


Chapter 2

Opportunities for Adopting Open Research Data in Learning Analytics

Katarzyna Biernacka

 <https://orcid.org/0000-0002-6363-0064>
Humboldt University of Berlin, Germany

Niels Pinkwart

Humboldt University of Berlin, Germany

ABSTRACT

The relevance of open research data is already acknowledged in many disciplines. Demanded by publishers, funders, and research institutions, the number of published research data increases every day. In learning analytics though, it seems that data are not sufficiently published and re-used. This chapter discusses some of the progress that the learning analytics community has made in shifting towards open practices, and it addresses the barriers that researchers in this discipline have to face. As an introduction, the movement and the term open science is explained. The importance of its principles is demonstrated before the main focus is put on open data. The main emphasis though lies in the question, Why are the advantages of publishing research data not capitalized on in the field of learning analytics? What are the barriers? The authors evaluate them, investigate their causes, and consider some potential ways for development in the future in the form of a toolkit and guidelines.

DOI: 10.4018/978-1-7998-7103-3.ch002

INTRODUCTION

The movement to publish datasets has been growing for some time now. Research institutions, funders, a growing number of publishers, and even the research communities themselves, promote the publication of research data (DCC (Digital Curation Centre); Deutsche Forschungsgemeinschaft, 2019; European Commission, 2016; L. Jones, Grant, & Hrynaszkiewicz, 2019; Kim, 2019). Although the benefits of sharing data are already known (Heather A. Piwowar & Vision, 2013), Learning Analytics data has still held back. One of the reasons for this could be the large amount of personal data collected by the Learning Analytics systems. The strict data protection regulations and the anonymization procedures seem to prevent scientists from sharing their data, or at least make it more difficult (Biernacka, & Pinkwart, 2020).

The Humboldt-Elsevier Advanced Data and Text Centre (HEADT Centre)¹ has set itself the goal of exploring the various facets of research integrity. The EU General Data Protection Regulation (GDPR) plays an important role for research integrity, as do the legal regulations of other countries and regions. One of the central topics of the initiative is therefore to investigate the legal regulations as an aspect of research integrity. The answer varies across disciplines and it is especially relevant when research data includes personal data. The degree of data protection, however, may interfere with transparency, which is a key value of research integrity. The goal of this research project is to investigate the conflict between publication of research data and the issues of privacy, and to identify and test solutions, considering both differences between disciplines and between cultural perspectives.

In this chapter the authors explore the handling of Learning Analytics research data with a focus on the publication process. It begins with a comprehensive introduction into the movement of Open Science, and then proceeds to the topic of Open Research Data. This foundation is necessary to understand the difficult situation in the field of Learning Analytics regarding this movement. The chapter continues with a look at the barriers of publishing research data in Learning Analytics, based on studies conducted in Germany, Peru, India and China. In the final part of the chapter, the authors intend to provide guidance to scientists in Learning Analytics. Furthermore, the authors offer possible practical solutions for the publication of research data in this discipline. The chapter ends with a conclusion.

BACKGROUND

What is Open Science?

The literature has not yet agreed on a definition of Open Science, as different actors within the scientific process have different ideas on what should be opened up. The most used and cited definition though, is the informal one from Nielsen (Gezelter, 2011): “Open science is the idea that scientific knowledge of all kinds should be openly shared as early as is practical in the discovery process.” Vicente-Saez and Martinez-Fuentes (2018) define it as “(...) transparent and accessible knowledge that is shared and developed through collaborative networks”. In general, Open Science means opening up the research by making all of its outcomes publicly available with the goal of dissemination and re-use of knowledge for a better world. Open Science is thus a welfare-enhancing phenomenon that enables transparent, accessible, shared, collaborative and rapid public disclosure of new knowledge.

The openness, as a key principle of Open Science, creates new opportunities for researchers, decision makers, platform programmers and operators, publishers and the general public (Fecher & Friesike, 2014; Open Science and Research Initiative, 2014). For each of these stakeholders the term Open Science has a different meaning and concerns different areas. There is often a confusion between the principles, practices, outcomes or processes regarding Open Science. Therefore, it was decided on a taxonomy, including nine terms used at a first instance: Open Access, Open Data, Open Reproducible Research, Open Science Definition, Open Science Evaluation, Open Science Guidelines, Open Science Policies, Open Science Projects and Open Science Tools (see *Figure 1*) (Pontika, Knoth, Cancellieri, & Pearce, 2015).

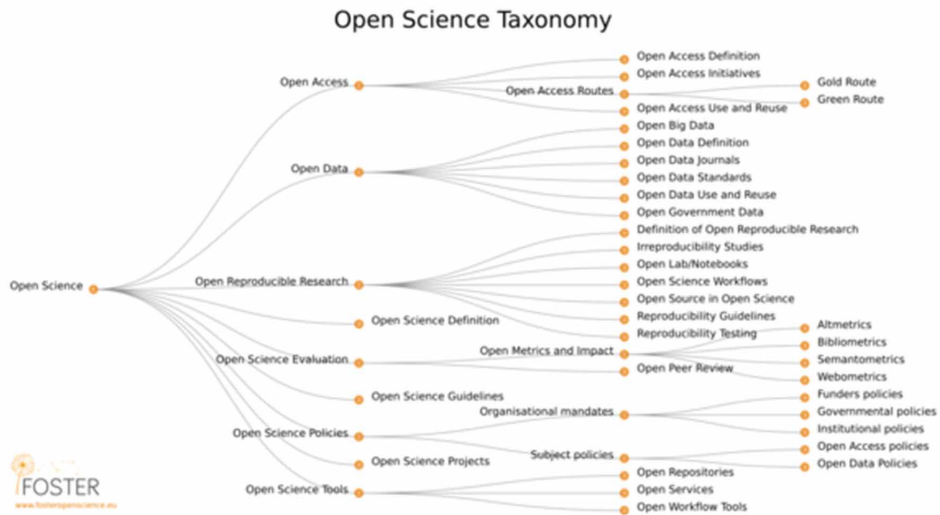
Fecher and Friesike (2014) decided to structure the discourse by proposing five Open Science schools of thought: the infrastructure school, the public school, the measurement school, the democratic school and the pragmatic school. Between these schools there is no clear cut, they can share some ontological principles. *Table 1* shows the central ideas of each school.

Table 1. Open Science Schools of Thoughts (Fecher & Friesike, 2014)

School of thought	Central idea
Infrastructure	Develop openly available platforms, tools and services for efficient research
Public	Encourage the public to collaborate in research through citizen science, and make science more understandable and accessible for the public
Pragmatic	Open up the scientific process and increase the effectiveness of research and knowledge dissemination
Democratic	Make knowledge freely accessible to everyone
Measurement	Find new standards for the determination of scientific impact

Figure 1. Open Science Taxonomy

Source: (Pontika et al., 2015)



The infrastructure school concerns, as the name already says, the technical infrastructure. The advocates of this school emphasize that openly available platforms, tools and services are needed for efficient research. They see Open Science as a technological challenge to enable research on a bigger, wider scale. The infrastructure is a key element in all the subsequent school of thoughts: repositories, collaborative writing tools or storage.

The public school encourages the public to collaborate in research. The advocates of this school argue that science needs to be accessible and comprehensible for a broader public and interested non-experts. The research process can be made open and accessible, the audience can participate in the research process or just observe/follow it. A very well-known example for this stream is Citizen Science (Catlin-Groves, 2012; Irwin, 1995), e.g. zooniverse.org, which enables everyone to take part in real research in many different disciplines. This stream is possible through the new technologies that have arisen since Web 2.0. Alternatively, the researchers can open their results to the public in more comprehensible way than in the common scientific article. Examples of science communication in the context of this tenet of the public school are (micro)blogs (Ebner & Maurer, 2008), articles in non-scientific journals or talks, e.g. TEDTalks (TED, 2020).

The pragmatic school wants to make research and knowledge dissemination more efficient in optimizing the research process, e.g. opening the scientific value chain, including external knowledge or allowing collaboration through online tools.

Nielsen (2012) shows on the example of the Polymath Project² how science can shift from closed to collaborative. Experts from different institutions and countries can work together using an online tool, e.g. Wiki.

