Data in the making : Temporal aspects in the construction of research data

Haider, Jutta; Kjellberg, Sara

Published in:
New big science in focus : Perspectives on ESS and MAX IV

2016

8. Data in the making: Temporal aspects in the construction of research data

Jutta Haider & Sara Kjellberg

Increasingly the material research deals with is cast as data, and more and more as digital data, a seemingly unproblematic concept with which to describe the matter of research at all stages of the research process, from the object of investigation to the output. The aim of this essay is to complicate this framing by investigating the ways in which notions of data emerge in the construction of new big science facilities, in order to explore some of the implications for how and when knowledge production is thought to occur. We study data and the making of data during the design and construction of two large-scale research facilities in southern Sweden, the ESS and MAX IV, and specifically of the necessary infrastructure for dealing with various aspects of research data management. The making of data does not refer here solely to the data produced during an experiment or an observation, but rather to how they are made possible by setting up and planning for the production, storage, and use of data, and even the limitations, strategic roles, and other effects.

'Rarely can a magic moment be established when things become data' writes Christine Borgman (2015, 62), and as she develops at length, the question commonly asked—What are data?—and which occupies policy makers, lawyers, university administrators, data service staff, and archivists, might not be the most interesting or even the most relevant. Rather,

considering the various ways in which data always exist in a specific moment and in relation to particular conditions, a more adequate and indeed more productive question, she suggests, is when are data? We take inspiration from Borgman (2015) and let this question guide our exploration. Hence, our intention is not to compare different understandings of data, even less to judge which is preferable, but to a gain a diverse and faceted understanding of the meaning of research data in the process of building a research facility, and of the temporal aspects to how these meanings are shaped.

'One of the founding myths of scientific practice is that science is carried out in an eternal present. From it all external influence has been banished', writes Geoffrey Bowker (2005, 32–33) in his book on memory practices in the sciences. Of course, the archive is one of the most central functions for science as an institution. As an organized collection of the records of past science, it introduces a temporal axis to scientific knowledge and knowledge production that is equally foundational. The archive, as a memory institution, articulates a specific relationship between the objects and the records of science. Bowker (2005, 36) further reminds us 'all things on earth can be seen as at once objects and archives'. As institutions, archives are involved in turning things into documents, which can then be stored, described, organized, accessed, and put into new contexts (see Briet 1951). However, with the increasing significance of computers and information and communications technology (ICT) for knowledge production in the sciences (Hine 2006), the process of documenting objects has been fundamentally complicated. The question of what it actually *is* that is turned into documents is getting increasingly difficult to answer, at the same time as the need to store data for longer periods and as openly accessible has grown exponentially. Furthermore, as the number of stakeholders involved in the process of documenting in the sciences increases, it has become more and more obvious that the relationship between object and memory institution—archive, library—has to be recast (see Hansson 2015). Our focus is on the apparatus, including the work, functions, and policies, that enable data collection and processing before the actual research can be carried out. As Borgman (2015, xviii) reminds us, 'data rarely are things at all. They are not natural objects with an essence of their own'. Data are records of something, and in this way they are born as

documents. At the same time they are also the very objects that need to be turned into documents to be added to the archive. This essay is intended to better the understanding of how this happens, and when.

Research data management: Between data-driven science and open data

At time of writing, two new big science facilities are being built just outside the city of Lund, the multinational ESS and the national MAX IV laboratory, located next to each other with a planned science village alongside. While MAX IV is a new facility that developed from an existing centre for synchrotron research at Lund University, the other facility, a neutron source, is an international effort spanning several European Union countries. In addition, both the ESS and MAX IV are multidisciplinary, with research covering physics, chemistry, geology, biology, and medicine, and are primarily intended to serve researchers from a wide community, including industry and different disciplines, and to develop an infrastructure to support the users as temporary visitors when performing experiments. The ESS will have a dedicated data management and software centre to handle, analyse, and possibly store research data emanating from the experiments. Interestingly, this centre is located in Copenhagen, Denmark, on the other side of the Öresund.

The digital aspects of doing research and its implications for handling data as part of the research process are important for all fields today, and interest in the role of computers in knowledge production in the sciences has grown in step with the emergence of concepts such as eScience, data-intensive research, and also big data analytics (see Borgman 2007, 2015; Hine 2006; Meyer & Schroeder 2015;). Data-intensive research can be described as being based on data sets that are analysed using computers, algorithms, and statistical methods. There have been attempts to distinguish between eScience, eResearch, data-driven research, and computational research (see Griffin 2013; Ray 2014), but without much success. Even though there can be differences in these categories of research, the main discussion turns on how changes of methodology and approach might shape contemporary scholarship or research (Borgman 2007, 2015; Ekbia

et al. 2014; Frické 2015). Dealing with huge data sets and large amounts of data is part of this development, and thus is linked to the phenomenon of big data (Borgman 2007, 2014; Boyd & Crawford 2012). Big data is debated in relation to with the epistemics of knowledge production and data-driven science, and put in opposition to problem-or theory-driven science (Frické 2015; Ekbia et al. 2014; Leonelli 2014).

