PaN-data Europe

Deliverable D7.1

Report on survey of publication repositories, cross-linking and long-term preservation

Grant Agreement Number	261537
Project Title	PaN-data Europe Strategic Working Group
Title of Deliverable	Report on survey of publication repositories, cross-linking and long-term preservation
Deliverable Number	D7.1
Lead Beneficiary	STFC
Deliverable Dissemination Level	Public
Deliverable Nature	Report
Contractual Delivery Date	31 May 2011 (Month 12)
Actual Delivery Date	30 Jul 2011

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

Abstract

This report presents the results of a survey of the current status and views on extending support for the research lifecycle within facilities, in particular, cataloguing and linking data and publications, recording provenance of derived data and digital preservation.

Keyword list

Metadata, Data catalogues, Publications, Research lifecycle, Provenance, Derived data, Persistent identifiers, Data reuse, Data storage, Linking, Preservation.

Document approval

Approved for submission to EC by all partners (July 2011).

Revision history

Issue	Author(s)	Date	Description
0.1	Brian Matthews, Simon Lambert, Juan Bicarregui	1 April 2011	Survey design
0.2	Brian Matthews	31 May 2011	Collation of survey results
0.3	Brian Matthews	28 July 2011	Final draft
1.0	Brian Matthews, Simon Lambert	30 July 2011	Final version
1.1	Simon Lambert	1 Sep 2011	Corrected error in cross-references

Table of contents

Page

1 Introduction

The aim of the integration work package is to foster the integration of the whole lifecycle of the science supported by the facilities participating in PaN-data. This would focus on such issues as cataloguing and publishing of the experiments carried out at facilities, with the resulting raw data; tracking publications arising from work at facilities and linking them to experiments; the interaction between institutional repositories of publications; recording and storing the provenance trail resulting from the subsequent analysis of the results, including the software used; and the packaging of the data and other research outputs for long-term preservation. We would further consider the services required for search and reuse of this provenance and preservation information.

1.1 Aim of the survey

The first task in this workpackage is to: review existing provision for publication repositories, cita*tion recording and long-term preservation in use across the facilities and in the user community, including facility libraries*.

The aim of this survey is to:

- Gather information on the current state of support for the aspects of the research lifecycle considered within this work package: experiment and publication cataloguing and tracking, provenance tracking, and preservation issues.
- Evaluate the awareness and view of the facilities to supporting these issues.
- Capture opportunities and costs associated with providing further sources
- Identify areas to develop immediate and longer-term support for these aspects within facilities.

Consequently, we have undertaken a survey if all PaN-data partners to establish the current status in facilities of support capture the requirements and perceived benefits in these areas.for:

- Publications
  - Why and how facilities track publications resulting from their experiments?
  - Do they support an institutional repository, or database of publications?
  - How do they collect data on publications?
  - Who carries out this work?
  - Do they systematically report the relationship of publications to experiments?
- Recording raw and derived data
  - Do they catalogue the data
  - Store derived data?
  - Record software

- Preservation
  - Retention policies
  - Persistent Identifiers (DOIs)

The questionnaire has some 36 questions, plus a free form question. This was circulated to facilities in April 2011, and results collated in May 2011. All partners responded, although Alba responded that as facilities which is not yet in operation, they felt that they were not in a position to respond to the questions, as there was no practice to report. Consequently, they are omitted in the following results.

2 Commentary

The responses to the questionnaire are reported in section 3. For clarity, we summarise some comments and recommendations in this section. For each topic area, there are questions on current status and future plans; in this summary, we cover both status and plans within the same discussion.

2.1 Publications

The first topic area of the questionnaire addressed how facilities are currently managing publications, both those of their own staff and also those arising from the use of facilities produced by the user community.

2.1.1 Current situation: commonalities across facilities

Recording the published output of staff and particularly users is a key common feature of all PaN-data partners as an important metric of the quantity and quality scientific work carried out at the facility. As consequence we note:

- All facilities require their own staff to record publications
- All facilities require their user communities to inform them of resultant publication.
- As a consequence, they relate the recording of publication to new proposals and will withhold further beamtime from users who do not inform them of previous publication.
- Almost all facilities provide some form of database or repository to record these publications for reporting purposes.
- Almost all facilities make an effort to harvest external services, in particular ISI Web of Knowledge, to discover peer review publications related to their facility. This presumably is to augment and to ease the addition of publications by the users themselves.

2.1.2 Current situation: differences between facilities

There are no common software or tools for recording publications, and each facility has largely adopted their own ad hoc system sufficient for their own purposes. This is a satisfactory approach as far as it goes; the main goal is reporting of published output to measure the value of the facility, rather than a preservation system for publications (these are assumed to be stored by publishers or other library repositories), or a dissemination mechanism (this is largely assumed to be the users' responsibility). However, there may be benefit in common formats.

There are wide variations in the estimates of how successful the publication tracking is, and this is a hard to evaluate.

A variety of different actors are responsible for the recording of publications, including user office, library and web team.

2.1.3 Opportunities and drivers

Clearly, there are strong motivations and support provided to support the recording of publications of staff and users. Three of the main drivers for providing publication repositories (in general) could be characterized as:

1. **Publication recording ,** i.e. a repository of metadata that can be browsed, searched and exported for reporting on the performance of the institution
2. **Publication dissemination** i.e. an open access repository of metadata and content that can be browsed, searched and exported by external researchers, thus disseminating the work of the institution and its users and increasing citation count and credit, and stimulating further research.
3. **Publication tracking**, i.e. maintaining a record of all institution associated published outputs, with all items uniquely identified and records linked with other outputs, (data, derived data, software) for provenance tracking, validation and reuse.

The first of these is the main driver for facilities to record publications; to determine the quantity and quality of the science arising from the facility (*is good science being carried out at the facility*?). This also gives a strong incentive to partiality respond to the third driver; there is a need to trace a publication to the user and user's experiment, to determine value from a particular allocation of beamtime, so that the allocation process can be evaluated (*are the right applications for beamtime being accepted ?*). That such information should be provided by users is in the beamtime application process of a number of facilities, and is proposed to be mandated within the PaN-data data policy framework. However, it appears that is it hard to assess how well this information is being collected, there are some quite labour intensive efforts to supplement this information already (e.g. by harvesting external services and thus there would be a major incentive to try to improve this process.

