

Sergio Anguita*, Carme Montaner**, Joaquim Oller***, Rafael Roset****

Digital preservation at the Institut Cartogràfic de Catalunya

Keywords: Digital files; preservation; cataloguing; metadata.

Summary: The ICC has been storing digital data since its creation in 1982, when its activity started, while the map library has been storing paper maps. The IT department has had the duty to store, protect, keep readable and accessible this huge amount of digital information that different departments, including the Map Library, have been generating during three decades. Until nowadays digital maps and paper maps were following different paths. The ICC is starting a digital preservation plan on which the IT department and the Map Library are working side by side to preserve the ICC map production as cartographic heritage.

Preserving paper maps in the map Library

The map library of Catalonia was created as a unit inside the ICC in 1986. Its aim was and still is to collect paper maps of Catalonia and the world, with special care of the map production of its own institution. In this 27 years the collection, which started with few records, now holds more than 300.000 sheet paper maps, 68.000 books, 350.000 vertical aerial photographs, 45.000 panoramic photographs, as well as more than 200 topographic instruments and an archive with documentation related with territory, meteorology and geography.

A significant portion of resources has been devoted to preserve all this material in good conditions: steel furnishings in air conditioned stores; paper restoration; renew of book bindings and a large number of actions aimed to preserve for the future this whole cartographic heritage.

To catalogue all these items, the map library started with manual cards but in a few years the automation of catalogues was introduced, using international standard description formats as MARC or MARC21. In the process of map cataloguing, the special format of the map series has led us to increase the level of cataloguing in a sense that the bibliographical standards did not contemplate (as different editions or reprints of any separate sheet). Some specific records as aerial photographs were described with internal ad-hoc catalogues because they were used in the map production processes.

At the beginning of the 21st century the introduction of digital environment in the map library meant a deep change. On one hand we started the digitisation of paper maps of the library, and on the other hand the irruption of the digital maps forced us to reconsider the internal process of cataloguing, preservation and dissemination: from CD's to geoservers, the maps, the map libraries and the users are not the same they were.

With the digital, the map library designed a new catalogue to include paper and digital maps and to have both related when the digital comes from the digitization of a paper map. This catalogue uses an in house database based on two aspects: on one hand describes two levels of cataloguing: general description of the map as well as any of the sheets (in a multisheet map) and on the other

* Sergio Anguita. Head of IT Area, ICC [sergio.anguita@icc.cat]

** Carme Montaner. Head of Map Library, ICC [carme.montaner@icc.cat]

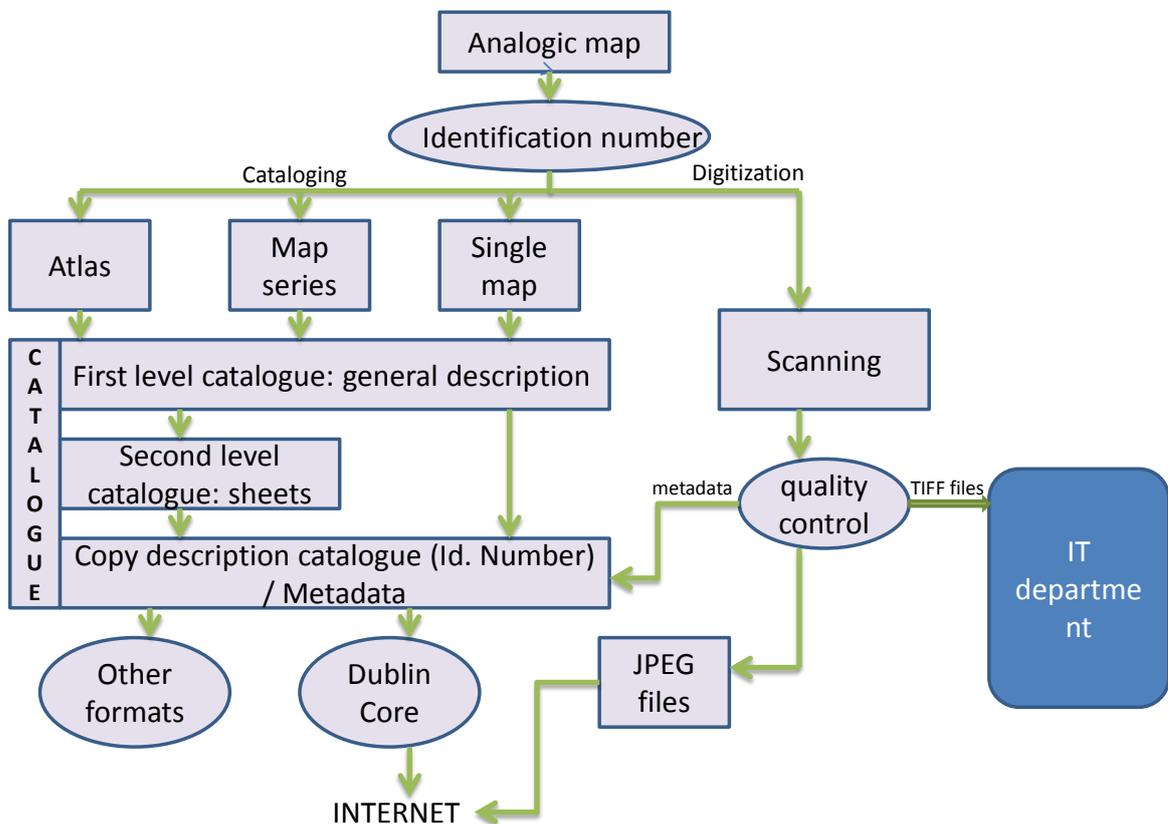
*** Joaquim Oller. Senior IT Data Manager, ICC [joaquim.oller@icc.cat]

