David Giaretta

Why is digital preservation so difficult?

slide 1 The preservation of digital objects implies several difficulties due to their peculiar characteristics.

slide 2 As opposed to traditional documents, digital objects, which are made up of sequences of ones and zeros, are not legible in a direct way but specific instruments are required to look at them.

slide 3 For instance, looking at the close up of a pressed CD by means of a magnifying glass, it is possible to see little pits on its surface; but obviously the pits do not correspond to bits.

slide 4 The real issue is much more complicated since various levels of details have to be unraveled, from the physical to the logical level, in order to read and to understand the content of a CD.
Looking into a digital object, we can see what is called bit stuffing, and though it may be good enough for a network transmission, there might be other aspects to consider, i.e. error correction codes, issue of logical addressing, the organization of the material on the disk and so on.

slide 5 The bits might be preserved by carving them in stone or writing them in titanium sheets or other long-lasting material, but in any case something about them would be missing.
The preservation of the bits is not enough, it does not encompass what they mean; more information is required.
This kind of information is often called metadata although this term does not actually mean anything, because it covers a range of materials too broadly. In preference, a more precise terminology ought to be used. According to the OAIS reference model, we can refer to this kind of information as Representation Information.

slide 6 The CASPAR project looks at different kinds of materials which are all encoded digitally: documents, obviously, but also pieces of scientific data, materials from cultural heritage and contemporary performing arts, etc. In order to test what has been done in the project, an extremely wide range of disciplines has been chosen.

slides 7-9 [The following slides show several examples of the sorts of data which have been chosen and of really complex materials that should be preserved.]

slide 10

Most of the projects involved in digital preservation only show interest in formats and how they can be preserved.
This is, indeed, a problem, but there are tools which help to discover what format some digital object is in. Using these kinds of tools, it should be possible, for example, to understand that a file is written in the Word format and so to read its content by a Word application.
Even obtaining this result, the certainty of actually having preserved the digital object and the ability to understand it, has not necessarily been gained. Not only do we need to know what format the file is in, but also what the semantics involved is.
Although this issue is not particularly evident for documents, it becomes very clear in the scientific environment where the semantics is absolutely necessary in order to interpret the data.

slide 11

For instance, adding a certain amount of semantics to scientific satellite data (which are clearly made up of sequences of ones and zeros) we succeed in transforming that raw data into a table of numbers organized in different columns. However, other information is required to interpret the table correctly: what the latitude, the coordinate system, the units are, etc.
All the questions that may be asked about those data, ought to be answered in order to go on to the following step which is to process the data to derive some scientific value out of it, for any eventual combination with other pieces of scientific data.

slide 12

Other aspects should be envisaged, namely the legal aspects which are associated with any piece of data.
(A separate presentation concerning digital rights management within CASPAR has been given by Marlis Valentini of Metaware which explains how to manage digital rights in order to understand and to preserve them on long term basis).

slide 13

The preservation of digital objects undergoes many different threats.

slide 14

In some ways, digital preservation is quite easy: it is done all the time, at least over short term scales, and it simply requires money.
With sufficient long-term financing, it should be possible to employ people who are constantly looking after resources to be preserved.
Unfortunately, even if some projects have a huge amount of founding, it is limited at a certain period. It is fairly sure that this will not continue forever.
For this reason, one of the tasks of the CASPAR project is to find a way to survive and to keep the digital objects useful even through period in which financial uncertainty neither allows continual preservation nor a large enough team to look after them.

slide 15

These financial aspects could be a big disincentive for preservation for any organization since generally the budget varies, whereas the cost of digital preservation increases constantly due to the continuous production and accumulation of new digital objects.
CASPAR is trying to provide tools and techniques to make it practical to survive without involving huge amounts of money, essentially by allowing people to share their efforts.
Furthermore, even if it is often said that resources are preserved for future generations, there is probably a more important reason to maintain things: digital preservation is very closely tied to the reuse of digital objects, and this is true especially for scientific and commercial data.
By combining scientific data from different disciplines, real benefits may be obtained, new products can be created, with consequent enhancements in the economies of our countries, not purely thanks to digital preservation, but because we preserve the ability to use and understand the material we have.
That is not to forget the cultural heritage which is a benefit to humanity of course. Even if it is more clearly understandable in terms of scientific and commercial digital information, the same sort of arguments concerning the reuse of data could be used for the areas of cultural heritage and contemporary performing arts as well.

slide 16

Besides the financial issue, there are also threats concerning the trustworthiness of the repositories; in fact there is the risk to entrust the preservation to someone that is not really up to look after it.
To overcome these threats the CASPAR project has tried to adopt the approach of OAIS – Open Archival Information System reference model – that provides a fundamental view of preservation which revolves around testability.

