

Rafael Roset*, Noelia Ramos**

Present and future of the Map Libray of Catalonia

Keywords: digitization, Dublin Core, digital collections, CONTENTdm, metadata.

Summary

In 2006 the ICC purchased new software for digital collections and a year later a huge scanner for the digitization project at the map library. The modified workflow to produce digital content and the future possibilities of interoperability of the data are summarized in this presentation.

Introduction

The ICC is the national mapping agency of Catalonia, started back in 1982 from the remaining of several different institutions with good expertise from many years of producing geographic information, and depends on the local Government of the *Generalitat de Catalunya*. In 2007 the ICC was employing some 232 workers, aged 40,83 years on average, with a yearly budget of 28.858.740,47 euro. The Map Library was started as a project inside the ICC to provide cartographic materials for everyday production and to serve as the reference Map Library of Catalunya.

ICC
Institut Cartogràfic
de Catalunya

Collection

Composition:
300.000 maps
300.000 aerial photographs
40.000 pictures
60.000 documents
41.000 books

2

Besides the cartographic products of the ICC, which include maps and aerial photographs, the funds of the CTC are mainly built on cartographic materials from all over the world, with a special focus on antique maps of Catalunya and Spain. A small fraction of these funds have been digitized, and some of them were already online but each with its dedicated application

and scope. A unique catalogue for all the materials has been envisioned, but development is taking longer than expected.



3

A specially conditioned room at the lower level of ICC facilities holds the archive of the CTC. It is composed of two separate areas: one for maps and paper materials, and the other for photographs. Along with these storage facilities two more separate rooms exist to provide specific conditions for the rarest and oldest maps, thus the more valuable. All these are paper based or film based materials, and will keep on growing except for the aerial photographs archive: two years ago cameras were replaced with Digital Mapping Cameras, creating a new scenario for the Map Library: born digital images. This means huge amounts of digital data that have to be put online.

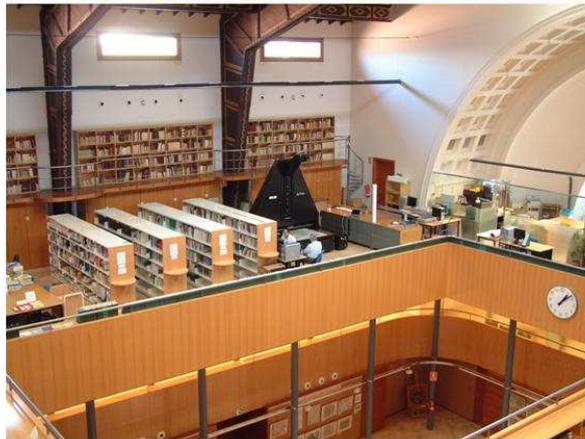
Map library main activities:

- Purchasing
- Cataloging
- Digitization
- Diffusion

For the last 20 years these have been the main activities, and in 2006 a huge digitization project was started. Nowadays, we've put online more than 5.000 images.

4

Started along with the ICC itself, the CTC has been growing its huge catalog by means of massive purchases of maps from all over the world, and by ingesting the yearly map production of the ICC. After so many years of being publicly accessible, time has come to put it online for everyone to enjoy. As the map library of reference other libraries and catalogues rely on the cataloguing performed by us for maps of Catalunya.



5

The CTC was installed at the main floor level of the ICC, with three separate areas: an open reading room where users have access to map materials and online catalogues; a technical library for ICC employees and the CTC staff facilities.

In 2007 a Metis DRS2A0 scanner was installed close to the technical library, along with the other scanners: it's the core of the Digital Map Library. Big maps and rare maps are fed to the Metis scanner, while other map sets are digitized by third parties. Technical specifications, procedures and quality control are run by CTC staff.

