

PaN-data ODI

Deliverable D7.4

D7.4: Report on evaluation of preservation mechanisms.

| | |
|------------------------------------|--|
| Grant Agreement Number | RI-283556 |
| Project Title | PaN-data Open Data Infrastructure |
| Title of Deliverable | Report on evaluation of preservation mechanisms. |
| Deliverable Number | D7.4 |
| Lead Beneficiary | STFC |
| Deliverable Dissemination Level | Public |
| Deliverable Nature | Report |
| Contractual Delivery Date | 30 September 2014 (Month 36) |
| Actual Delivery Date | 24 Oct 2014 |

The PaN-data ODI project is partly funded by the European Commission under the 7th Framework Programme, Information Society Technologies, Research Infrastructures.

Abstract

The present document goes through the different achievements of this work package and tries to evaluate the cost and benefits of experimental data preservation produced by analytical facilities.

After presenting the tools that have been necessary to set up data preservation, we try to evaluate the current outcome of this project and suggest a number of ideas for the facilities that are not currently involved in data preservation in order to help them to draw up their own processes.

Keyword list

Data preservation, data integrity, metadata, experimental logbook, Digital Object Identifiers.

Document approval

Approved for submission to EC by all partners on 17.20.2014

Revision history

| Issue | Author(s) | Date | Description |
|-------|------------------------------------|-------------------------------|-----------------------------------|
| 1.0 | Jamie Hall Jean-François Perrin | 16 th October 2014 | Initial release |
| 1.1 | Brian Matthews | 17 th October 2014 | Input on benefits of preservation |
| 1.2 | Jean-François Perrin | 24 th October 2014 | Integration of feedbacks |

Table of Contents

| | | |
|-------|---|----|
| 1 | Experimental raw data preservation..... | 4 |
| 2 | Mechanisms of data preservation..... | 4 |
| 2.1 | Tools..... | 4 |
| 2.1.1 | Contextual information | 4 |
| 2.1.2 | Archive..... | 6 |
| 2.1.3 | Data portal..... | 6 |
| 2.1.4 | Data Integrity | 6 |
| 2.1.5 | Persistent identifiers..... | 6 |
| 2.2 | cost estimation..... | 7 |
| 3 | Evaluation | 10 |
| 3.1 | Feedback from individual users | 10 |
| 3.2 | Uptake by the community | 11 |
| 3.3 | View from research infrastructures engaged in data preservation | 11 |
| 4 | Assessing benefits of preservation for facilities data..... | 12 |
| 4.1 | Introduction to Benefit framework..... | 12 |
| 4.2 | Utility | 13 |
| 4.2.1 | Desirability..... | 14 |
| 4.2.2 | Reusability | 15 |
| 4.3 | Substitutability | 17 |
| 4.4 | Reproducibility..... | 17 |
| 4.5 | Replaceability..... | 19 |
| 5 | Conclusion | 21 |

1 EXPERIMENTAL RAW DATA PRESERVATION

In the spirit of the work done in PaNData, experimental raw data preservation is strongly linked with the notion of provenance and open data. Preserving only data files make sense for the experimental team especially where the size of the data sets is important or becomes huge, the archives of the facilities represent a free, reliable and trust store. Nevertheless our aim is to go beyond this usage and to provide archives where every interested scientist (i.e. not only the ones that performed the experiment) could really use these data either to validate publications either to produce new analysis and new scientific outcomes even 5 or 10 years after.

That's why in this work we have tried to preserve not only data files but also all possible contextual information that is necessary and helpful to understand data. Whenever possible we tried to move from facility standards to open standards.

2 MECHANISMS OF DATA PRESERVATION

2.1 TOOLS

In the following sections we will try to follow the journey of data through the different tools that have been developed to ensure their preservation in view of making them shareable and usable by the community in the long run.

2.1.1 Contextual information

Proposal and more generally **administrative information** (abstract, sample name, formula, users ...) are registered during the proposal submission call. The information pertaining to successful proposals are ingested in the data catalogue alongside the other metadata of the experimental condition.

Proposal 7-05-388 Popout

Proposal **Links** **Members** **Data ranges** **Data folders**

Title
Diffusion of hydrogen on ZnO And Pd decorated ZnO surfaces: Is hydrogen spillover active at low Temperatures?

Abstract
The development of economical materials for use as adsorptive storage medium continues to be a major hurdle for meeting the US DOE requirements for acceptable storage and fuel cell systems. Two potential classes of materials for sorptive hydrogen storage are carbon(C) and metal oxide(MO) based where the adsorption can be broadly categorized as either: H₂-physisorption, based on weak Van der Waals interactions, and chemisorption, caused by H₂ dissociation. Numerous studies on C and MO based materials indicate that high adsorption capacity through H₂ physisorption is rather unlikely at near-ambient temperatures, required for practical applications. Significant enhancement of H₂ storage capacity can be achieved by doping/decorating the materials with small amounts of metals that act catalytically and create a spillover of the adsorbed molecules. Spillover refers to the transport of an active species adsorbed or formed on a first surface onto another surface that does not sorb or form the active species (under the same conditions). QENS will be used to characterize the onset and nature of translational diffusion of H₂ on pure or Pd decorated ZnO and MgO.

