JLSC

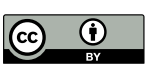
ISSN 2162-3309 | JLSC is published by the Pacific University Libraries | http://jlsc-pub.org

Volume 3, Issue 2 (2015)

Paving the Way For Data-Centric, Open Science: An Example From the Social Sciences

Astrid Recker, Stefan Müller, Jessica Trixa, Natascha Schumann

Recker, A., Müller, S., Trixa, J., & Schumann, N. (2015). Paving the Way For Data-Centric, Open Science: An Example From the Social Sciences. *Journal of Librarianship and Scholarly Communication,* 3(2), eP1227. http://dx.doi.org/10.7710/2162-3309.1227



ISSN 2162-3309 10.7710/2162-3309.1227

PRACTICE

Paving the Way For Data-Centric, Open Science: An Example From the Social Sciences

Astrid Recker
Preservation and Data Management Training Coordinator,
Data Archive for the Social Sciences, GESIS - Leibniz Institute for the Social Sciences

Stefan Müller
Researcher in the GeorefUm Project,
Data Archive for the Social Sciences, GESIS - Leibniz Institute for the Social Sciences

Jessica Trixa
Research Data Management Specialist,
Data Archive for the Social Sciences, GESIS - Leibniz Institute for the Social Sciences

Natascha Schumann
Preservation Specialist and Repository Manager,
Data Archive for the Social Sciences, GESIS - Leibniz Institute for the Social Sciences

INTRODUCTION Data has moved into the spotlight as an important scholarly output that should be shared with the scientific community for replication and re-use in new contexts. This has a direct impact on libraries, archives, and other service providers in the data curation and access landscape. **DESCRIPTION OF PROJECT** The GESIS Data Archive for the Social Sciences (DAS) has been curating and disseminating social science research data since 1960. The article presents tools, services, and strategies developed by the DAS to support the research community in adequately responding to the legal, ethical, and practical challenges that the transformation towards data-centric, open science presents. These include GESIS's Secure Data Center, the data publication platform "datorium" and a recent project to create a georeferencing service for survey data. **LESSONS LEARNED** The experiences gained through these activities show that getting involved—now, rather than further down the road—pays off in that it allows service providers to actively shape the ongoing transformation. At the same time, by cooperating with suitable partners, the effort and investment of resources can be kept at a manageable level for individual organizations.

Received: 02/27/2015 Accepted: 06/05/2015

Correspondence: Astrid Recker, GESIS - Leibniz Institute for the Social Sciences, P.O.Box 12 21 55, 68072 Mannheim, Germany, astrid.recker@gesis.org

INTRODUCTION

Data is at the center of the most significant transformation that the research ecosystem is currently undergoing. Our understanding of scholarly communication as primarily based on textual publications is moving towards including data as an important scientific outcome that should be accessible to the scientific community and beyond. This is, among other things , the result of a growing conviction that the results of publicly funded research should be publicly available (see, for example, Berlin Declaration, 2003) and concerns over quality and reproducibility of research (de Mesquita et al., 2003; King, 1995). At the same time, we are gaining better insight into the benefits of data sharing, which "includ[e] the ability to discover and re-use data which has already been collected, thus avoiding redundant data collection and saving time and money; and providing opportunities for collaboration" (Callaghan et al., 2012, p. 108).[1] Indeed, we are in the midst of a drift towards what The Royal Society has termed "intelligent openness," which hinges on the accessibility, assessability, and reliability of the data underlying research (2012, p. 7).

As service providers for researchers, libraries and archives offer an important part of the infrastructure needed to promote and accommodate this paradigm shift. This includes technical infrastructure in the form of repositories and, maybe more importantly, a support infrastructure for researchers who wish to document and share their data (see Reznik-Zellen, Adamick, & McGinty, 2012). The transformation to which these support services respond is ongoing: even in 2015 we are still in the process of "transitioning from a document-centric view of science to a data-centric view, and the infrastructure is not yet in place for the seamless sharing and re-use of scientific data" (Stuart, 2015). In consequence, rather than merely reacting to the changes and developments briefly sketched above, libraries and archives can still actively shape the transformation towards data-centric, open science.

In this article, we discuss some of the trends and developments that occur as part of this transformation and the challenges that they present for data curation and access. Following a literature review, the challenges that we discuss are: 1) Accommodating the legal and ethical framework for sharing data in the social sciences, 2) Creating tools and workflows for data publication, 3) The emergence of new data types, 4) The fragmentation of the data curation and access landscape. The discussion of each aspect will be complemented by

[1] It should be noted that data sharing is far from a recent phenomenon. Thus, some disciplines have a long tradition of data sharing and archiving (see Martinez-Uribe, 2014, for a historical overview of social sciences data archiving services). It is true, however, that both the (technical) possibilities and expectations of data sharing have a different quality today.

practical examples of services and products offered by GESIS and the GESIS Data Archive to illustrate possible ways of answering to the identified challenges. In conclusion, the article will briefly summarize lessons learned and practices that we believe any organization involved in providing infrastructure for an open, data-centric science should keep in mind as they join the ride.

