Data Curation in Cultural Heritage Institutions: Two Case Studies

Klaus Kempf[1], Anna Maria Tammaro[2] and Stefano Casati[3]

[1]*Bayerische Staaatsbibliothek Munich Germany*
[2]*University of Parma, Parma, Italy*
[3]*Museum Galileo Digital Library Florence Italy*

Abstract

The research analyzes the data curation practices carried out by two digital libraries: the digital library of the Bayerische Statsbibliothek and the digital library of the Museo Galileo. Four lines of data curation activities are analyzed: Access, Workflow, Data representation, Reuse. Some considerations on data curation and the problems that digital libraries need to improve are highlighted.

Keywords

Data curation, Bayerische Staatsbibliothek, Museum Galileo Digital Library

1. Introduction

In recent decades, several projects have been carried out to digitize the cultural heritage owned by archives, libraries and museums (LAM) that highlight the importance of data curation standards and good practices. In the present recovery phase after the pandemic, where digitization projects are financed to extend democratic access to cultural heritage, it is very useful to examine the developments of the data curation practices carried out so far by LAM and to consider the approaches that have been adopted, what are their theoretical frameworks and where there are gaps.

Rather than merely presenting the technical challenges, we intend to analyze organizational and social challenges not as separate considerations, but as integral parts of the structure of the whole. For this purpose we examine two case studies: the case of Bayerische Staatsbibliothek (BSB) and the case of Galileo Museum Digital Library (MGDL). The two case studies were chosen because they represent two pioneering digital library experiences that have adopted different models of digitization: the mass digitization of the cultural heritage of the BSB and the MGDL digitization of a specialized collection distributed in different libraries. The aim pursued by the two case studies was similar: to provide general access to scholars to hardly visible collections or to meet the needs of a specific research program.

The Bayerische Staatsbibliothek in Munich is the central library of the Free State of Bavaria and one of the most important universal libraries in Europe. The BSB was a pioneer in mass digitization, starting a collaboration with Google Book at the end of the 90s and at the same time opening an internal digitization center called Munich DigitiZation Center (MDZ) to manage the entire workflow from production to preservation.

The Digital Library of the Museo Galileo was born in 2004 as a specialized library for the history of science. The MGDL was one of the first digital library projects carried out in Italy with the contribution of the Ministry of Cultural Heritage, it integrates various archives of texts, images, 3D objects from the Galileo Museum and other partner libraries. The Digital Library of the Galileo Museum has set up an internal center for supporting the phases of creation, management and preservation of digital content.

The paper aims to summarize several current and emerging trends in data curation in heritage institutions, with a strong emphasis on the use of technologies, as tools capable of exhibiting, acquiring

IRCDL 2022: 18th Italian Research Conference on Digital Libraries, February 24–25, 2022, Padova, Italy
klauskempf@gmx.de (K. Kempf); annamaria.tammaro@unipr.it (A. Maria Tammaro); s.casati@museogalileo.it (S. Casati)
0000-0003-3674-5437 (K. Kempf); 0000-0002-9205-2435 (A. Maria Tammaro); 0000-0002-7943-7955 (S. Casati)

and transforming digital representation on multiple levels, along with organizational and social implications.

2. Data curation

Data curation includes "all the processes required for principled and controlled data creation, maintenance and management, along with the ability to add value to data." (see Wikipedia and [9])

Data curation is a broad term to indicate processes and activities related to the organization and integration of data collected from various sources, their enrichment as well as their publication and presentation so that their value is maintained over time and remains available for reuse and preservation. In the modern era of big data, the curation of data has become more prominent, particularly for software processing high volume and complex data systems. In science, data curation may indicate the process of research data management and extraction of important information from scientific texts, following FAIR principles.

In cultural heritage institutions, the transition from predominantly analogue to predominantly digital acquisitions requires significant changes in thinking and practices [6].

Organising cultural heritage institutions in the digital environment is addressed in four lines of data curation activities:

- **Access**: discovery/data retrieval;
- **Workflow**: maintenance and improvement of the quality of data;
- **Data representation**: addition of value to data;
- **Re-use**: re-use of data including preservation.

2.1. Access

Data curation allows and improves the accessibility and traceability of data. For example, it offers researchers the possibility of integrated research tools through different and heterogeneous data sets, using semantic web technologies. It also allows enrichment of the user interface by improving presentation and display techniques.

