Building together digital archives for research in social sciences and humanities

Benoît Habert, Claude Huc

Abstract: In order to help understanding the possible interplay between transmission and digitization, a pilot project of long term preservation for research data in SSH is presented by its two coordinators. The paper provides some background context on transmission in digital form of past and present research in SSH. It shows the discrepancy between the increasing role of digital information and the fragility of it. It presents the standard abstract model for archival information systems and the way it was instantiated in the pilot project. It ends with some reflexive remarks on the factors which are bound to act upon the future of such projects: organisational behaviours, role of data and knowledge, communities of users, institutional issues, status of collective memory in SSH.
Keywords: long term preservation, OAIS, Open archival information system, Information model, Representation information, digital archives, data and knowledge transmission, Large Research Infrastructures, sustainable data, knowledge communities

Transmitting past and present research: digital data and long term access to it

Current research in social sciences and humanities (SSH in the sequel) is often deeply rooted in the past. For instance, Claude Lévi-Strauss's death, in the fall of 2009, was the occasion for stressing the actuality of the framework he designed. *Tristes Tropiques*, one of his main contributions, is definitely part of what modern anthropologists and ethnologists, among others, (should) read, even if it was published in 1955, more than half a century ago. In the sixties, the sociologist Edgar Morin led an interdisciplinary study of a small village in Brittany (Bretagne): Plozévet. Nowadays, several actors – historians, sociologists, ethnologists, etc. – try and use the reports and data from this survey in order to get some background about the evolution of the region in the long term and to help designing hypotheses about its future. Meanwhile in France, the researchers who pioneered SSH in the sixties are now retiring: their results, the data and the archives they gathered are endangered. For past research, this situation results in an increasing request for digitizing. Even though transmitting in up-to-date – ie digital – form their field data, their corpora, and so on, represents an abstract duty for retiring researchers, it faces several obstacles. The first one is an overall individualist type of research (some disciplines such as archaeology rely more on team work than other ones, such as philosophy – within a discipline, some specializations favours more a collaborative work than others, for instance corpus linguistics as compared with intuition-based linguistics). A universe of craftsmen finds it difficult to agree on a common policy in digital matters. What's more, the average technical equipment and computing skills and knowledge are rather low. Thirdly, there is little awareness about long term digital preservation issues: digital data is supposed to be reliable in the long term. For present research, the projects in SSH supported by the French funding agency for research, ANR (Agence Nationale de la Recherche), since its creation, in 2007, are very often supposed to deliver digital results (corpora, web sites, prototypes, data, etc.). So far however, there is no policy, no common guidelines for long term preservation of these results. It is not even sure that the problems at stake (see next section) are really understood or even perceived: current digital data are bound to get lost in ten years time if nothing is done. SSH thus faces the risk

of (re)producing data and knowledge, either natively digital or digitized, which can become useless in the long run.
In order to help understanding the possible interplay between transmission and digitization, we are going to present a pilot project of long term preservation for research data in SSH. This project was designed and founded by a Large Research Infrastructure, the TGE Adonis (see third section). The first author of this paper was then deputy head of the TGE Adonis: he was leading the project. The second author had been in charge of digital archiving at the French space agency (CNES) and was responsible for the French workgroup on digital preservation [PIN]: he was in charge of the technical and organisational coordination of the project, as a consultant. This paper is therefore an account of this project by two actors, coordinators and leaders of it. In the context of the recent evolutions of collective memory in our society (see second section), we try and give some insights on the organisational and technology issues which are at stake, and on the relationship between the two of them. We are not able to provide a sociological analysis of the project as such but we hope nevertheless to bring some elements for such an analysis which is most needed.
The paper is organised as follows. The second section provides some background context on transmission in digital form of past and present research in SSH. It shows the discrepancy between the increasing role of digital information and the fragility of it. The third section presents the need for research infrastructures in SSH on the pattern of the ones developed for physics and other sciences. The fourth section explains the rationale for choosing oral data for a pilot project within a research infrastructure for SSH. The abstract model for archival information systems constitutes the fifth section. It is the framework for the main preservation projects, whether French or international. The way this abstract model was instantiated for the pilot project is detailed in the following section. The next section tries and understands the organisational behaviours to develop in order to get a sustainable framework. The last sections present some reflexive remarks on the factors which are bound to act upon the future of this project: organisational behaviours, role of data and knowledge, communities of users, institutional issues, status of collective memory in SSH.
NB: within the paper, web sites and pages are referred to via acronyms between brackets and in capitals (e.g. [PIN] above), so as to distinguish them from ordinary bibliographic references. The acronyms and the corresponding URLs are given in the reference section. The day of visit is March 2010 the 15th.

Digitization: impending "bad memories"?

An obsession with memory, whether individual or collective, seems to arise in the last quarter of the 20th century. It persists and develops in the very beginning of the 21st century. This phenomenon has been analyzed by R. Robin (2003) and E. Hoog (2009), among others. Let us give some evidence of this trend. Hoog (2009: 109) quotes a study stating that 420 thousand millions of snapshots were taken in 2007, that is about 50 millions per hour. Pierre Nora edited between 1984 and 1992 a 5,000 pages collection on French spaces of shared memory ("Les lieux de mémoire"), with a broad meaning of "spaces" (events, symbols, mottos are studied as well). It was a huge editorial success and it is now available in paperback. It is as if the French people, "overwhelmed with history" (De Gaulle), is nowadays reluctant to think of its future and is looking shelter in the past, its preservation, or even its "embalming". For instance, six national commemoration days were chosen in France between 1880 and 1999, ie in more than a century, and six others quite recently, in a span of only 6 years, between 2000 and 2006 (Hoog 2009: 52). This tendency – re-working or even re-inventing roots – seems to be shared among "old countries", including the USA, but as well among "new" ones, that is the countries emerging from the reorganization of East Europe. The pervasiveness of digital data and documents has major consequences on this obsession with memory. From now on,

an image, a sound, a video, a text, a program are represented in a uniform way as a sequence of bits, of 0s and 1s. Such sequences can be stored on hard drives, DVDs, CDs, and so on. There is no more predefined association between a given information and a medium, as it was previously the case for instance between text and paper, or sound/video and tape. The size of the storing devices is growing at an astonishing pace. On the one hand, preserving memory very often amounts to digitizing analog data or documents. Heritage institutions, such as the French National Library (BnF) or the French Archival Board (DAF) have been funding important digitization projects in the last twenty years. On the other hand, "native" digital data and documents are every day more central in our lives, as we sadly realize when we crash a hard drive or when our laptop is stolen. We are continuously producing and consuming data in a digital form. Its long term preservation should be of concern to us.

Surprisingly enough, this ever growing production of digital data, whether native or digitized, does not prevent most of us to be in ignorance of the real life expectancy of the resulting information. Peoples fear to loose "their past". However, digital memories could quickly transform themselves into bad memories. As a matter of fact, for several reasons, digital data is very fragile data. Media for digital information are not growing old gracefully. It is the other way round. For instance Google stated that 8% of their two or three year old hard drives have to be replaced as they are not working properly (Hoog, 2009: 114). As formats evolve rather swiftly, data gets locked away when its format is not usable any more. It is obviously the case for text processing, but this is more acute in other areas. Each of us possesses digital documents (data, texts, images, videos, programs) (s)he is not in position to use any longer. Software, which in fact represents our gateway to data, dies, as we will, but rather sooner than us. Data gets lost as well because there is no longer access to it: there are no metadata, or metadata are too terse, or the actual wording of them is not the current way of talking about the underlying facts. We enter a digital world in which actors find it difficult to state what is precious and thus should be preserved and what is not. Therefore, mega, tera and petabytes are stored as the underlying information "can prove useful". "Nowadays we do not preserve something because it is important, but on the contrary, because it is preserved, it can become important. Rather, its preservation allows it to become important someday" (Hoog, 2009: 121). That is what Robin (2003) calls *The overloaded memory*: the present and the foreseeable futures are not clear enough to help communities in choosing what is to be preserved from their past and in making a clean slate of the rest of it. Because we are not able to perceive the part of the past which actually can help us for the present and the future, we resort to huge amounts of unstructured data and powerful research engines to cope with them. That is the rationale for the size of the Web and for the processing capacities it requires. For instance Google data centres amount to 200 petabytes of mass storage, that is 200 millions of gigabytes. To help in grasping such figures, let us say that the 280 pages of the book (Banat-Berger et al., 2009) in text only represent 600,000 bytes and 3.5 mega-bytes in ODT format with images and layout: more than 57 thousand millions of the 3.5 mega-bytes version could be stored in Google data centres.