Although the discussion will be primarily from the perspective of the social sciences, the challenges we focus on are relevant to other disciplines as well. Similarly, while the decision to focus on these specific challenges was informed by our role and experience as a data archive, we strongly believe that in important aspects this experience is similar to that of, for example, academic libraries—because the roles that we fulfil, as providers of research support and infrastructure, are similar as well.

BACKGROUND

Founded in 1960 as one of the first archives for social science data worldwide, the GESIS Data Archive for the Social Sciences (DAS) looks back on a history of over 50 years of curating, preserving, and disseminating data. The archive offers long-term preservation for a comprehensive collection of digital social science survey data, which is reviewed, processed, and documented to provide re-use value to the scientific community. The collection can be accessed via an online catalog (https://dbk.gesis.org/dbksearch/).

Today the DAS is a department of GESIS–Leibniz Institute for the Social Sciences, the biggest social science infrastructure institution in Germany. In addition to carrying out social science research projects of its own, GESIS offers support services throughout the data lifecycle: from initial research and study planning to data collection and analysis to data registration and archiving (see http://www.gesis.org/en/services/). GESIS is also a service provider for CESSDA, the Consortium of European Social Science Data Archives (CESSDA) (http://www.cessda.net), an umbrella organization dedicated to fostering cooperation and the creation of synergies between the contributing archives.

The combination of active research with the provision of comprehensive services to the social science research community is at the core of GESIS's self-definition. In consequence, the ongoing changes in the culture of scholarly communication and the challenges associated with them deeply affect who we are, what we do, and how we do it—now and in the future. With the projects and services described in this article we aim to get actively involved in shaping the transformation: to promote a culture of open science and data sharing, and to ensure that GESIS and the DAS continue to be relevant in this environment.

LITERATURE REVIEW

Among the policy documents and reports that have shaped the European discussion around the transformation towards a more data-centric, open science, HLEG [High Level Expert Group on Scientific Data] (2010), Van der Graaf and Waaijers (2012), and The Royal Society (2012) have been particularly influential. All three formulate visions for an infrastructure that supports data sharing; discuss challenges, drivers, and barriers; and make recommendations on realizing the vision. The topics they primarily address are:

- Incentives for and barriers to data sharing: All reports emphasize the importance of a reward system making it possible for researchers to gain recognition for sharing their data.
- Data collection and preservation, especially the question of what is collected and who preserves it.
- Data access and use, with a particular focus on the issues of accessibility and interoperability, security and trust, as well as understandability (metadata).
- Who pays for the data infrastructure, and how much does it cost?
- The need for education and training for both researchers and professionals (demand for data scientists and data librarians).

Among the many publications describing the status quo of data sharing and the required data infrastructure, LERU [League of European Research Universities] (2013), Van Den Eynden and Bishop (2014), and LIBER [Ligue des Bibliothèques Européennes de Recherche] (2014) provide an overview of the current situation in Europe. Addressing research universities, LERU's *Roadmap for research data* emphasizes that establishing a data sharing culture requires "leadership at an institutional level" (2013, p. 4) and puts a particular focus on the need for support services and tools for data management and publication.

Van Den Eynden and Bishop (2014), a study based on 22 interviews with researchers, explores researchers' motivation to share their data, the benefits of doing so, and the impact that policies and existing support infrastructure have in this regard (p. 15). Among others, the study points to "cultural norms" and "intrinsic incentives" as factors that have the capacity to stimulate data sharing (pp. 27-29).

To complement its recommendations published in Christensen-Dalsgaard (2012), the Association of European Research Libraries recently published 11 Research Data Management case studies from European universities, each highlighting the approach taken by the respective institution to create an infrastructure that facilitates data management and sharing in HEIs (LIBER, 2014).

CHALLENGE 1: LEGAL FRAMEWORK

The current European discussion summarized above has led to changes in the framework conditions of science, e.g. in the form of changed funding policies. However, this policy-oriented part of the framework for research is complemented by another one, namely the legal conditions determining if, when, and how data can be shared. Two legal areas are especially relevant where social science research data are concerned: copyright and data protection law.