The production and use of digital data and the challenges this poses on research also call for new knowledge in order to handle the data successfully (Ray 2014). Research data management involves more than the individual researcher's work in managing, describing, sharing, archiving, and preserving research data; as has been pointed out, a multitude of supporting roles are required to cover all the different aspects of managing the complex of research data (Verbaan & Cox 2014). Studies show that new support services are being developed (Antell et al. 2014; Cox et al. 2012; Cox et al. 2014; Griffin 2013; Mayernik 2015; Ray 2014; Verbaan & Cox 2014). Research data management is part of a whole chain of research documentation both before, during, and after a project takes place. The organisation of documentation includes descriptions to discover the data and metadata about the data-sets, and also how data have been managed to make them trustworthy in order to prevent data loss and possible file corruption (Ray 2014).

Moreover, it is research data management that lies behind the idea of making data open—available and accessible—and of sharing data. The assumption is that making data freely available for others to use will benefit society and promote new ways of using the data and cross-connecting with other data sets. Over decade ago, Arzberger et al. (2004) explained the principle as follows: 'publicly funded research data should be openly available to the maximum extent possible'. This is now also spelled out in the policies of the National Science Foundation (NSF) in the US, the EU in its open access strategy, and various other national and international funding bodies and research councils. Increasingly, funders demand that data management plans be included in grant applications (see Arzberger et al. 2004; Tenopir et al. 2011).

Additionally, governments encourage researchers to share their data, and periodicals have begun to include requirements for uploaded data sets

when submitting manuscripts for publication (Borgman 2015, 8). To make data open you need not only a technical infrastructure, but also routines, which depend on organizational capacity (Meyer & Schroeder 2015, 184). At the same time several studies have made it clear that there are also challenges (see Axelsson & Schroeder 2009), given that there are no uniform practices for data-sharing (Beaulieu 2003; Hine 2006; Tenopir et al. 2011). Even within the same discipline, different approaches can be found (Mayernik 2015). The attitudes of researchers and the social shaping of research communities also have an effect on making data open, as do possible developments in both official policy and the technology as such (Meyer & Schroeder 2015, 186).

Material and analysis

Our aim is to understand the ways in which notions of data emerge in the construction of big science facilities, and specifically of the infrastructure for research data management. This is shaped by the expansion of data-driven science and the paradigm of making data freely available. Thus we chose to interview people working in the support and administrative organization at the new facilities in Lund and to collect and analyse documentation available from the facilities' websites. Our source material also included the slides for a presentation that one of our respondents shared with us. In addition we took part in a group meeting with several representatives from the ESS, where we heard various presentations and were able to ask questions. This helped us to identify possible interviewees and relevant organizational groups and to draw up the interview guide.

The potential number of interviewees was limited by the type of expertise relevant to our study. We began by contacting and interviewing people working at the ESS. During those interviews we found that, in order to better understand our preliminary analysis, we needed a more diverse material. Hence we decided to include an interview with someone in a similar position at MAX IV. This proved fruitful as it showed that data management at MAX IV is facing the same challenges, as the very similar views expressed in the interview would seem to indicate.

In total we conducted five interviews with seven respondents. The

interviewees worked in support functions—legal, communications, and curation/data management—and we chose to focus on the latter group. Our questions concerned their understanding of data and their views on research data as part of the development and construction of the facilities. The interviews were semi-structured, with a set of questions designed to elicit responses about the following three themes: disciplines and user groups; data and metadata; and sharing data. The questions were then slightly adapted in order to accommodate each interviewee's field of expertise.

The interviews lasted approximately one hour each. We recorded all but one interview, and subsequently listened to them repeatedly, took notes, and transcribed the relevant parts. The transcriptions and notes, together with the documents, formed the basis for our analysis. We constructed themes by repeatedly going back and forth in the material to identify commonalities and differences that emerged during the analysis. In what follows, we bring together the salient points from the interviews, before interrogating the material using our guiding question, 'When are data?'

Some emergent meanings, or, what are data?

Most of our interview material is derived from interviews with staff in leading roles at the ESS and MAX IV, who work with different aspects of systems development to enable research data management. While their views also dominate our analysis and provide the greatest detail, a study of research data from legal and public relations perspectives using other documentation can contribute to an understanding of the full range of demands and requirements that determine how data are envisioned and systems to handle them are built.

Below we present the most tangible understandings of research data as they emerged in the interviews, which also serve to set out our general findings and the ways in which different notions of research data are conceptualized (and in relation to what, where, and whom). The focus when talking about research data was largely determined by who the interviewees were. This is not to say that our interviewees were unaware of

other meanings; on the contrary, their roles as mediators between different groups in many ways demands a high degree of awareness, and in fact all our interviewees expressed concern at the way communication between groups and systems worked, and all reflected on their roles in this.

Data as a technical issue

The interviewees who work with research data management and software issues all had leading roles in developing the computer-based, technical infrastructure. All had a scientific background, and all had used similar facilities as researchers before going over to working with research data management. Their understanding of research data was shaped by seeing data as part of interlinked processes and other data such as metadata gathered using the instruments. In this perspective, data become a technical issue and are treated as such. Here, data never just are—they are always being dealt with, changed, reduced, moved, sent, or described, and always in relation to technology of some kind, whether a database, reduction program, metadata registry, analysis tool, storage device, fibre cable, or visualizing software and the like.