In sharp contrast with the "layman" rather blind confidence in the current digital "way of life", the vulnerability of digital information gave rise to several initiatives in the past ten years. They tried to associate technical, educational, methodological and organisational points of view. Space research played a pioneering role (see below), but culture and heritage entities soon joined the field, as exemplified by the project [InterPARES] in Canada or, in Great Britain, by the Digital Curation Centre [DCC] and the Joint Information Systems Committee [JISC]. Several European projects (see below as well) show an emerging awareness or even lucidity in those matters: Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval [CASPAR], Preservation and Long-term Access through Networked Services [PLANETS], Permanent Access to the Records of Science in Europe [PARSE], and so on.

Enabling long term research: the need for Large Research Infrastructures (LRI)

A Large Research Infrastructure (LRI) is designed and developed in order to put at the disposal of a given scientific community an "instrument" whose cost dramatically exceeds the limits of the short term and small or medium budgets available to every entity of this community. A LRI only makes sense when it is actually used by an important scientific community, which is often scattered all over the world (or within a country).

The first LRI were built for physicists, very often at an international level. The Large Hadron Collider, the world's largest and highest-energy particle accelerator, is one of the most famous LRI currently in use. It was inaugurated in October 2008. It is expected to cost about 3 thousand million Euro. It relies on the collaboration between more than 10,000 scientists and engineers from over 100 countries. The construction and operation of such a LRI requires much planning and cooperation, for instance to set the principles of use for the community. The actual LRIs resort to various legal frameworks and business models.

The very success of the LRIs in physics led to an extension of the concept for other scientific areas. For instance, in astronomy, were organized Virtual Observatories (VO). In this case, the instrument is not a huge telescope. A Virtual Observatory relies on software tools and interoperable data distributed among the nodes of a network of actual observatories in order to conduct astronomical research programs with a transparent access to the original data. Astronomers are in a position to combine and to analyze data coming from originally heterogeneous data collections. This interoperability implies agreeing on data formats, possibly on data standards, on ways of processing the datasets. Virtual Observatories had a leading role in developing unique identifiers for publications of the kind of the Digital Object Identifier [DOI]. Such a unique reference is necessary to prevent the "missing link" message so frequent with digital publishing: it allows for persistent identification, it gives access to the bibliographic resource even if its actual location has changed. On the whole, for the astronomy community, Virtual Observatories consist in progressively building standards for sharing data, publications and software tools.

The need for similar tools in Social Sciences and Humanities slowly came out, whether at an international or at a national level. The implicit model was the Virtual Observatories, as very few disciplines in this area require "physical" heavy instruments. The European Strategy Forum on Research Infrastructures [ESFRI] delivered a first roadmap in 2006. It was updated in the fall of 2008. It comprises several Research Infrastructures for SSH [RI-SSH-EU]. A Research Infrastructure named [SYNERGIES] was started in Canada. It relies on a network of 5 major universities, from coast to coast.

In France, in 2007, the CNRS (National Centre for Scientific Research), which still is the main research agency (see next section), with 26,000 tenured employees (11,600 researchers and 14,400 engineers and support staff), and which covers all the fields of research, decided to create a SSH Large Research Infrastructure (as opposed to a research program), called TGE (Très grand équipement – Very large equipment) [ADONIS]. The scope was broad: a unified access to digital data and documents in SSH, that is to scientific journals and primary data. TGE Adonis targets all SHS researchers: most of them are not CNRS tenured employees, but university lecturers or professors. Its intended audience is francophone but it is involved as well in a European project for building a LRI in humanities: DARIAH [RI-SSH-EU]. The contributors are teams and laboratories, the clients laboratories and researchers. The policy followed by TGE Adonis is then to try and organize operators sharing a common goal or functionality in order to get open, interoperable and shared solutions. Its lever is the available budget for such solutions (from 2007 to 2010, TGE Adonis budget was between 2 and 3 million Euro per year). For instance, TGE Adonis funded projects in which two models for publishing digital journals were associated, a "free" one – [CLEO] and a "paying" one

[CAIRN]. Because of the contrast between the need for long term archives in SSH (first section) and the vulnerability of the current digital data (second section), for TGE Adonis, long term preservation of digital data and documents was chosen as a major objective, in order to get reliable data and to reduce as well the costs by sharing equipments and human resources and skills.

A pilot project for long term preservation of spoken data in SSH (2008-2009)

The underlying motto of the past five years in France seems thus to be: big science is good, or even science must be big if it wants to be good science. It was advocated that, for instance, the CNRS should be reorganized in large institutes, following the model in physics of the IN2P3, Institute for nuclear physics and particular physics. The IN2P3 is a network of 20 research laboratories. It gathers 2,500 employees all over France and has a budget of 42 million euro. It heavily relies on a computing centre: the [CC-IN2P3], which is itself a Large Research Infrastructure. The universities are being reorganized as well. The implicit objective of the 2007 law about the liberty and responsibility of the universities is to get a small set of powerful universities able to appear in the international research context, with the remaining universities playing a more national, regional or even local role.

The "big science" policy came as a shock to the SSH. Its small or even tiny communities, its reliance on "craftsmen", its lack of technical skills and interests, its heterogeneity explains that steps must be taken to have the SSH adapt to this new and rather hostile environment. This analysis was the rationale for the TGE Adonis strategy for long term digital preservation. A global solution for each and every field in SSH was obviously not an option.

In November 2007, the TGE Adonis ordered an expert report on needs and offers in digital archiving for SSH. It was decided to resort to a foreign expert – O. Bärring from the European Organization for Nuclear Research (Geneva) – so as not to be involved in French peculiarities and divisions. O. Bärring delivered his report in February 2007 [BARRING]. His main conclusion was to rely on existing large computing centres, with already important resources, rather than to create from scratch a new entity dedicated to this function. The purpose of such a choice was an economy of scale. The SSH archiving would only constitute a tiny activity for these centres. Huge and costly investments in servers, hard drives, infrastructure, network and so on would not be necessary. Such a solution would as well benefit from the abilities of the engineers and technical staff of these centres, as far as storage or exchange of massive data were concerned. The specific skills involved in digital archiving (see below) would develop better in such a "rich" scientific and technical environment than in the SSH world.

In order to prove the feasibility of a digital archiving process in SSH, in March 2009, the TGE Adonis Steering Committee chose to fund a pilot project, limited to spoken data. Spoken data mainly consist in dialogs or monologs recorded for linguistics research (language learning, spoken language syntax and overall functioning – such as so called disfluencies or forms of interaction) and for language engineering (training data for speech to text systems). But they can as well be complemented with measures of physiological parameters involved in speech production (electroglottography, articulography, palatography, and so on). Most of the time, spoken data do not stand alone. In order to be of any use, they are annotated or enriched: textual transcriptions of the utterances, analysis in sounds (phonemes) or syllables, indications about rhythm, stress and intonation of speech (prosody), etc. In any case, each recording and each annotation is associated with metadata: place and time of recording; age, gender and social professional group; language(s) and dialect(s) in use. These metadata are mandatory for research purposes.