The democratic school of Open Science focuses on the accessibility of research products, in particular on the free access to research publications and research data. Thus, the two main streams emerging from the democratic school are Open Access and Open Data. In this section the authors will focus on Open Access, Open (Research) Data will be highlighted in the next section.

Open Access to research publications – in particular peer-reviewed journal articles - means the “free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited” (Chan et al., 2002). This term was established by the Open Access Budapest Initiative³ in 2002.

Since then, Open Access has grown in importance and a variety of full Open Access journals have been launched, e.g. PLOS. Still there has been a resistance to publish in these journals, as the subscription-based journals retained the highest impact factor and this measurement is still important for the evaluation of the scientific impact, and therefore for the reputation of the researchers. One of the solutions for this problem is the publication of unpublished works on preprints servers, e.g. arXiv⁴ or PeerJ Preprints⁵. In some domains, computer science and physics among others, this practice is already well established (Gentil-Beccot, Mele, & Brooks, 2009; Larivière et al., 2013). The benefits of submitting unpublished work to a preprint server include, free and fast dissemination and citeability. For the latter case DOIs are assigned, which gives a time stamp to the preprint, which can be important for priority claims too.

Another solution to the problem of impact factors is encompassed by the measurement school which aims to find new standards for the determination of scientific impact. Rentier (2016) successfully draws the comparison between social processes to achieve prestige, and peer review. He shows that heredity, courtship or clubbing can also occur in the latter case. This could be prevented e.g. with Open Peer-Reviews by also making this process more transparent. The value of a scientific publication is currently defined by the reputation of the journal or collection in which it is published (Journal Impact Factor) and not by the quality of the article itself. According to this school of thought, an alternative and faster impact measurement that includes new forms of publication is needed. The umbrella term for this new

impact measurements is altmetrics (Priem, Taraborelli, Groth, & Neylon, 2010). Altmetrics include tweets, blog discussions, bookmarks (e.g. on Mendeley or any research network), HTML views, citations and use in public APIs across platforms to gather data with open scripts and algorithms. According to the manifesto of Priem et al. (2010), altmetrics are great for measuring impact in this diverse scholarly ecosystem, tracking impact outside the academy, impact of influential but uncited works, and impact from sources that aren't peer-reviewed. Altmetrics expand the idea of what scientific impact nowadays consists of (see *Figure 2*).

Figure 2. Four ways to measure the impact of an article

Source: Priem et al. (2010) licensed under the Creative Commons CC BY-SA license



A different perspective to look at Open Science is throughout the research lifecycle (see *Figure 3*). From there, five main aspects of Open Science may be identified: Open Data, Open Methodology, Open Access (Open Paper), Open Peer-Review (Open Evaluation) and Open Source (Open Code). Additionally, as an important part of research: Open Educational Resources. In the next section the focus will be on Open Data.

The Push to Open Research Data

In the past, data was rarely public. There can be several reasons for this, but one of the most important was certainly the medium: paper is not a good data storage medium. The digital world has now opened up new possibilities and thus the call for open data. The change of technology has made data collection, storage, and sharing more feasible and the movement has been driven by increasing amount of data sharing policies and mandates from research funders and journals.2014)

Opportunities for Adopting Open Research Data in Learning Analytics

Figure 3. Opening up of the research process

Source: Based on European Commission (2014)



Open Data as one of the tenets of the democratic school of Open Science promotes the equal right to access knowledge – in this case to access data. The definition of “openness” is based on the Open Definition: “Open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness).” (Open Knowledge Foundation, 2015) and more specific: “Open data and content can be freely used, modified, and shared by anyone for any purpose” (Open Knowledge Foundation, 2015). This means that data has to be interoperable and give researchers the ability to interoperate - or intermix - different datasets (Open Knowledge Foundation, 2013). This type of data disclosure also makes it possible to create more transparency in science. Here we speak of Open Research Data.

Research transparency was already put into focus by government leaders and funders, to rebuild the trust in science. In 2011 the UK House of Commons Science and Technology Committee examined research integrity and the peer review process and concluded that “Access to data is fundamental if researchers are to reproduce and thereby verify results that are reported in the literature” (House of Commons

Science and Technology Committee, 2011). Frauds, such as those highlighted in a discussion of reproducibility issues by Ince (2011) can be avoided. The early publication of research data can thus help to reduce misconduct, facilitate replication, and support further research and collaborations.

Yet still, the data availability in many disciplines is not a common practice. For many years the quality of scientific work was judged on the conclusions drawn from the data, rather than on the data itself. This led to a poor understanding of data management along the scientists (e.g. missing descriptions, bad preservation of the data) and a general mistrust in the data produced by others. The concerns about data quality resulted in a reluctance to sharing or publishing research data.

For the purpose of this chapter the difference between sharing and publishing should be determined. *Sharing* describes making data available to other researchers (mostly on demand). No persistent identifier is assigned to the data, and it's hard to verify the provenience of the data, or to cite the data. Data can be shared personally, via repositories or through other communications platforms. Putting resources on a website would be public sharing, while sharing it internally with collaborators – private sharing. However, simply having data available or shared is not of much use. It is not guaranteed that data put on a website will still be there in 3 years. To raise overall research transparency, the transparency of the whole data creation process is needed. In the latter case, when the data is *published*, it should fulfil four criteria: it should be available, documented, citable and validated (Kratz & Strasser, 2014)⁶. To meet these criteria, it is important to document the research data extensively and to provide them with (subject-specific) metadata. This increases the traceability and findability of the research work among the peers. The next step is to choose a suitable, subject-specific repository that is relevant to the community⁷. In order to make the research data available and citable over the long-term, it is important to assign a persistent identifier to the data. Most often, the repository assigns a Digital Object Identifier⁸ (DOI) at this point. This makes the research data uniquely referenceable. The biggest hurdle to overcome is the data validation. It is difficult to decide what criteria can be used to evaluate the quality of research data, in particular because it can be distinguished between technical and scientific evaluation (Callaghan et al., 2012).

Besides the governments and funders, various institutions already demand data accessibility (publication) too. ALLEA (2017, p. 6) – All European Academies – requires in the European Code of Conduct for Research Integrity, that “Researchers, research institutions and organisations ensure access to data is as open as possible, as closed as necessary, and where appropriate in line with the FAIR Principles (Findable, Accessible, Interoperable and Re-usable) for data management” and “Researchers, research institution and organisations provide transparency about how

to access or make use of their data and research materials". The FAIR Principles were published in 2016 (Wilkinson et al., 2016) and are intended to act as guideline for enhancing the re-usability of data. Besides to the requirements for findability, accessibility and the assignment of persistent identifiers (see criteria for published data as above mentioned), interoperability is also important here. The data should be available in such a way that it can be exchanged, interpreted and integrated with other data sets (re-used).

Not all published data is FAIR data by definition, and not all FAIR data is open though. In order to open the data in the sense of Open Science is to minimize the *usability* restrictions. The minimum requirement of Open Data is to have open terms of use (open licenses). The most frequently used licenses for research data are the Creative Commons⁹. Care should be taken to ensure that the re-use conditions are as "open as possible and as closed as necessary" (ALLEA, 2017, p. 6). Of the seven licenses they offer, three are in line with Open Science: CC0, CC BY and CC BY-SA. The other four are too restrictive.

To achieve greater openness of data, paywalls have to be avoided and machine-readable, non-proprietary formats and open standards used. This higher degree of openness is, where FAIR data meets and overlap Open Data.

In this chapter the authors focus on these research data that fulfil the ideas of published and open research data that meet the FAIR Principles. In the following sections, the Open Research Data¹⁰ in Learning Analytics will be considered.

Advantages of Publishing Research Data

As already shown in the section about Open Data, informal data sharing still seems to be much more common in many disciplines than formal publication of research data (either on a repository or as a data paper). Even though there is evidence that publication of data leads to more citations (Gleditsch, Metelits, & Strand, 2003; Peng, 2011; Pienta, Alter, & Lyle, 2010; Heather A. Piwowar & Vision, 2013), researchers still seem unconvinced.