A lot of thought and effort has gone into preparing the facilities' user services. This chimes well with how the ESS describe its handling of software and data, for example—'The ESS is putting special emphasis on creating and using first-class software for instrument control, data processing, analysis, and visualisation' (ESS n.d. 'The unique')—and is also reflected in the design of MAX IV (MAX IV 2010). Here too data are presented as a technical issue, and dealing with them is seen as a service for researchers. As the MAX IV website explains, 'Companies wishing to solve their research needs at the MAX IV Laboratory can be offered initial discussions with expert laboratory scientists, sample preparation, assistance during measurements and help with data analysis and interpretation' (MAX IV n.d.). Both facilities will cater to academic researchers with different disciplinary backgrounds and users from industry. They are assumed to have different requirements regarding data management and processing, not least interface design and visualization. 'Everyone has the challenge of making it straightforward and giving the user community the

kind of analysis and the tools required to actually get the scientific information and impact out of the data, which traditionally has been the work of a Ph.D.' (Int. 2). Things are meant to be kept simple, 'otherwise they'll just get confused, because these are people who do biology. They are not neutron scatterists', said one respondent, and continued, 'the time when you could make a career out of just neutron scattering … those days are gone, it's more multidisciplinary, so you have to cater for a broader range of scientific disciplines … and provide them the tools they can reasonably access.' Later he added, 'The same goes for the data. That needs to be presented in a way that they understand what that is, rather than, well I mean obviously, rather than giving them neutron events, but also…' (Int. 2). Clearly, the expectation that there will be a growing and more diverse user base has implications for how software development and research data management are tackled. These expectations shape what future users will experience, see, measure, and interpret when they finally get to encounter 'their' data on a screen.

Data as a problem

From a legal perspective, data are mostly approached in terms of a problem to be dealt with. How to define data is a pressing issue, as is awareness of different interests that need to be accommodated, different user groups who have to be served, and laws and policies adhered to. The most important goal was said to be serving the so-called science community, a community that was terminologically cast as the legal department's most important client. However, it is less clear who exactly makes up that science community, as it features as a vaguely homogenous bloc. Specifically, the role of different disciplinary cultures or the relations between industrial research and university research seemed unclear or remained unexpressed. Attempts to circumscribe data in relation to this nominal client, as to laws, regulations, and policies, were felt by the legal team to be something of a challenge, with data being neither immaterial nor material, neither object nor archive—or both at the same time—and so escaping the existing legal terminology. While free and open access to research publications is quite unproblematic, at least in the sense that there is a widely shared

understanding of what a publication is and what we should encounter when we access one, what exactly should be regulated when data are freely available is a lot more opaque (see Wennersten & Maunsbach, in this volume). The conflation of data and intellectual property is common, and one of the thorniest issues is the role research facilities will have in opening up access to research data, which implies control of the data in the first place.

Data as a possibility and responsibility

From a public relations and communications perspective, research data was framed as both a possibility and a responsibility. As one interviewee put it, 'The data that we produce at this facility is our raw material. We have to help our users or create processes ourselves if we want to get the most of our raw material' (Int. 4). Research data are connected to a vision of science that highlights science's potential to solve societal problems and to advance knowledge for the common good. Industry and EU funding frameworks, which require industrial partnerships, function as categories that help describe research as useful. However, when it comes to research data, open data—which is thought of as something that might be problematic for the demands of industry—is seen as a means to make visible how the facility does what it is supposed to do, namely produce science. This is documented in the form of research data made available for others to see and reuse. This way, research data is inscribed into a double narrative of future opportunity and evaluatory control.

Managing the flow, or, when are data?

One of our respondents opened the interview by claiming that it was premature to talk about data. In hindsight, this introduced a number of complexities to the topic that we had not appreciated at the outset, for what it expresses is not so much the way in which people in different positions perceived the issue as more or less urgent, but rather the transient character of research data. If data can be premature, when are they mature? we have

to ask. The transience of research data that we encountered is expressed in the various temporally structured descriptions employed, and also in the way in which the digital materiality of data in use is constantly changing. Put bluntly, data are never in and of themselves, but exist only in relation to other data, software, and instruments, to people, measurements, interfaces, computers, or various other tools. We will sketch out some of the most tangible ways in which the temporality of research data emerged in our material, in relation to the processes of doing research, to ideas of an archive, to policy demands, and, probably most of all, to understandings of the future. We use these themes as pragmatic categories to outline the various ways in which data are imagined as having been configured over time. In that sense, they are neither mutually exclusive nor do we ascribe values intrinsic to these themes nor to the institutions they refer to.

