Preserving Transactional Data: Defining the Challenges

Sara Day Thomson

Digital Preservation Coalition

Abstract

This paper is an adaptation of a longer report commissioned by the UK Data Service. The longer report contributes to on-going support for the Big Data Network – a programme funded by the Economic and Social Research Council (ESRC). The longer report can be found at doi:10.7207/twr16-02.

This paper discusses requirements for preserving transactional data and the accompanying challenges facing the companies and institutions who aim to re-use these data for analysis or research. It presents a range of use cases – examples of transactional data – in order to describe the characteristics and difficulties of these 'big' data for long-term access. Based on the overarching trends discerned in these use cases, the paper will define the challenges facing the preservation of these data early in the curation lifecycle. It will point to potential solutions within current legal and ethical frameworks, but will focus on positioning the problem of re-using these data from a preservation perspective.

In some contexts, these data could be fiscal in nature, deriving from business 'transactions'. This paper, however, considers transactional data more broadly, addressing any data generated through interactions with a database system. Administrative data, for instance, is one important form of transactional data collected primarily for operational purposes, not for research. Examples of administrative data include information collected by government departments and other organisations when delivering a service (e.g. tax, health, or education) and can entail significant legal and ethical challenges for re-use. Transactional data, whether created by interactions between government database systems and citizens or by automatic sensors or machines, hold potential for future developments in academic research and consumer analytics. Re-use of reliable transactional data in research has the power to improve services and investments by organisations in many different sectors. Ultimately, however, these data will only lead to new discoveries and insights if they are effectively curated and preserved to ensure appropriate reproducibility. This paper explores challenges to this undertaking and approaches to ensuring long-term access.

Introduction

This paper derives from one of two studies commissioned by the UK Data Service (UKDS) and carried out by the Digital Preservation Coalition (DPC). A companion study looks into the long-term preservation concerns around social media data,[1] while this study addresses other forms of transactional data – any data generated from individual interactions with a database – that have value for academic or commercial research. This type of data often falls under the umbrella term 'Big Data', though this paper uses the term *transactional* to bring attention to the technologies and circumstances that create these data. The UKDS Data Impact blog features a useful definition of big data from the perspective of research support:

> 'Big data are larger or more complex than traditional datasets, so traditional processing applications may simply not be able to manage them. The sheer amount and diversity of information available make big datasets physically different to the typical data information that researchers are accustomed to handling' (Moody, 2015).

This study responds to the growing interest in exploiting these types of data generated by routine capture – for instance through government services, loyalty card points, or energy meters. Re-use of these data in academic research or commercial analysis reveal insights into previously invisible patterns and trends through computational processing. However, in order to process these data reliably, researchers and their supporting organisations will need to find new methods for curation and preservation.

In both the non-commercial and commercial sectors, the ability to process and analyse transactional data requires planning and adherence to best practice. In *Best Practice Guidelines: Big Data,* the Association for Data-driven Marketing and Advertising (ADMA) emphasises that "Big Data is less about size and more about quality" and that "these data sources may be unrelated, disconnected or un-matchable in their raw form" (2013). Though these data appear ubiquitous, they cannot be usefully exploited for further study without additional action. This paper will focus on articulating the challenges to managing these data for re-use from the perspective of long-term preservation. More detailed strategies for long-term preservation will be presented in the longer report but this paper focuses on defining common challenges.