Whenever archives and repositories accept data for preservation and dissemination, they have to make sure that no submitted materials infringe the copyright of third parties. It is not uncommon, for example, that survey instruments (e.g. certain combinations of items in questionnaires) are protected by copyright and may accordingly not be distributed without permission from the rights holder.

However, copyright not only affects the sharing of contextual materials, it also applies to the data itself. A comparison of copyright legislation in Germany, the Netherlands, the United Kingdom, and Denmark showed that different criteria for protecting and licensing research data exist (CIER, 2011). This concerns, among others, the threshold for protection (e.g. data are more likely to be protected in the UK than in other countries) or the question of who owns data that were created as part of an employment contract. This means that multi-national research projects may have to comply with different legal requirements depending on where the data is created and/or used.

In the social sciences and other disciplines that carry out research entailing human subjects, data protection regulations determine what can and cannot be done when collecting and sharing data. In Europe, the overarching framework for data protection is Directive 95/46/EC, adopted in 1995 (European Parliament, 1995).[2] In German legislation, the protection of personal data is governed by the Federal Data Protection Act ("Bundesdatenschutzgesetz") and the Data Protection Acts of the individual federal states ("Landesdatenschutzgesetz").

[2] Directive 95/46/EC does not adequately address the implications that the technological progress of the last years has on data protection. Therefore, and to put an end to the current fragmentation of country-specific regulations, the European Commission (EC) proposed a major reform of the data protection legislation in 2012 (European Commission, 2012). The proposed regulation, which strengthens privacy protection for individuals, has sparked considerable debate in the European research community because it appears to unduly limit the possibilities of research (see Kvalheim, 2014). It remains to be seen to what extent the resulting legislation will be capable of balancing the rights of individuals with the valid concerns and interests of the research community.

The purpose of these laws is to protect the individual's right to privacy (§ 1 BDSG) and accordingly they regulate the collection, processing, and use of personal data by public authorities, private organizations, and in scientific research. This includes, for example, that personal data must be anonymized as soon as the research purpose allows doing so. Until then, any characteristics that would make it possible to directly identify a person have to be stored separately from the other data (§ 40 BDSG). Thus, it is strictly prohibited to store data and address data of survey respondents in the same location (Metschke & Wellbrock, 2002). Moreover, the collection, processing, and use of personal data is only allowed if the person in question has freely given their consent (§ 4 BDSG).

As a rule of thumb, the less anonymized data is, the richer it is. To enable researchers to work with such rich but sensitive data in accordance with legal obligations some institutions offer Secure Data Centers or Data Enclaves. Providing a secure, restricted-access environment, these centers allow researchers to work with sensitive data and therefore help balancing the "trade-off" between easy access and tapping into the rich potential of weakly anonymized data. It enables the sharing of data that could, due to legal and ethical considerations, not be shared otherwise. At the GESIS Secure Data Center (http://www.gesis.org/sdc) researchers can either work on-site at a Safe Room workstation or, for selected datasets, off-site by signing a contract that binds them to fulfill special security requirements. In both cases trained staff supervises the analysis process by, for instance, consulting with guests and reviewing any research output guests have created.

To support researchers in planning and carrying out their research projects, the CESSDA Training team (http://www.cessda.net/training), located at the DAS, offers workshops in research data management. Our experience from teaching these workshops shows that researchers often lack awareness of the legal issues that arise from their research, especially where data protection and copyright are concerned. This often means that at the point when they offer data to an archive, it suddenly turns out that sharing the data is not possible —because they infringe the rights of third parties, or because informed consent was not sought from respondents. Thus, discussing these issues has become an important part of our workshops. However, we often find that researchers generally come to these workshops quite late in their research—sometimes too late to rectify certain problems. It is crucial that researchers are made aware of these problems as early as possible. This is something where libraries and other research support services in universities and research institutes come into play, because often they are a first point of contact for researchers as they plan their research.

CHALLENGE 2: DATA PUBLICATION

A large portion of the discussion around data sharing focuses on the question of how to incentivize it (see, for example, Van Den Eynden and Bishop, 2014; APA, 2011). Providing

one possible answer to this problem, the concept of data publication emphasizes that to make data a "first class object of scholarly communication" (JLSC, n.d.) the scientific community needs to establish mechanisms and procedures that allow for the publication of data in a similar way as publishing articles.

A formal publishing process for data amounts to more than just making data available somehow and somewhere. It entails measures of quality control and mechanisms to ensure that this data can not only be found but also be understood and re-used by others (see Callaghan et al., 2012). In addition, it "also provides a mechanism for allowing data producers to obtain academic credit for their work in creating the datasets" (Callaghan et al., 2012, p. 109). While this acceptance of data publications as scholarly outputs equal to text publications ultimately requires a change in the scholarly reward system, it can be facilitated by tools for the citation and sharing of research data, for example, persistent identifiers and citation conventions (see Helbig & Hausstein in this issue) and repository systems.