Data are what researchers will take back with them or access remotely after they have done an experiment using one of the instruments at the ESS or a beamline at MAX IV. Everything that is being built, installed, and programmed as the instruments and data centres are installed is meant to lead up to this. Yet the data that will be stored on a hard drive and taken back on a plane or accessed over a network will have undergone a series of reductions, translations, and contextualizations since the neutron or X-ray beam has met the sample. They will undergo numerous further treatments in order to be calculable, publishable, storable, describable, and accessible—or to be overwritten and deleted. Yet, there is one magic moment, to use Borgman's words, when researchers first encounter their data, when they see the data coming in from the instrument and displayed on a computer screen, or as one respondent envisaged it, 'so, we publish the data frame by frame on our computers. The people who are doing the experiment, they are sitting at a terminal' (Int. 1). He went on to describe it as 'a publish–subscribe system (it's like Netflix) … it's streaming. Multiple subscribers can subscribe to the same film' (Int. 1). There are two interrelated questions here. Firstly, what do researchers see when they look at data, as it was put repeatedly in our interviews, and how did this data get to be data at that very moment? In order for data to be something that can be looked at as it moves past on a screen like a film, as our respondents described it, entire series of translations must have occurred. At some point there must have

been a decision on how to visualize the way in which, say, neutrons hit a sample in meaningful way—as a graph, a scatterplot, a 3D image—and, indeed, how to deal with that data from that point on, not least when moving on from processing to putting records in a file for storage.

Data are anything but static in the accounts we were given. They do not arrive on the researcher's computer desktop ready and waiting to be handled. Interviewees likened data to a film that happens on a screen, and that has a clear temporal dimension to it, as its series of moving images elapse over time. This is also the way it is presented in an organizational chart entitled 'Data Acquisition, Reduction & Control' that one of our respondents showed us. Here we get a picture of how data are meant to be processed from the instrument to the screen. Numerous lines and arrows connect boxes with names of software, metadata standards, illustrations of instruments and storage devices, all illustrating a flow of data from the experiment via a series of automated data aggregation and reduction steps, involving time-stamping and metadata descriptions, to the instrument control room. This too is included: a little box showing a person sitting at a desk in front of a screen with colourful dots, seemingly watching a data visualization as if it was a film. Clearly, a great deal happens to the data before they are even encountered *as* data by the researchers. And none of this is forgotten, because the process is added to the archive and attached as metadata. As one respondent put it, 'the data framework used for data reduction keeps a history of reduction itself of everything that was done to the data' (Int. 2). Yet, while processes are kept as records, it is also a priority to reduce visible complexity and to speed things up. This can be connected to the question of how researchers are constructed as users, with research data management seen as a service for those users who need to be presented with a simple interface (see. Hine 2006).

Yet, time is important here in a different way—'Not in real-time, but we are pushing for it', and 'you can get it after a few minutes' (Int. 3), as one of the interviewees puts it. The issue here is the time that elapses between the experiment taking place in the instrument and the data becoming visible to the researchers, and thus is as much a matter of efficiency as speed in carrying out research, something which our interviewee touched on in different ways. Immediacy is positioned as the ideal. This, of course, would

allow for beamlines or instruments to be used more efficiently, which is financially beneficial. Yet the research process itself has a part to play here. Researchers, we learned from our respondents, might want to tweak samples and adjust the set-up of their experiments in direct response to the data, as they are visualized on screen. In a way this also aspires to immediacy, where the elapse between neutrons hitting the sample and the data being visible on a computer screen would ideally shrink to almost nothing, making the computer and software—after an enabling series of translations that are as rapid and invisible as possible—something to see through rather than see with. 'Absolutely at the top of the wish list, and what we are trying to get to work right now, is to get some type of integration. I have seen so many individual solutions that do not fit together in their context; it is at the very top of my list, and it is the thing which no one will see, it will just work' (Int. 3), said one interviewee, underlining the significance of the often invisible infrastructure. The ideal of offering a 'real-time' as well as a 'ready-to-use' interface for the data film, supported by an increasingly invisible technical infrastructure, ultimately also depends on immediacy, intended to maximize the impression of control for users.

Policies and regulations

Policies on open and free access to scholarly publications have become commonplace throughout the world, and many of the world's largest funders now demand open access to publications that result from research funded by them. Increasingly, this has also been extended to encompass open access to the original data too, usually labelled open data policies. Regulatory moves have been made to circumscribe and regulate research data and institutional responsibilities are being negotiated (Borgman 2015, 42–5).

Policies, regulations, guidelines, and data management plans all impact on the research facilities we studied, and on many levels. All relate to data, but it remains an elusive concept which, while clearly significant and laden with values, expectations, and even capital, is very hard to pin down as a policy category—and this despite the fact that it, like the all-important issue of time and timing, has implications for how data management

services and processes are prepared. Local guidelines are drafted to reflect the policies researchers have with them from their home institutions or funders, while industry users are considered to need specific regulations and possibly exceptions. Legal requirements, rights, regulation, and the question of ownership all play an important role, as do notions of how researchers treat and create value from their data—and when they do it (see Arzberger 2004).

As one of our interviewees said of research data, 'As a researcher, you own it in within the embargo period. It's your data for a period of time … Generally this is set by the data policy of the facility … after that it somehow becomes open access?' (Int. 2). Embargo periods, during which individual researchers have exclusive rights to their data, are common in most fields of research, and will likely also feature in the type of research carried out at MAX IV and the ESS. Data change meaning in relation to variously negotiated periods of the research process, defined by who has access to the data; however, with periods that can vary significantly between disciplines in both length and scope, it is unclear how they might tally with the requirements of access policies and funder demands, which are a lot more uniform. Equally unclear is the role of research facilities in negotiating what these timings should be, and how best to express the agreed times in their research data management plans.