Approach

Many organisations, from a range of sectors, have begun to develop programmes to perform analysis on transactional data collected, initially, for purposes other than research. In particular, the services and infrastructure underway at the ESRC-funded Big Data Network (BDN) and the Administrative Data Research Network (ADRN) illustrate the potential of how different forms of transactional data can be re-used in research. The new Big Data Network comprises three research centres across the UK: the Urban Big Data Centre (UBDC), the ESRC Business and Local Government Data Research

---

[1] For more information about the social media report, *Preserving Social Media* DPC Technology Watch report at: doi:10.7207/twr16-01

Centre, and the Consumer Data Research Centre (CDRC), and are supported by the UK Data Service.[2] These centres facilitate research based on forms of big data, such as urban data, local government data, and consumer data. They will deliver tools and services for access, training courses in new skills and methods, and public engagement to make wider use of new research. The Administrative Data Research Network, coordinated by the Administrative Data Service (ADS), "helps researchers gain access to de-identified administrative data so they can carry out social and economic research – research that has the potential to benefit society."[3] The ADRN has a particular remit to support the linkage (or merging) of data from different sources (such as health data with education data) that may hold the potential risk of disclosing individual identities. Both of these research networks exemplify the types of research and analysis that can be achieved through robust management of transactional data. This paper will look at the types of data these networks are built to manage in dedicated use cases. These examples will help to illustrate the characteristics of transactional data and the challenges to effective management and preservation.

The use cases presented aim to show the complex environment around the capture and management of transactional data. Often, legal and ethical concerns preclude the active preservation of these data. In particular, use of these data are often subject to the UK Data Protection Act (1998) and other ethical questions around the wider impact of archiving personal data without the express consent of the data subjects (or individuals represented in the data). In some cases, these data are held in large database systems that are still in use, which presents challenges of scale and completeness. Legal, ethical, and technical obstacles continue to evolve as institutions increase their capacity for capturing and processing these data, resulting in a number of potential solutions for mitigating these challenges. Centralising data discovery and harmonising practices for data capture, for instance, holds promising possibilities for streamlining the process of curation. On a local level, as institutions undertake more research projects that deal with transactional data, they will become better-equipped to establish and provide guidance for best practice. Furthermore, centralised efforts to archive and preserve these data can help lead to uniform standards for documentation and metadata that facilitate better access and security. While there are myriad impediments, there is great value in ensuring long-term access to transactional data. Reproducibility presents a crucial benefit, but also access to historical data and the capability of conducting longitudinal studies (ADT, 2012). Finding strategies for preserving these data is a shared challenge that will best be approached through cooperation and cross-discipline collaboration.

Use Cases:
Making Progress with Transactional Data

A report on the long-term preservation of transactional data may seem pre-emptive, as many forms of these data, in particular the types presented in this paper – from the Big Data Network and the Administrative Data Research Network – currently still face considerable obstacles to capture and sharing. A main function of the both the ADRN and the BDN research centres is to help negotiate the use of third party data, but the centres often do not have ownership or even possession of those data. The institutions

---

2   UK Data Service, Big Data Network Support – Purpose: https://www.ukdataservice.ac.uk/about-us/our-rd/big-data-network-support/purpose

3   Administrative Data Research Network: http://adrn.ac.uk

who own the source data used by the ADRN and the BDN research centres may or may not have an obligation to preserve these data. In some cases, these institutions will delete data fairly frequently in order to reduce institutional risk or to reduce storage costs. However, as new uses for these data and processes for managing them emerge, research institutions and data centres have the opportunity to preserve them to a standard required for high quality, reproducible research.

In order to develop effective preservation planning to support reproducibility, it is important to understand the characteristics of these data. In the following sections, three different examples of transactional data used or made accessible though the BDN and the ADRN are shown in order to illustrate some of the challenges facing long-term preservation. In this context, long-term preservation starts early in the lifecycle of these data. The view is taken in this paper that consideration for long-term preservation should occur at selection and capture (or acquisition), in order to plan for transition to archival storage and access over time. Though other conditions may prevent the preservation of these data at the moment, data curation benefits from long-term planning. The following case studies present data at different stages in their lifecycles, but all demonstrate long-term value.