**** Rafael Roset. Digital map library, ICC [rafael.rosset@icc.cat]

hand it includes the metadata coming from the scanning process. This database which captures all the data permits also the exporting of data in different formats: we export fields in Dublin Core format to use in the dissemination platform on the internet or basic fields for an IDE metadata catalogue.

One of the basics of this cataloguing process is the identification of any record that is any map and any different sheet. To do this, the CTC uses the registration number (sequential identification number) used traditionally in the map libraries and written in the verso. This number which identifies any separate map or sheet is also used to name the digital file coming from the digitization process. In fact this number is the connection between the paper, the digital and the catalogue and identifies only one item. This permits the control of duplicates, new versions, new editions, re-prints of each one of the maps and every sheet of the multisheet maps as well as its digital and paper versions. This database permits a description of each record in a level that it's not contemplated in a bibliographic catalogue.

Cataloging and digitization process in the Map library of Catalonia



The most significant aspect of this new cataloguing process is while paper maps go to the map library stores, digital maps are transferred to the IT department. Until this moment, despite the map library made the description, the digital file is stored in another ICC department with other digital files coming from the ICC activities.

The digital production of the ICC is a special case. Until 2004 only paper copies (plots used in the quality control process) of the digital maps (series 1:5 000 orthophoto and topographic) were deposited at the map library. After 2004, no more paper was plotted and the maps produced by the ICC are only available on the website (the last version). Vector files and raster copies are stored in

the IT department with descriptions provided by the production units. In the last two years an ICC version of the ISO 19115 was defined to normalize the description into the INSPIRE European project. In 2012 the ICC has a lot of “ancient” digital maps coming from a long tradition, more than 20 years, producing digital maps. That is the heritage of the next future. It is time to change the accumulation process of digital data into a preservation policy in which all the staff has to be involved.

Storing digital data in the IT department

The ICC started in 1982 and the digital environment was contemplated until the first moment. Since the beginning of our activity, the IT department has been storing information on different kinds of magnetic and optical supports. The need to store data, together with the duty of making information rightly readable, has led to research on an structure task in order to move the data from one support to another, taking into account different aspects such as confidence, performance, durability, suitability and storage costs. Besides, some supports, considered reliable by the IT community, happen to be not always suitable for our requirements, as some new products fail to stay updated and functional long enough. During these 30 years, the IT department has had to face up several challenges:

- Technology evolution.
- Data growth.
- Data life cycle.
- An easy way to catalog data.

Technology Evolution

We started writing our data over open reel tapes, eight millimeters cartridges, twelve inches laser disc, CD, DVD, and digital linear tapes with several capacities and writing densities, witch is finally the actual one.

Although computers, operating systems, software and hardware have changed during this time, we have been able to keep, at least, the final products suitable to be read. Not always The original information we had at the beginning of our production system is not always still readable, due to the normal software obsolescence, but we make efforts to keep final products readable all the time.

Data growth

When talking about our experience on storing data, there is a clear turning point. Before 2005, most of the devices used were analog, and the only information that had to be stored was the output data created during the production process. The most used IT unit measure was typically in megabytes (MB). From 2005, the ICC started using digital sensors as primary data collectors, which considerable increased the business possibilities as well as the information size. In consequence, the gigabyte (GB) and terabyte (TB) put aside other smaller measure units.

