

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES



Paolo Brambilla

**Digital Curation in the Italian context:
new roles and professions for Digital Librarians**

Master thesis
International Master in Digital Library Learning
2015

Abstract

The aim of this research is to analyse the activity of digital curation in the Italian context, focusing the attention on the IT skills and competencies of the digital curator.

The main hypothesis comes from a reflection on the creation of new roles in the practice of data curation. The first objective is to demonstrate that data curation can be the basis for the development of different professions, such as data librarians, data journalists, data analysts etc. Another objective is to individuate a common core of technical and IT skills that can help them to find a specific specialization. Last, within a specific theoretical framework that concerns the data curation lifecycle, these new roles in data management allow digital librarians to add value to the whole data lifecycle, since curators contribute to find different patterns of knowledge within data and datasets.

A qualitative approach has been adopted, using collective and instrumental case studies as research method. Five semi-structured interviews have been conducted among people that deal with digital curation in Italy, such as librarians, digital humanists, data experts.

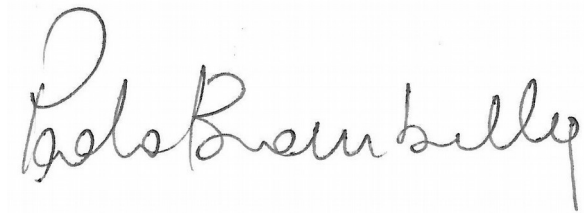
The results of the study show a varied context in which, although it is possible to individuate some basic IT skills, the specialization of the curricula calls for a more in-depth analysis. Additionally, a certain level of homogeneity among participants has been found in the perception and evaluation of the context.

Keywords: *Digital Curation, Data Curation, Digital Curator, Data Curator, IT Skills, IT Competencies, Data Management, Cyberinfrastructure, Open Data, Italy*

Declaration and Plagiarism Disclaimer

I certify that all material in this dissertation which is not my own work has been identified and properly attributed.

June 19th, 2015



Paul Brenbilly

Table of contents

Chapter 1: Introduction	1
1.1 Background.....	1
1.2 Motivation.....	3
1.3 Research question, aims and objectives.....	4
1.3.1 Statement of the problem.....	4
1.3.2 Research question.....	5
1.3.3 Research aims and objectives.....	5
1.4 Methodology.....	6
1.4.1 Data collection technique.....	6
1.4.2 Data analysis.....	7
1.5 Limitations of the research.....	7
1.5.1 Lack of time, economic and logistic limitations.....	7
1.5.2 Methodological flaws and possible biases.....	8
1.6 Ethical considerations.....	8
Chapter 2: Literature Review	9
2.1 An overview of digital curation.....	9
2.2 Data Lifecycle and added value.....	12
2.3 Outlining a technical curriculum for digital curators.....	17
2.4 New roles and new contexts for data experts.....	19
2.5 Digital curation in Italy.....	22
Chapter 3: Research Design	24
3.1 Conceptual framework.....	24
3.2 Research paradigm.....	26
3.3 Research methodology.....	26
3.3.1 Trustworthiness.....	27
3.3.1.1 Credibility.....	27
3.3.1.2 Transferability.....	27
3.3.1.3 Dependability.....	28
3.3.1.4 Confirmability.....	28
3.4 Research method.....	28
3.4.1 Target group and sampling.....	29

3.4.2 Key informants.....	30
3.5 Research techniques.....	31
3.6 Research instruments.....	32
3.7 Data analysis.....	33
Chapter 4: Analysis and Findings.....	34
4.1 Conceptualizing the work of the digital curator.....	34
4.1.1 A question of identity.....	35
4.1.2 Interdisciplinarity: a resource or a deficiency?.....	35
4.1.3 The value of Information Technology.....	36
4.1.4 The value of human contribution.....	37
4.2 Contextualizing the work of the digital curator.....	38
4.2.1 The perception of what cyberinfrastructure is.....	38
4.2.2 Open Data and collaboration.....	40
4.3 Technical competencies.....	41
4.3.1 Metadata standards and schemas.....	41
4.3.2 Extensible Markup Language.....	42
4.3.3 Ontologies.....	43
4.3.4 Data, database and data interoperability.....	44
4.3.5 Semantic tools: Linked Data and Semantic Web.....	45
4.3.6 Data visualization.....	47
4.4 New roles and professions for digital curators.....	48
4.4.1 Specialization.....	48
4.4.2 Teamworking.....	49
4.4.3 Critical issues for LAM networks.....	50
4.4.4 Back-end vs. front-end.....	51
Chapter 5: Conclusions.....	54
5.1 Research objectives.....	54
5.2 Implications for further research.....	58
References.....	60
Appendix 1: Interview guide.....	68
Appendix 2: Information letter and consent form.....	69

List of Abbreviations

API: Application Programming Interface

DCC: Digital Curation Centre

DILL: Digital Library Learning

FRBR: Functional Requirements for Bibliographic Records

IT: Information Technology

JISC: Joint Information System Committee

LAM: Libraries, Archives, Museums

LIS: Library and Information Science

NSF: National Science Foundation

OAI-PMH: Open Archive Initiative Protocol for Metadata Harvesting

OAIS: Open Archival Information System

OWL: Ontology Web Language

RDF: Resource Description Framework

RDFS: Resource Description Framework Schema

XML: Extensible Markup Language

Chapter 1: Introduction

1.1 Background

Recent years have seen an impressive rise in the amount of digital data in every sector of human life. Today, data is everywhere and businesses, industries, governments, universities, scientists, consumers and nonprofit organizations are generating data at unprecedented levels and at an incredible pace. (Gordon-Murnane, 2012). This global phenomenon has been defined in 2003 as *Data Deluge* by Tony Hey and Anne Trefethen. In their article, there is an attempt to preview “the imminent flood of scientific data expected from the next generation of experiments, simulations, sensors and satellites” (Hey and Trefethen, 2003, p.1).

As someone else has noted (Gray et al, 2002; Beagrie, 2006) in some subjects databases are supplementing or partly replacing journal publications as a medium of scholarly communication. In the academic context, research data are a vital aspect of the research activity that is growing very fast in terms of dimension, complexity and virtual space. The fact that computers and telecommunications have revolutionized the methods for collecting, storing, analyzing and disseminating information is now taken for granted (Heidorn, 2011). However, it is still important to stress the attention on the fact that nowadays science and society have shifted from data poor to data rich (Heidorn, 2011; National Science Foundation, 2007; Interagency Working Group on Digital Data, 2009). The IT revolution has radically changed the way in which research is done within the field of science and humanities. As Beagrie has noted (2006), in parallel of the growth of data some commentators have highlighted that the use and the value of data is also changing, and this is accordingly having a profound effect on how science is being conducted in some disciplines. “Science is now conducted in the digital realm, with scholars exchanging many terabytes of data among themselves” (Heidorn, 2011, p. 663) and academic, public and school libraries are logical repositories for society’s knowledge production (Heidorn, 2011).

However, research data are not the only kind of data produced nowadays. In parallel, government, institutions, media industries and other knowledge domains have started to free and share their data, often embracing the Open Data approach and using tools and instruments that allow to achieve this objective. The data deluge is affecting every aspect of society, and a 2010 article of *The Economist*,¹ by the same title, made this phenomenon explicit:

1 <http://www.economist.com/node/15579717> , retrieved at February 2015

Everywhere you look, the quantity of information in the world is soaring. According to one estimate, humankind created 150 exabytes (billion gigabytes) of data in 2005. This year, it will create 1,200 exabytes. Merely keeping up with this flood, and storing the bits that might be useful, is difficult enough. Analysing it, to spot patterns and extract useful information, is harder still. Even so, the data deluge is already starting to transform business, government, science and everyday life (...). It has great potential for good—as long as consumers, companies and governments make the right choices about when to restrict the flow of data, and when to encourage it.

In this perspective, the 2000s have witnessed the development of two important applications in the field of Information Technology. The first is the Semantic Web, that is “a web where the information can be read by computer as well as humans” (Stuart, 2010. p. 36) and where ontologies are designed in order to create specific knowledge domains. The second is the Linked Data movement, a way to use the web as a means to publish and connect data from different repositories. Semantic Web and Linked Data are encouraged by W3 Consortium and the latter is seen as the natural progression of the former.

This scenario is accordingly changing the domain of digital libraries. Research Data, Big Data, Open Data are indicators of the transformation in the field of Information Science, where we are turning from a web of documents into a web of data (Berners-Lee et al, 2001). This could be particularly true for librarians that are employed in research centers and universities, where the shift from libraries of documents to libraries of data could be disruptive. As a consequence of this huge amount of data, new professionals with both librarianship and computer science skills are needed to cope with increasing complexity.

On the other hand, one of the most powerful features of digital libraries is that they are not stand-alone applications, but are rather dynamic ecosystems that need to be integrated with other applications and services in order to develop and to survive. This “fluidity” of digital library services can potentially allow every institution that owns data or digital resources to build a technological system that is able to collect, organize and retrieve them. Of course, the complexity of the digital Library can vary highly from one context to another. This potential is enormous nowadays, and can be exploited in many contexts that could be seen far away from the LAM (Libraries, Archives, Museums) network. For all these reasons, digital libraries and digital library professionals can be employed in a great variety of working sectors.

In a *data deluge* perspective, the emerging role of the digital curator is becoming central in the data management activity. In a data-centric world (Lyon, 2012) digital curators are individuals “capable of managing digital objects and collections for long term access, preservation, sharing, integrity, authenticity and reuse. In addition, they have a range of managerial and operating skills, including domain or subject expertise and good IT skills.” More

specifically, the term *data curator* is used more and more to indicate the person in the organization responsible for all the activities connected with the management of research data.

This definition entails that, although the digital/data curator has a solid core of skills and competencies in the field of data management, he or she can be specialized in a great variety of job profiles. A recent study of Liz Lyon has made this concept explicit: the digital curator can develop his or her career in many directions, such as data librarian, data archivist, data analyst, data journalist etc. (Lyon, 2015). This makes clear one of the main characteristic of the digital curator, that is, without any doubt, interdisciplinarity: managerial, technical and collaborative competencies are essential to deal with this great amount of data.

In this digital context, the role of the librarian is evolving in directions that, up until few years ago, were unimaginable. The digital curator employed in the academic library can now be part of the research cycle and assist researchers and academics in the production of good research data. This new role for librarians has led academics to coin the term *embedded librarian* that is acquiring more and more importance in the field of digital librarianship.

However, like many other activities, digital curation needs a solid infrastructure that is able to support it. In research activities, but also in many other fields (Beagrie, 2006), the term *Cyberinfrastructure* has become the best way to define it. In very simple terms, “a cyberinfrastructure is a broad collection of computing systems, software, data acquisition and storage systems, and visualization environments, all generally linked by high-speed networks, often supported by expert professionals” (Tammara and Casarosa, 2015, p. 2). The healthier a national cyberinfrastructure, the more it can help the work of the digital curator and participate in building a solid global cyberinfrastructure.

1.2 Motivation

There are several reasons why the researcher has chosen the following topic of research:

- The first reason concerns the motives that led the researcher to choose the DILL course. When he got in contact with the world of digital libraries, he did not have a LIS background. Although he has a background in liberal arts, his professional training stems from the world of journalism. However, the researcher has found out that the potential of the digital library services is enormous and heterogeneous, and that it can be exploited in many, different ways. This has led him to investigate how digital libraries can go beyond the classical world of physical libraries, archives and museums, to be inserted in many domains of the Knowledge Society, such as media industry. In this perspective, the skills

and competencies of digital curation are a privileged starting point.

- The second reason comes from the interdisciplinary characteristic of digital curation activity. As said, technical, managerial and collaborative skills and competencies are essential for the work of a digital/data curator. On the other hand, the researcher believes that it is equally essential to isolate and analyze these competencies individually, in order to achieve a clear vision of the characteristics that a good data curator needs to do his or her job.
- The third reason is that this research is replicable. The model that follows could be used to extend the research to other skills and competencies in global, national or local contexts.
- The last reason concerns Italy, the country where the researcher comes from. The Italian cultural heritage is globally well known and it does not need to be mentioned here. Despite this enormous abundance, Italy has serious structural problems in the field of policies, cyberinfrastructure, data management and employment. Due to the lack of economical support, libraries and museums - despite being numerous - cannot satisfy the high demand of job positions. In this way, people with a strong humanities background and solid IT skills must turn to other working sectors. Digital and data curation can be seen accordingly as an important asset for the development of new kinds of professions.

1.3 Research question, aims and objectives

Based on these evaluations, the researcher has started to formulate a research hypothesis that has played a pivotal role in outlining the structure of the following research.

1.3.1 Statement of the problem

The main hypothesis of the research comes from a reflection on the creation of new roles in the practice of data curation. The researcher assumes that data curation can be the basis for the development of different professions, such as data librarians, data journalists, data analysts etc. The researcher also assumes that these professions need a common core of technical and IT skills that can help them find a specific specialization. Lastly, within a specific theoretical framework that concerns the data curation lifecycle, the researcher believes that these new roles in data

management allow digital librarians to add value to the whole data lifecycle, since he or she can contribute in finding different patterns of knowledge within data and datasets. Accordingly, the term *curation* assumes a meaning that goes beyond the term *preservation*: thus, digital curators can be placed not only in a *back-end* environment, but also in a *front-end* one. In addition, a specific context has been targeted, and the natural choice of the researcher is Italy.

1.3.2 Research Question

On the basis of this research hypothesis, there are various research questions that can be formulated. On the other hand, the researcher has chosen to isolate some factors and to apply them in a specific and verifiable context. Accordingly, the research question comes as follows:

- Can the IT skills and competencies of the digital/data curator in the Italian context be the basis for the development of new kinds of professions?

1.3.3 Research Aims and Objectives

Based on the research question stated above, research aims and objectives are outlined as follows:

Aims:

- To understand if this new scenario in the context of digital curation can support the development of new competencies and new opportunities for digital librarians.

Objectives:

- To analyse the perception of the practice of digital/data curation within the Italian digital communities, as well as other related concepts thank to which it could be developed.
- To identify and define basic IT skills and competencies that allow digital/data curators to perform their job, and to establish whether or not they can be used by many professions.

1.4 Methodology

The research is based on the epistemology of *constructivism*. This epistemology states that meaning is established in the interaction between the phenomenon and the observer. Accordingly, the meaning is constructed and not discovered. On the basis of this evaluation, the theoretical perspective is *interpretivism*, because only the interaction between the subject of the research and the world could lead to the explanation of the phenomenon. Moreover, the following research is qualitative, since it is deeply rooted in the nature of the research itself.

The researcher is aware of the fact that a quantitative approach could have been used likewise. On the other hand, the qualitative method will help the researcher to have a more holistic background of information about the digital curation community in Italy. Furthermore, qualitative data can help the researcher to go more in-depth in the investigation, and to have a clearer explanation of why some skill and competencies are so important in digital curation than a set of quantitative data would do.

In this perspective, research method is *collective and instrumental case studies*. The reason for this choice lies in the fact that a holistic vision of the digital curation community in Italy needs to gather data from different contexts and different profiles. Due to the interdisciplinary nature of the digital curation activity, data have been collected from different fields.

1.4.1 Data Collection Technique

On this basis, *five semi-structured interviews* have been conducted during this research. This technique has allowed the researcher to conduct a more thorough analysis than a simple questionnaire, and to understand the framework of information and experiences of the Italian digital curation community. Additionally – thanks to the interviews – the researcher has been able to fully comprehend the interviewees' opinions, with the aim of obtaining information and relevant suggestions from professionals in the field.

All interviews have been conducted online and individually and have been recorded; notes have been taken during the surveys. In order to obtain qualitative response from interviewees, some relevant questions about digital curation, IT skills and competencies, Cyberinfrastructure and Open Data philosophy have been asked.