Planning
 ▶ Cycle 121 (from 19/04/2012 to 06/06/2012), instrument IN6
 ▶ Cycle 122 (from 28/08/2012 to 14/10/2012), instrument IN6
 ▶ Cycle 123 (from 24/10/2012 to 09/12/2012), instrument IN6

Sample(s)
 ▶ Pd Decorated ZnO and pure ZnO

Data

Figure 1 Administrative information presented in the catalogue

Experimental conditions, including **instrument control logs** and **experimental user logbooks** are preserved in a database and linked to the data catalogue







| Logs | | My experiments | | Logout perrin | |
|---|--|---|--|---|--|
| Acquiring data | | Now | | Period | |
| Cycle 2014-2 | | Instrument CT1 | | Proposal Internal use | |
| 2014-10-13 16:00:12 | | Count | | Actions | |
|  | | Numer 046885 Subtitle moniteur IN14 Acq. time 2.04 Monitor 1 2.035e+03 (9.998e+02) | | Detector 3.913e+07 (1.922e+07) Monitor 2 1.017e+03 (4.997e+02) | |
| 2014-10-13 16:00:14 | | Count | | | |
|  | | Numer 046886 Subtitle moniteur IN14 Acq. time 2.03 Monitor 1 2.033e+03 (9.997e+02) | | Detector 3.996e+07 (1.965e+07) Monitor 2 1.016e+03 (4.996e+02) | |
| 2014-10-13 16:00:16 | | Count | | | |
|  | | Numer 046887 Subtitle moniteur IN14 Acq. time 2.01 Monitor 1 2.008e+03 (9.997e+02) | | Detector 4.079e+07 (2.031e+07) Monitor 2 1.004e+03 (4.999e+02) | |
| 2014-10-13 16:00:18 | | Count | | | |
|  | | Numer 046888 Subtitle moniteur IN14 Acq. time 2.00 Monitor 1 2.000e+03 (1.000e+03) | | Detector 4.163e+07 (2.081e+07) Monitor 2 1.000e+03 (5.000e+02) | |
| 2014-10-13 16:00:20 | | Count | | | |
|  | | Numer 046889 Subtitle moniteur IN14 Acq. time 2.04 Monitor 1 2.039e+03 (9.996e+02) | | Detector 4.246e+07 (2.081e+07) Monitor 2 1.019e+03 (4.995e+02) | |
| 2014-10-13 16:00:22 | | Count | | | |
|  | | Numer 046890 Subtitle moniteur IN14 Acq. time 1.64 Monitor 1 1.640e+03 (9.999e+02) | | Detector 4.329e+07 (2.639e+07) Monitor 2 8.200e+02 (4.999e+02) | |

Figure 2 Experimental logbook - acquisition.

| Logs | | My experiments | |
|---------------------|--|--|--|
| Sample environment | | Now | |
| Period | | | |
| Cycle 2014-2 | | Instrument CT1 | |
| Proposal | | Internal use | |
| 2014-10-09 14:00:31 | | OrangeCryostat --> 30 K setpoint ramp 0.1 K every 1 s (timed out = 20 m) | |
| 2014-10-09 14:00:31 | | OrangeCryostat --> waiting stabilisation time 300 s | |
| 2014-10-09 14:00:54 | | OrangeCryostat stopped | |
| 2014-10-09 14:00:55 | | OrangeCryostat 0 K | |
| 2014-10-09 14:01:00 | | OrangeCryostat --> 30 K (timed out = 20 m) | |
| 2014-10-09 14:01:25 | | OrangeCryostat --> waiting stabilisation time 300 s | |
| 2014-10-09 14:06:25 | | OrangeCryostat 30 K | |
| 2014-10-13 13:13:04 | | Furnace --> 200 C setpoint ramp 1 C every 20 s | |
| 2014-10-13 13:13:14 | | Furnace stopped | |
| 2014-10-13 13:13:14 | | Furnace 1 C | |

Figure 3 Experimental logbook - temperature control

2.1.2 Archive

The Instrument control process and data storage have also been developed. Even if the work done is not directly part of this project (i.e. it has been financed by the facilities) these components also had to evolve in order to permit a better identification of the data files (i.e. **identify the proposal that leads to these data**) and enforce data policies through **access control** for the duration of the non-disclosure period. Security in the archive is important in order to establish trust with the users.

2.1.3 Data portal

Data portals are the visible part of the archive and represent the main interface with the community, they have been modified in order to introduce the display of the file checksums but also links with DOI and logbook.

2.1.4 Data Integrity

Data integrity have been set in place through the use of file checksums. They are computed as soon as possible in the experimental workflow and stored in the manifest files of the archive as well as in the catalogues alongside the other metadata. Finally, they are presented on data portals in order to allow users to perform data integrity verifications on their own once they have downloaded the data.

