Intersection of Digital library and Data Science: learning data curation in a flipped classroom

Anna Maria Tammaro [1,*], Klaus Kempf [2]

[1] *University of Parma, Parma, Italy*
[2] *Bayerische Staatsbibliothek, Munich, Germany*

Abstract

Data curation is the core competency for both digital library and research data management. The paper describes a training course on "Data curation" and the challenge of designing a course at the intersection of digital library and data science. The survey of participants describes the gaps of skills and abilities that are perceived by LIS professionals.

Keywords

Data curation, Research Data Management, Continuous professional development, Flipped classroom.

1. Introduction

The word "Data curation" has been used for many years by the professional community but it remains a new and emerging practice for many librarians, novice and unprepared to face the digitization of collections, services and relationships. To date there is no shared definition of what data curation is, and agreement on the service and activities that the concept includes. The definition of Data curation we use in this paper is the following:

"*The active and ongoing management of data through its life cycle of interest and usefulness to scholarship, science, and education. Data curation activities enable data discovery and retrieval, maintain its quality, add value, and provide for reuse over time, and this new field includes authentication, archiving, management, preservation, retrieval, and representation.*" [1].

In the modern era of big data, curating data has become more important, particularly for processing complex, high-volume data systems. Data curation is an emerging field of theory and practice at the intersection of digital librarianship and data science. Both of these specializations are based on the common theoretical approach of Information Science with a user-centric focus. In science, data curation can indicate the process of managing research data during the research life cycle and extracting important information from scientific texts, following FAIR principles. In cultural heritage institutions, the transition from predominantly analogue to predominantly digital collections, requires significant changes in professional thinking and practices. The intellectual and practical framework developed for data curation to date highlights the concept of the importance of a data curator in adding value to any type of data including research data (science, social sciences and humanities). While arguing that data curation is a vital strategy for dealing with the so-called data deluge, there are key issues and debates in the digital libraries area, often confusing the simple application of technologies to the traditional flow of data and not considering digital transformation.

19th IRCDL (the Conference on Information and Research science Connecting to Digital and Library Science), February 23–24, 2023, Bari, Italy

* Corresponding author

EMAIL: annamaria.tammaro@unipr.it (A. Tammaro)
ORCID: 0000-0002-9205-2435 (A. Tammaro)

Data curation is an essential component of the digital object life cycle, which is not only created, managed, maintained, but can also be involved in transforming organizations and working with such data offering new services. Data curation is a fundamental and necessary skill in each stage of the digital object cycle. Data curation in digital libraries is about transforming the organization of cultural institutions and integrating data gathered from various sources. Data curation in research institutions involves annotating, publishing and presenting data so that the value of the data is maintained in its integrity over time. Often used for visualization of data such as a graph, dashboard or report, data curation is also beneficial to users engaged in data discovery and analysis.

In Data science, data curation may indicate the process of research data management and extraction of important information from scientific texts, following FAIR principles. In the Digital Librarianship context, the digital transformation requires significant changes in thinking and practices. While arguing that data curation is a vital strategy for digital transformation, there are key issues and debates in the sector professionals.

The paper is a reflection on the training course organized by Editrice Bibliografica, entitled Data curation, held online for four weeks starting from September 2022. The course aims to address the weakness of the professional skills of librarians engaged in the creation of digital libraries or in supporting research data management. The Data curation course described in this paper was the first experience to bring together the two contexts of Digital Library and Data science, combining digital object creation and research data stewardship. The course pursues the following learning objectives:

- to know the activities and infrastructures necessary in the various phases of the cycle of digital objects,
- to manage relationships with all stakeholders in the cycle of digital objects.

2. Learning data curation: a literature review

Traditionally cultural institutions have been responsible for the care and conservation of cultural heritage. However the digitization of digital heritage and the growth of vast amounts of research data has created new challenges for professionals to meet new community needs with new responsibilities and organizational change. Assuming these new responsibilities and also a leadership role in the digital transformation requires professionals to be competent and continuously updated in knowledge and skills.