In the context of scholarly communication the quality of publications is of particular importance. Accordingly, the traditional (i.e. document-centric) publishing system employs quality control measures—most commonly in the form of peer review. Although an increasing number of journals now ask for the submission—and sometimes publication—of data underlying a manuscript to check the quality of articles (Zenk-Möltgen & Lepthien, 2014; for policy examples see Silva, 2014; JLSC Editorial Board, 2014), the discussion of what suitable quality control and review mechanism are for data is ongoing (see Kratz, 2014; Lawrence, Jones, Matthews, Pepler, & Callaghan, 2011).

Whichever measures of quality control a repository, archive, or journal decides to employ, it is of the essence to create the greatest possible transparency concerning these measures. It has to be clear to data depositors and users which measures are employed to be able to comply with the requirements or to decide whether they trust the data or not. The DAS communicates its requirements and procedures through different channels. These include information on our webpages (http://www.gesis.org/en/services/archiving-and-registering/data-archiving/), policy documents and guidelines (e.g. GESIS Data Archive 2010; 2012), as well as publications (e.g. Jensen, 2012) and workshops on research data management and digital preservation. This denotes a significant shift from previous practice, when hardly any of this information was made available—partly, because it was deemed irrelevant and uninteresting to our stakeholders; partly also because for a considerable time, the DAS was not one among many services competing on the market but operated in a closely-knit community of "insiders" as the only European service provider in this field. It is only with the ongoing professionalization and standardization of digital curation and preservation practice that the expectation and requirement of increased transparency has arisen.

Our understanding of the archive as a service provider in an increasingly international and heterogeneous market (see Challenge 4 below) led us to consider how we can best meet the data publication demands and requirements of different stakeholder groups. Historically, the archive has offered a "standard" archiving and publication procedure involving extensive quality control and descriptive and subject cataloging to provide metadata and (unstructured) documentation, e.g. in the form of methods reports. Where necessary, value added services for special collections were offered, involving data harmonization and the creation of variable-level metadata.

However, in the light of the ongoing paradigm shift in scholarly communication, solutions are required for the so-called long tail of data—data generated by smaller projects and individual researchers. To provide more flexibility concerning the types of data accepted by the archive and to lower the threshold of data submission for this stakeholder group at the same time, datorium (https://datorium.gesis.org) was implemented as an additional GESIS service for the documentation, upload, and publication of research data. By creating a user-friendly offer for researchers and projects that do not have access to (or need for) significant resources for data management and curation, datorium helps to fill the gap between fully-fledged long-term data preservation and data lost on local hard drives.

datorium allows data depositors to determine the access conditions for the use of their data according to their needs, ranging from "open to everyone" to "access requires depositor permission." A persistent identifier (DOI®) guarantees that the published data can be reliably located, retrieved, and cited. To ensure that data is usable and understandable, and to address potential legal and ethical problems, all data and relevant documentation are reviewed by experienced data curators for completeness, coherence, and data privacy. This step has proven indispensable as submitted data frequently exhibits problems with the anonymization of participants which need to be addressed before publication. This is an issue that especially institutional and other non-subject specific repositories have to keep in mind when accepting data from empirical social science research (see CESSDA Training, 2013).

datorium also will serve as a data repository connected to social science journals, where research data related to publications can be published for replication and re-use purposes. datorium therefore also supports the creation of a more standardized, networked data publication infrastructure.

CHALLENGE 3: NEW DATA TYPES

Similar to most scientific disciplines, new data types emerge rapidly in the social sciences. This development is due to the technical progress and the increasing variety of disciplinary

methods, often based on new modes of data collection. Social media data are one of the most prominent new data types in the social sciences. Deriving from internet platforms such as Facebook or Twitter, they can, for example, be used as barometers for public opinion and forecasts in election studies (Tumasjan, Sprenger, Sandner, & Welpe, 2010). While social media data provides information on individual behavior, small-scale spatial data concerns the immediate living environment of respondents, for instance in regard to the neighborhood, housing, or municipal infrastructure. Merging survey data with spatial data—a process referred to as georeferencing—opens up new potentials for analyses, for example, by allowing us to explore whether correlations exist between the living environment and individual behaviors or opinions.

Both social media data and small-scale spatial data present challenges which are immediately relevant to how the data is managed, curated, and disseminated. These challenges relate to a) data acquisition and harmonization, and b) data privacy and consent.