Before data are even collected, there has to be a moment when they are described, however cursorily, in a research application, and more and more frequently in a data management plan too, as they are increasingly demanded by funders (see Mullins 2014). One interviewee used past experience to illustrate the problems with these plans: 'We had this in the US, because the NSF were asking people to hand in a data management plan, and, yes, users came rushing to the facilities and said, well, if you keep the data forever can we call that our data management plan' (Int. 1). Clearly, funders' requirements shape the demand for data management plans, and the focus is almost entirely on long-term preservation. Thus, by virtue of their mere existence, data management plans project data into a future, where they are primarily meant to be stored. Sure enough, each discipline's specific research culture works with the scholarly publishing system to shape how data are thought of, and here too time is a key

reference point. 'Research is an incremental process, essentially', as one interviewee said, continuing, 'So if you have more steps—and access to more data *will* give you more steps—you'll have a better stab at making some reasonable new understanding of [inaudible], which is the point of the literature. Literature only gives you one side of the story. It gives you published data after some period of distillation within their group … it doesn't give you everything else that they did.' His colleague added, 'Then journals require you to submit raw data … and when you're going to high-impact journals, there's "supplementary information", and people just put lots of stuff in supplementary information' (Int. 2). Here the idea of the scientific literature as continuously advancing, with each publication building on the ones before, means that research data are thought to take on different guises, depending on when in the publication process they feature. Raw data, published data, supplementary information, 'stepping stone' data—all make their appearance here, and all are conceptualized in terms of doing science as a linear process that plays out over a period of time, from the data from previous research, to raw data, to research data, to distilled and eventually published data, to supplementary information, and so on.

From data in use to data in waiting

From the framing of data as a technical concern and a service for researchers, it is a short step to accessibility and use, and from there to short- and long-term preservation. These are often constructed in relation to archiving as a question of disk space, with reference to vague temporal factors, the designation of different uses, and, importantly, the non-use of data.

The website of one of the Lund facilities offers this description of its computing centre: 'The primary activity is the operation of the high performance computing cluster, which is used by scientists who rely on computer modelling in order to support the design of the ESS facility and consists of two main parts: a high performance scientific computing cluster and a high performance storage and backup system' (ESS n.d. 'Computing'). Research processes and data storage are disconnected from each other, not only in time, but also in regard to the digital infrastructure (Leonelli 2014).

'We would keep the data … we have planned to sort of keep the data for ever' (Int. 1) said one of our interviewees:

> We are still some years out. I talked about the data being copied to multiple locations, the stream going to Copenhagen, going to Lund, also still being on the instrument in multiple copies, in case something went wrong. What I imagine is that we would also automatically copy this to one of the archive facilities, either something like EUDAT or possibly CERN. (Int. 1)

Data are seen to exist in multiple copies and on devices in different physical locations, with back-ups needed on the instrument as a safeguard. Interestingly, 'the instrument' is used as a stand-in for all the data processing and computing connected to the experiment. A long-term archive copy is more of a possibility than a certainty, and it is clear that the interviewees partly reflected on this because we posed questions about it; archiving, and specifically long-term archiving, was not an issue that came to the fore otherwise.

Archiving is mostly framed as a question of storage, not unlike an analogue archive, which is identified with its physical space: 'What determines how long we can save the data is how much money we have. We must have more disk space, that is what the question is about' (Int. 3). Another interviewee saw the costs as less of a problem, as 'storage is cheap' (Int. 1), yet here too archiving was predominantly described in terms of storage, and not so much a question of access or maintenance. Maintaining software or other means to access old file formats were thought less relevant, to the point of being almost speculative: '[Our] hope would be that the file format is still valid for the software in the future' (Int. 2). This also has to do with a vision of what constitutes long-term archiving for different purposes. In the public discourse, as in various official policies, archiving is presented as being synonymous with long-term preservation. With no time limits defined, the default appears to be 'forever'—indefinite preservation (see Kimpton & Minton Morris 2014). Routines for deletion, as a form of sanctioned and controlled forgetting, are not part of this framing of the archive or data's function in it. Yet, this long-term view is absent from the planning of research data management at the Lund

facilities, where data are described in terms of time frames that are more directly instrumental. One interviewee talks of mere months:

> Would it now be, now we are looking in the crystal ball, say that we would have a greater responsibility when it comes to open access and long term storage then we have basically the infrastructure to … we do not store, as we have said now we save data for 3 months in order to be able to bring home your data … (Int. 3)

Regarding a longer-term perspective, he continued, 'We must not rebuild the system to store it longer or to make it available for longer; at least as far as I know, there is nothing more going to happen to the data after 6 years than there is after 20 years' (Int. 3). Here, data are seen to have a 'use-by date', after which they become inactive and are irrelevant for the planning of the data processing necessary for future experiments. At the same time, this is also problematized, for as another respondent reasoned, 'it is not fair to compare the uselessness of 10-year-old data today for how it will be in the future … better metadata might make today's data more useful in the future' (Int. 2).