In addition, many projects are financed by third-party funds - whether from public or private funding agencies. The publication of the data, which is now increasingly demanded by the funders (Colavizza, Hrynaszkiewicz, Staden, Whitaker, & McGillivray, 2019; European Commission, 2016), can at this point also be seen as an investment in one's own reputation. The time spent on proper management, documentation, and the publication process itself pays off in the end, as this data publication can be presented to new potential funders. On the other hand, research data emerged from a public funded project, could be considered as public good that should be made open for the public.

Given the complexity of contemporary science, researchers have to act against fraud and misconduct. Publication of research data helps to promote research integrity and accountability. By making the data available to one's own peers for re-use, one receives direct feedback on the quality of the research, which is verified in this way. This can have a positive effect on researchers' reputations too.

Overall, the exchange of data with colleagues promotes new collaborations and also new insights. Van Horik, Dillo, and Doorn (2013) give examples on how fast the awareness and the practice of data management can positively change. The authors took Archeology, Oral History and Qualitative Social Science, Virology and Veterinary Medicine as an example, where data transparency and open access to data became the new scientific practice.

When publishing research data, the scientist may prevent unnecessary costs for gathering the same data twice. It allows a more efficient allocation of these resources in different projects and to gather more visibility. Furthermore, data put in a repository helps to prevent data loss.

Making data publicly and openly available facilitates therefore the re-use, verification, replication, meta-analysis and robustness check of the research. It supports more efficient and excellent science and leads to increase the trust and confidence in research processes.

Research Data in Learning Analytics

Similarly, digitization has helped to really bring Learning Analytics (LA) into existence. The use of Learning Management Systems (LMS) and Virtual-Learning-Environments (VLE) increased. Learning processes are increasingly taking place online, especially now during the COVID19 pandemic. As a result, large amount of learning and learners' data is generated every day. This information enables learning and teaching to become more personalized (Ferguson, 2012; Long & Siemens, 2011; Papamitsiou & Economides, 2014). These technical advances led to the development of a new field of research: Learning Analytics.

The range of research data in the field of Learning Analytics varies as much as the definition of the subject area itself. Scientists from computer science, educational research, psychology, as well as from all didactic subjects can identify themselves with this field of work. The community is roughly divided into three areas: Learning Analytics and Knowledge, Educational Data Mining and Academic Analytics. With different methods (such as data mining, qualitative analysis or statistics) research data is collected, which should help to model student behaviour, predict performance or make resource recommendations (Papamitsiou & Economides, 2014).

In semi-structured interviews (Biernacka, 2020a, 2020b, 2020c, 2020d), Learning Analytics scientists from Computer Science from Germany, India, China and

Peru have identified the following data types as their research data: process data, questionnaires, interview data, log data, audio-video data, multimodal data produced by sensors (e.g. ECG, EEG, GSR, vital data), assessment data, annotated text data, sociodemographic data, data from learning platforms (e.g. behaviour data), learning performance, online user behaviour, MOOC data, focus group observations, runtime data and many more. One can therefore clearly see the diversity of the research data, both qualitative and quantitative. A general research data management workflow will be only of little help here – all these types need different handling, in particular when legal aspects are considered. The data sensitivity shows large variation, but in most of the cases the scientist indeed deal with personal, or sometimes even sensitive data¹¹.

Barriers of Publishing Research Data in Learning Analytics

Despite the many advantages of publishing research data presented in the section before, in many disciplines data publication is still rare (Alsheikh-Ali, Qureshi, Al-Mallah, & Ioannidis, 2011; H. A. Piwowar, 2011; Schofield et al., 2009; Vanpaemel, Vermorgen, Deriemaecker, & Storms, 2015; Vision, 2010). Some studies already identified factors that prevent researchers from the publication of their research data. They include the “*fear for misuse and misinterpretation of data*” (Van den Eynden et al., 2016), “*the desire to publish results before releasing data*” (Schmidt, Gemeinholzer, & Treloar, 2016), “*lack of journal requirements*” (Lucraft, Allin, Baynes, & Sakellaropoulou, 2019) or “*not common in the community*” (Houtkoop et al., 2018). Besides the barriers mentioned, regular factors are connected to ethical concerns, legal constraints, not having the rights to make data accessible or to the anonymization process are identified (Cheah et al., 2015; Meyer, 2018; Schmidt et al., 2016). Already in these studies it becomes clear, that the different disciplines require different handling of their research data.

However, none of these studies have specifically addressed the concerns and needs of the scientists from Learning Analytics. In the HEADT Centre project, the researchers are looking in particular at the handling of research data and their publication in Learning Analytics in four different countries: Germany, India, China and Peru (Biernacka, 2019; Biernacka & Huaroto, 2020; Biernacka & Pinkwart, 2020). In addition to very different cultural perspectives, the different countries also show great differences in legal regulations. The authors consider distinctive issues that may arise considering these circumstances with the focus on the publication of data about learners’ behaviour and try to find out why the LA researcher are reluctant to publish their research data.

To understand the concerns about research data publishing in their domain, a semi-structured interview study with scientists in Learning Analytics was used. In

total 13 scientists from Germany, Peru, India and China were questioned (compare sampling in *Table 2*). The qualitative research was conducted between July 2019 and January 2020. Both, junior (2 to 4 years of experience) and senior scientists (more than 5 years of experience) were included. Researcher with longer professional experience seemed to be more willing to participate in an interview. Newcomer and scientists in early stages of their careers may have more inhibitions about expressing their opinion. The authors experienced some difficulties in finding researchers in this research field in Peru and India, where the awareness and understanding of Learning Analytics and the related issues in the data-driven society is still missing (Cobo & Aguerrebere, 2018). The terms “analysis of educational data”, “data-based feedback” or “data-based actions” seem to be more common.

Table 2. Sampling for the semi-structured interviews in Germany, Peru, India and China (2019-2020)

	Germany	Peru	India	China
No. of junior scientists	2	0	0	1
No. of senior scientists	3	2	1	4
No. of institutions	5	1	1	2
Total no. of interviews	5	2	1	5

The semi-structured interview study gave an insight of how the research data is handled in LA in general. Questions about their work and the research data their working with were asked. In the second part of the interview, the interviewer asked whether the researcher has published his/her data. Ten of the thirteen interviewed scientists answered “no” to this question, of which four are “*uncertain what is allowed*”. Another person who indicated this factor, published his/her data only aggregated as an evaluation in a paper. This result already shows one of the biggest challenges. This lies in an unclear legal situation with regard to research data. This problem seems to be not only in Germany (or Europe, where the GDPR ¹²applies), but worldwide: both India and China have mentioned this factor too. In the remained case that indicated “*uncertainty what is allowed*” we have no information about whether he/she published the research data.

A junior scientist from Germany concludes:

On the other side, especially because media is big on (unintelligible) about data security and data usage, everyone is very, very insecure: What can I do? (Junior Scientist, Germany (Biernacka, 2020c, p. 3 in os_013))

Opportunities for Adopting Open Research Data in Learning Analytics

In total, 27 different barriers to the publication of research data were mentioned by the scientists (see Figure 4). Those barriers and concerns can be clustered around five dimensions (see Table 3):

- Authority or practice considerations
- Technical or processing constraints
- Legal concerns
- Loss of control of data
- Resource constraints.

