Keeping the records of science accessible: can we afford it?

Report on the 2008 Annual Conference of the Alliance for Permanent Access,
Budapest, 4 November 2008

This summary aims to highlight those presentations and comments with the greatest relevance for the key theme of the conference, 'Business models for permanent access', leaving the reader to gain more detailed insights from the individual powerpoint presentations. Reported by Inge Angevaare, coordinator of the Netherlands Coalition for Digital Preservation.

The context: a complicated and imperfect market

At the start of the conference Alliance chair Keith Jeffery drew a compelling picture of the complexity of the preservation landscape. He quoted the Alliance's acting director Peter Tindemans who summed it up as follows: 'It is one thing to estimate the costs of preservation; quite another to estimate the value of preservation, based on the reasons why the records of science should be accessible, and then linking this to whom this value could be attributed in order to allocate costs. However, in many respects science remains what economists call a public good: everyone, not just its producer, reaps the benefits. Hence there must be a significant public role in creating and maintaining the infrastructure for long-term preservation of the records of science. That is, the costs will have to remain largely part of the normal funding models of science.' (Peter Tindemans, IJDC, 1-1-2006). In this context I was reminded of a rather worthwhile presentation by Paul Courant at the LIFE conference in June 2008, which dealt with the issue of public good and its consequences for funding permanent access in more detail [1].

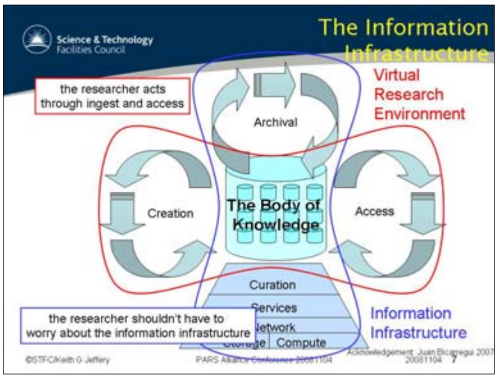
Alliance chair Keith Jeffery at the 2008 conference

Conference chair Ian Halliday of the European Science Foundation: 'The politicians are the key; it is our job to convince them to work for the common good rather than for their own nationals.'

Jeffery went on to quote Helen Hockx of the British Library: 'The situation is further complicated because costs for preservation are, in general, difficult to calculate and are poorly understood; also it is difficult to segregate costs for preservation from costs for access. Since the core funding for institutions does not grow in line with information growth, this makes it difficult to justify and make a case for digital preservation to institutional management.' (Hockx, JISC 2005)

Project funding of research, of course, adds to the problems, often leaving the resulting data to fend for themselves once a project is finished.

Where does the Alliance come in to all of this? Jeffery took his lead from the Alliance's mission: 'The Alliance aims to create a sustainable organisational infrastructure for permanent access to scientific information.' This involves: calculating operational costs, developing real business models, and developing a funding strategy for permanent access.


Science & Technology Facilities Council

The Information Infrastructure

the researcher acts through ingest and access

Virtual Research Environment

Archival

Creation

The Body of Knowledge

Access

Curation

Services

Network

Storage | Compute

the researcher shouldn't have to worry about the information infrastructure

Information Infrastructure

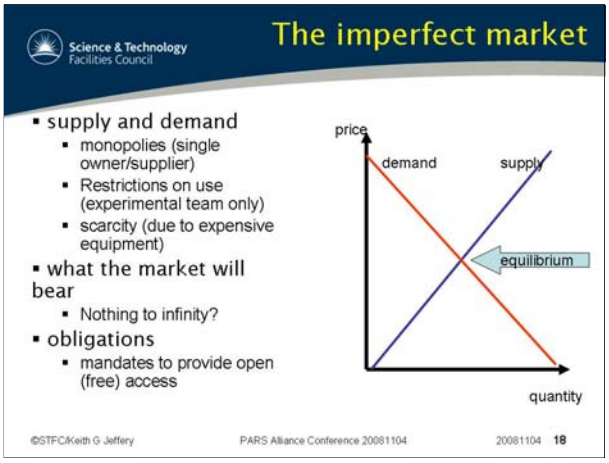
©STFC/Keith G Jeffery

PARS Alliance Conference 20081104

Acknowledgement: Juan Bicarregui 2007

20081104 7

Jeffery said that the researcher shouldn't have to worry about this infrastructure: he interacts with it through ingest and access. Later in the day, however, it was mentioned several times that the researcher is not without responsibilities when it comes to permanent access. First of all he has to be willing to share (which in some disciplines, such as psychology, is by no means a done deal), and secondly, he has to organise and document his data in such a way that they can indeed be properly preserved and accessed later. An recent study by the Research Information Network (RIN) on reasons why researchers do or do not share their data stressed that researchers will need help from data management experts for this [2].



Science & Technology Facilities Council

The imperfect market

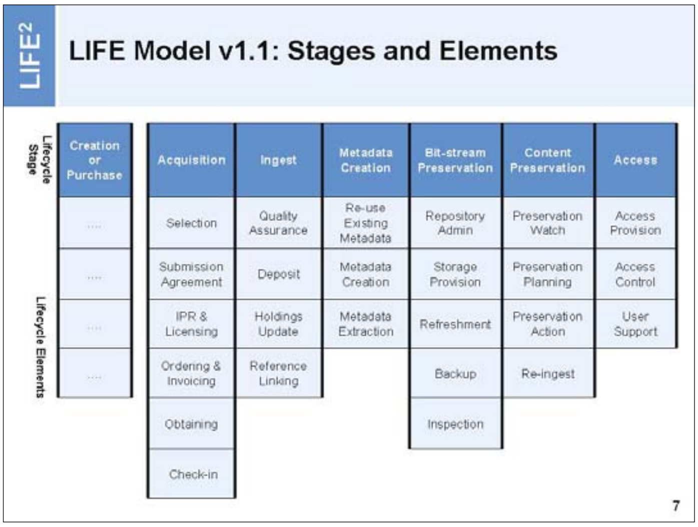
- supply and demand
  - monopolies (single owner/supplier)
  - Restrictions on use (experimental team only)
  - scarcity (due to expensive equipment)
- what the market will bear
  - Nothing to infinity?
- obligations
  - mandates to provide open (free) access

©STFC/Keith G Jeffery PARS Alliance Conference 20081104 20081104 18

Jeffery ended his presentation by concluding that while it may be possible to calculate the *costs* of preservation (see Beagrie below), it is almost impossible to calculate the *value* of the data – one can hardly put a price tag on the loss of unique data gathered from, say, an Etna eruption. Therefore, there is no market in the usual sense of the word where an equilibrium can be found for pricing. This means that there is a strong argument for public funding for the public good.