The main purpose is to understand if there are any relevant similarities and differences among interviewees and, as a consequence, to try to identify the reasons of these similarities and differences. Another important purpose is to define a general opinion of the sample, in order to draw the attention on this opinion in the presentation of the research. All these qualitative data

sets have been gathered and analysed for the purpose of this research. A process of individual revision followed the interviews, in order to give the interviewees the opportunity to confirm what they had said during the interview.

1.4.2 Data Analysis

As stated above, interviews have been recorded, transcribed and coded. In addition, this information has been integrated with notes that were taken during the interviews. All this material was submitted to a *constant comparative analysis*. This strategy is based on a bottom up approach and involves the examination and the comparison of data with all the similar or different data gathered during the research (Pickard, 2007).

1.5 Limitations of the research

In conducting this research, some problems or difficulties have inevitably arisen. The researcher has witnessed some limitations based on lack of time, of economic and logistic issues. At the same time, some possible methodological flaws and biases can be identified (Pickard, 2007).

1.5.1 Lack of time, economic and logistic limitations

Firstly, time has been the biggest limitation. The research has been conducted in four months, during which the researcher had to gather the relevant literature review, to make and transcribe interviews, to design and write the research. In addition, although some suggestions about other useful interviewees had been given during the interviews, the researcher did not have the time to conduct further interviews.

Secondly, due to the lack of economic resources for this project, the researcher has chosen to conduct online interviews. Despite technology having great potential nowadays, it cannot replace some relevant factors such as direct observation and interaction with the participants.

Furthermore, interviews have presented some logistic constraints. Firstly, distance made it difficult to arrange interviews, which have all been conducted via *Skype*. Due to the lack of time on the part of some participants, it has not been simple to arrange the majority of the interviews. This has affected the way in which interviewees have faced some questions and dealt with some issues, even during the revision process.

1.5.2 Methodological flaws and possible bias

As stated, the researcher has received some useful suggestions during the interviewing process. Some of these suggestions could have led him to find other relevant participants, who could have given different perspectives and useful integration in the research process. However, due to time and economic limitations, it has not been possible to conduct further interviews. This lack of consideration of the whole *spectrum* of the digital curation community in Italy inevitably brings to possible methodological flaws in the design of the research.

Moreover, a relevant bias could come from the object of the research itself. Digital curation is conceptually a recent topic of interest, and some concepts are still not clearly defined. The application of a specific conceptual framework in the research topic brings inevitably some preconceptions, especially in the making of the questions of the semi-structured interview.

Last, it is worth remembering that this is the first research project of the researcher, which could surely result in an evident lack of experience, especially in the adoption of a specific methodology and a specific data collection technique.

1.6 Ethical considerations

Due to preservation of the rights of privacy, confidentiality and autonomy of the interviewees to participate to the following research, all the interviews are anonymous. Additionally, a letter that states the purpose of the project, dealing with confidentiality issues, has been delivered to the participants at the beginning of the research phase.

An informed consent has also been delivered and signed by the participants, in order to make it clear that data have be utilized for the purpose of the following research solely (Pickard, 2007).

Chapter 2: Literature review

The following chapter is divided in five main parts. The first part analyses the major trends in the academic fields, trying to find a suitable definition of what digital curation exactly is. In the second part the DCC Curation Lifecycle Model is analysed in depth, since it will be used as the theoretical framework through which the following research will be structured. The third part is devoted to examine different approaches in the academic context that have tried to outline a curriculum for the digital curation. In the fourth section, new roles that digital curators can potentially play in different contexts will be analysed, giving an accurate definition of some concepts that are essential in this field. In the last part, the relevant literature concerning digital curation in Italy is analysed.

As far as the development of the following literature review, the researcher has selected peer reviewed articles and books from EBSCO Host databases, Public Knowledge Project platform and Google Scholar. Moreover, some articles, presentations and conference proceedings have been retrieved from the International Journal of Digital Curation (www.ijdc.net), D-Lib, the National Science Foundation and the Digital Curation Centre.

Due to the nature of the research, the language was restricted to English and Italian. Related terms such as *digital curation*, *data curation*, *digital and data curator*, *data lifecycle model*, *cyberinfrastructure*, *embedded librarian* and *digital curation curriculum* have been explored in depth, often combining these terms in different ways. Given that the context of the following research is the Italian landscape, the keywords mentioned above have additionally been translated in Italian, in order to retrieve the relevant literature review. Nevertheless, it has been found that the small literature review in Italian in this field uses mostly the English definitions, especially for *digital curation*. Moreover, the researcher has noted that the concept of *e-science* is preferred than *cyberinfrastructure* in the Italian (and sometimes in the European) academic research.

2.1 An overview of digital curation

In their 2003 article, Hey and Trefethen states that “generating data is one thing, preserving it in a form so that it can be used by scientist other than the creators is entirely another issue. This is the process of *curation*” (2003, p.13). Although this is a very clear concept, trying to define what digital curation is not equally simple. The concept of *data deluge* (Hey and Trefethen, 2003; Lord et al, 2004; Heidorn, 2011; Borgman, 2008) in academic research has emphasized

how urgent is the need for the management of the great amount of data that are produced nowadays. This makes the concept of digital curation a very trend topic in today's digital library studies.

Beagrie (2006) has focused the attention on the fact that the term digital curation is closely related to other terms, such as *digital preservation* and *digital archiving*, and that all these terms could be perceived in different ways from different individuals and disciplines. This leads sometimes to interpret the concepts of *data*, *preservation* and *archiving* in ways that are locally embedded and are difficult to eradicate (Beagrie, 2006).

Another reason for the complication in finding a clear definition of digital curation lies in the fact that this field of study is quite a brand new subject, which is still evolving and has just recognised officially (Beagrie, 2006; Higgins, 2008; Digital Curation Center; Borgman, 2015; Davenport, 2012; Lesk, 2011; Tammara and Casarosa, 2015). Authors from several academic fields - Library and Information Science, Computer Science, Data and Information Management, Digital Humanities - are discussing the development of the concept of digital curation (Beagrie, 2006 and 2008; Yakel, 2007 and 2011; Williams, 2009; Ray, 2002, 2009; Harvey, 2010; Cunningham, 2008; McDonald and Lord, 2003), focusing often the attention on the interdisciplinary nature of this activity (Kim, Warga & Moen, 2012; Gold, 2010).

Despite these difficulties, digital curation has been appropriately defined as “a truly 21st century term” (Choudhury et al, 2012), even for temporal reasons: the term was indeed used for the first time in October 2001, during a seminar in London named "*Digital Curation: digital archives, libraries and e-science*”, sponsored by the Digital Preservation Coalition and the British National Space Centre.

A useful introduction to the subject of digital curation is given by Elizabeth Yakel in her 2007 article with the same name. In her article Yakel emphasizes how, despite many librarians and information professionals have done for years many of the activities that are now ascribed to the digital curators (a reflection that is also in Ray, 2012), the emergence of the term has been prompted by some reports on the digital information landscape. Yakel also focuses the attention on four reports that have played a pivotal role in this field: two different reports produced by the National Science Foundation (NSF) in 2003 and 2007, the report produced by the American Council on Learned Societies (ACLS) in 2006 and the report produced by Liz Lyon in 2007 for the United Kingdom Office for Library and Information Networking (UKOLN). Surprisingly, despite their importance in the field, none of these reports give a clear definition of Digital Curation (Yakel, 2007).

Besides these important reports, Ray (2012) highlights a fifth document that could be added for its importance in the definition of the future digital management strategies: the 2009 report

by the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council, called *Harnessing the Power of Digital Data for Science and Society*, includes the recommendation that US federal agencies gives to “promote a data management planning process for projects that generate preservation data” (Interagency, 2009)

In this context, it is noticeable that *Science* represents the field in which the concept of digital and data curation started to arise. In the US the National Science Foundation represents the leading institution for the support of the curation of research data, and the two reports mentioned above (NSF, 2003 and 2007) have played a seminal role in the diffusion and the development of this approach. According to Gold, “in the US the NSF was the epicenter of influence. In the wake of Atkins report and other major policy documents, the NSF began to strongly encourage libraries to play a role in data curation.” (2010, p.5) Likewise, it is possible to see in these years the increasing number of “international research, conferences, and policy developments, notably in the UK and Australia, also exerted influences on the developing US research library perspective on digital data curation” (Gold, 2010, p.5). In this perspective, Choudhury defines data curation as “a means to collect, organize, validate, and preserve data so that scientists can find new ways to address the grand research challenges that face society.” (2010, p.195) Beagrie, again, highlights how this term is “used primarily by the scientific and digital library communities respectively.” (2006, p.4).

On the other hand, Beagrie also points out that “new forms of data publishing pose many challenges both technical and organizational. It is worth stressing that these changes and challenges in data publishing are not solely confined to research data. Similar trends can be seen in traditional publishing and in the Web as electronic publication increasingly evolve dynamic, on the fly generation rather than static fixed versions of context.”(2006, p.9) This concept has been furtherly analysed by Walter and Skinner (2011) when they try to make clear the distinction among digital curation, data curation and digital preservation, stating that these terms often correspond to different disciplinary contexts: “data curation is applied most often in science, engineering and social science fields; digital curation is used most frequently to describe digital humanities and arts environments; and digital preservation usually appears in library activities (Walter and Skinner, 2011, p.16).

An accurate and precise definition of the concept comes from the Digital Curation Centre,² probably the most important and useful institution that deals with digital curation nowadays. The aim of the DCC is “to provide expert advice and practical help to anyone in UK higher education and research wanting to store, manage, protect and share digital research data.”³ Founded in 2004

2 www.dcc.ac.uk

3 <http://www.dcc.ac.uk/about-us/about-site>, accessed February 2015

with the support of the Joint Information System Committee⁴ (JISC), this centre represents the first initiative of this kind in the world, whose aim is to be a centre of excellence in the area for UK and non-UK operators (Lord et al, 2004).

The DCC website gives a concise definition of digital curation, as something that “involves maintaining, preserving and adding value to digital research data throughout its lifecycle.”⁵ This short definition has been accepted universally and it is now used as the starting point from every research in the field. Going into details, it is important to highlight how the second part of this definition – *adding value to digital research data throughout its lifecycle* – is essential to understand the theoretical framework that lies behind a broader concept of digital curation and, accordingly, behind the structure of the following research.

2.2 Data Lifecycle and added value

As said, the concept of *data lifecycle* is essential to understand the activity of digital and data curation. Pennock asserts that digital curation is “the active management of digital information over its entire lifecycle for both current and future use.” (2007, p.1) Similarly, Shreeves and Cragin (2008) defines data curation as “the active and ongoing management of data through its lifecycle of interest and usefulness to scholarship, science, and education, which includes appraisal and selection, representation and organization of these data for access and use over the time.” (p.93)

The importance of a data lifecycle approach for the maintenance of digital resources has been long recognised by the JISC and other national institutions worldwide. This approach highlights how different stakeholders are involved with data resources at different stages and, at the same time, how the relationships between these stakeholders is vital for their maintenance and research values (Beagrie & Greenstein, 1998; Beagrie, 2004 and 2006).

For these reasons, the DCC has proposed a Curation Lifecycle Model, that can help stakeholders to visualize every stage – necessary or not – required for successful curation (Higgins, 2008). Using the words of Higgins, “a lifecycle approach ensures that all the required stages are identified and planned, and necessary actions implemented, in the correct sequence. This can ensure the maintenance of authenticity, reliability, integrity and usability of digital material, which in turn ensures maximization of the investment in their creation.” (2008, p.135)

The model that has been hypothesized by the DCC “provides a graphical, high level overview of the stages required for successful curation and preservation of data from initial

4 <http://www.jisc.ac.uk/>

5 <http://www.dcc.ac.uk/digital-curation/what-digital-curation>, accessed February 2015

conceptualisation or receipt through the iterative curation cycle.”⁶ Graphically, the curation lifecycle model appears in the following way:

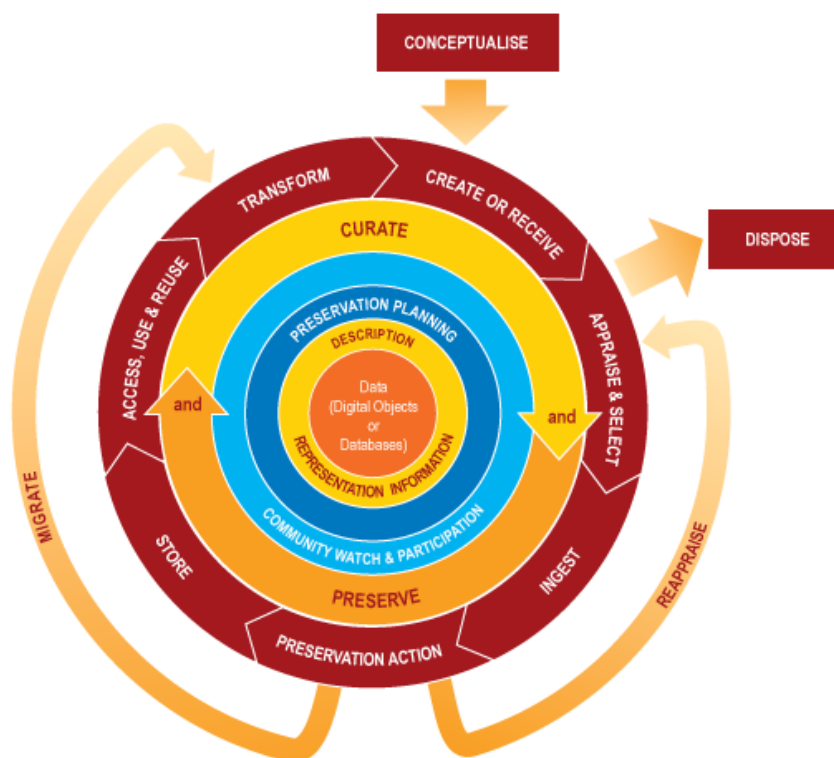


Figure 1: the graphical representation of the DDC Lifecycle model

At the core of the model there is the notion of *data*,⁷ which the DCC defines as “any information in binary digital form, is at the centre of the Curation Lifecycle.” A more exhaustive definition in this perspective is given by Reference Model for an Open Archival Information System (OAIS, also cited in Borgman, 2008, p.30), that is “a reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing. Examples of data include a sequence of bits, a table for numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen.”

Data can be simple digital objects (discrete digital items such as text files, image files or sound files, along with their related identifiers and metadata) or complex digital objects (discrete digital objects made by combining a number of other digital objects, such as websites); but data can also be databases, or, in other words, structured collections of records or data stored in a computer system.

It is also important to distinguish between full-cycle curation activities – that are activities

6 <http://www.dcc.ac.uk/resources/curation-lifecycle-model>, accessed February 2015

7 If not specified in another way, the following content has been taken from the DCC website <http://www.dcc.ac.uk/digital-curation/what-digital-curation>, accessed February 2015.

maintained in all steps of the cycle – and other sequential actions, which can or cannot be part of a specific curation activity. There are also other occasional actions that could be implemented in within the lifecycle model.

Full-cycle actions include:

- *Description and Representation Information*: assign administrative, descriptive, technical, structural and preservation metadata, using appropriate standards, to ensure adequate description and control over the long-term. Collect and assign representation information required to understand and render both the digital material and the associated metadata. The concept of Representation Information is used by the OAIS and analyzed in Patel and Ball (2008).
- *Preservation Planning*: plan for preservation throughout the curation lifecycle of digital material. This would include plans for management and administration of all curation lifecycle actions.
- *Community Watch and Participation*: maintain a watch on appropriate community activities, and participate in the development of shared standards, tools and suitable software. This activity can occur in multiple levels (Heidorn, 2011), since each class of data objects will have its own community of primary users.
- *Curate and Preserve*: be aware of, and undertake management and administrative actions planned to promote curation and preservation throughout the curation lifecycle.

