ascent
atos thought leadership

White Paper

Digital Preservation in the Age of Cloud and Big Data

Your business technologists. Powering progress

Atos

This White Paper describes the technical challenges and potential solutions for preserving digital artifacts, for long intervals of time (several decades, centuries) in a new world of massive, distributed data.

After introducing the basic concepts of Long-Term Digital Preservation, we show several examples of its importance in different sectors. Then, we explore the challenges that make this technical field a complex one, and in each case, we study the emerging state of the art that will define its evolution in the near future.

This is followed by a high-level vision of the architecture of Long-Term Digital Preservation implementation. To conclude, we analyze the status of these kind of solutions in enterprises, and provide some potential ideas for ICT companies interested in offering these services to their customers.

Published in September 2014

Digital Preservation in the Age of Cloud and Big Data

Contents

Authors

This Ascent White Paper was developed by the member of the Atos Scientific Community Celestino Güemes Seoane, Energy & Utilities Solutions R&D, Atos Worldgrid in Spain

About the Atos Scientific Community

The Atos Scientific Community is a network of some 100 top scientists, representing a mix of all skills and backgrounds, and coming from all geographies where Atos operates. Publicly launched by Thierry Breton, Chairman and CEO of Atos, the establishment of this community highlights the importance of innovation in the dynamic IT services market and the need for a proactive approach to identify and anticipate game changing technologies.

Prologue: A story of two travelers

Between 1831 and 1836, a young English scientist travelled all over the world, methodically annotating his observations in a series of paper notebooks. The information from those notebooks helped our brave scientist to develop one of the most revolutionary scientific theories, the Theory of Evolution. Darwin's notebooks for the voyage of the *Beagle*[1] are preserved, tainted by time but in reading order, by the English Heritage at Down House.

Let's move nearly two centuries forward. A satellite, the International Cometary Explorer (ICE), launched in 1978, is approaching Earth after decades of exploration[2]. The satellite seems to be in functioning order, as a carrier signal is detected by the Deep Space Network. But, almost in any sense, it is just a ghost, as NASA can't communicate with the satellite:

"Communication involves speaking, listening and understanding what we hear. One of the main technical challenges the ISEE-3/ICE project has faced is determining whether we can speak, listen, and understand the spacecraft and whether the spacecraft can do the same for us. Several months of digging through old technical documents has led a group of NASA engineers to believe they will indeed be able to understand the stream of data coming from the spacecraft.

NASA's Deep Space Network (DSN) can listen to the spacecraft, a test in 2008 proved that it was possible to pick up the transmitter carrier signal, but can we speak to the spacecraft? Can we tell the spacecraft to turn back on its thrusters and science instruments after decades of silence and perform the intricate ballet needed to send it back to where it can again monitor the Sun? The answer to that question appears to be no.

The transmitters of the Deep Space Network, the hardware to send signals out to the fleet of NASA spacecraft in deep space, no longer includes the equipment needed to talk to ISEE-3. These old-fashioned transmitters were removed in 1999. Could new transmitters be built? Yes, but it would be at a price that no one is willing to spend. And we need to use the DSN because no other network of antennas in the US has the sensitivity to detect and transmit signals to the spacecraft at such a distance."

Humble, cheap paper notebooks can keep information for centuries, but a big pile of complex and expensive telecommunication hardware and software can't (or, at least, not without considerable effort and passion, as we'll see in our epilogue). And this is just one example. Even in your personal life (who hasn't lost photos in a hard disk crash?) the ephemerality of hi-tech seems to be always present.

We can't deny that technology progress has provided manifold improvements: faster computation, ample connectivity, enormous data storage... But, with respect to long-term preservation of information, technology evolution brings tricky problems, with no easy solutions. And maybe we are blinded by all the wonders of tech, but we can't ignore these problems, as some of them may impair the long-term sustainability of mankind.

[1] Darwin's Beagle field notebooks, Darwin Online, http://darwin-online.org.uk/EditorialIntroductions/Chancellor_fieldNotebooks.html

[2] "ICE/ISEE-3 to return to an Earth no longer capable of speaking to it", http://www.planetary.org/blogs/emily-lakdawalla/2014/02070836-isee-3.html

Definition and general challenges

Long-Term Digital Preservation encompasses of the methods and technologies that ensure that valuable digital information can be correctly preserved in time, so it remains accessible and usable by interested parties, in spite of media failure and technological change.

Long-Term Digital Preservation can be considered the most complete form of all the approaches to preserve data in time, as it is shown in the following table (based in the definitions on the SNIA dictionary[3]):

Level	Description
Backup	A collection of data stored on (usually removable) non-volatile storage media for the purposes of recovery in case the original copy of data is lost or becomes inaccessible.
Disaster Recovery	The recovery of data, access to data and associated processing through a comprehensive process of setting up a redundant site (equipment and work space) with recovery of operational data to continue business operation after a loss of use of all or part of a data center.
Digital Archive	A storage repository or service used to secure, retain and protect digital data information and data for periods of time less than that of long-term data retention.
Long-Term Digital Preservation	Ensuring continued access to, and usability of, digital information and records, especially over long periods of time.

