Preserving Transactional Data

Sara Day Thomson

DPC Technology Watch Report 16-02 May 2016

Series editors on behalf of the DPC
Charles Beagrie Ltd.

Principal Investigator for the Series
Neil Beagrie

DigitalPreservationCoalition

This report was supported by the Economic and Social Research Council [grant number ES/JO23477/1]

Contributing Authors for Section 9 Technical Solutions: Preserving Databases

Bruno Ferreira, Miguel Ferreira, and Luís Faria, KEEP SOLUTIONS and José Carlos Ramalho, University of Minho

ISSN: 2048-7916

DOI: http://dx.doi.org/10.7207/twr16-02

First published in Great Britain in 2016.

Foreword

The Digital Preservation Coalition (DPC) is an advocate and catalyst for digital preservation, ensuring our members can deliver resilient long-term access to digital content and services. It is a not-for-profit membership organization whose primary objective is to raise awareness of the importance of the preservation of digital material and the attendant strategic, cultural and technological issues. It supports its members through knowledge exchange, capacity building, assurance, advocacy and partnership. The DPC's vision is to make our digital memory accessible tomorrow.

The *DPC Technology Watch Reports* identify, delineate, monitor and address topics that have a major bearing on ensuring our collected digital memory will be available tomorrow. They provide an advanced introduction in order to support those charged with ensuring a robust digital memory, and they are of general interest to a wide and international audience with interests in computing, information management, collections management and technology. The reports are commissioned after consultation among DPC members about shared priorities and challenges; they are commissioned from experts; and they are thoroughly scrutinized by peers before being released. The authors are asked to provide reports that are informed, current, concise and balanced; that lower the barriers to participation in digital preservation; and that are of wide utility. The reports are a distinctive and lasting contribution to the dissemination of good practice in digital preservation.

This report was written by Sara Day Thomson. The report is published by the DPC in association with Charles Beagrie Ltd. Neil Beagrie, Director of Consultancy at Charles Beagrie Ltd, was commissioned to act as principal investigator for, and managing editor of, this Series in 2011. He has been further supported by an Editorial Board drawn from DPC members and peer reviewers who comment on text prior to release: William Kilbride (Chair), Janet Delve (University of Portsmouth), Marc Fresko (Inforesight), Sarah Higgins (University of Aberystwyth), Tim Keefe (Trinity College Dublin), and Dave Thompson (Wellcome Library).

Acknowledgements

I would like to thank the UK Data Service (UKDS) and the staff at the UK Data Archive (UKDA) for their initiative in commissioning this report and for their on-going support. A special thank you to Nathan Cunningham, UKDA Associate Director and UKDS Functional Director for Big Data Network Support, and his Big Data team who helped shape this report. I want to acknowledge the contribution of practitioners from the institutions who serve as case studies in this report. Thank you to Sarah Sheppard from the Consumer Data Research Centre at University College London, to Leslie Stevens from the Administrative Data Research Centre-Scotland at the University of Edinburgh for her guidance on the legal issues around sharing government data, and to Aidan Condron for his input and perspective, and occasionally, his powers of persuasion.

Last but not least, thank you to Neil Beagrie for his work on the *Technology Watch Report* series and to my DPC family who keep me sharp, and full up on tea and cake.

Sara Day Thomson
June 2016

Contents

1. Abstract

This report tackles the requirements for preserving transactional data and the accompanying challenges facing companies and institutions that aim to re-use these data for analysis or research. Commissioned by the UK Data Service as part of their Big Data Network Support initiative, this report presents the issues and strategies which emphasize preservation practices that facilitate re-use and reproducibility. As with its companion report, *Preserving Social Media*, this publication explores the preservation concerns for novel forms of data. Transactional data – defined as any logical interaction with a database – challenge current approaches to long-term preservation. The scale and velocity of these data push current methods and tools for preserving databases to their limits. These data – from government data to environmental data – possess significant characteristics that require wider approaches to preservation. New approaches must consider an emergence of new uses for archived forms of these data. The meaning and value of these data derive not only from the raw content, but from the ways people interact with the technologies that create them. Through a range of use cases – examples of transactional data – the report describes the characteristics and difficulties of these 'big' data for long-term access. Based on overarching trends, this paper will demonstrate potential solutions for maintaining these data in a secure environment based on end user needs and regulatory frameworks.

