The *data librarian*: myth, reality or utopia?

Silvia Giannini, Anna Molino
CNR, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", Italy
[silvia.giannini, anna.molino]@isti.cnr.it

1. Introduction

It is undoubtedly true that we live in a more and more data-centric world (Cassella, 2016). As citizens and users of the Internet, we produce an enormous amount of data on a daily basis. This may be done consciously (e.g. when we deal with governmental and public organizations, universities and research centers, structured companies, etc.), as well as less intentionally, for instance when we use social networks, and, more generally, as users of the Internet for the widest variety of purposes.

In this scenario, data is becoming of greater importance not only for the average user when it helps making choices and decisions, but mainly for the growing number of companies and entities worldwide founding their activity in collecting and elaborating the exponentially increasing amount of information produced by the citizenry.

Therefore, we are now facing and coping with what has been defined by many commentators a "data deluge".

In the academic and, more generally, scientific world the need for management, long term preservation, and storage of research data has grown exponentially in the recent past. As the Turing award winner Jim Gray states: "a fourth paradigm [of science] is emerging, consisting of techniques and technologies needed to perform data-intensive science" (Gray, Szalay, 2007). Indeed, in almost all discipline areas "born digital" documents proliferate as files, spreadsheets, databases, digital notebooks, wikis, etc. As a consequence, the management, curation, and archiving of such data are becoming of crucial importance (Bell, Hey and Szalay, 2009).

In this context, the expressions *eScience* and *eResearch*[1] have been identified as umbrella terms describing converging sets of trends and technologies that are radically changing the way science is conduced. Librarians may bring important knowledge and skills complementary to the activities carried on by the eScience community, most notably in the management, preservation and archiving of information (Wright et al., 2007). Moreover, another essential component of eScience concerns the management of the scholarly communication lifecycle, being this one of the most prominent area of interest in the work performed by libraries.

The Open Science (OS) movement, in turn, is radically changing the perspectives adopted in the scientific production and dissemination, fostering new approaches to research and scholarly communication. The movement is growing considerably in academia and among scientists worldwide. Two fundamental aspects of OS are the Open Access (OA) to scientific publication and the possibility of discovery, sharing and exploit the data used for or produced during the research process. The need for creating Open Data (OD) is profoundly changing the perspectives adopted by researchers during the scientific production, as research data is increasingly recognized as a primary research output.

As far as OD is concerned, firstly it must be pointed out that after the regulation of the Open Access to scientific literature, the legal framework in Europe is recently supporting the approach to OA.

[1] *eScience* is the term preferred in Europe, while in other countries (e.g. Australia) the initiatives aiming at transforming the approach to science are labelled as *eResearch* (cfr. Wright et al., 2007).

Indeed, last European Recommendations (April 25 2018)[2] ask Member States to ensure that data management planning becomes a scientific practice and since 2017 the European Commission has made mandatory to open research data for all participants in Horizon 2020 and for any subject area, provided this is allowed from a legal or ethical point of view.
Thus, not surprisingly various funding bodies (e.g. the European Commission, the Wellcome Trust and the RCUK in UK, the Australian Research Council in Australia, or the National Institutes of Health in the U.S.) have been making mandatory the submission of a Data Management Plan (DMP) together with the project proposal. For instance, in the article 29.3 of the Annotated Model Grant Agreement[3], the EU asks all consortia submitting a proposal in H2020 program to declare: the type of data that would be produced during the project; the strategies for their management in order to guarantee their short- and long-term preservation; how much of the produced data would be openly available.
In this perspective, academic libraries are indeed increasingly involved in the management of research data across the lifecycle (Schmidt and Shearer, 2016), actively participating in tasks such as providing access to data, supporting researchers in managing their data and drafting DMPs, as well as managing data collections.
Given this context, we may ask ourselves: who is currently responsible for the management, curation, and archiving of (research) data? Fearon et al. (2013) observes that:
"the data management space in US in higher education is predominantly owned by the libraries […], whereas in the UK it is much more dependent on individual institutional cultures and circumstances whether it is the librarians, the academics, or the administrators who take the lead".
The Research Data Alliance (RDA)[4] recognizes that: "Many academic libraries are now extending their century-long track record in the professional management of knowledge resources towards the area of research data and therefore seek to maximize research data skills among staff in their organizations". They identify five main routes to achieve such goal, consisting in: training, expert recruitment, learning-on-the-job, online-courses, and (academic) degrees.
Swan and Brown (2008) in their report commissioned by the UK Joint Information Systems Committee (JISC) recognize a strategic role for libraries in data management, identifying three main potential roles: increasing data awareness; providing archiving and preservation services; developing a new professional strand of practice as data librarianship.
Many commentators have argued that the background knowledge of librarians may be essential in this scenario. For instance, the management of repository's contents may be seen as a *collection management issue* (Genoni, 2004), while the expertise in classification and description through cataloguing and metadata, as well as the experience in the selection of the information may be crucial for data curation (Witt, 2008).
Starting from this framework, we can consider academic libraries as "aggregator, collector and curator of external scholarship, be it printed or online". For this reason, in our work we will try to get an idea of the competencies required to a Data Curator, giving an overview of the features of Research Data Management and its conversion into an effective service (RDS), both from a theoretical point of view and through the observation of some concrete examples of data management by librarians. The aim is to understand if librarians are the most accredited candidates to fill the role of Data Curator, giving possible answers and outlining specific qualifications required to those currently operating in academic libraries with the purpose of possibly identifying the figure of *data librarian*.

[2] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32018H0790&from=EN
[3] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf.
[4] Research Data Alliance – Libraries for Research Data IG (2015). *How to maximize research data skills in libraries*, https://www.rd-alliance.org/how-maximize-research-data-skills-libraries.html.

2. What is (research) data?

2.1. Data: definitions and types

As anticipated in the introduction, the contemporary world may be defined as a data-centric reality, due to the huge amount of data we use and produce in our everyday life.
But what is the meaning of the word *data*?
The broader definition of the term given by the Cambridge Dictionary[5] is: "information collected for use"; more specifically, data can be seen as "information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer".
The Oxford English Dictionary[6] generally speaks about: "facts and statistics collected together for reference or analysis", providing the more discipline-oriented sub-definition: "The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media".
Merriam-Webster[7] specifies three different connotations of the word, seeing data as: "factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation"; "information in digital form that can be transmitted or processed"; "information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful".
It is noticeable that all three dictionaries cited take into account the shades of meaning concerning information science and technology, making reference to what is relevant for data-analysis as specific area of interest.
It is possible to identify different types of data, based on their exploitation, purpose, etc., being them identified and categorized in a more or less discipline as well as use oriented.
In our work, we focus on a specific typology of data, namely *research data*. The following subsection is dedicated to the description of this category, trying to outline a possible definition embracing the quantity of information produced in this specific context.