Acquisition and harmonization

Often social media and spatial data are generated and distributed by for-profit organizations. Harvesting and re-using such data can therefore be expensive or even prohibited by terms of service. Furthermore, in contrast to traditional social science data, little to no institutionalized mechanisms and workflows yet exist to deliver and to archive this data for the specific purpose of research. This affects the availability of such data and thereby makes it difficult to replicate the results of research based on it.

In the realm of social media data, the challenges that preservation and dissemination pose are illustrated by the Twitter archive hosted by the Library of Congress (LoC). Since 2010 the LoC is working towards building and providing access to its Twitter collection, which in 2013 contained 170 billion tweets. To provide access, both technical and conceptual issues have to be resolved. These relate, for example, to the response time when searching the collection, or to the question which information has to be provided to keep the collection accessible and understandable (Library of Congress, 2013). Even today this data cannot be easily accessed by researchers, which hints at the complexity of the issues that have to be resolved.

In Germany, a non-commercial source for small-scale spatial data is the public administration. However, access to this data is nonetheless difficult. As spatial data is often collected by small municipal agencies in order to fulfill certain legal requirements, there is not always an infrastructure in place that allows researchers to easily access and use the data.

In response to this challenge, the DAS currently carries out the GeorefUm project with the objective of building an infrastructure for the access and reuse of spatial data in the social sciences (http://www.gesis.org/forschung/drittmittelprojekte/projektuebersicht-drittmittel/georefum/). The long-term goal is to create a georeferencing service allowing researchers to merge their own survey data with small-scale spatial data. The challenges that the GeorefUm project will address are, again, technical and conceptual: Although standards for file formats exist (e.g. ESRI shapefiles), not all of them are compatible with each other. Thus, data originating from different sources often require harmonization, e.g. by migrating file formats, unifying attributes, or changing spatial units. The project now collects and harmonizes actual spatial data for a case study to develop the strategies and competencies required to solve these problems. As with the social media data, the following questions have to be answered: which context information is required to make the spatial data re-usable and how can this information be provided to data users? The GeorefUm project seeks to draw on DDI,[3] a well-established social science metadata standard (see Vardigan, Heus, & Thomas, 2008), to achieve this objective.

Privacy and consent

The preservation and dissemination of social media and small-scale spatial data also requires us to address a number of legal and ethical issues. Thus, analyzing social media data often means analyzing social networks. These are highly sensitive regarding data privacy since they can yield information on location, contacts, interests, socio-demographics, etc. of any person in the network. In consequence, there is a high risk of re-identifying individuals by combining certain facts. Moreover, social science researchers follow certain ethical guidelines that, among others, require that participants in research give informed consent. These requirements are often not sufficiently met in social media research and are therefore subject of an ongoing debate (Zimmer & Proferes, 2014).

The strict data protection legislation in Germany makes it necessary to develop strategies that enable archives and researchers to comply with this legislation without making research impossible. Addresses in form of coordinates are indispensable in the process of georeferencing data. Providing a technical infrastructure for merging survey data with spatial data therefore requires strong regulations to ensure that no direct identifiers of respondents are leaked into the merged data. Moreover, spatial data may contain values which are unique to a specific location. As a result, even after deleting direct identifiers, georeferenced survey data can bear the risk of re-identifying survey participants. Providing a secure, strongly regulated

[3] DDI: http://www.ddialliance.org/

environment such as GESIS's Secure Data Center is one way of addressing this problem. However, as we work towards conceptualizing and establishing a larger-scale georeferencing service, it is also necessary to develop criteria and general guidelines helping us to assess the risk for a re-anonymization of survey respondents in a given research project and to decide which protection measures and dissemination strategies are adequate.

CHALLENGE 4: FRAGMENTED CURATION AND ACCESS LANDSCAPE

As data sharing and publication become more and more important, the number of players in the field continues to grow. This results in an increasingly distributed and heterogeneous landscape of service providers in data curation and data access. For example, in February 2015, the Re3data registry listed just over 200 repositories for the subject category "Social and Behavioural Sciences" (Re3data, n.d.). In Germany, social science and statistical data are offered by a host of different organizations including government agencies on a national or federal level, archives, research data centers, and commercial services (for an overview see RatSWD, n.d.).

Among the many challenges that are associated with this "fragmentation" and de-centralization of the data services landscape, two seem particularly relevant from the perspective of both users and service providers: a) The more fragmented the landscape of data access is, the more difficult it becomes for users to find and access data that could be relevant to their research. b) At the same time, the strongly distributed data curation landscape makes it difficult if not impossible to ascertain that important data is indeed curated and preserved for re-use. Both challenges need to be addressed – to ensure that the scientific record is adequately preserved and to lower the barrier for researchers to publish or re-use data.

