View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by CERN Document Server

An update on electronic records at CERN (internal developments, collaboration and outsourcing)

Anita Hollier, CERN

Presented at the 'Future Proof IV' international scientific archives conference, Stockholm, April 2008

Abstract

Despite a lack of action in several important areas (recommendations made by CERN's working groups on electronic archiving, about e-mail management in particular, have not been adopted), some progress is being made. Interest from the research community in the long-term preservation of scientific data has been growing with the imminent start-up of the Large Hadron Collider. CERN has consequently joined the Alliance for Permanent Access, and will contribute to the EU-funded PARSE.Insight project. Use of CERN's digital repositories is also growing, and a facilitated self-assessment using the DRAMBORA (Digital Repository Audit Method Based on Risk Assessment) toolkit was carried out under the DELOS Pilot Digital Library Audit Programme. Archiving of CERN's Web pages has been outsourced for a trial period, and the captured pages will be made available via the European Archive.

Introduction

The European Organization for Nuclear Research (CERN) was founded in 1954 in Geneva, Switzerland. By 1959 CERN had built what was then the highest-energy particle accelerator in the world, and today it is the world's largest particle physics laboratory. The original 12 member states have increased to 20,[1] plus 8 observers.[2] The purposes of CERN are clearly stated in its Convention: "The Organization shall provide for collaboration among European States in nuclear research of a pure scientific and fundamental character, and in research essentially related thereto. The Organization shall have no concern with work for military requirements and the results of its experimental and theoretical work shall be published or otherwise made generally available."

Part of the vision in 1954 was that science would bring nations together, and around 8,000 physicists each year come from all over the world to use CERN's facilities. A further 2,500 people are employed on the CERN site, many of them currently engaged in the commissioning of the Large Hadron Collider (LHC), which is planned to circulate its first beams in mid-2008. CERN's Archive was set up in 1979, as part of the organisation's 25th anniversary celebrations, to support the writing of the "History of CERN". The Archive's scope has broadened since then and it now holds around one linear kilometre of records, including documents produced by the CERN Council and its subordinate Committees, files of previous Directors General and other senior staff, and records documenting the work of CERN Divisions and selected Experiments and

CERN-OPEN-2008-011
29/04/2008

[1] The current Member States are: Austria, Belgium, Bulgaria, the Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Italy, The Netherlands, Norway, Poland, Portugal, the Slovak Republic, Spain, Sweden, Switzerland and the United Kingdom.

[2] The European Commission, India, Israel, Japan, the Russian Federation, Turkey, UNESCO and the USA.

Committees. So far almost all accessions to the Archive have been on paper, and long-term preservation of the increasingly large number of electronic records produced at CERN has not been properly addressed.

Although CERN is a scientific research centre with a high level of computing know-how, the long-term preservation of digital records, whether scientific or administrative, has not been a high priority. This is not so unusual: as the UK's Digital Curation Centre observes, "The scientific record and the documentary heritage created in digital form are at risk from technology obsolescence, from the fragility of digital media, and from lack of the basics of good practice, such as adequate documentation for the data."[3] To begin addressing this issue at CERN a working group on long-term electronic archiving (LTEA) was set up in 1997 to look into "electronic documents at CERN" (scientific data were not included in its remit). Its report in March 2000 included an overview of existing CERN computer systems examined by the working group, criteria of a "Certified Information System", some weaknesses found in existing systems, and recommendations. Appendices included a draft Operational Circular on Electronic Records (these circulars are the organisation's internal 'laws'), and examples of electronic records management outside CERN. The second group, set up in June 2000 with a mandate to produce a set of implementable solutions, comprised mainly IT specialists. The group's report, which was presented to CERN's Director General in September 2001, was deliberately limited in scope. It conceded that 'it is premature to look for a general solution for LTEA', but pressed for urgent action in certain areas, in particular to implement a proposed in-house system for e-mail, to investigate Web archiving options (probably outsourced), and to develop a document handling policy. One of the main requirements of this policy would have been the use of one of CERN's electronic document management systems for certain records. This would not guarantee long-term preservation, but would at least keep the documents safe in the medium term and would be an improvement on the prevailing practice where important records had been found to be stored on the Web, on diskettes, on individuals' hard drives, and so forth.