2.2. Research Data

The definition of research data may be quite broad and open to different meanings, which can vary depending on the disciplinary field we consider. Indeed, what a researcher considers to be *research data* depends on the meaning of this data in the research process and this may differ for each scientific discipline. Therefore, different definitions of this concept have been suggested by various actors and entities operating in this field. Here we report those we find more significant, proposed by the "Essentials 4 Data Support" online course of the Research Data Netherlands (RDNL) [8].
The Queensland University of Technology in its *Manual of Policies and Procedures* defines research data[9] as: "data in the form of facts, observations, images, computer program results, recordings, measurements or experiences on which an argument, theory, test or hypothesis, or other research output is based. It relates to data generated, collected, or used, during research projects, and in some cases, may include the research output itself. Data may be numerical, descriptive, visual or tactile. It may be raw, cleaned or processed, and may be held in any format or media […]".

[5] https://dictionary.cambridge.org/dictionary/english/data
[6] https://en.oxforddictionaries.com/definition/data
[7] https://www.merriam-webster.com/dictionary/data
[8] https://datasupport.researchdata.nl/en/start-the-course/i-definitions/research-data/
[9] http://www.mopp.qut.edu.au/D/D_02_08.jsp

The UK Engineering and Physical Sciences Research Council (EPSRC)[10] consider research data as: "recorded factual material commonly retained by and accepted in the scientific community as necessary to validate research findings; although the majority of such data is created in digital format, all research data is included irrespective of the format in which it is created".
Crossley (2004) in her "Introduction to managing research data"[11] argues that they are "collected, observed or created for the purpose of analysis to produce and validate original research results".
Indeed, it must be taken into account that research data may be presented in a variety of formats, both digital or physical, e.g. electronic text documents; spreadsheets; laboratory notebooks, field notebooks, and diaries; audiotapes and videotapes; specimens, samples, and artefacts; methodologies, workflows, standard operating procedures and protocols; metadata, and so on (Scott & Cox, 2016).
Overall, the definition we prefer for its conciseness and effectiveness is the more general one, saying: "Research data is the material underpinning a research assertion", making reference to all the outcomes produced in the course of the research, from statistics to field observations and answers to questionnaires, in spite of their formats or media.

2.3. Research Data as Open Data

In the context of Open Science, the possibility of making research data as open as possible becomes of crucial importance. Indeed, there are considerable advantages in sharing materials supporting the research: sharing the research outcomes encourages the cooperation between scientific communities and favors a faster and more efficient research process, as it avoids useless data duplication and stipulates the collaboration between institutions and with the citizenry.
This may be accomplished following a series of practices and principles helping the scientific community in the correct production and reuse of the research results. For instance, the main goal of FORCE 11[12], a community of scientists, librarians, archivists, publishers and funders, is the promotion of the FAIR data principles. The acronym means *Findable, Accessible, Interoperable, Reusable* and it corresponds to a set of guidelines that enables a better realization and sharing of the data.
This brings to light the crucial issue concerning digital and, more specifically, data curation. The affirmation of digital products and services has in fact brought with it a set of strategies, technological approaches and activities that have taken the name of *Digital Curation*.
Digital Curation can be considered a transversal activity to various fields consisting in the creation, the maintenance and the preservation of a digital object throughout its lifecycle. The active management of research data reduces threats to their research value and mitigates the risk of digital obsolescence, enhancing the long-term value of existing data by making it available for further high quality research.
Digital curation and data preservation are ongoing processes, requiring considerable thought and the investment of adequate time and resources. This is the reason why the DCC[13] in UK has identified some steps to be followed during what has been named *digital curation lifecycle*, as represented in the picture below:

[10] https://epsrc.ukri.org/about/standards/researchdata/scope/
[11] https://www.scribd.com/presentation/138079216/Managing-Research-Data
[12] https://www.force11.org/group/fairgroup/fairprinciples
[13] http://www.dcc.ac.uk/digital-curation/what-digital-curation

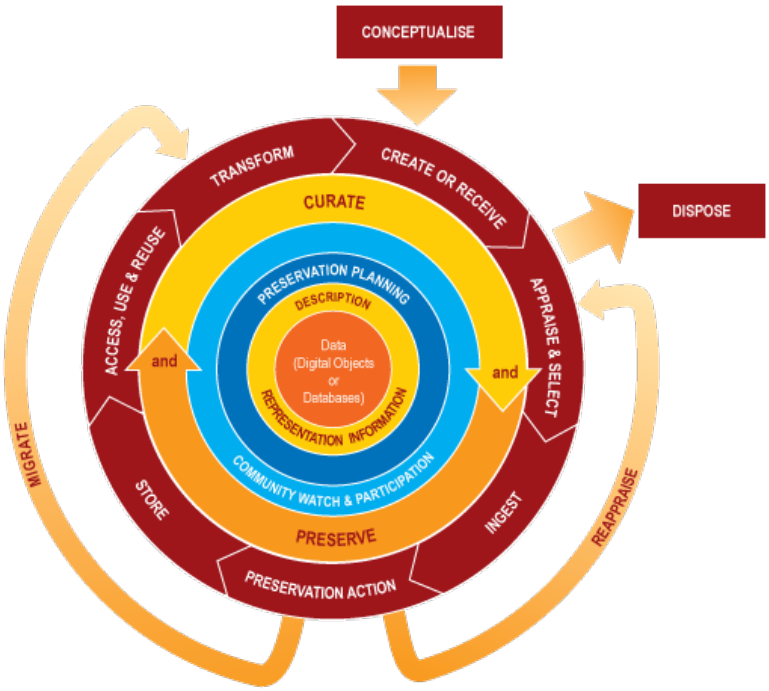

Figure 1: Digital Curation Lifecycle

The need to identify a professional figure who manage and store the growing amount of data in digital format has generated the role of *digital curator*. At the present time, this may not be considered a structured and well-defined character; as a consequence, the breadth of the definition implies that its reference community may include different actors and professional figures.
The concept of Data Curation can be seen as a sort of subset of the Digital Curation. Strictly connected with the academic and research world, it comes from the connection of the digital curation with the development and management of Open Access repositories. At the same time, the data curator becomes a specialization of the digital curator.
The reference community of Data Curation is very often limited to the researchers and the type of data taken into account are the research data associated to the scientific literature.
Thus, the professional figures move towards the librarians or, more exactly, towards the upcoming figure of the *data librarian*, for their wide experience in different disciplinary domains, their skills in the management of metadata sets, in the maintenance of collections and their involvement in the management of research information.
The following paragraph is dedicated to Research Data Management, giving particular attention to its transformation into a specific service granted by some established research entities, namely Research Data Service.