Equipment



Kodak IQsmart3

- Characteristics:**
- A3
 - 5500 ppi optical
 - negative / positive
 - reflective / transparent
 - high volume
 - xy stitching



6

The Kodak scanner is used for photographs and paper documents smaller than A3 in size. Although slow has a good optical quality, mandatory to provide good quality reproductions of photographs. Aerial photographs are no longer scanned in house, but still some 100.000 of them have yet to undergo the digitization process in a project estimated at 1.000.000 euro and 3 year span due to the tight requisites of aerial photographs.



HP Scanjet 8300

- Characteristics:**
- A4
 - 4800 ppi optical
 - negative / positive
 - reflective / transparent
 - high volume
 - single pass



7

Good for high volume paper jobs, the HP scanner is used mainly in those digitization projects involving documents and printed materials. Some photo sets that do not require the highest quality or those photo sets that need not preservation files are also scanned with this equipment since it provides good quality at high speeds.

Resolution	X cm	Y cm
300 ppi	182	121
350 ppi	156	104
400 ppi	137	91
450 ppi	121	81
500 ppi	109	73
600 ppi	91	61



8

The Metis scanner installed at the CTC has a maximum useable area of 121x182 cm at 300 ppi. Higher resolutions can be achieved at smaller areas, up to 600 ppi for areas smaller than 61x91 cm. In certain cases where originals are larger than the scanning area more than one shot is taken by placing parts of the map out of the table. Later these images are mosaicked together using commercial image processing programs. For that purpose the overlapping area between two images must be larger than 30% of the image to achieve good results. A typical image of the whole are at 300 ppi weighs 900 MB in uncompressed TIFF format.

Animation
-Resolution change
-Synchrolight
-Digitizing



9

Resolution is controlled by moving the CCD enclosure up and down the rails, where every resolution stop is marked. These spots already have a set of predefined parameters to ease the process of scanning documents of varying sizes. For each resolution a color control curve

exists, and all materials are scanned using a color patch card to ensure maximum color fidelity in the whole process from acquisition to printing. Color calibration is checked for every new scanning batch.

Coello, 1851
103 x 78 cm @ 450ppi
(~800MB)



10

The scanner was chosen for its ability to scan flat maps very precisely, and for its use of a pressure table that keep the materials, be that folded or rolled up, flat against the scanning glass. That way reflections and dark areas are minimized and an overall scanning quality can be achieved for every different map. Working one shift the scanner can produce up to 130 quality images per week, at some 700 MB average each, for this type of materials.

Damius, 1726(?)
34 x 49 cm @ 600 ppi
(~350MB)



11

Since the pressure table is controlled by a fine positioning mechanism we can also scan materials that would not initially lay flat against the glass. By using different materials to supple-

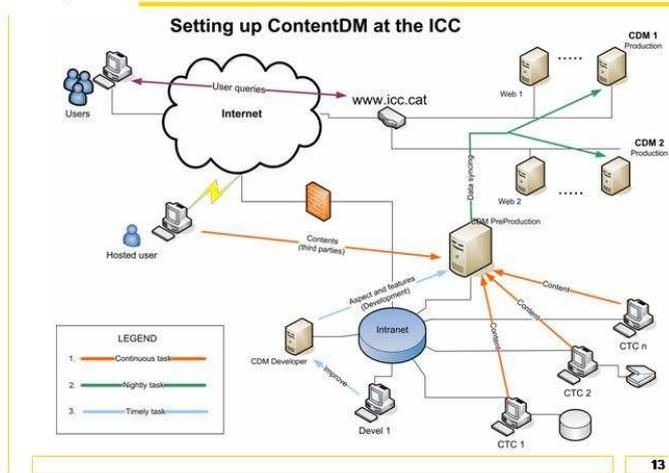
ment the necessary height, folded maps inside books can also be digitized with the same quality standards. This has been very useful when scanning manuscript pieces like the diaries of the duke of Darnius.

Productivity for this type of materials, and for rolled up maps is worse due to the manipulation of the originals necessary to load and unload them from the scanner. A few ancient maps really delicate were clocked at 15 minutes each before the scanner was ready. Once the scanner process begins it takes only 90 seconds for a 180x120cm scan at 300 ppi.