Although the groups' findings were well received by CERN's top management, no concrete actions resulted. Top-level support is essential for systematic progress in this area. Vague approval is not enough to change organisational culture or to encourage an already overcommitted IT department to invest scarce resources in long-term projects. Unfortunately, most of the IT specialists who formed part of the second working group, and who would have been valuable advocates with their colleagues, have since left CERN. Efforts to obtain explicit management support are continuing, and as background to these discussions a survey was carried out in 2007 on e-mail archiving practices in some similar institutions (international organisations, scientific research centres, etc) as well as examples of best practice elsewhere. This showed that several organisations were at a similar stage to CERN – aware of the problem but not yet ready with a solution. Many national archives, of course, have implemented solutions and made information and tools freely available on the Internet. And some organisations have recognised e-mail as a preferred means of communication and taken the necessary steps to ensure its good management. All this information was summarised in a report along with a brief

[3] http://www.dcc.ac.uk/about/ (accessed March 2008).

description of the issues involved in the long-term preservation of e-mail. Hopefully, this will be a useful aide-memoire in discussions with management, focusing on just one area but broadening the view to take more account of what is happening outside CERN. In the meantime, however, it seems more fruitful to concentrate on finding other approaches – in particular looking for collaborations or solutions outside CERN – and making at least piecemeal progress this way.

Data archiving and membership of the European Alliance for Permanent Access

It is always easier to capitalise on existing areas of interest rather than trying to raise awareness of a poorly understood problem, and at the moment interest from CERN's research community is growing in an area that had been specifically excluded from the mandate of the two working groups: the long-term preservation of scientific data. This is becoming an area of particular concern as construction of the new particle accelerator nears completion. The Large Hadron Collider will collide two counter-rotating beams of protons or heavy ions at an energy of 7 TeV[4] per beam, producing around 15 petabytes of data annually (enough to fill 100,000 DVDs a year). Around 40 million events are expected per second; of these, the 200 most interesting events each second will be selected on-the-fly to write to tape. The data are then converted for analysis. Thousands of scientists around the world will want to access and analyse this data, so CERN is building a distributed computing and data storage infrastructure: the LHC Computing Grid (LCG). However, the question of longer-term preservation also needs to be addressed to allow future re-use of the data, for example, by researchers working on similar experiments, wanting to reinterpret the data, or testing new ideas. There is growing recognition that this issue cannot be left until "afterwards". Attempts at CERN to preserve data collected from a former accelerator, the Large Electron-Positron (LEP) collider, have run into precisely this difficulty, and the longer progress is stalled the harder it will be to find people who worked on these experiments and have the necessary knowledge and expertise. These concerns formed part of the background to CERN's decision to join the European Alliance for Permanent Access.

The Alliance was initiated by the National Library of the Netherlands (Koninklijke Bibliotheek) and "aims to develop a shared vision and framework for a sustainable organisational infrastructure for permanent access to scientific information."[5] Key objectives and benefits include aligning and enhancing permanent information infrastructures in Europe, building collaboration and relationships, joint advocacy and representation, and increasing impact and mass.[6] Alliance members have proposed, amongst other things, the PARSE.Insight project (Insight into issues of Permanent

[4] One tera-electronvolt (TeV) is roughly equivalent to the energy of a flying mosquito; but in this case it is squeezed into a much smaller space, as a proton is about a trillion times smaller than a mosquito.

[5] Its members include the European Science Foundation; European Space Agency; CERN; Max Planck Gesellschaft; Centre National d'Etudes Spatiales; Science and Technology Facilities Council; The British Library; Koninklijke Bibliotheek (Netherlands); Deutsche Nationalbibliothek; Joint Information Systems Committee; International Association of Scientific, Technical and Medical Publishers; National Archives of Sweden; Centre Informatique National de l'Enseignement Superieur; Digital Preservation Coalition; NESTOR; Perennisation des Informations Numeriques; and the Netherlands National Coalition for Digital Preservation.

[6] http://www.alliancepermanentaccess.eu/ (accessed March 2008).

Access to the Records of SciencE), which will be supported by the Seventh Framework Programme (FP7) e-Infrastructures initiative, and seeks to provide:

- A roadmap for a support e-infrastructure for long-term accessibility of scientific and other digital information in Europe,
- Insight into current and planned research,
- Identification of gaps in the existing and planned infrastructure,
- The ability to share best practice,
- Better-informed investment decisions, and
- An international process for evaluating the trustworthiness of digital repositories.

CERN's main role concerns the community insight objective: carrying out a case study on issues of preservation, re-use and (open) access to High-Energy Physics data. This is a twofold problem, technical on one side and sociological on the other. The intention is to expose the sociological issues, sparking a debate about data ownership, (open) access, credit, accountability, reproducibility of results, and depth of peer-reviewing.