2.1 Data curator profile

The formal titles under which data curators typically operate (such as data librarian, data stewards, research data manager, data scientist, data services librarian, etc.), can vary considerably between different institutional settings, over time and across international borders. The different profiles and different roles in the different stages of the data lifecycle (Fig. 1) have been described by DataOne [2] (Data Observation Network for Earth), founded in 2009.

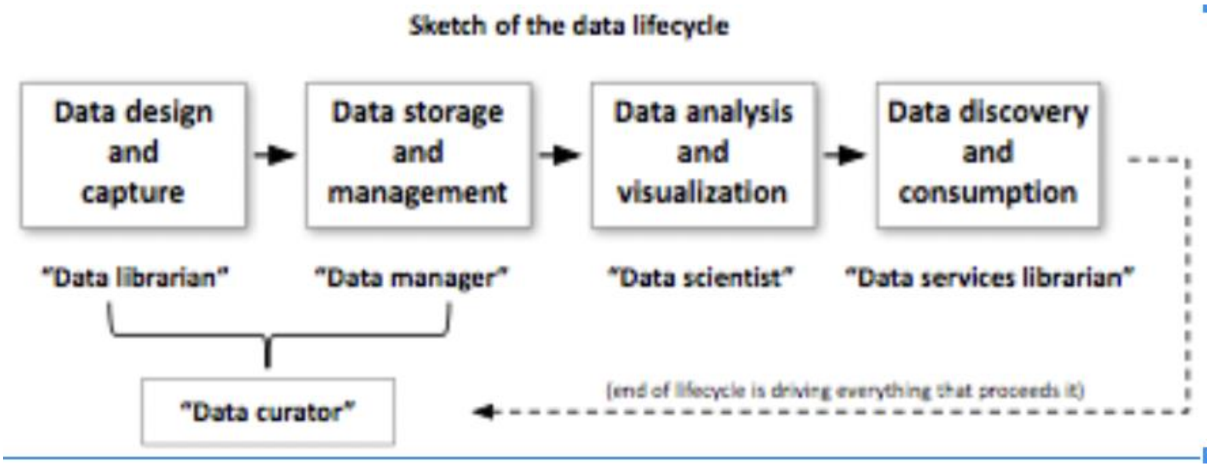
Sketch of the data lifecycle
Data design and capture
Data storage and management
Data analysis and visualization
Data discovery and consumption
"Data librarian"
"Data manager"
"Data scientist"
"Data services librarian"
"Data curator"
(end of lifecycle is driving everything that proceeds it)

Figure 1: Profiles on data lifecycle

There are different names for describing these roles in different steps of the data lifecycle, such as: Data Curators, Data Stewards/Data Librarians, DataScientists/DataAnalysts. In the annotation of their skills sets, DataOne authors include: they are always concerned with data quality, they are always working with people, including administrators and politicians [3, p. 5-15]. Similar results were obtained from the research of the Library Theory Section of IFLA which carried out a worldwide survey on the role of the data curator [4].

In Europe the EOSC Executive Board Skills and Training Working Group [5, p. 17] has developed a Framework that identifies ten roles from four distinct areas of knowledge: ICT, library and information science, research and citizen science. EOSC has also identified the different actors involved in data curation, to highlight the so-called “ecosystem” and the necessary collaboration between stakeholders, especially in support services.

In the context of digital scholarship, digital transformation is well advanced. The digitization of scientific activity is leading to a profound change in the conditions of production of scientific knowledge. This promotes the need for openness in terms of interoperability and freedom of access: the Open Science. To meet the needs of scientific advancement data curations is needed

The professional community has agreed on fundamental principles and some standards.

The standard called Open Archive Information System OAIS [6] provides an information model for managing digital assets as they pass through the archive system. The basis of this model is the Information Package (IP). It defines the following activities involved in managing digital objects: Ingest, Access, Data management, Archival Storage, Preservation planning, Administration and Management.

In the context of Open Science, the UK model of Data Curation Center (DCC) is the ideal vision of an Research Data Management (RDM) service. The DCC model includes two aspects: the infrastructure, including policies, planning, and training, and the more technical infrastructure requirements, which are based on the data lifecycle. The DCC definition of data curation evidences the dynamic data management challenge:

“Maintain and add value to a trusted body of digital information for current and future use; specifically, (it is) the active management and appraisal of data over the life-cycle of scholarly and scientific materials”[7].