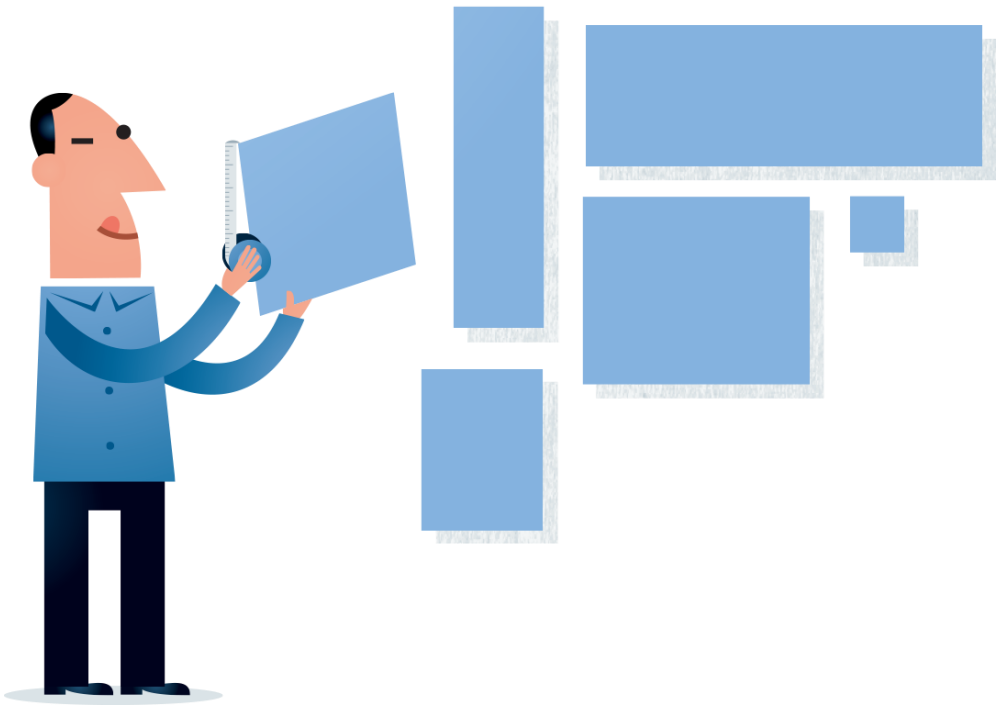

Figure 1: Where will we find a container to fit all this data?

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark.

2. Executive Summary

This report discusses requirements for preserving transactional data and the accompanying challenges facing companies and institutions that 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 report will demonstrate potential solutions for maintaining these data in a secure environment based on end user needs and regulatory frameworks.

The term 'transactional' will be used to refer to data that result from single, logical interactions with a database and the **ACID properties** (Atomicity, Consistency, Isolation, Durability) that support reliable records of interactions. In some contexts, these data could be fiscal in nature, deriving from business 'transactions' such as at an ATM or through a web service such as Amazon. This report, however, considers transactional data more broadly, addressing any data generated through interactions with a database system. Administrative data, for instance, are 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 organizations when delivering a service (e.g. tax, health, or welfare) and can entail significant ethical and legal challenges. Transactional data, whether created by interactions between government database systems and citizens or by automatic sensors or machines, hold potential for future developments in consumer analytics and in academic research. Ultimately, however, these data will only lead to new discoveries and insights if they are effectively curated and preserved to ensure appropriate reproducibility.

In some instances, the preservation requirements for transactional data held by government and other organizations will contradict the preservation needs of analysis or academic research. In particular, data collected by government institutions that contain personal information will fall under the regulations of data protection and will require strict retention schedules. Data containing personal information can often only be used for the purpose for which they were collected, which may preclude re-use in research. In order to ensure long-term access to these data in a secure environment, the maintenance of these data demands intensive database management.

