Research Data Literacy

René Schneider

Haute Ecole de Gestion, 7, route de Drize, 1271 Carouge, Switzerland
rene.schneider@hesge.ch

Abstract. This paper describes a pragmatic approach for the mediation and the teaching of research data literacy, i.e. those dimensions of information literacy that are dedicated to the creation, management, and reuse of research data. Based on prior work concerning the foundations of information literacy and curricula construction for data curation, the paper will begin with the definition of research data literacy, before describing an approach based on a fusion of core skills and a two dimensional matrix that reflects on the one hand the different student populations, and on the other hand a scale of various teaching modules. This matrix might serve as the basis for an operational implementation of different study programs.

Keywords: Research data management, data scientists, data curation cycle, research data literacy.

1 Introduction

Research data management, i.e. the processing of all types of raw or primary data, that are created along every research process, will not only play a crucial role for many scientists in the next years, but will also have strong implications for library and information science [1]: information specialists will become data librarians on the one hand; on the other hand, they will have to teach students, scientists, data managers and information specialists in order to prepare them for the new challenges of providing and using data infrastructures in almost all scientific disciplines.

Due to the fact that not only the current scientists, but also the actual and forthcoming generations of students of almost all disciplines will have to work with the research data, the paper will mainly discuss the need for teaching the students in a new sub-discipline of information literacy, namely "research data literacy". As this term shows, there are strong parallels with information literacy and the first can be seen as an offspring of the latter. This parallelism will be shown in reference to prior works concerning the definition of prototype curricula before striving towards the formulation of a complementary curriculum for research data literacy.

After establishing a fused list of core skills, the major competences of data management will be allocated to these core skills to be classified according to a scalable range of teaching units and different target groups.

S. Kurbanoğlu et al. (Eds.): ECIL 2013, CCIS 397, pp. 134–140, 2013.

2 Research Data Management and Information Literacy

In this section we will briefly discuss the definitions and relations that exist for and in between the terms "research data" and "information literacy," and present a new term proposed to combine them: "research data literacy." Basically, research data literacy is seen as a new sub-discipline within research data management that emerges from the need to educate students and scientists of all disciplines and to train information scientists from library and information science to do so.

Research data management is a method that enables the integration, curation and interoperability of data created during the scientific process, i.e. the production, access, verification, persistent storage and reuse of this data with the help of adequate and easy-to-use tools in virtual research infrastructures. These data are the essential part of the curation cycle [2] that comprises the following steps: the conceptualization, creation or reception, appraisal, selection, ingestion, preservation, storage, access, use and reuse, and transformation of research data. All data should be kept available in the three different domains that a scientist needs to do his work effectively [3]: a private, collaborative and public domain that are permeable for curation transactions.

The fact that almost the whole research process has to be transparent might lead to the assumption that research is on the way to becoming a utilitarian system of permanent control and evidence. As a matter of fact, the opposite is true, and research data management should be seen as “a liberal act” since it guarantees sustainable transparency for science through a “critical reflection on the nature of information itself, its technical infrastructure, and its social, cultural and even philosophical context and impact” [4]. Thus, the seven dimensions of information literacy curriculum in Shapiro and Hughes' paper can easily provide a basis for establishing a curriculum for research data literacy: tool literacy, resource literacy, social-structural literacy, publishing literacy, emerging technology literacy and critical literacy. As for the mediation of these literacies, a number of core skills have been defined: we refer firstly to the Big 6 defined by [5], which sought to teach how to “clarify, locate, select/analyze, organize/synthesize, create/present and evaluate” information. As for higher education, this list of skills was transformed to the seven pillars for information literacy, namely “identify, scope, plan, gather, evaluate, manage and present” [6]. As shall be seen later, these pillars build the starting point for the development of a curriculum for research data literacy.

