Big Data in the Digital Cultural Heritage

Antonella Fresa, Promoter Srl

DCH-RP Technical Coordinator

Table of Content

- Digitisation of Cultural Heritage
- Toward an e-Infrastructure for Digital Cultural Heritage
- Digital Preservation of CH data
- Sustainability of DCH
- Achieving impact

Digitisation of cultural heritage

Background

- The European amount of **digitized material** is **growing** very rapidly
- Museums, Libraries, Archives, Archaeological sites and Audiovisual repositories are committing to bring their heritage online
- National, regional and European authorities support the digitization processes with specific programmes
- Despite such wide mobilisation, still mostly each country, each sector and often each organisation have **different policies, procedures and guidelines** for accessing, sharing, managing and preserving the content

Recent studies commissioned by the EC showed that:

- 83% of cultural institutions said curatorial care is part of their mission
- 83% of cultural institutions have a digital collection or are currently involved in digitisation activities
- 20% of the collections that have been identified to be digitized, have been actually digitised
- Concerning born digital collections:
  - 89% of audio visual institutions have them
  - 43% of museums of art and history have them
- 34% of institutions have a written digitisation strategy
- About one third of the institutions are included in a national digitisation strategy (for national libraries, more than half are included)

Sources:

- **NUMERIC** Study Report: cordis.europa.eu/fp7/ict/telearn-digicult/numeric-study_en.pdf
- **ENUMERATE** Survey Report on Digitisation in European CH Institutions 2012: www.enumerate.eu/fileadmin/ENUMERATE/documents/ENUMERATE-Digitisation-Survey-2012.pdf
- **Digital Renaissance** EC Comité des Sages Report on Cost of Digitising Europe's CH: ec.europa.eu/information_society/activities/digital_libraries/doc/refgroup/annexes/digiti_report.pdf

Digitisation of cultural heritage imposes several reflections

1. To establish a Digital Cultural Heritage **research community** within a dedicated e-infrastructure
2. To address **digital preservation** as priority service for the digital cultural heritage
3. To enable the **dialogue** between sectors who are not used to work together
4. To **recalibrate relationships** between cultural heritage institutions, citizens, academies and creative industry
5. To engage **citizen scientists** in the research on cultural heritage

Towards an e-Infrastructure for Digital Cultural Heritage

Digital Cultural Heritage: ICT needs

- High quality information technology management, to ensure **trust, availability, reliability, long-term safety of content, security, preservation and sustainability**
- Enhanced **access** facilities
  - For the researchers who will look for contents into the DCH e-Infrastructure for their research
  - For the cultural institutions that will deliver their data to the DCH e-Infrastructure
- **Interoperation** among existing cultural heritage repositories, among cultural portals and among data from the digital cultural heritage and from the research

Main challenges

- High investment for the production of DCH data due to the need of **human intervention** of experts
- High costs of digital preservation, due to the use of **separate solutions** implemented by each memory institution
  - The estimated total cost of digitising the collections of Europe's museums, archives and libraries, including the audiovisual material they hold is approximately €100bn, or €10bn per annum for the next 10 years
  - The cost of preserving and providing access to this material over a 10-year period after digitisation would be in the order of €10bn to €25 bn, provided that "federated" repository infrastructure is made available for the purpose
- DCH content is **complex** and **interlinked** through many relations
- **Contextual** data are very important for cultural research
- The digitisation process is **unique** cannot be replicated unless the whole work is done from scratch

e-infrastructures and DCH

- 2 twin-projects (**DC-NET** and **INDICATE**)
- an ongoing international coordination action (**DCH-RP**)

brought together in the last years memory institutions and e-infrastructure providers from all over Europe to work for the future, in order to create a data infrastructure devoted to cultural heritage research

- These initiatives are contributing to pave the way towards an **Open Science Infrastructure for Digital Cultural Heritage**

Benefits offered by the e-infrastructures to DCH

- To allow for **cost reduction** in digitisation, cataloguing and metadata generation by substituting expensive human workforce with cheaper machine processes
- To support the **permanent identification** of digital cultural objects and providers
- To facilitate **storage and preservation**, ranging from short- medium- to long-term
- To improve **search facilities** supporting semantic search and linked open data
- To enhance **processing and visualisation of complex cultural data** (e.g. 3D modelling and VR representations) through the computing resources offered by research e-infrastructures
- To enable dynamic distributed **virtual organisations**, facilitating collaboration with information and resource sharing (e.g. virtual conferences, document sharing, blog, cooperation platforms, ...)