The professional and research community has created extensive documentation on RDM, we can therefore highlight good practices, especially for the reproducibility of research, and we can summarize the reason for the need for data curation: this is fundamental since the birth of the digital object, it improves access, improves quality, facilitates reuse.

2.2 Continuous professional development

While there are standards and best practices for data curation, there are still no shared standards for professional development and education of data curators.

The rapid evolution of the skills required by the information professionals makes conventional systems of education and training insufficient. There have been some successful training projects offered by universities such as DIGCCURR [8] and the International Master DILL project [9, 10], together with university Master's programs mainly in the computer science area. However, besides the higher education system of Master diplomas, that depend on a multiannual course, in and next to higher education, some projects and training courses have been implemented that focus more directly on skills and adopt transdisciplinary schemas.

Is the traditional librarian background sufficient to fill these roles? Are LIS professionals aware of the responsibility for data curation (or are there different curation concepts)?

In 2011, on the occasion of the Stelline Conference, Tammaro [11] investigated skills in five areas, identified as: 1. Information (and data) management, 2. Transversal skills (Organisation; Collaboration; Communication), 3. Support Services; 4. Technologies; 5. Management. The results highlighted that the majority of librarians believe that information organization skills and transversal skills are also fundamental for data curation and librarians may be able to offer support services by acquiring specific technological knowledge.

Recently the European project CHARTER [12] tries to identify skills gaps in the heritage sector and propose training and curricula for the development of new skills for cultural heritage professionals. The characteristic of this project is that it seeks to make alliances with all the stakeholders in the culture sector to identify skills needed for the overall sector-specific growth strategy. Six areas of competencies have been identified: Management, Governance and Policy making, Preservation and Safeguarding, Recognition, Engagement and use, Research and development - Education.

The Biblio [13] Project: boosting digital skills and competencies for librarians in Europe, funded by the Erasmus program, also tried to define some professional profiles necessary for cultural heritage institutions, preparing a MOOC to facilitate the acquisition of the highlighted skills.

For Research Data Management (RDM) training needs, there are many education and training experiences that offer very useful Open educational resources (OERs) and competency frameworks for creating courses. One example is FOSTER who created an ontology of Learning Objectives. A table with learning objectives lists services, tools and standards has been shared to facilitate course design. To agree on data curation terminology, the Terms4FAIRskills Project [15] has created a formalized terminology that describes the competencies, skills and knowledge associated with making and keeping research data FAIR (Findable, accessible, interoperable, reusable).

There are no data curation courses however that combine together the two aspects of cultural institutions and research data management. The Bibliografica professional development course presented here was intended to fill this gap.

3. Methodology

The challenge of the Editrice Bibliografica training course “Data curation” has been to design an innovative course at the intersection of Digital Library and Data Science. The authors have combined their respective experiences in cultural heritage and research data management.

The course has been designed in elearning and based on a mixture of experiences and theoretical fundamentals.

The course was the first offered in Italy to acquire data curation skills. Therefore the authors mainly relied on their own experience, in addition to the literature references mentioned above. At the end of the course, a questionnaire was distributed to the participants, involving them in defining their perceptions of skills gaps and the skills considered fundamental.

3.1 The Learning Program

Following the definition of data curation linked to the data lifecycle, the course Data curation covers four areas of competencies for expected results, that are common to digital libraries and research data management. These are: Data design and collection; Data storage and management; Data discovery and consumption; Data analysis and visualization. Of the four competencies areas within the course, the first three areas are fairly common to information management systems, especially when dealing with digital object creation, collection and organization and interoperable technical infrastructure. The remaining fourth area is more closely aligned with key dimensions of Library and Information Science (LIS) expertise and is central to the "purposeful curation" concept: user communities and their information behavior; networking and community participation, policy development, and intellectual property.

The syllabus of the course includes four Units: Workflow, Data Representation, Access, Reuse.

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 digital library portal or platform.
Learning objectives were:

- to plan a Data Management Plan,
- to be aware of Workflow optimization and reorganization,
- to choose a Digital Asset Management System,
- to understand Data quality.