During the Q&A session following Jeffery's presentation Niamh Brennan of Trinity College, Dublin, took issue with the conclusion that the value of research data cannot be estimated. She stressed that the community must do everything in its power to demonstrate its benefits if it is to insure political support for sustainable funding. Others agreed that proving benefits must be on everybody's agenda, but it will be a difficult task.

The costs of permanent access



LIFE² LIFE Model v1.1: Stages and Elements

Lifecycle Stage	Creation or Purchase	Acquisition	Ingest	Metadata Creation	Bit-stream Preservation	Content Preservation	Access
Lifecycle Elements	....	Selection	Quality Assurance	Re-use Existing Metadata	Repository Admin	Preservation Watch	Access Provision
	....	Submission Agreement	Deposit	Metadata Creation	Storage Provision	Preservation Planning	Access Control
	....	IPR & Licensing	Holdings Update	Metadata Extraction	Refreshment	Preservation Action	User Support
	....	Ordering & Invoicing	Reference Linking		Backup	Re-ingest	
		Obtaining			Inspection		
		Check-in					

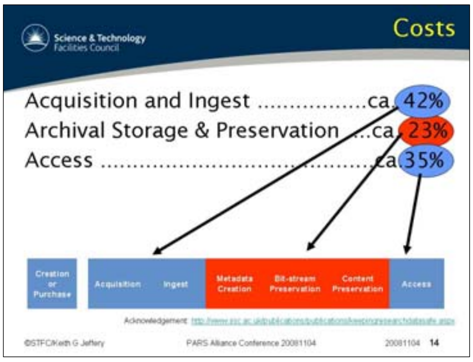
7

Conference chair Ian Halliday, President of the European Science Foundation, stressed that the research community must build up pressure to convince politicians of the need for funding permanent access. Essential to building such a case is insight into the costs involved. Jeffery quoted from the British LIFE project [3], which is in the process of developing a comprehensive costing model for digital preservation. Neil Beagrie of Charles Beagrie Ltd. took the debate a step further by reporting on a study his team recently completed for JISC, entitled 'Keeping research data safe: a cost model and guidance for UK universities' [4]. Beagrie began his presentation on a modest note, emphasizing that this was but the *first* research done on the subject and by no means the last word, if only because the scope of the study was rather limited and it focused primarily on UK Higher Education Institutions. This being said, the study does offer a number of insights that might well prove to have much broader validity.

Beagrie stated that it is of course not possible to calculate *THE* costs of digital preservation, as there is a multitude of variables involved – the price of first-mover innovation, collection levels, the type of data, their quality and quantity, the type of access that is required, etc. However, costs can be broken down to some degree and factors influencing costs can be identified. It is noteworthy that Beagrie's team calculated on the basis of Full Economic Cost (FEC), as is now required by the UK Research Councils.

Neil Beagrie

Beagrie's study indicates that preserving data is much more expensive than preserving publications, which is to be expected – as long as publications are mostly straightforward PDFs. Timeliness was identified as another important factor influencing costs: properly curating data at creation is by far the cheapest. Repairing batches of records with bad metadata, say, ten years on, is prohibitively expensive. This was confirmed by other speakers such as Erich Weichselgartner of the Institute of Psychology Information at Trier University.



Costs

Acquisition and Ingest ....ca. 42%

Archival Storage & Preservation ....ca. 23%

Access ....ca. 35%

Acknowledgement: http://www.jisc.ac.uk/publications/publications/keepingresearchdatasafe.aspx

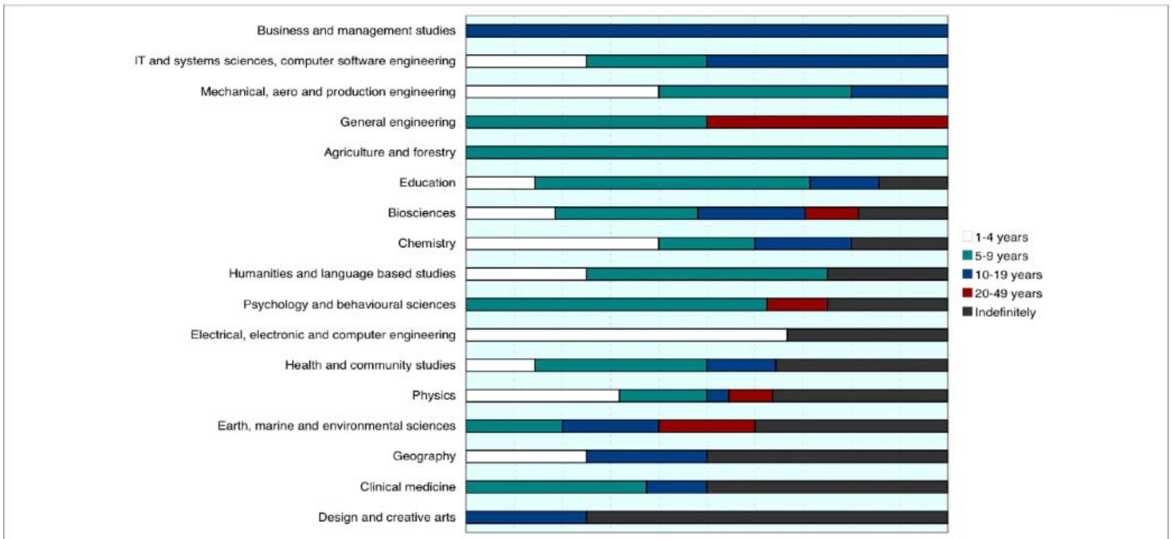
©STFC/Keith G Jeffery PARS Alliance Conference 20081104 20081104 14

Something can also be said about stages in the digital object's life cycle and concomitant costs. Data from the Archaeology Data Service (ADS, see below) suggest that most costs are incurred at the beginning of the process - at acquisition and ingest (42%), while storage and preservation accounts for only 23% of overall costs. ADS data also show that refreshment and cumulative costs decrease substantially with time. This could indicate that quite a large proportion of permanent access costs could in fact be calculated into research budgets and into routine operations – alleviating the burden on long-term special funding to some extent.