Data retention and storage

- In the DCH sector, data which are digitised are then retained. The main **retention requirements** are:
  - Separation of content and metadata
  - OASIS compliance
  - Accessibility through powerful retrieval and search system
- Generally, data and metadata are stored in **data centres/ repositories hosted by the memory institutions** themselves
  - this poses big maintenance issues due to the lack of ICT expertise

Access to data

- Access to shared resources is very appealing, but there are still problems in adopting this approach:
  - Issues related to **copyrights**
  - Cultural data are curated by many different persons: **data management and administration + user access control** are very important
  - Trust building is a key factor (particularly when it is not determined where data are stored)
  - Access to the e-infrastructure services should be simple without requiring **IT specialist knowledge**
- The creation of the **online presentation** of DCH materials is a central part of any digital heritage initiative (content management system, portal, digital library, digital repository.

Authentication and authorisation

- Access **requirements**:
  - Generally, most of the data should be available to others
  - Usually, access to these data is open for view only; protected instead for importing and updating data
  - Adding and editing data needs to be password protected and limited to known individuals authorised by the institution
  - Authentication mechanisms most in use: Open access, Password protected, IP-based, Shibboleth or equivalent
- **Federated access** can be a valid approach:
  - To reduce the number of credentials for the users,
  - To increase security
  - To improve users experience (sign in once, access more resources)

Digital Preservation of DCH data

Digital Preservation

- Digital preservation of cultural heritage data has been identified as the **highest priority** for the DCH sector (crf. Service Priority Handbook produced by DC-NET)
  - 23% of institutions have a written digital preservation strategy, figures range from 44% for national libraries to 12-25% for museums
  - About a third of the institutions are included in a national preservation strategy
  - 40% of national libraries say there is no national digital preservation strategy
  - 30% of the institutions are included in a national digital preservation infrastructure
- Type and size of content to be preserved vary from case to case
  - **Types** include: texts, still images, 3D models, publications, digital exhibitions, virtual reconstructions, etc.
  - **Size** range from 5 to 200 GB

Preparing for preservation of DCH data

- Standards and their reference implementation
- Persistent identifiers
- Roadmap for digital preservation of cultural heritage data

Use of standards

- The extensive use of **relevant and open standards** is a vital pre-requisite for the CH community to promote interoperability, encourage widespread access and control costs in its digital preservation programmes.
- Extensive reviews under the auspices of the Minerva (2008), Athena (2009), Linked Heritage (2011) and DCH-RP (2013) projects categorized and described many of the standards that are most applicable or recommended in this area.
  - Examples are: EAD, OAIS, ONIX, Indecs, EDM, Premis, CIDOC-CRM, FRBR, EPIDOC, Dublic Core, SKOS; MARC, METS, MAB2, MODS, Museumdat, ObjectID, SPECTRUM, LIDO, BIBO, etc.
  - Furthermore, the OAI-PHM/OAI-DC standards are used to aggregate data and to make data available for publication in other portals (as in the case of Europeana).

