Data literacy: in search of a name and identity

Tibor Koltay*
Department of Information and Library Studies
Szent István University, Faculty of Applied Arts and Pedagogy,
Jászberény,
Hungary

Abstract
Purpose: The role of data literacy is discussed in the light of such activities as data quality, data management, data curation, and data citation. The differing terms and its relationship to the most important literacies are examined.
Approach: By stressing the importance of data literacy in fulfilling the mission of the contemporary academic library, the article takes an interdisciplinary approach, with the main emphasis on information literacy. The characteristics of digital literacy, scientific literacy and academic literacy are also examined. The content of data literacy education is explained in the context of data-related activities.
Findings: It can be concluded that there is a need for data literacy and it is advantageous to have a unified terminology. The need of taking a critical stance on hopes and fears, related to the promises of widespread ability of (big) data is also emphasized. The importance of context in data is underlined, as well. The necessary competences, skills and abilities, required for successful data literacy are identified.
Originality, value: The paper clarifies the terminology and provides a comprehensive and balanced view of competences. It adds to the body of knowledge about information literacy and other literacies in the light of research data and data literacy.
Keywords: data literacy, information literacy, literacies

Article Classification: Viewpoint

Introduction
Digital research data is a hot topic today. The professional literature of the 2010s clearly shows a tide in appearance of articles on digital research data issues. For instance, the ACRL Research Planning and Review Committee (ACRL 2012) identified data curation (in a wide sense) as one of 2012 top ten trends in academic libraries. The increased interest in data characterizes not only the highly instrumented scientific and engineering research, but the social sciences and the humanities, as well. The vast amounts of data allow researchers to ask new questions in new ways, and – in the same time – also pose a wide range of concerns for access, management and preservation (Borgman et al., 2011). Data sharing is also an issue. To make data accessible, we need to develop appropriate technical and organizational infrastructures for storage and retrieval.

*koltay.tibor@abpk.szie.hu
Incentives and policies for researchers to share data are also indispensable (Kowalczyk & Shankar, 2011) and data literacy, which carries the potential of motivation, is one of the essential elements of this infrastructure. However, the literature shows that differing concepts and terms are used. Besides data literacy there is data information literacy (Carlson et al., 2011), science data literacy (Qin & D’Ignazio, 2010) and research data literacy (Schneider, 2013).

Before trying to solve this naming problem and finding the proper identity of data literacy, we have to take a quick look at the role that the library can and should play in coping with the phenomenon of the “data deluge”, as it is often called (Borgman, 2012; Little 2012).

**Research data and the library**

In the same way as the library has traditionally facilitated access to documents, now information professionals could facilitate access to data, even though data do not necessarily fit into the same, document formats that libraries used to offer (Stuart, 2011). It is clear namely now that technology-driven research, known as e-research, generates large quantities of data. However, the lack of tools, infrastructure, standardized processes and properly skilled personnel may impede the continued development of e-research (Carlson et al., 2011).

No doubt, if libraries want to ensure that they remain relevant, services related to research data offer an opportunity to put the expertise of information professionals to good use (Christensen-Dalsgaard et al., 2012). Realizing this, the Association of Research Libraries declared in 2012 that academic libraries are in a favourable position to help researchers to meet the challenges of a data-driven research paradigm for as libraries and information professionals:

- Have experience and skills in fostering cross-departmental, cross-campus, etc. communication and collaboration, required for effective research data management.
- Are familiar with the research data needs of researchers and have been among the supporters of innovative publishing models, including open access publishing.
- Are increasingly providing data consultation services and information professionals are already involved in acquiring necessary abilities to manage data (Hswe and Holt, 2012).

The fields of involvement of academic libraries in providing research data services address the full data lifecycle, including planning, curation, and metadata creation and conversion (Tenopir et al. 2011) as well as playing the critical role as data quality hubs on campus, by providing data quality auditing and verification services for the research communities (Giarlo, 2013).