Database management also involves the implementation of preservation strategies that account for the changing, fragmented nature of many database systems. The approaches outlined in this report support both the governments and organizations that generate transactional data as well as the institutions which must archive these data to a standard re-usable in academic research. These two contexts require different approaches but may utilize the same solutions. These shared solutions might include the technical steps for retiring database systems as well as standards and principles such as SIARD and CHRONOS (Lindley, 2013). In addition to relevant solutions, this report will address associated challenges, such as the problems with anonymization for performing analytics on large sets of data (OECD, 2013).

Though this report focuses mainly on preservation concerns, it examines issues of access and data creation to the extent that they help inform effective preservation. The use cases presented will also include accompanying issues of access and end user needs. While the report provides an overview of current approaches to preserving databases, it does not provide an exhaustive discussion of all technical strategies; it does not, for instance, present data warehousing. Rather, the report examines the broader framework surrounding the capture and preservation of transactional data, including the impact of new uses for archived datasets and the legal implications of re-using data captured for purposes other than research.

3. Introduction

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

> '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).

There is a 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. In order to process these data reliably, however, researchers and their supporting organizations will need to find new methods for curation and preservation. This growing interest in exploiting data – and the corresponding need to curate and preserve these data – is reflected in the requirements published by the European Commission in the *Guidelines on Data Management in Horizon 2020* (Directorate-General, 2016). These guidelines for EU-funded research require projects under the Open Research Data Pilot to make all research data and metadata freely available in an open-source repository (*ibid.*). From the commercial sector to research funding bodies, the importance of curation and preservation underlies an increasing number of initiatives to exploit big data, such as transactional data.

In both the non-commercial and commercial sectors, the ability to process and analyse transactional data requires planning and efforts for developing best practice. The Association for Data-driven Marketing and Advertising (ADMA) emphasizes 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 usefully be exploited for further study without additional action. New methods for processing and analysing these data will require careful curation to ensure data from different sources are compatible.

The organizations who collect transactional data – government departments, retail companies, other corporate organizations – may have different motivations for preserving these data. Primarily, organizations aim to manage and preserve routinely collected data for business purposes as part of their records management. Organizations often take measures to ensure the long-term preservation of data because of external laws and regulations. As Heiko Müller points out in his briefing paper on database archiving, 'compliance with government regulations on data preservation is the main driver for the majority of current data archiving efforts' (2009). Strategies for preserving this data are varied and could entail the preservation of derivative datasets, snapshots of databases at a particular time, or the retirement of entire database systems in archival formats.

This report will focus on articulating the challenges to managing these data for re-use from the perspective of long-term preservation. While some strategies for databases preservation will be presented – including tools and standards – the emphasis will be on preserving data for re-use in computational analysis either by the organizations that collect it or by external research institutions.

[1] *Preserving Social Media*: http://dx.doi.org/10.7207/twr16-01

Centralised efforts to archive and preserve these data can help lead to uniform standards for documentation and metadata that facilitate better access and security.

4. Background to the Study

Many organizations, from a range of sectors, have begun to develop programmes to perform analysis on transactional data collected, initially, for purposes other than research. In the UK, the development of services and infrastructure are underway at the ESRC-funded Big Data Network Support (BDNS), which includes the Administrative Data Research Network (ADRN) and research centres that form the Business and Local Government Data initiative.[2] These centres illustrate the potential for reusing different forms of transactional data in research. This report, resulting from a 15-month study, was commissioned to support the long-term preservation issues faced by these ESRC-funded centres.

The research centres that focus on business and government data across the UK – the Urban Big Data Centre (UBDC), the ESRC Business and Local Government Data Research Centre, and the Consumer Data Research Centre (CDRC) – are supported by the UK Data Service.[3] 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. As mentioned in the introduction, these UK initiatives mirror wider initiatives across Europe (Directorate-General, 2016). There is even a global precedent for re-using and adequately curating and preserving big data. Open Knowledge, for instance, is a network of people and organizations from across the world that advocates and supports 'open data'.[4] Open Knowledge promotes the definition of 'open' as: 'data and content [that] can be freely used, modified, and shared by anyone for any purpose'.[5] While not all transactional data can be made open, for various reasons including data protection regulations and commercial ownership, many forms of transactional data have the capacity to support research and government transparency if properly preserved and made 'open'.