3. Research Data Management and Research Data Service

3.1. What is Research Data Management (RDM)?

Research Data Management (RDM) is a general term to indicate a set of good practices concerning collecting, storing, using, sharing and preserving research data in an effective and productive manner. It involves services, tools and infrastructures that support the management of research data, which may significantly differ across the lifecycle. (Schmidt and Shearer, 2016; Schmidt et al., 2016).
Whyte and Tedds (2011) in their Briefing Paper for the DCC clarify some terminological distinctions between research data management, preservation and curation, arguing that: “Research data management concerns the organization of data, from its entry to the research cycle through to the

dissemination and archiving of valuable results. It aims to ensure reliable verification of results, and permits new and innovative research built on existing information. Preservation is about ensuring that what is handed over to a repository or publisher remains fit for secondary use in the longer term (e.g. 10 years post-project). Curation connects first use to secondary use. It is about ensuring that project results are fit to archive, and that valued research assets remain fit for reuse".
The various aspects of RDM should be seen as research support services distributed across various departments (e.g. Research Offices, IT Services, Libraries), as researchers need support in different areas, such as planning, organizing, documenting and sharing, preparing datasets for deposit and long-term preservation, not forgetting copyright issues (Schmidt and Shearer, 2016).
Therefore, RDM involves a wide range of activities across the data lifecycle, requiring a high level of interaction with both researchers and other support services (e.g. technical services and research officers), such as creating and collecting, processing, analyzing, publishing, archiving and preserving, and re-using data (Schmidt et al., 2016).
Figure 2 illustrates the major steps of the RDM lifecycle, highlighting the common points that might be shared by different scientific communities and realities worldwide:

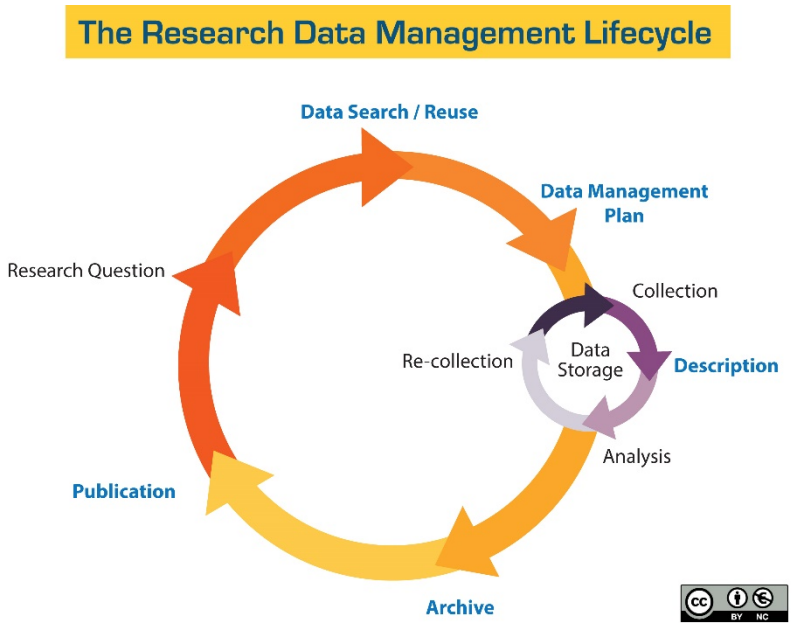

Figure 2: Research Data Management Lifecycle, diagram (University of California)

As for the Research Data Netherlands[14], those operating in RDM should guarantee support within the following influence spheres:

- Legislation and policy;
- Technological infrastructure;
- Culture (i.e. research practices and their exploitation);
- Knowledge of storing, managing, archiving and sharing research data by both researchers and support services;
- Skills (e.g. conversational and influence skills);
- Motivation for collecting and managing research data.

[14] https://datasupport.researchdata.nl/en/start-the-course/vi-data-support/influence-sphere/

3.2. Libraries' competencies in RDM

As anticipated above, the traditional competencies of librarians may be well-versed in RDM, due to the broad experience in wide-ranging disciplinary domains, the practice in the management of metadata sets, as well as in the creation and preservation of collections. More in general, we can recognize that librarians have always been familiar in dealing with data.

In their Final Report on Research Data Management (2012), the LIBER Working Group on eScience identified *Ten recommendations for libraries to get started with research data management*, underlining the crucial role of librarians in offering support in data management, in the development of metadata and data standards, their participation in the elaboration of institutional policies for data administration, as well as in assisting in the exploitation of interoperable infrastructures granting access and storage to research data (e.g. via the application of persistent identifiers), being possibly involved in subject specific RDM practices.

Schmidt et al. (2016) identifies three main groups into which library services for RDM can be broadly categorized: provide access to data; awareness and support to students and researchers in handling data; managing data collections. Despite the possible overlaps between them, authors also identified distinctive roles for librarians inside each area. Indeed, the first one reflects more traditional library services (e.g. consultation and reference for datasets); the second involves hands on support for researchers across the data lifecycle, e. g. in policy and advocacy on RDM and data sharing, as well as training. The third and final category includes the preparation of data for deposit, the management of metadata, and data preservation activities (Schmidt and Shearer, 2016).

Moreover, libraries have also the opportunity to act as point of contact with the public audience, supporting public engagement with science and acting as a hub to collate links and information about citizen science activities, as well as researchers advocating the use of guidelines and templates. In addition to this, librarians are crucially involved in the training of scientists in data management and reuse (Lyon 2012).

This favors the evolvement of RDM practices into Research Data Services (RDS), which find numerous and different applications throughout academic and, more in general, research realities around the world.

3.3. Research Data Management as Research Data Service

The fundamental importance acquired by data collection and reuse in the research process, as well as the establishment of data management mandates by funding bodies have motivated research libraries to develop a set of services that may be generally labelled as Research Data Services (RDS).

Indeed, RDM acquires substance when it becomes a service, i.e. when an infrastructure made up of people and tools is established, providing assistance and advice for RDM practices supporting all the phases constituting a research project.

The results of the studies conducted by Tenopir et al. (2014) on U.S. and Canadian research libraries shows that the provision of RDS would augment institutions' research impact as well as the perception of the library in terms of relevance and prestige, RDS fitting the traditional role of librarians as "stewards of scholarship". At the time of their investigation, the most common services provided by academic libraries in U.S. and Canada could be seen as "extensions" of familiar reference services into the realm of data, dealing mostly with access to and citation of datasets (Tenopir et al., 2014).