The Association of European Research Libraries (LIBER) published recommendations for libraries to get started with research data management support that involves – among others the development of metadata and data standards, creating data librarian (data specialists) posts, providing services for storage, discovery and permanent access and promoting data citation (Christensen-Dalsgaard et al., 2012). Mooney and Silver (2011, pp. 480) state that librarians “will not only need to learn how to apply traditional techniques of reference, instruction, and collection management to the unique format of data, but also how to raise awareness of this new service to the wider academic community.”

According to Calzada Prado and Marzal (2013), academic libraries can have different responses to eScience and the growing need to use research data. They can hire specialized staff (data librarians or data specialists) or further data management and analysis training for (generally reference) librarians. Intensifying the collection of data sources and providing access to them including the participation in the development of institutional data repositories is another option.
Last, but not least they can incorporate data literacy in their instructional programs and services. Data literacy also appears in the predictions about future directions for academic libraries in the Library 2.0 world (Merrill, 2011).

Information professionals have begun to understand these needs. For instance, an online survey of 140 libraries in Australia, New Zealand, Ireland, and the United Kingdom identified needs in guidance in the handling and management of unpublished research data, including data literacy education (Corral et al, 2013).

**In search of a name and identity**

What we stated above shows that speaking about data literacy is impossible without delineating the concept of literacy. Without aiming and exhaustive discussion (done to a substantial extent by Bawden, 2001), we can affirm that while literacy and its counterpart, illiteracy were once well understood and well defined (Buschman, 2009), the growing role of digital technologies changed their meaning. It meant originally being knowledgeable with the body of writing of aesthetic merit, while now it often signifies being able to cope with the socio-technological changes and challenges brought by the convergence between media, telecommunication, information and communication technologies (Livingstone et al. 2008).

When speaking about the issue of finding the proper name, it seems to be helpful to look at the difficulties experienced in the naming practices of information literacy (IL). As Hunt (2004) explains, there may be no agreement on the precise definition of information literacy. Nonetheless, most people use the term information literacy rather than library instruction or information fluency. However, if we do not use the same language, it will be difficult to convince our stakeholders about the importance of IL education. In the case of data literacy we may experience the same.

**Defining data literacy**

In our opinion, terminology should be unified around the concept and term of data literacy. We can adopt the approach of Calzada Prado and Marzal (2013), according to whom data literacy enables individuals to access, interpret, critically assess, manage, handle and ethically use data.

Critical assessment of data includes giving emphasis to the version of the given dataset, the person responsible for it (ACRL, 2013). It also may involve understanding what data mean, including how to read graphs and charts appropriately, draw correct conclusions from data, and recognize when data are being used in misleading or inappropriate ways (Carlson et al., 2011). Two additional features can be found in a report of the Research Information Network. They are data preservation and curation (RIN, 2011).

An important feature of data literacy is the emphasis on the creation of data. Even though both Carlson et al. (2011) and Schneider (2013) use different terms (data information literacy and research data literacy, respectively) instead of data literacy, they stress i.e. the necessity to accommodate not only the data consumer’s viewpoint, but the data producer’s, as well.

Data literacy education is not only a relatively new field for libraries, but it is one that cuts across disciplinary boundaries and across the traditional structures of academic library organizations. One sign of this is that humanities and social science scholars are on the process of becoming a new constituency for data literacy education (ACRL, 2013).

Data literacy education can have a dual purpose. One is rather self-explanatory, i.e. to achieve that students and researchers become data literate science workers. The other goal is to educate data management professionals (Qin & D’Ignazio, 2010, Schneider, 2013). In regard to the former, Haendel et al. (2012) speak about creating a culture of semantic scientists, stating that education in data literacy and information literacy should accompany scientific training to establish a new cultural standard, especially because researchers often do not realize that their own scholarly communications constitute a primary source of data.
The summary of employers’ requirements set against scientific data specialists (data librarians) by Si et al. (2013) shows high frequency of offering reference services for scientific research and data curation is regarded as one of the core duties of scientific data specialists, stating that in order to foster the data literacy of users, libraries require the instruction and training to users, including helping them understand the significance of scientific data curation and master the usage of various tools for data processing, data analysis and data statistics.