Table 3. The five dimensions of barriers to publication of research data emerged from the semi-structured interviews

Authority or practice considerations	Technical or processing constraints	Legal concerns	Loss of control of data	Resource constraints
No extrinsic motivation or obligation	Anonymisation – conducting the anonymization process	Anonymisation – no complete security	Anonymisation – loss of information	Costs
No sharing culture	Big data	Balancing privacy and openness	Competition	Missing infrastructure
Non-visible value	Complexity of the publication process	Consequences	Fear of misinterpretation	Time and/or work effort
Not established in community	Unclear which infrastructure	“I’m not allowed to publish”	Quality of the data	
	Unfamiliarity with the publication process	Legal regulations	Vulnerability	
		Licenses		
		Personal / sensitive data		
		Uncertainty what is allowed		
		Uncertainty who owns the data		
		Unclear responsibility		

Opportunities for Adopting Open Research Data in Learning Analytics

Figure 4. The occurrence of emerged codes for the barriers to the publication of research data in Learning Analytics in Germany, Peru, India and China

Codesystem	LA Germany	LA Peru	LA India	LA China	SUMME
Barriers to and problems with publication of research data					0
patents					0
embargo					0
big data				1	1
unfamiliarity with the publication process				1	1
unfamiliarity with the legal regulations				0	0
legal regulations			1		1
quality of the data			1		1
loss of control					0
licenses			1	1	2
stealing ideas					0
mistrust					0
missing peer-review					0
fear of misinterpretation		1			1
analysis is easier to understand than the raw data					0
no extrinsic motivation or obligation		1			1
no sharing-culture			1		1
costs	1				1
not established in community				1	1
vulnerability		1			1
uncertainty who owns the data				1	1
uncertainty what is allowed	1		1	1	3
"I'm not allowed to publish"	1			1	2
balancing privacy and openness	1			1	2
consequences	1				1
unclear responsibility	1				1
personal / sensitive data	1				1
non-visible value	1				1
complexity of the publication process	1				1
missing infrastructure		1			1
unclear which infrastructure	1			1	2
anonymisation					0
no complete security	1			1	2
conducting the anonymisation process	1			1	2
loss of information	1			1	2
time and/or work effort	1		1	1	3
competition	1	1			2
Σ SUMME	15	6	6	13	40

The “*uncertainty what is allowed*” is followed by the two codes that have to do with the anonymization process: “*no complete security*” and the “*loss of information*”. While the first one underpins the unfamiliarity with the legal regulations and the uncertainty with all that is associates with it, it also shows the concerns about the potential harm that might arise out of the identification of the participants. The interviewees expressed their concerns that the publication of data could compromise the participants’ confidentiality as the risk could not always be fully mitigated by the de-identification process of individual data:

What is behind it is of course, that anonymized data will never provide full security. There are enough examples where anonymous data sets has been combined with

Opportunities for Adopting Open Research Data in Learning Analytics

others and in the end you could filter out individual persons through certain features. (Senior Scientist, Germany (Biernacka, 2020c, p. 2 in os_002))

or

I can not take the risk of explosion [sic exposure]... some, eh.. some data of others. (Senior Scientist, China (Biernacka, 2020b, p. 3 in os_029))

The “*loss of information*” through conducting the anonymization process is relevant in Learning Analytics indeed too. As the discipline lives from analyzing personal and behavior data, it is exactly what is needed for the evaluation or for the training of the e.g. machine learning. If these data are anonymized too early it can have huge influence on the results of the research project:

(...) the question about anonymisation has to be looked at critically. Because... at the beginning of the research you don't really know, what are the important factors. (Junior Scientist, Germany (Biernacka, 2020c, p. 2 in os_013))

On the other hand, publishing anonymized data in a discipline that works on the personalization of learning arises doubts too:

I would at least secure that some kind of information can be gained from the data. If that is not the case, you have to ask yourself why you even work on it. (Senior Scientist, Germany (Biernacka, 2020c, p. 3 in os_024))

Research data in Learning Analytics is based on collecting information about the learner, his/her learning behavior and achievements. Since it's the ground for the personalization of the learning and teaching experience, these data are particularly subject to data protection laws and regulations. According to Pardo and Siemens (2014), a clear definition of *privacy* in LA is elusive and has to be addressed from different angles. Issues like confidentiality, trust or data ownership have to be dealt with (Drachsler & Greller, 2016; Pardo & Siemens, 2014). While in Germany and China the legal concerns form more than 50% of the named factors that prevent publishing of research data (64% and 55% respectively), it's only 33% in India and not a single one in Peru. In the latter case the publication process for research data is “*not established in the community*” yet. At this point one should bear in mind that this field of research in Latin America is still fundamentally new and all processes here are still being redefined (Biernacka & Huaroto, 2020). The researchers express their concerns:

They think that the data that will be shared will be measured or will have other results and will contradict the work that they are doing. (Senior Scientist, Peru (Biernacka, 2020a, pp. 3 in os_018, translated))

The re-analysis of their data can lead to different conclusions or in some cases, even identify mistakes in the raw research data or the original data analysis. Such a situation can possibly cause reputational damage to the researcher or their whole institution.

In India the researchers are more concerned about the “*quality of data*”. The published research data should be of good quality and therefore curated and validated:

I'm expecting to take that to a certain quality, a certain format before I publish. (Senior Scientist, India (Biernacka, 2020d, p. 4 in os_032))

Another big hurdle to overcome, is the “*time and/or work effort*” that has to be undertaken. Making research data accessible costs time and human resources for the preparation and publication of the data. Researchers prefer to invest this time in the research itself rather than in the management of the data:

And, so we think, a lot of work needs to be done before this kind of data can be published. (Senior Scientist, China (Biernacka, 2020b, p. 5 in os_007))

Moreover, funds are rarely made available for this purpose. In low-income countries this barrier is emphasized even more when it comes to costs for storage and archiving.

Five Dimensions of Barriers

In the previous sections the authors outlined on the one hand the benefits of the publication of research data, and on the other hand the factors that prevent scientist from publishing research data according to a semi-structured interview study with scientist in Learning Analytics from Germany, Peru, India and China. The potential advantages do not seem sufficient so far to motivate the researchers, in particular from low-income countries, to make their data publicly available; even though many of the interviewees stated that they had an interest in Open Data and Open Science in general.

In this section, the authors will give a wider explanation of the five dimensions of barriers to the publication of research data.

Authority or Practice Considerations

The main findings of this study show, that publication of research data among Learning Analytics researchers is not a common practice yet. Even though it is considered as desirable, the time and work effort needed to prepare good-quality data is too high. The processes of scientific article publication, which have been imprinted for decades as the highest mark of recognition in other, older disciplines, also leave their mark in this young research domain. Furthermore, the publication of Learning Analytics data requires in most of the cases an anonymization process that leads to loss of the most important information in the dataset.

There are different steps that have to be done to overcome these barriers. It is not easy to change thinking patterns and the scientific publishing world is not making it easier. The system has to be adopted. Increasing the value of publication of the research data or its citation would be of great help. It should not only be the publication of scientific articles that contributes to the reputation of the researcher, but also providing high-quality research data. Mandatory or rewarded data publishing, enforced by institutions, journals or funders would be highly efficient in increasing the motivation for the publication of research data.

Technical or Processing Constraints

Many of the barriers mentioned show clear problems with the technical system or related processes. Digitization brings with it a large flood of data. This data is on the one hand very complex and on the other hand very extensive and therefore requires a lot of storage space. Transferring many terabytes from the local laboratory server to a repository and at the same time making them available in a form that potential re-users can work with this research data is beyond the means of many researchers.

This brings up the question of which system or which repository is suitable for this. Many researchers would not even know where to start looking for a suitable repository, and what “suitable” means in this context.

Furthermore, researchers are often not sufficiently trained to publish research data in a way that complies with scientific publication standards. There are uncertainties in the processes involved, from the correct administration to anonymization and the publication of research data.

Legal Concerns

When publishing research data, many scientists face a number of legal challenges or uncertainties. Whether it is a matter of researchers working together in collaborations and not knowing what they are allowed to do by contract, or whether the question

of who owns the data has not yet been clarified. The question of legal liability also often arises: what happens if data is published illegally? Is it the scientist who is accountable, his or her supervisor, or the institution? The consequences of data publication can be difficult to assess and it is difficult to decide which license best protects the interests of the study participants and the scientist while at the same time acting in the spirit of Open Science.

Rarely is the publication of data planned from the outset and therefore not included in the informed consent. This leads to problems at a later stage, as it is often not possible or too costly to obtain a publication permit at a later stage. According to the GDPR, the processing of personal data is only possible if it follows the six principles: lawfulness, fairness and transparency, purpose limitation, data minimization, accuracy, storage limitation and integrity and confidentiality. This means that the participants need to know what will happen with their data, the minimum of data needed is collected and that this data stays confidential. Person-related information shall be eliminated before the data can be published. Some can argue though, that the process of anonymization is not sufficiently secure to guarantee protection of the research subjects as it is not possible to know what other data was already published or leaked about the participants.