The second and third (partially) are less strong motivations for facilities. In both these cases, the benefit largely arises for the science user community, and it could reasonably be said that these are best served elsewhere (for example via user institutional repositories, publishers, and libraries). Nevertheless, as scientific institutions which are at the centre of significant scientific communities and who would benefit from increasing the support available for those communities, there is value in facilities supporting this as much as possible.

Thus there is an opportunity within PaN-data to :

- Allow the sharing of publication data between facilities. This may allow a more complete record of publications to be established, especially on cross-facility projects and users, and for publication arising from *derived studies* (ie studies which exploit the findings of a previous experiment, indirectly (via a citation) or directly (by reusing data)).
- Provide an added value service to users to record and search for publications within the neutron and synchrotron science community ("PaNPubs").
- Provide linking to other outputs to trace provenance (discussed more later).

2.1.4 Barriers

There are a number of barriers for collecting and sharing:

- Facilities while seeing publication reporting as necessary, and provide systems to support this for their own purposes, may not see sharing and adding value as priorities to assign resources to.
- The overlap between publications useful to more than one facility may be small.
- Identification of unique identities of users may be difficult; sharing user information may assist in providing such user information.
- Licensing from Web of Knowledge may prevent the sharing of data accessed using it.

2.1.5 Possible ways forward

A number of actions might arise from the publications agenda, to make publication gathering more effective and also make the publication data more useful as added value.

- Propose common cataloguing standards, metadata and controlled vocabulary, including facilities specific concepts (see below and also WP6).
- Federate and share publication data to allow cross searches and harvesting of data between facilities (possibly PaNPubs – a federated catalogue of photon and neutron source publications).
- Provide common guidelines and framework to link publication to other parts of the facilities research lifecycle.
- Common policy for publication deposit, including derived studies.
- Citation tracking to assess the impact of the publications arising from facilities work

Given the level of support already provided in this area, there may be some quick wins here.

2.2 Data Cataloguing

The next topic of the questionnaire addressed how facilities are currently managing the cataloguing of experiments and the resultant raw data generated from use of the facilities.

2.2.1 Current situation: commonalities across facilities

Although not all the facilities in PaN-data carry out systematic cataloguing of experiments and associated raw data across the whole of the facility, there is a general recognition that this is a valuable activity. Several partners either have systems in place, fully or partially (DLS, ISIS, ILL, LLB, SINQ, LLB, SOLEIL), or currently under development (DESY, ELLETRA, HZB).

The resulting catalogues are almost always restricted to the facilities team and registered users (particularly data owners), so these are not a public catalogues, but rather of primary use to the facility and their users.

2.2.2 Current situation: differences between facilities

A variety of different solutions and packages are used for cataloguing data. ICAT is part of the solution for four facilities, although integrated into a bespoke architecture. Others use a variety of custom databases (although Oracle is a common database platform).

There is a mixed situation of recording annotation and lab note books to provide supplementary information on experiments.

2.2.3 Opportunities and drivers

The drivers for providing a facilities data catalogue would include:

- To keep a systematic record of experiments undertaken, their actors, the sample materials investigated and the results gathered, so experiments can be evaluated and unnecessary repetition avoided.
- To assist the management of data, so data collected is systematically stored and made accessible for further analysis to the investigation team.
- To provide a long term record and archive of the experiments and data undertaken at the facility to allow long term retrieval and reuse.

All three drivers are acknowledged to different degrees by the facilities, although data management is seen as the primary driver.

Thus by providing a more systematic means to catalogue data, via shared expertise and reference implementations, these drivers can addressed more effectively and efficiently.

If these data catalogues are then shared, further opportunities arise which extend these three drivers:

- If there is a record of experiments undertaken across facilities, their actors, the sample materials investigated and the results, then results can be evaluated more widely and repetitions between facilities avoided.
- If data is taken at more than one facility by the same team or on similar samples, then the data between them can be accessed from each for further analysis to the investigation team.
- Long term record and archives of the experiments and data can be investigated for reuse from a wider range of sources, allowing future recombination and reanalysis of data from different facilities.

2.2.4 Barriers

Barriers and obstacles to establishing and sharing data catalogues would include:

- Difficulty in integration of data catalogues into the software environments and processes, such as the different user office systems which need to work with.
- Problems of effectively collecting metadata accurately and with minimal interference with users (preferably none, automatic gathering being more reliable and accurate).
- Embargo periods and different data policies making sharing difficult.

- The culture within different organization and communities may not be in favour of data sharing and this may require advocacy of benefits and cultural change, clarifying the role and value of the experiment catalogue, and in sharing it more widely. This may become harder still once lab note books are

2.2.5 Possible ways forward

There is a lot of activity and a recognised need here, although benefits may need to be articulated better. But there is room for improvement in:

1. Making it easier for facilities to introduce data cataloguing.

This could be supported by:

- Providing reference metadata standards for data cataloguing
- Providing reference code for providing data catalogues
- Defining common interfaces and APIs to data catalogues, including to user office systems and analysis packages.
- Providing expertise and advocacy.
- Provide a quality assurance on the level of data cataloguing.
- Consider how to integrate other tools (e.g. LIMS, electronic lab note books) into the infrastructure

2. Providing infrastructure for federation sharing
   - Again common metadata standards and APIs would ease the interoperability.
   - Provide a common vocabulary for describing facilities, (e.g. beamlines, instruments, roles, samples etc), so they can be shared.
   - Provide a common portal?
   - Enforce facilities data policy within a common data policy framework.

2.3 Provenance and linking

The third topic of the questionnaire addressed how facilities are currently addressing issues around the management of derived data.

2.3.1 Current situation: commonalities across facilities

As noted in the publications section, all facilities recognized the value in linking between the experiment proposal and resultant publication, for recording and reporting, and most had or are putting systems in place to record this link. Most however are not necessarily linking publications to raw data, which would be necessary to complete the provenance chain for reanalysis.

2.3.2 Current situation: differences between facilities

Providing support for storing derived data can vary widely. If it is done, it tends to be done on a local scale on a instrument or instrument scientist level, providing ad hoc or temporary basis. A similar situation occurred with software recording provenance, and added value functions such as assisting with the generation of supplementary material for publication.

2.3.3 Opportunities and drivers

The drivers and opportunities here are seen as rather more long term. Managing derived data and linking to trace provenance could:

- Provide added value services to the user base, supporting and enhancing the community with managed infrastructure for derived data and software.
- Providing added-value services such as packages of supplementary material.
- Support the better management of the analysis process downstream of the raw data collection – for the benefit of the investigator.
- Provide mechanisms for validation of results.
- Reuse of parts of provenance for secondary analysis.
- Provide citation and credit for data collection and management (see below).