In summary, we can define data literacy as a set of skills and abilities related to accessing research data, understanding, interpreting, managing, critically evaluating and ethically using it. Managing comprises preservation and curation. Data literacy on the whole accommodates the data consumer’s viewpoint, as well.

**Data literacy and information literacy**

The work of today’s researcher mobilizes competencies on three levels:

- conceptual competencies that include among others innovative thinking, problem solving and critical thinking;
- human competencies, like social networking skills, self-management and cross-cultural interaction skills;
- practical competencies that include media literacy and information literacy (Lee, 2013).

The perhaps best known form of practical competencies is information literacy. IL education emphasizes critical thinking and the necessity to recognize message quality. It has strong positions among literacies despite some (well founded) scepticism, highlighting the fact that this concept and especially the *lack of information literacy* has always seemed to be of more importance to academic librarians than to any other players in the information and education arena (Bawden & Robinson, 2009).

A broader interpretation of IL recognizes that the concept of information includes research data (RIN, 2011). The reason for this is in the fact that information literacy is the best contender to draw together the other literacy movements into a single emphasis (Badke, 2009). Many writings, related to data literacy provide evidence of this. We already mentioned some. Besides of these, Fosmire and Miller (2008) speak about information literacy in the data world. Even though without referring to data literacy, Wang (2013) mentions that reference librarians frequently conduct IL sessions that educate the users about the existing data resources for their specific study areas. Calzada Prado and Marzal (2013) state that information literacy and data literacy form part of a scientific-investigative educational continuum, a gradual process of education that begins in school, is perfected and becomes specialized in higher education and becomes part of lifelong learning. In this later quality it is close to scientific literacy, which we will address shortly later.

Andretta et al. (2008) identified presenting, evaluating and interpreting qualitative and quantitative data as a learning outcome of IL. According to Hunt (2004), data literacy education should borrow heavily from information literacy education, even though the data literacy field is more fragmented than the domain of information literacy. Schneider (2013) also defines data literacy as a component of information literacy.

Si et al., (2013) state that data-related services should be supported by professionals with both excellent scientific literacy and information literacy skills.

Adapting the ACRL information literacy competency standards for higher education (ACRL, 2000) to data information literacy, as done by Carlson et al. (2011), offers an interesting perspective.
Obviously, we know that the ACRL standards have been criticized for representing a narrow and marginalized view of information literacy (Špiranec and Banek Zorica, 2010; Nazari & Webber, 2012). On the other hand the standards not only acknowledge the value of research data, but include a number of explicit references to them (Calzada Prado & Marzal, 2013, Magnuson, 2013).

Standard Three, of the standards, is especially relevant as it directs attention to evaluating information critically. If we translate this universal principle to data, we see that data information literate students are required not only to determine when and how to share data, but also document their own sources of data. This acquires significance in the light of the difficulty of documenting data, a fact that appeared in our writing several times. The reason for this is fairly simple. Critical examination of some information or data is practically impossible if we cannot rely on its documented feature, e.g. we cannot verify its provenance. Let us add that – as reflected by our definition – it is an imperative for data literacy to be built on critical approach towards data (Carlson et al., 2011).

In a typical IL course we can assume the students know how to read, use a web browser and word processor. Data-related skills and abilities, even relatively simple ones, are not always so obviously present (Hunt, 2004). Despite of this, data literacy education could rely on the experiences of the writing across the curriculum (WAC) movement. The WAC movement presented a model since the 1970s for encouraging collaboration between faculty and librarians in developing composition and critical thinking skills (Bronshteyn & Baladad, 2006). WAC can provide information literacy with ways to resolve the increasingly irrelevant theory-practice divide as it places the student at the centre of the educational process (Elmborg, 2003).

Unfortunately, it would be utterly naïve presume that researchers easily (and readily) accept the need for acquiring data literacy skills. There is substantial evidence that the people in general as well as researchers hold themselves competent and skilful in dealing with information (Nicholas et al. 2008). Data is probably not different in researchers’ perception. Even more, they may feel more confident with it.