Loss of Control of Data

A major problem in the publication of research data is competition and the fear of misuse of the data. Researchers do not want to lose control over their data and want to know who is using it and for what purpose. They are afraid that the data will be used unintentionally (e.g. by one of the well-known data power-houses for commercial purposes). Others consider their data to be of insufficient relevance or quality. The last point is difficult to determine because there is no established peer-review process for research data. The curation of data always has to take place on two levels (technical and content-related) (Callaghan et al., 2012) and thus costs a lot of effort. Furthermore, the publication of research data carries the risk that weaknesses in data collection and analysis may become visible and errors being exposed.

To guarantee a high security of anonymity, it is necessary to eliminate a large amount of information from the data. This does not happen without losing value of data, and then the question arises as to why one wants to do the effort at all.

Resource Constraints

The barriers that arise regarding resources refer mainly to four types of resources: human, time, financial and infrastructural.

Opportunities for Adopting Open Research Data in Learning Analytics

The time and human resources required for the preparation of research data are often ignored in project planning, which in turn is a major problem later on. Additional data managers cannot be paid for, but the scientists' time is too valuable at that moment to put it into administration. Researchers prefer to invest time in the research itself rather than in the management of their research data for later publication.

Often the supporting infrastructure is also missing at the institutions. There are no points of contact for support during the various stages of the research process and Data Protection Officers are often left to manage the high number of requests on their own (as shown in Ostendorff and Linke (2019) too). This problem is even more visible in the low-income countries and thus worse possibilities to guarantee additional personnel or technical systems.

RECOMMENDATIONS FOR THE PUBLICATION OF RESEARCH DATA IN LEARNING ANALYTICS

Solutions can be found for all five dimensions of barriers introduced in the section before. In this section the authors will give recommendations for tools and further reading tips for those researchers in Learning Analytics that want to publish his or her research data but faced the barriers mentioned before. During the semi-structured interviews, the participants suggested solutions on how to address these barriers and concerns which will be included here too.

The recommendations consist of two parts: a toolkit and guidelines.

The toolkit (see p. 14) is a collection of suggested and exemplary tools and services, as well as further reading suggestions. The resources are available (mostly) for free online and shall help the Learning Analytics researcher to overcome the barriers to the publication of research data. The proposed further sources for reading can be websites or scientific articles where the researchers can go into the deep of the topic.

The guidelines (see p. 14) can be regarded as an extension of the DELICATE checklist (Drachsler & Greller, 2016) and thus shows step by step what the researcher can and should do before publishing his or her research data.

FUTURE RESEARCH DIRECTIONS

The HEADT Centre research project consists of three phases: qualitative research, quantitative research, and findings implementation. Only the results of the first phase are presented in this chapter.

Table 4. Toolkit

	Tools & Services	Further reading
Authority or practice considerations	https://www.imsglobal.org/learningdesign/index.html - IMS Learning Design	Lee, B. D. (2018). Ten simple rules for documenting scientific software. <i>PLoS Comput Biol</i> , 14(12), e1006561. doi:10.1371/journal.pcbi.1006561
	http://dublincore.org/documents/dces/ - Dublin Core Metadata Element Set	Biernacka, K., Bierwirth, M., Buchholz, P., Dolzycka, D., Helbig, K., Neumann, J., Odebrecht, C., Wijjes, C., & Wuttke, U./Irike. (2020). Train-the-Trainer Concept on Research Data Management (Version 3.0). Zenodo. doi:10.5281/zenodo.4071471
	http://hisc.ieee.org/wp12/files/IOM_1484_12_1_v1_Final_Draft.pdf - Draft Standard for Learning Object Metadata	Tennant, J. (2020). The [R]evolution of Open Science.
Technical or processing constraints	https://rdmpromotion.rbind.io/ - Research data management Promotion material	
	https://re3data.org - Registry of Research Data Repositories	
	https://psldataa.shop.web.cmu.edu/ - DataShop – a data analysis service for the learning science community	
	https://zenodo.org - Multidisciplinary repository	
	https://choosealicense.com/ - Choose an open source license	https://www.ukdataservice.ac.uk/manage-data/legal-ethical/gdpr-in-research/consent.aspx - Applying GDPR in research
Legal concerns	https://creativecommons.org/share-your-work/ - Creative Commons tools to help share your work	Meyer, M. N. (2018). Practical Tips for Ethical Data Sharing. Association for Psychological Science, 1(1), 131-144.
	https://aircloak.com/top-5-free-data-anonymization-tools/ - Top 5 Free Data Anonymization Tools	Guibault, L., & Wiebe, A. (2013). Safe to be open. Study on the protection of research data and recommendations for access and usage.
		Drachler, H., & Greller, W. (2016). Privacy and analytics. Paper presented at the Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16.
Loss of control of data	https://osf.io/ - Platform to support your research and enable collaboration	Wallis, J. C., Borgman, C. L., & Mayernik, M. (2007). Know Thy Sensor: Trust, Data Quality, and Data Integrity in Scientific Digital Libraries. UCLA Papers. doi:10.1007/978-3-540-74851-9_32
	https://originstamp.com - Create secure timestamps using blockchain technology	
	https://citation.crosscite.org - Citation Formatter	
Resource constraints	https://ukdataservice.ac.uk/media/622368/costingtool.pdf - Data management costing tool	
	https://v2.sherpa.ac.uk/juliet - SHERPA/JULIET – Research Funders' Open Access Policies	
	https://dmponline.dcc.ac.uk/ - Helps to create, review and share data management plans	

Table 5. Guidelines for publication of research data in Learning Analytics

<p>1. Create a consent form before collecting data</p>	<p>Try to think of everything that you want to do with the collected data. Don't forget to mention the evaluation, archiving, and publication. Your participants need to be informed about all the steps you want to undertake. Here you can already put the information about the repository and the licenses you want to give your data. If you need to anonymize data, mention it as well and explain to your participants at which point of your research you want to do that, what happens to the raw data, and how will you provide data protection before you anonymize the research data. Let all your participants sign a consent before the study starts.</p>	<ul style="list-style-type: none"> • Why do you want to apply Learning Analytics? • What is the added value of your research? • What will happen to the collected data? How will you process the research data? • Do you need to collect personal or even sensitive data? • Why are you allowed to collect the data? • Where will you store your data? • How will you protect your research data? • Will you share your data and with whom? • Will you publish the research data? Where? Under what license? • Where will the data and its documentation be stored after the end of the project? • How long should the data be kept?
<p>2. Document all your steps</p>	<p>Most of the documentation is simply good research practice, so you are probably doing it anyway, just write it down step by step. Start with it as early as possible and document consistently throughout the project. Try not to leave the documentation at the very end of your research project. Write it down as long as you have it fresh in your mind and try to think about what information is needed to understand your data. You can write it down in a separate document called README, in a data dictionary or a codebook (or combining all three forms if needed).</p>	<ul style="list-style-type: none"> • What information would you need to understand the data? • For what purpose was the research data created/collected? • What does the dataset contain? • How was the research data collected? • Who collected the data? • When was the research data collected? • What data cleansing processes were undertaken? • How was the quality of the research data ensured? • In which formats is the data available? • How can the data be accessed?
<p>3. Think about where to publish your data as early as possible</p>	<p>Data Journals and repositories can follow specific metadata standards, controlled vocabulary or have other requirements to your data (e.g. formats). The most popular discipline-specific repository for LA is probably DataShop. You can decide for a generic repository like Zenodo or an institutional repository too. The choice depends on your needs, but all of those are a good choice. The important thing is that the repository provides a DOI (or another persistent identifier) to your data.</p>	<ul style="list-style-type: none"> • Where do you want to publish? • Are there specific recommendations or requirements coming from the repository or Data Journal? • Is the repository trustworthy? Does it have a seal? • Does the repository provide a DOI (or another persistent identifier)? • How durable is the service provider? • What terms of use of the data are possible? • What access is there to the data?
<p>4. Define different levels of processing¹ and de-personalize data</p>	<p>Before you can publish your data, you have to get sure that your data is de-personalized. Delete all person-related information. If this is not possible for you, you should consider publishing the metadata including the documentation instead of the dataset itself. You can define different levels of processing, e.g.</p> <p>0 Raw data: full data 1 Pseudonymized data (full data with redaction for direct and indirect identifiers) 2 Anonymized data (de-identification via direct or indirect identifiers not possible)</p>	<ul style="list-style-type: none"> • Is it possible to anonymize your research data? • When do you pseudonymize your data? • Are there technical procedures to guarantee privacy? • Does the data storage or externals fulfill highest international security standards?

