

## FAMILY ARCHIVES

Maurizio Lunghi Fondazione Rinascimento Digitale

Most of the documents produced by public administration, hospitals, university and also by professionals and citizens, are born digital and must be managed in an appropriate way/time. Libraries and archives started facing the challenge long time ago. Public administrations are now preparing policies and infrastructures to preserve digital material. The weakest component nowadays is surely the single citizen.

In this paper we propose a service of digital preservation targeted to citizens user requirements, the idea is under study and refinement, it's not yet an operative project. Innovation is not in technology but in the service definition and in the new relation between citizens and public institutions.

The proposed service is strongly related and built on the Magazzini Digitali project experience and results carried out by some national libraries in Italy and the Fondazione Rinascimento Digitale.

### Going digital towards information society

Why do we need such a service?

The citizens have to manage a lot of different digital contents every day, some of them issued by the public administrations (PA), some of them collected from Internet, some of them produced directly through photo camera, audio-video recorder, PC, mobile or other equipment. Some of these contents have legal relevance, some others are only private, they can vary so much as digital format and coding, they require different software, some of them are only temporary but other ones can be valuable for long term preservation.

Is the average citizen capable to manage by himself this task?

Obviously no.

It's true that Internet offers a lot of services to store e-mails, photos, files, videos and most of them for free, like Gmail, Facebook, Dropbox, Picasa and so on, so why do we need a new service?

First of all, no one of these applications is a real preservation service simply because they address a different request, they offer immediate use, flexible access, integration of various systems, independency from the workstation or space, but they don't care about issues like authenticity, provenance, integrity, context, rights and future render-ability.

Secondly, all these applications are provided by private commercial companies, in most of the cases by a single company not based in our country, and that is a risk factor for security, persistence, privacy and long term usability.

We propose a long term digital preservation service targeted to citizens user requirements, the service must be certified for issues like security, authenticity, provenance, integrity, identification of people, future usability, rights and privacy respect. It's not a file sharing box or e-mail or shared folder or back-up directory or blog or similar, it is something no yet available for citizens.

Our idea of service is in line and linked to the process of PA becoming digital and developing services towards the information society, so we are moving in the framework of the Digital Agenda initiative.

The Digital Agenda for Europe (DAE) <http://ec.europa.eu/digital-agenda/digital-agenda-europe> aims to help Europe's citizens and businesses to get the most out of digital technologies. It is the first of seven flagships initiatives under Europe 2020, the EU's strategy to deliver smart sustainable and

inclusive growth. The digital economy is growing at seven times the rate of the rest of the economy. Over 50% of Europeans use the internet daily – but 30% have never used it at all! As ever more daily tasks are carried out online, everyone needs enhanced digital skills to participate fully in the future information society and benefit of all the services developed by the PA to interact with citizens independently by age, education, language, etc. The Digital Agenda tackles the digital divide.

Launched in May 2010, the DAE contains 101 actions, grouped around seven priority areas:

1. Create a new and stable broadband regulatory environment.
2. New public digital service infrastructures through Connecting Europe Facility loans
3. Launch Grand Coalition on Digital Skills and Jobs
4. Propose EU cyber-security strategy and Directive
5. Update EU's Copyright Framework
6. Accelerate cloud computing through public sector buying power
7. Launch new electronics industrial strategy – an "Airbus of Chips"

Funding opportunities to implement the DAE are in the Horizon 2020 programme, criteria are related to the inclusion of citizen and participation to the information society, against digital divide for different areas or user communities, PA services and infrastructure diffusion, development of new services to facilitate the usage of digital content by all the users.

As implementation of the DAE in Italy, a new regulation for the digital PA has been released recently 'Codice della Amministrazione Digitale' available at the Agency for the digital Italy <http://www.agid.gov.it/> and it is creating a lot of positive debate and comments, we are facing in reality how to move our PA archives and services towards digital. The focus is on PA infrastructures and on services between PA offices and institutions or companies or individuals. In particular, this regulation draws also the scenario for the digital preservation for these archives and it starts to define the roles & responsibilities for the digital curators and digital repositories.

Hot topics are about the personal identifiers and the strategies to use this digital identity for the services, in particular the government is going to create a national registry (Sistema Per l'Identità Digitale) with a unique stable identifier for each person resident in Italy, based on the fiscal code and other anagraphic info, any service will use this identifiers register.

For the service definition and functions, we always refer to the national implementation of the Digital Agenda for all the services, registers and functionalities.

This work is evidently also in the sfera of the Personal Archiving that is a very wide area and it addresses an extremely large spectrum of topics, responsibilities, types of content and functionalities, mixing together seldom private-public entities. Actually it is still not clearly defined what we mean for that and what limits should be put, neither what applications should go under this category or not, so it is very subject-point-of-view. Up to now very few projects have faced this demanding challenge with low level of successful results.