Data types and formats

- DCH content is composed of **several different typologies** of information and **different formats**: texts, still images, 3D models, publications, digital exhibitions, virtual reconstructions, etc.
- Examples of standardised formats often used by memory institutions are:
  - **Documents**. Public authorities and other institutions producing electronic documents and media content on national level are normally using open standards adapted to specific requirements to produce their electronic files. **PDF/A**, and its different versions, is for example the standard mostly used by archiving institutions for electronic documents
  - **Images**. **TIFF** and **JPEG2000** are the preservation format most often used by memory institutions for still image digitisation
  - **Audiovisual** contents. The Material eXchange Format (**MXF**), the container format (developed and maintained by audio-visual industry, particular for postproduction and distribution purposes) + **M-JPEG200 and FFV1** for the actual coding

Reference implementation of standard formats

- PREFORMA: 

- Pre-Commercial Procurement project
  - Developing tools for implementing good quality digital archives based on standardised file formats
  - Giving to memory institutions full control of the process of the conformity tests of files to be ingested into archives
- Coordinator: National Swedish Archives
- Technical Coordinator: Promoter
- Running from 1/1/2014 until 31/12/2017
- Expected outcome: open source instruments for compliance tests, reporting and suggestor of digital archives

Persistent identification

- The **PID requirements** do not vary significantly from one DCH initiative to another and it represents a service useful to most DCH work.
- Primary candidate for use in Digital Cultural Heritage are:
  - general digital identifier: **URI** (Universal Resource Identifier), **URL** (Universal Resource Locator) and **URN** (Universal Resource Name)
  - service-associated digital identifier: **PURL** (Persistent URL) & Handle System, **DOI** (Digital Object Identifier), **OpenURL** and **ARK** (Archival Resource Key)
- Arguably, identifiers which are maintained and associated to services are likely to offer more comprehensive features to CH institutions, but issues relating to both **cost** and **policy** have reduced their widespread adoption.

A Roadmap for digital preservation

- Unlike digitisation, where common approaches and best practices are rather well developed, digital preservation is still an area where workflows and easily applicable universal toolkits are not widely available.
- The creation of a Roadmap can contribute to develop a commonly agreed **vision** of distributed digital preservation architecture relying on e-Infrastructures:
  - Harmonising data storage and preservation **policies** among cultural institutions
  - Including **integration** of preservation within the overall workflows for digitisation and online access
  - Fostering participation of DCH in larger data **e-Infrastructure** initiatives
  - Establishing the conditions for **aggregation and re-use** of digital resources

Sustainability

Data re-use

- 'Users' of digital cultural heritage data are **researchers, curators** and the **general public**
- 're-users' are **Cultural and Creative Industries**
  - **Cultural Industries** comprise museums, libraries, cultural tourism, as well as education and research in cultural domains
  - **Creative Industries** comprise arts (visual and performing arts), architecture, design, crafts, fashion, music, film, publishing, advertising, TV and Radio, toys, video games and serious-educational games.
- Use cases for the re-use of DCH exists for: **educational** products, **commercial** ventures (e.g. publishing, tourism), **collaborative social projects** and **digital exhibitions**

Limiting factors

- The use of cultural heritage content by the creative industries is still limited by factors including:
  - Issues around the **IPR** status of content
  - Poor metadata **quality**
  - Successful **business cases** demonstrating the potential for exploitation of digital cultural content
  - Lack of **awareness** by the cultural heritage sector about the exploitability of the cultural assets in the digital world

Business models for the re-use of digital cultural content

- Memory Institutions need to digitize their content primarily for **preserving** it in a digital format and for granting and enlarging **access** to them by researchers, teachers & students and citizens. These are public services that need public funding for their sustainability.
- Museums, libraries, archives should become **content providers** and **service providers**, exploring new audiences and markets and attracting further investment in digitisation of cultural content
- New projects have been funded in the last years by the EC (such as Europeana Photography and Europeana Space) to experiment with innovative applications and services for the **creative re-use** of cultural resources

About the re-use of DCH

EarlyPhotography

- A CIP ICT PSP Pilot B project to digitise 430,000 images from the most prestigious photographic archives, public libraries and photographic museums covering the length of time from the beginning of photography to the beginning of the Second World War (1839-1939)
- Special attention is devoted to the management of intellectual property, which is further emphasised by the involvement of content providers from both the private and the public sectors