Beagrie found strong arguments in his research for economies of scale. Handling large volumes of data is much, much cheaper than handling small quantities or individual items. This caused Peter Wittenburg of the Max-Planck-Institute for Linguistics to comment that the economy of scale rule might not always apply: in the case of data on endangered languages, for instance, storage in the country of origin, even in small quantities, might in the end be more cost-effective as local expertise is vital to proper preservation.

Peter Wittenburg of the Max-Planck-Institute for Linguistics in the Netherlands

Beagrie presented results from follow-up work done by the UK Research Data Service Feasibility Study (UKRDS), which surveyed 700 researchers' perception of longer-term value of current research data holdings. The data reveal marked differences between disciplines:



Perceived longer term value of current research data holdings by discipline (Q13)

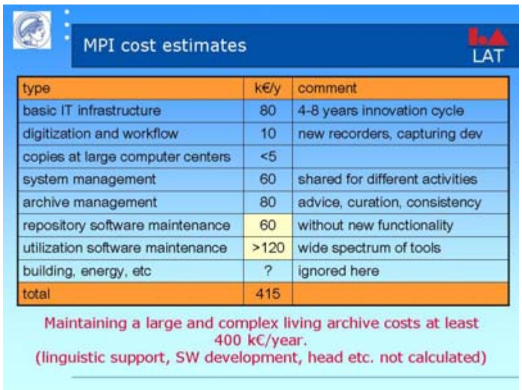
Source: Charles Beagrie survey 2008

Business models in various disciplines

Two parallel sessions highlighted examples from humanities and social sciences on the one hand, and the natural sciences on the other, to explore similarities and differences between disciplines.

Linguistics: CLARIN

The volume of data generated by linguistics pales in comparison with the volumes generated by natural sciences, said Peter Wittenburg of the Max-Planck-Institute, but terabytes are not everything: the complexity of the data must be taken into account, the all-important semantics. Wittenburg revealed that the very fragility of, e.g., unique sound recordings of endangered languages *and* the fact that these must also be regarded as important cultural treasures, prompted the community to start thinking about long-term preservation as early as 2000. As most countries prefer to archive their own cultural heritage, a single facility was not an option. The Max-Planck-Institute instead set up the technical infrastructure for an international network of data archives, depending on multiple copies and migration as preservation strategies. The MPI estimates that maintaining a large an complex living archive costs at least K€400/year.

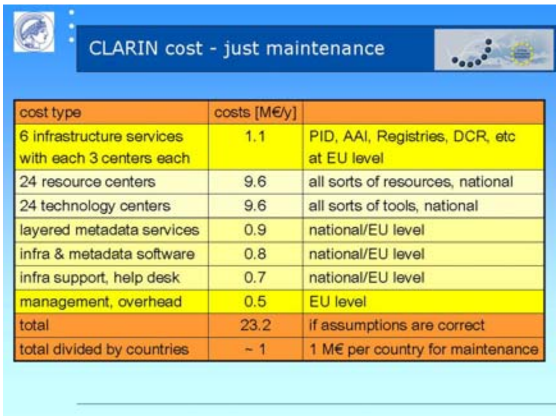


MPI cost estimates

LAT

type	k€/y	comment
basic IT infrastructure	80	4-8 years innovation cycle
digitization and workflow	10	new recorders, capturing dev
copies at large computer centers	<5	
system management	60	shared for different activities
archive management	80	advice, curation, consistency
repository software maintenance	60	without new functionality
utilization software maintenance	>120	wide spectrum of tools
building, energy, etc	?	ignored here
total	415	

Maintaining a large and complex living archive costs at least 400 k€/year.
(linguistic support, SW development, head etc. not calculated)

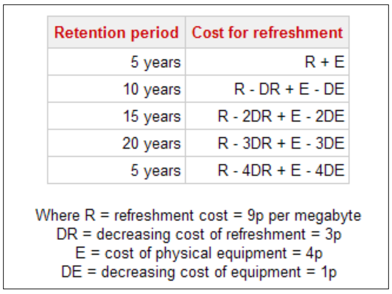


CLARIN cost - just maintenance

cost type	costs [M€/y]	
6 infrastructure services with each 3 centers each	1.1	PID, AAI, Registries, DCR, etc at EU level
24 resource centers	9.6	all sorts of resources, national
24 technology centers	9.6	all sorts of tools, national
layered metadata services	0.9	national/EU level
infra & metadata software	0.8	national/EU level
infra support, help desk	0.7	national/EU level
management, overhead	0.5	EU level
total	23.2	if assumptions are correct
total divided by countries	~ 1	1 M€ per country for maintenance

This work has now evolved into the CLARIN initiative, which is funded by the Seventh EU Framework Programme as part of the Preparatory Phase funding for the ESFRI Road Map research infrastructure projects. CLARIN is a *distributed*, heterogeneous research facility spanning at least 32 countries and 130 well-known institutes with a great variety of data facilities. The intention is to move on from accidental collaborations to a structured domain with clear responsibilties. Wittenberg presented some very provisional cost estimates. It is now expected that the cost per data centre in the system will amount to M€0.7/yearly. Some countries have already pledged resources for the initiative, while in other countries roadmap discussions are taking place and ministries are expecting the humanities as a whole to come up with joint funding proposals.