Sequential actions include:

- *Conceptualise*: conceive and plan the creation of data, including capture method and storage options.
- *Create or Receive*: create data including administrative, descriptive, structural and technical metadata. Preservation metadata may also be added at the time of creation. Receive data, in accordance with documented collecting policies, from data creators, other archives, repositories or data centres, and if required assign appropriate metadata.
- *Appraise and Select*: evaluate data and select for long-term curation and preservation. Adhere to documented guidance, policies or legal requirements.
- *Ingest*: transfer data to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements.
- *Preservation Action*: undertake actions to ensure long-term preservation and retention of

the authoritative nature of data. Preservation actions should ensure that data remains authentic, reliable and usable while maintaining its integrity. Actions include data cleaning, validation, assigning preservation metadata, assigning representation information and ensuring acceptable data structures or file formats.

- *Store*: store the data in a secure manner adhering to relevant standards.
- *Access, Use and Reuse*: ensure that data is accessible to both designated users and re-users, on a day-to-day basis. This may be in the form of publicly available published information. Robust access controls and authentication procedures may be applicable.
- *Transform*: create new data from the original, for example by migration into a different format, or by creating a subset, by selection or query, to create newly derived results, perhaps for publication

In the end, occasional actions include:

- *Dispose*: dispose of data, which has not been selected for long-term curation and preservation in accordance with documented policies, guidance or legal requirements. Typically data may be transferred to another archive, repository, data centre or other custodian. In some instances data is destroyed. The data's nature may, for legal reasons, necessitate secure destruction.
- *Reappraise*: return data which fails validation procedures for further appraisal and re-selection.
- *Migrate*: migrate data to a different format. This may be done to accord with the storage environment or to ensure the data's immunity from hardware or software obsolescence.

The model is not definitive, but will undoubtedly evolve. According to Higgins “the next stage of the project is the development of domain-specific variations to help further to contextualise training and resources, while providing more tailored advice. (Higgins, 2008, p.136).

In addition, there is a discrete corpus of literature that has started to discuss the current developments that concerns the DCC Lifecycle Model. Many authors have stressed the importance on the fact that the term *curation* involves the production of some added value that the term *preservation* does not. As Beagrie writes:

“The new term benefited from some existing usage of the term “curation” by both the library and museum sectors, and the biological sciences. In all three sectors the

term implies not only the preservation and maintenance of a collection or database but some degree of added value and knowledge.”

And again:

“In the biological science, the term curation had been applied to the maintenance and publishing of databases such as the human genome and was therefore already implicitly digital. In this context added value is derived from annotation, linkage, and the management, validation, and editorial input of domain specialists employed to curate and publish the database.” (Beagrie, 2006, p.5)

In this perspective, the publishing, the linkage and the visualization of data become relevant issues within the curation activity. The same concept has been highlighted by Rusbridge and others, when they claim that:

“Curation embraces and goes beyond that of enhanced present-day re-use, and of archival responsibility, to embrace stewardship that adds value through the provision of context and linkage (...) in way that ease re-use and promoting accountability and integration.” (Rusbridge et al, 2005, p.2)

This capacity of creating added value is intrinsic to the whole data lifecycle. In this context, not only preservation is just one aspect of the more complex system of data curation (Gold, 2007a; 2007b), but curation also affects the way in which data are used and published. Accordingly, the digital curation activity can include archivists for the preservation or librarians for adding metadata schemas, but also authors, publishers and professionals for the creation of new relationships among different datasets, or professionals who deal with the visualization of data. Digital curation becomes a holistic activity that involves different stakeholders. This concept has been emphasized by Gray et al. (2012) when they state:

“Unwillingly, and sometimes unknowingly, projects become not only Authors, but also Publishers and Curators. The Consumer interact with the projects directly. Scientists are familiar with how to be an Author, but they are just starting to learn, out of necessity, how to become Publisher and Curator. This involves building large on-line databases and designing user interfaces. These new roles are turning out to be demanding and require new skills.” (Gray et al, 2012)

In the same perspective, Choudhury et al (2009) discuss the way in which data management is revolutionizing the role of librarians and the publishers in academic research:

“So we have, on the one hand, a community, or a subset of several communities, that has been working on the “back end” of digital production from the generation of raw data to the construction of an organized product that can be accessed, and, on the other hand, another community – publishers – who work on the “front end” of scholarly communication, from manuscript to publication. This paper discusses some possibilities for bringing these communities together and demonstrate the role that libraries are playing in making this connection.” (Choudhury et al, 2009, p.477)

The way in which data are manipulated and published is becoming more and more relevant in this context, making libraries the *technical glue* to bring together the various components (Choudhury et al, 2009; Borgman, 2008). A similar concept can be found in Gold (2007b), using the words *upstream* and *downstream* rather than *front-end* and *back-end*:

“[t]he opportunities afforded by digital data include more dynamic and fluid linkages throughout a “stream” or life cycle, including the possibility of reusing data. Such reuse is arguably both downstream and upstream at the same time! Still the demarcation may be a useful one to consider, particularly since established library roles and capabilities tend to fall in the “downstream” direction and new libraries roles may deal more with the “upstream” part of the research cycle.” (Gold, 2007b, p.4).

As many authors have pointed out, these new roles for libraries and librarians call for new skills and competencies, especially from technical areas. The next paragraph will analyze some efforts that have been made in order to define these skills.

2.3 Outlining a technical *curriculum* for digital curators

According to Heidorn, “curation of the data is within the libraries’ mission, and libraries are among the only institutions with the capacity to curate many data types.” (2011, p. 663) In addition, it is noteworthy how the emerging relationship between library-run institutional repositories and national or global networks of data repositories suggests that the practice of digital data curation will make rapid progress - as a library practice - over the next decade (Gold, 2010; Baker and Yarmey, 2009; Witt, 2008). This raises some questions about the technical skills and competencies that are necessary for the practice of digital curation. Witt (2008) pointed out that librarians have many skills to bring to conduct research data curation: librarians have expertise in classification and description of information through metadata services such as cataloging, as well as reference and instruction assist in finding and using information effectively. Collection management selects, deselected and presents information in the appropriate

context (Witt, 2008).

Many authors have tried to answer to these question, trying to outline a curriculum for the digital curation activity and, in parallel, to design a specific educational background (Lee et al, 2007; Tibbo et al, 2008; Cragin et al, 2009; Addom and Stanton, 2011; Witt, 2009). Some remarkable efforts in the field come from the US, where some research libraries (Cornell, Purdue, the MIT and the University of Minnesota) “have taken the initiative to convene campus-wide e-Science initiatives or centers, or initiate data curation partnerships with domain researchers, computer scientists and campus IT.” (Gold, 2010, p.15) In 2006, the Graduate School of Library and Information Science at the University of Illinois, Champaign-Urbana obtained a grant to develop a Data Curation Education Program. In the same year, the University of North Carolina launched the first phase of DigCCurr,⁸ in order to “develop an openly accessible, graduate-level curricular framework, course modules, and experiential and enrichment components and exemplars necessary to prepare student to work in the 21st century environment of trusted digital and data repositories”. The focus of DigCCurr goes beyond scientific data and includes cultural artifacts and records, cultural heritage assets and teaching materials. A second phase, DigCCurr II, is developing “an international, doctoral-level curriculum and educational network in the management and preservation of digital material across their lifecycle.”

Other remarkable efforts come from Europe. Besides the already mentioned DCC based in UK, the most well-known and omni comprehensive project is undoubtedly the Digital Curator Vocational Education Europe (DigCurV),⁹ a project funded by the European Commission’s Leonardo da Vinci program to establish a curriculum framework for vocational training in digital curation. Besides many activities that include training opportunities, conferences and resources, the DigCurV offers a *Curriculum Framework*, which is “a means to identify, evaluate, and plan training to meet the skill requirements of staff engaged in digital curation, both now and in the future.”¹⁰ For the following research, it is relevant that the Curriculum Framework highlights some *skill identifiers*, in which a prominent place has been given to the *Knowledge and Intellectual Abilities* (KIA), which include some technical skills such as *Subject Knowledge* (KIA1), *Selection/Appraisal* (KIA2), *Information Skills* (KIA4) and *Data Skills* (KIA5).

Other remarkable projects in the field are the Digital Curation Unit based in Greece,¹¹ the Nestor project,¹² a transnational partnership of academic institutions in Germany, Switzerland

8 <http://www.ils.unc.edu/digccurr/>

9 <http://www.digcur-education.org/eng/>

10 <http://www.digcurv.gla.ac.uk/>, accessed February 2015.

11 <http://www.dcu.gr/>

12 http://www.langzeitarchivierung.de/Subsites/nestor/EN/Home/home_node.html

and Austria, and the Netherland Coalition for Digital Preservation.¹³

Nevertheless, finding a specific research on technical and IT skills and competencies for digital and data curation is not simple, since they are just one part of a bigger and more complex system. As many academics and not-academics have stated, “data curation development is not merely a technical/engineering issue, because data curation involves many different stakeholders, and data curators need to be aware of the organisational context and socio political issues.” (Tanmaro et al, 2014, p.3).

A study conducted by Kim, Warga and Moen (2012) has analysed 110 job advertisements in order to have access to some indicators of the competencies required by employers in the field of digital curation. The first result was that the position title has many variations, in which is possible to find terms such as librarian, data, digital and archivist. More interestingly, the ability to work in an information technology intensive environment emerged, that included knowledge of multiple operating systems and web architectures, programming and scripting, web development skills, the ability to work with data using relational databases, data analysis tools and some specifications like SQL, XML and RDF.

2.4 New roles and new contexts for data experts

Referring to research digital curation, Tammaro et al (2014) try to focus their attention on the competencies gap and on the role of information professionals in the research lifecycle. They also emphasize how more and more professionals within the LAM network seem to converge into an *information professional* (p.2), asking for a common core of knowledge/skills that can be taught. This raises also several questions about the identity of this figure: “should there be a data librarian, a data archivist or a data museum curator?” (Borgman 2010, 2012, cited in Tammaro et al, 2014, p.3). Tammaro et al also highlight how the role of “Information Professionals should evolve from one of holders and providers of knowledge resources to one of being an active partner in the research process.” (p. 4).

Beyond the LAM network, the specialization of curricula in the practice of data curation is an interesting subject of study. For the purposes of this research, the recent works of Liz Lyon are essential to understand how the curriculum of the data curator can specialize in many job profiles. Lyon also cites the analysis of emerging job trends proposed by Larsen et al. (2014), that highlights the growth in new positions which include some elements of *data* (Lyon, 2015). This study has highlighted the presence of “completely new designations and reflect the recognition that there is a requirement to build capacity and capability in a wide range of data

13 <http://www.ncdd.nl/>

areas.” (p. 113)

Referring first of all to the subject of *data science* in a broad sense, Lyon and Takeda (2012) emphasize how data scientists can follow varied career routes nowadays. In detail, in her 2015 article, Lyon sum up her concepts by means of a table of data scientists roles, that includes jobs such as data analyst, data archivist, data engineer, data journalist, data librarian and data steward (although this is not a complete list of possible roles). Every role has its peculiarities in term of tasks, locations and skills; however a core of common skills and competencies can be found in every profile. Lyon (2013) also highlights a shortage of data scientists at the present time: citing the McKinsey Global Institute’s report on Big Data, she reports that the prediction is a shortage of 190.000 data scientists by the 2019.

On the other hand, the term *digital curation* calls for other related keywords that are useful to fully comprehend this practice. As the DCC Lifecycle Model has shown, digital curation is indeed a complex activity, which requires specific skills and competencies, a well-disposed context in which it is possible to work and a solid infrastructure that allows digital/data curator to perform their job.

According to Anne Gold:

“One of the challenges of talking about “data curation” is that activities of curation are highly interconnected within a system of systems, including institutional, national, scientific, cultural and social practices as well as economic and technological systems. Data curation is a nascent set of technologies and practices emerging in the context of this complex and rapidly evolving socio-technical ecosystem.” (Gold, 2010, p.3)

This complex *socio-technical ecosystem* is useful to contextualize the work of the data curator. Beagrie (2006) observes this issue in a clear way:

“A substantial part of the cost-base of repositories consists of skilled staff and these human resources and many existing workflows and practices will not scale appropriately. There will be a need for more automation of processes and metadata generation, software tools for this, and potentially the development of greater collaboration and shared services to lower the entry and operational costs for institutions.” (p.8)

The manifestation of this need is the development of a global *cyberinfrastructure*, a concept that is mostly endorsed in the US by the National Science Foundation and refers to an infrastructure based upon distributed computers, information and communication technologies (2003 and 2007). Tammaro and Casarosa defines cyberinfrastructure as “a broad collection of

computing systems, software, data acquisition and storage systems, and visualization environments, all generally linked by high-speed networks, often supported by expert professionals” (2015, p.2) The value of the cyberinfrastructure in relation with research digital libraries and data curation has been underlined by Tyler Walter and Katherine Skinner (2011). They focus the attention on the importance of the trio of strong infrastructure, content and services: “infrastructure includes facilities, technologies and the human expertise applied to the organization” (Walter and Skinner, 2011, p.6). Content refers to the information that libraries make accessible and services include those in a traditional and virtual environment, such as information production, access and dissemination, long-term curation and preservation. In addition, Walter and Skinner individuate some roles and responsibilities that are recently emerging in the digital libraries environment, and are helping to redefine the library landscape of the twenty-first century. For the purpose of this research, the vision of *librarians as content producers and disseminators*, which recalls the idea of added value in the data lifecycle is noteworthy.

Within this context, identifying and articulating the specific competencies, skills and abilities required to perform a series of digital curation functions is an excellent basis for developing and tailoring training and educational programs to meet specific objectives. A useful overview in this direction is given by Anne Gold:

“Becoming literate in cyberinfrastructure means understanding cyberinfrastructure, E-science, collaboratories, collaboration science, computational and grid science, data curation, the Semantic Web, open data, open archiving, digital preservation, and data management, and how they relate to each other.” (Gold, 2007b, p.1)

And later on the same article, referring to how to manipulate data:

“Linking data in a rich and robust ways to support data reuse and integration will require understanding and documentation of the data provenance, the development of ontologies, expert annotations, and analysis. Further downstream, services enabled by this activities will include visualization, simulation, data mining and modeling, and other forms of knowledge representation and extraction.” (Gold, 2007b, p.3).

Another issue for the development of the cyberinfrastructure, and therefore, for a high level of data curation, is the *collaborative attitude* among universities, institutes and governments for the exchange of data and datasets. Reports such as the one by the National Science and Technology Council (2009), for instance, suggest the need for collaboration among organizations, entities, and individuals to carry out data management and data curation responsibilities. This attitude is

realized by means of open data approaches, which means “data that can be freely used, re-used and redistributed by anyone”. (Open Data Handbook, 2012, p.6). Walter and Skinner highlight how “the emerging digital humanities community emphasizes the use of open standards and open source solutions and has, from its inception, entered in strong partnerships with research libraries. (2011, p. 69). Furthermore, a document provided by JISC and DCC stresses the importance of the value of open source for digital curation: “In considering the value of open source for digital curation it is convenient and worthwhile to consider its merits through every stage of the data life-cycle model. Encompassing creation, active use, archiving, preservation, access and re-use, and disposal or transfer one can identify three main user roles. These are of data creator, data curator and finally data re-user.” (McHugh, 2005, p.17). Thus, open source is another fundamental issue for the development of the practice of digital curation at a high level.