Nowadays, we are using a Linear Tape Open, also called LTO, witch has 800 GB capacity with no data compression, and up to 1600 GB capacity, depending on the possibility to compress data. Although, this last value is usually not regular and so can't be considered as a precise one, due to its variability.

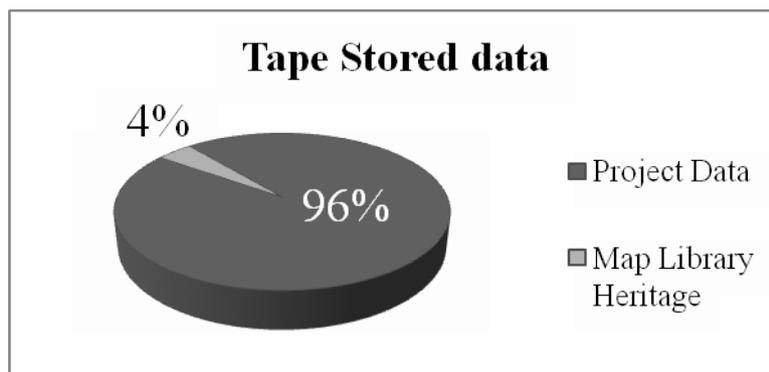


Figure 1. Relationship between stored data and Map Library digital assets

The above chart (Fig. 1) summarizes the up-to-date amount of information held on tapes. We have stored almost 700 TB (up to 1.4 PB, if copies for disaster recovery, in case of tape read error, are considered). The Map Library Heritage is stored in about 30 TB, which represents 4% of the total stored information.

The normal activity of the ICC produces digital data with different kinds of protection. Therefore, not all generated data are needed forever such as intermediate stages, temporary objects with time limited value, or even end customer deliveries that sometimes require their products to be kept several years as an added value to a project awarded.

Therefore, we are working on three data types, depending on its retention time and its use.

- Online data, relate to information that users are working daily on it with a high usage and criticality. This information should be restored “as is”, in case of server failure or human error. To protect this data we use the typically grandfather-father-son backup rotation scheme.
- Near-Line data refers to information of medium usage, that has to be restored as soon as possible without operator intervention. This data is stored on a Hierarchy Storage System (HSM), a system that keeps a link between the data on a server and that placed on a tape.
- Offline-Data, or data that is not going to be accessed in a short time but should be stored forever. This data is written on a tape and cataloged for later use if necessary.

Easy way to catalog data

As indicated, since the beginning of our activity, the generated information has been suitable for reading even when the mentioned technology changed. Our information is obviously produced on many servers, which are renewed when they become obsolete.

Most of the backup tools used, including the active one, have been “server-oriented” backup software.

Every project or stage project is spread to several servers. To this point, we found essential to be able to see our information not only in a “server oriented” structure but also as a “project oriented” one, getting a wider view over our data. In other terms, we did not want to worry about where the information was produced, but we really needed to know to which project it belonged.

This is the mainly reason why the ICC Tape Library Software has been developed (from now on Data Access Portal - DAP), so that end users can give to each entity¹ a tag² that has a meaning by itself as an element to classify the information.

¹ Unique element (file or folder) belonging to a project, subproject, project phase, etc...

Each time a user decides to store an entity on the DAP, the IT department picks up this entity, and writes it on a tape, making a new DAP record. The end user then, is able to make a data query to the DAP to ensure that his request has been correctly processed.

Current problems

The current processes we support are nearer to classical backup strategies used in System Administration Departments than preservation ones. This is the starting point of our project that will try to solve some important problems:

- Evolve our current Backup Catalog onto a Unified Data Catalog
- Data producers are catalog responsible (distributed system)
- Problems with different versions (no metadata associated)
- Backup data is mainly useless, if it is not shared as read-only live data
- Data duplication, increasing costs in case you want to share this info on the Internet.

A first step

As mentioned before, through the DAP it has been possible to access the data catalog stored by the IT Department. However, the data copy/recovery time, in this context, would still depend on manual operations. In order to avoid this delay, and making an effort to provide a better service to the users, we designed a project, called Digital Photo Library, whose core technology uses the automatic technology of archiving / unarchiving information to tape. This technology allows users or applications to have a single file system in which each file is stored on different support types: disks of different speeds and/or tapes.

The Digital Photo Library is the repository of the ICC Photogrammetric flights. Information of Photogrammetric flights carried out by the Institute, as well as information from the individual frames corresponding to these flights can be found in this repository. Given the amount of information, the system only displays the data on disk at low resolution (online data) and, only when the client validates that he wants to access the original information, then it is automatically read on tape (level near-line data) and temporarily stored on disk.