Archaeology:
The Archaeology Data Service



Retention period	Cost for refreshment
5 years	R + E
10 years	R - DR + E - DE
15 years	R - 2DR + E - 2DE
20 years	R - 3DR + E - 3DE
5 years	R - 4DR + E - 4DE

Where R = refreshment cost = 9p per megabyte
DR = decreasing cost of refreshment = 3p
E = cost of physical equipment = 4p
DE = decreasing cost of equipment = 1p

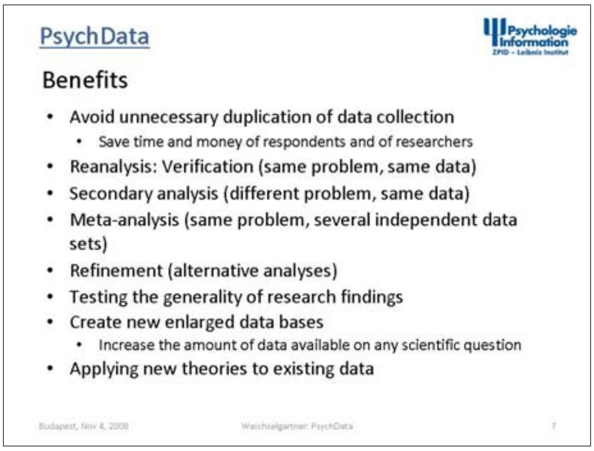
The Archaeology Data Service was one of the institutions surveyed by Beagrie's *Keeping Research Data Safe* study and it is particularly interesting in that it is one of the first such services to put a price tag on long-term archiving: a one-time, up-front fee of £0.30 per MB for storage and refreshment 'in perpetuity' – calculated at 25 years for the time being. The up-front payment model enables long-term preservation to be included in project funding. Catherine Hardman explained that this fee does not include charges for ingest and dissemination. These are calculated on a case-by-case basis depending on the type of data, the work needed on ingest (economies of scale), and the chosen dissemination option. All data are routinely made available online by the ADS, but complex access options such as integration in third-party services will come at a price. Lastly, prices may vary according to the profit or non-profit goals of the organisation archiving the data.

> ***'Use it or lose it!'***
> Catherine Hardman, ADS

At the end of her presentation Catherine Hardman stressed the need to make sure that the data stored are being used if funding is to be found. As for the sustainability of the business model Catherine Hardman admitted that there is, as yet, no guarantee as to the longevity of the ADS itself.

Psychology: PsychData, an archive for primary research data in psychology

Erich Weichselgartner of the Institute of Psychology Information at Trier University comes from a community where data sharing is still resisted by many researchers, mainly because of the sensitive (personal) data involved. And yet, it is within this very discipline that PsychData was developed by the community itself. PsychData aims at long-term, stable archiving of selected research materials, while developing tools to incorporate long-term care in the initial design of a study. Promotion of the benefits of data sharing is an explicit goal.



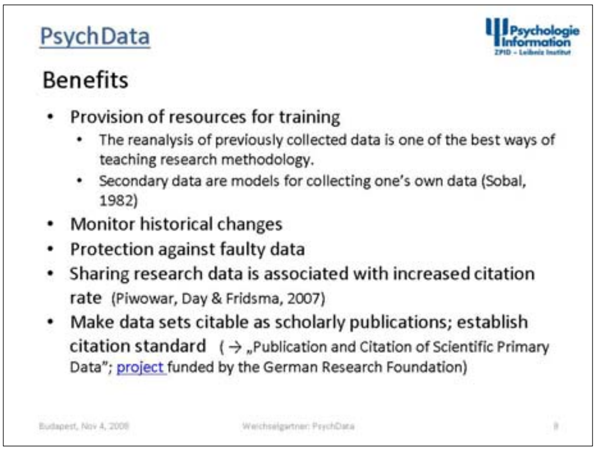
PsychData

Psychologie Information ZPID – Leibniz Institut

Benefits

- Avoid unnecessary duplication of data collection
  - Save time and money of respondents and of researchers
- Reanalysis: Verification (same problem, same data)
- Secondary analysis (different problem, same data)
- Meta-analysis (same problem, several independent data sets)
- Refinement (alternative analyses)
- Testing the generality of research findings
- Create new enlarged data bases
  - Increase the amount of data available on any scientific question
- Applying new theories to existing data

Budapest, Nov 4, 2008 Weichselgartner: PsychData 7



PsychData

Psychologie Information ZPID – Leibniz Institut

Benefits

- Provision of resources for training
  - The reanalysis of previously collected data is one of the best ways of teaching research methodology.
  - Secondary data are models for collecting one's own data (Sobal, 1982)
- Monitor historical changes
- Protection against faulty data
- Sharing research data is associated with increased citation rate (Piwowar, Day & Fridsma, 2007)
- Make data sets citable as scholarly publications; establish citation standard (→ „Publication and Citation of Scientific Primary Data"; project funded by the German Research Foundation)

Budapest, Nov 4, 2008 Weichselgartner: PsychData 8

PsychData only accepts data which are offered voluntarily, and while the metadata are freely available online, use of the primary data is restricted to research purposes; data sharing agreements must be signed to ensure protection of personal data. Although PsychData was initiated to ensure permanent access, the project itself has so far only been funded for four years by the Deutsche Forschungsgesellschaft (DFG).

Conclusions from the humanities and social sciences session

John Marks, the session's facilitator, summed up the main points made during this parallel session's presentations and discussions, including recommendations made for future Alliance action:

John Marks of the European Science Foundation

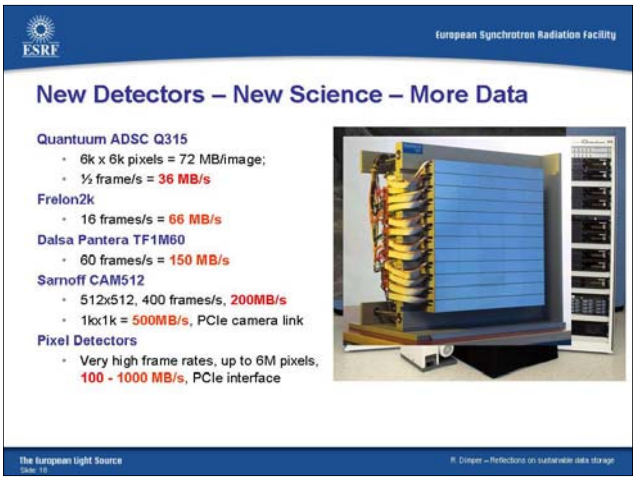
- Heterogeneity of types of data (pictures, films, texts etc)
- Digital data are disappearing faster than artefacts on which they are based (archaeology), in linguistics the opposite
- Community responsibility or national (European) responsibility for the funding?
- Tension with project related character of research funding (and often of data storage) versus need for long-term preservation, including the 'people side'
- Many data repository initiatives: need for structured networking and standardisation
- How long is long-term preservation and for which types of data sets?
- Data sharing is not generally accepted in communities (traditions and privacy); make data sets citable
- Up-front role of funding agencies in requiring deposition of data (NIH, DFG)
- Data preservation policies
- Optimisation of structure of repositories
- Development of accreditation
- Assisting in bringing together specific national data preservation and interconnecting
- Convincing governments in areas where there are no other overarching structures

- Creating global linkages (e.g., US National Science Foundation, but also others)
- Place to share experiences
- Venue to agree on decentral but harmonised action
- For future more real workshop type of meetings.