These opportunities are recognized by facilities as useful aspirations, as providing added value to the data collected and especially in increasing the scientific integrity of the resultant work.

2.3.4 Barriers

However, a number of issues were raised which were of concern.

- Concerns were raised on the size and unbounded scale of the storage and other support managing derived data would entail, with resultant costs of staff and resources.
- Recording raw data may be seen as a “science problem”, with any advantages arising to the investigators, and if there were advantages to be gained from it, it would be the responsibility of the users, not the facilities infrastructure providers.
- The technology is not mature enough to support this systematically and effectively.

2.3.5 Possible ways forward

There is a feeling amongst some that the case for this is not as yet proven in general for providing support for derived data, beyond the publication. The costs were seen as being unpredictable and the benefits within the “science” rather than the infrastructure provision, and therefore that was where the responsibility for providing support in this area lies.

Others were more enthusiastic and saw they had a role in supporting the downstream analysis process, providing software and software which could track provenance, assisting the user in generating and managing derived data and preparing for publication. There was a recognition that this was an area which would want to be pursued in the future.

Possible ways forward would be to:

- Concentrate on mechanisms to record the link between publication to experiment see the value
- Interact with the user community to establish use cases and need
- Provide advocacy to the community on benefits and costs.
- Monitor best practice and technology within the community.

- Undertake trials and use case studies into the value of the recording of provenance and the retention of derived data to develop a science case.
- Develop appropriate tools and technologies to support derived data and provenance within facilities tools and practices.

2.4 Preservation

The final topic of the questionnaire addressed how facilities are considered facilities attitudes and approaches to long-term archiving and preservation of data, in particular raw data from experiments.

2.4.1 Current situation: commonalities across facilities

All facilities felt that they had strong and well-managed provision for the storage of raw data, so storage of the raw bits are not seen as an issue. The storage is usually undertaken within a dedicated data storage facility, with specialist staff and managed process, including back-ups and some integrity checks on the data. There have been few serious losses of data, and little problems with loss of context resulting in un-interpretable data.

As only a few data formats used for the storage of data, which have been well documented, and now subject to standardization in NeXus, then managing data format change is not seen as a major issue, and there are a number of data converters available which are applied as needed.

Currently, there is little or no assignment of persistent IDs to data, although this is beginning in some cases. There is a recognition that this as useful approach to allow data citation, which is seen by most as a strong benefit.

2.4.2 Current situation: differences between facilities

Notable difference between neutron and synchrotron sources on the time scales they were willing to commit to the storage of data. While Neutron and Muon sources (ISIS, LLB, ILL, SINQ, SμS) stored all experimental data from the initiation of the facility, some synchrotrons were much less willing to make long term commitments, guaranteeing to store data only for a period of months. This is undoubtedly a result of the much larger amounts of data which are generated within synchrotrons with the resultant cost benefits.

Retention and disposal policies are not consistently applied across the facilities, though this is also covered within the data policy framework, which may result in clearer statements on these issues.

Representation information in the OAIS sense (often confused with metadata) is managed in a fairly mixed and unsystematic fashion.

2.4.3 Opportunities and drivers

Some drivers for preserving data in the long term within facilities:

- Access to the raw data for further analysis by the investigators. Investigators research may extend over many years, and they may return to study the data at a later date.

- The validation of results by accessing the raw data and checking the analysis; to carry this out fully, the entire provenance chain needs to be preserved, including software.
- The secondary analysis of raw data by other researchers.
- Save the possibility of inefficient repetition of experiments on the same sample.

However, they are nevertheless seen to have some issues.

2.4.4 Barriers

Some barriers to preservation identified include:

- Volumes of data being generated, particularly within synchrotron sources.
- Open ended commitment to a long term cost in storage and management, including staff costs.
- Uncertainty as to the value of storing data for the long term as it may not be reused.
- Raw data reuse may not be a strong driver in this community, as the data from a particular instrument is seen as specialised, only of interest to a small community.

2.4.5 Possible ways forward

Data preservation is a currently major issue in research, with a number of policies, initiatives and projects being developed at national and international level. The case for preservation in areas such as environmental science and astronomy, where observations are essentially non-repeatable are clear cut.

In photon and neutron facilities science, the case is less clear. Experiments can be repeated on a sample of the same material if necessary, potentially generating a better result, if the measurement was more carefully or better calibrated than previously. And with technological progress, new generations of instruments and facilities are likely in the future to produce a better result, so the data may have only a say 10 year useful status as “the best available data”. Thus the argument for preserving data is more nuanced.

As a consequence, there was mixed response to issues of preservation. Synchrotrons in particular are worried about data volumes and long term cost, while the much smaller volumes of neutron sources mean that they able to include a long term commitment as part of their normal operation, at least as long as the facility operates. Doubts were expressed about reusability of data either for its usefulness in new circumstances or realistic ability to recreate the original analysis. As a consequence, it would be valuable for PaN-data to explore the case for preservation.

Some possible ways forward might include:

- Recommendations on the use of persistent ID for experiments and data, and potentially services to support those ID (e.g. DOI services).
- Policies, formats and recommendations to support data citation by users within publications,
- Development of retention and disposal policies, as part of the common data policy framework.

- Guidance and tools for drawing up OAIS compliant Data Management Plans appropriate for PaN-data facilities.
- Trials and scenarios preservation within facilities to develop models for benefits and costs, for advocate within facilities to present a science case.
- Identification of “simple wins” on preservation such as integrity and format checks.
- Identification of suitable representation information for facilities data

2.5 Overall Future opportunities and costs

All these areas were seen as ones which PaN-data partners would need to pursue in the future, but there were clear priorities. There were particularly strong requirements around publications, and data cataloguing. Provenance and preservation issues were one which there were doubt around benefits and costs, and there is a clear need for trials and advocacy to prove the case. Some facilities are leading in these areas, and it would seem appropriate if they were to lead on demonstrating where the value lies.

3 Results from the questionnaire

The questionnaire has three parts: on current practices, on state of the art, and on desires, requirements, opportunities and constraints. We include here all the questions together with the collated responses to each.

