

Open Research Online

The Open University's repository of research publications
and other research outputs

Creating an Understanding of Data Literacy for a Data-driven Society

Journal Item

How to cite:

Wolff, Annika; Gooch, Daniel; Cavero Montaner, Jose J.; Rashid, Umar and Kortuem, Gerd (2016). Creating an Understanding of Data Literacy for a Data-driven Society. *Journal of Community Informatics*, 12(3) pp. 9–26.

For guidance on citations see [FAQs](#).

© 2016 The Authors

Version: Version of Record

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data [policy](#) on reuse of materials please consult the policies page.

oro.open.ac.uk

Special issue on Data Literacy: Articles

Creating an Understanding of Data Literacy for a Data-driven Society

Annika Wolff

The Open University, United Kingdom

Corresponding Author.

annika.wolff@open.ac.uk

Daniel Gooch

The Open University, United Kingdom

daniel.Gooch@open.ac.uk

Jose J. Caverio Montaner

The Open University, United Kingdom

jose.caverio@open.ac.uk

Umar Rashid

The Open University, United Kingdom

umar.mir@open.ac.uk

Gerd Kortuem

The Open University, United Kingdom

gerd.kortuem@open.ac.uk

Society has become increasingly reliant on data, making it necessary to ensure that all citizens are equipped with the skills needed to be data literate. We argue that the foundations for a data literate society begin by acquiring key data literacy competences in school. However, as yet there is no clear definition of what these should be. This paper explores the different perspectives currently offered on both data and statistical literacy and then critically examines to what extent these address the data literacy needs of citizens in today's society. We survey existing approaches to teaching data literacy in schools, to identify how data literacy is interpreted in practice. Based on these analyses, we propose a definition of data literacy that is focused on employing an inquiry-based approach to using data to understand real world phenomena. The contribution of this paper is the creation of a common foundation for teaching and learning data literacy skills.

Wolff, A., Gooch, D., Caverio Montaner, J.J., Rashid, U., Kortuem, G., (2016). Creating an understanding of data literacy for a data-driven society. *The Journal of Community Informatics*, 12(3), 9-26.

Date submitted: 2015-12-16. Date accepted: 2016-06-13.

Copyright (C), 2016 (the authors as stated). Licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 2.5. Available at: www.ci-journal.net/index.php/ciej/article/view/1286.

Introduction

Whilst the fields of machine learning, data analysis and visualization are rich, there is surprisingly little research into the human component, particularly as it applies to more complex data. What are the competences that people must acquire in order to learn from and solve problems with data? What new skills must humans learn in order to both design, interpret and critique complex data analysis and visualisation?

The term data literacy is used to broadly describe the set of abilities around the use of data as part of everyday thinking and reasoning for solving real-world problems. Data literacy is increasingly considered to be a life skill, as daily interactions with data become evermore commonplace [12] and individuals more frequently make judgments from data and make decisions regarding the use of their own personal data [11]. There is also an identified shortage of data scientists in the U.K. [18]. As a result there is a recent push to increase the teaching of data literacy in schools, where previously it was mostly undergraduates who would find a need to acquire more sophisticated data handling skills upon entering higher education.

Current definitions of data literacy are not fit for purpose - they don't account for changes in the nature of data sets, which are becoming larger and more complex. Nor do they account for the different roles in which people must apply data literacy skills. This has implications for increasing the data literacy of society through school education, since without a clear definition of what data literacy is, it is both hard to teach and to assess the outcome of teaching.

This paper examines the research literature relating to data literacy in order to identify commonalities between definitions and to disambiguate it from the more coherently defined statistical literacy, which has been the topic of research for a longer time and is a more established part of the teaching curriculum in schools. The aim is to demonstrate that there is a need for a clearly defined field of study around data literacy that is distinct from and covers a different set of competencies than statistical literacy.

Through mapping both the space of user needs and the space of individual competences or skills that comprise the definitions of data (and statistical) literacy, we develop a single framework to support the multiple perspectives of data literacy and form a common foundation for the teaching and learning of data literacy skills.

Definitions of Data Literacy and Statistical Literacy

All of the data literacy definitions we reviewed are based on a specific scenario, which delimits which data literacy skills will be applied. Mandinach and Gummer [14] propose a definition of data literacy in the context of supporting teachers to use student data to improve their practice, as a type of learning analytics. In their view, data literacy is:

“the ability to understand and use data effectively to inform decisions. It is composed of a specific skill set and knowledge base that enables educators to transform data into information and ultimately into actionable knowledge. These skills include knowing how to identify, collect, organise, analyse, summarise and prioritise data. They also include how to develop

hypotheses, identify problems, interpret the data, and determine, plan, implement, and monitor courses of action.”

Deahl [7] and Vahey et al. [29] both propose definitions of data literacy in the context of teaching data literacy in schools. In Deahl’s view, data literacy is:

The ability to understand, find, collect, interpret, visualize, and support arguments using quantitative and qualitative data.

Deahl further elaborates on this by defining a more specialized skillset for the data literate student, ranging from the more general ‘understanding data’ to checking data for ‘bias and inaccuracy’ and ‘taking measurements’.

Vahey et al. [29] propose that:

data literacy includes the ability to formulate and answer questions using data as part of evidence-based thinking; use appropriate data, tools, and representations to support this thinking; interpret information from data; develop and evaluate data-based inferences and explanations; and use data to solve real problems and communicate their solutions.