Natural sciences

Photon science: the European Synchrotron Radiation Facility

Rudolf Dimper brought home the magnitude of the term ‘data deluge’ when he described data policies at the European Synchrotron Radiation Facility (a synchrotron is a super microscope to examine condensed matter). Data production in 2007 alone amounted to a impressive 700 TB and multiplies by 300 times every 10 years. This exponential growth profoundly changes the landscape of information technology, Dimper asserted. High-speed access to networked information becomes the dominant feature of future computing and secure remote access to large-scale data sets will eventually become feasible.


ESRF

European Synchrotron Radiation Facility

New Detectors – New Science – More Data

Quantuum ADSC Q315
- 6k x 6k pixels = 72 MB/image;
- ½ frame/s = 36 MB/s

Frelon2k
- 16 frames/s = 66 MB/s

Dalsa Pantera TF1M60
- 60 frames/s = 150 MB/s

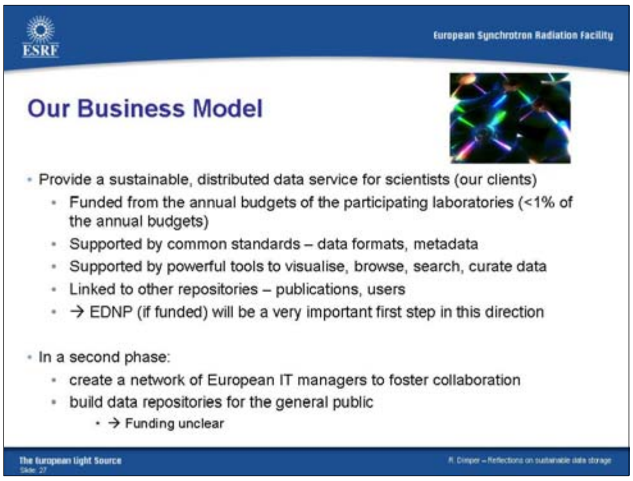
Sarnoff CAM512
- 512x512, 400 frames/s, 200MB/s
- 1kx1k = 500MB/s, PCIe camera link

Pixel Detectors
- Very high frame rates, up to 6M pixels, 100 - 1000 MB/s, PCIe interface

The European Light Source
Slide: 18

R. Dimper – Reflections on sustainable data storage

So far, however, little of this wealth of information has been kept for any length of time. According to the ESRF data policy, user data are kept for 30 days, in-house research data for 100 days, and a tape back-up for half a year, which means that after only six months all data are erased from central storage. Any long-term archiving is presently left to individual users. This practice results in dramatic data losses.


ESRF

European Synchrotron Radiation Facility

Our Business Model

- Provide a sustainable, distributed data service for scientists (our clients)
  - Funded from the annual budgets of the participating laboratories (<1% of the annual budgets)
  - Supported by common standards – data formats, metadata
  - Supported by powerful tools to visualise, browse, search, curate data
  - Linked to other repositories – publications, users
  - → EDNP (if funded) will be a very important first step in this direction
- In a second phase:
  - create a network of European IT managers to foster collaboration
  - build data repositories for the general public
    - → Funding unclear

The European Light Source
Slide: 27

R. Dimper – Reflections on sustainable data storage

To remedy this situation, a proposal for a European Data Infrastructure for Neutron and Photon Sources (EDNP) has been developed in the context of the EU Seventh Framework Programme. Within the project sustainable data storage and data management tools will be developed to allow for long-term preservation and data curation. A decision on the project is expected in January 2009.

Space physics: Lessons learnt on controlling long-term costs

Space physics have quite a long tradition in long-term data archiving, as the evolution of the Earth’s environment can only be studied by comparing many years’ worth of unique observation data. Claude Huc of the French inter-organisation PIN group (Pérennisation des informations numériques) reviewed a number of long-term preservation projects in space physics to try and extract those factors which can enable a data archive to maintain long-term activity with constant, or preferably decreasing, annual maintenance and ingest costs in the context of steady acquisition of information to be preserved and spectacular developments of data services.

Huc asserted that the experience in space physics has demonstrated that it is crucial to clearly define roles and responsibilities between data producers and archives, limiting the responsibilities of the latter. Data should be properly curated at creation according to established standards if they are to be efficiently archived for the long term.

Other recommendations include:

> ***'In the end we may have to turn around our thinking. Funders will say: Here is your budget, you decide what to keep and what to throw out.'***
>
> Ian Halliday, European Science Foundation

On archival storage:

- A storage infrastructure requires a high level of skills in data storage technology, and possession of such technology is viable only when the quantity of data archived exceeds a certain critical value
- European coordination is indispensable:
  - to develop a network of data storage facilities accessible to institutions and archives of sub-critical size
  - to create a environment within which the problems of storage on multiple sites can easily be solved in the most efficient way.

Conference venue: the elegant Gellért Hotel on the Danube in Budapest.

On software systems:

- The feasibility of generic multi-discipline software systems has been demonstrated
- The use of such systems at several sites, each for one or more disciplines, would be a major factor towards sharing and reducing the costs of maintenance and evolution
- It is clearly essential to establish coordination :
  - to prepare the development, using generic open source software, of a future system dedicated to archiving of scientific data,
  - which can be readily adapted to the needs of all such scientific data archiving centres.

On standardisation of data formats:

- In scientific fields where a standard data file format has emerged, open source software for processing, analysing, displaying and providing value-added services has developed rapidly
- Conversely, disciplines in which no such format has been adopted are heavily penalised. These disciplines should accelerate work in this area
- Strong need for international cooperation in the definition of standard data formats. This work needs strong scientific support from the community itself.