The most striking difference between the terms “research data” and “information literacy” may be the fact that the first focuses on data and the second on information. This distinction has not been done to separate the terms from each other, they were coined within their own contexts at different points of time, but the line separating them is rather thin: research data management is interested in raw data from creation until extinction or archiving; information literacy has always been interested in the proper understanding and use of data that - only through the ability to use it - is converted into information. The focus on *data* in research data management can be explained with the primary and simple intention of getting back or giving testimony of the basics of research. After having focused for a very long time and always giving

preference to theoretical hypotheses, these are nowadays considered as being only one side of the academic research process, as opposed to the data. Thus, research data management is never only interested in raw data or the pure archiving, but on the use and reuse of data and its embedding context, which once again wipes out the border between data and information. Therefore the argument to recognize research data as information was proposed in the 2011 report of the Research Information Network [7]. Hence, data management and data curation can be seen as a logical extension of information literacy concepts [8].

In this sense, both the data and the context create - due to their innate relationship - the famous "difference that makes a difference". The problem of dissociation of data from its context is at the heart of the data management problem. If the context is lost, the reuse becomes difficult, if not impossible, i.e. that preservation and the other activities of the data curation have their place in problem solving, but if the problem of losing the persistent connection between data and context is not solved, all efforts remain worthless.

This problem may be compared to a treasure box: the data is the treasure, preservation the box, and the context the map to find the way to the box, once it has been stored in its repository. The action of a person that knows how to decipher the map and find the way to the treasure box, brings us finally back to the topic of this paper: research data literacy, i.e. the human competence to locate, analyze, organize, present and evaluate the treasure, i.e. research data in its context.

Actually, it should not be forgotten, that until now most of the questions concerning research data management are about to be asked, not answered! We see that there are different stakeholders, i.e. data creators, data scientists, data librarians, and data managers [9], all familiar with their situation and experts in their domain but lacking to a considerable degree literacy concerning most of the aspects of the curation cycle. Therefore, the most important aim besides the creation and provision of infrastructures must be the mediation of the know-how needed to use them for an efficient collaborative curation of the data to be stored. Due to the variety of the stakeholders and the varying degrees of knowledge, there is a need for flexible and scalable approaches that take into consideration the diversity of the stakeholders.

3 Building a Flexible Curriculum for Research Data Literacy

The need for flexibility can be illustrated by the different lenses that reflect the different points of view that people have on a specific matter; this is definitely true if the matter is information or data: they are "different in different contexts and for different ages and levels of learner and also dependent on experience and information need" [6] and any literacy approach must take into account the personal context in which the individual operates.

In our case, the differences can be found first and foremost in the variety of topics that deal with research data management, ranging, for example, from legal issues to metadata, from storage to marketing and from disaster management to data modeling. These topics are in relation to the manifold types of stakeholders implied in research data management: creators, curators, users whose implications are related to the different roles they play - being professors, research assistants or students doing

research - librarians, data scientists and curators, computer scientists, editors etc. - which lead to the need for flexibility and scalability concerning the width and the depth of the curriculum. In other words: different people have different needs; we therefore propose an approach for the education of research data literacy that does not solely focus on data managers but on distinct student populations or target groups.

Table 1. Synopsis of information literacy and research data skills

Big 6	**Seven Pillars**	**DPOE curriculum**
Clarify		
Locate	Identify	Identify
Select /Analyze	Scope	Select
Organize /	Plan	
Synthesize	Gather	Store
		Protect
Evaluate	Evaluate	
	Manage	Manage
Create / Present	Present	Provide

To do so, we compared - in a first step - the previously mentioned core skills of the Big 6 in Information Literacy and the seven pillars of information literacy in higher education with the six major data curation skills taken from the DPOE curriculum that was established by the Library of Congress for their "Train the Trainer Program in Digital Curation", namely: "Identify, Select, Store, Protect Manage, Provide" (http://www.digitalpreservation.gov/education/curriculum.html) (see Table 1).

