Embracing the 'Data Revolution':

Opportunities and Challenges for Research

by Matthew Woollard[1]

This is an edited transcript of the keynote plenary speech I gave at the IASSIST Conference in Bergen on 1 June 2016.

Introductory remarks

I think that this is a really exciting time for those of us who work in data services. The so-called data revolution has the potential to change everything which we as data service professionals do, and we must adapt and develop to meet the challenges which are already facing us. And not only are there challenges which are caused by the new and novel forms of data, which seem to be making most of the headlines, there will be opportunities and challenges which relate to the integration of the 'old' with the new.

Before I warm to my theme, I need to provide some context, because this is a personal perspective. I'm the Director of the UK Data Archive[1], which is the lead organisation which runs the UK Data Service[2]. The Archive is involved in other projects, and the other organisations involved in the Service have other projects and responsibilities.

The primary goal of the UK Data Service is to ensure that the high quality data which researchers need in order to do research is got to them as speedily as possible. We have to pre-empt user demand by selecting and acquiring data rapidly; we have to process and document data in order for users to find it and use it, and ensure it can be reused into the future. And we have to make sure that the right access conditions are set up and met, so the right access mechanisms can be used.

About ten years ago the UKDA set up the Secure Data Service which has become the UKDS Secure Lab[3] – which allows us to distribute access rather than just distribute data. Stepping over the 'secure' hurdle coincided roughly with the starting bell for the data revolution. This shift in thinking towards distributing access needs to be carried forward in the new order. In our private lives we have the opportunity of doing just about everything in the 'cloud', why do we rigidly stick to the old paradigm of letting people download data to work on their own computers? (I'll return to answer that question in due course…)

Some further context: the remarks which I'm making here are not based on any empirical research, almost everything I say is based on my recent experiences at work, and in particular observations from a number of groups or committees I've been on in the last year or so. And I want to note that these activities have involved me with government officials, ethics researchers, research funders from across the globe, data service staff, technologists, statisticians, commercial data users and privacy activists. This is a very wide range of stakeholders, and this adds to the complexity of data service infrastructure, because these and other stakeholders have different roles and different priorities. This complexity is not going away, and it is a significant challenge as we move further into the so-called data revolution.

I don't want to focus on stakeholders. I want to discuss a few themes which seem to have coalesced in my work in the last year, but I want you to keep in mind that the data revolution means that the traditional three parties (data owners – data archives – data users) involved in data services are changing. That's a challenge.

Trust

My first theme is on Trust. Not on Trusted Digital Repositories[4], but in the more philosophical sense.

Traditionally, most of the data held by social science data archives is highly anonymised, and the people involved have given some form of consent for information about them to be used. For some surveys, especially censuses, there are explicit statements about confidentiality which are designed to give subjects a sufficient level of trust in the survey organisation to respond, and to respond truthfully. If we assume (and we shall for a moment) that the data revolution is typified by the availability of new kinds of data, in pretty significant quantities, which have research potential but are not being created for the purpose of research then one aspect of the data revolution should be in the broad change in the overall attitude of the data controllers. Is this true? I'm not sure that it is. Official data controllers protect because they have a legal obligation to; commercial organisations do too – and they often have an additional commercial interest. With more data breaches publicised people seem to be less trusting than ever before. And people are clever – they seem to reduce their requirements for trust when they receive something in return. (It would be nice to have some real evidence for this beyond the anecdotal.) Trust has become more complex. Data subjects (or customers of internet services) seem to be more acutely aware of the context in which they provide personal information.

But, with some sorts of data, for example, internet shopping sites, data subjects are not given an obvious choice – there is not what one would call informed consent, which means that re-use beyond the original service provided becomes more tricky – not impossible, but generally more ad hoc.

I believe that across the next decade or so, there will be a continued need to try and return to first principles. How do we best protect and respect the data subject while maximising the potential for use in research (and maximising the value that research can bring). My feeling is when we – as the data service community – demonstrate our commitment to the respect of data subjects and to the protection of those data subjects, access to many of the new forms of data will become easier. We also need to find ways of expressing that trust we have in the researchers and that we have in our secure settings and in ensuring that the data controllers and the data subjects are more aware of them. Great work has been done in the UK in terms of administrative data [5, 6], and in the use of personal data in commercial organisations, [7] but this is not the end of the story. The first big challenge facing us in the data revolution is giving citizens and data subjects more opportunity to trust the activities of researchers (wherever they are based, and whoever they work for) who use our services and to trust the activities of independent data archives.