So we must be very careful and conservative, we have decided to limit our project coverage to a very limited and simple area: private archiving of only few types of contents selected and described through metadata by the citizen at the ingest phase.

For the moment, we have not yet included the types of document that can have legal value like signed e-mail (PEC), contracts, diagnostic medical images or audio visual.

The present idea of service is directly linked and I could say, built on, the Magazzini Digitali project developed by the National Central Library in Florence to create a legal deposit national service for any publication afferent to the Italian Ministry of Culture and Education.

Magazzini digitali (hereinafter MD) is the national long term digital preservation infrastructure for legal deposit of Italian publications [www.depositolegale.it](http://www.depositolegale.it) according to the provisions of the new Italian legal deposit law (L. 106/2004, DPR 252/2006). MD was first implemented in 2006 by a cooperation among the National Libraries in Florence, Rome, Venice and the Fondazione Rinascimento Digitale, then a second phase started in 2010 for digital doctoral dissertation archiving project.

The ingestion can be made in Rome or in Florence. Data replication procedures ensures that six copies of a document will be stored. For the end user there is a single point of access to MD service.

Each institution (Florence, Rome and Venice) has selected three different data centers owned and managed by 3 different companies (to reduce the commercial risk of "domino" effects) and at least 200 km far away from each other (to reduce the risk of natural threats). Compliance certification to ISO 27001 (international security standard) has been the basic prerequisite for the selection. The selection was completed at the end of 2011.

As an example of an MD component we can have a look to an end user access module. MD is working with the virtualization and emulation technology not only as strategy for long term digital preservation purposes but also as technology that enable a "protected access" to deposited resources. If MD has to give guarantees to publishers that access to deposited publications will not allow unauthorized copies, you can physically lock the library workstations but in our experience this is not a viable solution. To "simplify" the management of library workstations and also to facilitate the user access MD is planning to allow users - using also their laptop - to access to deposited publications from the premises of National Central Library. MD "Remote browser" solution offers "protected access" to deposited publications.

Long term digital preservation is a public service to be provided by trusted digital repositories in order to ensure - for designated community:

- viability = bits are preserved in correct sequences
- render-ability = formats of files can be interpreted by a computer
- authenticity (identity + integrity) = specific metadata and PI
- availability = resources can be accessed

## The actors – The service – The trust elements

### The actors

We want to merge together the experience and credibility of some public institutions on digital preservation to the technical competence and infrastructures of the private sector. So the actors are:

1. A board of public institutions (PIB) defining the terms of service, the network architecture and executing a continuous monitoring of the system. In our proposal, the idea is that they can be the Tuscany Region, the Florence council and the National Library.
2. At least 3 service providers (SP) that can be public institutions or private companies.
3. The final users so the citizens archiving their contents.

The PIB defines the terms of service and functionalities of the system based on a trusted digital repository to store user digital contents, the institutions managing the repository are the national library in Florence, the town council, the Tuscany region, but the PIB is open to any other institution. Technical, organisational and legal specifications are defined by the PIB and implemented by the SP.

The contract, to be signed by the user, describing the terms of service is jointly defined by the PIB and the SP.

The PIB is responsible to define a preservation policy and risks register in order to grant safe control of the global system and to make it public and absolutely clear to users: written roles and responsibilities must be defined and accepted by all.

The SP can be public or private because in any case they implement the terms of service and respect the general functions of the system defined by the PIB. The SP manage the daily interaction with and support to the user, they are responsible for the service maintenance and communication.

### The service

We propose a long term digital preservation service targeted to citizens user requirements, the service must be certified and trusted by the target user community for issues like security, authenticity, provenance, integrity, context information, identification of people, future usability, rights and privacy respect. A new service to storage only selected digital memories of citizens (selected by them), with the appropriate description metadata, the service is offered by an infrastructure public-private based on trust and sustainability.

A fundamental component of this system is its training goal!

Making aware users about accepted types of content and digital formats, technology risks, legal constraints and responsibilities. Training of users will be about topics like selection criteria, possible strategies for long term curation, licencing and accessibility issues, right management, technology migration strategies.

The contract innovative points are:

- i) Persistent Identifiers are used for any person and for any digital content
- ii) the service is based on the ownership of the digital contents;
- iii) the account can be individual or related to a family;
- iv) the user can decide the type of accessibility of each single content;
- v) the user pays an annual fee for an active login, silent logins are free;
- vi) the contents are not cancelled for long time.

Any user of the system must first of all make a registration for the personal profile. In this phase the user is requested to provide information like:

1. first and last name
2. sex and nationality
3. fiscal code and the personal identifier in the national register (SPID)
4. e-mail and other contact

At the first creation of a login the user is asked if the login is individual or 'family' that means related to other members of a family with whom the user wants to share some information. The login is multiple in the sense that some members of the same family can share contents and information, they can have joint management and common area for publication of contents.