2.5 Digital curation in Italy

Due to the novelty of this field of study, finding a relevant literature review about the practice of digital curation in the Italian landscape has been difficult. Very few articles were found on this topic, and other few were closely related. The majority of studies in this field are focused essentially on research data management and open source, where the term *digital curation* rarely appears. Moreover, relevant articles in Italian language that deal with research data management tend to highlight the side of the preservation of data, paying less attention on the activity of curation. In this perspective, although the interest around the management of trusted digital repositories is becoming more and more important, the interest around the practice of digital curation cannot be mentioned.

Despite the work of few scholars (Tammaro, 2007, 2010, 2012; Tammaro et al 2014; Tammaro and Casarosa, 2015; Cassella, 2013; Testoni, 2013; Guercio, 2012) a theory for the development of data curation activity in Italy has not been outlined yet. Some scholars have also collaborated to the creation of the curriculum in the DigiCurV project, but focusing the attention on the European context. Some librarians (Cassella and Vivarelli, 2013) have highlighted the scarce attention to the issues of digital curation and digital preservation, focusing the attention on the necessity to raise awareness on digital curation both at national and institutional level.

Nevertheless, it is noteworthy the development of two different projects in the field. The first is *Magazzini Digitali*,¹⁴ launched in 2006 with the aim of preserving digital born resources and facilitating the access to trusted digital repositories. The project is linked to digital legal deposit

14 <http://www.bncf.firenze.sbn.it/pagina.php?id=212>, accessed April 2015.

project¹⁵ for doctoral thesis, and is coordinated by the National Library of Florence, supported by the *Fondazione Rinascimento Digitale*.¹⁶

A more recent project concerns a training course for digital curators at the University of Turin in October 2014. The aim of the course is to train information professionals at executive level, in order to support digitization projects related to university libraries. The course will end in May 2015, at the same time the following thesis will be ended. For this reason, it has not been possible to give any relevant feedback.

15 <http://www.depositolegale.it/>, accessed April 2015.

16 <http://www.rinascimento-digitale.it/>

Chapter 3: Research design

3.1 Conceptual framework

As stated in the first chapter, the main hypothesis of the following research is that data curation can be the foundation for the development of different and more specialized professions, such as data librarians, data journalists, data analysts etc. The researcher also assumes that these professions need a common core of technical and IT skills and competencies that can help them to find a specific specialization. Lastly, within the theoretical framework of the DCC Lifecycle Model analysed in the literature review, the researcher believes that these new roles in data management would allow digital librarians to add value to the whole data lifecycle, since he or she can contribute to finding different patterns of knowledge within data and datasets. The Italian context has been chosen in order to analyse a specific and controlled context.

Based on this hypothesis, a research question has been outlined as follows:

- Can the IT skills and competencies of the digital/data curator in the Italian context be the basis for the development of new kinds of professions?

Accordingly, research aims and objectives are the following:

Aims:

- To understand if this new scenario in the context of digital curation can support the development of new competencies and new opportunities for digital librarians.

Objectives:

- To analyse the perception of the practice of digital/data curation within the Italian digital communities, as well as other related concepts thank to which it could be developed.
- To identify and define basic IT skills and competencies that allow digital/data curators to perform their job, and to establish whether or not they can be used by many professions.

Bearing in mind the theoretical framework that has emerged from the literature review, the researcher has extracted some important concepts that are the grounds on which the research itself is based. Concepts arisen from literature review have been put in connection in order to find a specific network of relations that lies beneath them. This has led the researcher to structure a conceptual framework that has been graphically represented using an open source software named *Cmap*:

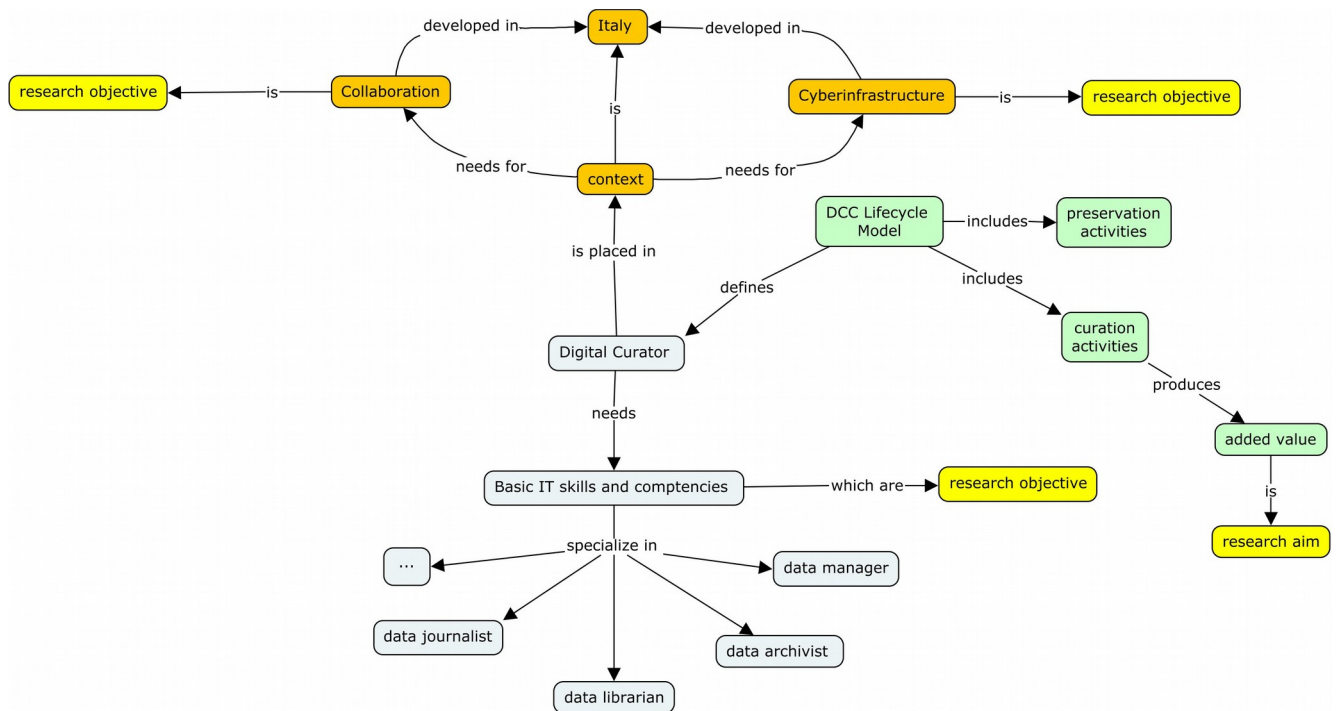


Figure 2: graphical representation of the conceptual framework

Bearing this conceptual framework in mind, and according to the research hierarchy model introduced by Lincoln and Guba (1985), this research will be structured as follows:

The research **paradigm** is *interpretivism*;

The research **methodology** is *qualitative*;

The research **method** is *collective and instrumental case studies*;

The research **techniques** are *interviews*;

The research **instruments** are *human beings and computers*.

3.2 Research paradigm

As said in the introduction, the following research is based on the epistemology of constructivism, which states that meaning is established in the interaction between the phenomenon and the observer. Following this epistemology, the paradigm of the research is *interpretivism*. According to this view, there is not a tangible and universal reality, but there are many and complex worlds given by the individuals (Pickard, 2007). In this perspective, reality is individual, it fits into the context and is opposed to the universal (Flick, 2002). Time and context play, therefore, a central role in this perspective, since they influence the way in which data are gathered and analysed. Accordingly, meaning relies on context, and the interpretation of the action or the opinion has to take into account the situation in which it is produced (Dey, 1993, cited in Pickard, 2007).

3.3 Research methodology

Following the paradigm of interpretivism, the methodology of research can only be *qualitative*, since it includes the alternation of the exchange among participants and the hermeneutic based on the knowledge of the researcher, whether it be tacit or explicit (Pickard, 2007). A qualitative approach connects the single participants of the research, the researcher – as instrumental to the research and the appropriate techniques for data collection. Accordingly, *human beings* are the research instrument that this approach privileges: on the one hand, participants are the source of valuable meaning; on the other the contribution and the overall understanding of the participants relies highly on the researcher itself. In this context, “human lives and their interpersonal relationships create complexities that need to be understood and the researcher acting as the research instrument allows for understanding and depicting these complexities” (Pickard, 2007, p.14). This leads to justify the nature of the qualitative approach, since if a person needs to be understood as a person and not as an object the relationship between the researcher and that person needs to be dynamic and mutual (Maykut and Morehouse, 1994).

Another useful consideration for the adoption of this approach comes from the *emergent design* of the research (Guba and Lincoln, 1994, Kumar, 1999), whose flexibility allows adapting appropriate techniques along the research process. The emergent design gives the participants the opportunity to interact with the researcher and to negotiate outcomes of the inquire. This adds additional insight into the case that has been investigated (Pickard, 2007).

3.3.1 Trustworthiness

In general terms, the trustworthiness of a qualitative research can be evaluated by means of different methods. In the following research trustworthiness is evaluated through the lenses of the model suggested by Pickard (2007), which is an adaptation of the model proposed by Lincoln and Guba (1985), which originally includes the four concepts of *truth value*, *applicability*, *consistency* and *neutrality*. Pickard adapts these concepts and propose the following criteria.

3.3.1.1 Credibility

In qualitative researches, *credibility* is proved by means of ever-long engagement with participants, triangulation of data collection techniques and member checking. As far as the engagement with the participants goes, the researcher has tried to interact with all the participants not only during the interview process, but also before and after the interviews through e-mails and phone calls. This helped the researcher to establish a common background of knowledge and to talk a language common to the field of research.

Due to the nature of the research itself, the triangulation of data collection has not been possible. The following research includes an in-depth analysis of the participants' thoughts and insights, and a relevant technique that could have been useful to the researcher to triangulate data has not been found.

On the other hand, great importance has been given to following the members. All the completed interviews have been submitted to each interviewee, in order to be checked, edited and integrated. This helped the researcher to clear some concepts, to gain further insight and to have a better understanding of the topic of the research.

3.3.1.2 Transferability

The aim of *transferability* is to apply the results of the research to another context. As said in the introduction, one of the motivations that has led the researcher to investigate this topic is that the following model can be replicated. The context of the research – Italy – could be seen as very specific, especially in the field of digital curation. Italy has some characteristics that other countries might not have, that come directly from its cultural and socio-technical context. On the other hand, the researcher believes that the homogeneity of the technical skills and competencies within the practice of digital curation could transfer the means of investigation of this topic to other contexts, at both national and local levels.

3.3.1.3 Dependability

During the process of the research, the researcher has been constantly in contact with the advisor, reporting on the status of the project every two weeks. The researcher and the advisor have frequently shared debates via e-mail, chat and video calls concerning the methodology and the best method to obtain samples and to collect data. A thesis seminar was arranged in May 2015, with the aim of informing the advisor, key informants and other students on the status of the project. In addition, during the thesis seminar the researcher had to deal with an opponent - another DILL student - who debated around theoretical framework, methodology, methods and data collection techniques.

3.3.1.4 Confirmability

The aim of *confirmability* is to validate the results and to establish whether or not data that have been collected have been used to analyse and produce the final results. This is essential to limit any bias that can be found in the research. Transcriptions of every interview (in Italian) is available upon request to whoever may be interested. The whole documentation has also been submitted to the advisor.

3.4 Research methods

Due to the specificity of the context that has been chosen for the following research, *collective* and *instrumental case studies* have been selected as research methods. As said in the introduction, the reason for this choice lies in the fact that a holistic vision of the digital curation community in Italy needs to gather data from different contexts and from different professional profiles. In addition, due to the interdisciplinary nature of the digital curation activity, data has been collected from different fields. In this perspective, the case study's definition given by Yin is still valid: "an empirical inquiry that investigates a contemporary phenomenon within its real-life context, where the boundaries between the phenomenon and context are not clearly evident, and in which multiple sources of evidence are used" (Yin, 2002, p. 23).

Moreover, the case studies presented here have been purposely selected to be *collective* and *instrumental*. The former has allowed the researcher to use more than one case to investigate the topic of this research. The latter has been chosen since instrumental case studies allow the researcher to have an in-depth investigation of a specific phenomenon or a hypothesis, making the case itself the means of the investigation (Pickard, 2007).

As far as the whole research process is concerned, the researcher has taken into account the background theory of the qualitative research based on case studies as an interactive process. In other words, “once in the field the researcher will allow the design of the study to develop as he gains into the salient issues” (Pickard, 2007, p. 87). Following the model proposed by Lincoln and Guba (1985) for naturalistic research and adapted by Pickard (2007):

- The *context* of the research is Italy;
- The *unit of analysis* are people that deal with digital curation. In designing the following project the researcher has been initially tempted to identify this group of people as a Community of Practice (Wenger and Snyder, 2000; Wenger, 2006). However, after serious considerations the researcher has been brought to state that there are not enough factors to define this group as a Community of Practice, since they are not naturally defined and they do not share the same language.

3.4.1 Target group and sampling

For the purpose of the research, the approach of *purposive* and *snowball sampling* has been selected. The former has initially helped the researcher to identify – with the support of the advisor and some key informants – some relevant participants among people that deal with the practice of digital curation in Italy. Pickard (2007) uses the words of Patton in stating: “the logic of purposeful sampling lies in selecting information-rich cases for study in depth. Information-rich cases are those from which one can learn a great deal about issues of central importance to the purpose of the research” (Patton, 2002, p. 69). In this purposeful perspective, the participants have been selected on the basis of the following *a priori* criteria:

- Participants must have a relevant position in the Italian digital landscape;
- Participants must have a strong background in digital and data management activities;
- Participants must deal with data on a daily basis in their professions.

Snowball sampling has emerged during the first interviews, since some participants have suggested further people that could potentially be interviewed. In addition, the advisor and key informants have given their contribution to building the sample. According to Pickard, “in any bounded system there are key informants who will have a great deal of knowledge about the case as a whole and what goes on at a variety of levels within the case” (2007, p.69). As the practice of digital curation in the Italian context is not simple to evaluate, the work of key informants has

been useful at the beginning. Following the theory of Lincoln and Guba (1985), the exit strategy of the sampling has been *redundancy*: the research has established autonomously when a certain level of saturation of the quality of data had been reached.

Accordingly, participants have been selected based on their professional context. As said, the purpose of this research is to establish a common core of technical and IT skills and competencies in the field of digital curation, in order to understand if these can be the groundwork for the development of different professions. Bearing the research question in mind, participants have purposively been selected from different contexts: none of them has the same professional background or comes from the same workplace or institution. Rather, participants come from the heterogeneous world of Digital Humanities, Academic Libraries, Open and Linked Data movements and private companies.

3.4.2 Key informants

As said, key informants have played an important role in the process of sampling. Moreover, they have also been useful in outlining the structure of the interview. The researcher has been put in contact with two different key informants, regarded as information-rich sources for a preliminary conversation on the topic (Pickard, 2007):

- The first one comes from Italy and is employed at the National Research Council of Pisa. His area of interest is in Computer Science in relation with the world of Digital Libraries.
- The second key informant is from Scotland, and is employed at the Information School at the University of Sheffield. His area of research is digital curation, research data management and information systems.

The Italian key informant has given his support mostly in defining the sample of the research and in defining a general background of IT skills and competencies that would have been useful to deal with and to investigate. The Scottish key informant has mostly given his support in the field of digital curation, especially in facing the current trends and developments of the conceptual model of the Digital Curation Centre. Key informants have helped the researcher to find a critical approach and a specific direction at the beginning of the project. For these reasons, their interviews have not been included as a direct source of data for the research.