Note that, according to SNIA, the definitions of Digital Archive and Long-Term Digital Preservation seem to be differentiated by an undefined time interval. However, we may say that a Digital Archive focuses mainly on the data being stored, while Long-Term Digital Preservation includes all the additional means needed to assure the access and usability of that data.

[3] 2014 SNIA Dictionary, http://www.snia.org/education/dictionary

There are several intrinsic characteristics of digital media that create important challenges for Long-Term Digital Preservation, both technically and economically:

- Digital objects have mediated access, as they need a hardware and software environment to render them. These environments can be quite complex and in some cases, distributed. And technological evolution makes them obsolete in a short span of time, so you need to find ways to preserve both media and their execution context.
- Digital objects can be quite complex, as sometimes they are the composition of different sub-components. This makes it quite difficult to guarantee the integrity of objects.
- Digital objects can be dynamic, so it is complex to "freeze" them in a concrete state. Also, issues on authenticity may arise due to these dynamic behaviors. Auditing acceptability of data is a critical aspect for legal and business environments.
- Digital objects are easy to create and they are growing in number very fast, so they represent a scaling problem for digital preservation activities. This is more problematic for "fat" content, like audio and video media.
- For some digital objects, like software programs, the absence of source code may be a problem, so they could be ported to newer environments. Legal aspects, like copyrights or copy protection mechanisms, can make this even more complex.
- Many tech solutions, nowadays, seem to be designed with a "planned obsolescence" mindset, intentionally or not. One example of this short-term thinking is the use of DRM schemes that require validation by external entities that may be unavailable in the future.
- Decisions about what to preserve are not easy. As the saying goes, "one man's garbage is another man's treasure'. Even data that, by itself, may seem of no intrinsic value, may provide important value by indirect means, like analytics.
- Digital Preservation may collide with other aspects related to data management. One example is the potential conflict with Privacy, as the evolving concept of "Right to be forgotten" shows. Also, aspects related to Intellectual Property Rights (IPR) may have a profound impact.
- Finally, economic challenges can be quite significant. Preservation can be costly, both in initial and ongoing costs, and benefits only accrue to future generations. As many initiatives in this field are backed by governments, they can be seen as an unnecessary expense in harsh economic times.

Why is Long-Term Digital Preservation important?

One may think that Digital Preservation only matters to rich, idealistic people that are worried about preserving history. But the problems that can arise when digital assets aren't correctly preserved can have a clear social and economic impact, as the following examples show:

- Obviously, historic research on our present times will suffer if online evidence about them vanishes without trace. Social media messages that played a significant role in historic events (like the Arab Spring revolutions) are disappearing at an alarming rate[4], so investigations about how they unfolded could become impossible.
- The legal system isn't immune to "link rot" either. In the case of the US Supreme Court, 49% of links point to nowhere[5]. The legal implications of this can be quite daunting. Documents with legal impact (contracts, official notary documents) need to be preserved for decades, with full liability.
- Contemporary Science is data-intensive, and one of its basic principles is reproducibility. And nowadays, this principle is in very bad shape: a study shows that 90% of 20-year-old studies' data is missing,[6] due to poor archives and impossible to find authors.
- Engineering isn't safe, as we have seen in the ICE Satellite Case. Sometimes, the efforts to recover information from NASA missions end well, but after looking for extremely rare hardware in museums, like tape readers from the 1950s[7].
- Healthcare institutions are increasing their demands on digital preservation. Not only for medical purposes, but also due to regulations or recommendations by Medical Associations, suggesting retention times of at least 25 years, or even "indefinitely"[8].
- Some countries are making an extensive (and costly) effort to preserve their cultural heritage. For example, Norway is digitizing all Norwegian Books, providing free access to all its citizens, irrespective of copyright status[9].
- Ironically, avant-garde Digital Art may be one of the more ephemeral forms of art, compared with traditional paintings and sculpture. Some authors have declared that the need to preserve digital art is already an emergency[10].
- Media organizations with a long history, like the BBC, have massive archives, and have implemented big digital preservation projects . But this has not been always the case, as the "Doctor Who" missing episodes can attest[12].
- It may sound extreme, but, in case of a planetary disaster, digital preservation may be essential to ease the burden of recovery for surviving generations. A more detailed view on this (grim) scenario is available in "Annex: Just in case you have to hit the "Civilization Restart" button".
- Finally, the eternal human desire to have some kind of impact in our descendants is, paradoxically, very difficult in the age that our "online footprint" is so big. Will our descendants be able to access our photos, our videos, our writings... stored in the cloud? Don't take that for granted.