Vahey et al. further stipulate that students must be anchored in the context of the data in order to be able to generate appropriate arguments and perform the analyses for solving problems.

Despite obvious overlap between these three definitions, it is clear that they each have a different focus – which tends to reflect the context in which it was derived. They also have a different level of granularity, not just between the definitions, but also within them. For example, the ability to ‘use data effectively’ is at a different level of granularity to the more specific skill of ‘collect data’ (both taken from [14]).

There is considerably more agreement amongst the definitions of statistical literacy. For example, Carlson et al. [4] defines statistical literacy as the ability to:

read and interpret summary statistics in the everyday media, in graphs, tables, statements, surveys and studies.

Callingham, 2006 [3] proposes that statistical literacy is:

the ability to understand and critically evaluate statistical results that permeate our daily lives” and to “appreciate the contributions of statistical thinking towards private, professional and personal decisions.

Taken as a whole, the definitions indicate a shared understanding of statistical literacy as being the ability to critically assess statistical evidence encountered within everyday life. This consistency makes the definition more useful when trying to teach statistical literacy, for example in school. However, it should be noted that these same competences appear, in some form, within many definitions of data literacy - but do not constitute the entire definition. This suggests that statistical literacy may form part of a wider set of competences that constitute data literacy. Resolving the overlap between the two literacies will lead to better provision for teaching both.

Identifying commonalities amongst definitions

A qualitative analysis of definitions of both data literacy and statistical literacy was undertaken, to identify commonalities between them. Nine definitions of data literacy and four definitions of statistical literacy were selected for the analysis. These were definitions that were somehow unique, i.e. they were not simply referencing or restating previous definitions. The full definitions can be found in Appendix A. The outcome of this analysis is summarized in Table 1.

Card sorting to Unify Competences

An open card sorting method [25] was chosen as a first step towards unifying competences. The aim was to first reach a consensus about which competences shared a similar meaning and could therefore be merged, then to understand how this reduced list of competences could be further categorized. A similar method was used by Petrie et al. [20] in creating a unified definition of web accessibility. The sorting process was led by the first author, with other authors participating. To give an example of how competences were merged, in Mandinach and Gummer's [14] definition of data literacy they refer to "identify problems" whereas Deahl [7] includes "skills necessary to ask questions that can be researched using data". This is merged into the competency "Identify problems or questions that can be solved with data". One term 'prioritise data' (from [14]) was omitted from the analysis due to lack of clarity of meaning. The process of grouping competences that could be merged was repeated a number of times for more accuracy and in the end a total of 23 unique competences were identified. These are shown in the column headed 'competence' in Table 1. The references after each competence reflect the definition(s) from which it was derived, and whether they were data literacy definition, or statistical literacy definitions (shown in bold). Next, the authors sorted the merged competencies into categories.

In the first iteration of this sorting task, two strong categories emerged. In the first category were competences that described a process through which data literacy skills could be applied. These included to 'undertake data inquiry process' and 'plan, implement and monitor courses of action' (see 'inquiry process' under Competence in Table 1).

In the second, larger, category were competences that described what we term foundational knowledge, such as 'understand how data can be produced or found' or 'interpret information derived from datasets'. There was noticeably more agreement amongst definitions with regard to the set of competencies found in this category. Out of 9 definitions of data literacy, there were 8 that referred in some way to the ability to create explanations from data.

Applying the PPDAC Inquiry Process

PPDAC is an approach to teaching statistical thinking that is used in schools in New Zealand, where statistics is emphasised as a subject in its own right [33]. PPDAC stands for Problem, Plan, Data, Analysis and Conclusion. Like other types of inquiry (for an example, see White [32]), the stages represent part of an iterative cycle in which the conclusions might prompt further questions and analysis, often of increasing complexity as the problem is being solved. Sometimes, in answering one question, a completely new question or problem is identified which triggers a completely new inquiry process. While PPDAC follows a fairly typical inquiry process it places less emphasis on the planning and conducting of scientific

experimentation and more focus on real-world problems that can be solved through an analysis of data. The PPDAC cycle is shown in Figure 1.

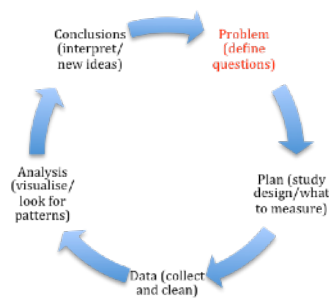


Figure 1. A representation of the PPDAC cycle

Using knowledge of PPDAC and other inquiry processes, the authors identified that it was possible to further categorise foundational knowledge under different stages of a PPDAC inquiry cycle. This is an important finding, since few data literacy definitions explicitly refer to undertaking an inquiry process, despite the stated competences reflecting one. The left hand column in Table 1 reveals this mapping.