To move from fragmentation to distributed but connected services, both technical and organizational measures are required. Among the more technical ones are interoperability and standardization in the form of common protocols (e.g. OAI-PMH, SWORD[4]), metadata standards (PREMIS,[5] subject specific standards), and adoption of standard procedures for curation and preservation (supported, for example by OAIS [CCSDS, 2012], and certification procedures such as Data Seal of Approval, DIN 31644 or ISO 16363[6]).

[4] OAI-PMH: http://www.openarchives.org/pmh/; SWORD: http://swordapp.org

[5] PREMIS: http://www.loc.gov/standards/premis/

[6] See http://www.trusteddigitalrepository.eu/Trusted%20Digital%20Repository.html for an overview of these procedures.

COAR [Confederation of Open Access Repositories] (2015) provides a comprehensive overview and discussion of interoperability issues currently faced by repositories.

To address this challenge, the CESSDA partners are in the process of planning a new portal, which will include a cross-catalog search for all CESSDA archives. For this purpose, an updated version of the shared CESSDA metadata schema based on DDI/XML is currently being developed.

The latter example points towards the importance of the organizational dimension in addressing the fragmentation of the curation and access landscape. Thus, collaborations in combination with a clear distribution of roles (on an institutional, national, and international level) can help to ensure that no significant gaps exist in the infrastructure and that at the same time duplicate efforts are minimized. The DAS is involved in a number of activities that help to promote cooperation between different players in data curation and access both nationally and internationally. For example, as pointed out above, the expertise of the DAS lies in quantitative survey data. Faced with more and more mixed methods surveys, which result in both quantitative and qualitative data, we have to develop solutions for distributed but connected data curation and access with partners who have expertise in qualitative data. With this in mind, the Verbund Forschungsdaten Bildung (http://www.forschungsdaten-bildung.de) was established as a cooperation between the DAS, the German Institute for International Educational Research, and the Institute for Educational Quality Improvement. The objective of this joined effort is to create a distributed preservation infrastructure with a single point of contact for producers and users of quantitative and qualitative data.

CONCLUSION AND LESSONS LEARNED

This article presented selected technical and organizational challenges that service providers face in the current transformation towards a data-centric, open science. In the face of the sometimes overwhelming amount and speed of changes, we would like to conclude with three messages:

Get involved! As service providers for the research community, we have to find ways of connecting to this community—be it as researchers ourselves, as liaisons to research groups or departments, or by relying on online and offline communication channels. Thus, a policy at the DAS and GESIS is that service staff should also be actively involved in a research community, e.g. in the social sciences or in the digital curation domain.[7] In our

[7] An example of such an engagement at the DAS are Bruns and Weller (2014) and Kinder-Kurlanda and Weller (2014), who are actively engaged in the emerging field of social media research.

experience, this active involvement in the community means that we are visible and credible to researchers. This helps us tackle the one challenge that could turn out to be the most difficult to address: researcher attitude. Talking about data sharing from the perspective of an insider rather than an outsider enables us to meet researchers on their own ground. This makes it easier to change their attitude towards data sharing and publication—from reluctance to acceptance to, maybe, enthusiasm.

Start now, but give it time! While we would caution anyone to jump into projects head first, it is our experience that sometimes we have to take a pragmatic approach to things: get started rather than getting stuck in what and ifs at the planning stage. Thus, to find out if something is going to work, sometimes we have to try it. When we started developing workshops and offering consultations on research data management for researchers in the social sciences in 2011, it was not yet a hot topic in Germany and many other countries of the CESSDA network. It was unclear what the demand was going to be and which topics would be relevant to the community. However, by actually teaching the workshop repeatedly, we were able to develop a much better understanding of what the burning questions and everyday research data management problems of the community are. This has allowed us to adapt and fine-tune the content of the workshops to meet the needs and expectations of our participants (e.g. by addressing legal issues).

At the same time that we urge service providers to start offering support and building an infrastructure sooner rather than later, it is our experience that sometimes offers take time to catch on. This was certainly true for our workshops, which only now—after over three years—are finally met with significant demand. It is, however, also an experience we made in building the datorium service. Seeing as there are hardly any mandates for data sharing from publishers or research funders in Germany, it was not surprising (albeit somewhat disappointing) that researchers did not line up in long queues to hand over their data as soon as datorium went live. But, after about a year of operation, we are noticing a slow but continuous uptake in the submissions and the interest in datorium. As publishers and funders are finally beginning to increase the pressure on researchers to share their data, we have an infrastructure already in place to meet the demand.