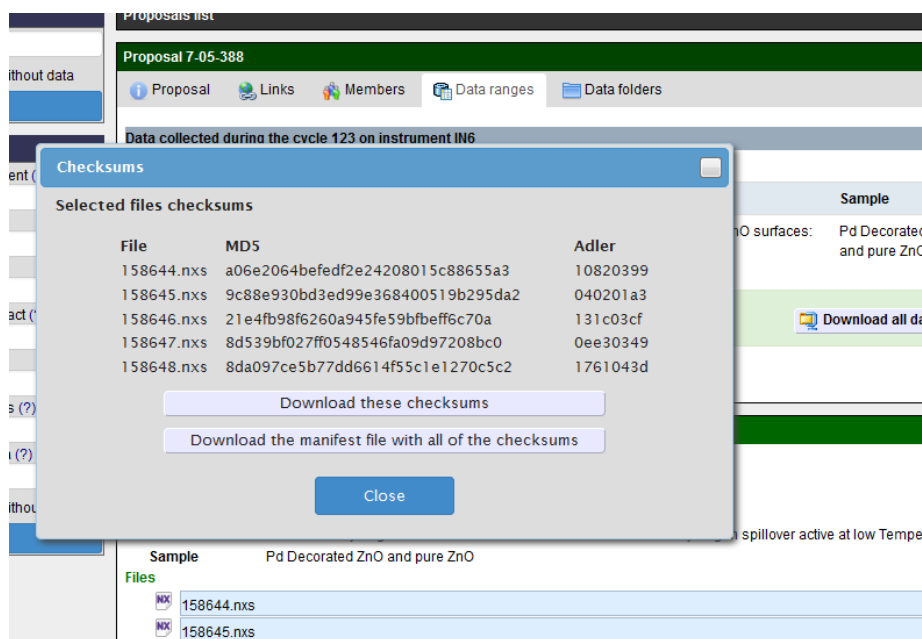


Figure 4 Data file checksums

2.1.5 Persistent identifiers

Persistent identifiers have been set in place in the form of Digital Object Identifiers, they are generated automatically using the DataCite web service APIs at the end of the first experiment.

Currently a single DOI is generated per proposal, it might change in the future in order to be able to identify individual datasets.

NEUTRONS FOR SCIENCE

DOI: 10.5291/ILL-4-01-1216

Please note
The full details of this dataset is not yet available to the public as it still under its embargo period. As such there are only a few details publically exposed. To find out more about how the ILL governs the release of data, please go [here](#). Thank you for your understanding.

Title
Doping-induced energy gap in a quantum spin chain

Abstract
This data is not yet public

Download
This data is not currently available to download

Data citation
The recommended format for citing this dataset in a research publication is in the following format:
Buechner, Bernd; Gvasaliya, Severian; Hess, Christian; Ivanov, Alexandre; Mansson, Martin; Mohan, Ashwin; Piovano, Andrea; Simutis, Gediminas and Zheludev, Andrey. (2012). Doping-induced energy gap in a quantum spin chain. Institut Laue-Langevin (ILL) [doi:10.5291/ILL-DATA-4-01-1216](#)

Instrument

Metadata
DOI: doi:10.5291/ILL-DATA-4-01-1216
Authors: BUECHNER, Bernd, GVASALIYA, Severian, HESS, Christian, IVANOV, Alexandre, MANSSON, MARTIN, MOHAN Ashwin, PIOVANO, Andrea, SIMUTIS, Gediminas, ZHELUDEV, Andrey
Publisher: Institut Laue-Langevin
Publication year: 2013
Cycle(s): 20123
Proposed:

DOIs represent a tool necessary for linking the publications to the data but they can also help to discover data through the use of DOI metadata standardisation and the external catalogues, such as the Thomson-Reuters Data Citation Index, that use OAI-PMH services to harvest them. In our case the OAI-PMH service is provided by DataCite.

Figure 5 DOI landing page

2.2 COST ESTIMATION

Currently the main cost element of data preservation is the storage system, it could easily represents the main part of the IT budget of a facility. Figure 6 represents the storage cost evolution of a standard analytical facility: the hard drive storage cost of classical infrastructure, expressed in net capacity with double disk failures protection, and three years warranty. Since 2008, the cost per TB of hard drive storage has stopped decreasing by the large factors we were used to some years ago. This is not a real problem for the data sets that represents small volumes (typically some dozens of GB) but for the largest one (some dozens of TB) this is a real concern.

The volume of the experimental data produced is also increasing very quickly due to improvements in detectors and the trend in the neutron scattering community to requests more and more events mode acquisition instead of the traditional histogram mode. This has been a clear concern for the synchrotron facilities which has also reached the neutron facilities recently, as presented by Figure 7 and Figure 8.

For the largest datasets, over 10 to 15 years, **storage cost could represent up to 10% of the total cost of the experiments.** This is a rough estimation based on ILL figures, where the largest dataset (35TB) has been produced during a 50 days run. The total cost of a day of experiment at ILL is estimated at 12K€, the cost of this experiment is therefore 600K€. The hardware cost for storing these 35TB over 15 years where we have to change 3 times the equipment (after 5 years most of the hardware storage components have to be replaced) over this period is approximately 48K€ what represents 8% of the cost of the experiment. This is only for the hardware part adding the man power we are very close to the 10% estimation. Less expensive solutions (e.g.; tapes, hierarchical storage) exist, but they also require an important investment and necessitate more data movement and management during the lifecycle of the data in order to cope with the technology changes and fit performance needs.