The way in which the archive is framed reflects the conflict inherent in an instrumental view of data's place in the research process, and when data are inactive and have passed their use-by date. Access and use are relevant during the initial period, but afterwards data are put on hold and reduced to a question of disk space and storage, where they are at best held in waiting, but generally are defunct. Yet, that said, the same data can also hold a different future, a future when they might be found useful, and this is framed in terms of an opportunity, an expression of hope—'what if' or 'just in case'. That hope that data might have a life in a near or distant future also emerged in our material, yet here again this was vague and contradictory at times.

Timewise, research is done on tight margins. Lack of time necessarily factors in when discarding certain questions or not following certain paths. By saving data, the assumption is that researchers can go back to it later, when they have more time to follow up on interesting points noted at the time: such future data serves to delay the present, offsetting some of the

pressures of delivering fast results while wanting to be thorough. Future data can 'provide opportunities to do things' (Int. 3) in new contexts, or 'in twenty years' time you might be interested in different effects or you could store the intended effect with the data. So you could replay the visualization' (Int. 2). Quite apart from this being good PR for the facility, it is exactly the point of open data policies that advocate long-term or perpetual preservation (see Arzberger 2014; Meyer & Schroeder 2015, 175–6). The hope is that technology will advance knowledge, almost by itself: 'You can imagine the future: that by the time by we get to 2020, 2022, that kind of time, that maybe you'll already have the algorithms available to have machine-learning tools that could help to qualify that data' (Int. 4). The hope, the possibility, that technology could be the driving force in the advancement of science in the long term, stands in contrast to the few months quoted above for the time data would be of real use to researchers. Similarly, data-mining was described as a far-off prospect—'I think it's far in the future. I've not seen anyone in this business who's looked into it' (Int. 3). The role of human researchers, close to their material and their sample, the physical artefact to be studied using the instrument or the beamline, was much more present. Interestingly, the sample is seen to be central for how it is imagined data will be useful outside their context of creation: 'If you don't have an understanding of the sample, it's a different story' (Int. 1) said one interviewee, highlighting the difficulty of making data meaningful through a succession of decontextualizations and recontextualizations.

Concluding remarks

We started from the question 'When are data?' in order to interrogate our material and pinpoint how data are made into objects of research and into documents in the archive of science (Bowker 2005). How data are created in the actual research process, as it is commonly imagined, is not our focus here; rather, all that surrounds and supports these processes. By bringing together the temporal and contextual notions of data, a richer, more diverse, but also more complicated, understanding of research data emerges. We find that research data are not only different things for different disciplines

and in relation to different functions or even policies, but also that they have different meanings depending on when in the research process they are approached, and that the research process starts long before individual researchers start work and extends long after they have finished.

We have investigated how the ESS and MAX IV, two large-scale research facilities, work with data and metadata standards, the software tools with which to handle data, and policy tools and communication strategies. The challenges of data curation, handling, and description are immense, and increasingly big science is discursively associated with big data. Notions of data—what it is, how it should be handled, stored, and accessed, and why—inevitably vary, but all relate to the idea that data are a fundamental, component in the processes that stabilize science. From having been seen as stepping stones in the production of scientific results, data are increasingly positioned as results in themselves (Leonelli, 2014). This shift was also seen in our material. Often the justification given for data preservation and openness is that they might be of some use for new discoveries in the future, although what this means in exact terms is put differently by different groups. Research data engage a multitude of stakeholders, tools, and policies, and so forth in its management, transforming them from an ephemeral procedural element into stable components of scientific research to be handled, stored, and passed on. Clearly, what data are anticipated as doing in different futures plays a role in how data are framed today. Yet, when exactly this future will occur is a lot less clear. The accounts we were given shifted between vague hopes for a time ahead when technology will drive knowledge production and old data will be useful in ways impossible to fathom today, and more cautious, down-to-earth descriptions of technical issues, researchers, the requirements of different disciplines, and actual samples, and issues such as backing-up, file transfers, metadata, and processing, for which the future is just around the corner. Data is framed as occurring in the present, but in passing and on various temporal axes: streamed past the researcher, data go through various processes of enhancement, description, visualization, or recalculation, always on-going, always in the making.

Concerning the handling of research data, the interviewees' focus was often on the perceived needs of users, and the various translation processes

required to enable communication between groups of people, but also between computer systems and software tools. Language metaphors were often employed to conceptualize the mediation of meaning or technical standards. Talk often turned to users—either individual researchers and groups, or industry as a more abstract category—with as many different ideas of who these users would be and what they might want to do. Those who had a background in the sciences imagined their users, and their data management goals and requirements, in greatly more diverse terms; they were alert to disciplinary cultures, policy or funder demands, users' computing skills, and publication or career demands. Across the board, users were portrayed as largely competent in expressing their demands, even when they lacked advanced computing skills. This is in contrast to what others have found elsewhere. Among our respondents from MAX IV and ESS, users were not described as a problem, existing only to disrupt an otherwise well-functioning system—a common way for users to be viewed by technical or other support staff in e-research and elsewhere (see Meyer & Schroeder 2015, 37). On the other hand, they are still seen to require a simplification of complex processes in order to be able to act at the level their qualifications and disciplinary background would indicate.