The workflow in data curation is iterative and semi-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 capable of extracting data from different service providers. An important aspect of the workflow is data quality and data enrichment to facilitate reuse.
This Unit discussed how to choose between systems, selecting out of box fundamental components. Even if DAMS is semi-automated, it is still by hand: who makes the algorithm?
Another discussion is : What is Workflow optimization? Reorganization of cultural institutions and research centers as a challenge and an ongoing task, a never-ending process.
The course brought some examples of concrete cases to highlight that digitization is not a separate section from the existing reality and not to reinvent itself, digitization is above all a learning on the job.

Data representation

Digitization in digital libraries is about the digital representation of cultural heritage. Data packets need to be modeled, not just texts and images, but different types of objects such as 3D models. It is possible to add value to digital objects, for example through in-depth indexing and the incorporation of Linked Open Data (LOD) subjects, and also by creating new contexts and by developing and offering new original services. Interoperability makes it necessary to harmonize different conceptual models for particular types of digital objects and different levels of granularity.
Learning objectives were:

- to apply schemas for cataloging and classification of digital objects
- to know how to add value and enrich data.

This Unit introduced geographic metadata applied in interactive maps and using geospatial coordinates. The Unit discussed also the added value of metadata and the possibility of data enrichment also from users according to standard annotation practices and crowdsourcing good practices

Access

Data curation enables and improves data accessibility and traceability. For example, it offers researchers the possibility of integrated search tools across diverse and heterogeneous datasets, using semantic web technologies. It also allows for the enrichment of the user interface by improving presentation and visualization techniques. Access services must take into account different types of data, formats and different granularity needs. Digital scholarship and user behavior change including related open science community needs, changes the traditional access services based on Information retrieval.
Learning objectives were:

- to know the principles for interoperability,
- ability to design services based on community needs.

We have prepared 2-3 questions for guided conversation among the participants on: programs design good practices, project management, infrastructure and cooperation.

Reuse

Data curation is essential for digital data preservation. Other features include error detection, documentation aggregation, ensuring data reuse, and in some cases even adding additional features and linking to external files. Reuse is based on the design and evaluation of interdisciplinary research and human-computer interfaces. Learning objectives were:

- be aware of ethical and legal issues,
- to know preservation methods.

Methods and measures to stimulate reuse have been described using some examples from the Digital library of the Bayerische Staatsbibliothek, such as the Austrian National Library Year for historical newspapers, and the Bavaria digipress.

3.2 Pedagogy (Andragogy)

The second important course characteristic concerns its pedagogy or andragogy.

The course used e-learning and the methods called flipped classrooms. Each Unit has been based on interactive lessons, video with interviews to experts, preparatory and in-depth readings. Participants are guided to use collaborative tools for discussion and group work.

Some videos and tutorials and specific tools have been indicated for each Unit to prepare the lesson. This teaching material had two purposes: 1) to enable participants to be active in class discussion, 2) to fill the gap of practical activities. The teaching material was therefore an enrichment of the lessons, available to the participants before the interactive lessons. Participants had to go through this teaching material to participate in conversations and interact with the teachers during the lesson.

4. Results and discussion

There were 20 participants, coming from different contexts, such as university libraries, archives, public libraries.

The majority of course participants answered the questions in the final questionnaire entitled “How do you become a data curator?”. The questionnaire wanted to investigate the skills of the data curator that the participants consider fundamental. The questionnaire also wanted to understand which skills the participants expected from the course were perceived as acquired and which insights they felt the need for. Finally, the last question asked participants what they would recommend to a young student or fellow professional to acquire data curator skills.

