When Patrimony Data meets Scientific Data

CERN’s Digital Memory

JY Le Meur
Live stats
HIGGS BOSON

WWW
The World Wide Web

The World Wide Web (WWW) is the universe of network-accessible information, an embodiment of human knowledge. It is an initiative started at CERN, now with many participants. It has a body of software, and a set of protocols and conventions. WWW uses hypertext and multimedia techniques to make the web easy for anyone to access, browse, and contribute.

What is W3?

A picture of the world online information, engines, W3 servers, etc.

WWW Architecture

WWW architecture, the layout for WWW clients, servers, databases, and tools.

Discussion

WWW mailing lists, news groups, and other WWW online discussions.

Technical Details

WWW technical details, the layout for W3 clients, servers, and tools.

Spreads, Tutorials, etc.

WWW spreads, tutorials, and other WWW resources.

First International Conference on the World-Wide Web

May 25-26-27, 1994, CERN, Geneva (Switzerland)

The Conference is over... This page remains as a report and as pointer to the next ones.

The third one will be:

WWW archive '94

CERN, Geneva, April 26-30, 1994, organised by the CERN

The second conference was held in the USA:

WWW Fall '94

Chicago, October 31-November 3, organised by W3C and CERN

The three days in Geneva were attended by 300 participants from all over the world.

My thanks go to many people, and also to many institutes and companies, especially to the co-organisers from the CERN (Centre European de Recherches Nucleaires), Oscar Niemeyer and Bertrand Brahman.

There were 49 formal presentations, 11 workshops and many discussions going on in parallel. The proceedings will be published in two forms, and until then, the supplementary proceedings are available.

The short of WWW archive '94 was handed out at the Conference Dinner.

The conference closed on Friday afternoon with a general panel discussing issues for the future.

You can submit your paper (but I do not guarantee publication).

User Society

Furthermore, we are considering creating a WWW User Society, which will serve the purpose of grouping the individual users, institutes and existing interest groups. It will run the conferences and suggestions on purpose and structure, as well as support of help with this Society, by e-mail to User Society at W3C via W3C.

Past History

You can still access the old information temporarily.

AC '93, June '94
FileSystem Phase-Out in 2018
→ 39'000 files dated between 1990 and 1997
"WWW", "Administration", "hypertext", "History", "ftp", "www0", "People", etc.

Donation Option
"CERN is not just another laboratory. It is an institution that has been entrusted with a noble mission which it must fulfil not just for tomorrow but for the eternal history of human thought."
(Albert Picot, 3rd Session of CERN Council, Geneva, 10 June 1955)
- Bit rot and redundancy failure
- Obsolescence of readers, formats, OS, HWs
- Lost in transitions
- Missing context
- Economic failure
- Corruption, mistake or attack
- Dissipation

...
OPERATIONAL CIRCULAR N° 3
Issued by Personnel Division

Applicable to:
- Members of the Personnel
- External users of CERN Archives

Person responsible for the matter concerned:
Director-General

Date:
October 1997

RULES APPLICABLE TO ARCHIVAL MATERIAL
AND ARCHIVING AT CERN

PREAMBLE

CERN is responsible for its documents and files and their preservation. It must formulate its own archiving policy and rules, taking into account legal obligations deriving from ordinary law. The aim of this document is to set out rules and procedures for archiving and access to archival material at CERN.
ARCHIVING POLICY AT CERN

PREAMBLE

Science does not exist in a social vacuum; therefore cultural, economic and political developments also find their expression in the records of scientific activity.

10. Pending the adoption of specific guidelines for preserving electronic documents, the archiving of a hard copy of important documents is essential and must be made compulsory in order to ensure that all such documents are archived. Nevertheless, their electronic form should not be destroyed.

11. There exist collections of files to which special archiving obligations apply.

1997
Digitizing the “Big Data” of the 20th Century
5'183 cassettes

3'648 tapes

1'740 Kg

15 TB
Ingested ~30’500 files (master and access) into 3’000 records - with 3’300 scanned timelines

- The digitally-born audio recordings (post-2006)
- Run speech to text?
- Open the access
  - → in ~100 years?