As can be seen, they do not differ considerably either in quantity or in quality; most of the differing terms can be seen as synonyms, such as "Scope" and "Select", "Gather" and "Store", and "Present" and "Provide", whereas the latter can be seen as another term for the original double concept of "Create/Present". Interestingly, data curation skills contain a further concept, namely "Protect", which is definitely at the heart of sustainable data management. On the other hand, the information literacy skills name explicitly "Plan", which is contained in the "Store" activity of data curation, and "Evaluate", which is not explicitly mentioned in the DPOE curriculum.

For optimal use, we decided to fuse the seven pillars and the curation core skills to an optimized list of eight core activities for research data literacy: identify, scope, plan, store, protect, evaluate, manage and provide, which gives us the variety needed for the flexibility and scalability of the program. We set aside the Big 6, since they are fully absorbed in the extended list.

In the second step, our list of core skills was compared to the core skills of data management as described by [9] (see Table 2). The discrete allocation of exactly one data management skill to exactly one research data literacy skill is certainly not sufficient: only a few do - as a matter of fact - deal solely with one of the information literacy skills, though most do show a certain interconnectivity to the neighbouring literacy skill; some are of higher importance than others and indispensable for any understanding, while others are of secondary interest and will only be taught when enough time is at one's disposal.

Therefore, we connected - in the third step of our curricula building process - the qualitative aspects of the skills to the quantitative aspects of the teaching units. This was done via a simple contingency matrix (see Table 3), that combines the core skills with several teaching units as listed hereafter: a) a two-hour unit: a short introduction to a matter, that might be taught to any clientele with the aim of providing a general overview to the novice who wants to learn the basic principles and methods; b) a full course or workshop: an either one-spot intensive workshop of one or two days or a consecutive course taught over 10-15 weeks that gives a broad theoretical overview and a first introduction to the methods and tools used; c) a full module: a teaching unit that consists of several courses and seeks to give a complete overview of the discipline in theory and practice, taught over a longer period of time, generally between six months to a full year; d) a specialization: being part of a larger study program in which the student specializes in almost all techniques and prepares himself for working in this field after completion of the studies; e) a full study program: a complete program to form research data managers and data curators during approximately two years, based on the foundations of information science and the new competences needed for research data management; f) a certificate: similar to a full study program, but directed to teach people who are already working in a job related to the matter taught, who want to acquire the knowledge needed to work as a data curator or data manager.

Table 2. Research data literacy and data management competences

Research Data Literacy	Data Management Competences
Identify	Documentation (research environmental, temporal) / Context / From Information Management to Knowledge Management
Scope	Monitoring Process / Extracting Information from Data Models (and People)
Plan	Data Modeling / Meta Data / Standards Development
Store	Data Analysis and Manipulation / Merging, Mashing, Integration
Protect	Data Preservation / Data Security / Access Authentication / Conditions of Use / Data Legislation
Evaluate	Data Appraisal and Retention / Value of Data / Economic Issues
Manage	Complaints and Expectation Management / Coordination of Practice across Institution / Negotiation Skills / Risk & Disaster Management / Contingency / Advocacy, Promotion, Marketing
Provide	Facilitation, Communication / Raising Awareness

This range comprises the whole academic ‘instrumentarium’ of teaching programs that is currently in use and has proven to be effective for the organisation of higher education as well as programs dedicated to train people on the job.

Similar to the preceding table (i.e. Table 2), the discrete allocation does not have to be interpreted in an absolute manner but should rather be seen as an indicator of where to place emphasis: Since “Provision” comprises understanding and publication, it is seen as a skill that is of importance for everyone, whereas in a two-hour course unit, the emphasis should be placed on the identification of research data and the understanding of the curation process. A full course could be dedicated to the

selection and integration of the data with a special focus on the metadata, i.e. the modeling of the context. A full module would be dedicated to the core skills of the curation process and only the full program would introduce the main skills of the management process.