3.5 Research techniques

As said many times before, the research techniques are interviews, that is one of the most frequently used data collection techniques in information research (Pickard, 2007). Interviews are also a valid instrument for gaining insight into the views of individuals (Kvale, 1996) and an in-depth investigation of the topic. These reasons have led the researcher to choose this data collection technique among others: the aim of this study is to collect relevant opinions from individuals, to collect in-depth information and to outline the context where this information is stored.

Going into details, *semi-structured interviews* has been chosen as the privileged type. The reason for this choice lies in the flexibility and the structure of this kind of interviews (Kvale, 1996). A closed interview with fixed questions would not have allowed the researcher to grasp the information-rich value of the environment.

Following the seven stages of the interview process outlined by Kvale (1996) and cited by Pickard (2007) – which are *thematizing*, *designing*, *interviewing*, *transcribing*, *analyzing*, *verifying*, and *reporting* – the researcher structured the interviews as follows:

1. The theoretical and the conceptual framework helped the researcher to identify initially some relevant concepts about the topic of the research. Key informants additionally supported this first step.
2. The interview included seven questions about the topic of digital curation and the relative context. The same questions were asked to interviewees (see Appendix 1), allowing them, at the same time, to expand or go beyond the topic of the research. This is the reason why a semi-structured interview system was chosen.
3. Five semi-structured interviews were conducted anonymously. All the interviews were also conducted online, using *Skype* software. This was the *synchronous* stage of the interview (Pickard, 2007). Interviews were recorded using a tool for Skype calls recording, named *Callnote*.
4. At the end of every interview, the researcher started the process of transcribing and coding.
5. The data was then analyzed, following a technique that will be described later in this chapter.
6. At the end of the transcribing process, participants verified their own interview: a *Google Document* was created and shared with each interviewee, so that he or she could check, edit and integrate the content of the interview. This represented the *asynchronous* stage of

the interview (Pickard, 2007).

7. The present research is an attempt to report the findings of the data analysis.

On this basis, it is now possible to sum up what has been said until now:

Interviewee	Professional context	Type of interview	Instruments
Interviewee 1	Digital Humanities	Online	Skype and Google Documents
Interviewee 2	Research Centre	Online	Skype and Google Documents
Interviewee 3	Academic Library	Online	Skype and Google Documents
Interviewee 4	Private Company	Online	Skype and Google Documents
Interviewee 5	Data Curation Centre	Online	Skype and Google Documents

Table 1. Type and mode of the interviews

3.6 Research instruments

The importance of the human as a research instrument in qualitative research has been already discussed. On the other hand, the researcher would like to stress the importance of the *computer as a research instrument*. Many types of software and tools have been used during this research: *Cmap* for the creation of the conceptual maps, *Skype* for online interviews, *Callnote* for recording the interviews, *Google Documents* for transcribing and verifying the content.

Pickard points out some licit doubts about conducting online interviews and the use of Computer Mediated Communication (Pickard, 2007). If on the one hand it is true that online interviews can carry some limitations and run the risk of losing the *sense of the other* (Mann and Steward, 2000), it is also true on the other hand that without the mediation of the computer most of the interviews here reported would not have been made. In conclusion, the computer has allowed the researcher to go beyond some critical logistic and economic constraints.

3.7 Data analysis

As far as data analysis is concerned, a *constant comparative analysis* has been applied. This strategy has been originally developed by Strauss (1987), and involves the examination and the comparison of data with all the similar or different data gathered during the research.

In addition, the following strategy involves a *bottom up* approach, where some categories have to emerge inductively and directly from raw data, and not to be *a priori* established. Following the more evolved model developed by Straus and Corbin in this perspective (1998), three different activities have been conducted:

1. *Open Coding*, through which concepts are identified by means of similarities and differences among raw data. This is the initial phase where the researcher has to identify some discrete concepts that are the “basic units of analysis of the emergent theory” (Strauss and Corbin, 1998, p. 101).
2. *Axial Coding*, through which categories and subcategories are linked together. At this stage of the analysis, the researcher has abandoned the raw data and he or she is manipulating the categories that have emerged during the previous stages of analysis (Pickard, 2007).
3. *Selective Coding*, through which the hypothesis is integrated and refined. In This final stage of data analysis, the theory has reached a certain level of saturation and no more connections, properties and relations are emerging from the analysis. The outcome of this stage is the development of the conceptual framework and the research hypothesis that emerge from data (Pickard, 2007).

The next chapter will introduce data analysis.

Chapter 4: Analysis and findings

As said in the previous chapter, a *constant comparative analysis* has been applied to data analysis. This approach allows the researcher to identify similarities and differences between raw data by means of a bottom up approach, thanks to which some categories initially emerge and are not established *a priori*: this is the phase called *open coding*. After this, *axial* and *selective coding* has followed, and brought directly to the presentation of the results.

There is not a single method to present, analyse and discuss data that emerges from a research. Pickard (2007) highlights how the analysis and the discussion can come in a separate section, after the presentation of data. As an alternative, analysis and discussion can be done together with the presentation. The choice is highly dependent on the kind of research, its purposes and the methodology that has been applied (Pickard, 2007). Information-rich interviews need an in-depth analysis of the topics, the concepts and the keywords that have emerged during the interaction between the researcher and every single participant. For this reason, the researcher presumes that a simple presentation of graphs and charts could not exhaustively show the richness of these qualitative results. The use of words - many words! - is the best method to present qualitative data (Pickard, 2007). Accordingly, these results are presented by means of a *descriptive* approach, since it has been evaluated the best way to present and discuss them. Moreover, analysis and discussion of data have been integrated together with their presentation. Every important topic that has emerged from the theoretical framework and from the interviews represents a paragraph of the following chapter. Furthermore, bearing in mind the phases of the constant comparative analysis, these categories lead to the definition of some subcategories that will help the reader to orient him or herself through these findings. Last, with the aim of helping the reader to have a graphic representation, as well as a general knowledge of the salient topics of these findings, a conceptual map will follow the qualitative analysis.

For convenience, interviewees have been labeled as I1, I2, I3, I4 and I5.

4.1 Conceptualizing the work of the digital curator

As the literature review has highlighted, it is not simple to find a specific definition of what is the practice of digital curation. Similarly, it is not simple to grasp who digital/data curators exactly are and what exactly they does. This is a relevant topic that has emerged from the conversations with the participants to the research. At the beginning of every interview, the

researcher tried to share a common framework with each participant, in order to focus on a definition for the work of the digital curator.

4.1.1 A question of identity

In his attempt to define the skills and competencies of the digital curator, the researcher has initially tried to answer the following question: who is a digital curator? During most of the interviews, a premise on the definition of *digital curation* has been necessary. In other occasions, a comparison between the concept of digital curation in the mind of the participants and in the perspective of the researcher has been necessary to establish a common background.

Despite the massive literature produced in this field during the last years, this term has been defined with extreme approximation. This has often led to some identity issues, since some interviewees recognized many common traits between their roles and the practice of digital curators. For instance when trying to give a definition of digital curation at the beginning of the interview, I4 stated, “In my opinion this is a word that gathers many meanings, and generally speaking I think about it as *working with a certain data type* and *applying a human side* to what machines cannot do.” Interestingly, the same interviewee stated after few minutes “nobody, in my field of work, uses the term digital curator. However, I see those competencies in my job and I think that part of my work should be labeled in this way.” This can easily be seen as a clear example of how little the concept of digital curation is rooted in the Italian context.

Furthermore, the *context* in which the digital/data curator is employed has emerged as an essential characteristic. According to I2, “the data curator should be placed exactly in the right point and in the context in which one is working”. In this perspective, the experience in the job could play an important role in the definition of the identity, and “the work of data curator should be considered in every situation, since someone can become a curator simply thanks to his or her experience.” (I2). The issue of the context has also been pointed out by I1, when he says, “[a digital curator] should see the whole picture, but he/she can’t invent this vision from himself. So, *contexts* that show such vision are necessary.”

4.1.2 Interdisciplinarity: a resource or an obstacle?

One of the most salient aspects that have emerged from the literature review discussed in this project is the interdisciplinary value of the competencies that a digital curator should possess. Although it could be considered as a positive value, as it highlights the potential of this practice in terms of employability, it can undoubtedly cause a certain level of uncertainty in drawing out

the skills and competencies needed to work successfully. Therefore, the aspect of *interdisciplinarity* plays a central role in writing an ideal curriculum vitae for digital curators.

From a learning perspective, I3 raised some issues about what is the appropriate level of education to teach digital curation competencies: “should it be a bachelor degree or master's degree?” (I3). She also raised some issues about how to train young students that choose this career.

Beyond the strict perspective of an educational path, participants were asked to outline a hypothetical curriculum vitae for digital curation. The question has been asked purposively in a general way, and all participants have highlighted the difficulty of making this reflection in general terms. For instance, I1 has stated:

“It seems difficult to imagine a curriculum for a single person who has all the competencies of the digital curator. I see this role as something like the *orchestra director*. Besides a knowledge in the field of music and some technical competencies in one or more instruments, he or she should know a little of everything in order to arrange and coordinate the work.”

Following this perspective, an interesting insight emerged from I4's reflection:

“Ideally speaking, a digital curator is fully bilingual, or even *trilingual*: he or she has competencies in graphics, in the field of Information Science and IT, and technical competencies.”

This beckons some positive issues in favor of the holistic vision for a curriculum in digital curation. Its interdisciplinary nature could mean that the practice of digital curation should apply to different contexts, and this is having an effect on a many different professions. This topic will be discussed in detail later on in this chapter.

4.1.3 The value of Information Technology

The role that Information Technology and Computer Science disciplines play in the field of Digital Humanities is still object of debate among academics, professionals and stakeholders; the same goes for the practice of digital curation.

Nevertheless, the *IT oriented* profiles that have been interviewed during this research have expressed the pivotal role that IT skills play in an efficient curriculum for a digital/data curator. I2 has pointed out that the technical aspect is essential to work successfully in the field of digital curation, whereas other aspects – such as collaborative and managerial competencies – are not

equally relevant. In line with this “geeky” perspective, I4 has declared that, in his opinion, “the digital curator should learn the ropes on IT, which means that he or she should have basic competencies, but very focused, on what *data, data format and metadata schema* mean.” Furthermore:

“I see the IT issue or the issue about competencies on data as the old issue about the knowledge of English language. A top manager who cannot speak English nowadays is not seriously considered. The same is for data, since nowadays data are everywhere and call for IT competencies, which in turn call for competencies in digital curation.”

A more focused opinion is shared by I5, referring to the creation of a curriculum for digital curation:

“First of all, [a digital curation curriculum] should be based on LIS disciplines, with a strong influence on Computer Science disciplines. Here, I am referring to text mining, data mining and knowledge representation activities.”

In a broader and lighter sense, also participants who have a background in LIS or in arts and humanities have stated the importance of technical competencies. For instance, I1 calls for ongoing training in this field: “few hours [of training] are not enough, since there are many specialist skills at stake, such as the comprehension of IT data and the context in which they are placed.”

4.1.4 The value of human contribution

The results of this reflection come from a specific question asked by the researcher during the interviews. This issue was raised by a comment made by one of the key informants discussing the level of contribution of the data curator in the data lifecycle. In conclusion, after a brief reflection, the key informant commented:

“Also remember that skills can also be replaced by technologies. People are trying to develop technologies that effectively provide support to data management, thus reducing the need for an expert. In addition, skills can be replaced by outsourcing. So probably many libraries will outsource the more technical aspects of data curation.”

This comment has led the researcher to ask himself, can technology totally replace the human contribution to data management? Could the role of the data curator disappear in a few years,

due to the implementation of machines and software solutions that deal with data more efficiently? These questions have been accordingly inserted in the structure of the interviews. The question was purposively asked after questions about technical and IT skills and competencies, in order to place this topic within the conceptual framework that had been set between the researcher and the interviewee.

In this perspective, a homogeneous response emerged. Participants believe that the human contribution will lose importance over the next years. I2 has clearly expressed this concept, stating, “I do not think that data curator will have a short life. Rather, I think that he/she will turn into something else.” This means that it is likely that, one day, the digital curator will deal with new skills and competencies, due to the increasing complexity of data management. This concept was mentioned by I3, referring especially to research data:

“As far as research data go, I see a problem in the fact that not all data can be stored, but it is necessary to choose which data have to be stored, for how long and how. I don’t see [the conclusion of the digital curation] as an issue, but I see change and an increase in the complexity of this profession.”

Another important topic is *automation*, which is the reason why many repetitive tasks are delegated to machines. Most participants have highlighted how automation is increasingly important in fields like Big Data. For instance, I2 and I4 both stress the importance of the human value in the field of knowledge representation (I5) and in the knowledge that can stem from the analysis of data, even when the curator starts from some uncertain premises given by raw or unclear data (I1 and I4).

4.2 Contextualizing the work of the digital curator

Due to identity issues that have emerged in the previous section, the context plays an essential role in the outline of a digital curator’s curriculum. Hence, some of the specific key elements in this field have been discussed with the participants. The aim is to give a relevant overview of the Italian scenario, in order to contextualize the practice of digital curation in Italy and to better understand the infrastructures and communities where digital curators work.

4.2.1 The perception of what Cyberinfrastructure is

At first, it is relevant to clarify that the concept of *cyberinfrastructure* was not entirely known to the participants. The researcher often had to explain the concept in order to clarify the

interviewees' doubts. As stated, a reason could be that the term *cyberinfrastructure* is used mostly in the US, whereas in the European context academics prefer the term *e-science*. On the other hand, the researcher believes that the term *cyberinfrastructure* is semantically richer than the term *e-science*, since it focuses the attention on the technical requirements for the development of digital innovation.

All participants have expressed similar thoughts about the poor quality of broadband connection at a national level. This is a highly debated issue in Italy, since institutions, companies and stakeholders have been complaining a lack of quality in this field for many years. I5, for example, defines the quality of broadband connection as “the minimal requirement to do our job.”

In a more general perspective, interviewees have highlighted many deficiencies in this field, often due to technological obsolescence (I3) and bureaucratic issues (I2). However, the most important reflection has emerged in talking about the gap between the IT facilities and the technical competencies that are needed. I2 has stated, “I see some cyberinfrastructure in Italy. What is needed is skills: if these are well focused, use of the technology solely for entertainment purposes will be avoided, and will become a useful instrument.” The same perception is shared by I5, when he says, “we don't lack machines, but the knowledge of how to use them.” Conversely, I3 shifts the attention on other external factors: “we have machines and skilled people, but we are struggling to build a solid cyberinfrastructure.” By contrast, this is not seen as a problem for I1, since, despite some deficiencies, relevant projects have been realized, nevertheless:

“If digital curation means taking care of digital data, our country – despite many delays in creating infrastructures – owns data collections that need competencies in digital curation. We have many examples of digital libraries in Italy, starting from the very important Italian Digital Library, which gathers all the volumes of copyright-free Italian literature. They have been created despite the low quality of the cyberinfrastructure.”

Participants have also highlighted a remarkable difference between the public and the private sector: these are seen as two different contexts that are struggling to communicate (I3, I4 and I5).

Moreover, the geographic location has emerged as a relevant aspect, since the quality of the cyberinfrastructure varies highly from one region to another. I2 and I5 have focused the attention on how much regions invest locally in the development of cyberinfrastructure. These local differences have inevitably brought to delays and deficiencies at a national level, since the development of the cyberinfrastructure is highly decentralized.

4.2.2 Open Data and collaboration

Decentralization could be seen not solely as a problem, but also as a resource. This is especially true for the Open Data movement, which is often endorsed by people that work in the field of digital/data curation. There are many digital communities that make collaboration the fundamental asset of innovation.

Participants have expressed the importance of the adoption of the Open Data philosophy within the practice of digital curation. I1 said, “there is little awareness of the fact that *openness* is a fundamental characteristic.”