At the ingest phase the user is requested to self-declare to be owner of the contents and to clear any problem for digital rights management and IPR, or in case he must declare shared ownership with some other members of the family. The ownership of the contents remains always of the users, even if some rights are extended to the service providers for the long term management of the digital archive. The ownership can be sold or passed to other people, also in case of death of the original owner.

In case of shared ownership all the actions, like in particular publication and cancellation, must be agreed by all the owners: the system checks that with an internal communication procedure.

At the ingest phase the user, in addition to a self-declaration about the ownership of the digital contents, can decide the type of accessibility for the single content: private, shared, public. In particular, the user can share the accessibility of that content with other members of the family. Of course the user can change accessibility for any content at any moment during all the life-cycle. For the public access we propose the CC 3.0 licence.

The PIB will investigate possible use and dissemination of the public contents, respecting always citation and limitation of the owner rights, likely thematic exhibitions, cultural and social history applications, educational purposes.

The system makes an extensive use of Persistent Identifiers (PI) both to identify in a secure/unique/stable way any user acting in the system and tracking all the activity, and to identify a specific digital content ingested by the user. As said the system uses the Italian national identifier (SPID) for any person resident in Italy, and one PI for digital object between DOI and NBN.

Of course only selected contents must be proceeded to this service, we offer a limited space up to 10 GB, so users are recommended to appraisal carefully contents and to provide during the ingest phase some metadata related for example to the context, ownership, access, authenticity, some of them will be updated, or created by the repository managers, during the life-cycle.

The user pays an annual fee to the SP who pass a part of that to the PIB for long term sustainability of the global system. The following years the user can continue to pay the annual fee to keep active the login and so ingesting other contents, or leave the login silent without paying the fee, in that case content are only visible and never cancelled: in any moment the user can reactivate the login and manage his archive.

User requirements have been analysed for the types of content that citizens receive from the PA (e.g., certifications, documents, medical reports) or contents created directly by them, we must defined digital formats, significant properties to be ingested, we must supposed duration and foreseen the possible future uses.

Table with some examples of content types and context information to be collected during the ingest phase for any content.

	photo	text	comments
accepted digital formats	JPEG, PNG	PDF/A, DOCX	for text only a list of fonts are accepted
PI of the owner ingesting the content	SPID	SPID	other PI systems are acceptable for people not resident in Italy
PI of the digital content	DOI, NBN	DOI, NBN	if existing otherwise the system will create one
self-declaration IPR	private-shared	private-shared	shared with other members of the family
self-declaration ownership & rights	private-shared	private-shared	shared with other members of the family
Title			free text
language	NA		only for text document
description of subject and context			free text
keywords			free text
date of ingestion or modification			The system assigns automatically, as well as any future change is tracked
forecast duration			
accessibility	private-shared-public	private-shared-public	In case of shared ownership the accessibility must be agreed by all the owners

### The trust elements

Trust is fundamental for this system, so it's crucial the role played by the PIB to grant super-partes management and long term preservation of the archive.

The PIB is responsible to manage technology migration and to define a preservation policy and risks register in order to grant safe control of the global system, the PIB makes users aware about this policy in the most clear and transparent way, likely it produces some recommendations about the related topics.

The system implements all the necessary functionalities/technologies to protect user privacy, security and rights management on all the contents.

The PIB manages duplication of data with cryptographic tools on the network, mapping schema of the duplicate is not known by the SP for security reasons.

From a technological point of view, the repository is compliant with the international standards for audit and certification of trusted digital repositories, like OAIS and ISO 16363, applying results from APARSEN project: specific attention is paid to persistent identifiers for objects and people.

Trust and Sustainability of the system are based on some basic elements:

1. the PIB oversees the global system architecture and executes a technology watch;
2. the SP are at least 3 independent and far enough;
3. the PIB grants to preserve the family archive for successors;
4. duplication of data is maintained safe on the network;
5. thanks to the agreement between the PIB and the SP, contents are not cancelled for long time even if the user stops paying the annual fee.

## Conclusions

The project is still under study and final refinement even if the user requirement would encourage to start a first trial phase that would be instrumental to test feasibility of the system and user satisfaction. Main open issues and risks are on the legal topics and possible risks for future management of the digital archive. A possible work plan for the first development can be forecast in the first part of the next year.

## Bibliography

1. "Cultural Heritage on line", Maurizio Lunghi, international conference in Florence in December 2012
2. "The National Bibliography Number Italia (NBN:IT) Project. A persistent identifier supporting national legal deposit for digital resources", Maurizio Lunghi, alii, Aug 2009
3. "Deposito affidabile per le fotografie di famiglia" anno accademico 2012/2013, Daniela Giglio, Università di Firenze
4. "The Future of Personal Digital Archiving: Defining the Research Agendas", Clifford Lynch, Executive Director, Coalition for Networked Information, Dec 2013