We identified some common basic services provided nowadays by research libraries participating in RDS; generally speaking, they are primarily involved in:

- Support in Data Management Plan (DMP) development;
- Digital Curation, i.e. data selection, preservation, maintenance, and archiving;
- Metadata creation and transformation.

Another fundamental aspect that must be taken into account when speaking about established RDS in universities and research bodies worldwide, is the presence of a formal RDM Policy, being this at institutional level (as in the case of many universities in UK, e.g. University of Edinburgh) or presented as codes of conduct at national level, as in the case of the *Australian Code for the Responsible Conduct of Research*, which covers a wide range of topics associated with research including the management of research data and associated materials[15]. The availability of such policies endorses the formal establishment of the roles involved in RDS, regulating the smooth functioning of the research support in each institution and contributing to the definition of upcoming figures such as the *data librarian*.
Finally, we can argue that RDS may be reasonably conceived as the conversion of RDM concepts into concrete practices. In order to validate such assertion, we reviewed some literature about RDS experiences worldwide and selected three case studies that we propose in the following paragraph. This helped us to understand how this process turns into reality.

4. Case Studies

4.1. University of Edinburgh Research Data Service (RDS)

The first case study we analyzed is the Research Data Service provided by the University of Edinburgh[16]. The aim of the service is to provide tailored support to researchers dealing with the production and/or reuse of research data, offering tools, support and training to university staff and students.
The Research Data Service, led by a *data librarian* whose background is in library services, is part of the Information Services of the University, whose experts contribute in delivering specific tools and software components for the management of the data produced during the research lifecycle. The main goal is to provide tailored services for researchers aiming to achieve good practices in RDM, according to their specific needs. Moreover, at any time of such process people working with research data may ask support for training, which will be delivered following specific programs.
It must be pointed out that the University of Edinburgh has a formal Policy for RDM, establishing that data must be "managed to the highest standards as part of the University's commitment to research excellence", granting that data will be made available as open as possible, protecting those considered as sensitive and giving the widest outreach to those that may be of public interest. Indeed, the last document's clause clearly states that: "Exclusive rights to reuse or publish research data should not be handed over to commercial publishers or agents without retaining the rights to make the data openly available for re-use, unless this is a condition of funding" [17].
The support offered may be divided into three major phases: planning, active research project phase, project conclusion. These represent the milestones of the project lifecycle and, subsequently, of the assistance provided by RDS.

Following this pattern, the RDM services delivered at the University of Edinburgh may be schematized as follows:

1) *Before*: this consists of the identification of existing datasets; the planning for the data collection and storage; the identification of possible sensitive data, as well as methods and terms for data sharing. We can reasonably argue that the crucial part of this step regards the creation of a Data Management Plan (DMP), required by the majority of funding bodies and universities, which

[15] https://www.ands.org.au/guides/code-awareness
[16] https://www.ed.ac.uk/information-services/research-support/research-data-service/about-the-research-data-service
[17] https://www.ed.ac.uk/information-services/about/policies-and-regulations/research-data-policy

accompanies the research project during its lifetime. The RDS at the University of Edinburgh assist researchers in its development, providing either tools and templates (e.g. DMPonline[18]) or personal consultation to discuss the DMP in detail and obtain expert advice.

2) *During*: in the active development of the project, consultation is offered about finding existing datasets containing data that might be reused and re-elaborated, some being freely available, others behind paywall. The Data Library here plays a crucial role, giving advices about possibility for exploitation by users, as well as helping researchers in the selection of the data resources based on their type (e.g. surveys, censuses, databases, etc.). Other contributions offered by the RDS regard solutions for data storage during the project lifetime, for the control and safeguard of sensitive data, for the sharing and versioning of data, keeping track of the changes while working with other researchers or research teams.

3) *After*: after the conclusion of the research project, RDS grant assistance in recording, sharing and archiving research data for the long-term. This is made via a set of specific tools recording descriptive metadata, providing storage in an open repository for the online discovery and re-use through the association of a persistent identifier (DOI) to researchers' data resources, and securing long-term archiving in order to keep data safe from accidental deletion or inappropriate access, meeting possible funders' requirements.

Besides RDM consultancy and support, particular attention is given to training. The RDS at The University of Edinburgh offer a wide range of courses for those unfamiliar with the fundamentals of research data management and sharing, in the form of online courses, classroom-based workshops and seminars.
Indeed, people dealing with research data have the possibility to select a suitable training option among the variety proposed, depending on their specific necessities.
For instance, a free five-week MOOC - created by the Universities of Edinburgh and North Carolina – has been designed to reach learners of various types across disciplines and continents. The subjects covered are: understanding research data; data management planning; working with data; sharing data; archiving data, following the stages of a generic research project.
In addition, a free, online course named MANTRA[19] has been realized with the purpose of understanding and reflecting on the management of the data collected throughout the research. This is particularly referred to post-graduate students, early career researchers, and also information professionals. It is composed of a series of interactive online units concentrating on the explanation of the terminology, key concepts, and best practices in RDM.
Inside this training path, a special focus is dedicated to librarians, underlining the central role of such figure in the RDM workflow. Indeed, a *Do-It-Yourself Research Data Management Training Kit for Librarians*[20] has been created in order to supply to the needs of academic liaison librarians. It is provided by EDINA and Data Library, University of Edinburgh, in association with the UK Data Archive, Digital Curation Centre (DCC), and Distributed Data Curation Center at the Purdue University Libraries. After an introductory "pre-training", the course is divided into five main sections, each containing a wide range of materials (e.g. podcasts, presentations, assignments, etc.) and specifically concentrated on: data management planning; organizing and documenting data; data storage and security; ethics and copyright; data sharing. Moreover, materials for post-training study are available, such as the *Data Curation Profiles,* which provide a complete framework for interviewing a researcher in any discipline about their research data and their data management

[18] https://dmponline.dcc.ac.uk
[19] https://mantra.edina.ac.uk/
[20] https://mantra.edina.ac.uk/libtraining.html

practices, giving practical overviews on what a librarian involved in RDM would face in the daily practice.
Finally, also bespoke group training and face-to-face classes and workshops are planned upon requests or periodically.

4.2. The Research Data Netherlands Front Office – Back Office model

The second case study we have taken into consideration regards the Front Office – Back Office Model of the Research Data Netherlands (RDNL FO-BO Model).
It consists of a federated infrastructure handled by DANS, 3TU.Datacentrum and SURFsara - the three organizations constituting RDNL, a coalition joining three archives in the area of long-term archiving, also open to other third parties - and modelled into a four-layer structure:

1) a basic technical infrastructure under the computer centers responsibility;
2) back office data services, providing facilities for long-term archiving and accessibility;
3) front office services, granting support and training to researchers and students in responsible data management;
4) data generators and data users.