slide 17

OAIS looks at data in a very general way; one of the key concepts is that it must be testable.

slide 18

The bits preservation can be tested in the sense that it is possible to verify whether the bits sequences have been unchanged. However, as explained above, it cannot guarantee understandability.
OAIS introduces tests about usability of data which include, among others, the concepts of Representation Information (RepInfo) and Knowledge Base of a Designated Community.
All of these could be encompassed in the term metadata, but it is better to use this more precise terminology.

slide 19

RepInfo is what we ought to add to the bits in order to make them understandable, and the key point is that RepInfo may itself be encoded digitally, so it goes through this same process and requires further RepInfo.

slide 20

For example: if we consider a FITS file (FITS is an astronomical data format), by means of the FITS standard we may be able to display the information from the FITS file, but not to actually understand it.
If the FITS standard is given as a PDF file, some PDF software must be used in order to read it, but if PDF is no longer used, then more information about this format is needed.
Some software that deals with FITS is necessary and that might be JAVA software; as it may not be available in future something about the JAVA Virtual Machine has to be explained. Even conserving all of these things, the information from the FITS file cannot be extracted yet because the FITS standard may be defined by a few dozen keywords. A typical FITS file formula observatory has many hundreds of keywords that tell how, where and when the data was taken, so a dictionary is needed too. In turn, it might be written in XML, that is fine now if one knows the specification, but more information concerning XML would be necessary to know in future.
Saying that the Knowledge Base of the Designated Community might change, OAIS explains this probable need of additional information to understand a resource.
In other words, if Java is no longer available, then the Knowledge Base of the Designated Community needs to be supplemented by RepInfo regarding the JAVA Virtual Machine and so on.

slide 21

CASPAR looks at digital preservation as involving creating many levels of information concerning the data: we have to look at access control, capture information about the semantics that is embedded in the resource, look at how it is laid out in the digital object, store the data and then, when it is used in future, we need to essentially follow the inverse process. At each level there is encoded metadata which must be preserved together with its explanation for future.
So, in order to preserve a digital object, techniques are needed that should allow one to preserve not only the digital object but also the digital rights, the semantic information, the description of the structure that go along with it, and so on.

slide 22

The other things which have to be guarded against are changes in software and hardware, changes in environment (there is a lot of interconnection especially because we rely on things that are available through the internet but a large fraction of them become unusable after quite a short time) and changes in people's knowledge base.
The preservation has to withstand all of these sorts of changes and this is quite difficult to do.

slides 23-24

[The speaker did not comment these slides.]

slide 25

Enquiries have been done about how these threats are perceived by the community.

Asking common questions about threats to different sorts of people who are involved in digital preservation, the majority of these different stakeholders thought that all of these threats were either important or very important.

When the survey was set out it was designed to have 5 options in a way to tempt people to choose the middle option, because it said “slightly important”, if people did not really know or care. A large majority of people were expected to tick the middle box.

What the numbers and the final results show is that the majority of people think that all of these threats are either important or very important.

slides 26-27

The CASPAR project tries to find solutions to face these threats.

Users may be unable to understand or use the data. The CASPAR solution (based on the OAIS model) is to maintain the Representation Information necessary to understand and use this data.

The non maintainability of hardware should essentially be faced by means of emulators, replacements in software terms, the rebuilding of systems or, as an alternative, keeping old hardware in a museum (there are some of these around but it does become very difficult to maintain them). The lack of evidence could be a concern for people and more certainly a concern for the state archives and for commercial companies that have legal obligations. What CASPAR is doing is to allow a coherent way of bringing together information that is evidence for preservation. This is tied with what OAIS and other projects say about the fundamentals of authenticity.

Access restrictions may make it difficult to reuse data or it may not be respected in future and difficulties may arise in its reuse if the rights involved are not clear (the way that CASPAR is dealing with digital rights management is explained in a separate statement).

The CASPAR project has not found all the answers, it has made a good start on how to address that need but, for example, not enough work has been done on the ability to identify data and there are a number of other areas that need further investigation.

Being concerned that the current custodian of the data may cease to exist in the future, something has been done about it and it is the Orchestration Management which could allow a brokering between organizations for handing on digitally encoded information that they can no longer maintain.

Finally, there are threats concerning the trustworthiness of people and archives to which the preservation of digital data is entrusted. They may not really preserve the information. A group of the researchers of CASPAR is working on producing the draft of a standard that is going into the international standards organization and hopefully, on that, a process for certifying archives will be built.

So, in all of these major threats that the majority of people, in a very wide

survey, have agreed to be important or very important, it can be said fairly confidently that the CASPAR project has done a satisfactory amount to address all of them.