3.1 Current practices

3.1.1 Publication tracking and Management

1. Do you have a current requirement to track publications:

	Produced by staff of your institution?	Produced from the use of your facilities?	Comments
DESY	y	y	
DLS	y	y	
ELLETRA	y	y	It is done through the ELETTRA Virtual User Office (VUO) system
ESRF	y	y	this is done by our library with a link to our proposal system (SMIS)
HZB	y	y	
ILL	y	y	
ISIS	y	y	a. Yes, enforced by management emails near annual report time b. Yes, and increasingly so.
LLB	y	n	a. yes from CEA b. no
PSI	y	y	
SOLEIL	y	y	

2. Do you have a policy for users to inform you of resulting publications or deposit papers in some place?

DESY	y	
DLS	y	
ELLETRA	y	If they don't inform us they can be penalized for further beam-time access
ESRF	y	this is a condition for obtaining beamtime the second time.
HZB	y	through our virtual user office GATE
ILL	y	Yes when referring to recent papers in a proposal they must enter the papers in the library database
ISIS	y	Yes, but not well enforced or followed. This will hopefully be improved in the next 12 months

LLB	y	Yes we request them by e-mail each two years when we inform on the deadline for beamtime request
PSI	y	Yes, users are asked to register facility related publications in the PSI Digital user office (DUO) database: https://duo.psi.ch
SOLEIL	y	Yes, users are requested to submit the references of their publication(s) in the SUN set application (based on PSI-DUO). Publication record is made available to the Peer Review Committee members. Publication submission is mandatory. Failure to provide publications may prevent the proposers from being allocated beamtime.

3. Do you have a central computerised system for recording publications? What platform is it based on?
    a. An institutional repository? (based on which software platform?)
    b. A database
    c. Other?

DESY	Yes, essentially everything is or will be managed through inspire, which is based on cds Invenio.
DLS	Yes, a customized database http://www.diamond.ac.uk/Home/ForUsers/academics/publications.html
ELLETRA	It is a single sign-on web-based system (Virtual User Office (VUO)) with Oracle as back-end DB. https://vuo.elettra.trieste.it/pls/vuo/publi_mgr.startup
ESRF	A commercial system called FLORA from the company EVER http://www.esrf.eu/UsersAndScience/Publications
HZB	Yes, based on an Oracle database http://www.helmholtz-berlin.de/pubbin/search.pl?sprache=en
ILL	A database in the library http://www.ill.eu/science-technology/scientific-publications/
ISIS	Use ePubs for storing some publications. EndNote used to maintain ISIS records and to produce annual report data. http://epubs.stfc.ac.uk
LLB	No - This is a request from the CEA
PSI	PSI uses the DUO system to record publications related to its user facilities SLS, SINQ, SμS. DUO is a web based system with php architecture and an Oracle DB.
SOLEIL	A dedicated EndNote-based tool is used for handling SOLEIL and users publications. This tool is fed by: - References of SOLEIL publications - References of the users' publications that are recorded in the SUNset database (Oracle DataBase) when submitted. - References of harvested publications (see next question, 2.1.1.4) http://www.synchrotron-soleil.fr/Recherche/Bibliotheque/DocumentationScientifique

4. Do you track/harvest publications recorded in external services? (e.g. Web of Knowledge (http://www.isiwebofknowledge.com/), ArXiv (www.arxiv.org) ? SPIRES (http://www.slac.stanford.edu/spires/)? Other?)

DESY	yes, all of those (+inspire)
DLS	Yes, through web of knowledge and searched via Google Scholar
ELLETRA	Only occasionally, not systematically.
ESRF	Not systematically
HZB	Yes, using ISI Web of Knowledge. We are currently implementing a routine which can be used to automatically imports datasets in our database.
ILL	Yes to make statistics that show how well we do compared to other facilities
ISIS	Not yet, but have done some pilot work and plan to soon
LLB	Yes - ISI
PSI	PSI uses ISI WoK. DUO is able to import ISI metadata and makes it easy for users to enter their publications. In addition 'normal' ISI campus license is available for each user of the PSI intranet. Fully automated publication tracking is not performed.
SOLEIL	Yes daily on ISI web of knowledge

5. What is the success rate of your system? Do you reliably know what publications result from your users' work on your facilities?

DESY	No, not 100% reliably. I'd guess that the success rate is ~80-90%.
DLS	We have no way of tracking the absences – any publications we find are included.
ELLETRA	As stated before users are required to update their records before requesting beamtime. The success rate depends anyhow on the users and may be error prone.
ESRF	Probably pretty good because most of our users are coming several times and have to make their publications known to us.
HZB	Up to now, the success rate is rather low. When it comes to reporting, staff members search the ISI Web for publications not registered so far.
ILL	We suspect that this is ~70% (with a large error bar)
ISIS	High for staff publications due to management emails (and in any case easy to track through e.g. Web of Knowledge by institutional address – nowadays >95% accurate compared to individual returns) Probably low for users due to no effective system or enforcement (estimate only 80% collected)
LLB	We have no definite information but estimate +/- 15%
PSI	The success rate is in the order of 90-95% (estimation).
SOLEIL	As explained above, the success rate depends on the users. But we are confident enough as users are required to update their records before requesting new beamtime, and can be penalized if they don't do.

6. Who (if anyone) in your organisation has responsibility for tracking publications?(e.g. Library, User Office, Admin?)

DESY	Library. But the User Office is of course reminding users to supply such information.
DLS	Managed by the web manager in the Communications team.
ELLETRA	Library
ESRF	Library
HZB	User coordination for user publications, the scientists themselves for publications by staff members.
ILL	Library
ISIS	User Office. Library staff do some trawling
LLB	Library
PSI	For publications related to the user facilities: PSI user
SOLEIL	Library with daily bibliographic search on ISI web of knowledge and User office with automatic reminder after 6 months of a carried experiment.

3.1.2 Data cataloguing

1. Do you systematically catalogue the experiments and associated data produced by your facility?

DESY	Experiments yes, but not exactly systematically. Associated data: in the very near future.
DLS	Yes
ELLETRA	No in operational level – but new systems are currently under test (ICAT).
ESRF	No
HZB	Information on the experiments is catalogued partly for in-house publications. Work on user publications is in progress. There is no catalogue for measured data!
ILL	Experiments and proposals via our 'visitor club' (Oracle database), all data since 1970 are saved/archived, since 1995 available via a simple catalogue, now available via ICAT
ISIS	Yes
LLB	No not centralized, but on each spectrometer
PSI	For some instruments at SINQ. Therefore the following answers apply only to these SINQ Experiments, not to the SLS.
SOLEIL	Yes, for most of the beamlines (15 out of 19 operating ones): the ones producing NeXus data files via the tools provided by the SOLEIL computing Division

2. If so, what software do you use for this purpose?

DESY	Self-made + dCache.
DLS	Bespoke in detail
ELLETRA	ICAT, VCR and VUO
ESRF	-
HZB	In-house solution relying on a Oracle Database, no commercial product.
ILL	ICAT

ISIS	ICAT
LLB	-
PSI	At SINQ we use an oracle database, some extraction scripts run by cron and a home grown WWW interface for this purpose. The system is old and up for replacement. Replacement does not happen due to lack of pressure and man power constraints
SOLEIL	All the NeXus data files and experiments are indexed into an Oracle database, using a dedicated “home-made” program.