E-Space

- A CIP ICT PSPS Best Practice Network to increase and enhance the creative industries' use of DCH content by delivering 6 pilot applications in the domain of performing arts, interactive TV, games, open publishing, apps and toolkits for museums
- The project aims to address the problems which limit the re-use of DCH data by creative industries, such as the IPR status of content and successful business models

Achieving impact

Steps to deliver impact

- To provide information, news, links to knowledge resources that can support the **dialogue** between different actors

- To understand the **changes** generated by digitisation in the society
- To **engage citizens** in digital cultural heritage
- To create environments for **collaboration**, where memory institutions, creative industries, ICT providers and users can meet and interact
- To develop spaces of **business and innovation** where products and services can be promoter, also through **public-private partnerships**

PROMOTER DIGITAL MEETS CULTURE .net PROMOTER

DIGITAL HERITAGE DIGITAL ART

Login

Username

Password

login Forgot? Register

Free text

search

Upcoming events

Taipei, Taiwan, 22-28 March 2014

International Symposium on Grids and Clouds (ISGC) 2014

Gothenburg, 25-26 March 2014

Chalmers Initiative Seminar on Big Data

Life through the lens of Europe's first photographers (1839 – 1939).
Pisa, 11 April - 2 June 2014

All Our Yesterdays. Europeana Photography exhibition

In Pisa, at Palazzo Lanfranchi, from 11th April to 2nd June 2014, a great photographic and multimedia exhibition based on the most advanced digitization and printing technologies, telling the stories of our grandfathers. The exhibition is organized in the framework of the EU-funded project Europeana Photography. Continue reading →

Using 3D printers to turn 18th Century etchings into real sculptures

Factum Arte brings Piranesi's etchings to life using 3D printers

The Italian design company that has produced these 3D versions of Piranesi's work, used the "largest stereolithographic printers in Europe" and "routing, milling and laser cutting" in their creation process.It remains unclear to what extent we can describe these works as Piranesi's own, if they were constructed hundreds of years after his death. Continue reading →

Digital Conversations Meetups prove useful for debate on key issues facing the digital world

DIGITAL HERITAGE SHOWCASES

Recalibraring relationships

- The shift of Cultural Heritage in the digital world is changing the relationship between cultural institutions and their users
- A socio-economic approach is needed
- RICHES, a project about change:
  - About decentring of culture and cultural heritage away from institutional structures towards the individual
  - About the questions which the advent of digital technologies is posing in relation to how we understand, collect and make available Europe's cultural heritage

RICHES

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 612789

- www.riches-project.eu
- info@riches-project.eu
- Coordinator: Coventry University
- Communication Manager: Promoter
- Running from 1/12/2013 until 31/5/2016

Citizen scientists

- Supporting the engagement of citizens in the research processes requires:
  - To assign roles and functions
  - To offer training and information
  - To provide instruments
- Broadening e-Infrastructure deployment to support the participation of citizens to digital cultural heritage and humanities research

CIVIC EPISTEMOLOGIES

- An FP7 CSA project under negotiation
- Coordinator: Italian Ministry of Economic Development
- Technical Coordinator: Promoter
- Expected start date: 1/6/2014
- Duration: 14 months
- Main outcome: a validated Roadmap of direction that the deployment of e-Infrastructures should take to support engagement of citizens e creative industries in the exploitation of the investment done in cultural heritage digitisation

Next Appointments

- **4 April 2014, Brussels**
  - Information Day to present the new procurement launched by PREFORMA (http://www.preforma-project.eu/info-day.html)
- **23-24 April 2014, Tallin**
  - DCH-RP Concertation meeting between cultural heritage institutions and e-Infrastructure providers (http://www.digitalmeetsculture.net/article/dch-rp-e-infrastructure-concertation-workshop/)
- **20 May 2014, Helsinki**
  - Workshop @ EGI Community Forum 2014: e-Infrastructures and services for data preservation and curation (http://www.digitalmeetsculture.net/article/e-infrastructures-for-data-preservation-and-curation-egi-cf-2014-helsinki/)

Thank you!

Antonella Fresa
Promoter Srl
fresa@promoter.it
www.digitalmeetsculture.net