Despite the apparent simplicity of this technology, the environment requires an accurate and thorough management that, in our case, is done through a specialized commercial tool. However, it should be noted, that this technology opens an interesting door in terms of cost reductions by integrating operation and data preservation environments, as they always will be accessed securely, not only through operators, but also by end-users and/or applications.

² Project, subproject, or project-phase name.

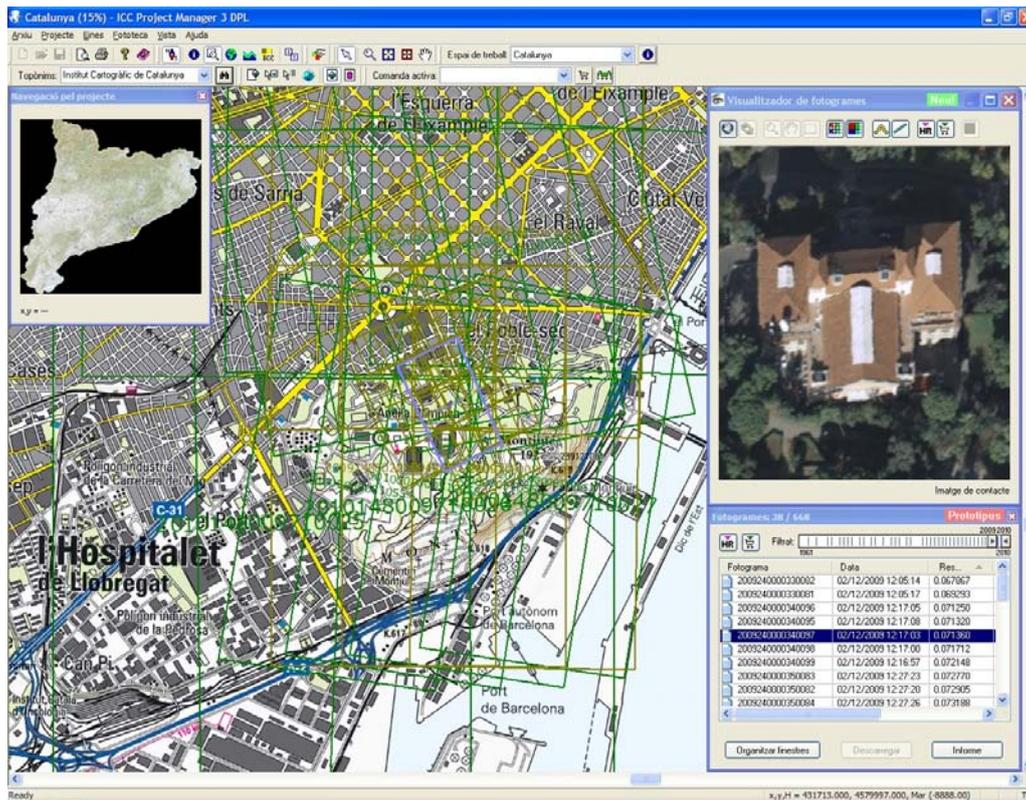


Figure 2: ICC Project Manager

Future strategy for digital preservation at the Institut Cartogràfic de Catalunya

One of the biggest problems of the current policy of digital preservation used at the ICC lies in the amount of duplicate data and its use, practically zero, as a result of being part of a circuit isolated from the rest of data exploitation systems.

The evolution proposed by the ICC requires a change of mentality regarding classical strategies. The digital data preservation shouldn't be considered as an independent process but as another requirement and functionality of the information exploitation systems.

Having terabytes of information stored on systems that are not exploited by applications and services through telematic networks except in case of disaster, implies a very high cost. Besides, it ends up only reducing the level of risk of loss until its residual value. At this point, investment in the current system generates a return virtually negligible, and therefore, dispensable. Thus, the main characteristic of the proposed system is the exploitation of the copies in operating environments.