3. If so, what is your primary motivation? For record keeping? For Data management? For archiving?

DESY	Data management, then archiving.
DLS	All three
ELLETRA	All of the above plus Data Preservation.
ESRF	-
HZB	Reporting and record keeping.
ILL	I think we have evolved naturally towards keeping everything we can
ISIS	Data Management
LLB	-
PSI	SINQ uses this for record keeping and statistics and as a tool to locate data files.
SOLEIL	Primary for data management. We’ve developed a tool allowing our users to browse and retrieve their data from outside of Soleil

4. Do you allow the catalogue to be searched to find experiments of interest? If so, is this restricted or made public?

DESY	Yes. Fully restricted for some time, and always restricted to registered users.
DLS	Not currently
ELLETRA	Restricted as it is on the pilot stage.
ESRF	No
HZB	Very limited search capabilities (Group, Author, Year, Title) for the in-house publications, restricted to staff members.
ILL	YES, public
ISIS	Yes. Subject to ISIS data policy. Basically public after 3 years embargo.
LLB	Restricted because not accessible but free on request
PSI	We do search but restricted to PSI staff
SOLEIL	Yes, we allow the catalogue to be used for retrieving data but it’s restricted to the experiment team

5. Do you capture and annotation or notes associated with the experiment (e.g. lab notebooks)?

DESY	No, but would like to.
DLS	Some but not really very effective
ELLETRA	Yes, extensively. Most are on traditional paper lab books but additional electronic means exist too (VCR portals Logbook)
ESRF	Some beamlines use an electronic logbook, but this is not linked

	to the data generated on the beamlines. The majority of beamlines still use paper log books. The MX beamlines capture meta data together with the data.
HZB	No
ILL	No
ISIS	No
LLB	Each spectrometer
PSI	At the SINQ yes. At the SLS yes, in a beamline specific way.
SOLEIL	Yes, with a weblogs, ELOG from PSI

6. Who (if anyone) in your organisation has responsibility for cataloguing data?

DESY	Different staff members have different roles in the process. Still under development.
DLS	Data Acquisition and Scientific Computing
ELLETRA	Scientific Computing Activity/team (scicomp@elettra.trieste.it) of the IT group and beamline managers.
ESRF	If beamline data is meant: nobody
HZB	The staff of the management board, if anyone….
ILL	IT Group
ISIS	ISIS computing group (along with STFC eScience) build and maintain the software and tools. Users and instrument scientists need to assign data by reference number
LLB	Each instrument responsible
PSI	There is no clear responsibility for this
SOLEIL	Cataloguing data is an automatic process implemented by the SOLEIL computing division

3.1.3 Provenance issues

1. Do you trace or link proposals at your facility to any publications based on them?

DESY	Yes, maybe. Though there is a link between a publication and a proposal, the correspondence is not exact, since publications are usually linked to a single proposal even if different proposals and experiments were involved.
DLS	As much as possible
ELLETRA	Yes even if the process can be improved.
ESRF	Yes
HZB	Yes
ILL	No
ISIS	Not yet. But this is a requirement
LLB	No
PSI	Yes, we do!
SOLEIL	Yes systematically (as soon as we know their references: via SUNset submission or ISI web of knowledge)

2. Do you trace or link raw datasets collected at your facility to any publications based on them?

DESY	Yes, in the near future, working on the final implementation.
DLS	Not really

ELLETRA	No – but experiments are currently performed (Nexus/HDF/ICAT)
ESRF	No
HZB	No
ILL	No
ISIS	Not yet. But we are issuing DOIs, in part, to facilitate this
LLB	No
PSI	Not yet!
SOLEIL	Indirectly through the “proposal number”: raw datasets and publications are referenced with the proposal number, but in two different databases.

3. Do you collect, store or allow the deposit of derived and analysed data within your facility? If so how do you record the connection to proposal, raw data or publication?

DESY	No
DLS	Done on a per visit basis
ELLETRA	Yes, derived and analysed data are often stored under the Beamtime's proposal number together with the raw.
ESRF	No link between data and publications yet
HZB	No
ILL	No
ISIS	Not yet
LLB	No
PSI	At SINQ and SLS this is the users responsibility. We provide only temporary storage for this kind of material.
SOLEIL	When processed via SOLEIL means, derived and analysed data can be stored in a dedicated sub-directory of the proposal directory,

4. Do you record the software used to derive data ?

DESY	No
DLS	No
ELLETRA	Yes - when the software is in house developed and works in automated ways concurrently with the experiment. For the rest, the Beamline responsible is aware of it.
ESRF	No
HZB	No
ILL	No
ISIS	Some software (Mantid) does record this in the analysed files
LLB	No
PSI	At SINQ: NO. At the SLS this depends on the beamline and experiment but in general no
SOLEIL	Yes, for In House developed software. Otherwise, it is of the responsibility of the scientists.