Data Access

My second theme is around data access and the conditions under which data can be accessed. In the last couple of months I've been working on designing a Data Access Policy for CESSDA. A few years back I constructed a data access policy for the UK Data Service [8], and earlier this year, I had the pleasure of representing the UK (or the rest of the world) on a working group which reported to the National Science Foundation in the USA.

We as representatives of the data service community need to be more involved in the policy side of data access conditions. Our collective experience in the implementation of the wishes of data owners gives us a brilliant opportunity to take more of a role in the development of data policies. We shouldn't just be implementing, we should be providing an advocacy role. In the past academic users, and, to a lesser extent, data creators are considered to be the most relevant people to be involved. Obviously all those with a vested interest in allowing making data accessible should be involved in policy development, but giving a whip-hand to any one group should be avoided.

I believe strongly that data produced as part of publicly-funded research and services should be re-usable by as wide an audience as possible. Obviously there need to be controls for some data, but access conditions should be applied only taking into account the sensitivity of the data, disclosure risk, intellectual property, and a reasonable first use period. Access conditions should not be driven by the availability of one secure mechanism or another. Access conditions should be set on the basis of the content of the data, and then the appropriate mechanism for access needs to be implemented (freely downloadable, usable through a secure mechanism, etc.). Data archives should help data controllers select the right access conditions for any particular data collection. It seems to me that sometimes data services are too nice, we sometimes allow data owners to call the shots in setting access conditions. We need to make the case for the widest possible data reuse subject only to the conditions I mention above.

Furthermore, data services should be lobbying data owners to be more consistent in their own practices. There are some hugely over the top mechanisms for access to sensitive data in place at the moment. Why can I freely download the UK's contribution to the cross-EU Labour Force Study from the UK Data Service after registration, but have to jump through some significant hoops if I want to access the same data via Eurostat? Why is there this difference in approach?

Data access conditions need to move together rather than move apart, and we need to advocate interoperability (not harmonisation), and it is the data archives which should be able to manage this. Over the last couple of years the UK Data Service has been intimately involved in the UK's Office for National Statistics review of the Approved Researcher Status [9], and I believe it is our knowledge and our reputation in this area which has allowed us to participate so intimately. Data services' involvement in this process should be a brilliant opportunity, but getting our foot in the door is sometimes a challenge.

Data Skills

My third theme is around data skills – and in fact the lack of them – within our broader community. I was involved in an activity earlier in the year where someone who described themselves as a data scientist produced a series of tables based on a survey. These were straightforward cross-tabulations, but they omitted the number of responses. A 'schoolboy' error. When our governments and funders talk about evidence-based policy the evidence has to be accurate. This is not someone misunderstanding correlation for causation, or misinterpreting their independent t-test. In the last year, I've also read some really appalling prose trying to explain some simple data analysis, and this was written by data service professionals! The data revolution should not simply be about new ways of interrogating new forms of data, but it should be about bringing writing about data into our domain as well. If data archives manage data, the staff within them should be more aware of the content. We can't provide the full service to our users unless we are more confident in understanding, manipulating and analysing the data which we hold on behalf of others; and we have to be better at interpreting and analysing the data which we produce ourselves. Typically, I think data services are quite good at this, but we should be better.

This is perhaps a minor skills gap next to the one which we face with big data. There are no fewer than four areas which need consideration. The first surrounds the intersection between traditional statistics and the new forms of data analytics. We can typify one as a branch of mathematics, and the other as a branch of computer science. This rather simplifies the distinction, but both need to understand the provenance and design of the data in order to apply it to 'real-world' situations. This has been part of statistics courses for decades, but is only starting to permeate into computer science. How can we make more data manipulators data-curious?

The second major skills gap relates to the data management and data handling. We are at the beginning of a major reconceptualization in the curation and the 'processing' of these types of data source for reuse. There has been little work in the academic sector about maximising the reuse value of these types of data. The Digital Preservation Coalition in the UK, with the help of the UK Data Service has just published two comprehensive 'white papers' on the preservation and curation of social media and

transactional data;[1] the UK Data Service's Big Data Network Support team are grappling with training resources for researchers around the use of big data.