The new proposal is based on three distinct levels:

- Management layer
- Operation layer
- Preservation layer

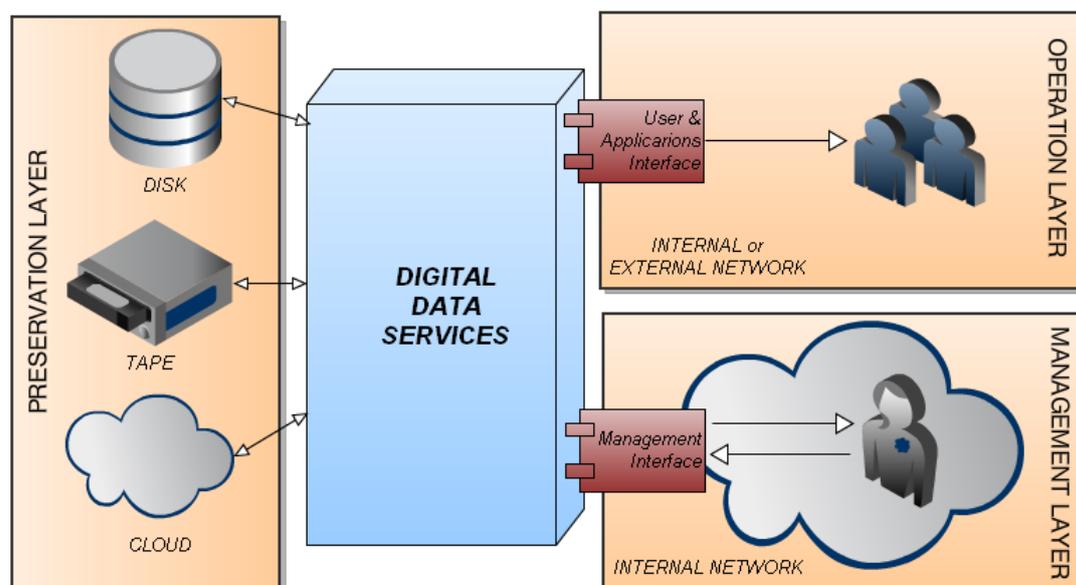


Figure 3: Digital Data Services Schema

The management layer aims at cataloguing the information, including the interface of consultation and intake of data. This layer should be managed by a single contact in order to ensure the validity of the information entered into the system, as well as the metadata associated with it. The application program interface will depend on each institution needs, requirements and peculiarities, but always should guarantee compliance with the different regulations applicable. That is why the Map Library of Catalonia should lead the definition.

The operation layer has the purpose of defining different interfaces of access to the data preservation layer, allowing the use of the information through company internal or external applications and systems. Access to information has to be multi-protocol, enabling to share repositories such as remote disks (CIFS, NFS) or the direct access to data (HTTP, HTTPS, S3, WebDAV,...). This access point should set up a first level of access control and security, as well as establishing the redirects to the available data according to the user profile.

The preservation layer aims at storing data following the defined preservation policies, and also at providing access to the information under the necessary restrictions, depending on whether the origin is from a management or an exploitation level.

There are several technologies and commercial solutions that can guarantee the data preservation, their immutability and strict access control, while enabling their exploitation through services and applications. The task of choosing one or the other depends on the particular needs of each environment and, above all, on the compatibility and interconnection with those already deployed at each centre. However, it is important to highlight certain aspects that we consider key in this kind of environment:

- Storage supports: it should have a minimum of two differentiated supports on the basis of its speed of data access (and, therefore, cost). Usually, it should have disk storage (SSD, SAS, SATA,...) and tape (LTO5, LTO4,...). However, it is recommended to consider the compatibility with cloud storage, as it offers a geographic distributed online access.
- The different storage supports should be integrated in a single system able to move the data according to its use (dynamic tiering).

- The system should be able to maintain a reporting of activity, which can identify any access and operation as well as optimize the stored data between the different layers.
- Finally, obviously it should be a minimum of two copies for data in geographically distant centers.

The digital preservation project is still at its beginning. The ICC has had in place cataloguing, archival and storage of digital information since it began operations, but not regulated preservation according to any standards and policies. Now the ICC aims to write and put to work a preservation plan that involves all parties at the company. In this way the IT department and the map library are in good position -as part of a cartographic agency- to look at digital preservation from two different angles: as a producer centre and as a map library. Trying to look into both disciplines we are in the process of designing a preservation policy combining experiences of the staff working on archiving data and staff working on heritage. The IT Department is the big store of the ICC map products as well as the maps coming from the scanning process of its map library, establishing technical processes to store, manage and recover digital files. The map library of Catalonia is the cataloguing team to describe, to document and to establish hierarchies between stored items. It is time to go outside of the classical divisions in the world of maps: data centre as part of the map libraries, map libraries as part of data centre, both working together as parts of the geographic information centre, describing, preserving and disseminating data, in our case geodata. All this effort will help preserve the cartographic heritage of Catalonia.