The reasons for the choice of spoken data as a first step were the followings. Firstly, spoken data are produced and analyzed by various scientific communities: psycholinguists,

phonologists, syntacticians, sociolinguists, computational linguists, language engineers, and so on. A given recording and the associated annotations thus can be studied and enriched from very different points of view. In these fields, the more the better. Giving access to all the existing data, when possible, with standard formats and metadata, helps in building upon previous results and in having a more cumulative type of research. It leads to new advances and in the same time to changes in habits and methods, since sharing for new purposes data "born" in a different subfield is not so far a common practice. What's more spoken data have a growing importance from a heritage point of view. At least since the Villers-Cotterêt edict signed in 1539 by King Francis the 1st, which gave French the status of the only official language for administration, other languages than French have be fought against, first during the monarchy, then since the Revolution and the Republic. Quite recently, however, in 2008, the revision of the French constitution included regional languages in the French heritage. Apart from their official status, regional languages and other languages spoken in France still play an important role in personal and collective identity. That's why the [DGLFLF], the French Delegation for French and languages spoken in France funded in the past six years corpus production and enrichment for those languages and funded as well a portal dedicated to the resulting resources [DGLFLF SPOKEN CORPUS]. The DGLFLF supported and edited as well a good practice guide for spoken corpora, which is being translated into English and which is freely available on-line (Baude, 2006). This guide gathers the point of views of the relevant actors or types of actors. It covers a broad range of topics: conventions for transcriptions, legal aspects, technical equipment, recording formats, metadata standards, etc. It shows an increasing level of maturity of the underlying communities, as they worked together in order to produce it, and, by doing so, precised their convergences and divergences. Secondly, the field of spoken data is rather well structured from an institutional viewpoint, as compared with other parts of SSH. The CNRS division for linguistics created two federations of laboratories, in order to foster cross-laboratories and cross-subfields projects. These federations lead a network of researchers, of teams and of laboratories which produce and analyze spoken corpora. They contributed to DGLFLF good practice guide. They led projects whose results were made public on DGLFLF portal. Furthermore in 2006 the CNRS created five centres for digital resources in SSH [CRN]. Each one of these Centres was devoted to a given type of resource: text, image, geographical data, manuscripts. The centre for spoken data [CRDO] – was distributed between two locations, Aix-en-Provence and Paris. During the period 2006-2008, the two locations mainly worked separately. However, they helped the spoken language communities and individuals in giving advice, in providing portals and tools, in transmitting standards and skills. Lastly spoken data represent a rather good test bed for an archiving project. There are neither too complex nor too huge, as compared for instance with the 3D simulation data produced in archaeology. At the same time, they are already a serious challenge. At the beginning of the project, in September 2008, the amount of data available via the CRDO was about 2 terabytes (about 1,500 recordings – mostly audio – with their annotations). There was sound recording, text transcriptions and annotation, but video recordings as well. Legal problems related to these data had been opened up in the DGLFLF good practice guide but were not entirely solved.

"Tractable" volume and complexity of data, institutional organization of the field, strength and dynamism of the communities of researchers and users, these three features were convincing arguments for choosing spoken data to set up an experimental process. In March 2008, the TGE Adonis Steering Committee launched the pilot project. Following one of O. Bärring's proposals, the Steering Committee wished that the actual archiving infrastructure could be hosted, in cooperation, by two large computing centres, the [CINES] in Montpellier and the [CC-IN2P3], in Villerbanne, nearby Lyon. Additionally the Steering Committee decided an external and formal evaluation for Easter 2009. Finally, the Steering Committee

asked the TGE Adonis to work on an agreement with the French Archival Board (DAF). As a matter of fact, the DAF is legally responsible for all the data produced during their work time by civil servants such as tenured CNRS researchers and university lecturers or professors. This situation implies that an entity archiving such data must get a formal delegation of powers from the DAF.
In spite of previous contacts between the intended actors, the actual kick-off of the pilot project was during the Summer school of the TGE Adonis, in September 2008, in which all protagonists were invited: CINES, CC-IN2P3, CRDO and other centres for digital resources. In June 2009, Yves Marcoux, from the University of Montreal, delivered his evaluation on the project. In October 2009, the French Archival Board assessed the project as well, in order to proceed towards a delegation of powers for spoken data.
In the next section, we present the reference model of digital archiving. The section afterwards explains how it was tuned for the pilot project.

Understanding what is at stake in digital preservation: the OAIS reference model

As early as the end of the 60's', space research pioneered the massive use of digital techniques. Information transmitted by spacecrafts necessarily was electromagnetic; its volume implied automatic processing. Harvested information was very often unique and irreplaceable. When a comet approaching the Earth or a Solar eruption are being observed or when a precise map of the forests around the Earth at a given date is being drawn, one cannot afford to loose the resulting information, as it will not be possible to reconstitute it. Right from the beginning of the 90's, after twenty years of building up observations and after the first trying technological transformation in digital processes, the space research community became aware of the urgency of organizing long term preservation of space observations. National space agencies, the NASA in the USA and the CNES in France, had already worked out temporary and pragmatic solutions. However, a really normative and comprehensive framework was dramatically required.
The Consultative Committee for Space Data Systems [CCSDS] provided the structure for working out such a framework. It is a forum of the major space agencies. It develops recommendations for data- and information-systems standards. It is as well the ISO (International Standards Organization) committee for spacecrafts. Engineers from the CCSDS were asked to propose a norm for long term preservation of space observations. Fortunately, their answer was preserving the future on two major issues:
i) As far as long term preservation of digital information is concerned, there is nothing special to space observations. On the contrary, researchers who face the same problem in other fields should be included in the normalisation process.
ii) As digital technologies, whether hardware or software based, constantly and quickly change and disappear, it is better to propose an abstract reference model. Such an abstract model will provide all the necessary concepts to understand and solve current and future problems. It will accommodate for the actual state of the technologies. As far as technology analysis is concerned, this model does not prescribe any specific artefact, whether hardware or software. It describes high-level entities and processes relating them. The two of them must be instantiated. That's why this model will be long-lasting. An implementation norm would too closely depend on the current state of the technology and would therefore become quickly obsolete. It is possible to use for this model the "term *interpretive flexibility* […] to refer to the degree to which users of a technology are engaged in its constitution (physically and/or socially) during development or use. Interpretive flexibility is an attribute of the relationship between humans and technology and hence it is influenced by characteristics of the material artefact […], characteristics of the human agents […] and characteristics of the context […]"

(Orlikowski 2000: 409). OAIS has been designed to have interpretive flexibility and it is one of its main strengths.
These positions led to the Reference Model for an Open Archival Information System [OAIS]. This OAIS Model was normalized by the ISO en 2003 (ISO 14721). We present its main features.
An archive (singular in the OAIS Model) consists of an organization of people and systems, that has accepted the responsibility to preserve information and make it available and understandable for a designated community over an indefinite period of time. The community should be able to understand the information in its current context, with its current experts, but without the assistance of the experts who originally produced the information. This means that metadata are of an utmost importance.
The OAIS model is twofold: an information model and a functional model.
The information model describes the types of information that are exchanged and managed within the archive. Information is any type of knowledge that can be exchanged, and is expressed by some type of data (in OAIS terminology, information works as the hypernym of data and knowledge). The archive must understand the knowledge categories of its designated community and thus must store significantly more than the contents of the data object it is expected to preserve. The content information is the set of information that is the original target of preservation by the archive. The archive must choose the minimum representation information that must be maintained. The representation information accompanying a digital object provides additional meaning by first mapping the bits into commonly recognized data types (character, integer, strings, records, etc.) and secondly associating these data types with higher-level meanings that are defined and inter-related in the designated community (for instance in an ontology). The content information is the content data object together with its representation information. There is as well reference information, which identifies and describes one or more mechanisms used to provide assigned identifiers for the content information and also provides those identifiers, context information which documents the relationships of the content information to its environment, provenance information which documents the history of the content information (origin or source, changes and custody, fixity information which provides the data integrity checks or validation/verification keys used to ensure that the particular content information object has not been altered in an undocumented manner).
The functional model states the repartition of responsibilities between the archive, the external participants – the producers of information, the consumers (OAIS terminology) or users – and what is called the Management entity. The Management defines the scope and the mandate of the archive. It often provides as well the funding and resources necessary to its functioning (OAIS terminology). Figure 1 shows the interactions and the functions which make the archive. The producer provides the information. The Management sets the overall policy (not the day-to-day operations). The consumer finds and acquires preserved information of interest to her/him.