Data processing



Once these old maps are digitized, they are grouped into collections by its geographical area and uploaded to a server hosting the *ContentDM* software on which digital collections are presented on the internet. The software runs internally on a flat file structure that makes it really easy to deploy. Programmed in PHP it's easy for a skilled developer to modify its functionality and to add new features. Quite solid too, no performance problems have been detected for these 9 months it's been online.

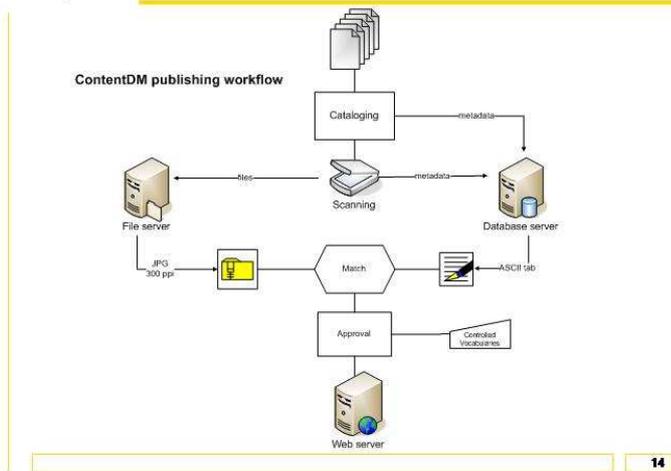


In order to provide a trouble free working environment, three layers of servers were deployed:

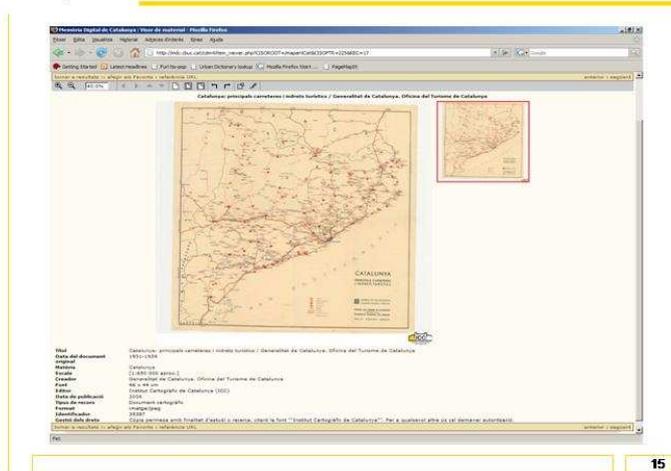
- The developer layer, where new versions of the software are tested along with minor changes to some functionalities.
- A preproduction layer where software functionalities and collections are loaded and tested.
- And a final production layer where a balanced server farm provides worldwide access to the digital collections.

Different roles of users have restricted access to each of the layers, to ensure everything performs as expected.

The images that make it to the preproduction server are down sampled to 300 ppi from the preservation images kept at the digital archive of the ICC. Larger maps, exceeding 1 meter in any dimension, are also reduced to half its size. As a rule of thumb JPG images for ContentDM can't exceed 20 MB.



The digitization workflow



The final aspect of a published document combines the overview and the metadata in the same view. A toolbar at the top of the page provides usual functions such as zoom, pan and rotate, and can be easily customized to add new functions.

Field name	DC map	Data type	Large	Search	Hide	Vocab
Títol	Title	Text	No	Yes	No	No
Autor	Creator	Text	No	Yes	No	No
Altres autors	Contributors	Text	No	No	No	No
Any original	Coverage-Temporal	Text	No	No	No	No
Any de cerca	Date	Date	No	Yes	Yes	No
Àrea geogràfica	Coverage-Spatial	Text	No	Yes	No	Yes
Àrea geogràfica alt	Coverage-Spatial	Text	No	Yes	Yes	Yes
Descripció	Description	Text	Yes	No	No	No
Escala	Description	Text	No	No	No	No
Dimensions	Description	Text	No	No	No	No
Registre	Identifier	Text	No	No	No	No
Època	Date-Issued	Text	No	Yes	No	No
Editor	Contributors	Text	Yes	Yes	No	No
Volum	Date-Issued	Text	No	No	No	No
Número	Date-Issued	Text	No	No	No	No
Any de publicació	Coverage-Temporal	Text	No	Yes	No	No
Sumari	Description-Abstract	Text	Yes	Yes	No	No
ISSN	Relation-Is Format Of	Text	Yes	No	No	No

General information

16

The necessary metadata fields and its mapping against the Dublin Core (DC) format.