Output Area Classification Data at the Consumer Data Research Centre (CDRC)

Background

The CDRC is part of the ESRC-funded Big Data Network and is based at University College London, University of Liverpool, and University of Leeds. The Centre provides "a national service to support a wide range of users to carry out research projects that provide fresh perspectives on the dynamics of everyday life, problems of economic well-being and social interactions in cities."[4] The CDRC acts as a liaison between consumer-oriented organisations and trusted researchers in order to promote innovation in the use of data. They provide a trusted infrastructure for accessing secure data through a process of registering and training researchers and through a state-of-the-art secure lab for data that requires controlled access. Though the CDRC do not hold these data themselves, with very few exceptions, they provide an online metadata catalogue and support for researchers looking to access data from third party sources, including retail and other service organisations.

Example of transactional data
CDRC 2011 OAC (Output Area Classification) Geodata Pack.

URL
https://data.cdrc.ac.uk/dataset/cdrc-2011-oac-geodata-pack-uk

Description
These datasets were created by the CDRC data analysis team in collaboration with the Office for National Statistics using 2011 Census data. The area classifications create 'clusters' of geographic areas using select characteristics of the population based in that cluster.[5] They are popularly used to create data visualisations using maps. In the next

---

4 Consumer Data Research Network: https://www.cdrc.ac.uk/research/research
5 Office for National Statistics, 2011 Area Classifications: http://www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/ns-area-classifications/ns-2011-area-classifications/index.html

year, CDRC hope to launch a free service on their website that allows researchers to create their own map-based data visualisations using data similar to these area classification datasets ('georeference' data). While these data come from open sources, the added value provided by the CDRC demonstrate how these data can be used in research and analysis. These derivative datasets, therefore, require careful preservation planning, including persistent identifiers to support citation, so that other researchers and organisations can build on the work done by CDRC data analysts. The UKDS provides data management support for all of the Big Data Network research centres in order to ensure the process of data management and preservation planning happen uniformly across the centres.

Energy Demand Research Project Managed by the UK Data Service

Background

The UKDS[6], based at the UK Data Archive with partner universities across the UK, delivers services and support for researchers, policy-makers, and data producers. Funded by the ESRC, the UKDS aims to improve the use of social and economic data for research and analysis. Part of this work includes the Big Data Network support project,[7] designed to coordinate data management at the three research centres as well as maximise the services and tools they develop. The UKDS provides guidance on using and archiving novel forms of data made accessible through the research centres for both researchers and data producers. Part of this guidance includes centralising discovery and over-seeing processes for long-term preservation.

Example of Transactional Data

Energy Demand Research Project: Early Smart Meter trials (2007-2010) (EDRP).[8]

PID

10.5255/UKDA-SN-7591-1

Description

The EDRP data derives from a set of trials carried out between 2007 and 2010 to monitor how households respond to knowledge about their energy use (UKDA, 2014b). The trials looked at energy data, including readings from household smart meters, provided by four different energy suppliers: EDF Energy, E.ON UK, Scottish Power Energy Retail and SSE Energy Supply. Significant measures were taken during the transfer of data from the energy suppliers to the Centre for Sustainable Energy (CSE), who compiled the data, to ensure reliability and privacy for the participating households. CSE received raw data but had no part or knowledge of the collection process and ensured anonymisation of the portions of the data sent to third parties for research. The data available through UKDS includes three datasets, subsets of the collected data: 1) Electricity smart meter half-hourly reads, 2) Gas smart meter half-hourly reads, and 3) Geography and Segmentation data. A metadata file is also available to describe the variables used in the datasets. The electricity dataset consists of 413,836,038 cases and is 12GB in size, the gas file consists of 246,482,700 cases and is

---

6 UK Data Service: https://www.ukdataservice.ac.uk

7 UK Data Service, Big Data Network Support: https://www.ukdataservice.ac.uk/about-us/our-rd/big-data-network-support

8 UK Data Service, EDRP Accompanying Documentation:
   http://doc.ukdataservice.ac.uk/doc/7591/mrdoc/pdf/7591_edrp_accompanying_documentation.pdf

9GB in size. Because of the large size of these datasets, they are provided in CVS format only. The catalogue entry provides advice to users on recommended methods to download and access the file.