Case studies:

BSB

The BSB is following the principle that the quality of access and of data retrieval is fundamentally determined at the stage of producing objects / digital content in the best quality and by adding a complete set of metadata. At the Bayerische Staatsbibliothek, for example, to enhance the visibility and accessibility of digital assets, metadata and/or object data itself aren!t only visible and accessible via the local OPAC, but they are integrated into regional, national and global catalogs and various portals such as:

- Deutsche Digitale Bibliothek[1]
- Europeana[2]
- World Digital Library[3]
- bavarikon[4]

In addition, the BSB participates with the data/metadata of its digitised copyright free holdings on a nationwide network of so called specialised platforms for the single disciplines of science ("Fachinformationsdienste"). Another possibility of access is the creation and offer of online / virtual exhibitions.

MGDL

[1] https://www.deutsche-digitale-bibliothek.de
[2] https://www.europeana.eu
[3] https://www.loc.gov/collections/world-digital-library/about-this-collection/
[4] https://www.bavarikon.de

The Museum Galileo Digital Library [1] allows free consultation of the content, except of course for the publications covered by copyright, which can only be consulted from the locations of the Library or remotely, by issuing authorization and relative password. The Digital Library also offers a free low-resolution download service. Some digital collections are also available at portals:

- Europeana[5]
- Internet Culturale[6]

The reading system is characterized by a dual navigation mode that allows the reader to "browse" the text or to use a structured index. For some works, the structured index provides links to related resources for in-depth study of the topics. The index constitutes an added value to the publication and, for the cases of stripped works, it is composed automatically. In the interface of the digital library there are other operating keys that also allow the display and searches of documents in textual form. To improve the readability of the text, a zoom can be activated; in addition, a gallery of illustrations can be browsed.

2.2. Workflow

The workflow, according to the OAIS model, includes all activities and tools from the initial creation of the digital object to the final presentation on the portal or on the digital library platform. The workflow in data curation is iterative and automated: each digital object that is digitized follows an automated workflow with a reduction in time and costs. The system that controls the workflow, or Digital Asset Management System, allows to realize the entire production process with a modular system that can extract data from different service providers. One important aspect of the workflow concerns the quality of the data and its enrichment, to facilitate the reuse.

Case studies:

BSB

The Bayerische Staatsbibliothek has a data curation policy which includes resource management for optimizing employee employment. Considering the large number of digital objects (over 2 million volumes) and above all the necessary migration of numerous sub-collections, which until now have more or less individual software solutions, an improvement in efficiency can only be achieved with standardization and strictly flow-oriented quality control of possibly all work steps in digital production. Each original is scanned once with the best possible quality and high quality and high resolution scanning requires permanent and systematic quality control with the use of the "Metamorfoze"[7] tool.

MGDL

The Digital Library Management System is a web application developed in house by the Museum Galileo Digital Library for using on the Intranet and managing all stages of the workflow process, from image acquisition to publication in the TECA Digitale. The Digital TECA is an application for the use of digital resources. The STORAGE component is the most critical system in terms of reliability and sustainability.

The workflow adopted for Galileo Digital Library requires the collaboration between computer scientists, systems engineers and librarians who design the service in a participatory approach with scholars and researchers. This collaboration has an impact on the acquisition policy, where more attention is paid to qualitative content than to 'quantity' (i.e for selection of object, specific training needs).

2.3. Data representation

Digitization is about the digital representation of cultural heritage objects. It is necessary to model the data packages, i.e. not only texts and images, but different types of objects such as 3D models. It is

[5] https://www.europeana.eu
[6] https://www.internetculturale.it
[7] Metamorfoze Preservation Imaging Guidelines are a tool for the production and preservation of images developed by the National Library of the Netherlands

possible to add value to digital objects, for example through in-depth indexing and incorporation of semantic structured data as Linked Open Data (LODs), and also by creating new contexts and developing new original services. Interoperability makes it necessary to harmonize different conceptual models for particular types of digital objects and at different levels of granularity.

Case studies:

BSB

To realize the potential of access and representation of information, Bayerische Staatsbibliothek adds a set of metadata as complete as possible (technical-administrative metadata, bibliographicstructural metadata, including a persistent identifier) and uses authority control services where and when ever available. The quality of the volumes scanned by Google is corrected and improved and permanent corrections are made on digital images and / or metadata. Electronic summaries (ToCs) are created and new collection contexts generated (Examples: German Reichstag Minutes + GND ADB / NDB)

An essential aspect of data curation is the inclusion of a quality policy and continuous quality control during the production of digital images: the resolution and sharpness of the image, as well as the color management, are essential parameters.

In this context, not only the reproduction technique available internally (scanners and digital cameras) is constantly being renewed. In addition to this, other quality assurance measures are taken. One of them is the systematic use of Metamorfoze.