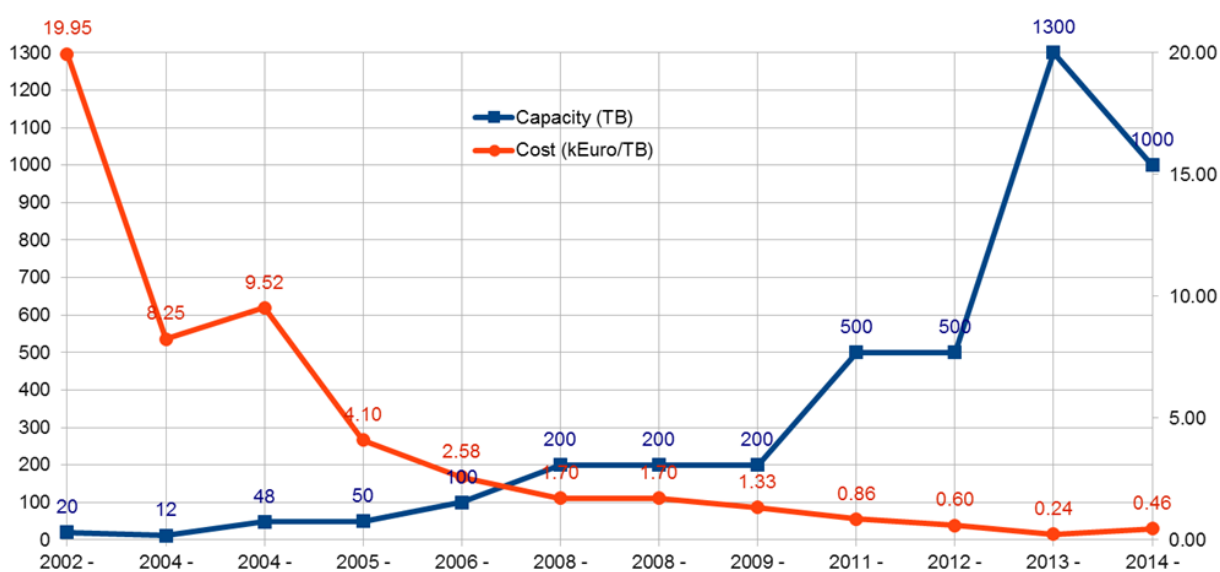


Figure 6 Storage cost evolution - Courtesy of B. Lebayle ESRF

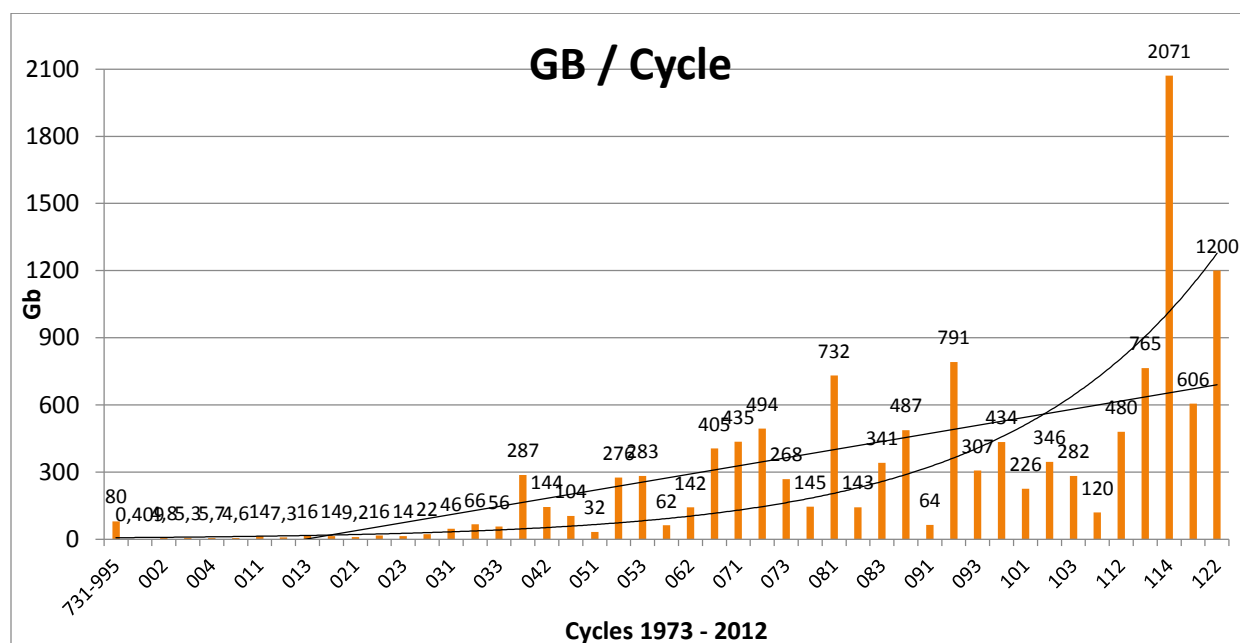


Figure 7 ILL experimental data evolution 1973 to 2012

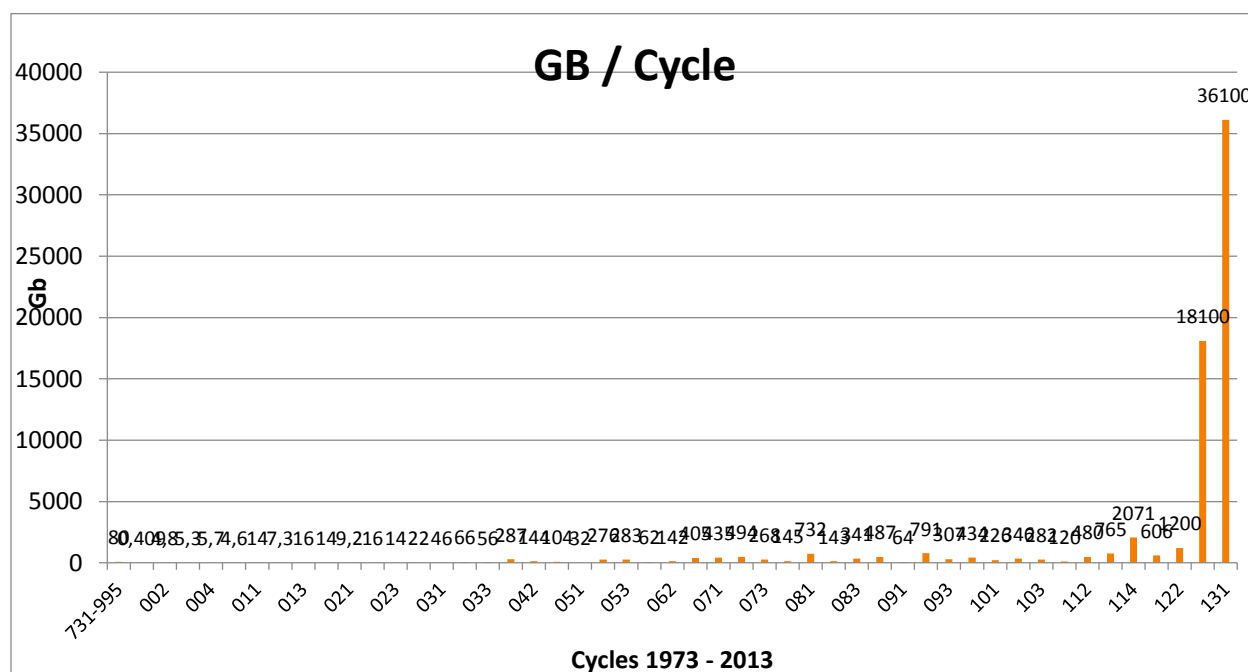


Figure 8 ILL experimental data evolution 1973 to 2013

Setting up mechanisms like metadata collection, data integrity, persistent identifiers attribution also represents a cost, especially for the facilities where it is important to spend time with the scientists in order to introduce the necessary underlying changes concerning user authentication and structuring of the archives. Nevertheless based on our experience and the fact that those processes have been automated and integrated into other operational systems such as proposal

management system or the experimental log management, we don't expect an important cost solely due to the maintenance of the preservation systems excluding storage cost.

Efforts on dissemination and uptake are probably also to be taken in consideration, but this only concerns early adopters and won't be significant when preservation and data sharing will become the standard of science.

3 EVALUATION

The following sections represent a very early evaluation more based on the tools' usage and feedback for individual users than the real uptake by the community.

3.1 FEEDBACK FROM INDIVIDUAL USERS

So far we didn't receive negative feedbacks. The real difficulty was the introduction of data policies but once published and after a period of scepticism, due to lack of tools, the attitude towards data preservation in view of opening the data after publication is generally positive and constructive.

The tools proposed to the facility users are seen as necessary, they even represent real improvements for the users' life even outside the strict scope of preservation. For example the fact that with the web log application, available from Internet, people who are not part of the experimental team (but part of the proposal or simply authorized by the team) can follow remotely the course of the experiment and comment on it easily, is seen as an important improvement.

The fact that data and metadata are properly organised and stored alongside, available on data portals, proposed in standard format are major improvements even for the facility users. They can come back to their experiment some years later and re-do analysis without having to worry where their paper notebooks are or even in some cases their USB disks.