<table>
<thead>
<tr>
<th># of tapes requiring cleaning</th>
<th>28</th>
</tr>
</thead>
<tbody>
<tr>
<td># of sticky tapes requiring baking</td>
<td>80</td>
</tr>
<tr>
<td># of ‘no signal’ (bande vierge)</td>
<td>677</td>
</tr>
<tr>
<td># with only 1 face recorded</td>
<td>688</td>
</tr>
<tr>
<td># of duplicates</td>
<td>125</td>
</tr>
<tr>
<td># of degraded audio signal</td>
<td>95</td>
</tr>
<tr>
<td># of failures (sticky, blocked, blank)</td>
<td>68</td>
</tr>
<tr>
<td># of failed continuity between tracks</td>
<td>1111</td>
</tr>
</tbody>
</table>

Treated: 5112 cassettes, 3289 tapes
The initial situation, offices and corridor storage
<table>
<thead>
<tr>
<th>Format</th>
<th>Bat-510</th>
<th>Bat-500</th>
<th>Bat-60</th>
<th>Bat-698</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>35 mm</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>16 mm</td>
<td>61</td>
<td>51</td>
<td>0</td>
<td>0</td>
<td>112</td>
</tr>
<tr>
<td>8 mm</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>D3</td>
<td>0</td>
<td>40</td>
<td>0</td>
<td>22</td>
<td>62</td>
</tr>
<tr>
<td>Digital Betacam</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>DVCAM</td>
<td>45</td>
<td>9</td>
<td>25</td>
<td>367</td>
<td>446</td>
</tr>
<tr>
<td>DV/miniDV</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>142</td>
<td>151</td>
</tr>
<tr>
<td>DVD</td>
<td>56</td>
<td>440</td>
<td>0</td>
<td>180</td>
<td>676</td>
</tr>
<tr>
<td>Betacam SP</td>
<td>188</td>
<td>385</td>
<td>600</td>
<td>195</td>
<td>1368</td>
</tr>
<tr>
<td>Betacam</td>
<td>23</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>33</td>
</tr>
<tr>
<td>1 inch C</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>U-matic S/SP</td>
<td>273</td>
<td>120</td>
<td>565</td>
<td>108</td>
<td>1066</td>
</tr>
<tr>
<td>U-matic</td>
<td>128</td>
<td>0</td>
<td>0</td>
<td>71</td>
<td>199</td>
</tr>
<tr>
<td>VHS</td>
<td>189</td>
<td>388</td>
<td>180</td>
<td>154</td>
<td>911</td>
</tr>
<tr>
<td>Others</td>
<td>27</td>
<td>0</td>
<td>1</td>
<td>16</td>
<td>44</td>
</tr>
<tr>
<td>Total</td>
<td>999</td>
<td>1438</td>
<td>1391</td>
<td>1275</td>
<td>5109</td>
</tr>
</tbody>
</table>
# Destination file formats

<table>
<thead>
<tr>
<th>Originals</th>
<th>Preservation Master</th>
<th>Access Master</th>
<th>Access Copy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>35mm films</strong></td>
<td>.mkv</td>
<td>.mov</td>
<td>.mp4</td>
</tr>
<tr>
<td>Video codec:</td>
<td>FFV1 - 10 bits RGB</td>
<td>Apple ProRes 422 LT</td>
<td>H.264 @ 5Mbps</td>
</tr>
<tr>
<td>Audio codec:</td>
<td>24 bits PCM, 48kHz</td>
<td>24 bits PCM, 48kHz</td>
<td>16 bits AAC, 44.1kHz, 256kbps</td>
</tr>
<tr>
<td>Definition / Aspect ratio:</td>
<td>4096x? / Original</td>
<td>1920x1080 / Pillar-letterbox</td>
<td>1920x1080 / Pillar-letterbox</td>
</tr>
<tr>
<td><strong>16mm films</strong></td>
<td>.mkv</td>
<td>.mov</td>
<td>.mp4</td>
</tr>
<tr>
<td>Video codec:</td>
<td>FFV1 - 10 bits RGB</td>
<td>Apple ProRes 422 LT</td>
<td>H.264 @ 5Mbps</td>
</tr>
<tr>
<td>Audio codec:</td>
<td>24 bits PCM, 48kHz</td>
<td>24 bits PCM, 48kHz</td>
<td>16 bits AAC, 44.1kHz, 256kbps</td>
</tr>
<tr>
<td>Definition / Aspect ratio:</td>
<td>2048x? / Original</td>
<td>1920x1080 / Pillar-letterbox</td>
<td>1920x1080 / Pillar-letterbox</td>
</tr>
<tr>
<td><strong>Analogue and digital SD video</strong></td>
<td>.mkv</td>
<td>.mov</td>
<td>.mp4</td>
</tr>
<tr>
<td>Video codec:</td>
<td>FFV1 - 10 bits YCbCr</td>
<td>Apple ProRes 422 LT - SD profile</td>
<td>H.264 @ 1Mbps</td>
</tr>
<tr>
<td>Audio codec:</td>
<td>24 bits PCM, 48kHz</td>
<td>24 bits PCM, 48kHz</td>
<td>16 bits AAC, 44.1kHz, 256kbps</td>
</tr>
<tr>
<td>Definition / Aspect ratio:</td>
<td>?x576 / Original</td>
<td>?x576 / Original</td>
<td>640x360 / Pillar-letterbox</td>
</tr>
</tbody>
</table>
● Documentaries and Seminars recordings, and all footages
● Inventory in collaboration with the Library & VideoLab
  ○ Systematic QR-coding of “everything”
  ○ Carrier analysis to either update existing records or create new ones
● QC by both contractor and CERN: with OpenRefine app
● Bidding: 121'000 euros (~23.7 €/carrier)
  ○ INA & Memoriav expertise
SCOOP :-)  https://videos.cern.ch/search?q=www94
34’000 albums since 1954