The third and fourth skills gaps surround the ethical and legal issues around these data. Ethicists are only just becoming aware of the complex issues, some of which are potentially unresolvable within existing privacy paradigms. Many of the risks surrounding the use of these data can be mitigated by training researchers in best practices in handling and using these types of data. There are ethical and 'secure' training courses in the handing and analysis of personal data, and in the UK these are increasingly being harmonised across multiple Research Data Centers , e.g., the Office for National Statistics' VML, Her Majesty's Revenue and Customs' Data Lab, the Administrative Data Research Network and the UK Data Service's Secure Lab. Further work is required on re-examining ethical and legal frameworks for the reuse of these types of data which are based on personal information

The last area here, in my mind, relates to context. Openness and transparency in the construction of big data needs to be driven by government initiatives. It is no use in simply publishing vast quantities of open data, or providing access to it, unless there is sufficient background information for researchers to be able to use them properly.

More on big data

The fourth area which I think is something which we should ignore at our peril. I've already alluded to big data. I'm not very keen on the term big data, and I tend to think of it as data which have not traditionally been used to produce research or insight, and data science, as a corollary, is a conflation the new methods of data manipulation and analytics which can be used on those data. Some of these 'new methods' are actually old methods warmed up, and some of them are truly new.

The particular forms of 'big data' which I am particularly interested in are those which contain information about people, and people who have not explicitly given their consent for the reuse of these data for any particular purposes (informed consent). We all know of examples of where 'big data' can have a potential for societal benefit: hospital admission records could help private care providers be more efficient; search activities could potentially assist law and order agencies in pre-empting civil unrest; tracking data from mobile devices may allow commercial organisations to target advertising more effectively; surveillance images have the potential to monitor power usage at a macro level. The possibilities are endless, and mostly positive, but each of the potential opportunities I've given above could cause considerable disquiet amongst a majority of people who are represented in the underlying data.

We know that there is significant potential for the use of these types of data in socio-economic research, but in many cases the cost/benefit ratio has not been measured or compared with other forms of data analysis. One should not assume that these types of data are always cheaper to draw conclusions from.

A recent consultation carried out by the UK's Cabinet Office, however, implied just that. Part of the text of the consultation said:

> "it is essential that official statistics and research draws on the wealth of data held by businesses and other bodies outside of the public sector…….the proposal is that a new power be created to broaden the scope of data that can be requested and allow more modern methods of data collection…..away from outmoded, burdensome and expensive surveys" [10]

No mention of the lack of documentation for administrative data; no mention of the lack of consent, no mention of the data quality, or representativeness of the data; no thought to the significant loss in re-use potential. A national statistical organisation may be able to estimate the gender ratio at a low geographical level on the basis of mobile phone records, but they'll almost certainly never be able to construct usable microdata with the richness of a traditional survey.

Data services need to advocate quality in data and in research. The same paradigm holds for big data. And we do need to keep in mind that part of the rationale for the invention of sample surveys was to be able to ask more from fewer. If big data offers us less from more, is it worth it? If it offers us more for less, or something entirely new (for example, data from smart energy meters) then it may be worth it.

There are some significant risks however, which we as data service providers can anticipate and mitigate in this area. These risks will depend on the type of data, the context in which it was created, the owner of the data, the method of 'analysis', and the purpose to which the analysis is put. If personal data from a government business information system is used by respected researchers, whose research is approved by data owners for public good, who understand the complexities of the data, work in a secure environment, have their outputs vetted, and are aware of the penalties under the law, the risks of anything "bad" happening is virtually zero. But, it seems from some of the public responses to the recent UK Data Sharing legislation that not everyone is quite so optimistic about the likelihood of something bad happening.

But there are some very specific risks surrounding the quality of these data. There is the possibility decisions may be based on insights from attractive new forms of big data, without the necessary work to understand and calibrate the extent to which it provides valid alternatives to more traditional forms of data collection. In the main, traditional census and survey type data sources involve considerable effort in the design of questions, sampling frames and definitions, allowing users to understand and quantify issues such as representativeness and bias. Most big data sources do not result from any explicit design considerations (for research) but are reflections of what has already been measured. Consequently they will often over- and under-represent different groups within the population in different and complex ways. For example, the very young and the very elderly do not and will not engage with the digital economy in the same way as young adults - they therefore leave very incomplete traces in the big datasets. All 'big data' has the possibility of suffering from some sort of bias which has the potential to affect results. Of course these can be corrected, but we need to keep in mind that they may indeed need to be corrected. These are both challenges and opportunities for data analysts and data archivists.