There are no insurmountable technical problems involved in the preservation of CERN's LHC data (though, as always, finding the necessary extra resources for this activity is unlikely to be straightforward); the difficulty has more to do with the complexity of the High-Energy Physics data model. HEP data does not have the same tradition of reuse that one finds in fields like astronomy or climate science. Exploiting the data depends on understanding all the contextual information about how the analysis was carried out, which is why the American Institute of Physics Study of Multi-Institutional Collaborations "agreed that technical data (especially the raw data) had virtually no value after its use by the collaboration,"[7] and said that use of data summary tapes "probably requires accessibility to collaboration members as well as documented software".[8] However, re-use of data is possible and does sometimes occur; one example is the re-analysis of data collected in the 1980s at DESY[9] together with data collected in the late 1990s at CERN in the light of improved theories.[10] This combined analysis was only possible because researchers from the JADE collaboration (DESY) had moved to OPAL (CERN) and had kept tapes of data and corresponding software.

One solution for long-term access could be to produce a "parallel" format for preservation and re-use in addition to the ones used internally by the experiments.[11] These would be high-level objects with the necessary additional knowledge embedded in them to produce a format understandable, and therefore usable, by researchers other than those involved in the original experimental collaboration. The example above shows that such an approach is feasible, but in this and similar cases the people able to provide the necessary inside knowledge were immediately available and had a strong incentive to prioritise the work. It could be harder to achieve if it were seen as an additional "archival"

---

[7] Phase I: High Energy Physics; Report No. 2, Part A, Section X-C 'Appraisal Guidelines'; http://www.aip.org/history/pubs/collabs/phase1rep2.htm#A10C (accessed March 2008).
[8] Ibid. Part D, Section V-E.
[9] Deutsches Elektronen-Synchrotron, Germany.
[10] The JADE and OPAL collaborations, Eur.Phys.J.C17 (2000) 19, arXiv:hep-ex/0001055
[11] More information is given in the presentation by Jos Engelen, CERN's Chief Scientific Officer, to the Alliance meeting on permanent access to the records of science held in Brussels, 15 November 2007, available here: http://www.alliancepermanentaccess.eu/index.php?id=7 (accessed March 2008).

activity in competition with more urgent research work. It is not clear exactly how much additional work would be required, and therefore how much this could cost; but even if it requires only a fraction of the human and financial capital invested in HEP experiments, a small fraction of a large number (thousands of person-years) still translates into a major project. Scientists would need enormous academic incentives or additional funds.

Quite apart from encouraging discussion of these issues within CERN, participation in an external project of this kind has the advantage of emphasising that such problems are not CERN-specific. Of course, specific solutions for HEP data must come from within this field, but it can still be useful to set them in the context of the experience and expertise on long-term preservation and access in general available outside CERN. Linking with a high-profile international collaboration facilitates the investigation of these challenges at CERN and increases the likelihood of the necessary actions being taken to solve them. Issues of preservation of, and access to, data can also usefully be tied in to another movement in which CERN is already heavily involved: the Open Access movement,[12] which promotes free access to scientific literature. It is possible that a culture shift may gradually lead to a similar approach for data, for example, publishing high-level data objects behind each scientific article. This is already done in some other fields, but the details of ownership and other practicalities (Voluntary or compulsory? After a time lapse?) are still to be worked out for HEP experimental collaborations.

Trustworthiness of digital repositories and the DRAMBORA self-audit

The Open Access (OA) connection is also useful for promoting long-term preservation issues in connection with CERN's digital repositories, since one aspect of OA involves the deposit of articles in "at least one online repository that is supported by an academic institution, scholarly society, government agency, or other well-established organization that seeks to enable open access, unrestricted distribution, interoperability, and long-term archiving".[13] CERN is currently collaborating with other HEP research centres (SLAC, Fermilab and DESY) to build a joint information database combining the best features of the existing systems: SLAC SPIRES and CDS (CERN Document Server). This will act as a subject repository for preservation of HEP scientific information, and in the future could include high-level data as well as documents. Promoting the use of CERN's electronic document management systems, including CDS, was one of the recommendations of CERN's second working group on long-term electronic archiving (LTEA), but the group stressed the need for improvements in these systems to increase their acceptability for users and to improve LTEA reliability.

Long-term preservation is often a neglected area of digital information management, faced with the immediate challenges of providing a good service to users. The emerging discipline of data curation may be seen as a response to this, and various criteria and methodologies are being developed to help identify and measure the trustworthiness of

[12] Open Access (OA) literature is digital, online, free of charge to the reader, and free of most copyright and licensing restrictions. For more on CERN's action on OA see: http://open-access.web.cern.ch/Open-Access/ (accessed March 2008).

