# Chapter 3 – Digital Science Use Cases: Enriching context and enhancing engagement around datasets

Christian Herzog1,a, Daniel W Hook1,2,3,b, Mark Hahnel1,c, Stacy Konkiel1,d, and Duane E. Williams1,e

1Digital Science, London, N1 9XW, UK

2Department of Physics, Washington University in St Louis, St Louis, Missouri, USA

3Centre for Complexity Science, Imperial College London, London, SW7 2AZ, UK

(a<https://orcid.org/0000-0002-9983-0033>, b<https://orcid.org/0000-0001-9746-1193>, c<https://orcid.org/0000-0003-4741-0309>, d<https://orcid.org/0000-0002-0546-8257> e<https://orcid.org/0000-0002-2111-3413>)

## 3.1 Introduction

The relationship between research, researchers and data is changing. Data has always played a critical role in scientific research, however in recent years it has taken centre stage not only for the so-called hard sciences, but also for the social sciences, and it has an increasing role in the humanities (Giuliano, 2019). We assert that this change is, at least in part, driven by two key factors: First, the increasing volume of data that is available to researchers either, for example, through the increasing sensitivity of instruments that aid experimental work, or through the ubiquity of computer systems with which we interact in our daily lives. Second, our ability to process and analyse these data is growing quickly as computers become faster and algorithms become more powerful. While some researchers welcome having more data to work with, others are challenged or marginalised in this new data-rich world (Eijnatten et al., 2013; Grusin, 2014; Leurs and Shepherd, 2017). These effects are often compounded by the tools that researchers must master to connect their research to data.

In the hard sciences, the CERN Open Data Portal contains 131 datasets describing particle collisions, each of which comprise around 300Gb of data at the time of writing (CERN, 2019). In the social sciences, the CISER Data Archive at Cornell is home to more than 1000 different social science data sets (CISER, 2019). These examples are individual instances chosen at random from many that could be used to demonstrate the variety and scale of data available for research. However, even these examples don’t begin to quantify the amount of detailed personal data available to companies such as Facebook. Data of this latter kind has already been used in academic studies as well as in more controversial contexts (Jordan and Weller, 2018; Kamp et al., 2019; Stark, 2018). Clearly, there is an increasing diversity and depth of data available for research from both traditional and from new sources.

Many researchers now work with large volumes of data. Fortunately, many facilitating technologies have become commoditised and are available at a fraction of their original cost: storage is cheap and data transfer is fast. But, Increasing the value of data to researchers is no longer about technology, rather it is about the information and culture around the data.

In this chapter, we take our lead from Chapter 1 in recognising not only that science is at a crossroads but that the whole of research is changing. We discuss two elements of infrastructure that, if enhanced, can make data more useful and valuable to the whole research community: information infrastructure and cultural infrastructure. The Rich Context project supports the development of tools that enrich not only information infrastructure around datasets, but which also enhance the cultural infrastructure. *Information infrastructure* includes details of the approach to data stewardship, context of usage, code applied to the dataset in its production, as well as code applied to the data to derive further results or translate it for practical uses. All these factors add critical elements to the research infrastructure. *Cultural infrastructure* includes creating the incentives, triggers and frameworks that encourage the dataset stewards, experts and users to contribute to these critical information elements.

## 3.2 Information Infrastructure

Information infrastructure can be defined as the collection of processes and artefacts that are foundational to today’s scholarly communications. A simplified model of scholarly communications would have artefacts such as journals, journal articles, article metadata and citations. In this case, the processes would be peer review and scholarly search.

When creating one of the first scientific journals *Philosophical Transactions* 350 years ago, the members of the Royal Society did not have today’s data-centric world in mind. While a clear line can be drawn from the articles of that time to the articles of today, infrastructures have grown up around research publications in the intervening years that have moved the structures and expectations of the research article forward. These norms are powerful and persistent through their ubiquity. For example, in the large majority of modern research literature, we continue expect articles to be grouped into journals and published on a specific date, and we expect there to be a version of record that constitutes a definitive record of a piece of research.