The Dublin Core metadata element set is an early standard (beginning in the mid-1990), for cross-domain information resource description. It's a simple set of elements for describing materials online such as image, text, sound, video and web pages. In 2003, DC was defined by ISO Standard 15836.

Furthermore, DC contains elements that allow cataloguers to use geographic referencing like specification of the spatial limits of a place (coordinate values of *northlimit*, *eastlimit*, *southlimit* and *westlimit*) and name for the place. This is an especially good place to start with metadata schemas, because there are ISO 19115-to-DC crosswalk, and DC-to-MARC crosswalk (a bibliographic data format used for library cataloging).

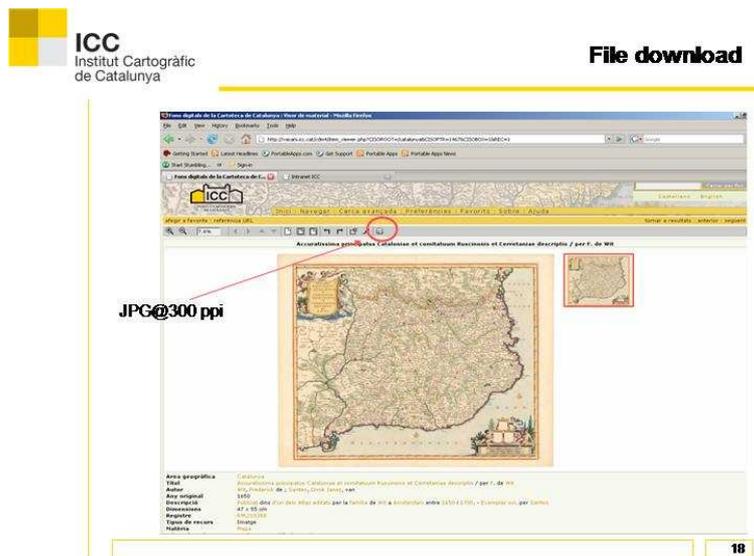
Metainformation about digital publisher

Field properties

Field name	DC map	Data type	Large	Search	Hide	Vocab
Tipus de recurs	Type	Text	Yes	No	No	No
Matèria	Subject	Text	No	No	No	Yes
Editor digital	Publisher	Text	No	No	No	No
Format	Format	Text	No	No	No	No
Disponible des de	Date-Available	Date	No	No	No	No
Permis d'ús	Rights	Text	No	No	No	No

17

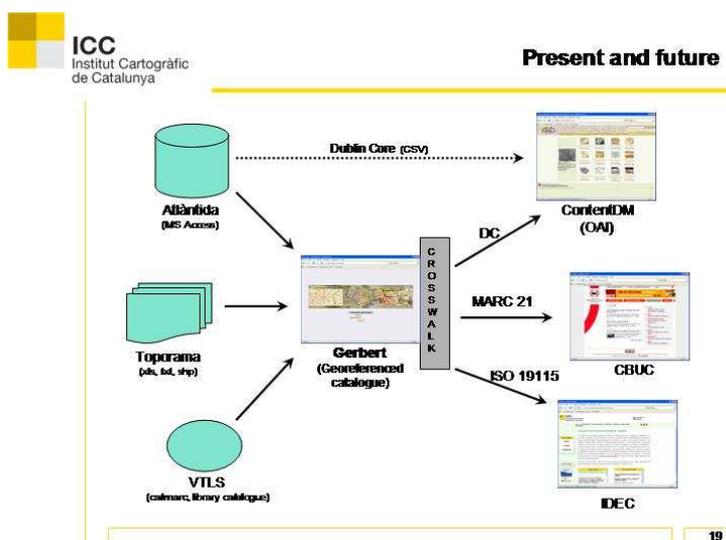
Also, some of the particular DC fields used for specific administrative purposes.