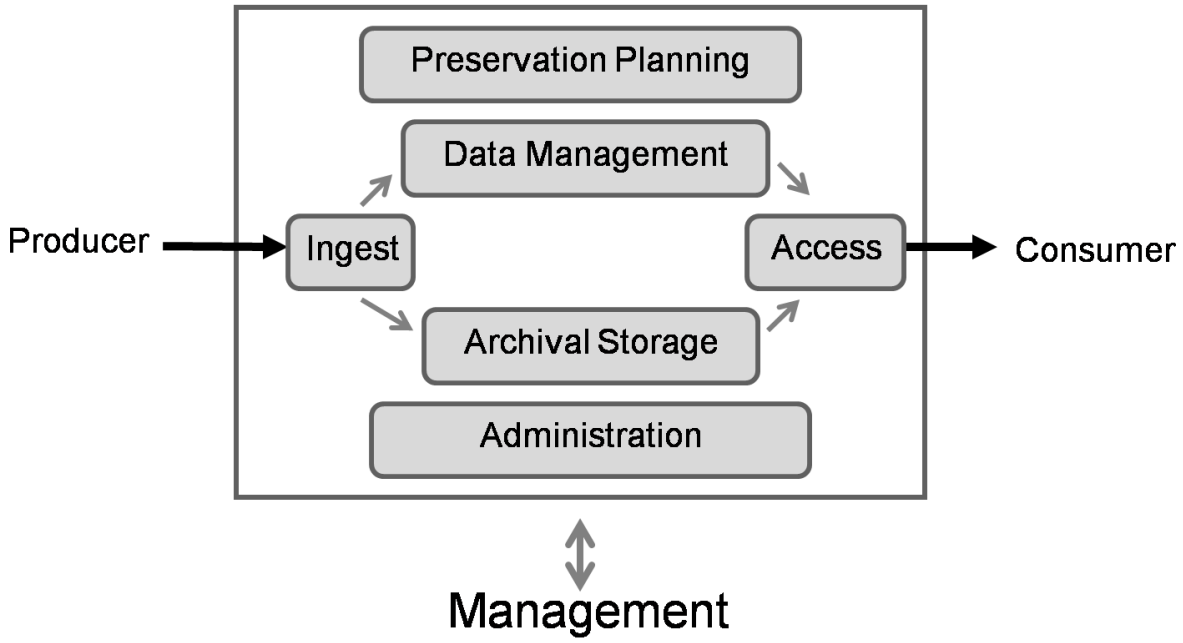


Figure 1 : The OAIS functional model

There are six functional entities and related interfaces: Ingest, Archival Storage, Data Management, Administration, Preservation Planning, Access.
The unit of exchange within an archive or between an archive and its surrounding environment is an Information Package. Between a producer and Ingest, it is a Submission Information Package, within the archive an Archival Information Package (and between Access and a consumer a Dissemination Information Package.
The Ingest entity provides the services and functions to accept Submission Information Packages from producers and prepare the contents for storage and management within the archive. The Archival storage entity stores, maintains and retrieves the Archival Information Packages. The Data management entity maintains both descriptive information which identifies and documents archive holdings and administrative data used to manage the archive. The Administration entity is in charge of the overall operation of the archive system, including: auditing submissions to ensure that they meet archive standards, and maintaining configuration management of system hardware and software. The Preservation planning entity monitors the environment of the archive and provides recommendations to ensure that the information stored in the archive remains accessible to the designated user community over the long term, even if the original computing environment becomes obsolete. The Access entity supports consumers in determining the existence, description, location and availability of information stored in the archive, and allows them to request and receive information products.
By the end of the 90s, and even before its formal ISO normalization, the OAIS Model was thought of as the conceptual reference for long term preservation of digital contents.

Implementing the OAIS Model

There is an important number of implementations for the OAIS Model. In France, the digital archiving infrastructure built by the French National Library (BnF) – the [SPAR] system – goes very far in complying as strictly as possible with the model.
The TGE Adonis implementation for the OAIS Model relies on two large computing centres. This choice permits to avoid major initial investments. The distribution of functions and of responsibilities is shown in Figure 2.

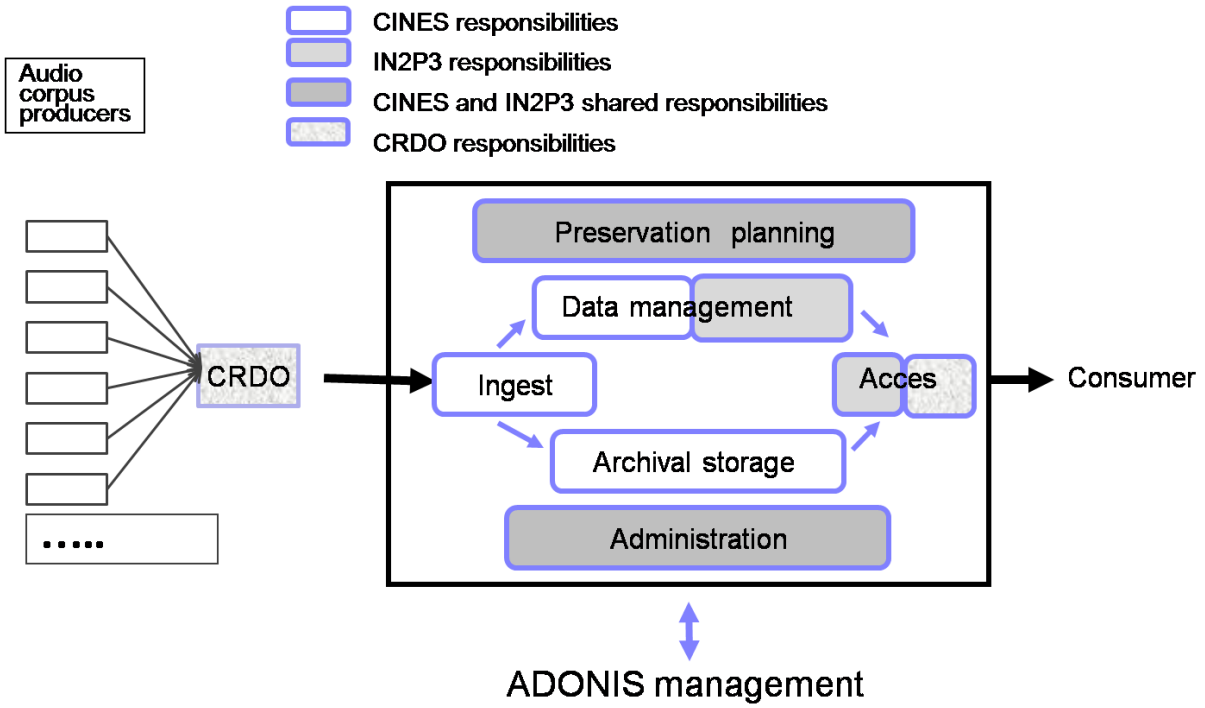


Figure 2 : The functional outline of the TGE Adonis shared infrastructure

In 2007, the National Computing Centre for Higher Education [CINES] has been given a mission by the Ministry of Higher Education and Research in the field of long term preservation of digital information. As a matter of fact, this mission is the recognition of the infrastructure, the skills and the knowledge the CINES had developed in this area since 2004. This competence is manifold: standardization of the structure of the Submission Information Packages transmitted by the producers; definition of mandatory preservation metadata; choice of admissible formats for long term preservation; in depth control and validation of the conformity to these formats of all the submitted data; multi-site replication of the archived information; migration strategies; renewing media; digital stamps to ensure the integrity of the data, and so on. The preservation platform of the CINES [PAC] brings together these abilities and the underlying technical infrastructure. The CINES is responsible for the archiving of digital versions of PhDs from various French universities, of the open-access platform for scientific publishing, [HAL], which is the French equivalent for arXiv, of digitized issues of journals in SSH which are no longer submitted to a temporary embargo [PERSEE] and of spoken data (TGE Adonis pilot project). Apart from the ability to give back to the producer its data or metadata "as they were" when they were submitted, the CINES has less experience on the access part of the OAIS model. The access function is played for PhDs thesis by the Agence bibliographique de l'enseignement supérieur, for HAL data by HAL, by Lyon 2 University for the [PERSEE] project and by the CC-IN2P3 for the pilot project on spoken data.