In the UK, 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'.[6] 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 report 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 the UK, use of these data is 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. Centralizing data discovery and harmonizing practices for data capture, for instance, hold promising possibilities for streamlining the process of curation. On a local level, as institutions undertake

[2] About the ESRC's BDN: http://www.esrc.ac.uk/research/our-research/big-data-network
[3] UK Data Service: https://www.ukdataservice.ac.uk/about-us/our-rd/big-data-network-support/purpose
[4] Open Knowledge: https://okfn.org/about/our-impact
[5] Open Definition: http://opendefinition.org
[6] Administrative Data Research Network: http://adrn.ac.uk

more research projects that deal with transactional data, they will become better equipped to establish and provide guidance for best practice. Furthermore, centralized efforts to archive and preserve these data can help lead to uniform standards for documentation and metadata that facilitate better access and security. A technology currently growing in importance among research and enterprise institutions alike is linked data – data published on the web in machine-readable languages that can be queried using programmed applications. The Semantic Web, devised by Tim Berners-Lee and developed by W3C, aims to enhance the re-use of open data through matching common events, objects, dates, persons, and other data.[7] In other words, they aim to promote a web of interrelated datasets connected through relationships, rather than isolated datasets. The library cooperative OCLC, for instance, has begun publishing bibliographic data on the web as linked data. By doing so, OCLC are enhancing library resources through connecting information on the web back to available library holdings and services.[8] While linked data promises a means of sharing and connecting data from a broad range of resources, this report focuses on institutional-level efforts to curate and preserve transactional data. With established policies and processes in place, institutions will be in a stronger position to publish their open data on the semantic web.

While there are myriad impediments, there is great value in ensuring long-term access to transactional data; reproducibility presents one crucial benefit, but there is 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.

5. Characteristics of Transactional Data

Transactional data – created through interactions with a database – can come from a wide range of sources and represent many different types of information. This section provides a description of transactional data and the technology used to create and interact with them. The methods used to create or collect transactional data have a direct impact on how they can be preserved. Approaches to preserving these data require an understanding, first of all, of what they are and how they work.

5.1. Defining Transactional Data

The term transactional data, as used in this report, applies to a great variety of types of data. Sometimes, transactional data are not by themselves 'big data'. It is the ability of transactional data – captured in databases – to be computationally combined with other sources of data that make it part of the discussion of 'big' data. Preserving transactional data, whether large or not, is imperative for the future usability of big data, which is often comprised of many sources of transactional data. Many forms of big data are transactional – generated by logical interactions with a database. Characteristics of these data, some common and others variable, help shape an understanding of how to preserve them. A primary characteristic to consider is how they will be used in the future. A number of definitions, arising from different sectors, provide a useful overview that helps indicate potential future uses. As with other types of archival material, which definition is more relevant will depend on the institution's main functions and their requirements for preserving these data.

It is the ability of transactional data to be computationally merged with other sources of data that make them part of the discussion of 'big' data.

In the commercial sector, emphasis is often placed on the most recent, up-to-date data. The majority of commercial organizations aim to re-use data for analysis that reflects consumer behaviour or provides insight into the efficiency of day-to-day business. ADMA, for example, gives the following definition of big data:

[7] W3C: https://www.w3.org/standards/semanticweb/data
[8] OCLC: http://www.oclc.org/data.en.html

> 'The defining dimensions therefore of what we know as Big Data are: 1- Data from multiple sources, 2- Data in a wide variety of formats, 3- Data that is in constant motion, and 4- Data volume can also be a feature. The value derived from Big Data often lies at the intersection of two or more different datasets' (2013).

ADMA's *Best Practice Guideline* further provides comparative definitions of big data to reflect its exploitation in commercial environments (*ibid*.). Big data in the commercial sector describes a variety of different types of data. Commercial organizations may rely on internal data, such as from sales or marketing information, or on web data harvested from their own websites or those of others (ADMA, 2013). This data may also include social media data, mobile data, or research data (ADMA, 2013). Increasingly, large retailers depend on data from customer loyalty programmes, such as Sainsbury's Nectar points[9] or the Boots Advantage card.[10] These programmes generate rich data about customers but also require careful management of personal data (Johnston and Henderson-Ross, 2012). Organizations which use this data for consumer or marketing analysis may not need to preserve data for as long as organizations interested in change over time, for example, social and economic research that helps inform public policy and governance.