Some interviewees have stated that the collaborative aspect of Open Data is struggling to succeed due to a frame of mind that is at the common with many private companies that supply software for data management (I2). I3, for instance, stated, “We are still lagging behind, since companies tend to create property owned software, and since we have many archives and repositories that are like silos.”

The university context has also been highly debated. I1 emphasizes how “The research university context is perhaps more open to acquiring new knowledge than financial and industrial businesses.” By contrast, I5 complains the fact that many universities see research data management as something that has to be held within the university and not to be shared with other institutions. In a more technical perspective, I3 sees the lack of data interoperability attitude among universities.

In terms of use of open source software, I4 has a negative view of the approach of institutions and public sectors, and shifts the attention on other contexts:

“I see a lot of ferment in other places, such as *GitHub*,¹⁷ where it is possible to find many open source software solutions that are used very often and rely on virality and on word of mouth. [...] The relationship between the library and the open world is also interesting. The world of libraries recognizes the importance of digital common goods like *Wikipedia*, *Wikisource*, *Wikidata* and *Wikicommons*, and I think that things will progressively get better.”

However, this attitude could also cause some problems: I2 highlights that despite the open world can concur to fix the infrastructure, “nowadays [the open source movement] is driven mostly by political reasons, or transparency reasons, in which systems and processes are ignored.”

17 <https://github.com>

4.3 Technical competencies

As stated previously, one of the objectives of the following research project is to identify and define basic IT skills and competencies that allow digital/data curators to do their job. In this perspective, the creation of a list of core IT skills and competencies that have emerged from interviews becomes essential to understanding the practice of digital curation in the Italian context. In the following section, there is an attempt to provide this list, in order to give a set of instruments that could be useful for digital/data curators.

4.3.1 Metadata standard and schemas

Metadata and *metadata schemas* represent the way in which data are structured and become information-rich data. They are also a fundamental aspect of the data lifecycle, since the way the digital resource is described by means of metadata can affect the way this resource will be retrieved. This aspect has been underlined by I2, when he says, “[the digital curator] should have competencies in metadata description, bearing constantly in mind the idea of *reuse*.”

Metadata and metadata schemas as a core knowledge have been highlighted also by I3 and I4. By contrast, I1 has stressed the importance of principle over the instruments, both for metadata and for other competences such as XML and interoperability:

“I agree that knowledge of XML, metadata standards and interoperability protocols are necessary. However, I maintain the importance of principles over the instruments as a fundamental issue.”

Another interesting insight has come from I5, regarding a detailed illustration of the type of metadata schemas that should be learned by digital curators:

“There is usually a big problem in the making of the curricula, since only textual descriptive standards are taken into consideration. Perhaps, standards for cataloguing images and graphic contents should be equally considered, and not only the Dublin Core or other Learning Object Metadata, but also those developed by Getty¹⁸ for the cataloguing of works of art, or those of the CARARE¹⁹ project for the cataloguing of 3D models.”

The latter is a relevant issue, since it raises some questions about which metadata schema a digital curator should definitely learn. If it is true, on the one hand, that the adoption of a specific

18 http://www.getty.edu/research/publications/electronic_publications/cdwa/

19 <http://www.carare.eu/swe/Resources/CARARE-Documentation/CARARE-metadata-schema>

metadata schema is highly dependent on the *context* in which the digital curator is placed, it is equally true, on the other, that metadata schemas that stem from LIS disciplines are probably not enough to deal with the work of the data curator nowadays.

Moreover, the purpose of this research is to establish if the IT skills of the digital curator could be the basis for the development of other related professions. Knowledge of metadata schemas is surely one of these skills, but it should also apply outside the context of Libraries and Information Science. The example given by I5 is especially valid for the Italian context, where there is an impressive cultural heritage and curators must deal with a huge amount of images or graphic-related contents.

4.3.2 Extensible Markup Language

Tim Berners-Lee has defined the eXtensible Markup Language (XML) as the *lingua franca* of the Web (2007). Its importance is globally recognized, and it is now taken for granted that its manipulability allows to express and deal with a variety of metadata schemas. However, although the knowledge of XML is required in many fields of Digital Humanities, should it be considered an essential competence for the practice of digital curation?

The common view has explicated that a competence in the use XML is a requirement for the job of a digital curator. On the other hand, some level of uncertainty has emerged around the *essentiality* of this competence. I4 has stated “a good digital curator should know what XML is and how to handle it”. The same view is shared by I3 and I5. In particular, the latter says:

“Starting from the foundations, I would take into consideration the language for data representation, and so every XML based language. A good competence of XML and its TEI²⁰ derivation should be achieved from the beginning.”

However, the essentiality of XML has been a matter of discussion for others. Even though he has recognized its importance, I2 has stated:

“I do not totally agree, since I see a lot of people working with XML non-compliantly. However, I would teach at first what structuring data means, and then I would teach XML or RDF, but not first. Instruments such as XML, RDF, SQL and JSON are just part of a path and rely on someone’s needs.”

The last concept has been analysed in a more theoretical way by I1: although he considers XML as very important, he does not think it is *essential*. Similarly, in his opinion on metadata

20 Text Encoding Initiative

competencies, XML represents the form through which other and more important principles (Linked Open Data, in his opinion) are expressed.

An explanation of these different visions can come from the importance of the use of XML in their daily job. The more they use XML on a daily basis, the more they consider it as essential. Furthermore, it is important to outline that participants that have considered XML as essential have some basic knowledge in LIS disciplines.

Interestingly, I5 has indicated knowledge of XSLT²¹ as the foundation (together with XQuery) of a good curriculum in digital curation. Furthermore, he sees the knowledge of XSLT as something that should come before the knowledge of other languages, such as RDF, RDFS and OWL.

4.3.3 Ontologies

The importance of competencies in *knowledge representation*, and hence, of ontologies, has emerged more theoretically than in practical terms. Here, I5 has stressed the importance of the theory of knowledge representation in the definition of a curriculum for digital curation. Going into details, he has also identified some technical skills and competencies that can help in the development of ontologies:

“[Ontologies] are the foundation for the development of the essential competencies of the data curator: the first is, as I said previously, knowledge representation, which can be developed in many ways. In my opinion an introduction to UML²² (Unified Modeling Language) is very important.”

This technical perspective on the field of representation of knowledge is also shared by I1, who sees ontologies as the point where technicians and humanists are bound together and can share their background:

“The technician who writes the ontology analyses the contents in depth; on the other hand, a humanist who knows what ontology allows by means of contents, must fully comprehend ontologies. Eventually, the distinction between technician and humanist is lost.”

By contrast, I2 and I3 see ontology-related competencies as something that should be learned after competencies that are more basic.

21 Extensible Stylesheet language, a language for the transformation of XML documents:

<http://www.w3schools.com/xsl/>

22 <http://www.uml.org/>

4.3.4 Data, Database and Data Interoperability

Competencies in this field seem to be the most difficult to outline and identify. The reason lies in the fact that dealing with data and dataset can require highly specialized competencies, that most of the times are not initially possessed by a digital curator. As a consequence, the context and the kind of work where the digital/data curator is employed plays a central role in the definition of these skills. Hence, participants have answered based on their professional background, which led to the identification of different instruments and tools.

Starting from *data* as a unit of analysis, I2 has stressed the importance of the difference between the structure and the representation of data. I3 has added a general knowledge of the processes of digitization and of digital formats, such as JPEG, TIFF and PDF. In addition, the following analysis highlights how the purpose of use is essential to understanding how to manage data.

Secondly, knowledge of *databases* management is required. From the perspective of a curriculum for digital curators, this is a fundamental issue. Even though being a data curator does not mean being a database expert, knowledge in this field has emerged as an essential competence.

I3 and I5 have underlined the importance of knowing software for data storage and management of data repositories. In particular, I5 has focused the attention on software for the design of databases, as well as software products like XQuery²³ (together with XSLT) for their interrogation. I5 has also added GIS competencies, since, in his opinion, “the work of a data curator does not involve texts solely, but also images with a lot of information: their localization and structuring at the level of storing is thus fundamental.” I5 carries on with other competencies that call into question the educational background of the digital curator:

“In my opinion, the use of tools like R,²⁴ competencies in text mining or web mining, tools for clustering that allow someone to catch information in huge datasets are fundamental. Obviously, this includes competencies in statistics, in linguistics, and in token differentiation within Computational Linguistics.”

This perspective is typical of *data science*, where skills, competencies and educational background are more oriented towards the Big Data and slightly different from the skills of a digital curator. To support this statement, a 2012 article of the *Harvard Business Review* defines

23 <http://www.w3schools.com/xquery/>

24 A software for statistical analysis, <http://www.r-project.org/>

the data scientist as *the sexiest job of the 21st century*,²⁵ calling for a solid foundation in mathematics, statistics, probability and computer science, and certain habits of mind. This will raise some serious issues about the identity and the education of the data curator in the next few years.

The use of JSON²⁶ format deserves a distinct analysis. This file format for data exchange has been cited by those participants that manage, use and download datasets on a daily basis. In more than one occasion (I2 and I4), JSON has been judged as the format that can naturally replace other and more complex languages and formats, such as RDF and RDFS. I5 has highlighted the importance of the use of JSON in data management, stating that it is used very often. In this field, I4 has also highlighted the importance of the knowledge of Python language:

“A good digital curator knows what XML is and how to handle it [...], he is able to use *OpenRefine* [...] and *Sublime* [...] and he has basic knowledge of Python, which allows him or her to write some codes to download and modify a JSON file.”

Obviously, this statement acknowledges the existence of other tools like *OpenRefine*²⁷ and *Sublime*,²⁸ which refer to data cleaning and text editing.

JSON plays an important role also in *data interoperability*. I2, for instance, defines data interoperability as an essential competence when data becomes *machine readable*. In this view, he makes the distinction between “standard *de jure* and standard *de facto*, which means that some of them, like JSON, does effectively what it should do.”

Talking about data and metadata interoperability, all the participants (except for I1, who believes in the importance of principles over instruments) have underlined their importance at a theoretical level. More interesting is the fact that only I3 has explicitly referred to OAI-PMH protocol, which could be explained by means of her professional background (research data management).

4.3.5 Semantic tools: Linked Data and Semantic Web

As far as the following skills and competencies are concerned, the researcher has chosen to put them under the same umbrella and to define them as *Semantic Tools*. The reason for this choice lies in the fact that Semantic Web and Linked Data are concepts that are strongly related. Making a distinction between these concepts would have meant isolating competencies that

25 <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/>, accessed April 2015

26 JavaScript Object Notation: <http://json.org/>

27 <http://openrefine.org/>

28 <http://www.sublimetext.com/>

cannot be considered separately, Linked Data can be seen as the natural consequence of the Semantic Web.

At first, the Semantic Web is naturally seen as the next level after the competences on knowledge representation and metadata schemas. Most participants have underlined the important difference between *human readable* and *machine readable* data. Semantic Web is a competence that encompasses other basic skills and, even if it is considered essential, it comes after other basic competencies such as data format (I3), knowledge representation (I5), and tools for data mining and data management (I4). In this perspective, I3 was the only participant who stressed the importance of the knowledge of *persistent identifiers*, such as DOI.

Surprisingly, competencies in the field of Linked Data deserve a distinct discussion. The reason is that a heterogeneous perception of the importance of this competence has emerged. First, there are those who granted Linked Data a primary role in the making of a curriculum for digital curation. I1, for example, pointed out that “Linked Open Data are an essential aspect, both theoretically and as a tool for working”, since “the first essential aspect is that data are linked, which means that we think about data only in terms of Linked Data”.

However, there are other participants that have expressed more than a doubt about the use and the potential of Linked Data. Although they have used different words, they seem to share the same vision. For example, I2 has pointed out, “It is true that in the world of Linked Data there is someone that makes extraordinary things, but it is highly difficult to create Linked Data, since many competences are required.” Later on in the interview, he has remarked the concept:

“Ontologies, Linked Data and semantics are very delicate topics, since they have to be inserted in the context where the data curator works. Most of the times, I see people that make great effort to gain little results, which means that they use Linked Data to do what they previously did.”

The same vision has emerged during the interview with I4:

“Paradoxically, the topic of Linked Data could even be set aside. Even though we like it, it is used to do few things. It should be used in huge projects, but in little projects, an API could give the same results. Knowing how to use and supply API is sometimes sufficient and the concept of Linked Data is not necessary.”

Later on, I4 underlined this concept during his interview, saying, “These are all things that need a context, but many times JSON might not work, and RDF is just a more complicated way to supply metadata.” This raises some relevant issues about Linked Data as an essential competence for the digital curator.

4.3.6 Data visualization

Skills in data visualization have emerged as a very controversial topic. As said, data visualization is now emerging as a field of work for the digital curator and a current development of the data lifecycle. From a quantitative perspective, skills in data visualization have curiously been evaluated in turn as *essential*, *more essential than many others*, *not essential and not taken into consideration*. The following description could be seen as a quantitative method of data analysis, but it seems the best way to get, eventually, to a qualitative interpretation.

I2 and I4 have evaluated data visualization as an essential skill in the practice of digital curation. While I2 has defined “data visualization as fundamental”, I4 stated that he was “not sure if the visualization of data can be included as a competence of the digital curator. Surely in my mind it can.”

I1 sees the issue of data visualization as something placed in an intermediate position between principles and instruments:

“I’m not sure to place [data visualization] at the same level of principles. I see the theme of data visualization at a higher level than instruments, but I would not place it alongside the principles and concepts. Maybe, I should place it at an intermediate level.”

By contrast, I5 does not judge data visualization as an essential competence of the data curator:

“I see the work of a data curator as something related to *data science*: he should have competencies on data visualization, but I don’t think it is essential. More fundamental are, in my opinion, the use of tools like R, competencies in text mining or web mining, tools for clustering that allow someone to catch information in huge datasets.”

Lastly, I3 has not cited data visualization as a competence.

This broad *spectrum* of evaluations is justified by each participant’s different idea about the identity of the digital curator. Interviewees who think about digital curation as an activity focused mostly on the preservation and administration of data see this as a non-essential requirement. Conversely, interviewees who have shown a more dynamic and *result-oriented* view of the activity of digital curation see data visualization as an essential element.

4.4 New roles and professions for digital curators

This research project is based on the assumption that the basic IT skills and competencies of the digital/data curator can be the basis for the development of new kinds of professions. From this point of view, data curators can specialize in many professional profiles, such as data librarians, data journalists, data archivists and so on. The researcher has specifically asked the participants to give their opinion in this field. Some similarities and some differences have hence emerged.

4.4.1 Specialization

A precise question about an opinion on the fact that digital/data curators could specialize in different professions was asked to every participant. Answers have reached a high level of homogeneity, since they all agree with this hypothesis. The need for a *specialization* in a specific field has emerged as an essential tool. I5, for example, stated “a high level of specialization is usual in the everyday work of the digital curator”, and:

“In my opinion, there is a common background. Like every discipline, one specializes on a certain view of reality. For instance, the data librarian takes the text as a landmark and specializes in giving knowledge, techniques and choices at textual level. [...] A data librarian does not focus on these areas but will concentrate on others, like structured data in FRBR and the linkage of these entities with others that come from institutes with which there is a collaboration.”

Here, it is important to outline that participants have highlighted how specialization could be reached in various ways. I3 has pointed this out explicitly:

“Many philosophers deal with ontologies, since it is interesting to see that the digital curator is not just a specialist in librarianship, but should become a domain specialist like the philosopher that deals with Linked Data. The latter and the Semantic Web are topics that are becoming diagonal in many disciplines.”