Data literacy and academic literacy

The relationship between data literacy and academic literacy has many faces. The latter has a grammatical dimension that information literate students must be taught and researchers must master. It involves the comprehension of the entire system of thinking, values, cultural identity and information flows of academia, which results in the ability to read, interpret, and produce texts valued in academia (Elmborg, 2006).

Academic literacy is more closely associated with formal learning, than data literacy. This is especially true in higher education.

Data literacy, affects the undergraduate student population to a lesser degree than academic literacy. Instead, data literacy has a true lifelong-learning character. The reason for this is in the fact that researchers in the past did not have to possess many of the data-related skills, at least not the ones related to data sharing and needed for citing data. In addition to this, the majority of the researchers have to acquire data literacy in the workplace, up to the point, when it becomes a standard literacy in graduate and doctoral education.

In his paper, Weideman (2003) enumerates some requirements set against academic literacy. Interpreting and using metaphors and idioms, or word play, rely mainly on linguistic abilities. The ability to understand academic vocabulary is different. It must be acquired together with the system of concepts and ways of thinking, characterizing a particular discipline or academia in general. Being aware of the logical development of an academic text is similarly two-faced. If we substitute data for information, we can identify a number of requirements that are relevant for data literacy.
Sensitivity for meaning that and intended audience, interpreting, using and producing information presented in graphic or visual format, as well as making distinctions between essential and non-essential information, fact and opinion, propositions and arguments; distinguishing between cause and effect are examples of this. The requirement to classify, categorise and handle data and to do simple numerical estimations and computations are directly connected to data literacy.

Other literacies

Data literacy shares some distinguishing features with media literacy, especially in regard to the use and reuse of content in ways not imagined by the content creator (ACRL, 2013). Mentioning media literacy has two reasons. First, we know that there is a convergence among literacies caused by the (already mentioned) convergence between different forms of media and information and communication technologies (Livingstone et al., 2008). This provides a reason, why media literacy has relevance for a wide array of media, including research data. Second, we are aware of the fact that the capacity and interest in data-related issues is to a substantial extent result of the appearance of the web 2.0. In this environment, users and their interests are represented in mediated spaces, which also serve as an environment to activate engagement with others (Jarrett, 2008). This mediated nature influences data indirectly and to a lesser extent. Nonetheless, it requires that we approach data literacy not only as a manifestation of information literacy, but look at it through the lens of media literacy.

Digital literacy cannot be bypassed, either. As Qin and D'Ignazio (2010) put it, information literacy mobilizes the abilities and skills, related to finding, retrieving, analyzing, and using information. The same object, i.e. information, is aimed by digital literacy, which not only accentuates creating it (as we already mentioned this), but emphasizes the use of digital technology.

When we underlined the importance of the data producer’s viewpoint, we connected data literacy to the competences of digital literacy, which include “being comfortable with publishing and communicating information” (Bawden, 2008, pp. 20).

Data literacy is also connected to scientific literacy, which comprises methods, approaches, attitudes and skills, related to thinking scientifically and doing scientific research (National Academy of Sciences, 1996). This implies that everybody should be scientifically literate, even if only a small number of graduate students become scientists.

All the above literacies fit well into the framework of metaliteracy, which is overarching by providing the foundation for media literacy, digital literacy and other literacies, and emphasizes content (Mackey & Jacobson, 2011). With this the boundaries between information in information literacy from data in data literacy become blurred. In fact they never have been rigid, as IL has always been interested in the proper understanding and use of data that is converted into information (Schneider, 2013).

Statistical literacy is often mentioned in connection with data literacy. Schield (2004) is of the opinion that information literacy, statistical literacy and data literacy are tied together by a common set of problems and a similar level of approach. All three are more general than specific, they each involve interdisciplinary study and they deal with fundamentals. Besides of this, data literacy sets the overall context for evaluating the sources of data and is needed to access, manipulate and summarize the data, while statistical literacy guides that process.