[4] "The disappearing web: Information decay is eating away our history" http://gigaom.com/2012/09/19/the-disappearing-web-information-decay-is-eating-away-our-history/

[5] "In Supreme Court Opinions, Web Links to Nowhere", http://www.nytimes.com/2013/09/24/us/politics/in-supreme-court-opinions-clicks-that-lead-nowhere.html

[6] "The vast majority of raw data from old scientific studies may now be missing" http://www.smithsonianmag.com/science-nature/the-vast-majority-of-raw-data-from-old-scientific-studies-may-now-be-missing-180948067/?no-ist

[7] "Ancient IBM drive rescues Apollo moon data" http://www.computerworld.com/s/article/9119960/Ancient_IBM_drive_rescues_Apollo_moon_data

[8] ENSURE FP7 Project Use Cases, http://ensure-fp7-plone.fe.up.pt/site/uses-cases

[9] "Norway To Digitize All Norwegian Books, Allowing Domestic IP Addresses To Read All Of Them, Irrespective Of Copyright Status" http://www.techdirt.com/articles/20131210/08174425520/norway-to-digitize-all-norwegian-books-allowing-domestic-ip-addresses-to-read-all-them-irrespective-copyright-status.shtml

[10] "The Challenges of Digital Art Preservation" http://www.e-conservationline.com/content/view/884/296/

[11] "Digitisation and Digital Preservation Challenges at the BBC" https://www.prestocentre.org/blog/digitisation-and-digital-preservation-challenges-bbc

[12] Wikipedia: Doctor Who missing episodes, http://en.wikipedia.org/wiki/Doctor_Who_missing_episodes

The dimensions of Data Preservation problems

Like many other cases in technology, digital preservation is a layered problem, with the following layers:

- **Physical storage:** the physical media where information is recorded.
- **Data formats:** the logical structure of the stored information.
- **Software and applications:** the executable context where information is managed.
- **The Web:** the most important model of distributed information.
- **The Cloud:** the most significant model of distributed computation.

However, while in some other technology cases the layered structure hides complexity from one layer to the next, in the case of digital preservation, there is a "multiplication" effect that makes the problems harder.

Physical storage

The first layer of any data preservation system is the physical medium where data is stored. Any digital media is susceptible of physical decay: CDs & DVDs use dyes that degrade, and magnetic media is susceptible to magnetic remanence decay, for example[13]. And mechanic elements in readers of both types of media are prone to failure.

Even so, data about the average life expectancy is hard to determine, as it depends on many factors: how many times it is accessed, the care with which it is handled, the storage conditions, and so on. The next table shows some average values on the average life of different media[14]:

Media	Average Life
Digital Media	
Floppy Disk	10-20 years
Magnetic data and cassette tapes	10-30 years
CDs and DVDs	5-10 unrecorded, 2-5 recorded
Blu-Ray	10-15 years
Rewritable Magneto Optical (MO) Disks	50 years [15]
Hard Disk	3-6 years
Flash Storage	Depends on write cycles, 5-10 years or more
Game Cartridge (Nintendo)	up to 10 years (battery)
Analogue Media	
Newspaper	10-20 years
Microfilm	10-500 years
Photographic slides	100 years
Archival-grade acid-free paper	100-500 years
Egyptian stone tablet	2,200 years

(Note that the Average Life is just one characteristic of storage media. Other features should be taken into account, like information density, or read-write times, when selecting a specific technology for data storage)

One special case that has been widely studied is one of the most used mechanisms for storing information nowadays, hard disks. Backblaze, a cloud storage company, has done extensive research on the failure rate of the hard disks they use, and it has found that the expected median life of hard disk is 6 years[16].

As we see, life expectancy can be worse than analog media, like paper. This explains that there has been a lot of investigation around physical storage media with much longer life expectancy, in different degrees of development:

- Sony and Panasonic have defined the Archive Disk[17], an evolution of the Blu-Ray disks that may store up to 1 Terabyte of information.
- M-Disk, a private company, has worked on a writable DVD media based on inorganic mineral layer, that enhances its durability, up to 1000 years[18].
- Hitachi has announced the capability of storing information in slivers of quartz that may last for hundreds of millions of years[19].
- The French nuclear waste agency, ANDRA, has developed a hard disk prototype based on sapphire with information visually engraved in platinum that may last for a million years[20].
- The University of Twente has done research in another "million year" storage disk, using tungsten and silicon nitride[21].
- Researchers at the University of Southampton (UK) propose an optical data storage system, where data is stored in nanogratings in quartz, created by ultrashort pulses from a femtosecond laser, with a life of 3 x 1020[22] years .
- But the prize for the most exotic form of long-term storage goes to DNA. Researchers are encoding information using DNA, and taking advantage of its stability and self-reproducibility to store information securely for tens of thousands of years[23].

Another important point to have in mind is that all storage media needs hardware or software to be read. Sometimes, the trickiest part is not that the medium is in good shape to be read, but that the reading mechanism still exists. Some of the research in the previous list use ingenious methods, like codifying the information using visible QR codes, which in addition provide error-correction capabilities. Or, like the solution for ANDRA, data is stored in analogue form (readable page images), being readable just using optical magnification.

These new extended storage capabilities, and the explosive growth in media stored by Internet giants, have demanded them to automate their capabilities in what is called "cold storage". For example, Facebook uses self-built massive robot systems, using cheap Blu-ray disks[24]. And Amazon, with its Glacier storage solution[25.] offers similar capabilities. These kinds of automated setups could be used for long-term archiving purposes.