Table 1. categorization of data literacy skills across multiple definitions

PPDA C	Competence	Foundational competence
<i>Inquiry process</i>		
	Plan, implement and monitor courses of action [14]	
	Undertake data inquiry process [34]	
<i>Foundational knowledge</i>		
	Understand the ethics of using data [34]	Ethics
	Use data to solve (real) problems [29] [2] [34] [14] [7]	Real-world problem-solving context
	Understand the role and impact of data in society in different contexts [7]	
P	Identify problems or questions that can be solved with data [14] [7]	Ask questions from data
P	Develop hypotheses [14] Identify data [14]	Develop hypotheses and identify potential sources of data
D	Collect or acquire data [6] [34] [21] [14] [7] <i>Critique data [21][7]</i>	Collect or acquire data

A	Transform data into information and ultimately actionable knowledge [29] [34] [14] [7] Create explanations from data [29] [33] [14] [7] [2] [4] [6] [21] <i>Access data [22]; Analyse data [2] [34] [21] [14] Understand data types [7]; Convert data [22]; Prepare data for analysis [7]; Combine quantitative and qualitative data [34] [7]; Use appropriate tools [29]; Work with large data sets [6]; Summarise data [14]</i>	Analyse and create explanations from data
C	Interpret information derived from datasets [29] [4] [34] [4][22][3][30] Critique presented interpretations of data [29] [4] [3][30]	Evaluate the validity of explanations based on data and formulate new questions

Placing Specialist Skills into the Developing Framework

There was a third group of competences which related to what could be termed specialist data handling skills, such as ‘create representations of data’ (e.g. create a chart or visualization). There was very little overlap between definitions of these skills, they are quite ad-hoc and tend to reflect different levels of expertise with data, for example one definition includes ‘understand data types’ whilst another refers to the ability to ‘convert data’. The specialist skills are shown in italics beneath the foundational competency to which they relate, which in nearly all cases is data analysis. It should be noted that it would also be possible to fill the space of specialist skills, creating some sort of hierarchy, for example distinguishing between different types of analysis (visual, statistical, machine-learning) or even further to talk about specific methods and tools. However, it would not be possible to create an exhaustive list, since new techniques, methods and tools emerge regularly.

Placing Higher-level Skills into the Developing Framework

Three competences could not be placed within the PPDAC framework. These concerned the ethical use of data, understanding the role and impact of data in different contexts and the use of data to solve problems - which we have taken to mean ‘real-world’ problems based on the context provided through literature from which definitions were taken. In understanding what a real-world problem is, we draw on notions of experiential learning, as popularized by thinkers such as Dewey [8] and Freire [9], which puts forward the idea that learning should be situated within a real-world context and with reference to ones own experience rather than reliant on rote learning of a collection of facts. Further, such learning experiences should be responsive to cultural differences that might affect individual learner’s view of the world.

We propose that given their importance these three competences should in fact cut across the inquiry cycle: for example, that ethical considerations, including concerns towards security and privacy of peoples’ data, should be in place in the planning, execution, and analysis of any data-driven inquiry, and that to be effective, an inquiry should address real-world problems and be responsive to the differing needs of learners.

Identifying the Foundational Competences for Data Literacy

In the final stage of analysis, the intermediate definitions of foundational knowledge, that were derived from existing definitions, were merged and re-described with the goal to reduce overlap of meaning and to create a set of foundational competences derived from the original definitions (right hand column of Table 1). At this stage, one concept was added which was

felt to be missing from definitions with regard to describing foundational knowledge. This was the ability to formulate new questions based on the outcome of an inquiry process (i.e. start a new cycle of inquiry). This step is standard across most models of inquiry, such as those proposed by White, 1999 [32] or Scanlon et al, 2011 [23].

Disambiguating Statistical Literacy

From the analysis it is also apparent that the foundational competencies associated with statistical literacy appear within the conclusions stage of an inquiry process. This is the point at which an inquiry has already been undertaken by someone in order to present findings. Thus statistical literacy – as previously identified through analyzing the definitions - seems most often to be associated with the skills required for assessing the validity of statistical data presented to support arguments, or used as the basis for decision-making. However, we would argue that in order to be able to fully critique presented statistics, it is important to have at least a working understanding of the process through which such explanations and representations of data are created, in order to understand how the choice of analysis and visualisation can bias the interpretation, or how incorrect selection of data from a population can skew results. This suggests that data literacy is important to support the understanding and application of statistical literacy. In other words, knowledge of the overall inquiry process and the activities it entails is a pre-requisite to applying individual competences.

Therefore, we propose to make the inquiry process more explicit within definitions of data literacy as a way of framing a set of data literacy competencies. We propose that within the inquiry framework there is a hierarchy of knowledge and skills related to working with data within a real-world context, ranging from a foundational level of understanding of inquiry, through to very specialist technical skills for hands-on data handling. Figure 2 shows a first step towards mapping the space of data literacy skills, within which the set of competencies can be formed.

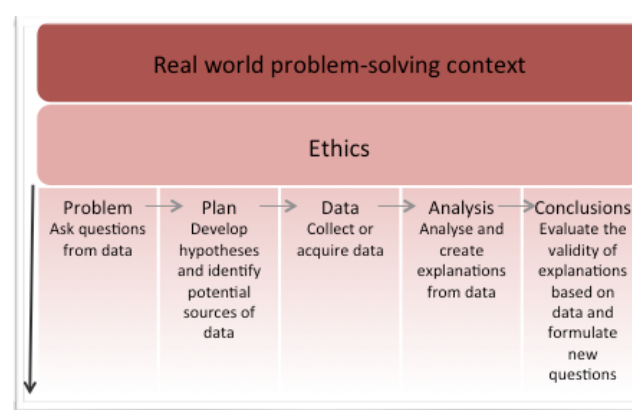


Figure 2. The space of data literacy skills. The arrow within PPDAC activities reflect more specialised data handling skills

Interactions with Data

Having begun to understand better the relationship between different definitions of data literacy and their place within a framework for inquiry, we now consider the situations in which data literacy skills are applied. This is necessary to ensure that our proposed definition of data literacy fulfills the needs of users.