Don't be an island! Service providers, especially in smaller institutions, should always be keenly aware of their environment and the other players in the field. They should actively seek out partnerships and form alliances, both within their own organization and with other institutions. Rather than feeling pressed to reinvent the wheel, all of us should remind ourselves to tap into the existing infrastructure already out there. In this way we can use the limited available resources most effectively.

At the DAS, cooperation has been one answer to the difficult question of how to reconcile demands for increasing the depth and the breadth of our services, voiced by internal and external stakeholders. We have reached a point where technological progress allows us to do amazing things with data, especially if we enrich them with machine-actionable context information, harmonize, and merge them with other data. However, in the light of a data landscape which becomes more and more interdisciplinary, diversified, and heterogeneous, we have to decide where to invest our resources. Striking a balance between breadth (preserving as much as possible) and depth (generating as much added value as possible) has proven much easier since aligning with partners who have expertise in areas where we do not. It is these partnerships which bring us closer to the vision of a distributed but connected data infrastructure formulated in HLEG (2010)—an infrastructure enabling "seamless access, use, re-use, and trust of data" in such a way that "the physical and technical infrastructure becomes invisible and the data themselves become the infrastructure—a valuable asset, on which science, technology, the economy and society can advance" (p. 4).

REFERENCES

Alliance for Permanent Access (APA). (2011). *The ODE project: Ten tales of drivers & barriers in data sharing*. Retrieved from http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/10/7782_ODE_Brochure_v5.pdf

BDSG, Bundesdatenschutzgesetz. (1990). Amended and promulgated on 14.1.2003 I 66; last changed by art. 1 Gv. 14.8.2009 I 2814. Retrieved from http://www.gesetze-im-Internet.de/bdsg_1990/

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. (2003). Retrieved from http://openaccess.mpg.de/Berlin-Declaration

Bruns, A., & Weller, K. (2014). Twitter data analytics – or: the pleasures and perils of studying Twitter (guest editorial for special issue). *Aslib Journal of Information Management*, *66*(3), 246-249. http://dx.doi.org/10.1108/AJIM-02-2014-0027

Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., … 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. http://dx.doi.org/10.2218/ijdc.v7i1.218

CCSDS. (2012). *Reference model for an Open Archival Information System (OAIS). Recommended Practice* (No. CCSDS 650.0-M-2). Retrieved from http://public.ccsds.org/publications/archive/650x0m2.pdf

Centre for Intellectual Property Law (CIER). (2011). *The legal status of research data in the Knowledge Exchange partner countries*. Retrieved from http://www.knowledge-exchange.info/default.aspx?id=461

CESSDA Training. (2013). *Self-archiving platforms and data verification*. Retrieved from https://admtic.wordpress.com/2013/11/12/self-archiving-platforms-and-data-verification/

Christensen-Dalsgaard, B. (2012). *Ten recommendations for libraries to get started with research data management. Final report of the LIBER working group on E-Science / Research Data Management.* Retrieved from http://www.libereurope.eu/sites/default/files/The research data group 2012 v7 final.pdf

COAR. (2015). *COAR Roadmap. Future directions for repository interoperability*. Retrieved from https://www.coar-repositories.org/files/Roadmap_final_formatted_20150203.pdf

De Mesquita, B. B., Gleditsch, N. P., James, P., King, G., Metelits, C., Ray, J. L., … Valeriano, B. (2003). Symposium on replication in international studies research. *International Studies Perspectives*, 4, 72-107. http://dx.doi.org/10.1111/1528-3577.04105

European Commission. (2012). *Data protection*. Retrieved from http://ec.europa.eu/justice/newsroom/data-protection/news/120125_en.htm

European Parliament. (1995). *Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data*. Retrieved from http://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX:31995L0046

GESIS Data Archive. (2010). *Usage regulations: Access categories*. Retrieved from http://www.gesis.org/en/services/data-analysis/data-archive-service/usage-regulations/#3_Access_categories

GESIS Data Archive. (2012). Digital preservation policy. *Principles of digital preservation at the Data Archive for the Social Sciences*. Retrieved from http://www.gesis.org/fileadmin/upload/institut/wiss_arbeitsbereiche/datenarchiv_analyse/DAS_Preservation_Policy_eng.pdf

High Level Expert Group on Scientific Data (HLEG). (2010). *Riding the wave. How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data*. Retrieved from http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

Jensen, U. (2012). *Leitlinien zum Management von Forschungsdaten: Sozialwissenschaftliche Umfragedaten*. GESIS Technical Reports, 2012/07. Retrieved from http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2012/TechnicalReport_2012-07.pdf

Journal of Librarianship and Scholarly Communication (JLSC). (n.d.). *Call for Papers: Special issue on sharing, publication and citation of research data* (Summer 2015). Retrieved from http://jlsc-pub.org/jlsc/cfp.html