The EDRP data were collected from energy suppliers by CSE and deposited to UKDS by the Department of Energy and Change. Under the care of UKDS, the data have been assigned a PID and provided with a reliable citation. The datasets are protected by the UK Data Archive's preservation policy, which deploys robust processes and technology to maintain digital content for long periods of time (2014a). This dataset provides a useful model for the effective management and preservation of transactional data collected by a third party through ensuring quality of data, adhering to data protection laws, and providing documentation and discovery to facilitate further research and analysis.

Higher Education Data at the Administrative Data Research Network

Background

The ADRN, as described in the introduction, is an ESRC-funded network of centres across the UK designed to facilitate researcher access to linked (or merged) administrative datasets. 'Administrative' is not an established legal or technical category but refers to the data collected routinely by government departments, such as health data or education data. The ADRN is comprised of four centres, one each in England, Scotland, Northern Ireland and Wales. The centres do not hold data themselves, but negotiate with government departments on behalf of trusted researchers who request access. The Administrative Data Research Centres (ADRCs) help to improve access to administrative data and linked administrative data, traditionally hindered by the legality of re-using these data in research and for policy-making.[9] In addition to acting as a liaison between researchers and government data sources, the ADRN provides a central metadata catalogue of administrative data held by different UK government departments. The research support and services made possible through the ADRN promote the re-use of these data to improve public well-being. Work achieved through the ADRN will hopefully influence a culture of caution currently inhibiting the sharing of valuable data (Laurie and Stevens, 2016).

Example of Transactional Data
Student Record, 1994/95- (not held by ADRN, but metadata available in catalogue).

URL
http://adrn.ac.uk/catalogue/cataloguepage?sn=888013

SN
888013

Description
The Higher Education Statistics Agency (HESA) have been collecting detailed information about students entering any programme of higher education since 1994. Held by HESA, these datasets include student home addresses, dates of birth, ethnicities, previous qualifications, and main sources of funding, though the data

---

9   Administrative Data Research Network – Administrative Data – Getting Data for Research:
    https://adrn.ac.uk/getting-data/admin-data

variables collected have changed over time. Each year, the dataset contains more than 2.25 million records. Though there are limitations to how this data can be used and linked, it can be merged with the Destination of Leavers survey and can be acquired alongside data from the National Pupil Database, HESA Student Records, and Individualised Learner Records. This dataset cannot be merged with any external data sources, however, if permission is negotiated, researchers may be able to apply probabilistic matching techniques to merge with a few designated datasets.

Because the ADRN do not hold this data, they do not have any archival or preservation accountability. However, through their services and training for researchers, they can encourage uniform processes to access these data and support reliable data management during the research process.

Summary

These examples demonstrate forms of transactional data – information generated through the interaction of individuals with third-party organisations – that have been extracted from their original environments and re-used for research and analysis. These data derive from a range of capture and storage technologies, from large government databases to electronic meters. In all three cases, the data has been changed and re-formatted for access by researchers. Data sources, the government departments and companies who collect these data, also face challenges to long-term preservation. These institutions also need guidance and benchmarks for best practice to cope with their growing data holdings. While some solutions do exist, such as standards for database preservation (e.g. SIARD and CHRONOS) and data warehousing, this paper focuses on the challenges to data management and preservation needed to support the re-use of these data for research.[10] In particular, it focuses on the legal, ethical, organisational, and technical challenges to preserving these data once they enter the research process.

On their own, no one example from the cases above necessarily meets the general definition of 'Big Data' (such as the one quoted into the introduction). They do represent new uses for transactional data, however, because of the technologies and circumstances surrounding their capture, format, and use in research. As researchers develop new computational approaches to performing research and data analysis, these forms of transactional data serve a new function. They can be adapted and processed by computational analytics to reveal new insights, often in conjunction with more traditional methods of social science research and analysis. They can be merged with other data sources and processed to the specifications of particular research questions. The increasing availability of routinely captured data provide new opportunities for these new approaches to research and analysis. As a result, data managers and archivists face new challenges to curation and long-term preservation in order to maximise and build on these opportunities moving forward.