In the public sector, transactional data possess value for research but also create liabilities for those who hold and work with them. While the UK government may increasingly value open data[11], public organizations have a responsibility to ensure the security of data containing confidentiality or privacy risks. David Rhind, in a report issued by the Advisory Panel on Public Sector Information (decommissioned in October 2015), provides a definition of big data based on the 'implications for exploitation' (2014). He separates data into three categories: 1 - Open Data (and the National Information Infrastructure); 2 - personal data held by government; and 3 - data held as a commercial asset (*ibid.*). This definition reflects a concern with balancing, on one hand, potential value in re-using forms of big data with, on the other hand, the legal and financial frameworks around these data. In order to illustrate what is meant by 'big data', Rhind gives a set of characteristics such as the distinction between data ('numbers, text, symbols') and information ('implies some degree of selection, organization, and relevance to a particular use'). He also describes the economy of digital data:

> 'Even though the initial cost of collecting, quality assuring, and documenting data may be very high, the marginal cost of providing an additional digital unit is close to zero'.

Furthermore:

> 'Unlike fuel, data and information are not consumed by use. They may get out of date (like census information) but rise again in value in comparisons with new information' (*ibid*.).

This characterization emphasizes the need to think of data as an asset, but with a value measured differently from traditional material assets.

The scientific community (and academic community more widely) emphasizes the need to curate and preserve data to support citation and reproducibility in research. In their 2012 report *Science as an Open Enterprise*, the Royal Society describes the value of data as the 'bedrock on which scientific knowledge is built', and that 'its accessibility to scrutiny by others than the originators is essential to the progress of science' (quoted in edited form in OECD, 2013). The report further describes the role of data:

> 'But disclosure of data has little value in itself. It must be communicated effectively, which means that it must be accessible, so that it can be readily located. It must be intelligible to those who wish to scrutinise it. It must be assessable so that judgements can be made about its reliability and the competence of those who created it. And it must

[9] Sainsbury's Nectar: https://www.nectar.com/collecting-points/sponsors/sainsburys
[10] Boots Advantage card: http://www.boots.com/en/Advantage-Card
[11] Data.gov.uk: https://data.gov.uk/about and Open Data policy paper: https://www.gov.uk/government/publications/open-data-white-paper-unleashing-the-potential

> be usable so that it can be re-used, which requires the provision of appropriate metadata' (quoted in edited form in OECD, 2013).

This assertion stresses the difference between data and information and information and knowledge. Transactional data is valuable, but only if it can be transformed into information and knowledge. As the Royal Society asserts, in order to exploit this value, further action must be taken on the raw data. The management and curation of these must include information about the data, such as context and provenance. Detailed information about the rights and restrictions attached to this data, including any licensing agreements, must remain closely associated with them as they move through the curation lifecycle. Long-term preservation requires knowledge such as the form of the data and the software and hardware needed to re-deploy them in future. Raw data – records of interactions with a database – mean very little or nothing without sufficient metadata, management and preservation actions.

As demonstrated, different sectors have different priorities when it comes to defining big data, much of which is transactional, and thus different priorities for preserving it. However, these data (whether commercial, public sector, or scientific) are often generated by, and stored in, similar technologies. Therefore, when it comes to long-term preservation, they share a number of important attributes. The majority of these data, for instance, are held in relational databases, with a growing trend in the use of non-relational database (or NoSQL database) models for storing data. These two types of database possess their own benefits and drawbacks for organizing data, depending on the form of the data and purpose of the database. Although relational databases are still predominantly used for most types of transactional data, new forms of big data are increasingly held in non-relational databases. Both types of database require different approaches to preservation. Non-relational databases, in particular, pose significant challenges to current approaches to archiving databases.