Our findings make it plain that data are not fixed and never can be. Data exist only by way of mediation, through their descriptions and the various digital tools that make them 'happen'. We explored some of the ways in which this is thought to occur, depending on when in the research process data are assigned a role. Data need to be rendered and related to other sets of data every time they are made manifest. They are emergent, relational, and shaped by their use—and use includes the preparation for data collection as much as archiving. The intricate relationship between data as object and data as archive is complicated further by the data being constantly relocated and redescribed in new contexts in order to function as research data in the first place. This way the archive is continuously delayed as new data objects emerge each time data are processed and made to exist. This brings us back to the question of how objects are made into documents, the central concern of the documentalist movement in the twentieth century (Bowker 2005; Hansson 2015). To conclude, thinking of research data as emergent not only through its entanglement with different

user needs and data-processing tools, but also through the various temporal factors and across time-scales, can enrich our understanding of data as an object of research, an object of memory, and a cultural object of a continuously suspended future.

Interviews

(Int. 1) Interview, 27 January 2015. (RDM)
(Int. 2) Interview with 2 people, 17 February 2015. (RDM)
(Int. 3) Interview, 10 February 2015. (RDM)
(Int. 4) Interview, 20 March 2015. (PR)
(Int. 5) Interview with 2 people, 27 January 2015, not recorded. (Legal)

References

Antell, Karen, Jody B. Foote, Jaymie Turner & Brian Shults (2014), 'Dealing with data: Science librarians' participation in data management at Association of Research Libraries Institutions', *College & Research Libraries* 75(4), 557–74.

Arzberger, Peter, Peter Schroeder, Anne Beaulieu, Geoffrey Bowker, Kathleen Casey, Leif Laaksonen et al. (2004), 'Promoting access to public research data for scientific, economic, and social development', *Data Science* 3, 135–52.

Axelsson, Ann-Sofie & Ralph Schroeder (2009), 'Making it Open and Keeping it Safe: e-Enabled Data Sharing in Sweden', *Acta Sociologica* 52 (3) 213–26.

Beaulieu, Anne (2003), 'Research Woes and New Data Flows: A Case Study of Data Sharing at the fMRI Data Centre, Dartmouth College, USA', in Paul Wouters & Peter Schröder (eds), *Promise and Practice in Data Sharing*, pp 65-88. Amsterdam: NIWI-KNAW.

Borgman, Christine (2007), *Scholarship in the digital age: Information, infrastructure and the Internet*. Cambridge, Mass.: MIT Press.

Borgman, Christine (2015), *Big data, little data, no data: Scholarship in the networked world.* Boston, Mass.: MIT Press.

Bowker, Geoffrey (2005), *Memory Practices in the Sciences*. Cambridge, Mass.: MIT Press.

boyd, danah & Kate Crawford (2012), 'Critical questions for big data', *Information, Communication & Society* 15(5), 662–679.

Briet, Suzanne (1951) *Qu'est-ce que la documentation?* Paris: EDIT.

Cox, Andrew M., Stephen Pinfield & Jennifer Smith (2014), 'Moving a brick building: UK libraries coping with academic data management as a "wicked" problem', *Journal of Librarianship and Information Science*. doi:10.1177/0961000614533717.

Cox, Andrew M., Eddy Verbaan & Barbara Sen (2012), 'Upskilling liaison librarians for research data management', *Ariadne* (70), 1–12.

Ekbia, Hamid, Michael Mattioli, Inna Kouper, Gary Arave, Ali Ghazinejad, Timothy Bowman, et al. (2014), 'Big data, bigger dilemmas: A critical review', *Journal of the Association for Information Science & Technology*. doi: 10.1002/asi.23294.

ESS (n.d.), 'The unique capabilities of the ESS', <http://europeanspallationsource.se/unique-capabilities-ess>, accessed 11 June 2015.

ESS (n.d.), 'Computing centre', <http://europeanspallationsource.se/computing-centre>, accessed 11 June 2015.

Frické, Martin (2015), 'Big data and its epistemology', *Journal of the Association for Information Science & Technology* 66: 651–661. doi: 10.1002/asi.23212.

Griffin, Stephen (2013), 'New Roles for Libraries in Supporting Data-Intensive Research and Advancing Scholarly Communication', *International Journal of Humanities and Arts Computing* 7, Supplement 59–71.

Hansson, Joacim (2015), 'Documentality and legitimacy in future libraries: An analytical framework for initiated speculation', *New Library World*, 116 (1/2), 4–14.

Hine, Christine (2006), 'Databases as Scientific Instruments and Their Role in the Ordering of Scientific Work', *Social Studies of Science* 36(2) 269–98.

Kimpton, Michele & Carol Minton Morris (2014), 'Managing and Archiving Research Data: Local Repository and Cloud-Based Practices'. In Joyce M. Ray (ed.), *Research data management: Practical strategies for information professionals*, pp. 223–38. West Lafayette, Ind.: Purdue University Press.

Leonelli, Sabina (2014), What Difference Does Quantity Make? On the Epistemology of Big Data', *Biology, Big Data & Society*, 1(1). doi: 10.1177/2053951714534395.

MAX IV (n.d.), *Industry Opportunities*, <https://www.maxlab.lu.se/industry>, accessed 11 June 2015.

MAX IV (2010), 'Detailed design report', <https://www.maxlab.lu.se/sites/default/files/DDR_MAX_IV_First_Edition_2010–08–25.pdf>, 11 June 2015.