The model is graphically represented in figure 3 below:

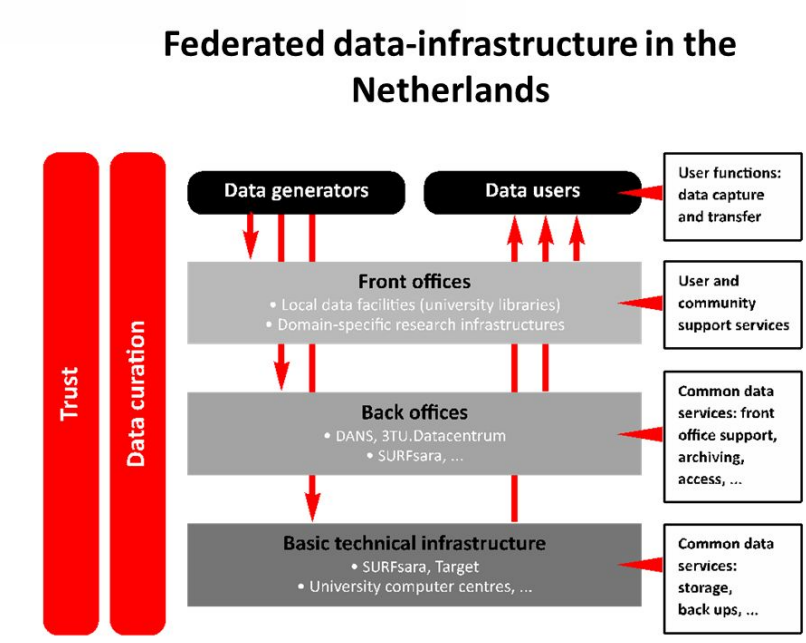

Figure 3: The federated data-infrastructure in the Netherlands

RDNL infrastructure moves from the model proposed in the EU *Riding the wave* report (2010), giving a general impression of how the various actors, data types and services should be interconnected in a global e-infrastructure for science. It must be also noticed that Netherlands have recently published the *Netherlands Code of Conduct for Research Integrity* (2018), which delineate specific directions for data management (pp. 20-21).
Globally, the services provided fall into three main categories: information provision, training, and data curation, management and storage.
More in details, the front-offices are situated mainly locally, in most of the cases in research libraries; their focus is on supporting their own research organizations, being primarily responsible for the quality assurance of the data produced.

The front-office is accountable for data collection and acquisition, as well as awareness raising in best practices for data management within its research community of reference, providing information and training to their research personnel. Moreover, the front office hosts so-called Virtual Research Environments or Data Labs, offering research tools and securing mid-term storage facilities for the organization's researchers. In consultation with the back office, the front office also facilitates the transfer of data to a trusted back-office digital repository after the research has been completed. Facilities that are shared by several universities can be hosted and supported by the back offices.
The back-offices are constituted mainly of computer experts, who provide data stewardship, guaranteeing long-term storage and accessibility to the collected data. Their functions are performed mainly by organizations such as DANS, 3TU.Datacentrum, and SURFsara, having a nationwide coverage and expertise on data from various discipline areas (humanities, social studies and sciences).
The back-office also provides consultation, training and support to front-office employees, acting as a center of expertise and innovation. Its fundamental duty is to ensure a sustainable and secure storage and retrieval upon completion of the research project.
It must be pointed out that since the division of labor between front-office and back-office is not always necessarily sharp, especially because organizations may differ in size, staffing capacity, etc. there might be institutes performing front-office tasks only, while outsourcing the back-office ones to a data archive. In this perspective, we can see how the figure of the librarian in this kind of RDM model is always present and plays a relevant role.

4.3. eResearch at Griffith University

The third and final case study we analyzed regards the activities conducted by the eResearch unit at Griffith University (Australia), where the Division of Information Services (INS) integrates what they have named e-research, library, and information and communication technology into a single organization (Brown et al., 2015).
As in the previous case studies, Griffith University responds to the *Australian Code for the Responsible Conduct of Research* (2007) - assigning researchers and their institutions a shared responsibility to manage research data and primary materials well – and to the *Griffith University Code for the Responsible Conduct of Research* (2012). As stated in this policy, researchers are required to manage their data – using methods appropriate to the discipline and to the nature of the data – to the highest standards, while the University is required to provide infrastructures, opportunities to develop professional skills, and access to advice and expertise that enable researchers to meet these standards. Finally, the *Best practice guidelines for researchers: Managing research data and primary materials* (Richardson, 2016) were published, aiming at expanding the Code for the Responsible Conduct of Research in relation to specific aspects of research data management; outlining practical steps that researchers can take; highlighting technology, advisory services and professional enabling development opportunities.
In this context, the first thing to point out is that faculty librarian roles have been modified to address data management. Indeed, while traditional librarian's duties are kept, and their core capabilities (e.g. structured thinking, knowledge of information management theory, etc.) are considered of great value for positioning them in the process of research data management (Brown et al., 2015), the librarians operating in the eResearch team are required to develop additional, more technical and discipline-oriented skills. More in detail, on the one hand librarians within the INS portfolios of Library and Learning Services and Information Management support researchers in well-established areas such as acquisitions, collection development, copyright advice and information literacy training, and are moving into newer areas such as open access advocacy,

publications repositories, research assessment exercises, and bibliometrics. On the other, Griffith's eResearch Services team operates within another portfolio in INS, building and managing technical research infrastructure for supporting researchers. Part of this infrastructure is targeted at specific research needs, while some is associated with university-wide management and discovery. Therefore, librarians in eResearch must explicitly demonstrate how their skills can be combined in productive ways with technical specialties, including software development and business analysis (Simons and Searle, 2014), as they work in close relationship with ICT experts and, for this reason, need to undergo specific and constant training to develop the specific and necessary skills to supply RDM demands.

Librarians working in this unit tend to consider themselves as "generalists", as they are required to have such a broad range of skills, knowledge and expertise that is difficult to acquire a specialization in any of these (Simons and Searle, 2014).

However, as anticipated above, a core set of skills and knowledge for librarians have been identified. As far as skills are concerned, they can be considered a sort of enhancement of the traditional librarian's competences. In fact, they deal with advanced metadata skills, high level communication skills, as in such context the role of the librarian foresees the "translation" of information between research groups. In addition, high level documentation skills are necessary for the production of documentation addressing a wide variety of audiences and purposes.