The Computing Centre for the Institute of Nuclear Physics and Particular Physics [CC-IN2P3] is a Large Research Infrastructure. It provides computing and storage resources to a very large community of users at a national and international level. One of its main contributions consists in abstracting computing and storage. Its users have access to data and computers all around the world in a homogeneous way, without having to pay attention to location, operating systems, file management systems, and so on. They are not bothered any more with interoperability between systems and with the associated technicalities. They can concentrate on the problem at hand and develop solutions which are truly independent of the current tools and equipments. Because of the amount of data produced by experiments in physics (for

instance, the Large Hadron Collider is expected to produce a total data output of 15 petabytes per year), the CC-IN2P3 has huge storage capacities: 5 petabytes for on-line data (2.5 thousand times the volume of the spoken data pilot project), 4 petabytes in hard-drives and 30 petabytes on storage cartridges. Within the TGE Adonis pilot project, the CC-IN2P3 develops and maintains the generic functions to access and use the data. It relies on the Open source software Fedora Commons [FEDORA] which adds a layer of abstraction on the access functions. This repository software for instance makes it possible to associate to a given archive format several access formats which can be more relevant to the end user. An image can be archived in TIFF format, which is space-consuming and which has too many options to be sure to get an appropriate driver for each of them. Fedora Commons can produce beforehand or on the fly thumbnails or JPEG versions of it, which will be easier to consult on-line or to download. Fedora Commons as well provides tools to manage fine grained access rights. As it offers rich metadata (based on a W3C standard, RDF), this layer of abstraction is potentially relevant for all fields in SSH.

The CINES and the CC-IN2P3 do not have specific insights on the data and the underlying notions and representations of the spoken data communities and of the SSH as a whole. The CRDO comes into play precisely as a mediator between the relevant research communities and the archiving infrastructure.

As the other Digital Resource Centres presented above, the CRDO was created to gather and develop technical competences about a certain type of resources, spoken ones in this case, in order to compensate for the overall context in SSH. Most of the time, researchers do not have the technical equipment, the staff and the collaborations necessary to comply with the standards in their area (which often they are not aware of). Before the pilot project, the CRDO knew precisely how researchers used to produce and analyze data (habits, formats). It was therefore in a position to define the relevant terminology and processes with the computing centres in charge of the submission and the access parts of the archiving infrastructure.

Somehow, in French, the very word ‘archives’ has a “dusty”, bookworm connotation. It reminds of long forgotten books, papers or even manuscripts, which are accessed now and then only for history researches. It is not the case in English for ‘archive’ (see the role of arXiv) or ‘repository’. As a matter of fact, these words, at least for digital contents, do not imply long term preservation, but rather focus on easy access. Anyway, in SSH, a long term preservation infrastructure needs to ensure that it does not deliver a “still life” of the research. On the contrary, current research must be able to use past research, as secondary data for instance. New analyses of primary data must be linked to it, in order for them to be falsifiable. It is often necessary to mend an annotation (when transcription conventions are updated, for instance) or to change metadata (for instance when access rights change). New corpora must be made available as soon as possible. Therefore archived data are living, ever evolving material.

The common definition of the archiving process between the actors of the pilot project led to some evolutions for the CINES procedures. It was necessary to certify new video recording formats, to cope with new versions of the annotations or of the metadata. One major change was adding relationships between data. For spoken data, it is crucial to link an annotation to the audio or video recording it refers to. As a matter of fact, this possibility, which was not present in the original version of the CINES archiving process proved recently useful for HAL archiving, as an author can upload several related versions of the same paper.

The main assets of this pilot infrastructure are its reliability, its genericity and its extensibility. The CINES and the CC-IN2P3 are experimented and stable large computing centres. They are not bound to disappear in the near future and will suffer less than other entities from the current reorganization of higher education in France. The solution, whether on the hardware side or on the software side, is a generic one: it can be used by other fields of SSH with little

changes. It thus relieves researchers and technical staff in SSH from problems which do not really belong to their domain and for which they should not have to develop an expertise. On the contrary, they can concentrate on their job and be better at it. Extensibility: each domain of the SSH can keep on using its specific metadata while at the same time relying on common agreed ones, which are intended to be a shared core, in the spirit of the [DUBLIN CORE]. As a matter of fact, at the very beginning of the project, a meeting with the other existing Digital Resource Centres was organized so that the initial choices of metadata for spoken data could be compatible with this extension perspective. There is evidence for another adaptability dimension. For the time being, each part of the [CRDO] chose a different location for their access interface, the CC-IN2P3 for the first one, a distant site for the second one. This flexibility helps researchers and technical staff in laboratories in choosing the level of sharing with the infrastructure with is the most appropriate and in changing it whenever necessary. The other Digital Resource Centres thus can make their choice according to their projects, their forces and the type of resources they are dealing with.

Transmitting digital content: building shared representations

In attempting to build a lasting archiving infrastructure, the main difficulty is building shared representation between all the actors who are involved. They need to agree on the way the data and the metadata are organized, on how it is going to be accessed and used. Even more crucially, the overall process and the precise division of responsibilities must be agreed upon.

As we both were in charge of the pilot project, our main task was to try and build a well-knit and motivated team. Each member of it was to continuously get both a clear and precise overview of the project and its evolution and of the functions and parts he was responsible for. This objective implied a precise and up-to-date definition of tasks and schedule. An adequate rhythm was as well necessary for meetings. The team had audio-conferences twice a month and met every month and a half. Minutes of each meeting were always taken and quickly made available on the project wiki [ARCHIVE WIKI] so as to plan future tasks. The wiki helped as well in specifying each one's role and in sharing a common vocabulary in a glossary section.

As a matter of fact, in such a project, the agreement on technicalities (network requirements, authentication mechanisms, storage devices, servers), on norms such as OAIS and on metadata standards – [DUBLIN CORE], etc. – represents only the outcome of a long and difficult process: sharing a common solution in which each participant willingly fulfils his role. What is at stake is not "implementing the OAIS Model", but finding together a possible meaning for it in a specific context. In the end, the solution will be reliable if and only if a deep agreement is obtained, on the overall scheme as well on the detailed procedures. That is why it is a lengthy operation which cannot be shortened. In our case, more than a whole year was necessary to get to a stable test state and to be able to launch the production phase (March 2010). Obviously, from time to time, there were and are hot discussions and disagreements. The functional and technical coordinator must therefore remain as neutral as possible so as to be in a position to help in solving potential conflicts. This position prevents the coordinator from being as well a partner.