Data as Evidence

In modern life, people are interacting with data on a daily basis. People contribute data through their actions, leaving digital traces of their lives and habits. Data, by itself, has little value. The value is provided through interpretation of one or more data sets in a given context. Collected data is processed and presented in a variety of different ways, to support news articles, advertising, consumer advice, political debate, or policy-making. The act of preparing and communicating this information can be undertaken by various groups such as journalists, advertisers and marketers and community advocacy groups, to name a few. In turn, people use data to help them to make decisions about what to buy, who to vote for, where to live, how to invest money or where to send their children to school [26]. In this scenario, the citizen acts as a reader of data, gaining knowledge from data and from other peoples presented interpretations of it for decision making, but not manipulating the data directly themselves. The data literate reader has the ability to properly evaluate the evidence that is presented in these scenarios, so that they can make critical judgments on the reliability of the information presented and can better understand how their own contributed data is being utilized and make more informed decisions when deciding what data to make available [29] [11]. Without data literacy, there is a risk that the data reader accepts biased interpretations of data as fact, leading to incorrect knowledge, or worse, to bad choices. Without data literacy, those communicating about data can inadvertently introduce bias. Of course, conversely, data literacy can be used to present an ‘on the surface’ plausible, biased, account with the intent to deliberately mislead.

Data as a Tool for Innovation

Large amounts of data are being made available as part of the Open Data movement [15][19]. There is a growing interest in how this data can be used, particularly (although not exclusively) within Smart City applications that use data for urban innovation and to find solutions to improve sustainability of expanding urban spaces. City officials might use this data to inform city planning and policy making. Companies create marketable products and apps. These are top-down approaches to smart city development. However, the real intention of open data is that it is a resource for anyone. Citizen use of urban data is of particular relevance for smart city innovation. In bottom-up smart cities, citizens are drivers for change, better placed for understanding their own local problems and proposing solutions that take citizens needs more fully into account [27][10].

Moving from top-down to bottom-up smart city innovation shifts citizens from a role as passive users/consumers of technology and contributors of data, through to active participants in consciously contributing data to drive smart city applications and identifying problems that could be solved with data and finally to innovators who shape and implement solutions to

urban problems [35]. These active participants and innovators are analogous to makers, or hobbyists (e.g. see [27]), using data as part of a broader DIY tool-kit to individually, or collaboratively, design solutions to urban problems through the process of bricolage. In addition to the data literacy skills of a consumer, the data literate maker needs a range of ad-hoc competencies depending on the focus of their innovation. These include (but are not limited to) knowing how to find and use the open data that is published about their environment, how to generate and use data, e.g. from sensors and how to integrate data with other technologies, for example building apps, dashboards or eco-visualisations. Data literate makers have an awareness of how their own data contributions impacts on the reliability of smart city technologies [5]. They are sensitive to concerns of privacy and ethical use of data. They need to understand and possibly produce visualisations of complex urban data sets. We define urban data as data that informs about environmental, social and economic aspects of urban life. It is typically generated by the activities of people living in a city. The data relates to a variety of topics, such as air and water pollution, energy and water consumption, crime, public and private transport, or car parking to give some examples. Some of the data is available in the form of live data streams that are real-time indicators of the state of the city system (such as traffic flow). Such data streams can generate large data volumes that show variations across days, seasons and years, and – more often than not - are affected by data quality issues. They can cover a large geographic area and/or time-span. Urban data may be collected through sensors, smart meters, satellite imagery or mobile phones, or derived from surveys and questionnaires.

Therefore, the data literate maker is as comfortable with using external data as their own personally collected data. Without data literacy, citizens run the risk of trying to solve problems that are not borne out by available data, that cannot be solved through data analysis at all, or where they miss the opportunity to back up their arguments. They might take and combine data sets that reveal sensitive information about other citizens, or share inappropriate information.

Data Literacy for Job Opportunities

In the U.K. there is an identified skills shortage for data scientists in the jobs market. Twenty seven percent of the biggest employers (with over 250 employees) now use data analytics to support their work [28]. But businesses also report that they cannot hire the data scientists they need to analyse their growing datasets and ensure their competitiveness [1]. A similar deficit has been identified in other countries, particularly the U.S. Unlike the consumer or maker, a scientist needs some formal qualifications to demonstrate the level and area of their expertise. They might specialise in a particular aspect of data, such as visualisation, or machine learning. A data scientist must combine data literacy skills with in-depth knowledge of the company in order to provide meaningful insights and reports. The data scientist is tasked to communicate findings to non-data experts and must also be adept at communicating with data and engaging the audience, for example through novel visualisations. The audience, who may not be data experts, are in the role of a consumer and therefore also need data literacy skills in order to critically assess the validity of the information being presented. Without data literacy, a scientist might apply complex methods and analysis to large data sets, yet fail to address the needs of the company.