Data is more fluid than a standard research article: it is produced and updated more frequently and iteratively; it needs to be shared with many in a collaborative context; it is processed and versioned by different colleagues. Data does not fit into the normalised research publication. Research fields that rely on data are beginning to publish data as a distinct output from a research article. Data is becoming a principal research output, while the technological challenges of publishing data are being addressed, the format and necessary fields of the metadata that describe data, the file format in which the data resides, the resource to annotate the data to make it useful to others, the way in which data should be cited in a paper or by another dataset, the description of the processing that has been applied to the data, the details of the ethical review process behind the exercise that gathered the data, and many other norms do not yet exist homogeneously across subjects and geographies. There are not yet strongly established norms that help researchers to have trust in data.

A dataset can change with time for many reasons: data may be added over time, corrections may be issued, and so on. In these cases, it may be appropriate to “version” the dataset (by issuing a persistent identifier for a point-in-time snapshot for the dataset, allowing subsequent changes to receive their own “versions”). But changes to a dataset may have a knock-on effect on the interpretation of the data and may fundamentally alter the research result that was originally reported. Moreover, in many fields “Big Data” is so central that it not only puts pressure on the community to establish an acceptable model of data publication, but also puts significant stress on how we read, interpret, and review research as a whole.

Many datasets are now so vast that we lack the ability as humans to consume them in an easy way. Visualisation technologies and other tools that allow us to interact with and sample data dynamically have received significant attention in recent years, and have helped with the interpretation of data in online environments. But it is simply impossible to reduce some types of data to a single figure or printable table, as would be the case for “traditional” journal publishing. By attempting to do so, we miss the essence of the data and risk failing to communicate data-driven conclusions accurately. This limitation of current publication formats (e.g. static PDF files for articles) is an issue that relates to the reproducibility crisis of modern research.

Peer review is another process that is not easy to apply to data as a “first class” research object. Historically, peer reviewers have ensured that a piece of research is well-communicated and correct in the sense that it is reproducible. This level of peer review is difficult to apply in the context of research data. If data is being published as a primary output, then it may be possible to perform a kind of peer review by applying some statistical tests to a sample of the data, or by using some other appropriate technique. However, it is no longer practical in most cases to set up a parallel experiment to reproduce data, as had been the case in years past. Across all contexts there are good reasons for these challenges: the experiment may be too costly to repeat, or the conditions of the original data collection may not be replicable (for example, surveying stress levels of the populace during a specific political event). In addition, ethical considerations such as the anonymity of those being surveyed may make certain types of data difficult to review. Thus, we need to develop robust and accepted approaches to peer review, not only for data itself but also for those publications that are heavily based on data. Without peer review or some suitable proxy for peer review that makes sense for data, it is difficult to know whether a dataset can be trusted. Without trust, a dataset has no value to a researcher who seeks to build upon it.

Several publishing innovations have made journal articles more discoverable and accessible in recent years, such as preprint servers, the widespread use of Digital Object Identifiers (DOIs), and centralized search engines. However, while some of these infrastructures do enhance a researcher’s ability to find research data, they do not fully translate from the realm of journals to data. There are multiple reasons for this lack of translation, some of the key features include: a) weakness of a homogeneous metadata infrastructure for datasets; b) inhomogeneity in the types of data that can be shared; c) proliferation of different platforms to store data; d) lack of standardised publication practices; e) lack of adoption of standards across fields. When compared with the “shape” of an academic article for which there is a standard structure (e.g. DOI, abstract, title, authors, keywords, etc) specifically designed to facilitate human search, it is clear that datasets are contextualised by an immature information infrastructure.

Datasets are more complex to classify and annotate than articles, yet some progress is being made. The core fields required to create a valid DataCite record are identifier, creator, title, publisher, publication year and resource type (DataCite Metadata Working Group, 2016). All other data fields are optional (e.g. location, funder, subject, contributors) due to the fundamental uncertainty in what might constitute research data in the future. This flexibility limits how data can be discovered. It has taken some years for Web of Science, Google and others to introduce functionality to search for datasets in their discovery systems.

Technological infrastructure for data--or lack thereof--has huge implications for the discovery, peer review, citation practices, interpretation, and availability of data. These challenges are interconnected with challenges we face when thinking about the cultural infrastructure for data, as well.

## 3.3 Cultural Infrastructure

There are two main aspects to cultural infrastructure: incentives and capability. Both aspects affect how researchers engage with research data, and their behaviours relating to sharing it with others and making it available to external scrutiny.