With respect to core knowledges, there must be a deep knowledge of the broader research environment where the librarian is acting, as well as of the mechanisms and processes of scholarly communication, and of the legal and regulatory framework, concerning mainly contract law and copyright issues, with a specific focus on licensing and data re-use.

Generic technical and managerial skills also play a distinctive role, as research teams usually work on goal-oriented projects, and since in eResearch project teams are comprised largely of software developers.

As far as we understand, the librarian acts as an advisor, even though technical skills make the difference in understanding how eResearch projects are run and in liaising with researchers and research managers.

5. The *data librarian* profile

In the light of the case studies, we tried to identify some basic characteristics of the *data librarian* profile.

Schmidt et al. (2016) briefly describe this upcoming figure in the research scenario as consisting of "Traditional librarian competences and skills into renewed organizational structures". Authors like Sada et al. (2013) highlight necessary technical skills, underlining the fact that many of the competencies of such professional are adopted from the ICT domain. In this perspective, the *data librarian* is seen mainly as responsible for the implementation of collaborative infrastructures for data access and reuse, fostering protocols for data interoperability and dedicating special attention to digital preservation.

Indeed, as argued by Cassella (2016) and anticipated previously, just a few steps separate the *digital curator* from the *data librarian*, the latest being a RDM specialist who constantly collaborates with other professionals.

In her presentation for the 2nd DCC/RIN Research Data Management forum, Rice (2008) defines the *data librarian*s as "people originating from the library community, trained and specializing in the curation, preservation and archiving of data". As a follow-up of the same event, the diagram reported below has been published in the DCC Data Management Forum[21]:

[21] http://data-forum.blogspot.com/2008/12/rdmf2-core-skills-diagram.html

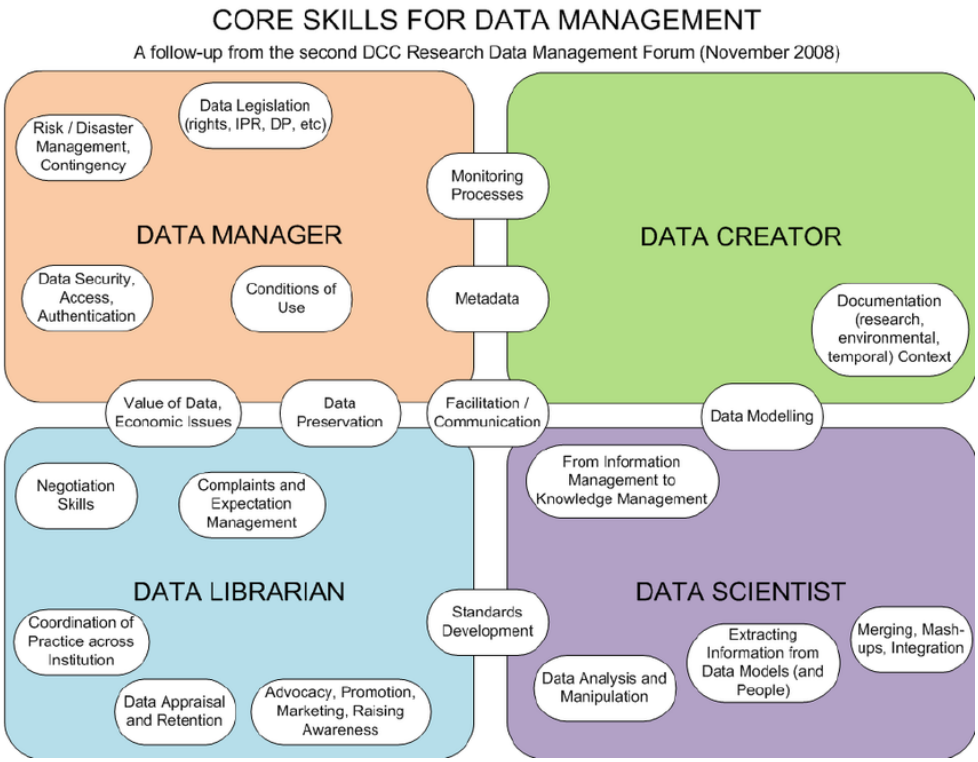

Figure 4: Core skills for data management. Chris Rusbridge and Martin Donnelly, November 2008.

The picture schematically describes the fundamental characteristics and tasks performed by the professional figures operating with data. Even though making a clear distinction between the four, a number of tasks overlap between them, as in the case of the preservation of data or the development of standards. It emerges from this model that at the time the major characteristics of the *data librarian* consisted in:

- Skills in communication and facilitation
- Standards development
- Data selection and evaluation
- Negotiation skills
- Advocacy, promotion and marketing
- Economic issues related to data value
- Preservation
- Complaints and expectations management

Despite its clarity and effectiveness, this model has undergone some critics regarding, for instance, the singular choice of placing the specialization in metadata at the boundary between the roles of data manager and data creator. Another remark concerns the absence of any reference to training for all data professionals and for *data librarian*s in particular (Cassella, 2016), which, as we reported in the previous paragraph, is now considered as crucial, thus representing a fundamental aspect of any RDS.

In this perspective, Carlson et al. (2011) identified a list of twelve core educational objectives for a data information literacy program, which might be of interest also for librarians approaching RDM:

- Databases and Data Formats
- Discovery and Acquisition of Data
- Data Management and Organization
- Data Conversion and Interoperability
- Quality Assurance

- Metadata
- Data Curation and Re-use
- Cultures of Practice
- Data Preservation
- Data Analysis
- Data Visualization
- Ethics, including citation of data

In the following years, the debate as well as the reality concerning professional roles operating with data has had various developments, as we illustrated in the case studies. Indeed, RDM policies and practices have evolved through time, allowing the growth of RDS in many academic realities.
In such scenario, Schmidt et al. (2016) identified core competencies for *data librarian*s. Besides having a basic understanding of the specific disciplinary landscape, as well as being aware of norms and standards, they would be in charge of:

1) Provide access to data
2) Advocacy and support for managing data (e.g. knowledge of DMP templates and tools, data sharing options, licenses, data citation and reference practices, etc.)
3) Manage data collections

More in details, librarians are required to have a good knowledge of existing data centers, repositories and data discovery mechanisms, funders' policies and publication requirements of journals, metadata standards and schemas, data formats, domain ontologies, discovery tools, and so on.
In addition to this, *data librarian*s would cooperate in related services such as collections' development and curation, assistance in OA and copyright policies, information literacy, digital curation and preservation, etc.