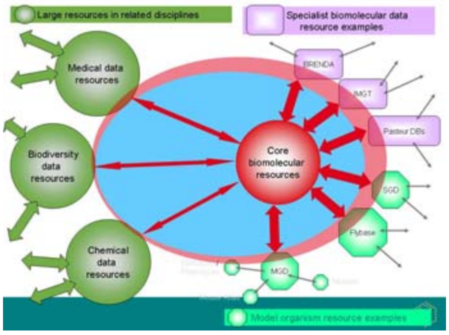
Biology: the European Bioinformatics Institute and the ELIXIR project

The European Bioinformatics Institute (EBI) of the European Molecular Biology Laboratory (EMBL) forms a hub for biological information that has an immense societal impact, e.g., on human health and medicine. Graham Cameron reported that in biomolecular research an ethos of open sharing is firmly established and global collaborations on data provision stretch over three decades.

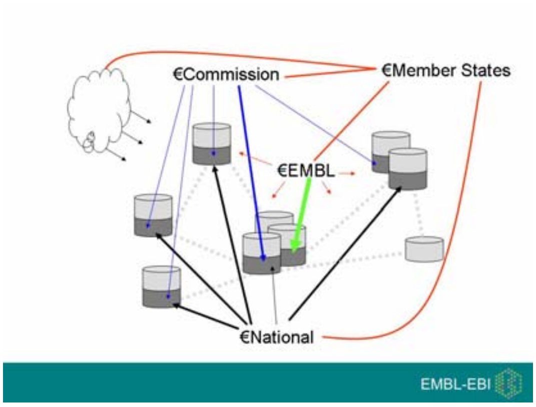
Cameron gave another twist to the debate about the costs of digital preservation by asserting that permanently preserving the data is really the *cheap* way of doing science.

In his view the bioinformatics infrastructure:

- has captured the data from several billion euros worth of science
- serves a community of perhaps a million users
- supports science on which the UK alone spends €3-4 billion a year
- cuts years of lab work down to hours of computer work
- is crucial to human well-being from medicine to agriculture
- sees data volume and usage growing exponentially
- might cost a few tens of millions (at most a couple of percent of the cost of the science it supports).


Large resources in related disciplines
Specialist biomolecular data resource examples
Medical data resources
Biodiversity data resources
Chemical data resources
Core biomolecular resources
BRENDA
IMGT
Pasteur DBs
SGD
Flybase
MGD
Model organism resource examples

Current funding structures are not well adapted to the task of preserving data, Cameron reported. Rather than being project funded, data preservation should be funded more like libraries, which enjoy much more long-term commitment from funders. Therefore, new initiatives were developed, and in May of this year, the European Commission awarded €4.5 million to a pan-European consortium to decide upon the best way to unite Europe’s biological data resources into a sustainable, integrative bioinformatics network for the life sciences. This is another Preparatory Phase project from the ESFRI Road Map list.



The European Life-science Infrastructure for Biological Information (ELIXIR) project is led by the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) and involves 32 partners from 13 countries. The consortium ultimately aims to establish a sustainably funded infrastructure for biological information in Europe, to support innovation in life science research, knowledge generation and its translation to medicine, the environment, the bio-industries and society.

An Irish Overview

Niamh Brennan, Programme Manager, Research Information Systems & Services, Ussher Library, Trinity College Library Dublin, offered the audience a general overview of developments in Ireland in the past few years. She indicated that Ireland is in the process of catching up with the rest of the world, and is keen to seek collaboration wherever possible.

During the panel discussion Niamh Brennan stressed that it must be possible to show the value of the information we preserve – especially if we succeed in offering it to users with seamless interoperability.

Conclusions from the natural sciences session

Neil Williams of the ESF, the session’s facilitator, summed up the main points made during the natural sciences session:

Neil Williams of the European Science Foundation

- Massive and rapidly increasing data flood
- Responsibility to preserve the *correct* data, for future need
- Maximising the common good
- Importance of unified policy either at (inter)national or disciplinary level
- Mandate stakeholders, e.g., funders or universities, so they have leverage to implement
- Must show benefit to users and those providing the data, so they participate and engage
- Strategic decisions are essential. Data and preservation must be included in initial decision making on scientific priorities, and their funding stabilised
- Plan expansion of preservation needs and capacity, but with a level budget
- Ensure maximum level of commonality and network archives and databases internationally and between disciplines. Maximise access.
- Access must be free and open. To charge is complex, time consuming and off-putting – it is against the premise of maximising the use of knowledge.
- Preservation is a specialist skill in its own right. Be clear, and separate who is responsible for archiving data (the provider) and preserving and organising the archive (the archivist).

Issues to address further:

- Does one size fit all – how far can commonality go?
- Reward to originator for use of data
- Some disciplines value *others'* data, some disciplines do not
- Inter-relationship of publishers and editors, open access to data, publications, researchers in archiving, science researchers etc.

Afternoon programme

How ESF is contributing to permanent access to scientific data

John Marks of the European Science Foundation outlined seven initiatives in which the ESF is presently engaged to further the cause of permanent access:

1. EUROHORCs-ESF Roadmap
2. ESF – DFG initiative on coordinated data sharing policies (of European research funding agencies)
3. EUROHORCs – NSF – ESF Working Group on Scientific Data Sharing and Interoperability
4. ESF member Forum on Research Infrastructure
5. Research Infrastructure Portal
6. European Marine Observations and Data Network (incl. Marine Observation and Data Expert Group)
7. Support to the Alliance for Permanent Access

The ESF supports a common policy on open access and permanent access to research data: ´The collection of research data is a huge investment. Permanent access to such data, if quality controlled and in intercomparable formats, allows other researchers to use them, allows re-analysis of for example long-time series and could play a role in ensuring research integrity. EUROHORCs and ESF will address how to best promote and ensure such permanent access to data generated with their funding.´

Marks concluded that the main role the ESF can play is to mobilise member organisations to support the agenda for permanent access.