Also note that not only the physical storage medium needs to last, but also all the surrounding infrastructure, like datacenters where data is stored, needs to be designed taking a long term view on its survivability. There is no gain in storing costly sapphire disks in a building that may be crippled by earthquakes or floods.

[13] Wikipedia, "Media Preservation" http://en.wikipedia.org/wiki/Media_preservation#Magnetic_media.2C_video_cassettes.2C_tapes.2C_hard_drives

[14] Based in "Data storage lifespans: How long will media really last?" (http://www.storagecraft.com/blog/data-storage-lifespan/) and "Digital Media Life Expectancy and Care" (http://www.caps-project.org/cache/DigitalMediaLifeExpectancyAndCare.html)

[15] Data from Maxoptix, http://www.maxoptix.com/rewritable-mo-disks/

[16] "How long do disk drives last?" http://blog.backblaze.com/2013/11/12/how-long-do-disk-drives-last/

[17] "'Archival Disc' standards formulated for professional-use next-generation optical discs" http://www.sony.net/SonyInfo/News/Press/201403/14-0310E/index.html

[18] "The 1,000 year DVD is here", http://www.zdnet.com/the-1000-year-dvd-is-here-7000009771/

[19] "Hitachi Plans to Store Data for a Few Hundred Million Years", http://blogs.hds.com/hu/2012/09/hitachi-plans-to-store-data-for-a-few-hundred-million-years.html

[20] "A million-year hard disk", http://news.sciencemag.org/2012/07/million-year-hard-disk?ref=hp

[21] "Million-year data storage disk unveiled" http://www.technologyreview.com/view/520541/million-year-data-storage-disk-unveiled/

[22] " Optical data storage has virtually unlimited lifetime", http://phys.org/news/2014-02-optical-storage-virtually-unlimited-lifetime.html

[23] "DNA Storage: the code that could save civilization" http://www.bbc.com/future/story/20130724-saving-civilisation-in-one-room

[24] "Facebook's prototype cold storage system uses 10,000 Blu-ray discs to hold a petabyte of data" http://www.theverge.com/2014/1/29/5359628/facebook-blu-ray-storage-system-uses-10000-discs-for-petabyte-data

[25] https://aws.amazon.com/glacier/

Data formats

Regarding data file formats, there are some desirable traits that make for a good file format for digital preservation:

- The specification should be freely available.
- There should be no patents or licenses on the format.
- The format should be ubiquitous, that is, in wide use.
- The format should have an extensive feature set.
- The format should be endorsed by other established repositories.
- There should be a variety of writing and rendering tools available for the format.
- The format should be interoperable, with tools that allow the conversion to other formats.
- The format should include error-correction capabilities.
- For image, video and audio information, lossless formats are desirable.
- The format should support a stable mechanism for metadata management.
- Data integrity features, based on flexible digital signatures for existing and future cryptographic methods.
- Self-describing capabilities are very desirable, especially for files that store big data sets.

As it is easy to see, many of these traits are closely related to the Open Source philosophy. It is not rare, so, that many Digital Archiving initiatives have selected open source backed formats as their "normalized" formats. But there are many pitfalls that need to be taken into account:

- Even as some open source software can read proprietary formats, like those of Microsoft Office, their rendering or conversion of the documents may be incomplete or erroneous.
- Many audio and video formats, like MP3 and MPEG-2, even though seen as "free", include proprietary licenses.
- Tooling for "Lossless" formats is not as widely available as for compressed ones that may imply data loss (for example, JPEG or MP3). Multiple passes due to system migrations may "pollute" the data, rendering it unusable.
- Metadata management can be tricky. Even as there are several metadata standards like PREMIS, METS, Dublin Core, EAD, and ISAD(G), the integration with existing standards can be difficult, so sometimes simpler, external methods based in ASCII or XML may be used.
- For massive scientific datasets, self-describing formats are a must. Some emergent formats like Avro, ORC, or HCatalog are trying to cover this need.

One very important trend is to extend the concept of Data File to an extended concept, Portable Information Objects[26] that complements the information object with additional metadata, audit logs, other linked objects, readers, etc. so the object becomes self-described and self-contained, and improves its portability. Standards like SIRF[27] may contain multiple information objects. Note that these kind of standards may be used in other contexts, not only for data preservation, but for portable data repositories in cloud solutions.

[26] Long-Term Digital Preservation Model – Portable Information Objects http://www.ltdprm.org/reference-model/architecture/preservation-objects/portable-information-objects

[27] SNIA SIRF, http://snia.org/SIRF

Software & applications

Preservation of Data files is a basic step for digital preservation. However, for some kinds of digital content, like interactive art installations, or videogames, there is the need to preserve the execution environment for the software. In some rare cases, if the source code is preserved, there could be the possibility of rebuilding the software in modern computing environments. But more usually, there is the need to preserve the running environment with the data.

The usual solution is the use of virtualization, or more properly, emulation technologies. However, the focus is a bit different to the usual approach to virtualization in business environments. The digital preservation community depends on long-term availability of old hardware in virtualization/emulation software. This hardware has to be an exact replicate and be correct and accountable in regards to its code-base stability. So, projects like QEMU[28] , MESS[29] , Dioscuri[30] focus more on stability than on performance. And, by means of using the Portable Information Objects seen before, one could link preservation objects to their emulated running environment.