From Data Source to Research Citation

Transactional data, as they are collected by government and other organisations, are not immediately ready for re-use. Before these data are useful for researchers or data analysts, they must meet a number of requirements. To begin with, researchers and data centres must negotiate the legal and ethical conditions attached to the original data.

---

10 For more information about SIARD and CHRONOS, see Lindley (2013).

Issues of ownership and intellectual property may pose issues, but more often, transactional data useful for social or economic research contain personal data and must comply with the Data Protection Act (1998) and ethical standards. In many cases, the legal and ethical issues can be resolved, but organisational mechanisms or institutional culture may prevent the use of these data. The adaptation of these data for research often also pose technical issues, such as incomplete data or datasets too large for most archival repositories to handle. This section defines the challenges to long-term preservation of transactional data from the early stages of curation.

The use cases presented in this paper represent data collected by third parties and subsequently made available for re-use. Using data for a purpose other than the one for which it was originally collected creates a number of legal and ethical concerns. If the original data contain personal information about individuals or households, sharing that data could be prevented by the Data Protection Act. To ensure compliance, data must first be de-identified or other actions must be taken to prevent accidental disclosure of individuals, such as using a trusted third party to replace personal identifiers (ADT, 2012). Once data controllers, or data owners, have assessed that data can legally be shared, it should be assessed whether long-term preservation creates any further risk of disclosure. Digital preservation itself is often an exercise in risk management – issues of preserving personal data are not new to curators and information managers. Risk assessment at the stage of sharing or linking data should also entail further assessment for any additional risk posed by preservation.[11] Additionally, at this early stage, it is crucial to gain necessary permission to preserve data and derivative datasets for the necessary amount of time.

Recently, new EU regulations were approved (December 2015) that will come into effect in 2018 to replace the current Data Protection Directive 95/46/EC and the UK Data Protection Act.[12] The new European General Data Protection Regulation ('GDPR') includes some changes that potentially benefit the preservation of transactional data, and particularly administrative data. Two separate articles contain exceptions that allow for the preservation of these data when found to be in the public interest (Stevens, 2015). It is uncertain how UK adoption of these new regulations will directly impact the preservation of transactional data, but the outlook is positive.

Beyond the legal questions around transactional data, ethical concerns arise over the re-use of personal data, even de-identified personal data, when the data subjects may not be aware. In the context of academic research, many university ethics committees may impose more stringent requisites for consent than the law requires. These requirements may limit new research using transactional data. Similarly, many organisations who own such data err on the side of caution when it comes to making decisions about sharing data. In their research and surveys, the Administrative Data Taskforce found that "the value of using administrative data for analytical purposes inside and outside government is well understood" (2012). Unfortunately, the complexity of legal and ethical issues prevent data owners from sharing data, even legally. Laurie and Stevens quote from the 'Data Sharing Review Report' to illustrate the situation at most organisations considering whether or not to share data:

---

[11] For more information about digital preservation and risk management, see the DRAMBORA audit framework created jointly by the Digital Curation Centre and Digital Preservation Europe at http://www.dcc.ac.uk/resources/repository-audit-and-assessment/drambora or the SPOT Model for Risk Assessment created by Sally Vermaaten, Brian Lavoie, and Priscilla Caplan in 'Identifying Threats to Successful Digital Preservation', doi:10.1045/september2012-vermaaten

[12] Information about the decision on 15 December 2015 can be found on the European Commission's website: http://ec.europa.eu/justice/data-protection/reform/index_en.htm. The proposal itself can be found at: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52012PC0010

'Despite the current availability of lawful means to link or share identifiable personal data or de-identified data for research in the public interest, "…in the vast majority of cases…the complexity of the law, amplified by a plethora of guidance, leaves those who may wish to share data in a fog of confusion"'[13] (2014).