5. Do you assist users in the gathering of supplementary information for publications for submission to publishers?

DESY	Good point. No
DLS	Not directly

ELLETRA	Yes, the Beamline personnel does this.
ESRF	No
HZB	Not in a formalised way.
ILL	No
ISIS	No
LLB	No
PSI	At SINQ and SLS: not systematically, but users can ask for help with this
SOLEIL	Sometimes if it is a collaboration with the Beamline staff

6. Who (if anyone) in your organisation has responsibility for linking experiments to data and publications?

DESY	The user.
DLS	No
ELLETRA	At the moment, no one.
ESRF	Nobody
HZB	Does not apply.
ILL	No one
ISIS	No one as yet. Plan is for users to do this as only they really know.
LLB	No one
PSI	Regarding publications: Beamline managers, users, user office
SOLEIL	Linking experiments to data = automatic process implemented by the Computing Division Linking experiments to publications = the user has to link publication and proposal/experiment when he submits the publication reference and abstract

3.1.4 Preservation Issues

1. Do you keep the data in an archival store? As a backup or dark archive copy?

DESY	Yes, dCache.
DLS	Yes
ELLETRA	Yes
ESRF	All data is backed up to tape for 6 to 12 months. Some data is archived on tape.
HZB	Not for user data: The responsibility for the safekeeping of measured data lies with the users.
ILL	Data is in a readily accessible file system
ISIS	Yes. Layered system with 3 local checksummed copies on mirrored spinning disk, a tape backup and as a dark archive.
LLB	Each instrument is responsible, it is technically available
PSI	SINQ, muSR: yes, online and archived SLS: beamline specific, in most cases the in-house data are archived and/or backuped whereas the data of external users are only stored for a short period of 60 days.
SOLEIL	Yes for Disaster Recovery Plan (DRP)

2. Have you ever encountered problems of reading or interpreting archived data?

DESY	Yes, HEP-data, but not so much on the archiving site.
DLS	Yes
ELLETRA	On recent years no, but older tape systems were more problematic.
ESRF	No
HZB	There have been issues with broken tapes and tape-drives but none that eventually couldn't be resolved. Interpretation of the data lies with the user-groups and problems may be unknown to us.
ILL	I have one example of minor format issues with data from 1970!
ISIS	Yes - had problems with corrupt disks which have been resolved using local copies (2 of the 3 disks). Never actually needed to restore from the dark archive
LLB	Yes
PSI	SINQ: no SLS: The tape archive is in general quite reliable over a period of several years but data loss has been encountered in rare cases.
SOLEIL	not applicable up to now

3. Do you assign persistent identifiers (e.g. DOIs or managed URIs) to data to assist in the citation and retrieval of the dataset?

DESY	In some sense yet. The identifiers are unique, but not citable. URIs are persistent in dCache.
DLS	No
ELLETRA	No – but it is under study.
ESRF	No
HZB	No
ILL	Not yet
ISIS	Yes – to data at the experiment level
LLB	No
PSI	DOI numbers are only linked to registered publications, not to data SINQ: no, but data file names are reveal a lot SLS: no, but in some cases data file names are a unique, experiment specific identifier
SOLEIL	We don't store a specific information equivalent to a DOI, but we easily can provide an URI from the information stored in our data catalog, we currently think about this through the Common Data Model project

4. Do you have a retention policy on how long you commit to keep data? Do you have a process to decide which data to retain?

DESY	Not yet.
DLS	Retention policy: Yes; data retention process: No.
ELLETRA	*Not an enforced one.*
ESRF	Yes. Data is deleted automatically beyond the defined retention time.
HZB	Data (as part of the backup-system) in general has a retention

	time of 6 weeks. Data characterizing the facilities operations has a retention time of 12 month.
ILL	We keep all data
ISIS	Data is currently sufficiently small to keep it all.
LLB	No
PSI	At SINQ: not officially. But all data since the start of SINQ is retained. SLS: Data of external users is stored for 60 days. For in-house data the individual groups decide. Typical storage times may be on the order of 3-5 years, strongly varying with experiment and research group
SOLEIL	The policy defined at SOLEIL is to keep data: - 100 days on central disks, - between 1 or 5 years on tapes depending on the amount of data that the beamline is producing. Due to the Active Circle software, data on tapes can be read in the same way as data on disks without any computing staff action (but obviously with a slower access time). After this delay, data can be archived, and the beamline leader is deciding which data to keep or not (he only knows the scientific issue of these data

5. Do you have a disposal policy for the selection of data for deletion?

DESY	Proposed, but not yet implemented.
DLS	No
ELLETRA	No
ESRF	No, because data is deleted automatically when expired (after the defined retention time).
HZB	The facility does not actively delete data. As long as users stay within the allocated disk quota data and his/her account is still valid data will be kept indefinitely.
ILL	No, see above
ISIS	No – see above
LLB	No
PSI	SINQ: no SLS: For external data 60 days after they have been recorded.
SOLEIL	See previous question

6. Do you manage checks (e.g. Checksums) on data to ensure that its integrity is maintained whilst in storage?

DESY	Yes
DLS	No
ELLETRA	Yes but more advanced ways than checksums. Integrity is ensured at system level.
ESRF	No
HZB	Inherent feature of the backup-system (all backups are checksummed). No checksum are being used for the active (on-disk) data store
ILL	I don’t think so.
ISIS	Yes on the spinning disks, but not on tape copies
LLB	No

PSI	SINQ: checksums: no. But data is protected read-only and we have backups. SLS: in general no
SOLEIL	Not yet. But it could be done in the future; due to a new feature in the file system we are using (active Circle).

7. Do you undertake any transformation processes to convert old formats into new ones?

DESY	No
DLS	Software maintained
ELLETRA	Yes, there's a plethora of in-house developed converters – and their design is a regular activity.
ESRF	No
HZB	Yes, when storage technology is being replaced. As there is no long-term store no special measures are required (the old format mustn't last longer than retention times)).
ILL	No, our format has not changed!
ISIS	ISIS RAW format files can be converted into NeXuS my Mantid. This is not done as part of the archive process.
LLB	No
PSI	SINQ: has not yet been necessary but will be done when required SLS: in general no
SOLEIL	No need for the moment

8. Do you maintain any supplementary information (metadata or “representation information") to retain context and understanding of datasets? Who is responsible for adding this supplementary information?

DESY	Not yet. Responsibility lies with the data producer.
DLS	Yes
ELLETRA	Yes, often stored in text or HDF5. This info is added by the beamline (automatically or manually entered by the personel)
ESRF	Yes, some metadata is added automatically, some manually, but globally the metadata information is insufficient to correctly describe the data.
HZB	No, the responsibility for this lies with the user groups, who in general keep run-books with information on their data-files.
ILL	The CS group set up a simple database by extracting a minimal set of metadata from raw data files, no metadata is added. ICAT will make a much better job of this.
ISIS	Yes. Automatically entered from proposal or experiment control system. Users can enter limited additional metadata at experiment time, but not afterwards.
LLB	On the responsibility of each spectrometer responsible
PSI	SINQ: yes, in data files. Users and instrument scientists are responsible for entering data SLS: yes, in separate data files. Beamline scientists are responsible for deciding on the beamline specific mechanisms for this and therefore quite different standards are in place. Additionally an automated archiving system is available for storing in pre-defined intervals or upon changes above a certain

	threshold values like temperatures, beam positions, motor positions etc.
SOLEIL	Raw data and Metadata are integrated in NeXus data files. Beamline staff defines the required metadata. The acquisition process software which is provided by the Computing staff collects them and integrates them into the NeXus files