Strict confidentiality and IT security enforcement, during the non-disclosure period, is also seen as an important element for scientists to ensure the preservation of valuable information. We have collected a number of scientist statements declaring that before these security measures they were trying to hide their work (by not declaring the exact formula of the samples or by trying to destroy data files, etc.) when the experiments were concerning hot topics.

DOIs also raised interest, especially the principle of citing data: the main use case mentioned is when scientists are reading publications and want to get access to the data easily in order to do their own analysis.

We have also recently received requests for new functionalities, which is a very positive sign of the attractiveness that represents the data preservation for the scientists. The recurrent ones are listed hereafter:

- “We should have the possibility to cite, through DOI, only a subpart of the dataset, this is important when a set of samples has been used for the experiment and the publication referred to a single one.”
- “Concerning the web logbook, it will be perfect if we could have a view where each user could select the elements he wants to be displayed, this is really user and instrument dependant.”
- “Concerning the web logbook, it allows to monitor some valuable parameters like cell temperature, pressure, etc. It would be nice that this tool could graph those parameters and preserve these graphs.”
- “Could we improve this logbook to be able to use it also during analysis and not only during the acquisition?”

Those requirements are precious, the development will take place as soon as possible in order to keep the momentum with the users.

3.2 UPTAKE BY THE COMMUNITY

Only two facilities (ILL and ISIS) have adopted an open data policy and preserve the data over a sufficient period of time for permitting different analysis than the one performed by the experimental team. In both cases these data policies were adopted in 2011, the first experiment under these new policies were performed in 2012 and the resulting data are still under the 3 years embargo period.

Since the mean time between experiments and publications in the community is 3 years the actual impact is very difficult to assess. This is difficult to say that a general uptake by the community exists today, but as mentioned earlier individual uptake exists and should soon lead to the community one.

3.3 VIEW FROM RESEARCH INFRASTRUCTURES ENGAGED IN DATA PRESERVATION

The cultural and technical changes introduced by the preservation and open data concepts have represented a significant effort for the facilities that have adopted them. The expectation is that this project will be useful for the whole community and should lead to more peer reviewed publications, and ultimately lead to better impact of the Research Infrastructure as a whole.