In natural sciences, a measuring device is the result of a complex evolution in a community which progressively agrees on what is to be measured, and on the kind of errors and measurement uncertainties which are to be accepted. Afterwards the device is used as a "black box" (Latour), without further questioning about its way of working. This situation is similar to the one described by A. Desrosières (2000: 406) for the development of statistical notions and indicators from the 18th century up to now. A contradictory deliberation about the available choices for a city-state implies a common language to represent the entities, to word the aims of action and to discuss its results. This language does not exist before the debate. On

the contrary, it is negotiated then stabilized for a given period. Afterwards, it can be altered or dropped. This language creates distinctions which did not exist as such before and offer new ways for measuring and changing the world. For instance, there is currently a hot debate in France about ethnic statistics which so far are not legal. The mere existence of such statistics and the range of distinctions which would be used would change the way French people think and behave in a lot of domains: education, employment… In other countries, the USA for instance, such a "language" is part of everyday life. Such languages are conventions which need to be stabilized at least for a while to be of any use, to help organizing collective action. But these conventions are nevertheless questionable and can be entirely changed.
The development of long term archiving in the past fifteen years can be considered as the progressive conventionalization of such a language. We certainly are at the first step of such an evolution. The production of the OAIS Model as a CCSDS standard and as an ISO norm contributed to this evolution. The participants of the pilot project adapted to their context the OAIS glossary and relationships. It took months. Like a measurement device, the actual implementation, if it succeeds, should become a black box, for the data producers and for the end users, with clear requirements and protocols for submitting data and for using it. However, it is still a glass box as its notions and distinctions are not shared by the layman in SSH. Therefore it calls for a patient work within the SSH communities

From data to knowledge in SSH

Neither of us is a sociologist, nor even a SSH scholar. However each of us was involved in cultural translation activities, between computer science and linguistics, space research or physics. That's why we try and pay special attention to differences of meaning for "storage", "preservation", "information", "data" and "knowledge" between computer science and SSH.
From a computer science point of view, there is an opposition between non structured data (raw texts, audio and video recordings) and structured ones. A relational database, for instance, corresponds to a conceptual analysis of a sub world, its entities and their possible relationships. It records facts about this sub world. In this respect, this database stores knowledge about this sub world. In the closed word hypothesis, it records all which is known and which is relevant about it (Habert 2009). In the late 1990s there was a trend for capturing, formalizing tacit and explicit knowledge in specific areas or more generally. The key words were Knowledge Management, Knowledge Management Systems (Galliers & Newell 2000), terminologies and ontologies, such as the Unified Medical Language System [UMLS]. The tools were expert systems, conceptual graphs, logic programming languages (such as Prolog), and the like. The first version of the vision of a nearly full "semantic web" by Tim Berners-Lee in 2001 was somehow an extension of that dream. The recent versions of the project are more modest [SEMWEB]: "providing a common framework that allows data to be shared and reused across application, enterprise, and community boundaries". As a matter of fact, it is possible to analyze the failure to produce large-scale and usable knowledge systems as the origin of two complementary and recent trends for formalizing knowledge. The first one tries and modelizes not an entire domain but some parts of it. This is the realm of XML. It is a meta-language which permits to state and to valid constraints on a "document" (in a broad meaning): its structure, what is compulsory, and so on. For instance, a submission information package for the pilot project is an XML document following a pre-defined XML "grammar". The system can then valid this document according to this grammar: it makes sure that all needed information is in fact provided, in the expected form and at the correct place. There is as well an XML "grammar" for the commonly agreed upon set of metadata [DUBLIN CORE]. This makes it possible to certify the metadata attached to a given digital object. The second trend helps in formalizing not a whole document but "bits of information". That is the role of RDF [SEMWEB], which is as well an XML grammar. This Resource Description

Framework, supported by W3C, the consortium which organizes the Web, has three features. Firstly, information is split in triplets: <entity>, <predicate>, <value>, such as 'CRDO', 'produces', 'spoken data'; 'TGE Adonis', 'supports', 'CRDO' and so on. Secondly, these triplets can be connected by common entities, such as 'CRDO' in the example, in order to build up knowledge bases, which are (possibly non-connected) graphs. Lastly it is possible to state in it that whoever/whatever (it can be a program – see below) wrote the triplets has her/his/its doubts about it. The frontier between truth and knowledge can be drawn, as Galliers and Newell (2000) advocate. With these two trends, the distinction between data and knowledge (Galliers & Newell 2000) becomes fuzzier. An XML document is a model of a part of a sub world. Documents which are valid according to a given XML grammar combine data and knowledge. The ratio between the two of them depends on the precisions of the constraints from the grammar. More specifically, the structure inserts an interpretation, which can be coarse or precise. On the other hand, a document can be unstructured on the whole but some parts of it can be automatically formalized as RDF triplets, as "molecules" of knowledge. That is the rationale for the [DBPEDIA] projects which aims at extracting RDF triplets from Wikipedia so as to build a huge knowledge base. If we compare these two compromise solutions between data and knowledge with the situation in the 1990s, let's make clear the difference. Nowadays data and knowledge processing heavily rely on XML, as opposed to the less central role of knowledge management and knowledge management systems at that time, which in the end did not scale up. For instance, RSS feeds on web sites are in RDF, which helps in merging ('syndicating') contents.

The data which are and will be preserved by the archiving infrastructure fully belong to this paradigm (Habert 2005). They are not "raw data". On the contrary, they use XML grammars to constrain their metadata and the annotations and enrichments. The amount of knowledge such semi-structured – or partly interpreted – data contain does not depend only on the precision of the associated grammars. It depends more crucially on the size and on the strength of the corresponding knowledge community, which through the XML grammar shares categories to organize the data. As a matter of fact, since 1994, a consortium, the Text Encoding Initiative [TEI] collectively develops and maintains a standard for the representation of texts in digital form, chiefly in the humanities, social sciences and linguistics. Its Guidelines are the result of a huge collaborative process, involving curators, computer scientists and SSH scholars from a large range of disciplines. They have been widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation. For instance, the British National Corpus [BNC], makes available since 1995 100 million words – the equivalent of 1.000 medium size novels – in TEI format in which parts of speech are added (it is possible to tell for a given word in context if it is a verb, a noun, and so on). These Guidelines provide general conventions (very rich metadata, corpus structure, character sets…). They provide standard ways of encoding certainty, precision, and responsibility, which helps in distinguishing between truth and knowledge (Galliers & Newell 2000). They have precise conventions for names, dates, people, and places, which are crucial "atoms of knowledge". But they offer very refined conventions for several types of "information": verse; performance texts; dictionaries; manuscript descriptions; representation of primary sources; critical apparatus; tables, formulae and graphics… For spoken data, they have conventions for the following components: utterances, pauses, vocalized but non-lexical phenomena such as coughs, kinesic (non-verbal, non-lexical) phenomena such as gestures, entirely non-linguistic incidents occurring during and possibly influencing the course of speech, shifts or changes in vocal quality… Sharing some of these general or specific conventions makes it possible to share annotated data, to build upon data enriched by another team. New types of grammars, such as

(Biber et al. 1999) can then give an empirical account of spoken English which was out of reach before.

Let's make clear to conclude this section that a long term preservation organization is not a digital information storage system. Storing information does not imply that the system is in a position to return to the producer the original information "as is" and so that the producer can still understand and use it. Special provisions must be made to meet such a goal. First, a digital "fingerprint" is returned to the producer when an information package has been successfully archived. In the case where the producer wants its package back, it is therefore possible to compare the actual fingerprint to the original one and to prove that the underlying digital object has not changed. Secondly, the package includes contextual information (compulsory metadata) which makes it understandable and usable in the long term. In an ordinary storage system, there is no metadata as such. The name and the type of the file are the only "contextual" information, and it is a poor one. The purpose of the file has to be reconstructed via the analysis of its content, which can lead to a somehow frustrating "digital archaeology". Thirdly a storage system does not check the validity of the information it gathers. It just stores sequences of bits. On the contrary, an archive validates each element according to the format it is supposed to follow. For instance, a PDF file can be visualized on one computer and not on another, because some resources (e.g. fonts) that should be present in the file are available on the first computer and not on the second one. In this case, the archive makes sure that the PDF file is really "self-contained". Fourthly, an archive keeps several copies of each archived package, in distinct locations, to prevent accidental loss. It regularly checks the integrity of the physical devices. When a format becomes obsolete, the archive migrates all the packages resorting to this format to a new one, with as little loss of information as possible.

Long term preservation of a long term preservation system

Paradoxically, some conditions must be met in order to obtain a lasting and effective infrastructure.

Firstly, several levels of choices must be made explicit. The detailed technical choices must be accounted for and memorized. Three software play a central role in the current architecture. The CINES relies on Arcsys from Infotel for its preservation platform. [iRODS] is a layer of abstraction used at CC-IN2P3 to monitor the exchange of archival packages with the CINES and to transform them according to the access uses. [FEDORA] adds research and transformation possibilities for users. Each software has been tuned for the pilot project. For instance, the CINES procedures were changed to take into account the relationships between primary data and annotations. As these tunings were the result of many debates, trials and errors, the resulting "solutions" must be archived as well. The organisational choices stem as well from a long maturation. For spoken data, it became quickly clear that researchers and laboratories could not be considered as producers: their awareness of the problem and of its solution, their resources, their skills were not sufficient for such a role. The CRDO would fit better. This analysis should be available in the next step, in order to assess the possibility for the other Digital Resources Centres to play a similar role.