9. Who (if anyone) in your organisation has responsibility implementing your archival policy?

DESY	IT staff.
DLS	DASC
ELLETRA	IT group, User Office, Admin.
ESRF	The IT group.
HZB	There is no archive for measured data. The responsibility for archives of operational data lies with the specific organizational unit.
ILL	IT Group
ISIS	ISIS computing group. Hope to outsource to STFC eScience soon (SDB project)
LLB	Direction
PSI	No clearly identifiable person responsible SLS: Archival of in-house data is decided individually by each group. Tape archive facility is provided by the central IT department
SOLEIL	Computing Division

10. Who is responsible for data storage?

DESY	IT staff.
DLS	CASTOR/e_Science
ELLETRA	Scientific Computing Activity/team (scicomp@elettra.trieste.it) of the IT group.
ESRF	The IT group.
HZB	The IT-department far as storage-systems (hardware) are concerned. User-groups as far as usability of measured data is concerned.
ILL	IR Group
ISIS	ISIS computing group. Hope to outsource archive store (but not immediate local copies/store) to STFC eScience soon (SDB project)
LLB	Each instrument responsible
PSI	SINQ: NUM computing staff SLS: Data storage is provided by accelerator department computing staff
SOLEIL	Computing Division: - the network and system group is in charge of the data storage architecture, eg hardware and file system layer. the Control and Data acquisition group is in charge of the applicative level

3.2 Best Practices

Can you recommend any methods, tools and projects inside and outside the photon and neutron community relevant to publications, data cataloging, managing derived data, provenance and preservation which you would like to see considered within the survey?

DESY	Data Management and Storage Systems (e.g. Fedora, dCache). Document Management Systems (e.g. CDS Invenio)
DLS	
ELLETRA	
ESRF	The Australian National Data Service.
HZB	
ILL	
ISIS	
LLB	
PSI	
SOLEIL	

3.3 Desires, requirements, opportunities and constraints

1. What would be your requirements:
   - Publication management and tracking
   - Derived data
   - Tracing provenance (including software)
   - Preservation

DESY	• Publication management and tracking: we presumably will use inspire / invenio. Fullfills our requirements. • Derived data: no concrete plans to manage derived data. Strongly depends on the input from the user. • Tracing provenance (including software): tightly connected to derived data, so no clear position on that one. • Preservation: Bit stream preservation: I think, dCache fullfills our requirements
DLS	All
ELLETRA	All plus: • Tracing of Algorithms • Privacy management • Tracing of HW (i.e. stating the HPC requirements etc)
ESRF	• Publication management and tracking – linking to data • Tracing provenance (including software) – Yes • Preservation – Yes, also preservation of software
HZB	At the moment there is no agreed position on this.
ILL	• Publication management and tracking: more reliable

	tracking of ILL output. DOI's for data should help here. • Derived data: - • Tracing provenance (including software): something for data treatment software, which PaNsoft and PaN-dataODI are addressing. Most of the other pieces are structured and in databases that could be linked • Preservation: OK for raw data, nothing structured for e.g. software
ISIS	• Publication management and tracking A system our users can use to easily find their publications (e.g. WoS search) and link them to proposals and data. Must also be able to cover other experiment outputs such as conference talks, other grants held, EU funding, industrial links. • Derived data The ability to upload derived data and link it to raw data and a proposal in our ICAT. Must be many to many (i.e. derived data could be from n experiments) • Tracing provenance (including software) Yes - it is desirable to be able to show what analysis programs were used and their inputs. Analysis software should support this and I believe nexus can record this. • Preservation
LLB	• Publication management and tracking yes • Derived data yes • Tracing provenance (including software) yes • Preservation yes
PSI	• Publication management and tracking Tracking is still quite labour intense. We encourage users to register their publications by making this process easy and by requiring recent publications for beamtime applications but this is still not sufficient. • Derived data • Tracing provenance (including software) • Preservation There are several synchrotron beamlines where long term preservation of user data would cause considerable costs, i.e., this can not easily be provided by the facility.
SOLEIL	The whole

2. What benefits do you see accruing in these areas? Can you give any explicit scenarios of use in these areas?

DESY	-
DLS	-
ELLETRA	The benefits can be both immediate and long term. A possible outcome could be that of a well organised safe repository where all the future disseminated results (papers etc) can link to the corresponding data.
ESRF	Secure on-line storage of data for the scientists doing the experiments, persistence of data for long-term projects, data reuse, re-analysis, proof in case of scientific fraud, correlation

	with other data sets, combination of data for experiments done in several labs with different methods, statistics for funding bodies, etc.
HZB	Benefits would be on the scientific side.
ILL	Retreating data to verify results but also to derive new results from existing data.
ISIS	Publications – full two way linking to proposal (via raw and analysed data)
LLB	-
PSI	– Scientific integrity – Taking the increase of data rates and volumes in several synchrotron applications into account it could be potentially helpful for users to provide a data storage and analysis center with remote access for them. – But there is no consensus on the potential benefits. E.g. one side argues that the costs of such a managed infrastructure are too high. Researchers are afraid of losing resources for their research. The other side argues that the management of the data costs in any case, but today the costs are well hidden and spread over many places. In their view it is very likely that a real cost estimate taken into account all hidden costs will reveal, that a better managed approach saves resources, which can then be returned to the research community for improving their research. – Unfortunately no systematic analysis of the pros and cons is known to us right now. Depending on their own experiences people either weight the risk higher than the chances or vice versa. Therefore a solid collection of arguments concerning pro and contra of a managed data (and/or analysis) infrastructure ,which are backed up by facts, would be very helpful. May be PaN-data could provide such a summary?
SOLEIL	Improvement of the current process and support. Knowledge sharing. Getting the most of the experiments: correlational researches, etc.