Types of Data Literate Citizen

Through the above analysis we have revealed four types of citizen according to the situations in which they would need to use data intelligently for solving real world problems. These are:

- (1) Communicators – who make sense of and tell stories about data for others to digest.
- (2) Readers – who need skills to interpret data that is increasingly presented as part of their every day life.
- (3) Makers – who need the skills to integrate data into broader overall strategies for identifying and solving real-world problems and to be actively conscious of their own data contributions that drive smart city applications.
- (4) Scientists – who need to combine strong technical data skills with communication skills and in-depth knowledge of the domain of the data.

We accept that this list is almost certainly not exhaustive, but suggest that it is a good basis for considering whether the type of data literacy required is the same in every case. Does a citizen using data literacy skills for understanding their environment have the same needs as a data literate citizen applying for a job as a data scientist?

Bridging the gap between definitions and citizen needs

In order to bridge the gap between existing definitions of data literacy and the needs of citizens, we draw upon our merged definitions of Foundational competencies in Table 1 and our space of data literacy skills in Figure 2. We propose that the foundational competences represent the essential skills of a data literate citizen in the role of reader, who can apply their knowledge of a data-driven inquiry process, in which the data is a reflection of real-world phenomena, and can use this knowledge to critically assess data-related arguments. As a general rule, the complexity of data skills required to act in the different roles increases from the role of reader through to the role of scientist. This is shown in figure 3.



Figure 3. Complexity of skills for differing roles

What is most important to note, however, is that within each role the citizen must build upon foundational knowledge to acquire the specialist skills they need to achieve their individual goals. Thus, the skills acquired in each role are dependent on the needs of that role and to some extent on the goals of an individual.

For example, the ‘maker’ citizen may require more specialist skills in some areas that are pertinent to a smart city project. Let’s suggest a citizen who has no prior experience of

working with sensor data is interested in discovering how the traffic is affecting pollution in different parts of their neighbourhood. In this scenario, they would need to either develop expertise in setting up pollution sensors themselves, or use their working knowledge of sensors to communicate their needs to someone with the technical knowledge, or find this information from an open dataset. In order to identify their own knowledge gap in the first place, they must have an understanding of how to plan for data collection, know how data can be identified and obtained and conceive how this data might eventually provide an answer to their initial question. These skills can be identified and learned within the context of our proposed data literacy framework.

In this case, we can reimagine the space of data literacy (figure 2) as we have in figure 4, surrounding a pool of data skills that can be increasingly specialized, and which can be drawn upon as needed to support a data driven inquiry process. The deeper the pool of knowledge of an individual citizen, the more capable they are to act in the role of maker or scientist. This reflects also that specialist skills cannot always be neatly identified with specific foundational competences or stages of inquiry. For example, some selection and analysis of data could form part of evaluating the validity of statistics presented to support an argument, or in an exploratory phase of planning an inquiry.

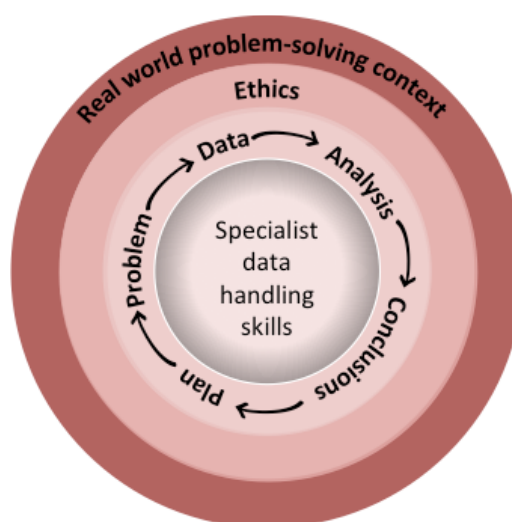


Figure 4. Data literacy pool

Teaching data literacy

How do citizens become data literate? The general view is that the foundation for creating a data-literate society starts in school. However, the importance of data literacy is yet to be reflected in current teaching practices in either schools or colleges [16]. Whilst a typical mathematics curriculum does include provision for teaching data handling and statistical analysis, it is suggested that a more effective approach would be to teach data literacy as a cross curricular subject, incorporating it into subjects such as science and geography [29] [21]. In this way, the concepts and methods learnt would be better contextualized and this would lead to more transferable skills. In order for this to occur, there is a need for a common vocabulary [31], such as through our proposed data literacy framework and competencies.

Similarly, datasets used in school are typically small and do not properly reflect the size and complexity of the types of data that are becoming available, some of which are generated and analysed in real-time [24]. Watson [31] points to the need to choose at least some classroom examples from the real-world setting rather than from artificial settings in text books. Whilst Williams et al. [34] propose that students collect their own data to analyse on the basis that students better understand data when they are an active participant in collecting the data, this does not provide experience of handling data that is neither small nor personally collected. Modern data collection methods, such as sensors, are not yet common in the classroom. However, sensors could act as a stepping stone between small personal data collection and large external data sets, since they allow the collection by a student of a large amount of data in a relatively short time [13].

At the same time, techniques for visualizing data are increasing rapidly. The line between visualization and art becomes blurred, as new visualisations are designed to firstly draw the reader in with aesthetic appeal and then to offer an interpretation of the data behind the visualization. It is not always the case that skills learnt with small datasets, based around fairly traditional graphs and charts, can be adapted and scaled up to working with these more complex types of data [17], or extended to producing and interpreting these new types of visualization. There is therefore an argument for using complex data sets and visualizations as part of classroom teaching materials for data literacy.