The content of data literacy education

No doubt, data literacy’s learning outcomes need to be codified and (when acquired) best practices for data literacy programs have to be articulated (Hunt, 2004). There are a number of reasons, why we need data literacy. However, the perhaps most important factor is that witness a widespread belief that the existence and access to research data, in particular to big data offers a higher form of intelligence and knowledge.
There is an aura of truth, objectivity, and accuracy around it, as well. While big data is seen as a solution to many burning questions, it is regarded by many as a tool that threatens privacy, decreases civil freedoms, ushering increased state and corporate control. The shifts to be expected of big data are probably more subtle than these, even though we cannot see this clearly among our current hopes and fears (Boyd & Crawford, 2012). This leads to the conclusion that one of the most important goals of data literacy education should be to foster critical thinking that keeps us away from the pitfalls of being overly optimistic or unduly pessimistic, behaving in an excessively critical or uncritical way.

Data literacy should take into consideration the framework of future work skills and abilities, outlined by Davies et al. (2011). One of these abilities is the one to translate vast amounts of data into abstract concepts, as well as to understand data-based reasoning. This may lead to a situation, when we will start seeing almost everything that we come into contact, through the lens of data, and will regard it as computational, programmable.

Calzada Prado and Marzal (2013) emphasize the importance of knowing how to select and synthesize data and combine them with other information sources and prior knowledge. They also enumerate the following abilities:

- To identify the context in which data are produced and reused (data lifecycle);
- To recognize source data value, types and formats;
- To determine when data are needed;
- To access data sources appropriate to the information needed;
- To critically assess data and their sources;
- To determine and use suitable research methods;
- To handle and analyze data;
- To present quantitative information (specific data, tables, graphs, in reports and similar);
- To apply results to learning, decision making or problem-solving;
- To plan, organize and self-assess throughout the process.

It is not by accident that context is mentioned in the above list in the first place. Context is utterly important. Dissociation of data from its context and the loss of context make reuse difficult, or impossible (Schneider, 2013).

From the content of a series of instructional session on socioeconomic data, described by Wong (2010), we can single out three aspects. One of them is exploring data evaluation and use. The other one is guiding students in understanding data-collection methods and dissemination channels. Introducing students to the rich varieties of data fulfilling different information needs should also be emphasized.

As service providers, data librarians should be acquainted with quantitative research methods, which enable them to process and analyze research data. To be able to provide support for researchers, they have to possess an extensive understanding of scientific data sources, which will enable them to recommending comprehensive and reliable data sources (Si et al., 2013). From among the fields, where data literacy competencies are used, following ones have to be mentioned:

- Discovery and acquisition of data;
- Data management;
- Data conversion and interoperability (dealing with the risks and potential loss or corruption of information caused by changing data formats);
- Metadata;
- Data curation and re-use;
- Data preservation;
- Data analysis;
• Data visualization;
• Ethics, including citation of data (Carlson et al., 2011).

Data literate graduate students and researchers have to be familiar with some of the questions that data curation poses:
• Who owns the data?
• What requirements are imposed by others (e.g. funding agencies or publishers)
• Which data should be retained?
• For how long should data be maintained?
• How should digital data be preserved?
• Are there ethical considerations?
• What sort of risk management is needed for research data?
• How are data accessed?
• How open should the data be?
• What alternatives to local data management exist? (Erway, 2013).

Data literacy should include answers to the question about openness, asked above. It has to include the notion of open data, which has been advocated by a number of researchers in order to make science more open and accountable (Stuart, 2011). While doing this, we have to be aware of the legitimate boundaries of openness set by commercial interest, the protection of privacy, safety and security (The Royal Society, 2012).

The need for reviewing data periodically is also worth of attention as it forms the basis of responsible decisions about disposal of “unnecessary” data, and preventing hardware or software obsolescence (Pryor, 2012).

A data literate researcher does not need to be a data curator. However, being familiar with the competencies of the curators and the fields, where the competencies, listed below, are of good use helps in developing data literacy:
• The ability to collaborate and work in teams;
• Familiarity with scientific data sources;
• Familiarity with quantitative research methods;
• Knowledge of general metadata standards (Si et al., 2013).