However, emulation is not a perfect solution. In addition to limitation regarding the capabilities of emulators (for example, support of very specific input-output devices, or performance), modern applications tend to be based in distributed architectures, very complex to emulate. So it seems that some kind of continuous mechanism of auditing and monitoring should be included, so it is executed periodically to check the validity of the access mechanism to stored data, and takes action if problems are detected (for example, recommending a potential migration).

[28] QEMU http://wiki.qemu.org/Main_Page
[29] MESS, http://www.mess.org/
[30] Dioscuri http://dioscuri.sourceforge.net/

The Web

The web is based on simple but powerful concepts, like uniform identifiers, or hyperlinking. Sadly, content preservation isn't one of them. This explains that maybe the only HTTP return code that most people know is the (in)famous "404 Page Not Found". It is difficult to know how much of the Web is preserved in one format or another[31], but some anecdotal evidence implies that the status is not good. For example, in 2006, the Million Dollar Page was a viral hit, where people linked pages to specific pixel. In 2014, 22 percent of it has rotted away[32].

Some initiatives are trying to solve some of the problems that the ephemerality of the Web brings:

- The Internet Archive (Archive.org[33]) is one of the best known initiatives that tries to preserve web content. Although it is most known by the WayBack Machine, that saves periodically websites, it encompasses a lot of activities for the preservation of different kinds of media[34].
- Sometimes, even if the original materials may have disappeared, some kind of "Digital Archeology" can be done, so part of it may be recovered from auxiliary sources, like search engine caches[35].
- Perma.cc is a service that tries to avoid the problems that unidirectional hyperlinks represent for material that requires persistent links (for example, legal materials). When a user creates a Perma.cc link, Perma.cc archives a copy of the referenced content, and generates a link to an unalterable hosted instance of the site[36].
- Academic Torrents[37] is an interesting initiative that uses the distributed capabilities of BitTorrent to store and distribute massive research data sets.
- One promising field of activity is the application of Semantic Web Technologies (like RDF, Ontologies, SPARQL, ...) for digital preservation solutions. For example, Atos has developed for the Bibliothèque nationale de France (BnF) the SPAR system (Scalable Preservation and Archiving Repository[38]) that uses a RDF triple store for metadata management[39].

[31] "Computer Scientists Measure How Much of the Web is Archived" http://www.technologyrev iew.com/view/509411/computer-scientists-measure-how-much-of-the-web-is-archived/
[32] "The Million Dollar Homepage still exists, but 22% of it has rotted away" http://qz.com/191794/the-million-dollar-homepage-still-exists-but-22-of-it-has-rotted-away/
[33] https://archive.org/
[34] "How to preserve the web's past for the future" http://www.ft.com/intl/cms/s/2/d87a33d8-c0a0-11e3-8578-00144feabdc0.html#axzz300WADwqG
[35] "Internet Archaeologists Reconstruct Lost Web Pages", http://www.technologyreview.com/view/519391/internet-archaeologists-reconstruct-lost-web-pages/
[36] http://perma.cc/about
[37] http://academictorrents.com/
[38] "SPAR, le système de préservation numérique de la BnF", http://www.bnf.fr/fr/professionnels/spar_systeme_preservation_numerique.html
[39] http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-401/iswc2008pd_submission_14.pdf

The cloud

Even as cloud delivery models represent a lot of advantages for business and end-users, the considerations about digital preservation tend to be secondary, if at all present, in the design of cloud services. Similar problems to those mentioned previously in the case of web content extend (and get even worse) in the case of cloud solutions. We can mention:

- In some cases, the user is unable to have copies of the data stored in the cloud service.
- Control over data content, format, and associated metadata is limited.
- The execution environment can be very complex (or just impossible) to reproduce in a controlled manner.
- Execution environment can be distributed, as a combination of different services that may evolve in different ways.
- Cloud services may disappear, without previous warning. And, in some cases, they can take user's data with them.

Obviously, some of these problems may have an impact that goes beyond the needs of digital preservation. For example, some governments are trying to avoid "cloud lock-in" using SaaS solutions built over open source solutions, known as OpenSaaS[40]. And some companies are looking to extend the concept of "software escrow services" to cloud applications[41].