Existing Approaches to support data literacy teaching

A number of approaches and online platforms exist for teaching data literacy in schools. We have evaluated four of these systems against our proposed data literacy framework with a view to understanding how the teaching of data literacy is interpreted in practice and how they might inform or support a new definition of data literacy.

Maine Data Literacy Toolkit (MDLK)

The MAINE data literacy project (<http://participatoryscience.org/project/maine-data-literacy-project>) defines a data literacy framework for teachers and students to use data in schools to improve data literacy. The project provides real data sets, from online sources simplified for the classroom. The Data Literacy Framework is built upon two foundational skills, firstly to be able to visualize variability in a data set using frequency plots and secondly to describe variability in data, e.g. in terms of range, measures of central tendency, etc. Students are then expected to demonstrate what they have learned in the context of a stated question. Finally, they should be able to explain how their choice of visualisation supports their interpretation of the data with respect to the question. Students repeat the process with increasingly complex data, for example starting with small datasets that can be analysed by hand to more complex data that requires special software. The resources themselves are provided as multiple set of very detailed instructions for visualising or analysing specific datasets, with little guidance or support for students to guide their own explorations of the data.

Kids' Survey Network (KSN)

Kids' Survey Network (www.kidsurvey.org) provides a set of games and videos designed to teach children about conducting survey research and aims to promote 'the responsible and

successful use of data by future citizens and workers'. Students design and conduct questionnaires in order to answer their own questions or address issues of concern in their communities. Students learn to analyse survey data, interpret their results and present their findings. In this context, students are encouraged to think about how to use data responsibly.

Tuva Labs (TL)

Tuvalabs (<https://tuvalabs.com/>) provides access to numerous data sets linked to learning resources. Each data set has a set of standard graph based visualisations. Teachers and students first select a data set, then a visualization type and then choose the attributes to place on the chart. Teachers and students can use suggested questions, or else make up (and optionally add to the system) their own question. Datasets are available for download, which gives the opportunity for a student to use the data to create visualisations from external software. However, Tuvla is limited to questions that can be answered through graph rather than map-based visualization. Whilst the Tuvla data browser provides immediate interactivity with each data set, allowing students to explore the data in a very hands-on way, it is unclear to what extent the approach emphasises the inquiry aspects. For example, supporting students in using data as part of a self-directed real-world inquiry, in which they use the data to formulate and answer their own questions.

Citydigits (CD)

Citydigits (<http://www.citydigits.org/>) was a project that encouraged students to answer questions about their local neighbourhood through geo-spatial data analysis. In Local Lotto, students combined interview data with data about lottery spending in different neighbourhoods to form an opinion about whether the lottery was good or bad. They prepared digital storyboards to communicate their findings, using the evidence from their analysis to reinforce their opinion. In Cash City, students use maps to analyse the distribution of pawn shops and combine it with real data about the cost of pawning and also their own survey data, in order to understand the role of pawn shops in their community. Similar to the Local Lotto, students communicate their overall opinions at the end of the analysis and also have the opportunity to comment on the opinions of other groups of students in the same task. In Citydigits, students use real data within a structured inquiry task, but do not have the opportunity to formulate their own questions from the data.

The Urban Data School (UDS)

The Urban Data School is an initiative that aims to use real, complex, urban data sets that are being collected as part of a smart city project (www.mksmart.org) as a resource for teaching data literacy in school [36]. The approach is designed to teach students to ask and answer questions from data through a data-driven inquiry process, with consideration to the ethical use of data, and to understand how data can be used to drive urban innovation.

The additional specialist skills reflect the origin of the project in Smart City technologies and focus on skills required to create data literate 'makers', comfortable with both small, personal data and large, complex externally sourced ones. In UDS sessions, students have been asked to analyse and create explanations from real data sets, to collect and visualize their own data and to design phone apps that used data to solve a problem, such as helping them to be more

considerate about home energy consumption (see figure 5 for one example). Students present findings to the group describing what they have discovered through data analysis.

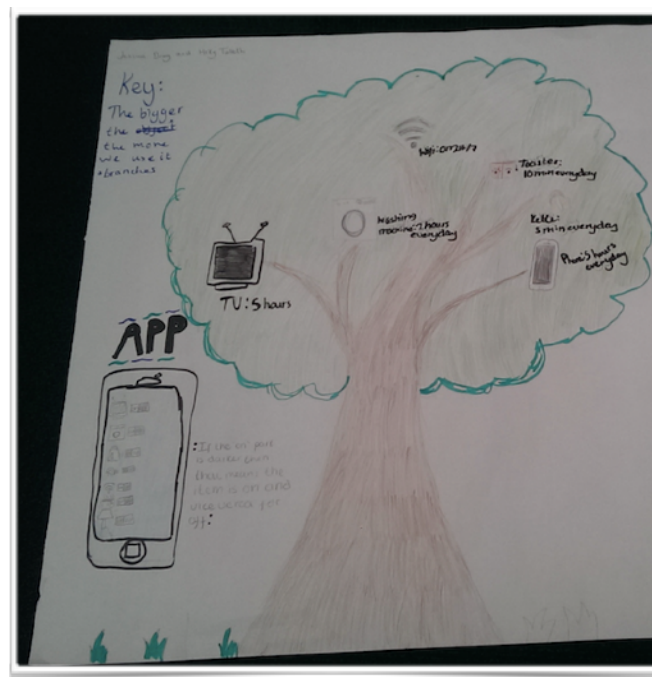


Figure 5. Visualization of home energy data as a tree. Bigger branches indicate bigger use of the appliance