Cassella (2016) identifies five major areas of reference for the role of the *data librarian*:

1) Library science
2) Scholarly communication
3) Technology
4) Disciplinary law, copyright and licenses
5) Communication and management

Another fundamental aspect emerged from the analysis of the case studies is the importance of domain-specific competencies. Indeed, the knowledge of the research mechanisms of specific scientific areas would represent an advantage, as the *data librarian* always operates in teams, being part of a staff composed by different researchers and technical figures.
In this respect, Brown et al. (2015) recognize the increasing complexity of the roles of the librarians supporting what they have named eResearch, underlining again that it is in the network of specialists they closely work with that *data librarian*s acquire the domain specific competencies differentiating them from other library professionals.
Thus, *data librarian*s would present advanced skills in those areas where their colleagues not operating in eResearch units have general or basic knowledges.

More in details:

- Advanced understanding of discipline-based research process, outputs and scholarly communication (e.g. data types and formats)

- Advanced knowledge of ethics, intellectual property, copyright and licensing
- Advanced knowledge of discipline-specific metadata schemas and related standards (item and collection level)
- Knowledge of repository certification schemes and standards
- Knowledge of semantic web standards
- High level communication and documentation skills, project management and business analysis skills

In these regards, we are facing a composite reality, being the role of the *data librarian* one of the most complex to define and identify in such composite scenario, as it emerges firstly from the case studies we observed.

We may reasonably argue that at the present time there is still no overall agreement on the competencies and, most of all, on the tasks that a *data librarian* operating in RDS should perform, this being reflected also in the terminology currently in use to describe and identify such role. As a matter of fact, the *data librarian* may identify a wide range of different context-related professionals. However, most of the commentators agree on stating that *data librarian*s usually work in team with other specialists and for this reason having or acquiring some basic domain knowledge would be an advantage. Moreover, the traditional librarians' ability to count on both existing capabilities and newly acquired skills favors their establishment as core members of a research support team.

6. Conclusions

The emergence of e-science and e-research has opened new paths and trends in scholarly communication and management. In the academic environment, the need for opening research products to a wider audience has become increasingly urgent. For instance, many funder bodies now request Open Access to scientific publications and require the presentation of a Data Management Plan along with the project proposal.

In addition, it has been widely recognized that data sharing would bring major benefits to the scientific community, avoiding useless duplications and saving the researchers' time and resources.

In order to pursue such aim, an efficient management of research data has become essential. This is the reason why in the recent years an increasing number of institutions has been adopting specific policies dedicated to the effective management of the data produced during the research process, leading to the creation and clear definition of dedicated services, namely Research Data Services.

In this scenario, perspectives and concrete realities are quite different, as we observed in the case studies, even though some common aspects may be identified, as the importance of data access, curation, preservation, and, last but not least, the fundamental importance of training either for the professionals operating in this field, or for researchers and students producing and collecting data in their daily work.

In such context, in the recent years the figure of the *data librarian* is acquiring importance, as it may represent one of the possible evolutions of the traditional librarian in the contemporary academic world. Many definitions describing this role are available in the literature, making quite difficult outlining unique skills, knowledge, competences, and tasks.

However, it is quite clear that it is a role that would not develop based on the classic skills of the librarian only, although these represent an extreme value and a concrete base for the development of a constantly evolving career. Indeed, it must be pointed out that, due to the recent advancements

in science, technology and scholarly communication, almost all professionals working in the research field had to reshape their attitude towards the research processes and scholarly communication. In such prospect, librarians occupy a privileged position either for their conventional background, or their successful adaptation to the transformations and evolutions their profession has undergone over the years.

Finally, we can argue that the traditional competences of the librarian should not be idealized, assuming that they are sufficient for becoming a *data librarian*. On the other hand, the *data librarian* should not be seen as a utopia, since numerous academic experiences show how this role is becoming a concrete reality for many professionals around the world.
As a conclusion, we may say that the *data librarian* is neither myth, nor utopia, but a composite reality.

Bibliography[i]

1. AGA – Annotated Model Grant Agreement, version 5.1, 6 December 2018. *Horizon 2020 Programme, European Commission*, http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf.
2. Australian Government (2007). *Australian Code for the Responsible Conduct of Research*, http://www.nhmrc.gov.au/index.htm.
3. Australian National Data Service. *Research data policy and the Australian Code for the Responsible Conduct of Research*, https://www.ands.org.au/guides/code-awareness.
4. Ball A. (2012). *Review of Data Management Lifecycle Models* (version 1.0). REDm-MED Project Document redm1rep120110ab10. Bath, UK: University of Bath.
5. Bell G., Hey T., Szalay A. (2009). *Beyond the Data Deluge. Science,* 323 (5919) 1297-1298; http://science.sciencemag.org/content/323/5919/1297.
6. Brown R.A., Wolski M., Richardson J. (2015). *Developing new skills for research support librarians.* The Australian Library Journal, 64 (3), 224-234.
7. Carlson J., Fosmire M., Miller C., Sapp Nelson M.R. (2011). *Determining data information literacy needs: a study of students and research faculty.* Portal: Libraries and the Academy, 11 (2), 629–657.
8. Cassella M. (2016). *Dal digital curator al data librarian*. Biblioteche oggi, 34 (4) 13-21.
9. *Commission Recommendation (EU) 2018/790 of 25 April 2018 on access to and preservation of scientific information*. Official Journal of the European Union*,* https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32018H0790&from=EN.
10. Crossley J. (2014). *An introduction to managing research data*, https://www.scribd.com/presentation/138079216/Managing-Research-Data.
11. Digital Curation Centre. *What is digital curation?* http://www.dcc.ac.uk/digital-curation/what-digital-curation.
12. Digital Curation Centre. *DMPonline,* https://dmponline.dcc.ac.uk.
13. European Union (2010). *Riding the wave. How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data. A submission to the European Commission*, https://www.fosteropenscience.eu/content/riding-wave-how-europe-can-gain-rising-tide-scientific-data.
14. Fearon D. Jr., Gunia B., Pralle B.E., Lake S., Sallans A.S. (2013). *Research Data Management Services*, SPEC Kit n. 334.
15. FORCE11, *The FAIR data principles*, https://www.force11.org/group/fairgroup/fairprinciples.
16. Genoni P. (2004). *Content in institutional repositories: a collection management issue*. Library Management, 25 (6/7), 300-306.
17. Gray J., Szalay A. (2007). *eScience – A Transformed Scientific Method*. Presentation to the Computer Science and Technology Board of the National Research Council, Mountain View, CA, 11 January 2007; http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt.
18. Griffith University (2012). *Griffith University Code for the Responsible Conduct of Research*, https://policies.griffith.edu.au/.
19. KNAW; NFU; NWO; TO2-federatie; Vereniging Hogescholen; VSNU (2018): *Nederlandse gedragscode wetenschappelijke integriteit*. DANS, https://doi.org/10.17026/dans-2cj-nvwu.
20. LIBER (2012). *Ten recommendations for libraries to get started with research data management. Final Report of the LIBER Working Group on E-Science/Research Data Management*, https://www.fosteropenscience.eu/content/ten-recommendations-libraries-get-started-research-data-management.
21. Lyon L. (2012). *The Informatics Transform: Re-Engineering Libraries for the Data Decade.* The International Journal of Digital Curation, 7 (1), 126-138.
22. McMillan D. (2014). *Data Sharing and Discovery: What Librarians Need to Know*. The Journal of Academic Librarianship, 40, 541-549.