In this perspective, I2 and I4 have cited *data journalism* as a possible specialization: while I2 has initially stated that, in his opinion, “the data curator is not a person who works in libraries and museums solely”, he developed this concept later, as follows:

“The data curator does not work in few professional sectors. He could work in

research centres, for instance, or become a data journalist. In the latter field a lot of skills are needed in this job, but one could specialize in one or two skills. Data curation is a very important profession that is becoming more and more important, and that transforms in many different ways.”

Interestingly, a critical approach to this vision comes from I1, when he states that IT requirements can raise some problematic issues in the definition of the profession:

“On the one hand these competencies (data librarian, data archivist, data scientist, data journalist) are defined by the IT, because the instruments for working with data come from that field; on the other, these competencies remain rather incomplete, since many people think that these instruments are good as they are, without a detailed analysis of the needs of those who work with them.”

Lastly, it is noteworthy that, even though the researcher has explicitly asked participants to make further examples of specialization, no relevant answers were given.

4.4.2 Teamworking

Specialization calls for *segmentation*, which calls for the importance of *working in a team*. Using the words of I1, “vertical competences are needed, since we can’t think that a single person can do everything.” Once the digital curator has found the right context, he or she has to deal with it. Again, some interesting insight come from I1:

“Initially, the dream of Digital Humanities was that one person possessed all the competencies needed to work in the digital sector, so the philologist became a programmer. As time goes by, [...] I think that it is impossible that a single person is highly skilled at all levels. I think that today’s logic of research is correct: jobs are team based, where teams mean gathering different people with different and specialized competencies.”

In this perspective, the work of a digital curator is something that should be assigned to many people with many competencies (I1). I2 reiterates these concepts when he says, “the data curator can be employed in many ways, but this is not a path that someone can do on their own.”

Hence, collaboration between data curators and other professionals plays a central role. An interesting insight comes from the words of I4:

“What I have learned from my job is that if you learn to talk with technicians you get better results, since you are talking a common language. If you can talk successfully

to a database administrator and you succeed in saying what you would like to do, he will forecast some potential problems, and the job will be easier.”

This topic is indeed highly related to the context and the way in which the digital curator is employed. This is another evidence of the *fluid identity* of the digital/data curator.

4.4.3 Critical issues for LAM networks

In his attempt to demonstrate that the role of the digital curator can produce other professions, the researcher has surprisingly found himself dealing with some critical reflections on the current role of the LAM networks.

Contrarily to the hypothesis of this research thesis, the practice of digital curation could even result in the disappearance of some professions. In particular, this has been underlined by I3, saying that she sees serious employment issues for archivists. In her view, and referring to research data management, “it might be that research communities become so autonomous in the curation of their data that they could eventually bypass the profession.” Again:

“The problem with academic libraries is that they integrate with research communities. The more there is integration, the more research communities run the risk of imposing themselves on the work of the academic librarian. This is an evolving scenario, in which I am not saying that the academic librarian will not do data curation one day, but it is likely that he will not be the only one that do it.”

From another perspective, I4 analyses critically the role of LIS professionals and their discipline, especially when they deal with the *data deluge* issue. Even though I4 has a LIS background, he introduces a severe judgment:

“Today, many competences in data are acquired by many people and applied to many other contexts. Paradoxically, data journalists have emerged before data librarians, or at least data librarians have been locked in their garden, with their MARC, while the Web was evolving and leaving them out. I think that, apart from the GLAM context, everyone else will become a data curator, since I see a serious delay.”

In addition:

“The Information Scientist thought to hold the knowledge in this sector, but they have been overtaken by people that know how to manipulate data, create databases and IT tools, how to program and so on. Automation is in charge in the field of data, and whoever is able to manage it wins, whereas those who just know FRBR lose.

Everything that is about data will need curation.”

Even though this concept has been used to criticize the LIS world, it can be used positively for the purpose of this research. The researcher has therefore shared with the interviewees the hypothesis according to which the role of the digital librarian (and therefore, of the data curator) could explode in many directions. I4 has answered:

“Absolutely. It depends on how we look at these things. If we lift our gaze, look at the competencies of the librarian and call them IT competencies, they can blow up and really be used where there is data.”

4.4.4 Back-end vs. front-end

Going back to the DCC Lifecycle Model, another purpose of the research is to establish if these new roles in data management allow digital curators to add value within the data lifecycle. Thus, digital curators can contribute to finding different patterns of knowledge between data and datasets, a fact which raises further relevant questions: can digital curators be employed even in front-end environments, rather than back-end context solely?

The following question has been perceived as rather controversial among the participants. Whilst they all agreed with the hypothesis of new roles and professions that the practice of digital curation can yield, several doubts about the ability to produce new contents have been raised. Firstly, the majority of the participants (I1, I2, I3, I4) have pointed out that this ability relies mostly on the technical competencies of the digital curator. I3, for example, sees this opportunity only with the help of Linked Data, whereas I4 states that, in this initial phase for the practice of digital curation, he or she can work in many different environments. In a more detailed way, I5 states that in his job he often deals with data that come from different sources and different descriptions:

“I can’t talk in the name of everyone, but I had to do this and it is part of the job of the digital curator. Once that someone manages the data of one or more institutions, it is usual to predict how that data can be linked together, and how, at a query level, not only a single piece of data, but also the connections among them, can be retrieved.”

By contrast, I1 highlights the difficulties of this process:

“I partly agree with this, but I am also dubious. Adding new knowledge means not

only knowing data digitally, but also culturally. A data curator can achieve this, but on the other hand they might not. This is related to teamwork's topic, where knowledge can come from the *content expert* or from the digital curator.”

These issues have been further discussed when talking about the ability of adding value in terms of *back-end* and *front-end*. Interestingly, participants have highlighted how the line between front-end and back-end is very thin and, sometimes, does not even exist. I2, for example, has stated:

“If [the digital curator] starts his or her activity bearing in mind the context in which he or she is working, this will have effects even on the front-end. I have seen people who have started from the back-end to get better in the front-end, and conversely. The losers are those who have started from the back-end of the front-end, and there they stay.”

A similar vision is shared by I4, who even contextualizes this issue:

“On the basis of the context and aim that he or she has to reach, he or she will be placed in turn in the front-end or the back-end. For example, in data journalism it might be that someone else will write the storytelling, but the digital curator will be an essential component in any case, since he or she will supply data, will give it the right aspect and could even take care of its visualization.”

Lastly, talking about their job as data curators, I5 and I4 does not see relevant differences in these two activities. The former declared, “I see the production of the content even as a back-end activity. Work done just in the back-end is not a good job, because I believe that the work of a data curator is, first of all, talking with users and with communities.” At the same time, I4 states, “Thanks to an existing infrastructure; my back-end job turns into a front-end job just with a click of the mouse.”

At the end of this analysis, a map of the main concepts that have been here discussed can give the readers a graphic support.

Chapter 5: Conclusions

At the conclusion of this research, the most relevant findings that emerged from data analysis are aligned with the objectives of the study and concisely presented. Research objectives have initially been individuated as follows:

- To analyse the perception of the practice of *digital/data curation* within the Italian digital communities, as well as other related concepts thank to which it could be developed.
- To identify and define basic IT skills and competencies that allow digital/data curators to perform their job, and to establish whether or not they can be used by many professions.

Moreover, some implications for further research are suggested.

5.1 Research objectives

As said, the first objective of this research study was an analysis of how the activity of digital/data curation is perceived by Italian digital communities. In this perspective, other related concepts that helped to outline what is digital curation have emerged during the interviews. To sum up:

1. The idea of the tasks and activities of the digital curator varies highly among the participants. In addition, a certain level of approximation has been found. During the investigation process many *identity* issues have raised, since some interviewee has found out to have lot in common with the practice of digital curators. Furthermore, the *context* in which the digital/data curator is employed has emerged as a fundamental issue in the definition of the identity of the digital curator.
2. The aspect of *interdisciplinarity* plays a central role in the designing of a hypothetical curriculum for digital curators. This has highlighted some positive aspects, such as the vision of the digital curator as the *orchestra director*, that means someone who knows a little of everything in order to arrange and coordinate the work. By contrast, due to interdisciplinarity, a participant has raised some difficulties in the arrangement of an

educational path in the right way.

3. Generally speaking, the pivotal role that *IT skills* play in an efficient curriculum for a digital/data curator has been underlined by the participants. A more radical view in this perspective has been endorsed by those who have a more IT oriented curriculum. Nevertheless, the importance of IT skills has been highlighted even by participants with a background in LIS disciplines and Digital Humanities.
4. In parallel, the *human contribution* has been evaluated as fundamental, and as something that will not run out over the next years. A homogeneous response has come from the participants, focusing especially the attention on the future increasing of complexity that data managers will deal with. Moreover, *automation* has been judged as the reason for why many repetitive tasks are delegated to the machine, and for why it is increasingly important in fields like Big Data.
5. The idea of the *cyberinfrastructure* as something that can practically support the work of the digital/data curator has been vaguely analysed by the participants. This can be partly due to the term itself, which is not fully recognised within the European context. However, interviewees have mostly focused their attention on the deficiencies of the Italian cyberinfrastructure, especially for what concerns the quality of the broadband connection and the effect of the decentralization, which in turn remarks upon a relevant gap from one region to another. In addition, it is noteworthy that, despite the deficiencies, the quality of the Italian cyberinfrastructure has been perceived positively.
6. It has been observed that people that work in the field of the digital/data curation endorse *Open Data* movements, and participants have expressed the importance of the adoption of this attitude within the practice of the digital curation. In this perspective, some interviewees have observed that collaboration is struggling to succeed due to the private logic at the basis of many companies that supply software for data management. This lack of *collaboration* has been evaluated as a serious topic especially within university contexts, where research data management are sometimes not shared with other institutions and there is no interoperability among repositories.

The second objective of the following research was to identify and define some basic IT skills and competencies thanks to which digital/data curators can perform their job. Furthermore, the

researcher has tried to establish if these IT skills and competencies can be the basis for the development of many professions that are related one another. As far as IT skills and competencies, a list can be synthesized as follows:

1. *Metadata* and *metadata schemas* have been recognised as a core knowledge in the field of digital/data curation. Interestingly, it has been observed that the current developments of curricula in digital curation take into consideration only textual descriptive standards, rather than standards for cataloguing images and graphical contents. Accordingly, metadata schemas from the world of LIS disciplines are probably not enough to deal with the work of the data curator: due to the many contexts in which he or she can be employed, a knowledge of specific metadata schemas out of the context of the Libraries and Information Science is needed.
2. A certain level of uncertainty has emerged around the essentiality of the knowledge of *XML* and *XML based languages* (such as RDF and RDFS). While some interviewees have steadily defined this competence as essential, others have expressed some doubts, because see XML as just the expression of higher principles or because they have observed an inappropriate use of this tool. It is important to highlight that the importance of the knowledge of XML has been stressed by people that actually work as digital/data curators, and use XML on a daily basis.
3. The importance of competencies in *knowledge representation* and *ontologies* has emerged more theoretically than in practical terms. Moreover, some participants have evaluated ontology-related competences as something that should be learned after more basic competencies. In practical terms, a basic knowledge of *Unified Modeling Language* (UML) has been suggested by a participant.
4. *Data* and *database-related* competencies have been evaluated as essential, but they require a highly specialized knowledge in the field. Furthermore, these competencies are highly dependent on the kind of work the digital curator is doing. However, knowledge in the management of databases is required, as well as knowledge in data mining. Some specific tools have been mentioned, such as *Xquery*, *R*, *OpenRefine* and *Sublime*. Competencies on the use of JSON format files are here noteworthy: it has emerged that these can practically replace the use of RDF and RDFS in many circumstances, when curators deal with different types of data and datasets. Moreover, some competencies in

the field of statistics, linguistics and computational linguistics have emerged as useful for the work of data curator, but these are highly related to a data science vision. Last, interoperability among data and datasets has been evaluated as fundamental, but the use of the OAI-PMH protocol has been explicitly mentioned just one time.

5. *Semantic Web* and *Linked Data* have been judged as essential in the moment in which data become machine readable. In this perspective, Semantic Web is a competence that encompasses other basic skills and, even if it is considered as essential, it comes after other basic competencies. Surprisingly, a heterogeneous perception about the importance of competencies on Linked Data has emerged, and someone has expressed more than a doubt on the use and their potential. In detail, a relevant gap between the effort to produce Linked Data and the insufficiency of the results has been emphasized: sometimes the use of JSON format is more effective than the use of RDF.
6. The emerging skill in *data visualization* has had a particular reception: it has been evaluated in turn as *essential, more essential than many others, not essential and not taken into consideration*. Differences in the evaluation of this competence can be explained by the different idea of the identity of the digital curator given by the participants.

Given these basic skills and competencies, finding regarding to the development of new kind of professions can be described as follows:

1. *Specialization* is an essential aspect of the practice of digital curation. Participants have agreed with the hypothesis that data curation can bring to the creation of related professional profiles, such as data journalists, data librarians, data managers and so on. However, the development of these professions calls for a specialization of the curriculum, which could call in turn a re-arrangement of the basic IT skills and competencies. Moreover, participants have seemed to agree more theoretically than practically, since any relevant example has been made.
2. *Working in team* has emerged as essential for the work of the digital curator. This is because it is highly unlikely that a single person can perform all the tasks that this activity requires. Hence, issues about the collaboration with other professionals and the context in which the digital curator is employed has re-emerged, making it clear the *fluid*

identity of this profession.

3. Some critical reflections about the current role of the *LAM networks* has been observed. Someone has raised issues on the fact that the practice of the digital curation could even bring to the disappearance of some professions, especially to those related to research data and research data repositories. Another participant sees a serious delay in the LIS context, making it clear that Information Scientist have been overtaken by people with many skills in data manipulation and dataset management. The same participant has also pointed out that, with the help of IT skills, the profession of digital librarian could explode in many directions.
4. The idea of digital curator as a producer of *added value* has been perceived as rather controversial among the participants. Whilst they all agreed with the hypothesis of new roles and professions that the practice of digital curation can bring, several doubts on the ability in the production of new contents have raised. This issue has been discussed even in terms of *back-end* and *front-end*: here, participants have highlighted how the difference between these two activities is not very marked and, sometimes, does not even exist.

5.2 Implications for further research

The results have shown a complex scenario in the field of the digital curation in the Italian landscape. Some issues have been analysed superficially, and call for further investigation. In particular:

- A more specialized research can be conducted. For instance, IT skills and competencies for highly specialized curricula - data librarians, data journalists, data scientists and so on – can be investigated. Furthermore, it is possible to analyse the differences among these curricula, in order to outline highly specialized profiles.
- The following research can be replicated to other contexts at a national and local level. It would be possible, in this way, to understand the specific IT skills and competencies that need in every country in the field of digital curation.
- A quantitative research can be conducted among digital and data curators. In this way, it would be possible to have quantitative data from a broader sample by means of other

research instruments.

- As said, a training course in digital curation has been completing at the University of Turin. It would be interesting to investigate how the course has been structured and what are the opinions of the students that have completed the course. At the same time, it would be interesting to make a comparison between what has been taught and what is requested by professional sectors in the field

References

Atkins, D. (2003). Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure.

Baker, K. S., & Yarmey, L. (2009). Data stewardship: Environmental data curation and a web-of-repositories. *International Journal of Digital Curation*, 4(2), 12-27.

Beagrie, N., & Greenstein, D. (1998). A strategic policy framework for creating and preserving digital collections: a report to the Digital Archiving Working Group.

Beagrie, N. (2006). Digital curation for science, digital libraries, and individuals. *International Journal of Digital Curation*, 1(1), 3-16.

Beagrie, N. (2008). Digital curation for science, digital libraries, and individuals. *International Journal of Digital Curation*, 1(1), 3-16.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific american*, 284(5), 28-37.