The link between peer reviewed publications and data is also extremely interesting for the internal metrics and governance of the facilities. For instance, they could be very helpful in making decisions about the construction of new instruments.

The scientists who produce publications using the experimental data, sometimes forget to list as author the facility staff scientists who help during the preparation, the acquisition or the analysis of the data and even sometimes forget to mention explicitly the name of the facility in their scientific texts. This lack of reward is a real issue for facility staff scientist careers especially for the youngest ones being in short term contracts who have no time to conduct their own research.

Expectations are strong that when the uptake will be large, publishers and funders will request explicit data citation in the papers and even make it mandatory.

This project has been generally been welcome by the users, the first signs are positive but are not yet visible in the usual facility metrics (i.e. number of peer reviewed publications per year and facility). In order to foster this uptake we currently discuss the possibility to enforce the citation of DOIs for data in publications by taking this practice as an element of the review of proposal (give a bonus to the users that have cited data DOIs in their publications).

4 ASSESSING BENEFITS OF PRESERVATION FOR FACILITIES DATA

In PanData-Europe Deliverable 7.1¹, a survey of facilities was undertaken which included a consideration of the role of preservation in facilities. This concluded that while data storage and management were of major concern to facilities, doubts were given about the long term preservation of data. This was driven by the costs involved in storing and making accessible data for the long-term, especially at the scales of data as generated by synchrotron sources. But also, the benefits of preservation were not clear; what is the real value of keeping data for the long-term in terms of supporting and generating new science.

Some of the work of PanData-ODI has been addressing some of the barriers identified in the survey. Data publishing with DOIs has been proposed and implemented in some facilities. There has been a consideration into the mechanisms of integrity checking and also capturing provenance and contextual information. And some facilities have put in place data policies including retention and disposal. However, the value of keeping data still remains an issue.

In this section, we give an approach to assessing the value of data which could be applied to the scenario of facilities data in the form of a benefits framework. This framework could be applied in conjunction with the more established cost evaluation technique to give a likely case for preserving facilities data.

4.1 INTRODUCTION TO BENEFIT FRAMEWORK

The Keeping Research Data Safe (KRDS) model of benefits discussed in the guide to the framework² defines three dimensions: outcomes, timescales and beneficiaries as a framework to evaluate the benefit of a data product. Outcomes are then divided into:

- Direct benefits: positive impacts obtained in a data curation activity.

¹ <http://wiki.pan-data.eu/images/GHD/8/84/PaN-data-D7-1.pdf>

² Charles Beagrie Ltd. Guide to the KRDS Benefits Framework. Keeping Research Data Safe. report, v3. July 2011. http://www.beagrie.com/KRDS_BenefitsFramework_Guidev3_July%202011.pdf

- Indirect benefits: negative impact avoided by investing in a data curation activity.

The guide to the benefits framework then goes on to discuss how this framework might apply in particular instances. This gives particular instances of outcomes which might apply; however, these are rather ad hoc lists of potential outcomes.

In this section, we propose a more systematic characterisation of the *outcomes* which could be applied to a data product within a research data scenario such as a facility's data. This approach can then be combined with the rest of the KRDS approach to provide a more detailed analysis of the potential benefits accruing from the preservation of a data product.

This approach can also be compared with that of Whyte and Wilson³ who identify seven general criteria for retention (*Relevance to Mission; Scientific or Historical Value; Uniqueness; Potential for Redistribution; Non-Replicability; Economic Case; Full Documentation*). Again, while these are useful, they are not comprehensive, and do not in general capture the intentionality behind the criteria which may lead data archivists to identify additional benefits not covered within these definitions, or provide measurable criteria.

We analyse the benefits by considering two main categories of benefits: Utility and Substitutability. These categories approximately correspond to KRDS's direct and indirect benefits.

A third category could be seen as legal compliance; sometimes data needs to be preserved due to either legal necessity (e.g. accounting records, pharmaceutical trial data or engineering manuals), or through the mandate of an organisation (e.g. the purpose of a data archive). Thus the motivation of not breaking legal liability becomes a strong incentive. However, this does not cover the *intention* behind the legal compliance (e.g. revisiting or rechecking trial data), which is a better measure of the benefit for legal compliance and the purpose behind the legal enforcement.

4.2 UTILITY

How useful is the data likely to be in the future?

Utility refers to the value of the data for re-examination and reuse in the future. Thus if the Utility of the data is high, then the benefit of the data is high.

We further sub-divide Utility into two categories:

³ Angus Whyte and Andrew Wilson. Appraise & Select Research Data for Curation. Digital Curation Centre and Australian National Data Service "working level" guide, 25 October 2010. <http://www.dcc.ac.uk/resources/how-guides/appraise-select-data>

4.2.1 Desirability

How much is the data re-examined and reused?

Clearly the data is more valuable if the data is requested, re-examined and reused in the future, especially in new contexts and new situations.