Secondly, as they were presented in the second section, research infrastructures correspond to long term investments. Short term contracts are neither economical nor efficient for such projects, as skills and experience are lost when qualified people stop working for them. However at least in France, a policy of short term savings is the current wisdom. That's why a sustainable business model must be studied. It would take into account the necessary services and volumes, the complexity of the process for different types of documents (text versus 3D simulations, for instance) and it would then determine the foreseeable costs. It is necessary to evaluate hidden costs in hardware, software, and work time which come from the current

anarchical situation. Such an evaluation would help to make the case for the infrastructure. As a matter of fact, the savings will not be immediate, as economies of scale and transfers will be progressive.
Finally, the archiving infrastructure will last only if the French context allows for it. So far, the main weakness of the project is related to the hazardous status of the CRDO and of the other Digital Resource Centres. The CRDO brings a lot of energy and competence in the project, but its medium or long term status within the CNRS or the overall research organization still needs some clarification. Long term preservation policy implies a minimal stability for the concerned communities. At the moment, it is not the case for SSH, which pay lip service to the global aim of digital archiving without necessarily having the strength to make the necessary decisions and to stick to it. We presented in the fourth section the current reshaping of French research, which gives rise to two questions or rather two fears. Since the 17th century at least, France has been a very centralized state. The past thirty years dramatically changed this tendency, with the law on state decentralization in 1982 and the law in 2007 granting more autonomy to universities. To make a long story short, there is now a contradiction between a centralized state and centrifugal forces. The first fear is that this contradiction could be detrimental to the creation of infrastructures. The second fear concerns the real intention and/or capacity for the French state to manage in a continuous and coherent way the construction of large infrastructures for the SSH. International partnerships and huge investments protect "historical" Large research infrastructures, in physics, for instance. This is not the case in SSH. In the French roadmap for Large Research Infrastructure (December 2008), the provisional budget for SSH represented 1.5% of the total budget… As a matter of fact, apart from the TGE Adonis, the Large Research Infrastructures which were announced in the December 2008 roadmap do not yet exist and are not founded. The case for long term preservation still needs to be made at a political level.

Building communities of users

We were actors of the project in its initial phase (from its design to the beginning of the test of the archiving process). We left the project in October 2009. The production phase started in March 2010. This means that data producers, in this case the Centre for spoken resources, started then to submit to the archive all the research data they possessed. The actual use of the archived data is just starting at the moment of writing. We have therefore little feedback on it. However let's make some comments on the status of the users.
In other sciences, a Large Research Infrastructure (LRI) is targeted to a specific community of users, astronomers for instance in the case of Virtual Observatories. Because of its budget and of its complexity, a LRI needs to be developed top down, but it implies that a mature community of users already exists, that this community is going to provide feedback as soon as possible and to help in quickly adjusting the LRI operation. In France however, for long term preservation, such a community is still on the make. The language of long term preservation is so far a foreign language for SSH. What's more, digital SSH do not exist really: they are more a motto than a shared vision. Leading scholars in SSH have little awareness and knowledge of digital issues. The average scholar often is not wiser in those matters, even when (s)he is young(er). There is as well a lack of technical staff and engineers in SSH. This situation is rooted in a sharp separation between natural sciences and SSH, as well as between "theoretical science" and applied science, which is supposed to "follow" theoretical science. This opposition is related to Auguste Comte's positivism (Lecourt 2001: 22). There are different universities, often in different places, for natural sciences and for SSH. This situation leads to inadequacies for interdisciplinary projects such as the archiving one: people (computer scientists, curators, SSH scholars…) representing the necessary "flavours" for such projects do not happen to work in the same place. It is not even sure that

SSH could represent a single community of users for the archiving infrastructure. As a matter of fact, the LRIs planned or decided at the European level [RI-SSH-EU] are specialized according to broad categories which do not correspond to the usual fine-grained domains, such as the ones used for evaluation purposes in the European Reference Index for the Humanities [ERIH]. A first LRI corresponds mostly to heritage data (relevant disciplines: history, archaeology, fine arts, but philosophy and literature as well), a second one to human geography, sociological, political and economical data, the last one to language data, the main target being language engineering, linguistics, but as well all the disciplines which can make use of language engineering tools and resources for their own purposes (such as sociology, for instance). The communities of potential users seem to be more mature in the last two cases than in the first one.

The lack of mature communities of users for a long term preservation infrastructure and the uncertainties about frontiers between disciplines in the context of digital SSH certainly are a pity. However the high entropy of the SSH and its way of sticking to its "specificity" seem to prevent it to make a more useful and expected contribution to society at large. A top down approach based on the model for natural sciences is certainly surprising. One can wonder whether there was really an alternative approach to try and build an archiving infrastructure for SSH.

Resorting to an archiving infrastructure would lead to different ways of doing research in SSH. For instance, as it is the case for Virtual Observatories, persistent links can be made between archived data and publications, leading to new ways of comparing methods and approaches and of verifying hypotheses and conclusions. However the infrastructure must first prove to the communities its added value. Producing data and results in a long term preservation perspective changes the actual scientific processes and brings some standardization and some "industrialization", with norms, division of work, and so on. New skills and habits have to be developed, which can be thought of as a burden, or as even an improvement of efficiency at the expense of innovation (Galliers & Newell 2000). Training must therefore be organized to increase the awareness of what is at stake and of what it requires. Help must be provided as well, building upon the example of CC-IN2P3 which makes available to the biology community grid facilities (computing, storage, Web services, virtual environment, tools): a full-time engineer helps the biology community. The production of "sustainable data" will get more attention from researchers and from laboratories as well only if these archived data are evaluated as such for individuals and laboratories, just like papers, if they are really made part of the scientific production. Researchers will then be in a better position to plan the archiving process, to decide what is precious and how to document it. Ideally, a long term preservation infrastructure would have a scientific board in which researchers and engineers would be associated in order to have both scientific and technical expertise.

Choosing our past in order to better face the future

As we were actors of this archiving project and as we are not sociologists, our contribution is bound to be limited to an insider point of view et to some reflexive remarks on the interplay between human actors and structural features of organizations: "Through the regular action of knowledgeable actors, patterns of interactions become established as standardized practices in organizations […] Over time, habitual use of such practices eventually becomes institutionalized, forming the structural properties of organization" (Orlikowski 2000:404). We do not know at the time of writing if it is going to be the case, as what we were involved in was more the design mode than the use mode (Orlikowski 2000:408) and as the overall context is somehow unpredictable.

"Today, memory and its cult serve as 'liaison officers' between a fantasized past, a disquieting present and an inscrutable future" (Hoog, 2009: 48). The facilities of digitization and digital content (re)production intensify the current obsession for memory. About the role of memory in individual and collective lives, P. Ricoeur (2000: 83-97) builds on Freud's opposition between repetition of the past and its rememoration. Repetition leaves past strictly as it was, it petrifies it. Rememoration, on the contrary, reorganizes the past according to the present and to the foreseeable futures. We should be aware as well that "Oblivion is necessary to individuals and to society" (Augé, 1998:7) and that "[it] is the living force of memory". Even in the framework of long term digital preservation for SSH, we should learn as well to forget in order to memorize really what is precious to us, as "Memory is the collective organization of selective oblivion" (Rony Brauman).

Acknowledgments

The paper has benefited from the insights of the two anonymous reviewers and of the editors of the issue. We thank as well for their most helpful comments and suggestions our colleagues from the pilot project (Bernard Bel – LPL ; Pascal Dugénie – CINES ; Michel Jacobson – DAF ; Thomas Kachelhoffer – CC-IN2P3 ; Nicolas Larrousse – RISC ; Olivier Rouchon – CINES) and from EDF R&D Thierry Chauvier and Martine Le Corroller.