Big data may offer exciting insights on topics which may previously have been hard to measure, but also they also provide enormous opportunities to misrepresent and misunderstand important social and economic questions. Nowhere near enough methodological work has yet been done to bridge this gap. The nature of some

of these data and the allied techniques used to analyse them are such that validation and research integrity may be extraordinarily hard to perform. Two statisticians may be able to take the same raw survey and end up with the same result; two data scientists doing the same with big data will almost certainly not. The golden thread of integrity from data to policy (evidence-based policy) has the potential to be broken. Again there has not been enough work done to bridge the gap between the two.

A further challenge is that 'traditional' sources of socio-economic data become less reliable and/or less accessible. The consultation for the 2021 UK census proposes the exploration of administrative data (really no more than another term for big data) to supplement the 2021 census. The exploration of this topic is warmly welcomed, but these sources should not be seen as the magic bullet to reduce costs. Obviously data collection costs might fall, but data analysis costs may rise, and the richness of available data may become diluted. Perhaps less important, but not a negligible problem will be access by a third party. The UK has one of the longest running and most successful data archives which make data accessible to researchers. The replacement of traditional surveys with less robust big data may, for various rights-related reasons, reduce the opportunity for reuse. We note that government has been keen on the idea of increasing efficiency by sharing and linking administrative data but we are still at a very early stage in effective research access to and linkage of these extraordinary well-regulated forms of data; any policy developments towards the further use of big data needs to be cognizant of the fact that effective access and use of existing administrative data and other types of 'big data' is still far from being achieved. So, again, here are both opportunities and challenges for data service professionals.

Data infrastructure

At the UK Data Archive, we are currently building an Open Data Platform for the Social Sciences (the Open refers to the fact that the software is open source, not that the data is 'open') on a hybrid model, which provides the information security required for sensitive data, the computing power required for 'big' data analytics, and hopefully the memory and the storage capacity too. But, to ensure that those users who are able (as well as capable) to undertake analysis within the context of their 'personal' computer, the option of downloading (for convenience (as well as information security)) should be possible. End users' wishes still need to be taken into consideration.

Scalable services and infrastructure needs to exist to maximise the benefits provided by big data opportunities. Infrastructure and services for data, especially data which are personal are not simply capital costs. Infrastructure and services for these data include hardware and software, but also the skills and human resources to make them function. Shared infrastructure is likely to be more efficient. In the social sciences the ratio of pure capital investment and operational costs may be in the region of 1 to 10; in the hard sciences this ratio may be in the region of 10 (or more) to 1. This imbalance needs to be taken into consideration in investment. Data services will need to demonstrate that investment in this area has (or will have) a return. The challenge is to get the funding to build these data platforms, make them work, and maintain them sustainably.

CESSDA

I want to start winding down making some remarks about CESSDA. CESSDA may not be the hosts of this conference, but the main office of CESSDA is located here in Bergen. Most of the European representatives here will be aware of the transformation of CESSDA from a council into a consortium. Some of the bureaucratic detail and formalities are taking a long time to finalise but there is now real progress.

For those who don't know the old CESSDA was like a club. Members learned from each other, and in dozens of research projects worked together. The new CESSDA aims to take this collaboration one stage further, and to develop an open, extensive and evolvable data service system, which all of the data archive members can benefit from single pieces of development work. I've mentioned the CESSDA Data Access policy which is an example of this, a single policy which can be used by each of the archives as aspirational or immediately implementable, and with no single archive investing directly into its construction. More technical developments are taking place at the moment, and CESSDA will be developing tools and services for data discovery, data integration and data curation in the next couple of years. CESSDA's ambition is to organise and coordinate activities across its membership, but still act as a network.

Put simply CESSDA is here to support national research and enhance European research, and to do this with the support of national ministries and funders. CESSDA needs to look at some of the challenges and opportunities which have arisen because of the so-called data revolution, and apply some of the responses which national data services have made at a European level. And, most importantly from a funding point of view, CESSDA should also provide the means for spreading the cost.