continues on following page

Table 5. Continued

<p>5. Choose a license¹⁴</p>	<p>There are different licenses you can choose from for your research data depending on the format of the data. Try to choose a free and open license like Creative Commons, Open Data Commons or MIT.</p>	<ul style="list-style-type: none"> • Do you have software as your research data? • Do you want attribution for your work? • Do you want to allow others to use your work commercially? • Do you want to allow others to change, remix, adapt or build upon your work?
<p>6. Choose the right format</p>	<p>Choose open formats and try to avoid proprietary formats to support re-use. Not everybody can afford to buy the use licenses. Use a format that is readable by machines and humans. Learning Analytics comes with a wide variety of formats, so it's difficult to make clear recommendation. Compare it with the actual recommendation for long time archiving.</p>	<ul style="list-style-type: none"> • Is the format you're using proprietary? • Is the software you used widely distributed? • Is the format well established in the community? • Is it a compressed file format? • Would it be better to make the research data available in the source format and additionally in a widely used export format (e.g. PDF/A)?
<p>7. Consider all the related legal aspects¹⁵</p>	<p>Get sure that all the legal aspects are clarified before you publish.</p>	<ul style="list-style-type: none"> • Which legal provisions exist in general? • Are there any patents pending? • Are you allowed to publish the data? • Who owns the data? • Does the research data fall under the Copyright Act? • Are there any agreements on the intellectual property of the research data? • Are there any predetermined requirements of the funding agencies?
<p>8. Define the access levels¹⁶</p>	<p>Before you publish your data, you should think about different access levels to control the re-use, e.g you can define different access levels: Open – Data is freely available for re-use Restricted – Data is available for re-use with access restriction Controlled – Data can be shared after approved by the researcher Closed – Data cannot be shared. Data can only be used by the researcher or for archival purpose</p>	<ul style="list-style-type: none"> • What data can be published openly (Open Data without any restriction)? • What data should have restricted access (data available when a user meets standard criteria)? • Were any agreements made regarding data accessibility? • What data should have controlled access (data available only when e.g. a user is approved by the original researcher)?
<p>9. FAIR Data Principles¹⁷</p>	<p>Check your data complies with the FAIR Data Principles. If you cannot assure the privacy protection of your participants consider preparing your metadata for publication. You don't have to make your data open. Maybe conditional access is the better choice: the metadata record is available to the public but access to the research data themselves occurs only after pre-determined conditions are met.</p>	<ul style="list-style-type: none"> • Is your data described by rich metadata? • Can you put your metadata online in a searchable repository or catalogue? • Does the metadata record specify the persistent identifier? • Can the metadata be accessible, even if the data aren't? • Does the metadata provided follow relevant standards? • Did you use controlled vocabularies, keywords, thesauri or ontologies? • Are qualified references and links provided to other related data?
<p>10. Publish</p>	<p>You are ready: Take the step, and publish your research data (or metadata if nothing else is possible)!</p>	

Opportunities for Adopting Open Research Data in Learning Analytics

In the second step, the hypotheses that emerged from the semi-structured interviews will be revised through a wide-spread online-survey. Thus, by incorporating a quantitative analysis, the authors wish to resolve some of the limitations of the qualitative phase of the study.

In order to better understand the influence factors on the publication of research data in general, two more disciplines should be considered: medicine and climate impact research. These disciplines show a wide variation in the research data types, particularly in terms of the data sensitivity.

In the final phase of research, in addition to the guidelines and recommendations, technical implementations for repositories will be proposed and best practices for researchers will be developed.

CONCLUSION

Learning Analytics present significant opportunities for a change of teaching and learning experiences. It is particularly useful because it incorporates computational analysis techniques to the already established research on evidence and improvement of teaching and learning. It is also based on algorithms and methods that require and produce a lot of data. According to Drachsler and Greller (2016) researchers and institutions dealing with Learning Analytics are seeing privacy as a big concern. The authors emphasize that most of the people are not aware of the legal boundaries. The semi-structured interview study of the HEADT Centre underpins this observation and focuses on the publication of data in LA that would be so important in this area. It can be extrapolated from the research trends in other disciplines, that the scientists in Learning Analytics put their focus on the publication of scientific articles, including the results of their research, rather than publishing the underlying research data.

In 2014 Scheffel, Drachsler, Stoyanov, and Specht (2014) it is already shown that two of the most important topics in Learning Analytics are: the openness and transparency of the used data, and the data privacy . But still, it can be said that the process of publishing research data in Learning Analytics has not yet been fully established. A complete openness of data also seems quite unlikely in this case due to the processing of personal data. Although the participants in the interviews and surveys from related research come from different countries and are therefore subject to different data protection regulations, they agreed that “*uncertainty about what is allowed*” or legal issues in general (data privacy in particular) is the biggest factor preventing them from publishing their research data. Ignoring these fears can lead to a lack of acceptance from the research participants and to the hesitation of publishing research data from the researchers.

In all phases of research data management, the most diverse areas of law must be considered. This fact alone overwhelms many researchers even before they start preparing research data for publication. Rights of use, science law, fundamental rights, international law, patent law, competition law, copyright law, contracts, policies, labor law and above all - concerning almost every phase of the research data life cycle – the data protection law.

Learning Analytics as a subject has a difficult starting position, because research here is based on individual data to enable personalized teaching and learning in order to achieve even better learning results. Basically, beginning with the planning of a research project and through to the collection of the research data, it must be considered whether these data have a personal reference and whether this personal information is important for the research to be conducted. If this is the case, the informed consent of the research participants is essential. The data should be made anonymous as soon as the research purpose allows it. If it is not possible from the beginning, other protective measures must be conducted (e.g. pseudonymization). Anonymization should only be postponed in research projects if those features that can be used to identify a person are really needed to achieve the research purpose or individual research steps. Anonymization can be seen as an enabler for the publication of data and it reduces the fear of privacy breaches too. However, caution must be paid: in many cases of automated anonymization it is at best a pseudonymization. In this case, the data, in combination with other data sources, can lead to the de-anonymization or identification of the persons (Drachsler & Greller, 2016).

The publication of research data is still a tough issue in some areas. This is also true for Learning Analytics, the value of such data publication is not yet apparent to researchers. The frequent barriers associated with the many legal aspects create uncertainty. With this chapter, the authors launch a call to break through these fears and show the benefits of publishing and citing data. Other ways are also pointed out in the very difficult cases where complete opening of research data is not possible. In many cases the data is not validated or not all information can be shared, but perhaps new collaborations or meta-analyses can emerge from FAIR metadata alone.

The road to truly open and FAIR published data is still long and certainly challenging. The basic data protection regulation rightly protects the participants in the research, but at the same time it spreads a large degree of uncertainty among scientists when publishing research data.