Anecdotally, academics do not typically take up research careers for financial gain. Rather, they choose to dedicate themselves to understanding a specific problem or field partially in the hope of making a discovery. For most researchers, success is not strongly coupled to prize winning, but rather by winning the freedom to determine their own research agenda. Researchers in many fields are promoted by publishing in specific high-impact journals, leading to funding success, which in turn usually leads to greater control of your research.

Sharing data is often not well-aligned with the current model of incentives. Parting with the data that underpins your research gives rise to two concerns. Firstly, that someone may find an error in your work and discredit what you have done. Secondly, that someone else may not share their own data but will gladly reuse yours if you make it available. This is especially the case in fields where success is based on having more data to analyse.

A further level of inequity exists in which data-related jobs are valued by the Academy. If a researcher happens to be particularly talented in working with data curation, data analysis or data processing, there is no track for recognising these talents. They are unlikely to be a first author on a publication in a major journal due to their data wrangling talents, and hence they have less of a chance of career progression than researchers who take a more traditional “publish or perish” path with their work as described above.

This set of perverse incentives means that people with the capability to handle data are often incentivised to leave research. Hence, not only do we have a problem of incentives in sharing and communicating data, but we also have a problem in retaining people who have the capability that we need to structure data so that it can be shared and built upon.

Capability for sharing data is the second aspect of the cultural challenge that academia continues to wrestle with. Making data available to others is generally accepted as a key part of the research communication process. However, there are certain established norms around when the data should be shared, and to what depth it is shared (Linek et al., 2017); for example, in fields where human subjects research is prevalent, there is a much more conservative attitude towards open data than in fields like astronomy where data sharing is widely practiced, given that data can be collected by only a handful of observatories and telescopes worldwide.

In fields that are more applied, ensuring that data generated as a result of a commercial relationship is protected is crucial. In such fields, academics often have a better understanding of copyright, intellectual property rights and licences (Treadway et al., 2016). But outside of this context, there is a general lack of understanding of these issues and hence data are often not shared over concerns for a perceived legal barrier.

Other concerns are ethical—for example, should these data be shared if it might infringe the rights of the subjects of the research? Researchers are beginning to become aware that, through the use of algorithms, some data is not as well anonymised as it may first appear (Siddle, 2014). Anonymisation of data is a research field in and of itself (Li et al., 2007).

The degree and nature of ethical issues and industry-proximity vary greatly between different research fields and give rise to different cultures of data usage and re-usage across fields and even within fields. Some researchers are motivated to engage with the open access community and hence choose approaches to sharing data that include granting permissive licences, association of unique identifiers with data, adherence to data standards and training students to adhere to similar approaches. Other researchers are motivated to ensure that data are not shared due to the information that can be inferred by processing the data.

The power of the newest algorithms, or of algorithms yet to come, mixed with constantly developing ethical nuance means that it is difficult to pre-empt what may or may not be acceptable to share in the future. Hence, some may feel that it is simply better not to share, especially in the social sciences, where many of these issues are more prone.

Other concerns are simply practical—how does one make data available in a way that is meaningful to others? The work associated with making a dataset generically machine-readable is challenging for many researchers, who are not to be experts in data handling. The work associated with making a dataset human-understandable, reproducible and fully contextualised is often significant. Funding constraints may make it impractical to share data or to add useful, valuable or even critical annotations to a dataset. However, funders are beginning to prioritise these activities in their grant programs (Jisc, 2019; NNLM, 2019). All these factors lead to uncertainty exacerbated by different levels of confidence and understanding and consequently an uneven landscape in what is shared, how it is shared and where it is shared.