Keeping the Records of Science Accessible Can we afford it?
Business models for an infrastructure for permanent access to scientific data and publications

Relations with the United States

During the Q&A session, attention was drawn to the fact that the research network is a global one, and especially the relationship with the United States should be cherished. Unfortunately, the US point of view could not be heard during the conference as Lucy Nowell of the Office of Cyberinfrastructure, US National Science Foundation, had to cancel at the last moment, because she was not well. She has, however, submitted her presentation, which is summarised below:

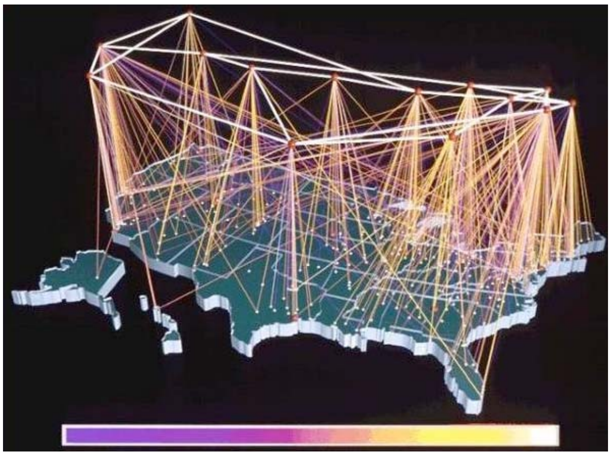
Data Preservation and Access:
a View from the US National Science Foundation

Lucy Nowell of the Office of Cyberinfrastructure (OCI), US National Science Foundation (NSF, with a budget of over six billion dollars) made her annotated powerpoint presentation available after the conference. In it, she reviewed the available policy documents stimulating sharing of data and permanent access in the United States. Lowell added another valid reason for sharing data to the ones already mentioned (by, e.g., Weichselgartner): 'Broadening participation in the scientific process is a core value of the NSF, and sharing data enables participation by those who do not have access to major observatories, sensor networks, and laboratories.'

Nowell stressed the need to choose what we keep, as we can no longer keep everything we create. This involves appraising the future value of information, which is a tremendous challenge.

'The Office of Cyberinfrastructure's first major salient in the area of data preservation and access is the program on Sustainable Digital Data Preservation and Access Network Partners, known as DataNet. There are four primary goals for DataNet:

- To provide reliable digital preservation, access, integration, and analysis capabilities for science/engineering data over decades-long timeline.
- To achieve long-term preservation and access capability in an environment of rapid technology advances.
- To create systems and services that are technologically and economically sustainable. And we mean long-term sustainability from funding by a constellation of partner investors -- not just the NSF, and
- To empower science-driven information integration capability on the foundation of a reliable data preservation network. DataNet is intended to support data collection at many scales.


DataNet: Building a network of data networks

'The greatest need for DataNet may be in medium to small size projects, especially as government agencies that fund research increase their expectations for more open sharing and systematic stewardship of data objects.

'DataNet is an ambitious and critically important exploratory initiative. DataNet Partners will be motivated by a powerful vision for multidisciplinary science. Each will integrate library and archival sciences to meet the challenges of appraisal and curation, and each will engage at the frontiers of computer and information science and cyberinfrastructure with research and development to drive the leading edge forward.

'Our approach to achieving the goals of DataNet is based upon building networks of diverse services and institutions. We plan five awards, each with sub-awardee partners, that will explore, demonstrate and understand diverse approaches to developing and providing DataNet services in sustainable ways. Each DataNet awardee will have an initial focus on several disciplinary areas but each will have active outreach to more communities and more disciplines over time.'

At the end of her presentation, Lucy Nowell referred to the international efforts in which the OCI is involved, ending in a cooperative spirit: 'We look forward to working with you.'

The role of Open Access

Norbert Kroó of the Hungarian Academy of Sciences, speaking on behalf of the European Research Council, took a broad view of the many changes the digital age has brought to science. Both he and the next speaker, John Smith of the European Universities Association, expressed strong support for the cause of Open Access to scientific information. This prompted Eefke Smit of the STM Publishers Association (which is also a member of the Alliance for

Norbert Kroó (left) and John Smith (middle) during the panel discussion, with Neil Beagrie between them. Next to Smith John Marks and Claude Huc.

Permanent Access) to mention the growing number of publishers who are offering open access models to their authors and to speak up in favour of the freedom of researchers to publish where they want, a value that publishers offer over certain other publishing arrangements of a more mandatory nature. All agreed, however, that a constructive dialogue is now ongoing which includes the experimental investigation of various models.

On a slightly critical note: as important as Open Access is to the community, in and of itself it seems to have little bearing on business models for *permanent* access. Whether the user or the author pays for information, it is clear that neither are likely long-term custodians of digital data.

Towards a global e-science infrastructure: are repositories enough?

László Kovács of MTA SZTAKI, the Computer and Automation Research Institute of the Hungarian Academy of Sciences, raised a number of fundamental issues with regard to preservation of digital data in a scientific environment. He argued that science itself undergoes a fundamental paradigm shift every twenty to thirty years, and as preservation of semantics and semantic tagging is still a serious research issue, the idea of long-term preservation is seriously challenged.

In line with other speakers earlier in the day, Kovács showed that those who have to do the work on permanent access (e.g., the authors), seldom reap the benefits. He argued that for the system to become sustainable a more balanced distribution of work and benefits must be found at both the individual, institutional and community level.

László Kovács

Kovács called upon the Alliance to revisit its early decision to deal only with public research, arguing that industrial research is such a major stakeholder in any infrastructure, that it cannot be left out. Such inclusion would, however, have consequences for the organisation of the infrastructure, as the value systems of classic academic science and industrial sciences differ. Commercial companies quite often cannot afford long-term preservation: unless obliged by law to keep data (such as in pharmaceutics), most data are discarded within two years after creation.

In response to his own question in the title of the presentation, Kovács was of the opinion that simply establishing repositories is *not* enough. A complex, sophisticated information infrastructure is to be built with repositories, value-added service providers and registries. Different business and activity models might coexist in the same infrastructure, thus a neutral approach towards the diversity of business models is required. Clarification of the cross-border market between academic science and industrial science is needed.

Peter Tindemans, acting director of the Alliance, commented that the Alliance might well consider inclusion of industrial research data, but only if and when such data are made publicly available.

Current initiatives by the European Commission

The last speaker of the day was by no means the least: Patricia Manson of Directorate General Information Society and Media of the European Commission had come to Budapest to underline the Commission's commitment to the cause of sustainability of digital information. She said she was struck by how the issues had moved on from last year's conference. Although the challenges are much the same (in fact, she used last year's slide to illustrate this), Manson reported that political awareness has grown since then. She referred to the i2010 Digital Libraries initiative of the EU, stressing that libraries should take a role in developing the infrastructure.