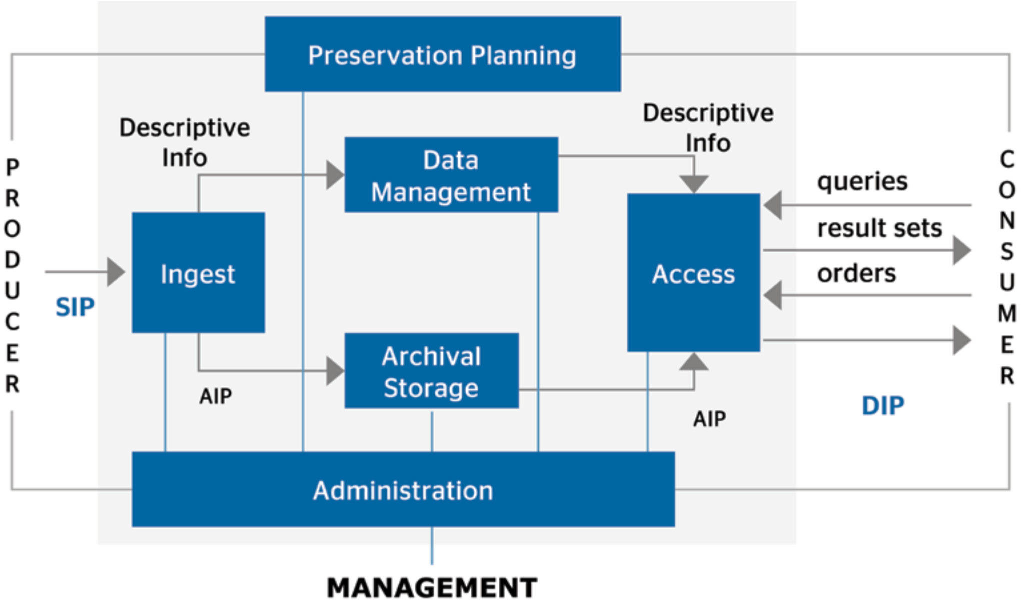
But there is a brighter side to cloud regarding digital preservation: to provide Long-Term Digital Preservation services in a cloud-based model, or, following the typical cloud naming scheme, LDPaaS[42]. It may sound a bit contradictory, but the model has some important advantages:

- It brings the advantage of cloud solutions (elasticity, virtualization, pay-as-you-go) to a complex field, as we have seen.
- It will bring Long Term digital preservation capabilities to smaller archives, that couldn't deploy them internally.
- In fact, big initiatives (at the national or transnational level) could be the providers of LDPaaS solutions for these smaller entities, and that would help to achieve a certain level of standardization.

[40] "OpenSaaS and the future of government IT innovation" https://opensource.com/government/14/1/opensaas-and-government-innovation?sc_cid=70160000000c0zjAAA
[41] "How to protect your company against vanishing cloud services" http://gigaom.com/2013/05/03/a-few-ideas-for-protecting-your-company-against-vanishing-cloud-services/
[42] "Digital Preservation Cloud Services for Libraries and Archives" http://www.diglib.org/forums/2011forum/schedule/digital-preservation-cloud-services-for-libraries-and-archives/

Architecture standards

Long-Term Data Preservation represents what it could be the most advanced form of data management, so it is pretty clear that architecture & governance models are essential for it to work. The starting point is the OAIS ISO 14721: 2012[43], that is the Reference Model for Open Archival Information Systems (OAIS). OAIS, as reference model, provides a conceptual framework for the understanding and increased awareness of archival concepts needed for the long term digital information preservation. Being a conceptual reference, it doesn't imply any specific technologies or methodologies, but more the definition of the functional entities that take part in these kind of systems, that can be seen in the following figure:



SIP: Submission Information Package
AIP: Archival Information Package
DIP: Dissemination Information Package

Figure 1 - OAIS Functional Entities

From this conceptual overview, more concrete architectures are derived. One example of a derived architecture from OAIS is the one defined in the Project CASPAR (Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval)[44].

[43] http://www.iso.org/iso/catalogue_detail.htm?csnumber=57284
[44] http://www.casparpreserves.eu/caspar-project.html