In *Table 1* we give some instances of the types of evidence for the desirability of data, together with some guidelines on metrics which might be used to measure such evidence. Often such metrics are subjective and difficult to measure, especially for a long time in the future.

| Evidence | Description | Metric |
|-------------------------------------|--|---|
| Data requests | Number of requests for the data arising from the user community. | Number of user requests. This can be also seen as a percentage of the funding which is supporting the user community (e.g. future research grants). |
| Data Citations | Citations of the data within refereed published literature. | Number of citations to data (or a reference paper for the data), weighted by the impact factors of the papers. |
| Research grants | Future research grants which cite or request access to the data. This is evidence that the data remains relevant in an active research area. | Percentage of the value of research grant. |
| Requests for instrument time | Future requests for instrument time which cite a previous experiment and in particular the data generated by it. | Percentage of the value of instrument time. |
| Commercial data access | Sales of access to the data or added value products using the data. | Value of sales of the data or derived products |
| Patents | Use of the data leads to commercial patents. | Number of patents (and an indication of their value e.g. use in products) |

| | | |
|-------------------------------------|--|--|
| Products | Use of the data leads to commercial products. | Value of sales of products |
| Influencing decisions makers | Use of the data by government or other agency to either: <ul style="list-style-type: none"> - influence policy (e.g. included in government) - directly influence action | Citation of data in policy documents. Estimate of value of policy or action. |
| Used for academic assessment | Use of the data in submissions to research evaluation exercises. | Citation of data in research evaluation. |

Table 1: Data Desirability Metrics

For these metrics, the other dimensions of the KRDS framework (timescale and beneficiaries) also need to be taken into account; thus for each instance of the evidential criteria, we can determine and monitor:

1. The number of instances in a desired time period (e.g. over a period of funding of the archive, or the projected useful lifespan of the data), and whether the number of instances is increasing or decreasing.
2. Which key stakeholder groups are involved in each instance (e.g. researchers, public policy makers, general public etc).

Many of these metrics are captured are similar to those captured in a Facilities impact assessment exercise; these should be extended to cover use of data.

4.2.2 Reusability

What characteristics does the data have which can make it more reusable?

Data may have more beneficial impact if it is presented in a manner which encourages re-examination and reuse; if it is easier to comprehend and integrate with other data and computing systems, it is likely to be reused, and thus have a higher utility. Reusability factors which would encourage reuse might include:

| Evidence | Description | Metric |
|--|---|---|
| Standard data formats | The data uses well-documented and widely used data formats in a consistent manner. | Validation of conformance to data standards |
| Metadata quality | The data is provided with comprehensive and accurate metadata describing its provenance and usage. | Validation of conformance to metadata standards. |
| Persistent Identifiers | The data is assigned with persistent identifiers for consistent and trusted reference to the data. | Use of standard identification schemes, e.g. DOIs. |
| Integrity and format validation | The data has periodic actions to check its quality, such as integrity and format validations. | Use of integrity and validation checks. |
| Availability | The data is available within in clearly stated rules for access (e.g. open-access, user registration, licencing, and price as appropriate). | Access policy clearly stated, and consistently applied. |
| Data quality | The data has quality metrics available for its accuracy, error range, coverage etc. | Quality audit on the data available. |
| Preservation planning and actions | The data is supported by a process which maintains the reusability for the designated community over time. | Preservation certification |

Table 2: Data Reusability Metrics

For these metrics, the timescale and beneficiaries aspects are manifest in that the data should be presented with formats and metadata *suitable for use by the likely stakeholders (in OAIS, the “designated community”)*. Thus it should be presented in the terms the envisage users are likely to be able to understand and reuse. Also, to maintain reusability, these standards need to be maintained as the knowledge of the designated community changes, e.g. as formats or metadata standards change; thus we have expectation of reusability over particular timescales.

Many of these factors are within the scope of a “well-operated” data management process, which is typically the case within facilities science, where there are dedicated resources to maintaining data. PanData has been developing the use of data format standards (e.g. HDF5, NeXuS), metadata formats (e.g. CSMD), persistent identifiers (e.g. DOIs) and other aspects of data reusability.

4.3 SUBSTITUTABILITY

Can an acceptable substitute be found for the data?

Substitutability factors are those which assess whether an alternative data set of an acceptable quality which can be used in place of the data can be accessed if it is needed, if the archive’s copy is not available. If a reasonable substitute can be accessed at a reasonable cost (for example at a lower cost than preserving the data), then the benefit of keeping a copy of the data within the archive is likely to be lower.

Substitutability factors are often overlooked in research and development work on digital preservation, which often emphasises that there are disciplines where there either no substitute or an adequate substitute is hard to identify, such as environmental and other observational sciences, where an observation at a point in time is a unique event and cannot be recapitulated. However, in Facilities science, where experiments can in theory at least be repeated, substitutability is a serious consideration.

Again we subdivide Substitutability into two subfactors: reproducibility and replacability.

4.4 REPRODUCIBILITY

Can the data collection be re-enacted to regenerate the data?

If the data can be re-collected effectively, then that may be more efficient or simpler approach than that of keeping the data. In the case of observation, that is a measurement of a phenomena at some particular location and time, it is not typically possible to reproduce data. However, for experiments, that is a measurement of natural phenomena on a sample (e.g. a geological sample, a chemical compound analysis, an atmospheric chemistry analysis) or of conditions within an experimental apparatus. These are reproducible, given the appropriate apparatus and sample. However, this may be at significant cost. This is typically the case within a Neutron or Synchrotron facility.