Berners-Lee, T. (2007). Linked data. Retrieved February 2015, from <http://www.w3.org/DesignIssues/LinkedData.html>.

Borgman, C. L. (2008). Data, disciplines, and scholarly publishing. *Learned Publishing*, 21(1), 29-38.

Borgman, C., (2010). "Why data matters to librarians – and how to educate the next generation", *The Changing Role of Libraries in Support of Research Data Activities: A Public Symposium*, National Academy of Science, Board on Research Data and Information, 12 February, Washington DC, USA,

Borgman, C., (2012). IS289 "Data, Data Practices, and Data Curation", *Graduate School of Education and Information Studies*, University of California, Los Angeles

Cassella, M., & Morando, M. (2012). Fostering new roles for librarians: skills set for repository managers—results of a survey in Italy. *Liber Quarterly*, 21(3/4), 407-428.

Cassella, M. (2013). Il digital curator: Tra la tutela della memoria digitale e la gestione dei dati della ricerca. *Biblioteche oggi*, 31(6), 3-10.

Choi, Y., & Rasmussen, E. (2009). What qualifications and skills are important for digital librarian positions in academic libraries? A job advertisement analysis. *The journal of academic librarianship*, 35(5), 457-467.

Choudhury, S. (2010). Data curation An ecological perspective. *College & Research Libraries News*, 71(4), 194-196.

Choudhury, S., Furlough, M., & Ray, J. (2012). Digital curation and e-publishing: Libraries make the connection.

Cragin, M. H., Smith, L. C., Palmer, C. L., & Heidorn, P. B. (2009). Extending the data curation curriculum to practicing LIS professionals. *Proceedings of DigCCurr2009 Digital Curation: Practice, Promise and Prospects*, 92.

Cunningham, A. (2008). Digital Curation/Digital Archiving: A View from the National Archives of Australia. *American Archivist* 71:2, 530-543.

Cyberinfrastructure vision for 21st century discovery. National Science Foundation, Cyberinfrastructure Council, 2007.

Davenport, T. H., & Patil, D. J. (2012). Data scientist. *Harvard business review*, 90, 70-76.

Dey, I. (1993). *Qualitative data analysis: A user friendly guide for social scientists*. Routledge.

DigiCurV (2010) Training Opportunities. Retrieved April 2015, from <http://www.digcur-education.org/eng/Trainingopportunities>

Digital Curation Centre. (n.d.). What is digital curation? Retrieved April 2015, from

- Flick, U. (2002). Qualitative research-state of the art. *Social science information*, 41(1), 5-24.
- Gold, A. (2007). Cyberinfrastructure, Data, and Libraries, Part 1: A Cyberinfrastructure Primer for Librarians. *D-Lib Magazine*, 13(9/10).
- Gold, A. K. (2007). Cyberinfrastructure, data, and libraries, part 2: Libraries and the data challenge: Roles and actions for libraries. *D-Lib Magazine*, 13(9/10)
- Gold, A. (2010). Data curation and libraries: short-term developments, long-term prospects. *Office of the Dean (Library)*, 27.
- Gordon-Murnane, L. (2012). Big data: A big opportunity for librarians. *Online*, 36(5), 30-34.
- Gray, J., Szalay, A. S., Thakar, A. R., & Stoughton, C. (2002). Online scientific data curation, publication, and archiving. *Astronomical Telescopes and Instrumentation*, 103-107.
- Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. *Handbook of qualitative research*, 2(163-194).
- Guercio, M. (2012). Le discipline del documento e l'innovazione tecnologica nelle iniziative di formazione degli archivisti. *DigItalia*, 1, 9-28.
- Harvey, R. (2010). *Digital Curation*. NY: Neal-Schuman Publishers, 39.
- Heidorn, P. B. (2011). The emerging role of libraries in data curation and e-science. *Journal of Library Administration*, 51(7-8), 662-672.
- Hey, A. J., & Trefethen, A. E. (2003). The data deluge: An e-science perspective.
- Higgins, S. (2007). Draft DCC curation lifecycle model. *International Journal of Digital Curation*, 2(2), 82-87.
- Higgins, S. (2008). The DCC curation lifecycle model. *International Journal of Digital*

Curation, 3(1), 134-140.

Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council (2009). *Harnessing the Power of Digital Data for Science and Society*

Kim, Y., Addom, B. K., & Stanton, J. K. (2011). Education for eScience professionals: Integrating data curation and cyberinfrastructure. *The International Journal of Digital Curation*, 6(1), 125-138.

Kim, J., Warga, E., & Moen, W. (2012). Digital curation in the academic library job market. *Proceedings of the American Society for Information Science and Technology*, 49(1), 1-4.

Kumar, R. (1999). *Research methodology: A step-by-step guide for beginners*. Sage Publications.

Kvale, S. (1996). *InterViews: An Introduction to Qualitative Research Interviewing* (illustrated ed.). Sage Publications

Larsen, R.L., Palmer, C., Lyon, L., Hedstrom, M., & de Roure D. (2014). Preparing the workforce for digital curation. Panel presented at the 9th International Digital Curation Conference, San Francisco

Lee, C. A., Tibbo, H. R., & Schaefer, J. C. (2007, June). Defining what digital curators do and what they need to know: the DigCCurr project. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries* (pp. 49-50). ACM.

Lee, C. A., & Tibbo, H. (2011). Where's the archivist in digital curation? Exploring the possibilities through a matrix of knowledge and skills. *Archivaria*, 72(72).

Lincoln, Y., & Guba, E. (1985). *Naturalistic inquiry*. London: Sage Publications.

Lord, P., Macdonald, A., Lyon, L., & Giaretta, D. (2004, September). From data deluge to data curation. In *Proceedings of the UK e-science All Hands meeting* (pp. 371-357).

Lyon, L. (2007). *Dealing with Data: Roles, Rights, Responsibilities and Relationships*.

Consultancy Report.

Lyon, L. (2012). The informatics transform: Re-engineering libraries for the data decade. *International Journal of Digital Curation*, 7(1), 126-138.

Lyon L., & Takeda, K. (2012). What is a data scientist? (Data scientists in the wild). *Presentation at the Microsoft eScience Workshop*, Chicago

Lyon, L., & Brenner, A. (2015). Bridging the Data Talent Gap: Positioning the iSchool as an Agent for Change. *International Journal of Digital Curation*, 10(1), 111-122.

Macdonald, A., & Lord, P. (2003). Digital Data Curation Task Force: report of the Task Force Strategy Discussion Day. JISC.

Mann, C., & Stewart, F. (2000). Internet communication and qualitative research: A handbook for researching online. Sage.

Maykut, P. S., & Morehouse, R. E. (1994). Beginning qualitative research: A philosophic and practical guide (Vol. 6). *Psychology Press*.

McHugh, M. A. (2005). Open Source for Digital Curation.

National Academy of Sciences, Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age (2009). Ensuring the integrity, accessibility, and stewardship of research data in the digital age

Open Knowledge Foundation (2012), Open Data Handbook Documentation

Patel, M., & Ball, A. (2008). Challenges and issues relating to the use of representation information for the digital curation of crystallography and engineering data. *International Journal of Digital Curation*, 3(1), 76-88.

Patton, M. Q. (2002). Qualitative evaluation and research methods. SAGE Publications, inc.

Pennock, M. (2007). Digital Curation: A life-cycle approach to managing and preserving usable

digital information. *Library & Archives*, January.

Pickard, A. (2007). *Research methods in information*. Facet publishing.

Ray, J. (2009). Sharks, digital curation, and the education of information professionals. *Museum Management and Curatorship*, 24(4), 357-368.

Ray, J. (2012). The rise of digital curation and cyberinfrastructure: From experimentation to implementation and maybe integration. *Library Hi Tech*, 30(4), 604-622.

Rusbridge, C., Burnhill, P., Ross, S., Buneman, P., Giaretta, D., Lyon, L., & Atkinson, M. (2005, June). The digital curation centre: a vision for digital curation. In *Local to Global Data Interoperability-Challenges and Technologies*, 2005 (pp. 31-41). IEEE.

Shreeves, S. L., & Cragin, M. H. (2008). Introduction: Institutional repositories: Current state and future. *Library Trends*, 57(2), 89-97.

Strauss, A. L. (1987). *Qualitative analysis for social scientists*. Cambridge University Press.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage Publications.

Stuart, D. (2010). Linked data and government data: more than mere semantics. *Online*, May/June, 36-39.

Tammaro, A. M. (2007). A curriculum for digital librarians: a reflection on the European debate. *New Library World*, 108(5/6), 229-246.

Tammaro, A. M. (2010). *Biblioteca digitale per l'informatica umanistica*.

Tammaro, A. M. (2013). Integrating Digital Curation in a Digital Library curriculum: the International Master DILL case study.

Tammaro, A. M., Madrid, M., & Casarosa, V. (2013). Digital Curators' Education: Professional Identity vs. Convergence of LAM (Libraries, Archives, Museums). *Digital Libraries and*

Archives, 184-194.

Tammaro, A. M., Ross, S., & Casarosa, V. (2014). Research Data Curator: the competencies gap. *BOBCATSSS 2014 Proceedings*, 1(1), 95-100.

Tammaro, A. M., & Casarosa, V. (2015). Data Curation Glossary: a survey on terminology and interdisciplinary perspectives.

Testoni, L. (2013). Digital curation e content curation: due risposte alla complessità dell'infosfera digitale che ci circonda, due sfide per i bibliotecari. *Bibliotime*, 16(1).

Tibbo, H. R., & Duff, W. (2008). Toward a digital curation curriculum for museum studies: A North American perspective. In *Annual Conference of CIDOC [en línea]*, Athens, September (pp. 15-18).

Vivarelli, M., Cassella, M., & Valacchi, F. (2013). The digital curator between continuity and change: developing a training course at the University of Turin.

Walters, T., & Skinner, K. (2011). New roles for new times: Digital curation for preservation. *Association of Research Libraries*.

Wenger, E. (2006). Communities of practice: A brief introduction.

Wenger, E. C., & Snyder, W. M. (2000). Communities of practice: The organizational frontier. *Harvard business review*, 78(1), 139-146.

Williams, P., Leighton John, J., & Rowland, I. (2009, July). The personal curation of digital objects: A lifecycle approach. In *Aslib Proceedings* (Vol. 61, No. 4, pp. 340-363). Emerald Group Publishing Limited.

Witt, M., Carlson, J., Brandt, D. S., & Cragin, M. H. (2009). Constructing data curation profiles. *International Journal of Digital Curation*, 4(3), 93-103.

Yakel, E. (2007). Digital curation. OCLC Systems & Services: *International digital library perspectives*, 23(4), 335-340.

Yakel, E., Conway, P., Hedstrom, M., & Wallace, D. (2011). Digital curation for digital natives. *Journal of Education for Library and Information Science*, 23-31.

Yin, R. (2002). *Case study research: Design and methods*. Sage Publications.

Appendix 1: Interview guide

- 1) Talking about your working environment, how would you sketch out a curriculum for the creation of digital/data curators in Italy? How much importance would you give to technical, managerial and collaborative competencies?
- 2) On the basis of your professional experience, what are the essential technical competencies that a good digital/data curator must have?
- 3) Carrying on with technology, a key informant has told me that IT skills and competencies of the digital curator should be replaced, one day, by the technology itself. In other words, that the technical contribution of the digital curator should soon disappear. In your opinion, how important is the human contribution within the technical competencies?
- 4) The aim of my research is to establish whether or not the IT skills and competencies of the digital/data curator should be the basis for the development of different and related professions. Do you think that data curator could specialize in different professions, and be placed in different workplaces? Could you please make some examples?
- 5) In this perspective, another objective of my research is that the digital curator is able to add value to the data lifecycle and produce new patterns of knowledge among data and datasets. In other words, he or she is not only able to manage and preserve data (working in back end) but also to work in front end and to produce new content. Do you agree with this statement?
- 6) now we shift the attention on the cyberinfrastructure. How do you evaluate the cyberinfrastructure in the italian context, both globally (global cyberinfrastructure) and nationally (national cyberinfrastructure)?
- 7) In this perspective, how do you evaluate the level of collaboration among institutions in the italian context, both technically (data interoperability) and ideologically (Open Data movements)?

Appendix 2: Information letter and consent form



UNIVERSITÀ DEGLI STUDI DI PARMA

Information Letter and Consent Form for Invitation to be Interviewed

Date:

Dear XXX,

This letter is an invitation to consider participating in a study I am conducting as part of my Master's degree DILL Digital Library learning in the Department of Information Engineering at the University of Parma under the supervision of Prof. Anna Maria Tammaro. I would like to provide you with more information about this project and what your involvement would entail if you decide to take part.

The aim of my research is to analyse the Digital Curation activity in the Italian context, focusing the attention on the IT skills and competencies of the digital curator. I take the DCC Data Lifecycle Model as theoretical framework and as a starting point to individuate key services that help to define a specific cyberinfrastructure in this context and help digital curators to do his or her job (such as Linked Data, Semantic Web, Ontologies, Data visualization etc.).

This would lead me to define the basic IT skills and competencies that are necessary for this profession.

The objective is to demonstrate that Digital Librarians can play an active role in data curation and add value during the whole data lifecycle. In other words, digital librarians can now be part of the both back-end and front-end activity.

I would like to include your organization as one of several organizations to be involved in my study. I believe that because you are actively involved in the data curation of your organization, you are best suited to speak to the various issues, such as open and linked data, data management, cyberinfrastructure.

Participation in this study is voluntary. It will involve an interview of approximately 30-45 minutes in length to take place virtually. You may decline to answer any of the interview

questions if you so wish. Further, you may decide to withdraw from this study at any time without any negative consequences by advising the researcher. With your permission, the interview will be tape-recorded to facilitate collection of information, and later transcribed for analysis. Shortly after the interview has been completed, I will send you a copy of the transcript to give you an opportunity to confirm the accuracy of our conversation and to add or clarify any points that you wish. All information you provide is considered completely confidential. Your name will not appear in any thesis or report resulting from this study, however, with your permission anonymous quotations may be used. Data collected during this study will be retained for one year in locked office in my supervisor's lab. Only researchers associated with this project will have access. There are no known or anticipated risks to you as a participant in this study.

If you have any questions regarding this study, or would like additional information to assist you in reaching a decision about participation, please contact me at +39 329.3466525 or by e-mail at pol.brambilla@gmail.com. You can also contact my supervisor, Anna Maria Tammaro at the e-mail annamaria.tammaro@unipr.it. I very much look forward to speaking with you and thank you in advance for your assistance in this project.

Sincerely,

Paolo Brambilla
International Master in Digital Library Learning
Parma University
Parma – Italy

CONSENT FORM

I have read the information presented in the information letter about a study being conducted by Paolo Brambilla of the Department of Information Engineering at University of Parma. I have had the opportunity to ask any questions related to this study, to receive satisfactory answers to my questions, and any additional details I wanted.

I am aware that I have the option of allowing my interview to be tape recorded to ensure an accurate recording of my responses.

I am also aware that excerpts from the interview may be included in the dissertation and/or publications to come from this research, with the understanding that the quotations will be anonymous.

I was informed that I may withdraw my consent at any time without penalty by advising the researcher.

With full knowledge of all foregoing, I agree, of my own free will, to participate in this study.

___ YES ___ NO

I agree to have my interview tape recorded.

___ YES ___ NO

I agree to the use of anonymous quotations in any thesis or publication that comes of this research.

___ YES ___ NO

Participant's Name (please print) _____

Participant's Signature _____ Date _____

Researcher's Signature _____ Date _____

Researcher's Title _____ Department _____

Faculty Advisor Signature _____ Date _____

Faculty Advisor Title _____ Department _____