some ESFRIs dealing to this regard: CLARIN as a living set of languages of services and preservation policies; and DARIAH, among others.

Recommendations

It is clear that they will arise as new versions of the document are released being based on previous sections analysis. But some interesting preliminary ones are:

- ✓ **Not every data can be preserved:** It is necessary to understand which data could/should be preserved ("Data cemeteries" paradigm). Not only every data must be preserved and replicated ("Data reproducibility" paradigm), as some data doesn't have any scientific value. Where lies the ownership of data? Who owns the right of data ?...

- ✓ **Data re-usability:** This means that data have to be prepared properly: "Metadata robots", clarifying which type of robots specialized in the management of tapes that can restore some (critical or not) data, etc ...

- ✓ Recommended reading:
  https://www.beagrie.com/krds/

PART IV

CONCLUSIONS: EXPECTED IMPACT & NEXT STEPS

Expected IMPACT: **These guidelines are intended to show a set of technical recommendations, methodologies and standards, providing technical details on the recommended practical implementation by addressing five main "themes" consisting of "guiding principles" that should be applied to guarantee the preservation, accessibility, and usability of research data in the long term:**

- ✓ State-of-art, what is long-term preservation
- ✓ What is available
- ✓ What is needed
- ✓ Costs associated to long term preservation
- ✓ Possible solutions and further recommendations

e-Science Research Data initiatives & principles

- EOSC
- GO-FAIR
- OpenAIR
- RDA (incl. GEDE WG)
- OECD
- EUDAT
- LIBER Europe
- ...

Member & Associate States

e-IRG
e-Infrastructure Reflection Group

Research (e-)Infrastructures

ESFRI(CLARIN, DARIAH, etc.)
ERIC (LifeWatch, etc.)...

OPERATIONAL INTEROPERABILITY by providing the proper GUIDELINES containing technical recommendations, methodologies and standards FOR RESEARCH DATA LONG TERM PRESERVATION

SUGGESTED NEXT STEPS TO COMPLETE THE "LIVING" DOCUMENT

- **Consolidation of bilateral contacts with Member & Associate States.**

- Reinforcement of the 2 sub-working groups created dealing with sections:

  "4. Costs of long-term preservation" section, composed by Erik (Fledderus), Trøels (Tvedegaard Rasmussen), Fotis (Karayannis), Jesús (Marco de Lucas) and Fernando (Aguilar).

  "5. Possible solutions" OR "Examples of long-term Research Data Preservation" section, composed by Françoise (Genova) and Rosette (Vandenbroucke).

e-Infrastructure Reflection Group

Thank you very much !

¡ Muchas gracias !

Any questions ?

juanmiguel.gonzalez@mineco.es