Licenses: CERN or CC-BY-4.0
● PhotoLab store organized in albums
● B&W: 120’000 treated in 2014
● Approximate count of Colours:
  ○ 24x36 negatives: 180’000
  ○ 24x36 slides: 40’000
  ○ Medium & Large formats: 80’000
● Keys:
  ○ TIFF 48 bits (RGB) & 4800 ppi (for 24x36 size images)
  ○ File naming pattern to enable album identification
● Challenges:
  ○ Metadata enrichment
  ○ Merge with digital-born content
CERN World-Wide Web Days 8-9 March 1995

Over two hundred people, journalists and communication experts from all over Europe came to CERN for the World-Wide Web Days. Day one was devoted to practicalities, with talks from different speakers who all had a hand in the first years of its development.

Date: Mar 1995
14 photos
The challenge of preserving photos

https://youtu.be/CsmLb4cPZ0Q

Abstract

Data from High Energy Physics (HEP) experiments are collected with significant financial and human effort and are mostly unique. An inter-experimental study on HEP data preservation and long-term analysis was conducted as a panel of the International Committee for Future Accelerators (ICFA). The group was formed by large collider-based experiments and investigated the technical and organisational aspects of HEP data preservation. An intermediate report was published in November 2009 addressing the general issues of data preservation in HEP. This paper models and extends the intermediate report. It provides an analysis of the research and data preservation and a detailed description of the various projects at experiment, laboratory and international levels. In addition, the paper provides a concrete plan for an international organisation in charge of the data management and policy for high-energy physics.

This status report of the DPHEP collaboration details the progress from 2013 – 2015 inclusive.

CERN and ISO 16363 Metrics

CERN produces and preserves a variety of data that can be broadly grouped into the following categories:

- Scientific data from the various experiments conducted at CERN;
- Papers and publications, many of which derive directly from the above and / or are related to R&D, service deployment and operations etc;
- The “Digital Memory” of the organisation, which includes minutes of meetings, videos, digitised images and so forth.

This list is non-exhaustive and may be amended in future revisions of this document.

This document contains a set of responses to the metrics contained in ISO 16363 - itself based on ISO 14721, the reference model for an Open Archival Information System (OAIS).

Many of the metrics below refer to the host organisation whereas some require detailed discussion that varies according to the type of data being preserved, the associated services, as well as the designated communities who might access or re-use this data.
An Archival Information Package (AIP)

- E.g. title, author, abstract, etc
- E.g. checksum, digital signature
- E.g. directory structure, filenames, tape marks
- E.g. access control, finding aids

- Must have redundant copies
- Must be regularly checked
- Must be supported by preservation plan
- Must be sustained by an organizational policy
Unique and 
Expensive

Different

Massive

- 100 TB/experiment at LEP
- 1-10 PB/exp at HERA, TEVATRON at BaBar
- X00 PB/exp at LHC
- >10 EB planned at HL-LHC

and complex!
(1) VIRTUALIZING Archivematica TO RUN ON CERN CLOUD

   github.com/CERN-E-Ternity/archivematica-puppet

(2) CREATING A SIP STORE WITH Invenio

   BagIt File Packaging Format (IETF)

   github.com/inveniosoftware/invenio-sipstore

(3) AUTOMATING THE SIP PROCESSING INTO AN AIP

   github.com/inveniosoftware/invenio-archivematica
Difficulties with revised files and with huge size files

“Weak Archiving”
Conclusion

- Moving from Preservation by chance to Preservation by mission
- The Digitization of 20th Century Multimedia
- The Collision with DPHEP
- The Attempt to Converge into an OAIS Dark Archive

- Enlightening the digital archive....
PHOTO DISPLAY: the EPFL Memory Collider
Docu-fiction with CERN
Historical Videos

TIME MACHINE
A journey through time and space at CERN

A movie by Giulia Grossmann
Provided by the CERN Digital Memory project
An artistic surprise

The VolMeur collection of the CERN Digital Memory project

http://volmeur.org

Currently Exhibited at New-York MFTA Gallery
Thanks for your attention!