References

[ADONIS] TGE Adonis site http://www.tge-adonis.fr/
[AERES] French Agency for higher education and research evaluation - Agence d'évaluation de la recherche et de l'enseignement supérieur http://www.aeres-evaluation.fr/
[ANR] Agence nationale de la recherche) – French National Research Agency http://www.agence-nationale-recherche.fr/Intl
[ARCHIVE WIKI] Wiki for the TGE Adonis pilot project on spoken data long term preservation http://www.tge-adonis.fr/wiki/index.php/Accueil_Projet_pilote
Augé, M. (1998) *Les formes de l'oubli*. Paris: Payot.
Banat-Berger, F.; Duplouy, L.; Huc, C. (2009) *L'archivage numérique à long terme – les débuts de la maturité ?* Paris: La Documentation Française.
[BARRING] O. Bärring's proposal on archiving for SSH http://www.tge-adonis.fr/?Le-point-de-vue-d-Olof-Barring-du
Baude, O. (ed) (2006) *Corpus oraux – Guide des bonnes pratiques 2006*. Paris : Presses universitaires d'Orléans & CNRS Éditions. http://hal.archives-ouvertes.fr/hal-00357706/fr/
Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. (1999) *Longman Grammar of Spoken and Written English*. Harlow: Longman.
[BNC] British National Corpus: a 100 million word collection of samples of written and spoken English language http://www.natcorp.ox.ac.uk/
[CAIRN] Private plateform for publishing digital journals http://www.cairn.info/
[CCSDS] Consultative Committee for Space Data Systems http://public.ccsds.org/
[CC-IN2P3] Computing Centre for the [IN2P3] http://cc.in2p3.fr/
[CINES] National Computing Centre for Higher Education – Centre Informatique National de l'Enseignement Supérieur http://www.cines.fr/. See http://www.cines.fr/-D-I-S-T-.html for long term preservation
[CLEO] Public platform for publishing digital journals http://cleo.cnrs.fr/
[CRDO] Spoken data resource centre - Centre de ressources pour la description de l'oral, in Paris area: http://crdo.risc.cnrs.fr/ and in Aix-en-Provence: http://crdo.fr/

[CRN] Centres for digital resources in SSH – apart from the [CRDO], for manuscripts http://www.cn-telma.fr/, for texts http://www.cnrtl.fr/, for images http://www.cn2sv.cnrs.fr/, and for geographical data: http://www.m2isa.fr/
[CROSSREF] Links between bibliographic references and digital publications http://www.crossref.org/
[DBPEDIA] 'Community effort to extract structured information from Wikipedia and to make this information available on the Web' http://dbpedia.org/About
Desrosières, A. (2000) *La politique des grands nombres - Histoire de la raison statistique*. Paris: La Découverte.
Desrosières, A. (2008) *L'argument statistique I - Pour une sociologie historique de la quantification*. Paris: Presses de l'École des Mines.
[DGLFLF] French Delegation for French and languages spoken in France - Délégation Générale à la Langue Française et aux Langues de France http://www.dglf.culture.gouv.fr/
[DGLFLF SPOKEN CORPUS] Resources founded by the [DGLFLF] http://www.corpusdelaparole.culture.fr/
[DOI] Digital Object Identifier – Unique identifier for objets on the Web and specially publications http://www.doi.org/
[DUBLIN CORE] Minimal and recognized set of metadata http://dublincore.org/
[ERIH] European Reference Index for the Humanities http://www.esf.org/research-areas/humanities/erih-european-reference-index-for-the-humanities.html
[ESFRI] European coordination for Large Research Infrastructures http://cordis.europa.eu/esfri/
[FEDORA] Open source software for digital repositories http://www.fedora-commons.org/
Galliers, R. D & Newel, S (2000) Back to the Future: From Knowledge Management to Data Management, Working Paper Series #92, Department of Information Systems, London School of Economics and Political Science
Habert, B. (2005) *Instruments et ressources électroniques pour le français*. Paris: Ophrys.
Habert, B. (2009) *Construire des bases de données pour le français*. Paris: Ophrys.
[HAL] http://hal.archives-ouvertes.fr/
Hoog, E. (2009) *Mémoire année zéro*. Paris: Seuil.
[IN2P3] Institute for nuclear physics and particular physics – Institut national de physique nucléaire et de physique des particules - http://www.in2p3.fr/
[iRODS] Layer of abstraction on distributed systems https://www.irods.org/
Lecourt, D. (2001) *La philosophie des sciences*. Paris, Presses Universitaires de France.
[OAIS] CCSDS, 650.0-B-1, *Reference Model for an Open Archival Information System (OAIS)*, ISO 14721, janvier 2002, http://public.ccsds.org/publications/archive/650x0b1.pdf
Orlikowski, W (1992) The duality of Technology : Rethinking the Concept of Technology in Organizations, Organization Science, (3)3, p.398-427
[OVS] Virtual observatory in Strasbourg http://cdsweb.u-strasbg.fr/CDS-f.gml
[PAC] Archiving platform at the CINES http://www.cines.fr/-l-application-PAC-.html
[PERSEE] Access to digitized issues of journals in SSH which are no longer submitted to a temporary embargo http://www.persee.fr/
[PIN] Workgroup on digital information preservation - Préservation de l'Information Numérique http://www-pin.aristote.asso.fr/
[RI-CNRS] CNRS Site on Research Infrastructures http://www.cnrs.fr/fr/recherche/ups3019/feuilles-route-infrastructures.htm
[RI-MESR] French Ministry of Higher Education and Research Site on Research Infrastructures http://www.roadmaptgi.fr/
[RI-SSH-EU] Main European Research Infrastructure for SSH : DARIAH http://www.dariah.eu/ ; CESSDA http://www.cessda.org/ ; CLARIN http://www.clarin.eu/

Ricoeur, P. (2000) *La mémoire, l'histoire, l'oubli*. Paris: Seuil.
Robin, R. (2003) *La mémoire saturée*. Paris: Stock.
[SEMWEB] W3C Semantic Web Activity http://www.w3.org/2001/sw/
[SPAR] Digital archiving infrastructure built by the French National Library – Système de préservation et d'archivage réparti de la Bibliothèque nationale de France http://www.bnf.fr/pages/infopro/numerisation/num_spar.htm
[SYNERGIES] Canadian Research Infrastructure in SSH http://www.synergiescanada.org/
[TEI] Consortium which collectively develops and maintains a standard for the representation of texts in digital form, chiefly in the humanities, social sciences and linguistics http://www.tei-c.org/index.xml http://www.tei-c.org/Guidelines/P5/ Conventions for transcribing speech: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TS.html
[UMLS] Unified Medical Language System http://www.nlm.nih.gov/research/umls/

About the authors

Benoît Habert is a Professor in Computational Linguistics at the Ecole Normale Supérieure de Lyon. He is currently working at EDF Research and Development where he leads a project in digital archiving. He was deputy-head of the TGE Adonis [ADONIS] where he was responsible for the pilot project for long term preservation of spoken data which is described in the paper. Part of his research was devoted to building and using digital corpora and on the role of "instruments" in/for SSH: *De l'écrit au numérique : constituer, normaliser, exploiter les corpus électroniques*, avec C. Fabre et F. Issac, InterEditions/Masson, 1998; (Habert 2005); (Habert 2009) – see as well: http://www.revue-texto.net/Corpus/Publications/Habert/Habert_Portrait.html.

Claude Huc graduated in physics and worked as an engineer at the Centre National d'Etudes Spatiales (CNES), the French Space Agency, from 1973 to 2007. At the CNES, he was first a mediator between computing engineers and the space scientific community. He was afterwards head of the department for highlighting space data use, project manager for the creation of the Centre de données de la physiques des plasmas (CDPP), then senior expert for preserving space data. From 1995 to 2005, he was a member of the international normalization group for digital archiving (OAIS and related norms). He created in 2000 and leaded until 2009 the French work group on long term preservation [PIN]. He works as an expert for the European Commission and is a consultant since 2007.