## 3.4 Enriching context

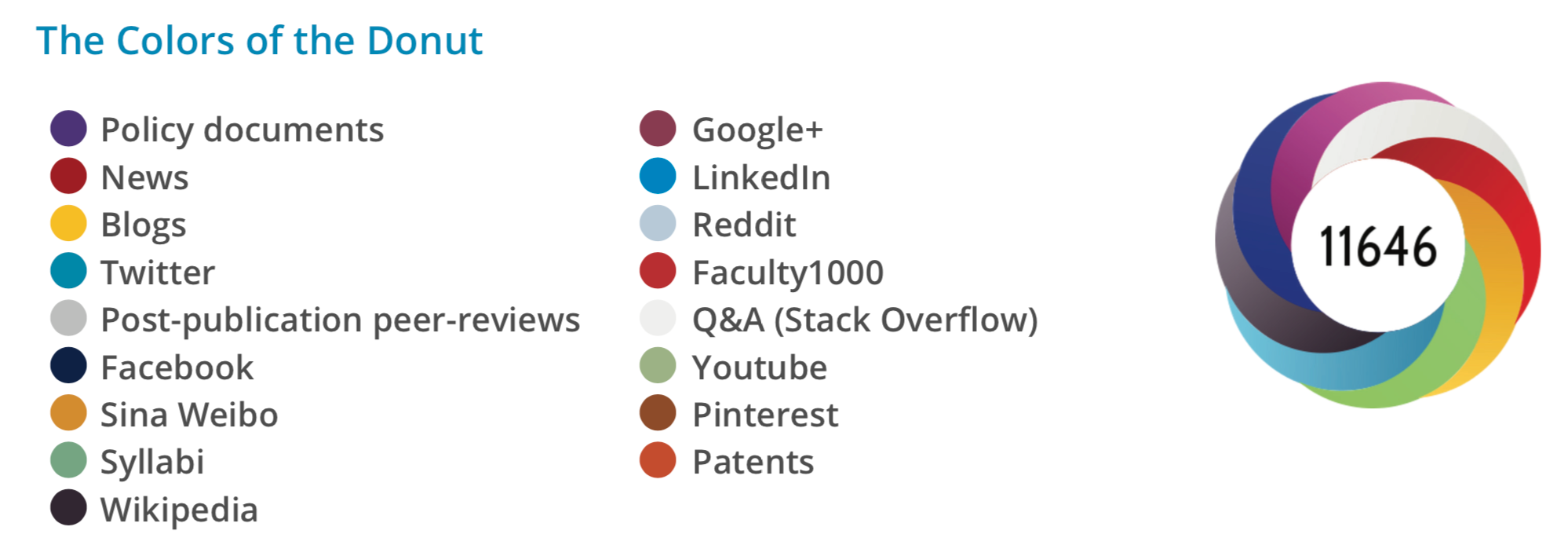
The points discussed above offer some indication of what would be needed to improve the value of research data. Firstly, to address issues of cultural infrastructure, we need to adopt an expanded version CredIT (Allen et al., 2014) that focuses on datasets. This expansion would ensure that all contributors to a dataset’s creation, development and maintenance over time are stored in a machine-readable format. Such a record is central to the facilitation of culture change across research. Only with this structure in place can the activities around datasets be readily recognised and incentives created that would support data sharing. Secondly, to address the deficits in information infrastructure, a set of tools that allow research data to be discovered and contextualised needs to be introduced. In this section, we focus on this second challenge.

The ability to add context any piece of research was a strong driver for the creation of Dimensions(Hook et al., 2018). The idea that all research happens in a particular place, at a particular time, carried out by a set of people, some of whom may be affiliated with a research institution, gives a set of metadata that allows us the “weak context” of a piece of research. By “weak context” we mean that the context being provided gives no deep understanding of the context of an article to a non-expert and is essentially indistinguishable from standard metadata. But with modern data mining approaches, it is possible to add a “strong context”.

Strong contextualisation of research should provide a user with rich information about the research including funding, other research produced as part of the larger project (e.g. related publications, clinical trials, etc), and details of the research that was built on top of it. This information should also fit into, trends and graphical representations that offer a more complete, more rapid understanding of how research fits into the larger field, related fields, or the context of the publishing journal or supporting institution. For example, for a research article, we should be able to quickly understand how many researchers are in a related field, whether the field is growing, how old the field is, how much funding has been deployed in the field, which countries have provided that funding, whether the field has begun the translation to application through patents or clinical trials, or whether it has been used as a basis for the formulation of policy.

Context can also be offered in the data that we provide to understand the reach and influence of research.

Alternative metrics (“altmetrics”) are data from the social web that run orthogonal to classic citation measures, which can be seen to add significant context to an article – extending our understanding of how different cohorts of potential users of the research are engaging with it. For example, altmetrics can be used to understand if an article is being mentioned in the news, in which geographical regions it is being noticed, whether it is being used as part of a teaching syllabus, and many other kinds of public and non-traditional scholarly engagement. These data can then be visualized in creative ways to add instant additional context to engagement with a research article (see Fig. 3.1).