4.1 Core competencies

The first question of the questionnaire distributed to the participants sought to know the perceived priorities of the data curator's core competencies. It is interesting to note that the ability to organize and manage the collection (creation, description, organization, preservation, discovery) is considered fundamental, but the role of the data curator is not limited to a technical aspect, instead the more social aspects of knowing the community, planning services, applying the policies and legal aspects characterize the profile in the same way. Knowledge of the technologies and tools needed in every single phase of the cycle are also considered fundamental. Only the promotion and guidance and data literacy instructor role of the data curator is perhaps underestimated. The profile of the data curator that is drawn is therefore a multidisciplinary and complex profile with a wealth of knowledge and skills that goes beyond the traditional organization of knowledge (see Fig. 2),

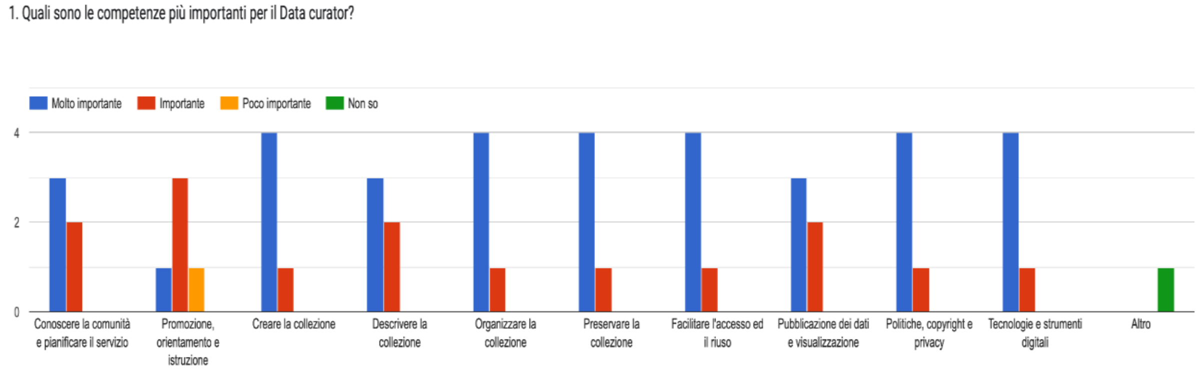

Figure 2: Perceived data curator competencies

When asked what the participants learned from the course, the answers collected show that they learned the importance of adapting the workflow to the data life cycle, together with the importance of networking with the community and the various stakeholders. Particularly:

"*The need to pursue interoperability and the possibility of reuse; ability to design a digitization and data curation project, also if small; ability to use tools for Digital Curation*"

"*Knowledge of technologies, of existing networks, of the possibilities of creating and using a database or of connecting to existing ones, possibility of updating*"

"*In particular, I learned the importance of data curation planning, especially through the use of supporting tools such as the various types of Data Management Plans*"

"*I learned the various stages of the workflow, exemplified by the case study of the Digitization Center of the Bayerische Staatsbibliothek and the various portals they manage*"

"*I understood that the creation of networks of various institutions/stakeholders is fundamental both for planning projects and for the effectiveness of their implementation*".

4.2 Competencies gap perception

Participants perceive that they need to deepen the skills learned in the course, which was only basic. In particular, the following tables identify the needs for further training that have been highlighted for each Unit.

For the workflow, the importance of project management is perceived, together with awareness of specific software tools to facilitate data management.

Table 1

Workflow gap perceptions

Learning outcomes - Workflow	Participants skills gap perceptions
Data Management Plan Workflow optimization and reorganization Digital Asset Management System Data quality	Data management throughout their life cycle Prior knowledge of the material being digitised Knowledge of specific software for the collection, their advantages and limitations Project management skills

For data representation, they would like to learn more about different metadata schemas for different types of digital objects and dynamic data, improving discovery.

Table 2
Data representation gap perception

Learning outcomes - Data representation	Participants skills gap perceptions
Cataloging and classification of digital objects Adding value and enriching data (e.g. improving data quality through correction, etc)	Capability to attribute metadata to digital objects Knowledge of metadata standards Capability to improve searchability

Service design is participatory and participants would like to enhance their skills to improve access and extend services to new functions.

Table 3
Access gap perception

Learning outcomes - Access	Participants skills gap perceptions
Interoperability Ability to design services built on created/collected data, based on community needs	Capability of data analysis skills Creation of cooperation networks with external partners and with the community

Reuse highlights the social role of the data curator and participants perceive that this is the area of expertise in which they have the greatest training needs. Being able to collaborate online and create networks seems to be the priority for acquiring new skills.