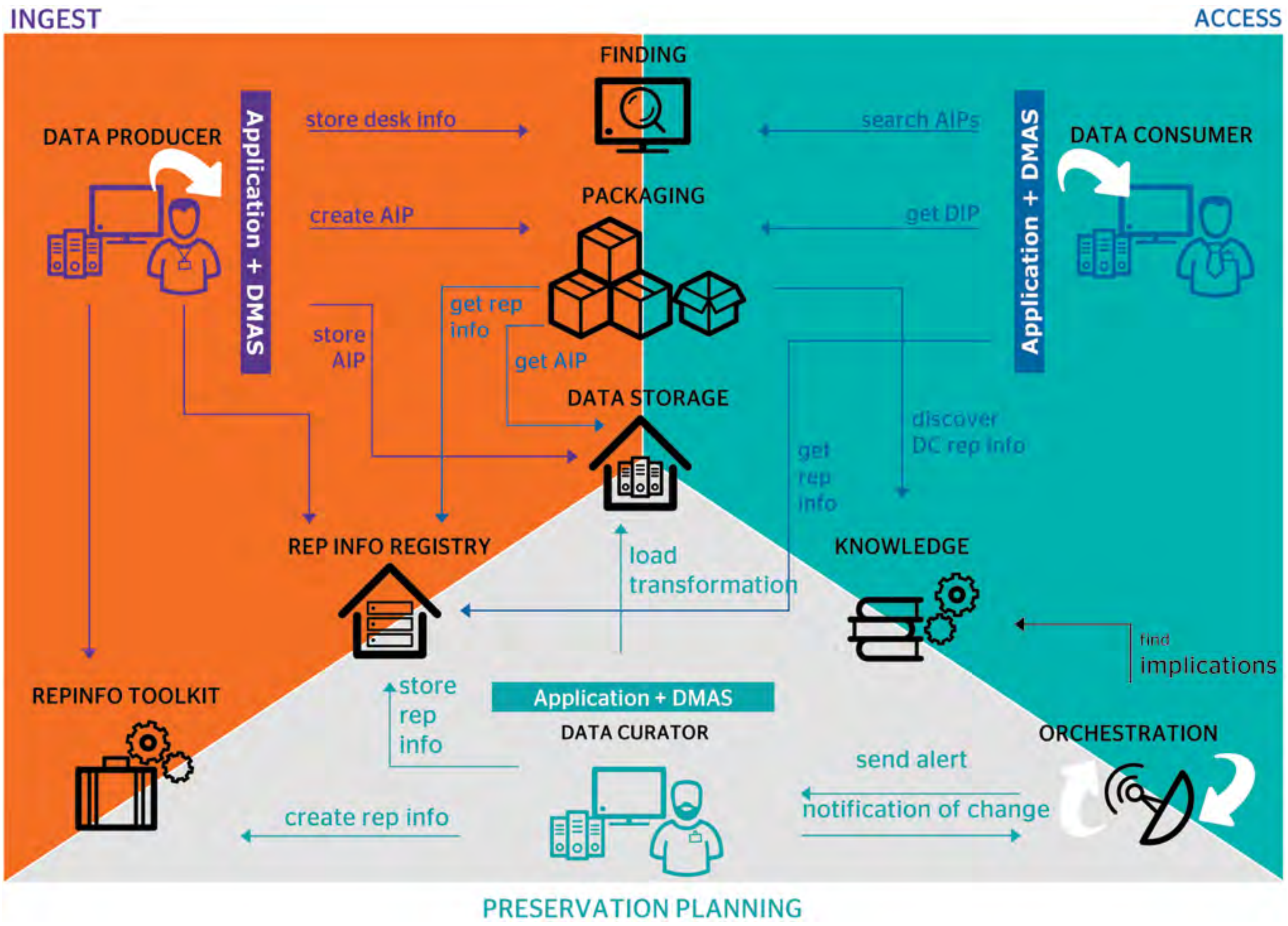


Figure 2 - Architecture of CASPAR Project

From a processes perspective, Long-Term Data Preservation is a topic being covered in different groups that cover Information Lifecycle Management, like ILM2.0[45].

[45] http://www.ilm20.org/

Use of Long-Term Digital Preservation in Business

As it may be evident from the previous points, it is quite clear that the Public Sector (Governments, Universities, and Scientific Groups) is the more active regarding Long-Term Digital Preservation. There are several causes for this: the "moral obligation" to preserve our cultural heritage in the case of Government, or the demands of Scientific Methods for reproducibility in the case of Science.

But the case in Private companies is quite different. Obviously, most of them have implemented Data Archiving functions (as they are necessary due to regulatory and assurance requirements), but they don't usually go to the full spectrum of activities involved in a Long-Term Digital Preservation solution. For example, data may be preserved, but the execution environments needed to have access to them are not. We may say that a different "cultural mindset" helps for this sorry state of Digital Preservation in companies: very long-term thinking (going beyond 5 years, for example) in companies is not very common, as markets impose burdening short-term demands, like "satisfying shareholders with the next quarterly results".

The big question is: is full-fledged Long-Term Digital Preservation a necessity for Businesses? As organizations are more and more Data-Centric and Data-Driven, we think that that is clearly the case. In order to achieve that, some aspects may help them to embrace Long-Term Digital Preservation:

- Due to the emergence of Data Science, some activities are becoming more "scientifically" oriented. Reproducibility, like in other scientific cases, becomes important.
- Also, as Data Science is embraced, the value of "old data" grows, as it may be used to find "old" business patterns that may provide valuable outcomes in the future.
- Digital Preservation processes may help them to preserve Legacy Systems, including preserved execution environments.
- Digital Preservation may extend their digital archiving capabilities, for example, with auditability and authenticity mechanisms.
- For some sectors, like Health, laws are extending regulatory retention times well beyond the capabilities of existing archiving solution, clearly demanding advanced Long-Term Digital Preservation features.
- Some concepts in digital preservation may be useful in other contexts. For example, Portable Information Objects may be applied in cloud solutions as a data portability mechanism.
- And, enterprises being a social institution, some of their cultural artifacts need to be preserved for posterity.

Business opportunities in Long-Term Digital Preservation

As we have seen, many solutions for Long-Term Digital Preservation come from public bodies (like National Archives), public R&D projects, and standards bodies. But, as for private ICT companies, are there potential business opportunities in this field? Some potential ideas that may be worth exploring are:

- As we have seen, the potential exists for offering Cloud Services for Long-Term Digital Preservation for Government and Enterprise, especially in partnership with big archival institutions.
- Multimedia-intensive industries, like TV companies, will drive a lot of demand, not only for preservation purposes, but also for using their old video archives as an additional revenue source.
- For the science and academic sector, digital preservation services oriented to scientific reproducibility could have a big demand. Especially with massive initiatives, like CERN's LHC, or future High-Resolution Space Telescopes.
- Some of these services may be scaled down to be offered to individuals, to preserve their "digital heritage".
- Open, Linked Data solutions may be built over digital preservation repositories, and combined with other data sources.
- Solutions that enforce strong authenticity – auditability capabilities will see a strong demand in business, health and legal environments.
- Specialized Data Science / Data Mining solutions could be used over preserved content, for social analysis and even marketing studies.
- The enablement of existing solutions (for example, Content Management Systems) so they could become "Preservation-Aware".