Conclusions

National data archives do need to respond to the significant changes which are taking place at the moment.

First, of course, we need to argue for funding for data services to be provided on a more sustainable basis. In order to maximise the reuse potential of any type of data, they need to be managed effectively all through the whole of the data life-cycle, and if this management can be centralised, harmonised controls (covering consent, privacy rights, the maintenance of ownership, (rights) and research ethics/integrity) can be applied consistently, and independently from either the data controller or the data users. Second, data service infrastructure also needs to have the authority to ensure that data and linked data can be curated, and that long term access can be provided, and managed together rather than in silos.

Third, organisations which host big data access facilities need to unequivocally set themselves apart from organisations which commodify personal data without proper ethical safeguards, and maintain levels of independence which the public or customers (as data subjects) have confidence in.

Fourth, and related to this, we need to help – and understand more about – public understanding of the opportunities which data from the data revolution can bring – which are of benefit to the whole of society and not just an individual's YouTube watching pleasure.

Finally, at the intersection of 'research' and 'ethical behaviour' there is a growing conundrum: a key feature of data analytics in the business sector is to uncover hidden patterns, or things which are

unknown; in the social science (and medical) domains 'informed consent' is to provide sufficient information on all aspects of research which may be carried out (or use to which the data may be put.) In the big data world, real informed consent may become a contradiction. Baroness O'Neill has argued eloquently that "genuine consent for the reuse of highly complex data for highly complex purposes is unworkable."[2] Big Data allows us the opportunity to reimagine the complex interplay of the ethical reuse of data for all data, and may also allow for a real data revolution.

The data revolution will only be mature when lots of the data created as part of that revolution is able to be reused by researchers, and this will only happen when we've met some of the challenges which I've outlined in this talk.

Reference websites

[1] UK Data Archive: *http://www.data-archive.ac.uk*

[2] UK Data Service: *http://ukdataservice.ac.uk*

[3] UK Data Service Secure Lab: *https://www.ukdataservice.ac.uk/get-data/how-to-access/accesssecurelab*

[4] Data Seal of Approval: *http://datasealofapproval.org/en/*

[5] Administrative Data Research Network: *https://adrn.ac.uk*

[6] Ipsos Mori report: *https://www.ipsos-mori.com/DownloadPublication/1652_sri-dialogue-on-data-2014.pdf*

[7] *http://www.esrc.ac.uk/files/public-engagement/public-dialogues/public-dialogues-on-the-re-use-of-private-sector-data-for-social-research-report/*

[8] UK Data Service Data Access Policy *http://staging.ukdataservice.ac.uk/media/455247/dataaccesspolicypublic_2_00.pdf*

[9] Office for National Statistics (12 July 2016) , Updated Response to the ONS Consultation on the Approved Researcher scheme: *http://www.ons.gov.uk/file?uri=/aboutus/whatwedo/statistics/consultationsandsurveys/allconsultationsandsurveys/approvedresearcherconsultation/updatedapprovedresearcherconsultationresponse.pdf*.

[10] The full text of the original consultation document (the source of the quote) is available at: *http://www.gov.uk/government/uploads/system/uploads/attachment_data/file/503905/29-02-16_Data_Legislation_Proposals_-_Con_Doc_-_final__3_.pdf*. The government response to the consultation is available at: *http://www.gov.uk/government/consultations/better-use-of-data-in-government https://www.gov.uk/government/consultations/better-use-of-data-in-government*.

Notes

1. Matthew Woollard is Director of the UK Data Archive at the University of Essex. He is also Director of the UK Data Service, the national social science data service for the UK. matthew@*essex.ac.uk*.
2. Sara Day Thomson, Preserving Transactional Data, DPC Technology Watch Report 16-02 May 2016. Available at: *http://dx.doi.org/10.7207/twr16-02* and Sara Day Thomson, Preserving Social Media, DPC Technology Watch Report 16-01 February 2016. Available at *http://dx.doi.org/10.7207/twr16-01*
3. This quote is from an unpublished note by Baroness O'Neill circulated to members of an OECD Expert Group on Big Data and Ethics. Her paper (2pp) is entitled: Notes on the Draft Data Protection Regulation and on alternative ways of securing workable, ethically robust protection for personal data in biomedical and research contexts.