Summary of Existing Approaches

Taken collectively these existing approaches offer a suite of methods, resources and tools for teaching important data literacy competences. Each approach has individual strengths and weaknesses, for example Tuvalabs offers access to a large quantity of open data, but lacks a structured inquiry approach. Both City Digits and the Urban Data School approach provide a strong context within which the task is conducted and combines multiple types of data. Students are very active in data collection, analysis and in interpreting and presenting their findings with respect to the questions. However, both are very focused on the specific domain of the data and the neighbourhood in which the data is located. It is not clear to what extent the teaching materials would be useful in other neighbourhoods, cities, or countries. City Digits does not provide opportunity for students to learn how to pose their own questions and plan for data collection, but this is an important part of the Urban Data School approach. KSN focuses too much on personally collected data, whereas MDLK provides a lot of external data but is difficult to navigate and does not appear to give students much leeway in learning how to plan their own research questions and data analysis.

What these approaches have in common is that they demonstrate that it is possible to successfully engage students to interact with real world data sets across a broad spectrum of data literacy competences.

Table 2. Analysis of existing approaches to support data literacy

	MDLK	KSN	TL	CD	UDS
Data	Real world (environmental) data	Personally collected survey data	Multiple real world open datasets	Real world (urban) data	Real world (urban) data
Real world context	Yes	Yes	Yes	Yes	Yes
Ethics	No	Yes	No	No	Yes
Inquiry process?	Yes	Yes	No	Partial	Yes
Additional data literacy skills	Large data sets, visual analysis, statistical analysis	Combining quantitative and qualitative data	Large data sets	Combining quantitative and qualitative data Geo-spatial data analysis	Novel visualisations, large data sets

A definition of data literacy

Based on the analysis of definitions of data literacy, the mapping to user needs and the survey of existing approaches to teaching data literacy in practice, we propose this following definition of data literacy:

Data literacy is the ability to ask and answer real-world questions from large and small data sets through an inquiry process, with consideration of ethical use of data. It is based on core practical and creative skills, with the ability to extend knowledge of specialist data handling skills according to goals. These include the abilities to select, clean, analyse, visualise, critique and interpret data, as well as to communicate stories from data and to use data as part of a design process.

This definition both encompasses and extends previous definitions of data literacy. It highlights the importance of learning the process of a data inquiry, of obtaining experience with both small and larger, more complex, data sets and that data literacy skills should be acquired through real world data and tasks. This definition is applicable across the identified user scenarios and is reflected in the approaches that have been evaluated for teaching data literacy in schools.

Future work

The definition of data literacy as a space within which a number of competencies can be formed opens up several research areas for further investigation.

Of particular interest is the use of data as a resource for formal and informal learning of data literacy skills. Topics of investigation might include how to make data more salient in order to help learning, for example through tangible data, visualizations, or better tools for interacting with data. The increasing use of complex data and visualisations raises many questions in terms of how people identify the important dimensions of data from these larger, more complex datasets or their non-standard visualisations.

Schools are an obvious starting point for teaching data literacy skills. Also of interest is how to provide learning resources for ‘maker’ citizens for the ad-hoc acquisition of skills necessary to identify and solve local problems. How is it possible to raise awareness of data that is (or could be) available within an environment to support citizens in innovating solutions to local problems?

Conclusions

The contribution of this work is a framework for defining data literacy that draws upon existing definitions and extends them to explicitly include the process of inquiry through which data literacy skills are applied. This framework allows for the identification of a set of foundational competences, against which data literacy expertise can be assessed and through which more specialist skills can be learned according to the needs and goals of the data literate citizen. Four types of citizen are currently identified, these are communicators, readers, makers and scientists. We have suggested that it is important to include teaching of data literacy as part of the curriculum in schools to ensure that future citizens achieve the required level of data literacy. We demonstrate examples in which real data sets have been combined with learning resources to teach data literacy.

References

1. Bakhshi, H., Mateos-Garcia J., & Whitby, A. (2014). Model Workers: How leading companies are recruiting and managing data talent. Retrieved from: <http://www.nesta.org.uk/publications/model-workers-how-leading-companies-are-recruiting-and-managing-data-talent>
2. Bhargava, R., & D’Ignazio, C. (2015). Designing Tools and Activities for Data Literacy Learners. Workshop on Data Literacy, Webscience 2015.
3. Callingham, R. (2006). Assessing Statistical Literacy: A Question Of Interpretation? *International Conference on Teaching Statistics (ICOTS7)*
4. Carlson, J., Fosmire, M., Miller, C. C., & Nelson, M. S. (2014). Determining data information literacy needs. *Data Information Literacy: Librarians, Data, and the Education of a New Generation of Researchers*, 11.
5. Celino, I., Contessa, S., Corubolo, M., Dell’Aglia, D., Della Valle, E., Fumeo, S., & Krüger, T. (2012). Linking smart cities datasets with human computation—the case of urbanmatch. In *The Semantic Web—ISWC 2012*(pp. 34-49). Springer Berlin Heidelberg.
6. Data Journalism Handbook (n.d.). Retrieved from: http://datajournalismhandbook.org/1.0/en/understanding_data_0.html
7. Deahl, E. (2014). Better the Data You Know: Developing Youth Data Literacy in Schools and Informal Learning Environments. M.S. Thesis, Massachusetts Institute of Technology.

8. Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educational process*. Lexington, MA: Heath
9. Freire, P. (2000). *Pedagogy of the oppressed*. Bloomsbury Publishing.
10. Gooch, D., Wolff, A., Kortuem, G., & Brown, R. (2015, September). Reimagining the role of citizens in smart city projects. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers* (pp. 1587-1594). ACM.
11. Haddadi, H., Howard, H., Chaudhry, A., Crowcroft, J., Madhavapeddy, A., & Mortier, R. (2015). Personal data: Thinking inside the box. *arXiv preprint arXiv:1501.04737*.
12. House of Lords Select Committee on Digital Skills. 2015. Make or Break: The UK's Digital Future. Retrieved from: <http://www.publications.parliament.uk/pa/ld201415/ldselect/lddigital/111/111.pdf>
13. Lee, V.R., & Thomas, J.M. (2011) Integrating physical activity data technologies into elementary school classrooms. *Educational Technology Research and Development*, 59(6), 865-884.
14. Mandinach, E. B., & Gummer, E. S. (2013). A systemic view of implementing data literacy in educator preparation. *Educational Researcher*, 42(1), 30-37.
15. McAuley, D., Rahemtulla, H., Goulding, J., & Souch, C. (2014). How Open Data, data literacy and Linked Data will revolutionise higher education. Retrieved from: <http://pearsonblueskies.com/2011/how-open-data-data-literacy-and-linked-data-will-revolutionise-higher-education/>
16. Nam, T., & Pardo, T. A. (2011). Conceptualizing smart city with dimensions of technology, people, and institutions. In *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times* (pp. 282-291). ACM.
17. National Research Council. (2013). *The Mathematical Sciences in 2025*. Washington, DC: The National Academies Press
18. Nesta. (2015). *Analytic Britain Securing The Right Skills For The Data-Driven Economy*. Retrieved from: <http://www.nesta.org.uk/publications/analytic-britain-securing-right-skills-data-driven-economy>
19. Ojo, A., Curry, E., & Zeleti, F. A. (2015). A Tale of Open Data Innovations in Five Smart Cities. In *System Sciences (HICSS), 2015 48th Hawaii International Conference on* (pp. 2326-2335). IEEE.
20. Petrie, H., Savva, A., & Power, C. (2015). Towards a unified definition of web accessibility. In *Proceedings of the 12th Web for All Conference* (p. 35). ACM.
21. Calzada Prado, J., & Marzal, M. Á. (2013). Incorporating data literacy into information literacy programs: Core competencies and contents. *Libri*, 63(2), 123-134.
22. Schield, M. (2004). Information literacy, statistical literacy and data literacy. *IASSIST Quarterly*, 28(2/3), 6-11.
23. Scanlon, E., Anastopoulou, S., Kerawalla, L., & Mulholland, P. (2011). How technology resources can be used to represent personal inquiry and support students' understanding of it across contexts. *Journal of Computer Assisted Learning*, 27(6), 516-529.
24. Schutt, R. (2013). Taking a Chance in the Classroom: Embracing the Ambiguity and Potential of Data Science, *CHANCE*, 26:4, 46-51
25. Spencer, D. & Warfel, T (2004) "Card sorting: a definitive guide." *Boxes and Arrows* (2004).
26. Steen, L. (1999). Numeracy: The new literacy for a data-drenched society. *Educational Leadership*, 57(2), 8-13.
27. Tanenbaum, J. G., Williams, A. M., Desjardins, A., & Tanenbaum, K. (2013). Democratizing technology: pleasure, utility and expressiveness in DIY and maker practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2603-2612). ACM.
28. Tech Partnership. (2015). Employer insights: skills survey. Retrieved from: https://www.thetechpartnership.com/globalassets/pdfs/tec_employer_skill_survey_web.pdf
29. Vahey, P., Yarnall, L., Patton, C., Zalles, D., & Swan, K. (2006). Mathematizing middle school: Results from a cross-disciplinary study of data literacy. In *Annual Meeting of the American Educational Research Association, San Francisco, CA*.

30. Wallman, K. K. (1993). Enhancing statistical literacy: Enriching our society. *Journal of the American Statistical Association*, 88(421), 1-8.
31. Watson, J.M. (2011). Foundations for improving statistical literacy. *Statistical Journal of the IAOS*
32. White, B. Y., Shimoda, T. A., & Frederiksen, J. R. (1999). Enabling students to construct theories of collaborative inquiry and reflective learning: Computer support for metacognitive development. *International Journal of Artificial Intelligence in Education (IJAIED)*, 10, 151-182
33. Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review/Revue Internationale de Statistique*, 223-248.
34. Williams, S., Deahl, E., Rubel, L., & Lim, V. (2014). City Digits: Local Lotto: Developing Youth Data Literacy by Investigating the Lottery. *Journal of Digital Media Literacy*
35. Wolff, A., Gooch, D., Mir, U., Caverio, J., & Kortuem, G. (2015). Removing barriers for citizen participation to urban innovation. In: *Digital Cities 9*, Limerick.
36. Wolff, A., Kortuem, G., & Caverio, J. (2015). Urban Data in the primary classroom: bringing data literacy to the UK curriculum. In: *Data Literacy Workshop*, Oxford.