Among the additions made to the core functionalities of ContenDM we added a new button to the image toolbar to help downloading the images. After registering as a user at the ICC website, anybody can download the JPG images at 300 ppi available in our digital collections. Registration is necessary for us to keep track of statistics both to improve the usability of the web site and to generate activity reports. Since most of our real world users are migrating to this new virtual environment and other users that would never get to the visit the CTC in the real world are entering our website we need to collect visitor statistics just like a real world map library does.

Other additions made to the Content DM code were shared by users at the product mailing list, among which the Levensthein distance being the most eye catching. It consists of a simple algorithm implementation for providing alternative spelling to search terms. It has proven to be really useful in our case, since the main search box is restricted to geographic area because almost 90% of the searches conducted by users are on places.

The future



Managing data in different environments makes the process error prone. In order to centralize all data and metadata a specific cataloguing application is being developed at the ICC with the special needs of the CTC in mind: handling of series and atlases, single documents, maps and photographs, in different languages and with its corresponding georeferencing metadata to be able to spatially query the collections.

In the near future this application will also export the metadata in other formats to make it available to other platforms. Interoperability is a must to fulfill our role as reference map library. We want to be able to export:

- *Dublin Core*: to feed our Digital collections and share contents with the world.
- *Marc 21*: to join University Libraries union catalogue (CBUC).
- *ISO Standard 19115:2003* of “Geographic Information Metadata”: to join Spatial Data Infrastructure of Catalonia (*IDEC*). *ISO 19115* defines the schema required for describing geographic information and services. Although this standard is applicable to born-digital data its principles can be extended to other forms of geographic data such as antique maps and charts. Also, being the library of a cartographic data producer we need to create metadata records that follow ISO standards.

"The beauty of standards is that there are so many to choose from." Richard Tannenbaum.



Georeferencing of maps is key to the new application: a spatial catalog will be built upon the geographical area field to help users in discovering maps and retrieving them using a point and click interface. This new application will accommodate maps, series, atlases, books and photographs, both analogue based and born digital.

- Born digital materials
- Improved image browser in ContentDM
- Georeferencing of maps
- Interoperability with other catalogues



Nowadays we must face some important challenges:

- Born-digital materials will be part of our workflow: purchasing, cataloguing, storage and retrieval.
- Improved image browser in Content DM through Zoomify, Openlayers or Google API interface.

- Also, we'll try to customize each collection to provide extra value for every customer: researchers, librarians, students and general public.
- Georeferencing of maps to provide geographic search. The natural interface to such a catalogue is a point and click approach. Google generation users are moving from text entry to graphical queries.
- Interoperability with other catalogues: data harvesting will create new scenarios to data sharing where users will be able to launch federated searches from a single access point.

References

A Framework of Guidance for Building Good Digital Collections [on line]. 3rd edition. National Information Standards Organization (NISO), December 2007 – [Consulta: 11/04/2008]. Available: <http://www.niso.org/publications/rp/framework3.pdf>

Dublin Core Metadata Initiative [on line] – [Consulta: 11/04/2008]. Available: <http://dublincore.org/>

Estàndard ISO 19115 – Informació geogràfica – Metadades – Perfil IDEC. Esborrany fina [on line]. Infraestructura de Dades Espacials de Catalunya (IDEC). Generalitat de Catalunya. – [Consulta: 11/04/2008]. Available: <http://www.geoportal-idec.net/geoportal/cat/docs/perfilidec.pdf>

Marc to Dublin Core Crosswalk [on line]. Network Development and MARC Standards Office. Library of Congress, February 2001. – [Consulta: 11/04/2008]. Available: <http://www.loc.gov/marc/marc2dc.html>