MGDL

The Galileo Digital Library is a new generation thematic digital library, which collects texts, images, documents, bibliographic references, chronological repertories, lexicons, thematic indexes, catalogs of objects and experiments, research aids, etc., on every aspect of Galileo's life, cultural activity and fortune. MGDL consists of two systems: Galileo's personal library and Galileo//thek@[8].

Other systems are used for integrating collection of digital resources, such as Sinapsi[9], for accessing the Leonardo's Library, the Iconographic Collection Portraits of the members of the Georgofili Academy and Bibliotheca Perspectivae (in progress).

Galileo Digital Library has also created many digital born resources, increased by the Multimedia Laboratory and "Wiki projects" [3]. A part of the Cumulative Database of MGDL, entitled "Galileo Museum database: tools, books, photographs, documents" has been selected for conversion into LOD (and therefore into RDF - Resource Description Framework) [1]. The project MINERV@[10] added a dataset of MGDL to Datahub (Open Knowledge Foundation) and to OpenData (Regione Toscana), following the principles of Linked Open Data.

2.4. Re-use

Data curation is essential for the preservation of digital data. Other features include helping in detecting errors, aggregating documentation, ensuring data reusability, and in some cases even adding additional features and files. Reuse is based on the design and evaluation of the interdisciplinary research approach of the human-computer interfaces.

Case studies:

BSB

[8] https://galileoteca.museogalileo.it
[9] http://www.progettosinapsi.it/soluzioni/
[10]https://www.museogalileo.it/en/news-archive/121-news-archive-2015/1590-museo-galileo-dataset-in-datahub-okf-and-in-opendatatuscany-b-en.html

The Bayerische Staatsbibliothek makes available new services (like ever gratis) such as "Data for scientific research" (Daten für die Forschung / DaFo) and provides for the main part of its (historical copyright free) digitized holdings using IIIF standard[11].

MGDL

The Galileo Digital Library put great attention to interoperability and preservation but this aspect requires further research. Two projects with WIKI and Google are working for the reuse of digital objects [3].

3. Conclusion

Digitization is an important step towards digital transformation and has a huge impact on the organization of cultural heritage institutions and their traditional procedures. In conclusion, we can highlight the following issues on which more research is needed:

- Data curation is an ongoing - or rather a never-ending process with always new challenges – due to (changing) technologies, costs and changing - even growing - user needs.
- To realize the full potential of access to the digital library, the challenge is to get to know users better and to be able to create a participatory approach and transdisciplinary collaboration.
- An essential part of re-use problem solutions is the collaboration not only between data holding institutions, but also the close interaction with the users.

4. References

[1] S. Casati, La Biblioteca digitale del Museo Galileo, Biblioteche oggi, Gennaio-Febbraio (2015), pp. 45-51

[2] S. Casati, I. Rolfo, Online la nuova versione della Galileo//thek@, Galilæana: journal of Galilean studies, A. 13 (2016), pp. 181-186

[3] S. Casati, C. Rotoli, La Biblioteca digitale del Museo Galileo e il progetto GLAM, Biblioteche oggi Luglio-agosto (2017), pp. 33-36

[4] S. Casati, A. Pocci, Le collezioni digitali tematiche del Museo Galileo: esperienze e nuove prospettive , in: Storie d'autore, storie di persone : fondi speciali tra conservazione e valorizzazione, a cura di F. Ghersetti, A. Martorano, E. Zonca, Roma, AIB (2020), pp. 273-280

[5] S. Casati, F. Butini, F. Viazzi, Redazione e uso di mappe strutturali: un esempio di cooperazione fra biblioteche digitali: la biblioteca digitale del Museo Galileo e la Biblioteca europea di informazione e cultura, Digitalia (2018), pp. 51-63

[6] P. Gerth, A. Sieverling, M. Trognitz, Data Curation: How and Why. A Showcase with Re-use Scenarios. In Studies in Digital Heritage (2017), 1(2), 182–193. https://doi.org/10.14434/sdh.v1i2.23235

[7] K. Kempf, Data curation oder (Retro-)Digitalisierung ist mehr als die Produktion digitaler Daten. In: o-bib Das offene Bibliotheksjournal (2015) Nr.4, Bd. 2/2015

[8] K. Kempf, Curated content come un aspetto centrale della politica delle raccolte nell'epoca digitale. In: La biblioteca che cresce. Contenuti e servizi tra frammentazione e integrazione. Milano 14-15 marzo 2019. Relazioni del Convegno. Milano: Editrice Bibliografica (2019), pp. 140-149

[9] R. J. Miller, Big Data Curation in 20th International Conference on Management of Data (COMAD) (2014), Hyderabad, India, December 17–19, 2014

[11] https://iiif.io