Mayernik, Matthew S. (2015), 'Research data and metadata curation as institutional issues', *Journal of the Association for Information Science & Technology*. doi: 10.1002/asi.23425

Meyer Eric T., & Ralph Schroeder (2015), *Knowledge Machines: Digital Transformations of the Sciences and Humanities*. Cambridge, Mass.: MIT Press.

Mullins, James L. (2014), 'The Policy and Institutional Framework'. In J. M. Ray (ed.), *Research data management: Practical strategies for information professionals*, pp. 25–44. West Lafayette, Ind.: Purdue University Press.,

Tenopir Carol, Suzie Allard, Kimberly Douglass, Arsev U. Aydinoglu, Lei Wu, Eleanor Read et al. (2011), 'Data Sharing by Scientists: Practices and Perceptions', *PLoS ONE* 6(6): e21101. doi:10.1371/journal.pone.0021101.

Ray, Joyce M. (2014) (ed.), 'Introduction to Research Data Management'. In J. M. Ray (ed.), *Research data management: Practical strategies for information professionals*, pp. 1–21. West Lafayette, Ind.: Purdue University Press.

Verbaan, Eddy & Andrew M. Cox (2014) 'Occupational sub-cultures, jurisdictional struggle and third space: Theorising professional service responses to research data management', *Journal of Academic Librarianship* 40(3–4), 211–19.

New big science in focus
Perspectives on ESS and MAX IV

JOSEPHINE V. REKERS AND
KERSTIN SANDELL (EDS.)

LUND STUDIES IN ARTS AND CULTURAL SCIENCES 8

Lund Studies in Arts and Cultural Sciences is a series of monographs and edited volumes of high scholarly quality in subjects related to the Department of Arts and Cultural Sciences at Lund University. An editorial board decides on issues concerning publication. All texts have been peer reviewed prior to publication.

Lund Studies in Arts and Cultural Sciences can be ordered via Lund University:
www.ht.lu.se/en/serie/lsacs/
E-mail: skriftserier@ht.lu.se

ISBN 978-91-981458-4-7 (print)

Lund Studies in Arts and Cultural Sciences 8
ISSN 2001-7529 (print), 2001-7510 (online)

Cover design Johan Laserna
Layout Gunilla Albertén
Images Louise Wester | Photograph Kerstin Sandell
Printed in Sweden by Media-Tryck, Lund University, Lund 2016

Contents

1. NEW BIG SCIENCE: OPPORTUNITIES AND CHALLENGES 7
Josephine V. Rekers & Kerstin Sandell

2. FROM THE GROUND UP?
LAUNCHING THE ESS, FERMILAB, JLAB, AND THE APS 25
Catherine Westfall

3. HOW CLOSE IS CLOSE ENOUGH FOR INTERACTION?
PROXIMITIES BETWEEN FACILITY, UNIVERSITY, AND INDUSTRY 45
Josephine V. Rekers

4. CAN BIG BE MADE SUSTAINABLE?
ENVIRONMENTAL CONTESTATIONS OVER THE ESS AND MAX IV 71
Anna Kaijser

5. LOOKING AT VALUE-MAKING:
COD AND SCIENTISTS SWIMMING THEIR OWN WAY 97
An interview with Kristin Asdal by Anna Kaijser

6. HOW NEW THINGS COME INTO BEING 105
An interview with Hans-Jörg Rheinberger
by Kerstin Sandell & Catherine Westfall

7. HOW TO DESIGN A QUESTION-GENERATING MACHINE FOR THE FUTURE: INSTRUMENTS AS PART OF EXPERIMENTAL SYSTEMS AT MAX IV 119

Kerstin Sandell

8. DATA IN THE MAKING: TEMPORAL ASPECTS IN THE CONSTRUCTION OF RESEARCH DATA 143

Jutta Haider & Sara Kjellberg

9. DATA AND THE LAW 165

Ulf Maunsbach & Ulrika Wennersten

10. INSTITUTIONAL CHANGE IN SCIENCE ACTIVITIES: THE CASE OF HUMAN SPARE PARTS IN FINLAND 189

An interview with Markku Sotarauta by Josephine V. Rekers

ABOUT THE CONTRIBUTORS 197

ACKNOWLEDGEMENTS 199

This anthology is about new big science, approached through perspectives from law, sustainability studies, sociology of science and technology, history, human geography and information studies. In focus are the two large experimental facilities being built in Lund: the European Spallation Source (ESS) and MAX IV. Put centre stage are the communities that launch, build, use, host, and benefit from these science facilities.

New big science facilities are large in their physical footprints, their costs, and their ambitions. Expectations abound and are in constant production. This anthology captures the opportunity to study these complex science facilities in the making – at the juncture where expectations meet reality – asking questions about what possibilities, constraints, and risks these projects entail.

This anthology is the outcome of an interdisciplinary research theme at the Pufendorf Institute for Advanced Studies at Lund University. The editors Josephine V. Rekers and Kerstin Sandell are both social scientists, affiliated with the Department of Human Geography and the Department of Gender Studies respectively.

LUND STUDIES IN ARTS AND CULTURAL SCIENCES
ISBN 978-91-981458-4-7
ISSN 2001-7529