In this case, we also need to take into account other factors in evaluating the cost of repeating the experiment include:

| Cost | Description | Metric |
|------|-------------|--------|
|------|-------------|--------|

| | | |
|--|---|---|
| Sample acquisition | <p>The cost of re-acquiring the sample may be high, including:</p> <ul style="list-style-type: none"> - the costs of collecting a rare geological or biological sample at a remote or difficult to get to location (e.g. meteorites, samples from the deep ocean); - the cost of the raw materials of a chemical compound (e.g. precious or rare elements), - the difficulty of chemical synthesis (e.g. difficulty in protein crystallization). | Cost of sample acquisition |
| Sample handling | Cost of handling the sample needs to be taken into account, as it may require specific conditions (e.g. cryogenic material at low temperature), or safety precautions (e.g. radioactive, bioactive or toxic material). | Cost of sample handling |
| Experimental running costs | Synchrotron or Neutron sources are highly expensive and specialized facilities. The use of the facility to set up and run should be costed. | Cost of setting up and running experiment at facility |
| Experimenter expertise | The experiment at the facility will require highly specialised expertise to be effectively carried out, both from the facility and the experimental team. | Cost of paying for highly skilled experimental staff. |
| Data quality | The new experiment may result in better data than originally collected. | Relative data quality of new to old data |
| Data Analysis Software and hardware platform. | The data analysis software required to process the experimental data may be costly to reacquire, configure and run on high performance resources. Conversely, more | Cost of rerunning software and relative better quality of resulting data set. |

| | | |
|--|---|--|
| | sophisticated software and hardware platform may result in better quality data. | |
|--|---|--|

Table 3: Data Reproducibility Cost Metrics

The timescale factors for these metrics are such that certain costs may vary over time. The cost of raw material for chemical synthesis may change (e.g. rare or precious elements may become more expensive), whilst expertise and experimental technique at the facility may become more “standard”; many facilities offer standardised or “express” services for areas such as crystallography for frequently used data collection processes which have a common configuration. Similarly, the cost of a similar *quality* result may decrease as laboratory equipment becomes more standardised.

Further, as techniques and technology improve, instruments become more accurate and sensitive, and software becomes more powerful, so a better quality result may be available by rerunning the experiment on new instruments (thus reducing the cost of reproducibility when scaled to the same level of quality).

In the case of facilities, as new facilities and instruments are developed and become available, the power and resolution of the resulting data may improve to such an extent to negate the value of previously collected data, and becomes likely that rerunning the experiment will get a better result than re-examining the old data; however, this assumes that the cost other factors (such as synthesising the sample) do not outweigh the advantages.

4.5 REPLACEABILITY

Can the data be satisfactorily replaced by other existing data?

The data set can be seen as being replaceable if there is data held elsewhere which can form an effective substitute for the data. The alternate dataset may be a copy of the dataset, or it may be other data which can be used within analysis scenarios as an adequate substitute, perhaps with some re-processing (e.g. data collected on another instrument from which the similar information can be deduced). Clearly an adequate substitution is likely to depend on the context in which it is used, and it may in general be difficult to determine whether a data set can be adequately substituted – data archives are likely to err on the cautious side and retain data.

Some factors in evaluating whether an alternate data set is a suitable substitute, thus reducing the benefit of keeping a copy of the data set, would include the following.

| Cost | Description | Metric |
|------|-------------|--------|
|------|-------------|--------|

| | | |
|---------------------------|--|--|
| Data Quality | What is the data quality and coverage of the alternate data set compared with the subject data set? | Comparative evaluation of the data quality of the alternate data set |
| Data management | Is the alternate data set (e.g. the copy) kept to as high a standard of data management (e.g. remit, staffing, resources, certification) as the subject data set? | Comparative evaluation of the relative data management standards of alternate to the subject data. |
| Data Accessibility | Is the alternate data set as accessible as the subject data set? E.g., does the alternate data have copyright, pricing, security, or other restrictions (e.g. restricted to access from particular geographical areas) which make it less accessible than the subject data set to the designated community of the subject data archive? It may be more open if the access limitations are less strict. | Comparative evaluation of the relative data access conditions of the alternate to subject data. |
| Data Format | Is the alternate data set in a data format which is useable by the designated community of the subject data archive? | Comparative evaluation of the relative data access of alternate to the subject data |
| Metadata Format | Is the alternate data set in with sufficient metadata which is useable by the designated community of the subject data archive? This would include the natural language of the metadata; it needs to be readable by the designated community | Comparative evaluation of the metadata of the alternate to the subject data |

Table 4 Data Replacability Cost Metrics

The timescale factors for this depend on the confidence of the continued availability of the alternate data source for the envisage lifetime of the data, and whether there is confidence that the alternate data source will continue to maintain its current conditions (for example, will not begin to charge for access to the data, or reduce level of funding and thus quality of data management).

5 CONCLUSION

We are absolutely convinced that this work is useful and necessary for modern science. Electronic tools have been developed and opened to users during this project. These tools are really in use and not only permit to preserve research data with their contextual information but often have also improved the day to day workflow of facility users.

Today tools are in place for preserving all elements of experimental raw data, only samples preparation which often takes place at users' home organization is not covered. This is not always necessary for understanding experimental data but its preservation would constitute a consistent achievement. We will try to address it in the near future. Preserving raw data reduction and analysis is also of high interest especially in the context of the strong increase of data volume often referred as "data tsunami".

We are really at the beginning of this new era, only initial evaluation has taken place and the real reward for the facility did not yet happened, nevertheless we are confident that open data and preservation represents the future of analytical facilities. ILL and ISIS in this project have acted as pioneer, in section 4 our suggestions have been presented that can help other facilities in their decision process regarding data preservation.