Table 4
Reuse gap perception

Learning outcomes - Reuse	Participants skills gap perceptions
Ethical and legal issues Preservation	Learn to networking Knowing how to create cooperation networks with external partners and with the community

4.3 Final consideration

Finally to the question: "What would you recommend to a young person or professional who wants to become a data curator", the answers show that the results of the course have been achieved:

"*First of all, I would recommend attending a course or training activity like this, which clarifies upstream the stages and development of the workflow, standards, good practices, and the study of the huge amount of material made available by various institutions in terms of supporting learning and structuring projects*"

"*First, don't take anything for granted, be open to new possibilities and technologies, be very attentive to your users and the collection you want to incorporate into this service*".

In conclusion, data curation is the fundamental service of the "hybrid" library of today. The push for digital transformation is driven by the emancipation of the user from the library, which had the effect of a paradigm shift in the library from media- or collection-centered service to user-centered service. Other conditions are represented by the competition or competitive situations with commercial information providers, as well as the dissolution of boundaries between memory institutions, such as libraries, archives and museums. Libraries and librarians are not the main actors anymore [16], but they are driven by the increasingly rapid development of information technology and have new opportunities. New opportunities for digital libraries are highlighted for example to improve access to information and fight against populism [17]. Data curation is the professional background that will be needed to seize these opportunities and help improve our evolving environment and communities.

5. References

[1] University of Illinois- CLIR https://www.clir.org/initiatives-partnerships/data-curation/

[2] DataOne https://www.dataone.org

[3] Michener W. K., S. Allard, A. Budden, R. B. Cook, K. Douglass, M. Frame, S. Kelling, R. Koskela, C. Tenopir, D. A. Vieglais (2012) Participatory design of DataONE—Enabling cyberinfrastructure for the biological and environmental sciences, Ecological Informatics, Volume 11 https://doi.org/10.1016/j.ecoinf.2011.08.007

[4] Tammaro, A. M, Matusiak, K., Sposito, F. A. and Casarosa, V. "Data Curator's Roles and Responsibilities: An International Perspective " Libri, vol. 69, no. 2, 2019, pp. 89-104. https://doi.org/10.1515/libri-2018-0090

[5] EOSC Executive Board Skills and Training Working Group (2020) From Digital skills for FAIR and Open Science. https://eoscsecretariat.eu/news-opinion/digital-skills-fair-open-science-report-eosc-skills-training-working-group

[6] OAIS https://www.iso.org/standard/57284.html

[7] DCC Curation lifecycle model https://www.dcc.ac.uk/guidance/curation-lifecycle-model

[8] University North Carolina (2010) DIGCCURR https://ils.unc.edu/digccurr

[9] Tammaro A. M. (2013) Integrating Digital Curation in a Digital Library curriculum: the International Master DILL case study. Final DigCurV conference Framing the Digital Curation Curriculum, 6-7 May 2013, Firenze http://www.digcureducation.org/eng/Resources/DigCurV-2013-proceedings/Tammaro-paper20

[10] Tammaro A.M., Casarosa V. (2014) Research Data Management in the curriculum: an interdisciplinary approach, Procedia computer science 38 (2014): 138–142. doi:10.1016/j.procs.2014.10.023

[11] Tammaro, A.M. (2012) Una proposta (non?) sovversiva. Le competenze del bibliotecario dei dati, Biblioteche oggi, vol. 36, p. 65-71

[12] Charter https://charter-alliance.eu

[13] Biblio MOOC https://mooc.cti.gr/biblio-sc.html

[14] Foster. Learning objectives Table https://docs.google.com/spreadsheets/d/1UwsYf8fEFZzK8IPfK-7rFE3BO_VbjvOjQm3CigqBqyk/edit#gid=0

[15] Term4FAIRskills https://terms4fairskills.github.io

[16] Kempf K. (2022) Challenges of the new library world and opportunity for librarians. Presentation at Stelline Conference not published https://docs.google.com/presentation/d/1ec1oyw_dxGYKtvfj35cz1nCdZbSKq8NK/edit?usp=sharing&ouid=117467067778643130859&rtpof=true&sd=true

[17] Schüller-Zwierlein, A. (2022) Die Fragilität des Zugangs: eine Kritik der Informationsgesellschaft. Berlin Boston, DeGruyter/Saur