23. OECD (2007). *Principles and Guidelines for Access to Research Data from Public Funding*, http://www.oecd.org/sti/inno/38500813.pdf.
24. Osswald A., Strathmann S. (2012). The Role of Libraries in Curation and Preservation of Research Data in Germany: Findings of a survey, https://www.ifla.org/past-wlic/2012/116-osswald-en.pdf.
25. Perrier L., Blondal E., MacDonald H. (2018). *Exploring the experiences of academic libraries with research data management: A meta-ethnographic analysis of qualitative studies*. Library and Information Science Research, 40 (3-4), 173-183.
26. Queensland University of Technology (2015). *Manual of Policies and Procedures. D/2.8 Management of research data*, http://www.mopp.qut.edu.au/D/D_02_08.jsp.
27. Research Data Alliance – Libraries for Research Data IG (2015). *How to maximize research data skills in libraries*, https://www.rd-alliance.org/how-maximize-research-data-skills-libraries.html.
28. Research Data Netherlands. *Essentials 4 Data Support*, https://datasupport.researchdata.nl/en/start-the-course/i-definitions/research-data/.
29. Research Data Netherlands (2018). *A federated data infrastructure for the Netherlands: the front-office – back-office model. UK web version,* https://researchdata.nl/fileadmin/content/RDNL_algemeen/Documenten/RDNL_FOBOmodel-UK-web.pdf.
30. Rice R. (2008). *Roles and Responsibilities for Data Curation: the Data librarian.* RDMF2: Roles and Responsibilities for Effective Data Management, Manchester, Chancellors Hotel and Conference Centre, 26-27 November 2008. Slides, http://www.dcc.ac.uk/events/research-data-management-forum/roles-and-responsibilities.
31. Richardson J. (2016). *Best practice guidelines for researchers: Managing research data and primary materials.* Griffith University, https://www.griffith.edu.au/library/research-publishing/best-practice-guidelines-for-researchers.
32. Sada E., Gregori L., Sirito P. (2013). *Un bersaglio mobile: l'evoluzione dei profili degli "information professionals" alla luce dei nuovi scenari accademici.* AIB studi, 53 (1), 92-99.
33. Schmidt B., Calarco P., Kuchma I., Shearer K. (2016). *Time to Adopt: Librarians' New Skills and Competency Profiles. Positioning and Power in Academic Publishing: Players, Agents and Agendas.* Proceedings of the 20th International Conference on Electronic Publishing, Loizides F., Schmidt B. (eds.), IOS Press, 1-8.
34. Schmidt B., Shearer K. (2016). *Librarians' Competencies Profile for Research Data Management.* Joint Task Force on Librarians' Competencies in Support of E-Research and Scholarly Communication, https://www.coar-repositories.org/activities/support-and-training/task-force-competencies/.
35. Scott M., Cox S. (eds) (2016). *Introducing Research Data.* University of Southampton, fourth edition, https://eprints.soton.ac.uk/403440/1/introducing_research_data.pdf.
36. Simons N., Searle S. (2014). *Redefining 'The Librarian' in the Context of Emerging eResearch Services.* Paper presented at VALA 2014, http://www.vala.org.au/vala2014-proceedings/vala2014-session-15-simons.
37. Swan A., Brown S. (2008). *The Skills, Role and Career Structure of Data Scientists and Curators: An Assessment of Current Practice and Future Needs*. Truro: Key Perspectives. http://www.jisc.ac.uk/publications/documents/dataskillscareersfinalreport.aspx.
38. Tammaro A.M. (2016). *Libraries in Digital Age (LIDA): la trasformazione delle biblioteche in era digitale*, https://annamariatammaro.com/2016/06/21/libraries-in-digital-age-lida-la-trasformazione-delle-biblioteche-in-era-digitale/.
39. Tammaro A.M. (2018). *IFLA Global Vision Project: Barcelona Kick-Off Workshop*, https://annamariatammaro.com/category/ifla-global-vision/.
40. Tammaro A.M. (2018). *L'evoluzione della biblioteca digitale: dall'accesso alle risorse elettroniche alla creazione e alla cura dei dati*, https://annamariatammaro.com/2018/05/09/levoluzione-della-biblioteca-digitale-dallaccesso-alle-risorse-elettroniche-alla-creazione-e-alla-cura-dei-dati/.
41. Tenopir C., Sandusky R.J., Allard S., Birch B. (2014). *Research data management services in academic research libraries and perceptions of librarians.* Library & Information Science Research, 36 (2), 84-90.

42. UK Engineering and Physical Sciences Research Council. *EPSRC policy framework on research data – Scope and benefits*, https://epsrc.ukri.org/about/standards/researchdata/scope/.
43. University of Edinburgh *Research Data Service*: https://www.ed.ac.uk/information-services/research-support/research-data-service/about-the-research-data-service.
44. University of Edinburgh. *Research Data Management Policy,* https://www.ed.ac.uk/information-services/about/policies-and-regulations/research-data-policy.
45. University of Edinburgh – EDINA. *MANTRA. Research Data Management Training*, https://mantra.edina.ac.uk.
46. University of Edinburgh – *EDINA*. *Do-It-Yourself Research Data Management Training Kit for Librarians,* https://mantra.edina.ac.uk/libtraining.html.
47. Whyte, A., Tedds, J. (2011). *Making the Case for Research Data Management. DCC Briefing Papers*, Edinburgh: Digital Curation Centre, http://www.dcc.ac.uk/sites/default/files/documents/publications/Making%20the%20case.pdf.
48. Witt M. (2008). *Institutional repositories and research data curation in a distributed environment*. Library Trends, 57 (2), 191-201.
49. Wright M., Sumner T., Moore R., Koch T. (2007). *Connecting digital libraries to eScience: the future of scientific scholarship.* International Journal on Digital Libraries, 7 (1-2), 1-4.
50. https://dictionary.cambridge.org/dictionary/english/data.
51. https://en.oxforddictionaries.com/definition/data.
52. https://www.merriam-webster.com/dictionary/data.

[i] URL last access: December 2018.