Epilogue: Preparing for the Long Travels of Data

As we saw in the prologue, a tech-savvy organization like NASA didn't seem to have the means (or the money) to bring back into operation the ISEE-3/ICE mission. Happily, this wasn't the end of the mission. After the news that the problems were known, a group of space science fanatics formed and launched a citizen science project to recover it[46], called the ISEE-3 Reboot Project[47].

Getting funds through a crowdfunding campaign, establishing a Mission Control Center in an abandoned McDonalds ("McMoons") and developing creative solutions to mimic old communications systems using Software-Defined Radio (SDR), they were able to contact and control the satellite. Sadly they could not change its orbit because fuel was exhausted, but they could activate the scientific instruments to be used in future missions. The ISEE-3 Reboot Project is a nice example of a "digital archaeology" project that, with a lot of creative effort, restores part of our digital history. But sadly it seems more of the exception than the rule.

Mankind has been living in such an accelerated rate of change for so long, that sometimes it seems that we have lost any capability of long-term thinking. We have become accustomed to ephemeral products and services that aren't guaranteed to survive us. This impoverishes our societal heritage, and, in the worst case, it may even put our long-term survivability at risk.

In this context, the consolidation of Long-Term Digital Preservation is not only a desirable technological capability. It goes beyond mere technology: it should become a mindset, where the goal of long-term sustainability is one of the driving forces for mankind's progress.

[46] NY Times, "Lost and found in space", http://www.nytimes.com/2014/07/19/opinion/Rebooting-ISEE-3-Space-for-All.html
[47] ISEE-3 Reboot Project, http://spacecollege.org/isee3/

Just in case you have to hit the "Civilization Restart" button

Maybe the more extreme form of thinking about digital preservation may be the "Tabula Rasa" scenario: what could we do to help mankind to prosper again, after a global catastrophe? This may sound way too pessimistic, but studies funded by reputable institutions like NASA point out that our industrial civilization may be headed to an "irreversible collapse" [48], as happened with the fall of the Roman Empire. Or something completely unpredictable like a deadly pandemic, a solar storm, or a regional nuclear war, may destroy or severely damage our techno-centric civilization.

The main point is that mankind is quite resilient, but civilizations come and go. It may be possible that in many of the scenarios (excluding the 'total annihilation' ones), the essential elements (food, shelter, water, sanitation and medical care) of civilization could be restored in a (relative) short time[49]. But the deeply interconnected world that makes most of advanced economies in a profusion of services will be much more complex to restart. Even some basic elements, like electricity, may be hard to get.

The basic approach in this scenario is to preserve information (not only in electronic format) that may be useful for the survivors, and that may expedite their recovery. In some sense, they are the ICT equivalent of the Svalbard Global Seed Vault, an extremely secure seed bank set up by Norway near the Arctic Circle.

Some examples of these "Doomsday Manuals" are:

- The Rosetta Project[50], a global collaboration of language specialists and native speakers to build a publicly accessible library of human languages.
- CD3WD[51], a torrent-downloadable archives with "the Information necessary to Rebuild Society"
- The Global Village Construction Set[52], a series of Open Source industrial equipment that could be used to build a basic civilization.

[48] "Nasa-funded study: industrial civilisation headed for 'irreversible collapse'?" http://www.theguardian.com/environment/earth-insight/2014/mar/14/nasa-civilisation-irreversible-collapse-study-scientists

[49] "How to reboot civilization" http://vinay.howtolivewiki.com/blog/other/how-to-reboot-civilization-1816

[50] http://rosettaproject.org/

[51] http://www.cd3wd.com/cd3wd_40/cd3wd/index.htm

[52] http://opensourceecology.org/gvcs/

About Atos

Atos SE (Societas Europaea) is an international information technology services company with 2013 annual revenue of € 8.6 billion and 76,300 employees in 52 countries. Serving a global client base, it delivers IT services through Consulting & Systems Integration, Managed Operations, and transactional services through Worldline, the European leader and a global player in the payments services industry. With its deep technology expertise and industry knowledge, it works with clients across different business sectors: Manufacturing, Retail & Transportation; Public & Health; Financial Services; Telcos, Media & Utilities.

Atos is focused on business technology that powers progress and helps organizations to create their firm of the future. It is the Worldwide Information Technology Partner for the Olympic & Paralympic Games and is listed on the Euronext Paris market. Atos operates under the brands Atos, Atos Consulting, Worldline and Atos Worldgrid.

Interested in our Ascent - Thought Leadership publications?

Stay connected with the latest forward-looking and inspirational publications on business & technology
www.atos.net/ascent