JLSC Editorial Board. (2014). The article is not enough: Introducing the JLSC data sharing policy. *Journal of Librarianship and Scholarly Communication*, *2*(3), eP1186. http://dx.doi.org/10.7710/2162-3309.1186

Kinder-Kurlanda, K., & Weller, K. (2014). "I always feel it must be great to be a hacker!" The role of interdisciplinary work in social media research. *Proceedings of the 2014 ACM Conference on Web Science*, 91-98. http://dx.doi.org/10.1145/2615569.2615685

King, G. (1995). Replication, replication. *PS: Political Science & Politics*, *28*(3), 443-452. http://dx.doi.org/10.2307/420301

Kratz, J. (2014, May 8). Fifteen ideas about data validation (and peer review). *Data Pub: Blog about all things data from the California Digital Library*. Retrieved from http://datapub.cdlib.org/2014/05/08/fifteen-ideas-about-data-validation-and-peer-review/

Kvalheim, Vigdis. (2014). EU Parliament vote on new data protection legislation.

Lawrence, B., Jones, C., Matthews, B., Pepler, S., & Callaghan, S. (2011). Citation and peer review of data: moving towards formal data publication. *International Journal of Digital Curation*, *6*(2), 4-37. http://dx.doi.org/10.2218/ijdc.v6i2.205

LERU Research Data Working Group. (2013). *LERU Roadmap for research data*. Retrieved from http://www.leru.org/files/publications/AP14_LERU_Roadmap_for_Research_data_final.pdf

LIBER Scholarly Communication and Research Infrastructures Steering Committee. (2014). *Research data management case studies*. Retrieved from http://libereurope.eu/committee/scholarly-research/research-data-management-case-studies/

Library of Congress. (2013). *Update on the twitter archive at the Library of Congress*. Retrieved from http://www.loc.gov/today/pr/2013/files/twitter_report_2013jan.pdf

Martinez-Uribe, I. (2014). *Chronology of data library and data centres*. Retrieved from http://iassistdata.org/blog/chronology-data-library-and-data-centres

Metschke, R. & Wellbrock, R. (2002). *Datenschutz in Wissenschaft und Forschung*. Retrieved from http://www.datenschutz-berlin.de/attachments/47/Materialien28.pdf?1166527077

RatSWD. (n.d.). *Further data resources*. Retrieved from http://www.ratswd.de/en/data-infrastructure/other

Re3data. (n.d.). Repository search. Filter: Social and behavioural sciences. Retrieved from http://service.re3data.org/search/results/filter?term=&d=25&filter_subjects_active[social%20and%20behavioural%20sciences]=Social%20and%20Behavioural%20Sciences

Reznik-Zellen, R., Adamick, J., & McGinty, S. (2012). Tiers of research data support services. *Journal of eScience Librarianship*, *1*(1), 27-35. http://dx.doi.org/10.7191/jeslib.2012.1002

Silva, L. (2014, February 24). PLOS' new data policy: Public access to data. *The PLOS ONE Community Blog*. Retrieved from http://blogs.plos.org/everyone/2014/02/24/plos-new-data-policy-public-access-data-2/

Stuart, D. (2015). Libraries could play key role in managing research data. *Research Information*. Retrieved from http://www.researchinformation.info/features/feature.php?feature_id=497

The Royal Society. (2012). *Science as an open enterprise*. Retrieved from http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 178-185.

Van Den Eynden, V., & Bishop, L. (2014). *Sowing the seed: Incentives and motivations for sharing research data, a researchers' perspective*. Retrieved from http://www.knowledge-exchange.info/Files/Filer/downloads/Primary Research Data/Incentives/Incentives_for_Sharing.PDF

Van der Graaf, M., & Waaijers, L. (2012). *A surfboard for riding the wave. Towards a four country action programme on research data*. Retrieved from http://www.knowledge-exchange.info/Default.aspx?ID=469

Vardigan, M., Heus, P., & Thomas, W. (2008). Data Documentation Initiative: Toward a standard for the social sciences. *The International Journal of Digital Curation*, *1*(3), 107-113. http://dx.doi.org/10.2218/ijdc.v3i1.45

Zenk-Möltgen, W., & Lepthien, G. (2014). Data sharing in sociology journals. *Online Information Review*, *38*(6), 709-722. http://dx.doi.org/10.1108/OIR-05-2014-0119

Zimmer, M., & Proferes, N. J. (2014). A topology of twitter research: disciplines, methods, and ethics. *Aslib Journal of Information Management*, *66*(3), 250- 261. http://dx.doi.org/10.1108/AJIM-09-2013-0083