3. What costs and constraints can you foresee in the provision for support in these areas?

DESY	Costs are unclear. It requires a clear definition of support levels and reliabilities, which are not available yet. Costs for staff will presumably largely exceed costs for storage/archive, but that's much more difficult to acquire.
DLS	-
ELLETRA	The main constraints will be “political” and administrational – rather than financial and organisational.
ESRF	Mainly in developing the necessary software. The hardware costs will be relatively modest in comparison to the cost required to generate the data.
HZB	Implementation of a data archive for all measured data would incur major investments.
ILL	PaN-data gives some idea of costs and effort. I think that there are cultural constraints with respect to make the whole scientific

	process transparent.
ISIS	Neutron data is small but keeping all derived data could be almost infinite. Policies or quotas would be required.
LLB	1 or 2 full-time
PSI	The constraint is clearly the limited funding available for scientific computing. Thus data management issues will be in competition with DA tasks, which have a potentially higher ROI, and simply the running of the facility. In future it will be very important that the estimate of the total costs for IT resources needed to run an experiment are an integral part of the research funding process
SOLEIL	Main constraints are political and administrative as emphasized by ELETTRA but also financial and human resources. In all the above fields, collaborations between facilities, such as the PaN-data one, are crucial to make accept the concepts and to share the efforts

4. Can you see a benefit from the possibility of reconstructing the scientific process performed on datasets obtained at your facilities - for example, for validation or reworking with different analysis techniques?

DESY	Yes. There were a number of rather spectacular cases of scientific fraud in the last years, which probably could have been prevented by the availability of the raw scientific data. Also quite valuable for educational purposes and application development.
DLS	-
ELLETRA	Yes, such benefit is highly desirable. It could be a great boost for correlational research.
ESRF	Yes
HZB	Not really, as continuous changes in the instrumentation, specific characteristics of e.g. samples in experiments, etc. make it unlikely to repeat an experiment without close interaction of the scientists.
ILL	Yes, for example when publications show the derived data which is not quite what is wanted so it would be good to go back to the raw data and derive a different kind of result.
ISIS	In theory yes. In practice the current metadata is probably too poor to allow this in most cases
LLB	Not necessarily
PSI	The data reduction and analysis must be documented to ensure scientific integrity. In rare cases it is useful to reanalyze data using a modified procedure.
SOLEIL	Yes

5. Can you see opportunities for data becoming citeable in the same way as publications?

DESY	Yes
DLS	-
ELLETRA	Yes. Policies (PaN-data) and cataloguing technologies (ICAT) combined with other techs (VUO, VCR) could create such an opportunity.
ESRF	Yes
HZB	Not really.
ILL	Yes. This would allow experimental teams to get credit for the

	experimental work in the event that they are not publishing the results, unlikely but it may happen when usable data is made publically available.
ISIS	Yes, via DOIs
LLB	Yes
PSI	Yes, in case of well-documented, standardized experiments, i.e., in some but cases but seen over all research areas in the minority of cases.
SOLEIL	Yes

6. Would you say that potential reuse of data is an important consideration in running your facility?

DESY	Difficult to answer, strongly depends whom you ask. We are for example collecting accelerator data, which are an important input to work in simulations and theoretical physics. A significant data are however never going to be reused at all.
DLS	-
ELLETRA	For the Synchrotron ELETTRA and the Free Electron Laser FERMI@ELETTRA, such a reuse of data is of crucial importance.
ESRF	No, not for the time being. This still needs to be seen once adequate metadata capture and data management tools are in place.
HZB	Not really
ILL	Not at the moment because we don't have any experience of this. I think it is something that will start small and grow to be non-negligible in the future.
ISIS	No with regard to raw data. Yes with regard to derived data (e.g. the field of crystallography is largely dependent on databases of derived data).
LLB	No
PSI	No, not yet
SOLEIL	Yes

7. How much of a concern is long-term preservation of your datasets? What do you consider to be the main threats to stored data?

DESY	Bit-stream preservation is not so much of a concern. The main threat seems to be the loss of know-how. The most interesting and challenging projects are often connected to a limited number of scientists, which tend to vanish in the long term.
DLS	-
ELLETRA	One of the main threats is the Retrieval due to inadequate cataloguing (thus difficulty to know what should be deleted and what should not).
ESRF	Long-term preservation is currently only a concern for the scientists doing an experiment because the data remains "private" if only because there is not enough meta data adequately describing the data. Our in-house scientists start asking for a centralised archiving facility to have their data professionally managed to overcome the problems of quantity

	and loss of data. Threat: fire
HZB	So far the HZB isn't obliged to provide long-term storage. For a potential 10-year storage duration, technical problems wouldn't cause too much of a concern.
ILL	None for us but this is an IT issue.
ISIS	Not a major concern as we already do this for raw data and there are tools available to convert our data formats to NeXuS
LLB	Yes, too much data produced
PSI	This is a real issue at PSI mainly regarding costs and manpower. A real concern are the huge data rates produced by the latest generations of Pixel detectors. Presently (2011) there is an ongoing debate within the institute how to deal with this matter in the future. That means the main threats for stored data are – the costs for storing large volumes for extended times – spread of the people who performed and documented the experiment causing a high barrier to make use of data after none of the team members is available anymore.
SOLEIL	Unreadable media due to old technology; unusable cataloguing; ...

8. Do you have any current plans to develop further support in the areas of:
   - Publication management and tracking
   - Derived data
   - Tracing provenance (including software)
   - Preservation ?

DESY	• Publication management and tracking yes • Derived data no • Tracing provenance (including software) not concrete, more long-term. • Preservation ? Yes.
DLS	-
ELLETRA	Yes, to all the above. We intent to invest manpower and money for the purpose – while collaborating with the rest of the photon and neutron community (projects etc).
ESRF	Yes, but mainly in the framework of PaN-data and CRISP. In addition there is an effort to further develop the ISPyB database.
HZB	Not until this will be raised as a requirement from the (scientific) governing bodies.
ILL	See Question 1 above.
ISIS	• Publication management and tracking Yes. In planning/specification stages (discussed briefly with Brian already) • Derived data Initially for some beamlines (Xpress) and published data sets (using ICAT) • Tracing provenance (including software)

	Mantid does this to some extent. ICAT should be extended to support it. • Preservation ? Safety Deposit Box will allow this to some extent.
LLB	• Derived data yes • Tracing provenance (including software) yes
PSI	– Publication management and tracking The present publication database (DUO) is under discussion to be used institute wide as a general publication database. One of the latest DUO developments has been the export to html and the online view of publication lists on the PSI web-pages. – Derived data For many synchrotron experiments users are provided with data analysis software by the beamline staff since the experiments are highly specialized and no easy standard solution is available. The limitation for this support are our resources. – Tracing provenance (including software) No – Preservation ? Not yet. There is an ongoing survey on management level
SOLEIL	Thinking about