Presentation at the 'Future Proof III' international scientific archives conference, Strasbourg, April 2006 - E-projects at CERN

Anita Hollier, CERN

The European Organization for Nuclear Research (known as CERN) was established in 1954 in Geneva, Switzerland. By 1959 it 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 eight observers.[2] CERN's aims 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.'[3]

Part of the vision in 1954 was that science would bring nations together, and around 6,500 physicists come from all over the world each year to use CERN's facilities, which include a 27km circular particle accelerator. A further 3,000 people are employed on the CERN site, many of them currently engaged in the construction of the new Large Hadron Collider (LHC), which is due for completion in 2007.

The LHC is designed to collide two counter-rotating beams of protons or heavy ions at an energy of 7 TeV per beam.[4] Its construction is a major feat of physics and engineering, but also a huge IT challenge. The new accelerator will produce roughly 15 petabytes of data annually, which will be analysed by thousands of scientists around the world. CERN IT specialists are currently helping to develop a Grid-based data storage and analysis infrastructure to cope with this, and on 15 February 2006 the Worldwide LHC Computing Grid collaboration (WLCG) officially announced the successful completion of its latest service challenge. This involved sustaining a continuous flow of physics data from CERN to 12 major computer centres in Europe, Asia and America on a worldwide Grid infrastructure at up to 1 gigabyte per second.[5]

This high level of computing know-how made CERN a natural site-visit for experts from the Swiss Federal Archives when they were researching their e-Government project in 2001.[6] However, they were disappointed to find that current and future IT achievements do not necessarily imply an equally impressive record on the issue of long-term

---

[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 current observer bodies are: the European Commission, India, Israel, Japan, the Russian Federation, Turkey, UNESCO and the United States of America.
[3] 'Convention for the Establishment of a European Organization for Nuclear Research', signed 1 July 1953, entered into force 29 September 1954
[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] More information is available at: http://lcg.web.cern.ch/LCG/
[6] Archivage des données et documents numériques sur supports électroniques de l'administration fédérale aux Archives fédérales suisses (ARELDA)

preservation. From the perspective of an archivist, this area is rather neglected at CERN, and almost no progress has been made since the first 'Future Proof' meeting in 2003. It is pleasing, therefore, to start by reporting one new project, namely the efforts being made to preserve, at least for the medium term, the data collected from a former accelerator, the Large Electron-Positron collider (LEP). LEP operated from 1989 to 2000, and was removed to make way for the new LHC. The aim is to preserve these data for around 20 years to enable analysis to continue, which involves not only preservation of the data and necessary metadata to make it usable, but also maintenance of the computer programs needed to process it, and (even more difficult) preservation of the detailed knowledge of the detector that is needed to understand the data.

The original approach, outlined in 2003, was a 'museum system', with regular checking and copying of software and a pool of retired hardware for spare parts. The loss of knowledge after the departure of the people who understood the data was to be made up by 'exhaustive documentation'. However, the project leader now feels that software emulation would be a better approach than trying to maintain a hardware museum. Due to pressure of other work little real progress has been made over the last three years. The data are currently stored in the CERN storage management system, and former experiment collaborations are trying to ensure that their data are in a suitable form for whatever solution is found. In addition to the raw data, various metadata and calibration data are needed to convert it into something with which researchers can do physics.

Even if raw data are of limited value for long-term historical use in the field of particle physics,[7] the same cannot be said for electronic records in general, and the lack of progress here is more discouraging. CERN set up a first long-term electronic archiving (LTEA) working group to look into the issue in 1997. 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 some 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. This seemed important since these are the people who need to be sensitised to the issues, and who are in a position to do something about them. It was also felt that their opinion on this subject would carry more weight than that of archivists or librarians. This group's report, which was presented to the Director General of CERN 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 an in-house system for e-mail, investigate Web archiving, and develop a document handling policy.

[7] See, for example, the American Institute of Physics Study Of Multi-Institutional Collaborations; Phase I: High Energy Physics; Report No. 1, Part A, Section III 'Guidelines for Appraisal of Records': 'Another important finding is that raw data, and even data summary tapes, are not useful to researchers once the needs of the particular collaboration have been met.'

A relatively cheap in-house approach was proposed for e-mail, which advocated archiving all messages for selected users, and selected messages for all users. CERN lacks an organisation-wide records management policy, so the concept of identifying all 'records' (in the strict sense of the word), regardless of their physical form, and implementing records retention based on the intrinsic value of these records, was unlikely to be acceptable, or to be implementable in the short or medium term. As an example, it was suggested that all e-mails of the CERN Directorate should be kept permanently, though an opt-out button (and filtering techniques) would be offered in view of the widespread perception that business e-mails are somehow still 'personal' records. The new system would be integrated into the current e-mail system by a combination of mail forwarding and listbox archives. Solutions remained to be found to provide a reliable write-once/read-only repository of sufficient size and reliability, and the report briefly mentioned some advantages and disadvantages of proprietary 'archiving' systems, the use of a dedicated IMAP server, and the use of a dedicated Web server.

Further investigation of Web archiving was also recommended, with an emphasis on learning from the many excellent public sector developments, since meetings organised by the working group with companies dealing in this area had not been promising. Quite a broad approach would be needed, as the work of CERN is entirely based on worldwide collaboration; the whole CERN domain and links beyond it would have to be followed. It seems scarcely fitting that the place where the Web was born should not have a reliable Web archiving system.[8] The third recommendation was for the development of a document handling policy.

Presentations to top management were well received, but coincided with the start of a budget crisis and no action has resulted. Ongoing efforts to revitalise management interest are focusing on e-mail, since this is the most urgent problem. Although most archivists agree that it is wrong to treat e-mails in a uniform way, rather than on their merits as records, it seems better to try to achieve some improvement in the current situation at CERN, even in the absence of clear organisation-wide guidelines on records management. And to end on a more positive note, it is encouraging that even without such guidelines, use of the organisation's two electronic document management systems is increasing. Copies of all published articles by CERN authors (and certain other scientific documents) are supposed to be placed on the CERN Document Server (CDS), though this has been hard to enforce, and all engineering documents and equipment data relating to the accelerator currently under construction must be placed in the CERN Engineering Data Management Service (EDMS) database. Not only is compliance increasing, but both services are gradually becoming more widely used for other types of documents.

[8] The World Wide Web was invented at CERN by Tim Berners-Lee as an efficient way of using the Internet to share information quickly; more information at: http://intranet.cern.ch/Public/Content/Chapters/AboutCERN/Achievements/Achievements-en.html

Making documents available from an institutional repository such as CDS is one of the ways CERN contributes to Open Access (OA).[9] The OA movement aims to change the current publishing model and to promote open access to scientific research articles and results. A number of discussions and public meetings about it have taken place at CERN, and at these there have always been questions about the long-term preservation of electronic records. This shows that there is considerable concern about this issue, but also a lack of information, which archivists, librarians and others must try to remedy. It is not always generally understood that the 'archiving' of electronic journals is a issue *now*, affecting all journals, whether they are open access or not. Nor are people usually aware of the measure that have already been taken by national libraries and many publishers to deal with it. And fewer still are those who have looked beyond electronic journals to consider the possible fate of the huge mass of unpublished records.

[9] The following definition of Open Access is taken from the Budapest Open Access Initiative 2002: '…free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright … should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited.'
More information is available at the Open Access Bibliography:
http://www.digital-scholarship.com/oab/oab.htm