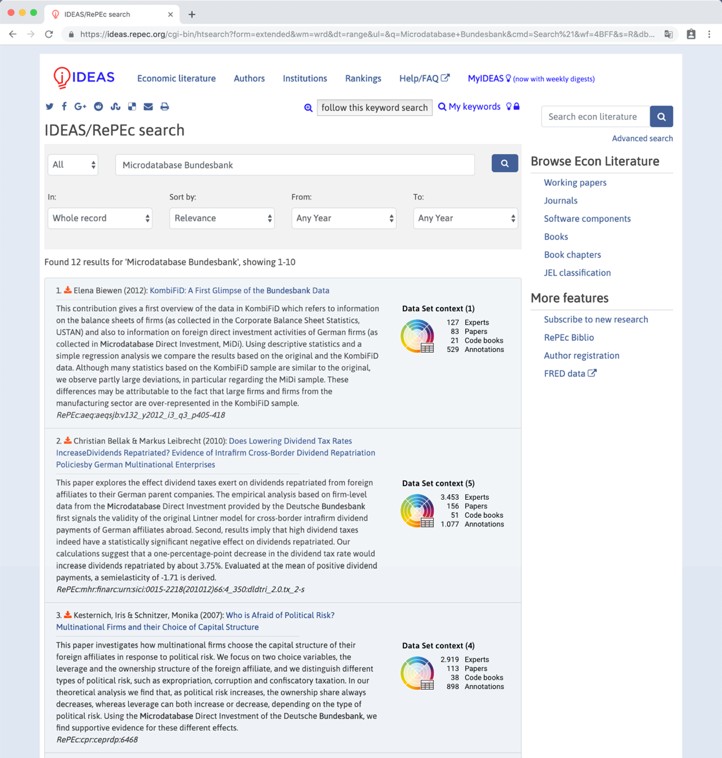
*Figure 3.1: Different types of context tracked by Altmetric.com for any research output.   
(Reproduced by permission of Altmetric.com)*

How datasets are used in research more broadly is another important piece of context that data search engines lack that would significantly enhance discoverability of a dataset and that would consequently increase the value of the data. This is where the Rich Context project can add significant value to a broad research community.

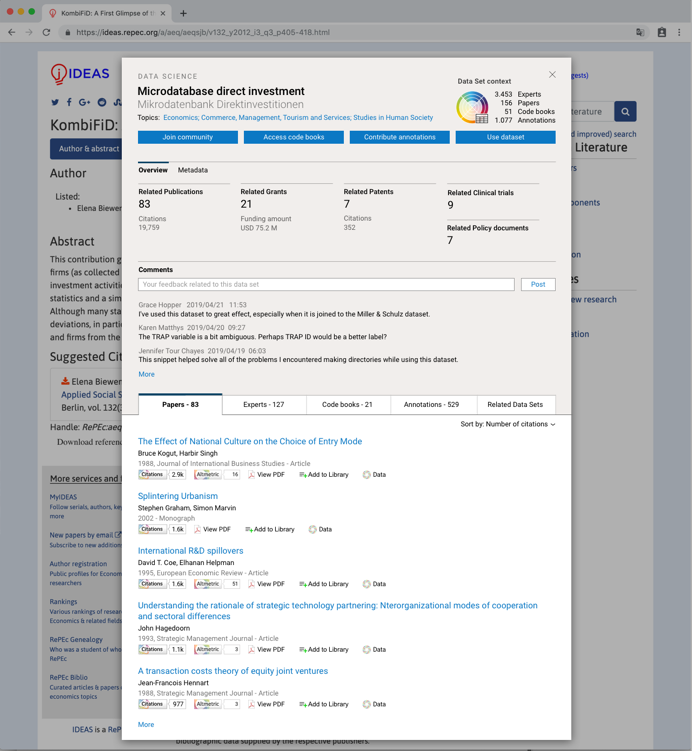
Enhanced context for research data and its impacts could be offered to users in the form of in-app badges and other “signposts” that connect data with its larger context. Such a contextualizing badge could bring together existing data, including not only the number of citations that the dataset has received, but also whether the data has been versioned (through Figshare’s repository metadata), discussed online (through Altmetric data), and what kind of tools and insights have been built on top of the data (through rich mining of full-text and citation data available in the ReadCube reference management corpus and in Dimensions).