[13] Bethesda Statement on Open Access Publishing, 20 June 2003, http://www.earlham.edu/~peters/fos/bethesda.htm (accessed March 2008).

digital repositories.[14] One example is DRAMBORA (Digital Repository Audit Method Based on Risk Assessment), jointly funded by the Digital Curation Centre (DCC) and Digital Preservation Europe (DPE). "DRAMBORA characterises digital curation as a risk management activity. The DRAMBORA toolkit provides a metric to enable an auditor to establish the organisational context and goals of a repository and then to assess how it is achieving these in terms of risk. Risk is used as a metric: it can be expressed quantitatively, thereby supporting comparisons across repositories and over time within a repository."[15]

DRAMBORA is designed to be a self-audit, but in 2007 CERN participated in a facilitated self-assessment carried out as part of the DELOS[16] Pilot Digital Library Audit Programme. This involved two members of the assessment team (including one of the DRAMBORA authors) spending three days at CERN to conduct a series of meetings with relevant members of the library, archive and IT staff. The main stages of the audit were:

- Structured and documented organisational self-awareness (identify policy and regulatory framework, activities, assets, etc),
- Identify, analyse and assess risks,
- Consider risk management strategies.

Considerable preparation was required in order to support the process and ensure it went as smoothly and efficiently as possible. The authors of DRAMBORA estimate the required effort for a self-assessment at 24-30 hours, though they stress that "this may vary, occasionally substantially". In our case it took longer. The most fundamental challenge was to ensure buy-in from key staff whose participation and knowledge were essential to the success of the exercise. These issues, as mentioned above, tend not to be regarded as a priority given the many other more pressing issues that they face. Another challenge was to locate adequate documentation about the repository and the parent organisation. This was necessary to enable the audit team to understand the repository, of course, but it is also a fundamental part of the DRAMBORA approach. One of the authors' criticisms of existing methods of assessing trustworthiness of digital repositories is that they place too little emphasis on evidence in the auditing process. It quickly became apparent that quite a lot of things at CERN happen thanks to goodwill and "gentlemen's agreements" rather than written mandates.

The assistance of the DRAMBORA auditors was of particular value in assessing and quantifying potential risks, which was quite a challenging activity. Having identified the repository's activities and assets, the vulnerabilities associated with them are characterised as risks. The toolkit includes tables to describe and assess each risk in terms of its probability, its potential impact, and its relationship with other risks, and provides a

---

[14] For example:
Trusted Digital Repositories: Attributes and Responsibilities (An RLG-OCLC Report)
http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf (accessed March 2008).
The nestor Catalogue of Criteria for Trusted Digital Repositories
http://www.langzeitarchivierung.de/modules.php?op=modload&name=PagEd&file=index&page_id=18
(accessed March 2008).

[15] http://www.repositoryaudit.eu/ (accessed March 2008).

[16] DELOS is a Network of Excellence on Digital Libraries: http://www.delos.info/ (accessed March 2008).

scoring system to help with this; nonetheless, it proved quite hard to reach agreement on quantifying the risks, or even to identify them consistently. This is an area where it is particularly important to have full buy-in from all relevant stakeholders, otherwise brainstorming potential areas of vulnerability is apt to be misinterpreted as unjustified criticism! Of course, it is difficult for outsiders to fully understand an organisation's system in such a short period of time, but the involvement of external partners was of very great benefit in raising awareness of long-term digital preservation issues and putting CERN's experience into a broader perspective. A summary of the audit process and findings will be published in the DELOS Pilot Digital Library Audit Programme report. An individual risk register was also compiled for each of the repositories that participated in the project. These are tools for internal use by the organisation concerned, and may be updated to take account of changes in risk management strategies, etc, thus allowing comparisons over time.

Web archiving

Another of the recommendations of CERN's second working group on long-term electronic archiving (LTEA) in 2001 was to investigate Web archiving options. It was felt that outsourcing would probably be the best route, but meetings held with companies offering such services at that time were unsatisfactory as they were unable to offer all the functionality required. Web content constitutes an increasingly important part of the records of most organisations, but it has particular significance at CERN because this is where the Web was born. The first proposal for the World Wide Web, showing how information could be transferred easily over the Internet using hypertext, was made at CERN by Tim Berners-Lee in 1989, and was further refined by him and Robert Cailliau in 1990. This made it seem all the more shameful for CERN to rely on the Internet Archive[17] for the only archiving that was carried out of its Web pages (not counting routine back-ups for operational purposes, of course).