From the fields, where these and other competencies can be used the following ones can be singled out:
• The data structure of different digital objects;
• The ways to assess the digital objects” authenticity, integrity and accuracy over time;
• Storage and preservation policies, procedures and practices;
• The risks of information loss or corruption of digital entities (Madrid, 2013).

Data-literate persons must be also aware of the limitations of computational thinking, mentioned earlier. This requires that they understand the limitations of data and remain able to act in the absence of data (Davies et al., 2011).

Data literacy education should not forget about data citation, which allows the identification, retrieval, replication, and verification of data underlying published studies. Standardized forms of data citation could provide a motivation for researchers to share and publish their data, thus it has the potential to become a tool of reward and acknowledgment for them (Mooney & Newton, 2012). Unfortunately, despite numerous initiatives, at the moment, there is no standardization and consistency in data citation.

The quality of data is a driving force of data literacy, as well (Carlson, 2011, Madrid, 2013). It is determined by multiple factors. The first one is trust that depends on subjective judgments on authenticity, acceptability or applicability of the data. Trust is also influenced by the given subject discipline, the reputation of those responsible for the creation of the data, and the biases of the persons who are evaluating the data.
The next factor is authenticity, which measures the extent to which the data is judged to represent the proper ways of conducting scientific research, including the reliability of the instruments used to gather the data, the soundness of underlying theoretical frameworks, the completeness, accuracy, and validity of the data. In order to evaluate authenticity, the data must be understandable. The condition for the evaluation of understandability of data is the presence of sufficient context in the form of documentation and metadata, and it requires the data be usable (Giarlo, 2013). If we want data to be usable, it has to be discoverable and accessible; and be in a usable file format. The individuals judging data quality need to have at their disposal an appropriate tool to access the data; which has to show sufficient integrity to be rendered. Integrity of data assumes that the data can be proven to be identical, at the bit level, to some previously accepted or verified state. Data integrity is required for usability, understandability and authenticity, thus it influences overall quality (Giarlo, 2013). Trust is related to cognitive authority, which has two levels. At an operational level, cognitive authority is the extent to which users think that they can trust the information. On a more general level, cognitive authority refers to “influences that a user would recognize as proper because the information therein is thought to be credible and worthy of belief” (Rieh, 2002, pp. 146).

At an operational level, cognitive authority is the extent to which users think that they can trust the information. On a more general level, cognitive authority refers to influences that a user would recognize as proper because the information therein is thought to be credible and worthy of belief (Rieh, 2002).

We already mentioned an essential aspect of data literacy. It is the profound understanding of the big data phenomenon, especially acknowledging that the decisive factor is not the quantity of data, but the capacity to search, aggregate, and cross-reference large data sets by virtue of the processing power of computers and networks. Data literate persons also have to know that data is no longer an exclusive issue for the sciences, but it is present in the social sciences, the humanities, arts and culture, as well (Boyd & Crawford, 2012).

**Conclusion**

The success of data literacy will depend on how well we train information professional and make faculty and administrators understand why data literacy is imperative (Hunt, 2004). Reskilling library staff is utterly important in this process, as only a few libraries are able to hire new, specialized professionals (Christensen-Dalsgaard et al., 2012). Those, who will use data, will need education about how to find it, how to understand, interpret, and apply what they find. So, educating students alongside with future and actual scholars to data literacy is a new challenge to libraries, which cuts across disciplinary boundaries.

Information professionals thus need to be involved in developing ways to handle key issues of data literacy as they are precisely in the intersection of scholarly communication and information literacy (ACRL, 2013). Last, but not least, we have to take into consideration that every research project is unique, so the training needs of every researcher are unique (East, 2005). This is true in the case of data literacy education, as well.

**References**

Andretta, S. et al. (2008), “Information Literacy Education in the UK”, *Communications in Information Literacy*, Vol. 2 No. 1, 36-51.