REFERENCES

- Alexander, S. M., Jones, K., Bennett, N. J., Budden, A., Cox, M., Crosas, M., Game, E. T., Geary, J., Hardy, R. D., Johnson, J. T., Karcher, S., Motzer, N., Pittman, J., Randell, H., Silva, J. A., da Silva, P. P., Strasser, C., Strawhacker, C., & Stuhl, A. (2019). Qualitative data sharing and synthesis for sustainability science. *Nature Sustainability*. Advance online publication. doi:10.1038/41893-019-0434-8
- ALLEA. (2017). *The European Code of Conduct for Research Integrity*. ALLEA - All European Academies.
- Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H., & Ioannidis, J. P. (2011). Public availability of published research data in high-impact journals. *PLoS One*, 6(9), e24357. doi:10.1371/journal.pone.0024357 PMID:21915316
- Biernacka, K. (2019). *Research Integrity and Privacy*. Retrieved from <https://headt.eu/Research-Integrity-Technology-and-GDPR>
- Biernacka, K. (2020a). *Perspectiva de los Investigadores sobre la Publicación de Datos de Investigación: Entrevistas Semiestructuradas de Perú*. edoc-Server, Humboldt-Universität zu Berlin. Berlin, Germany. doi:10.18452/21394
- Biernacka, K. (2020b). *Researchers' Perspective on the Publication of Research Data: Semi-structured Interviews from China*. edoc-Server, Humboldt-Universität zu Berlin. Berlin, Germany. doi:10.18452/21330
- Biernacka, K. (2020c). *Researchers' Perspective on the Publication of Research Data: Semi-structured Interviews from Germany*. edoc-Server, Humboldt-Universität zu Berlin. Berlin, Germany. doi:10.18452/21644
- Biernacka, K. (2020d). *Researchers' Perspective on the Publication of Research Data: Semi-structured Interviews from India*. edoc-Server, Humboldt-Universität zu Berlin. Berlin, Germany. doi:10.18452/21378
- Biernacka, K., & Huaroto, L. (2020). *Learning Analytics in Relation to Open Access to Research Data in Peru. An Interdisciplinary Comparison*. Paper presented at the LALA 2020, Cuenca, Ecuador.
- Biernacka, K., & Pinkwart, N. (2020). *Barriers and Hurdles to the Publication of Learning Analytics Data*. Paper presented at the 10th International Learning Analytics and Knowledge (LAK), Frankfurt (Oder), Germany.

Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., Ault, L., Bell, P., Bowie, R., Leadbetter, A., Lowry, R., Moncoiffé, G., Harrison, K., Smith-Haddon, B., Weatherby, A., & Wright, D. (2012). Making Data a First Class Scientific Output: Data Citation and Publication by NERC's Environmental Data Centres. *International Journal of Digital Curation*, 7(1), 107–113. doi:10.2218/ijdc.v7i1.218

Catlin-Groves, C. L. (2012). The Citizen Science Landscape: From Volunteers to Citizen Sensors and Beyond. *International Journal of Zoology*, 2012, 1–14. doi:10.1155/2012/349630

Chan, L., Cuplinskasm, D., Eisen, M., Friend, F., Genova, Y., Guédon, J.-C., Hagemann, M., Harnad, S., Johnson, R., Kupryte, R., La Manna, M., Rév, I., Segbert, M., de Souza, S., Suber, P., & Velterop, J. (2002). *Budapest Open Access Initiative*. Retrieved from <https://www.budapestopenaccessinitiative.org/read>

Cheah, P. Y., Tangseefa, D., Somsaman, A., Chunsuttiwat, T., Nosten, F., Day, N. P., Bull, S., & Parker, M. (2015). Perceived Benefits, Harms, and Views About How to Share Data Responsibly: A Qualitative Study of Experiences With and Attitudes Toward Data Sharing Among Research Staff and Community Representatives in Thailand. *Journal of Empirical Research on Human Research Ethics; JERHRE*, 10(3), 278–289. doi:10.1177/1556264615592388 PMID:26297749

Cobo, C., & Aguerrebere, C. (2018). Building Capacity for Learning Analytics in Latin America. In C. Ping Lim & V. L. Tinio (Eds.), *Learning Analytics for the Global South* (pp. 58–67). Foundation for Information Technology Education and Development.

Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2019). *The Citation Advantage of Linking Publications to Research Data*. <https://arxiv.org/abs/1907.02565>

DCC (Digital Curation Centre). (n.d.). *Overview of funders' data policies*. Retrieved from <https://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies>

Deutsche Forschungsgemeinschaft. (2019). Guidelines for Safeguarding Good Research Practice. Code of Conduct. In (pp. 29). doi:10.5281/zenodo.3923602

Drachsler, H., & Greller, W. (2016). Privacy and analytics. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16*. 10.1145/2883851.2883893

Opportunities for Adopting Open Research Data in Learning Analytics

Ebner, M., & Maurer, H. (2008). *Can Microblogs and Weblogs change traditional scientific writing?* Paper presented at the E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2008, Las Vegas, NV.

European Commission. (2014). *Background Document. Public Consultation 'Science 2.0': Science in Transition*. Author.

European Commission. (2016). *H2020 Programme. Guidelines on FAIR Data Management in Horizon 2020*. European Commission.

Fecher, B., & Friesike, S. (2014). Open Science: One Term, Five Schools of Thought. In S. Bartling & S. Friesike (Eds.), *Opening Science. The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing*. Springer Open. doi:10.1007/978-3-319-00026-8_2

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5/6), 304. Advance online publication. doi:10.1504/IJTEL.2012.051816

Gentil-Beccot, A., Mele, S., & Brooks, T. C. (2009). *Citing and Reading Behaviours in High-Energy Physics. How a Community Stopped Worrying about Journals and Learned to Love Repositories*. <https://arxiv.org/abs/0906.5418>

Gezelter, D. (2011). *An informal definition of Open Science*. Retrieved from <http://openscience.org/an-informal-definition-of-openscience>

Gleditsch, N. P., Metelits, C., & Strand, H. v. (2003). Symposium on Replication in International Studies Research. *International Studies Perspectives*, 4(1), 89–97.

Hartmann T. (2019). *Rechtsfragen. Institutioneller Rahmen und Handlungsoptionen für universitäres FDM*. Frankfurt (Oder): Europa-Universität Viadrina Frankfurt (Oder). doi:10.5281/zenodo.2654306

House of Commons Science and Technology Committee. (2011). *Science and Technology Committee – Eighth Report. Peer review in scientific publications*. Retrieved from <https://www.publications.parliament.uk/pa/cm201012/cmselect/cmsctech/856/85602.htm>

Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V. M., Nichols, T. E., & Wagenmakers, E.-J. (2018). Data Sharing in Psychology: A Survey on Barriers and Preconditions. *APS*, 1(1), 70-85. doi:10.1177/2515245917751886

Ince, D. (2011). The Duke University scandal - what can be done? *Significance*, 3. doi:10.1111/j.1740-9713.2011.00505.x

Irwin, A. (1995). *Citizen Science: A Study of People, Expertise and Development (Environment and Society)*. Routledge.

Jones, L., Grant, R., & Hrynaszkiewicz, I. (2019). Implementing publisher policies that inform, support and encourage authors to share data: two case studies. *Insights the UKSG Journal*, 32, 11. doi:10.1629/uksg.463

Jones, S., & Grootveld, M. (2017). *How FAIR are your data?* (2nd ed.), doi:10.5281/zenodo.1065990

Kim, J. (2019). Overview of disciplinary data sharing practices and promotion of open data in science. *Science Editing*, 6(1), 3–9. doi:10.6087/kcse.149

Kratz, J., & Strasser, C. (2014). Data publication consensus and controversies. *F1000 Research*, 3, 94. doi:10.12688/f1000research.3979.3 PMID:25075301

Larivière, V., Sugimoto, C. R., Macaluso, B., Milojević, S. a., Cronin, B., & Thelwall, M. (2013). *arXiv e-prints and the journal of record: An analysis of roles and relationships*. <https://arxiv.org/abs/1306.3261>

Long, P., & Siemens, G. (2011). Penetrating the Fog: Analytics in Learning and Education. *EDUCAUSE Review*.

LucraftM.AllinK.BaynesG.SakellaropoulouR. (2019). Challenges and Opportunities for Data Sharing in China. In (Journal contribution ed.): figshare. doi:10.6084/m9.figshare.7718441.v1

Meyer, M. N. (2018). Practical Tips for Ethical Data Sharing. *Association for Psychological Science*, 1(1), 131–144.

Nielsen, M. (2012). *Reinventing Discovery. The New Era of Networked Science*. Princeton University Press.