This confusion belies a predominant 'culture of caution' at the organisations in a position to share valuable data. The best remedy for this 'fog of confusion' is education. The ADRN and the BDN, positioned as a liaison between data sources and researchers, could provide information and guidance about the legislation that regulates the sharing of transactional data for non-commercial research in particular. Furthermore, in their role as intermediary, these networks are in a position to advise on the necessity of preserving these data for long-term access when appropriate.

The challenges facing the re-use of these data do not end after the legal and ethical issues have been resolved. Often, the size and fragmented nature of many of these data cause further problems for ingest into data repositories or onto researchers' machines. As Moody highlighted in her definition of big data, these datasets are often simply larger and more complex than the datasets researchers or data managers are accustomed to handling (2015). This problem of scale means that repositories face a growing issue of storage capacity as well as processing power. The issue of broken or incomplete data also pose difficulties. In a recent study, the GESIS Leibniz Institute for the Social Sciences in Germany linked environmental noise data with spatial data in order to assess how this type of data linking could support social scientists. They found that:

'… some states publish maps of existing health infrastructures, whereas in other states these data are published at the municipality level. Consequently, in Germany one can see that there is a huge amount of spatial data that are publicly available for free, however, these data are often fragmented and therefore incomplete' (Schweers, Kinder-Kurlanda, Müller, and Siegers, 2016).

This scenario could easily occur in the UK as well, where similar records are collected by different government departments within different jurisdictions (PHRDF, 2015). These data are not collected with the intention of merging them with other data sources and therefore may be incomplete, fragmented, or in incomparable formats.

The types of research services provided by the ADRN and the BDN reflect the larger trend toward improving the re-use of transactional data for research. As these networks and other similar programmes continue to develop, a coordinated effort to establish processes for curation and preservation will make future research far more streamlined and supportable. Some institutions have already begun to build infrastructure to capture, process, and store these data – the best time to integrate preservation planning and long-term access is now.

---

[13] Richard Thomas and Mark Walport, 'Data Sharing Review Report', 11 July 2008, Foreword. http://webarchive.nationalarchives.gov.uk/+/http://www.justice.gov.uk/docs/data-sharing-review.pdf

Conclusion:
The Future of Transactional Data

Relatively speaking, computational research methods using transactional data, compared to traditional methods of research, are in their infancy. The speed at which these data are being created prompts a sense of urgency to capture and exploit new sources of social information. At the same time, the rapid availability of new data creates confusion among data owners and the general public about the real impact of re-using these data. The ADRN and BDN have created some useful services and infrastructure to help on both of these fronts. For example, though it may seem small, the publication of metadata catalogues, publicly discoverable, allows the research and the wider community to see the range of open, safeguarded and secure data available for re-use. In some cases, catalogue records include information about the types of studies already using particular datasets to show how they can be used. In combination with training and public engagement, these networks help to further the education of researchers and the wider public about the processes and policies surrounding the re-use of transactional data.

The establishment of the BDN and the ADRN by the ESRC demonstrate growing initiative to exploit transactional data for research to improve services and policies. These two networks provide a model for how this type of research might be facilitated and supported. In particular, the work undertaken by the networks has the potential to foster a relationship of trust between data owners and researchers. As public-facing networks, they are also in a position to build trust with the larger population when it comes to re-use of data. Institutions who traditionally preserve digital content (e.g. repositories, libraries, archives) have also faced the need to demonstrate their trustworthiness to the general public. Elaborate accreditation frameworks have been developed to help archival institutions demonstrate their ability to maintain digital content, often critical digital records, over time.[14] Ultimately, archival institutions have had to learn how to communicate effectively the principles of digital preservation to non-experts and to foster understanding with the users who stand to benefit from well-maintained collections.[15] The trustworthiness of accredited repositories in the UK, such as the UK Data Archive, could provide useful assurance to a public concerned about the security and privacy of their data. The extent of this support will depend on how substantially the ADRN, the BDN, and other institutions integrate data management and preservation into their processes.