Patricia Manson

Patricia Manson made the following observations:

- Digital preservation should be promoted not as a drain on funding but as an *investment* in future research.
- Europe will invest in developing a skills base for digital preservation (the European Masters initiative)
- Responsibilities between the many stakeholders must be more clearly defined.
- A 'chaos' of repositories has come into being, and there are many questions with regard to business models, interoperability, trust and persistence. All of these issues require study, including persistent identifiers and certification of repositories.
- The community needs progress from digital preservation research. There is a great need for scalable, cross-domain solutions.
- The i2010 Digital Libraries programme will be funded by M€69. A substantial portion of this could go to digital preservation programmes. The community, according to Manson, 'needs to exploit commonalities to improve attractiveness of the field for sustainable funding.'

In response to Patricia Manson's presentation, John Smith of the EUA mentioned the complexity of Brussels: there are so many initiatives that it is difficult to find one's way. Patricia Manson agreed that the EU must work on concerting its actions, but she also called on the Alliance for Permanent Access to bring stakeholders together in a joint agenda for permanent access.

A panel discussion with (almost) all speakers concluded the day: from left to right Norbert Kroó, Neil Beagrie, John Smith, John Marks, Claude Huc, László Kovács, Niamh Brennan, Catherine Hardman, Rudolf Dimper, Graham Cameron, Keith Jeffery and Neil Williams. Time constraints (airplanes would not wait) necessitated panel members to get right down to the very essence of their conclusions.

Conclusions: a way forward for the Alliance

After a well attended and fruitful conference, the following picture emerges for Alliance action:

- The Alliance's major value lies in its ability to bring all the stakeholders together and act as an umbrella organisation.

- The Alliance has a major task in the area of advocacy, promoting the cause of working together for the common good rather than individually, and raising awareness of digital preservation issues with governments, the EU and funding agencies. In this way the Alliance will contribute also to increased alignment and coordination between governments, the different branches of the EU and the funding agencies.

- One role for the Alliance is to generate consensus on an optimal infrastructure of repositories and archives. A suggestion worth investigating is that connecting archives in a network could provide an upstream incentive for international collaboration to cut costs, as not all repositories need to include long-term preservation facilities.

- The advocacy role of the Alliance must involve promoting the value of preservation for the users, whether scientists or societal bodies or companies. Identifying and maybe quantifying the value is as much needed as further investigation of the costs. Funding models must reflect both, and the Alliance should continue working with the funding agencies to develop such models.

- The Alliance has a very useful role to play in ensuring that common provisions and facilities become available. To facilitate data sharing and seamless interoperability work must be done on registries, terminology and standards. The Alliance is poised to take a coordinating role here. Other examples are a European solution for accreditation of repositories and archives, as well as agreement on persistent identifiers.

- Clearly the Alliance is well placed to work with the various stakeholders on matters of policy. These range from policy within the scientific community to the policies of governments. Sharing data and fair use is an example of the first category. This must somehow become part and parcel of a researcher's workflow, stimulated not only by funding agencies but also by other, less tangible rewards such as citations. As an example of the second category the Alliance must make a case in Brussels for less restrictive Intellectual Property Rights regimes.

- Cross-community exchanges of experiences is a very valuable field of action for the Alliance, as the Budapest conference has shown. For example, it was demonstrated that when it comes to permanent access, the humanities and social sciences on the one hand and the natural sciences on the either share much the same problems, only the scale differs.

- European solutions must be part of, and therefore be designed with a perspective to, global solutions. Here the Alliance is in a good position to work with key stakeholders elsewhere in the world, such as the National Science Foundation (NSF) in the United States and organisations in Asia and Australasia.

PS from the chair: 'Let's not get *too* sustainable'

At the end the day it was the chair's prerogative to speak the last word. Ian Halliday said that building cooperation takes a lot of time, especially when the stakeholders are so heterogeneous. He showed himself much impressed by the progress made, but also struck a sobering note when he warned his audience that the need for 'sustainability' must not be turned into a bureacratic stumbling block. He challenged the Alliance and the scientific community to think less in terms of 'we should', and more in terms of 'we *do* – today'.

References

The slides were taken from the respective presentations. Powerpoints, abstracts and authors' biographies can be accessed at <http://www.alliancepermanentaccess.eu/index.php?id=3> . Photographs Inge Angevaare.

[1] See Courant, Paul N. (2008) *Some economics of digital preservation.* In: LIFE2 Conference, June 23, 2008, British Library, London, UK, <http://eprints.ucl.ac.uk/9051/>. More on Courant's presentation at <http://www.alliancepermanentaccess.eu/item.php?id=27>

[2] 'To share or not to share: publication and quality assurance of research data outputs', commissioned by the UK Research Information Network (RIN), <http://www.rin.ak.uc/data-publication>. More on the study at <http://www.alliancepermanentaccess.eu/item.php?id=29>.

[3] Ayris, P., R. Davies, R. McLeod, R. Miao, H. Shenton and P. Wheatley (2008), *Lifecycle information for E-literature: an introduction to the second phase of the LIFE project.* In: LIFE2 Conference, 23 June, 2008 British Library, London UK, <http://eprints.ucl.ac.uk/8608/1/8608.pdf>

[4] Beagrie, Neil, Julia Chruszcz and Brian Lavoie (2008), 'Keeping Research Data Safe: a cost model and guidance for UK universities', final report and executive summary at <http://www.jisc.ac.uk/publications/publications/keepingresearchdatasafe.aspx>.

Websites referred to in the text

ADS, Archaeology Data Service, <http://ads.ahds.ac.uk/>
Alliance for Permanent Access, <http://www.alliancepermanentaccess.eu>
CLARIN, <http://www.clarin.eu/ >
DataNet, <http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141&org=OCI&from=home>
FEC, Full economic cost, <http://www.rcuk.ac.uk/aboutrcs/funding/dual/fec.htm>.
ELIXIR, <http://www.elixir-europe.org/page.php?page=home>
Netherlands Coalition for Digital Preservation (NCDD), <http://www.ncdd.nl/en/>
PIN, Pérennisation des informations numériques, <http://vds.cnes.fr/pin>
PsychData, <http://www.zpid.de/index.php?wahl=products&uwahl=frei&uuwahl=psychdatainfo&lang=EN>
UKRDS, <http://www.ukrds.ac.uk >

Some eighty representatives of research institutions, digital archives, libraries, funding agencies and publishers from all over Europe (and one from the United States) joined the 2008 Alliance debate on business models for permanent access to the records of science.