In the fourth and final step of our curricula building procedure, the different teaching units are allocated to the corresponding target groups or student populations. This matrix (see Table 4) defines the groups of people that might, should or must become literate in the field of research data and combines them with the qualitative and respectively quantitative aspects of the study programs discussed above. The lines of this matrix represent firstly the four different student populations that might be implied to different degrees in the research data management process and four different target groups of people that are already working in their job.

Table 3. Research data literacy - Organization matrix

	Provide	Identify	Scope	Plan	Store	Protect	Evaluate	Manage
2 hour unit	*	*						
Full course	*	*	*	*				
Full module	*	*	*	*	*	*		
Specialization	*	*	*	*	*	*	*	*
Full study	*	*	*	*	*	*	*	*

Table 4. Research data literacy - Curricula matrix

	2-hours unit	Full course	Full module	Specialization	Full program	Certificate
Any Bachelor student	+	*	-	-	-	-
Any Master student	-	+	*	-	-	-
LIS Bachelor Students	-	-	+	*	-	-
LIS Master Students	-	-	-	+	*	-
Data Creators	+	*	-	-	-	-
Data Scientists	+	*	-	-	-	*
Data Librarians	-	+	-	-	-	+
Data Managers	-	+	-	-	-	+

The contingency cells of the matrix are filled with three different markers that indicate the intensity of the contingency between the instances of the two dimensions: these markers range from ‘compulsory’ (+) over ‘optional’ (*) to ‘not an issue’ (-), in order to further classify the needs and demands that the different student or target groups may have.

4 Conclusions

In this paper we described a modular approach for research data literacy in a four-step procedure that can be used to develop curricula for all types of participants implied in the research data management process. The approach aims for granularity concerning the teaching components and flexibility to put them together. It could be shown that the core skills of information literacy can be used as a starting point to build two comparative tables and two contingency matrices that combine different levels of literacy with different activities to represent the different lenses of the stakeholders.

Beside the technological process that fosters research data management and that is already "on the run", we are in need of study programs for the different stakeholders, hence the motivation to write this paper. We do consider that these programs are founded in the core skills of information literacy and can be built upon them by a new arrangement of already existing components in library and information studies as well as their slight modification and adaption to the specifities of research data management plus the amendment of some complementary components that are only relevant for research data literacy.

References

1. Lewis, M.J.: Libraries and the Management of Research Data. In: McKnight, S. (ed.) Envisioning Future Academic Library Services, pp. 145–168. Facet Publishing, London (2010)
2. Higgins, S.: The DCC Curation Lifecycle Model. International Journal of Digital Curation 3(1), 134–148 (2008)
3. Treloar, A.: Private Research, Shared Research, Publication, and the Boundary Transitions (2011), http://andrew.treloar.net/research/diagrams/data_curation_continuum.pdf
4. Shapiro, J.J., Hughes, S.K.: Information Literacy as a Liberal Art. Educom Review 31(2), 31–35 (2011)
5. Eisenberg, M.: Information Literacy: Essential Skills for the Information Age. DESIDOC Journal of Library & Information Technology 28(2), 39–47 (2008)
6. SCONUL Working Group on Information Literacy: The SCONUL Seven Pillars of Information Literacy: Core Model for Higher Education (2008), http://www.sconul.ac.uk/sites/default/files/documents/coremodel.pdf
7. Research Information Network: The Role of Research Supervisors in Information Literacy (2011), http://www.rin.ac.uk/system/files/attachments/Research_supervisors_report_for_screen.pdf
8. Carlson, J., Fosmire, M., Miller, C., Sapp Nelson, M.: Determining Data Information Literacy Needs: A study of Students and Research Faculty. Portal: Libraries and the Academy 11(2), 629–657 (2008)
9. Donnelly, M.: RDMF2: Core Skills Diagram. Research Data Management Forum 17 (December 2008), http://data-forum.blogspot.ch/2008/12/rdmf2-core-skills-diagram.html