Correctly developed and accepted by the community, this type of information can make a contribution to solving many of the problems highlighted in this article. If the correct contextual facets can be developed, then recognition would be easier to assign to those who have contributed to the process of creating and maintaining datasets. With greater context around them, datasets become easier to locate, understand and value. This in turn could lead to a broader evaluative environment and more engagement from academics.

Engagement across academia, however, is not uniform. Mechanisms need to be provided to engage data science-focused researchers from whom more details of their tools, scripts and codebooks could be drawn, adding further value to research data. At the same time, engagement tools need to allow data scientists to leverage this information so that it is valuable to them when they are the consumers of search results. These are subtly different use cases from those of standard researchers. By mining ever more open research systems wherein data is being analyzed (e.g. Gigantum, Github, etc), we can start to integrate these other crucial engagement contexts as well.



*Figure 3.2: Mock-up of a research data badge helping to contextualise a set of search results.*



*Figure 3.3: Mock-up of a research data badge helping to contextualise a specific dataset.*

In Figures 3.2 and 3.3, we have visualised some early concepts for how a contextualized research data badge could look. This visualisation is based on insights from the Rich Context project and uses data that could be mined from articles that use a specific dataset. In particular, we suggest four facets of context that both data science-focused researchers and others could find helpful when viewing a dataset:

* **Experts** **who have made use of the data**, sourced from references made to the dataset in a professional context such as an industry whitepaper or policy document
* **Academics** that **cite the data**, mined from citation of the dataset or ancillary data in the peer reviewed literature
* **End users of the data**, sourced from code book references included in public code repositories
* **Enhancements of the data**, vis-à-vis annotations and comments made on the data in public forums.

In summary, we believe that, if deployed across the many environments in which researchers discover data, the thinking behind the Rich Context project can overcome both the cultural and information-based infrastructure challenges that we highlighted. If these challenges can be overcome by the methods developed, for example in Chapter 13 of this volume, then this will significantly extend the use and discoverability of datasets. The number and variety of datasets in use in academia will certainly expand in the future, and we can only see data becoming even more central to contemporary research efforts. As such, it is critical to invest in robust infrastructures, not only to support the production and sharing of these data, but also to change the culture and evaluative environment around research data. It is only through initiatives such as these that we will be able to solve the vast and complex sociotechnical challenges that face academia today.

## References

Allen, L., Scott, J., Brand, A., Hlava, M., Altman, M., 2014. Publishing: Credit where credit is due. Nature 508, 312–313. https://doi.org/10.1038/508312a

CERN, 2019. CERN Open Data Portal [WWW Document]. URL http://opendata.cern.ch/search?page=1&size=20&subtype=Collision&type=Dataset (accessed 11.30.19).

CISER, 2019. CISER. URL https://ciser.cornell.edu/data/data-archive/ (accessed 11.30.19).

DataCite Metadata Working Group, 2016. DataCite Metadata Schema 4.0 [WWW Document]. URL https://support.datacite.org/docs/schema-40 (accessed 11.30.19).

Eijnatten, J. van, Pieters, T., Verheul, J., 2013. Big Data for Global History: The Transformative Promise of Digital Humanities. BMGN-LCHR 128, 55. https://doi.org/10.18352/bmgn-lchr.9350

Giuliano, F., 2019. Humanités numériques et archives : la longue émergence d’un nouveau paradigme. Documentation et bibliothèques 65, 37. https://doi.org/10.7202/1063788ar

Grusin, R., 2014. The Dark Side of Digital Humanities: Dispatches from Two Recent MLA Conventions. differences 25, 79–92. https://doi.org/10.1215/10407391-2420009

Hook, D.W., Porter, S.J., Herzog, C., 2018. Dimensions: Building Context for Search and Evaluation. Front. Res. Metr. Anal. 3, 23. https://doi.org/10.3389/frma.2018.00023

Jisc, 2019. Research Data Management Toolkit | Jisc [WWW Document]. URL https://rdmtoolkit.jisc.ac.uk/plan-and-design/data-management-planning/ (accessed 11.30.19).