Open Knowledge Foundation. (2015). *Open Definition 2.1*. Retrieved from <http://opendefinition.org/>

Open Knowledge Foundation (Producer). (2019). *Open Data Handbook*. Retrieved from <http://opendatahandbook.org>

Open Science and Research Initiative. (2014). *The Open Science and Research Handbook*. Retrieved from <https://www.fosteropenscience.eu/content/open-science-and-research-handbook>

Ostendorff, P., & Linke, D. (2019). Best-Practices im Umgang mit rechtlichen Fragestellungen zum Forschungsdatenmanagement (FDM). *Bibliotheksdienst*, 53(10-11), 717–723. doi:10.1515/bd-2019-0098

Opportunities for Adopting Open Research Data in Learning Analytics

- Papamitsiou, Z., & Economides, A. A. (2014). Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence. *Journal of Educational Technology & Society*, 17(4), 49–64.
- Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, 45(3), 438–450. doi:10.1111/bjet.12152
- Peng, R. D. (2011). Reproducible Research in Computational Science. *Science*, 334(6060), 2. doi:10.1126/science.1213847 PMID:22144613
- Pienta, A. M., Alter, G., & Lyle, J. (2010). The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data. In Inter-university Consortium for Political and Social Research. Institute for Social Research, University of Michigan. <http://hdl.handle.net/2027.42/78307>
- Piwowar, H. A. (2011). Who shares? Who doesn't? Factors associated with openly archiving raw research data. *PLoS One*, 6(7), e18657. doi:10.1371/journal.pone.0018657 PMID:21765886
- Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, (1), 25. doi:10.7717/peerj.175
- Pontika, N., Knoth, P., Cancellieri, M., & Pearce, S. (2015). Fostering open science to research using a taxonomy and an eLearning portal. *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business - i-KNOW '15*. 10.1145/2809563.2809571
- Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). *Altmetrics: A manifesto*. Retrieved from <http://altmetrics.org/manifesto>
- Rentier, B. (2016). Open science: A revolution in sight? *Interlending & Document Supply*, 44(4), 155–160. doi:10.1108/ILDS-06-2016-0020
- Scheffel, M., Drachsler, H., Stoyanov, S., & Specht, M. (2014). Quality Indicators for Learning Analytics. *International Forum of Educational Technology & Society*, 17(4), 117-132. Retrieved from <https://www.jstor.org/stable/10.2307/jeductechsoci.17.4.117>
- Schmidt, B., Gemeinholzer, B., & Treloar, A. (2016). Open Data in Global Environmental Research: The Belmont Forum's Open Data Survey. *PLoS One*, 11(1), e0146695. doi:10.1371/journal.pone.0146695 PMID:26771577

Schofield, P. N., Bubela, T., Weaver, T., Portilla, L., Brown, S. D., Hancock, J. M., David, E., Tocchini-Valentini, G., Hrabe de Angelis, M., & Rosenthal, N. (2009). Post-publication sharing of data and tools. *Nature*, *461*(10), 171–173. doi:10.1038/461171a PMID:19741686

TED. (2020). *TED. Our Mission: Spread ideas*. Retrieved from <https://www.ted.com/about/our-organization>

Van den Eynden V. Knight G. Vlad A. Radler B. Tenopir C. Leon D. Manista F. Whitworth J. Corti L. (2016). Towards Open Research. Practices, experiences, barriers and opportunities: Welcome Trust. doi:10.6084/m9.figshare.4055448

Van Horik, R., Dillo, I., & Doorn, P. (2013). Lies, Damned Lies and Research Data: Can Data Sharing Prevent Data Fraud? *International Journal of Digital Curation*, *8*(1), 229–243. doi:10.2218/ijdc.v8i1.256

Vanpaemel, W., Vermorgen, M., Deriemaeker, L., & Storms, G. (2015). Are We Wasting a Good Crisis? The Availability of Psychological Research Data after the Storm. *Collabra*, *1*(1). Advance online publication. doi:10.1525/collabra.13

Vicente-Saez, R., & Martinez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research*, *88*, 428–436. doi:10.1016/j.jbusres.2017.12.043

Vision, T. J. (2010). Open Data and the Social Contract of Scientific Publishing. *Bioscience*, *60*(5), 330–331. doi:10.1525/bio.2010.60.5.2

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 9. doi:10.1038/data.2016.18 PMID:26978244

ADDITIONAL READING

Bezjak, S., Clyburne-Sherin, A., Conzett, P., Fernandes, P., Görögh, E., Helbig, K., Kramer, B., Labastida, I., Niemeyer, K., Psomopoulos, F., Ross-Hellauer, T., Schneider, R., Tennant, J., Verbakel, E., Brinken, H., & Heller, L. (2018). *The Open Science Training Handbook* (Version 1.0): Zenodo. doi:10.5281/zenodo.1212495

Opportunities for Adopting Open Research Data in Learning Analytics

Corti, L., Van den Eynden, V., Bishop, D. V. M., & Woollard, M. (2014). *Managing and Sharing Research Data: A Guide to Good Practice*. Sage.

European Commission. (2018). *Ethics and Data Protection*. European Commission.

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5/6), 304. Advance online publication. doi:10.1504/IJTEL.2012.051816

Jensen, U., Netscher, S., & Weller, K. (2019). *Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten. Grundlagen und praktische Lösungen für den Umgang mit quantitativen Forschungsdaten*. Verlag Barbara Budrich. doi:10.3224/84742233

Kalkman, S., Mostert, M., Udo-Beauvisage, N., van Delden, J. J., & van Thiel, G. J. (2019). Responsible data sharing in a big data-driven translational research platform: Lessons learned. *BMC Medical Informatics and Decision Making*, 19(1), 283. doi:10.1186/12911-019-1001-y PMID:31888593

Majid, S., Foo, S., & Zhang, X. (2018). Research Data Management by Academics and Researchers: Perceptions, Knowledge and Practices. In *Maturity and Innovation in Digital Libraries* (pp. 166-178).

KEY TERMS AND DEFINITIONS

Altmetrics: An alternative way to record and document the use and impact of science.

Metadata: Structured data that provides basic description of other data.

Metadata Standard: Used for the standard definition of related data in terms of content and structure.

Open Data: Data that can be freely accessed, modified, processed and re-used by everyone for any purpose.

Repository: Infrastructure and the corresponding service that enables digital resources (e.g. data, code or documents) to be permanently, efficiently and sustainably stored.

Research Data: Data that are produced during the research process. It includes all data from the planning of the process to the outcome thereof.

Research Data Management: Includes all activities related to the collection, storage, preservation and publication of research data.

Research Integrity: Research Integrity refers to a set of principles that lead to good scientific practice. These include: reliability, honesty, respect and accountability.

ENDNOTES

- 1 <https://headt.eu/>
- 2 A collaborative project among mathematicians started in 2009 on Timothy Gowers' blog.
- 3 The Budapest Open Access Initiative was formed during a meeting convened in Budapest by the Open Society Foundations (OSF) on December 1-2, 2001.
- 4 <https://arxiv.org/>
- 5 <https://peerj.com/preprints/>
- 6 Kratz and Strasser (2014) distinguish between “published” data and “Published” data. Their definition of “published” data matches the term “shared” data in this chapter. However, when talking about published data in this chapter, this refers to “Published” data in the sense of Kratz and Strasser (2014)(formal publishing).
- 7 e.g. <https://www.re3data.org/>
- 8 <https://www.doi.org/>
- 9 <https://creativecommons.org/>
- 10 Although a complete opening of the data would be desired, it is not always possible due to personal data. Therefore, in this definition of Open Research Data it is considered that for those data that cannot be de-personalized, limited access or only the publication of metadata may be required.
- 11 Sensitive data are particular personal data, which require an increased protection: racial and ethnic origin, political opinions, religious or philosophical beliefs, union membership, genetic and biometric data, health data, data on sex life or sexual orientation.
- 12 General Data Protection Regulation (GDPR) valid from May 2018 in the European Union.
- 13 Based on Alexander et al. (2019)
- 14 Questions based on the Creative Commons Chooser <https://chooser-beta.creativecommons.org/>
- 15 Compare with Hartmann (2019)
- 16 Based on Alexander et al. (2019)
- 17 Compare with S. Jones and Grootveld (2017)