The aim of the Internet Archive is breadth rather than depth, so coverage of any individual organisation is necessarily incomplete. An obvious first approach to improve Web archiving at CERN was to contact Archive-It,[18] the Internet Archive's subscription service that allows institutions to build and preserve their own Web archives with whatever depth and frequency they wish. It seemed worthwhile to liberate funds from the Archive section budget (mainly by cutting back on a digitization project) in order to make some progress here, as this is exactly the sort of service CERN's working groups were looking for but did not find in 2001 (Archive-It was developed four years later). However, another opportunity arose in September 2006 when the European Archive Foundation was officially launched and began collecting and making freely accessible public domain collections and large Web archives via a multilingual website.[19] As CERN is a European organisation (even if its users come from all over the world) it seemed very appropriate to collaborate with them. A commercial Web archiving service particularly suited to the needs of businesses and other organisations is offered by Hanzo Archives,[20]

[17] http://www.archive.org/index.php (accessed March 2008).
[18] http://www.archive-it.org/ (accessed March 2008).
[19] http://www.europarchive.org/ (accessed March 2008).
[20] http://www.hanzoarchives.com/ (accessed March 2008).

who work closely with the European Archive and who are also the Web archiving business partner for the Europe-wide (FP7-funded) Living Web Archives project.[21]

Hanzo Archives began crawling the CERN Web sites in 2007, harvesting publicly accessible pages across the whole CERN domain plus one 'hop' outside. Quarterly crawls will be made for all sites, with more frequent crawls for fast-moving content, e.g. news pages. The archived pages will be made available on the European Archive and Internet Archive. Some difficulties were encountered due to the sheer size and complexity of CERN's Web sites, but once again the main difficulties within CERN were not so much technical as financial and sociological. There was some perceived embarrassment that CERN – "where the Web was born" – should outsource its Web archiving. On the other hand, waiting patiently for this to become an in-house priority did not seem a good option either. In any case, it provided a good opportunity for more discussion about the issues involved in long-term digital preservation: in the IT world "archiving" is too often taken to mean "moving off-line" or "backing up". There are some benefits in a specialist task, like web archiving, being undertaken by specialists. Another advantage of Hanzo Archives is that they offer a range of solutions, allowing scalability and flexibility. If the trial period proves successful we could choose to continue with our current service, or the licensed software could be taken in-house and run by our IT service, if they wished.

Conclusion

The lack of progress following submission of the working groups' reports was disappointing (though perhaps not surprising, as CERN continues to grow scientifically while shrinking in terms of resources). The momentum that had been built up, with IT staff ready to implement ideas, was lost. As more time passed, most of these experts left CERN, so their knowledge and support as "champions for the cause" was lost too. It is important to keep trying to enlist high-level support, particularly since CERN's top management changes fairly frequently; but this should perhaps be done in a way that shifts the focus more outside CERN, towards all the developments that are underway in the fast-moving emerging discipline of digital curation and preservation.

The problems of digital preservation are not primarily technical, but financial, political and cultural. It is all a question of priorities. The high level of IT skill at CERN is not necessarily a direct advantage, as it can discourage collaboration. CERN works with many partners, of course, for key projects like the LHC Computing Grid, but for a "peripheral" area like long-term preservation it is less likely to pay sufficient attention to developments outside. An organisation without these IT skills might have to go to the acknowledged experts in the field if they want to make progress, but at CERN the attitude tends to be: there is no need – we are perfectly capable of doing whatever is required. However, other factors generally mean that this capability is not translated into action.

The involvement of external partners is of huge benefit in bringing new expertise and broader experience in the field of digital preservation, as well as raising the profile of the

[21] http://www.liwa-project.eu/ (accessed March 2008).

issues involved. And CERN, partly thanks to the high level of IT skill mentioned above, is in the fortunate position of being attractive to various partners. The LHC will produce huge quantities of complex scientific data, which makes it an interesting case study for a project like PARSE.Insight. The size and importance of CERN's digital repository means it is of interest in the context of a pilot programme aimed at the development of audit and certification mechanisms for digital libraries. Archiving the birthplace of the World Wide Web is good publicity for a Web archiving company. We are lucky to have been able to take advantage of these opportunities.

Acknowledgements: The main workers at CERN on the PARSE.Insight project are Salvatore Mele and Peter Igo-Kemenes; and for the DRAMBORA audit, Tullio Basaglia, Tim Smith, Jean-Yves Le Meur, Tibor Simko and Nick Robinson. I would like to thank Jens Vigen, Salvatore Mele and Tullio Basaglia for their comments on this text.