References

Administrative Data Taskforce. (2012). *The UK administrative data research network: Improving access for research and policy.* Retrieved from http://www.esrc.ac.uk/files/publications/themed-publications/improving-access-for-research-and-policy

---

14 Two examples of these accreditation frameworks include TRAC (Trustworthy Repository Audit and Certification) developed by the US RLG and NARA and the Data Seal of Approval developed by the Dutch organisation DANS.

15 For a more detailed discussion of the issue of digital repositories and trust, see Yakel, Faniel, Kriesberg, and Yoon (2013).

Association for Data-Driven Marketing and Advertising. (2013). *Best practice guidelines: Big data*. Retrieved from http://datascienceassn.org/sites/default/files/Big%20Data%20Best%20Practice%20Guideline.pdf

Johnston, D. & Henderson-Ross, J. (2012). The new data values: Securing customer data as a renewable resource. *Aimia Insights*. Retrieved from http://www.aimia.com/content/dam/aimiawebsite/CaseStudiesWhitepapersResearch/english/WhitepaperUKDataValuesFINAL.pdf

Laurie, G. & Stevens, S. (2014). *The administrative data research centre Scotland: A scoping report on the legal and ethical issues arising from access and linkage of administrative data.* Research Paper Series, No 2014/35. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2487971

Laurie, G. & Stevens, L. (2016). Developing a Public Interest Mandate for the Governance and Use of Administrative Data in the United Kingdom. *Journal of Law and Society, 43*(3). Retrieved from http://onlinelibrary.wiley.com.ezproxy.lib.gla.ac.uk/doi/10.1111/j.1467-6478.2016.00759.x/full

Lindley, A. (2013). Database preservation evaluation report – SIARD vs. CHRONOS. Lisbon Technical University (IST), Portugal. Retrieved from http://purl.pt/24107/1/iPres2013_PDF/Database%20Preservation%20Evaluation%20Report%20-%20SIARD%20vs.%20CHRONOS.pdf

Moody, V. (2015). Thinking big: Making the most of big data in scientific research. *Data Impact Blog.* Retrieved from http://blog.ukdataservice.ac.uk/thinking-big-making-the-most-of-big-data-in-scientific-research

Organisation for Economic Co-operation and Development (OECD). (2013). *New data for understanding the human condition.* OECD Global Science Forum Report. Retrieved from http://www.oecd.org/sti/sci-tech/new-data-for-understanding-the-human-condition.pdf

Public Health Research Data Forum. (2015). Enabling data linkage to maximize the value of public health research data: Full report. Retrieved from https://wellcome.ac.uk/sites/default/files/enabling-data-linkage-to-maximise-value-of-public-health-research-data-phrdf-mar15.pdf

Schweers, S., Kinder-Kurlanda, K., Müller, S., & Siegers, P. (2016). Conceptualizing a spatial data infrastructure for the social sciences: An example from Germany. *Journal of Map & Geography Libraries*.

Stevens, L. (2015). The proposed data protection regulation and its potential impact on social sciences research in the UK. *European Data Protection Law* Review, *1:2*, 97 – 112. Retrieved from http://edpl.lexxion.eu/article/EDPL/2015/2/4

UK Data Archive. (2014a). Preservation policy. Retrieved from http://www.data-archive.ac.uk/media/54776/ukda062-dps-preservationpolicy.pdf

UK Data Archive. (2014b). *Smart meter energy demand research project: Data release.* UK Data Archive Study Number 7591 – Energy Demand Research Project: Early Smart Meter Trials, 2007-2010. Retrieved from http://doc.ukdataservice.ac.uk/doc/7591/mrdoc/pdf/7591_edrp_accompanying_documentation.pdf

Yakel, E., Faniel, I.M., Kriesberg, A., & Yoon, A. (2013). Trust in digital repositories. *International Journal of Digital Curation*, *8*(1), 143-156. doi:10.2218/ijdc.v8i1.251