Jordan, K., Weller, M., 2018. Academics and Social Networking Sites: Benefits, Problems and Tensions in Professional Engagement with Online Networking. Journal of Interactive Media in Education 2018, 1. https://doi.org/10.5334/jime.448

Kamp, K., Herbell, K., Magginis, W.H., Berry, D., Given, B., 2019. Facebook Recruitment and the Protection of Human Subjects. West J Nurs Res 41, 1270–1281. https://doi.org/10.1177/0193945919828108

Leurs, K., Shepherd, T., 2017. 15. Datafication & Discrimination, in: The Datafied SocietyStudying Culture through Data. Amsterdam University Press, Amsterdam. https://doi.org/10.1515/9789048531011-018

Li, N., Li, T., Venkatasubramanian, S., 2007. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity, in: 2007 IEEE 23rd International Conference on Data Engineering. Presented at the 2007 IEEE 23rd International Conference on Data Engineering, IEEE, Istanbul, pp. 106–115. https://doi.org/10.1109/ICDE.2007.367856

Linek, S.B., Fecher, B., Friesike, S., Hebing, M., 2017. Data sharing as social dilemma: Influence of the researcher’s personality. PLoS ONE 12, e0183216. https://doi.org/10.1371/journal.pone.0183216

NNLM, 2019. Data Management Plan | NNLM [WWW Document]. URL https://nnlm.gov/data/data-management-plan (accessed 11.30.19).

Siddle, J., 2014. I Know Where You Were Last Summer: London’s public bike data is telling everyone where you’ve been. I Know Where You Were Last Summer. URL https://vartree.blogspot.com/2014/04/i-know-where-you-were-last-summer.html (accessed 11.30.19).

Stark, L., 2018. Algorithmic psychometrics and the scalable subject. Soc Stud Sci 48, 204–231. https://doi.org/10.1177/0306312718772094

Treadway, J., Hahnel, M., Leonelli, S., Penny, D., Groenewegen, D., Miyairi, N., Hayashi, K., O’Donnell, D., Digital Science, Hook, D., 2016. The State of Open Data Report. Digital Science. https://doi.org/10.6084/M9.FIGSHARE.4036398.V1

## Biographical information

Christian Herzog is CEO of Dimensions and Chief Portfolio Officer at Digital Science. A medical doctor by training, Christian also studied economics and started in 2005 Collexis, a software company focused on text-mining based software applications for the research space. In 2010, Collexis was acquired by Elsevier where Christian spent the following two years as the VP for Product Management SciVal. in 2013, Christian and his co-founders started ÜberResesarch as part of Digital Science which led to the launch of Dimensions as a large-scale research information infrastructure in 2018.

Daniel Hook is CEO of Digital Science. He co-founded Symplectic while studying for his PhD in theoretical physics at Imperial College London in 2003. Symplectic became one of Digital Science’s first investments in 2010. Daniel continues to be an active researcher and holds visiting academic positions at Imperial College London and at Washington University in St Louis. He has written more than 30 academic papers and has co-authored a book on Quantum Theory. Daniel is a Fellow of the Institute of Physics, a Policy Fellow at CSaP in Cambridge and serves on the ORCID board as its treasurer.

Mark Hahnel is the CEO and founder of Figshare, which he created whilst completing his PhD in stem cell biology at Imperial College London. Figshare provides research data infrastructure for institutions, publishers and funders globally. Mark is passionate about open science and its potential to revolutionize the research and has led the community in the development of research data infrastructure. Mark sits on the DataCite board, the DOAJ advisory board, the judging panel for the National Institutes of Health (NIH), Wellcome Trust Open Science prize and acted as an advisor for SpringerNature’s masterclasses.

Stacy Konkiel is the Director of Research Relations at Dimensions and Altmetric. Stacy’s research interests include incentives systems in academia and informetrics, and she has written and presented widely about altmetrics, Open Science, and research data services. Previously, Stacy worked with teams at Impactstory, Indiana University & PLOS. You can learn more about Stacy at [stacykonkiel.org](http://www.stacykonkiel.org/).

Duane Williams is VP of US Government at Digital Science. Duane earned his doctorate in theoretical chemistry from the Quantum Theory Project at University of Florida. His current work focuses on improving strategic research investment decisions through new data sets, new tools and metrics to gain insight into the global research landscape. Prior to joining Digital Science, he served as Senior Scientific Analyst and Project Manager for the IP and Science division of Thomson Reuters (now Clarivate Analytics). There he designed and led custom analyses and software development to facilitate data-driven objective assessments of research programs.