5.2. Relational (SQL) Databases

Relational databases organize data into tables of columns and rows. Each table contains an isolated category of data, such as addresses or dietary requirements. The rows represent individual entries (records) and the columns describe the attribute captured ('address' or 'dietary requirements'). A table might contain addresses for a customer base, with rows for each customer (one row for Harry, another row for Sally, and so on) and with columns for a 'unique key' to identify each record, for example, 'customer name', and 'address'. The unique key is used to link data in separate tables. Because of how these databases arrange content, the data are sometimes referred to as 'tabular data'.

Relational databases are usually accessed, or manipulated, in the programming language SQL (Structured Query Language), which is a standard under the American National Standards Institute[12] and under the International Standards Organisation.[13] It is important to note, however, there is significant debate as to the effectiveness of SQL standardization with regard to facilitation of interoperability between systems (Gorman, 2005). While support by database application vendors for early versions of SQL is good, by later versions, as early as 2001, many vendors stopped complying with the full standard (*ibid*.). This decline in compliance is largely a result of vendors developing features outside of the standard in order to appeal to a wider customer base. The development of bespoke features has restricted interoperability with other vendor products, thereby limiting the usefulness of the SQL standard (*ibid.*).

Users can search for data held in a database using SQL queries in order to get information or compare records. A database might have an index, or indices, to filter data more quickly, which is particularly useful with large databases. Users typically interact with a relational database through a database management system (DBMS), a software application that provides an interface for viewing or manipulating a database and other applications. Common SQL DBMSs include MySQL,[14] PostgreSQL,[15]

[12] American National Standards Institute: https://www.ansi.org
[13] International Standards Organisation: http://www.iso.org/iso/home/store/catalogue_ics/catalogue_ics_browse.htm?ICS1=35&ICS2=060&
[14] MySQL: https://www.mysql.com
[15] PostgreSQL: http://www.postgresql.org

Microsoft SQL Server,[16] and Oracle.[17] The different components required for running a database – the DBMS, an operating system, the database, and any other software applications used to operate the database – are usually referred to as the 'stack'. Users can also access databases using cloud computing, either independently through a virtual machine or by purchasing Database as a Service (DaaS), such as Amazon Relational Database Service[18] or DynamoDB.[19][20] Databases using cloud computing allow users to interact with databases through the web. Relational databases (SQL databases) are not always well-suited to cloud computing because they do not scale as well (Lake and Crowther, 2013).

Databases have been preserved since the 1980s (Lappin, 2011). In a blog post from Northumbria University's Records Management Today series, James Lappin interviews Kevin Ashley, the director of the Digital Curation Centre. The post discusses the challenges of archiving databases now versus then:

> 'An individual database in an organization today is easier to understand and extract data from than an equivalent database in the 1990s. But the challenge today is that the databases in an organization tend to be integrated with each other. For example all or most databases in an organization may use the organization's people directory to hold information about their users. As soon as you try to archive data from one database you are faced with the challenge of archiving data from all the other databases that it drew data from' (*ibid*.).

The interconnected nature of databases makes the use and manipulation of databases much more efficient for users, but more challenging for archivists. Furthermore, as large data increase and organizations face the challenge of collecting and managing big data, some are beginning to turn to alternative types of databases and technologies, namely to NoSQL (or 'non-relational') databases. As of November 2013, relational database systems account for about a third of all systems used, and remain by far the most popular form of database model.[21] Rather than a market takeover, NoSQL database models represent a new trend in the development of systems for coping with big data. MongoDB, a NoSQL document store model, for instance, currently stands as the fourth most popular Database Management System.[22]

Relational database models are much more established and supported and, therefore, provide a more useful technology for organizing data when consistency and reliability are higher priorities than performance. A number of specific conditions, however, may make relational databases a difficult method for storing data. The situations in which relational databases do not provide the best model are subject to opinion. Some programmers might advise against using a relational model, for instance, if the relevant data are structured as a hierarchy or a network with variations in depth. Similarly, if there is a greater need for reading than for writing, a relational database may not be the best solution. One perspective on the shortcomings of the relational database model is discussed in the Data Ops blog post, 'Why Programmers Don't like Relational Databases' (http://dataops.co/why-programmers-dont-like-relational-databases). It's important to remember that the choice of database model should be fundamentally informed by the nature of the data at hand and how users want to access that data.

5.3. Relational Databases and the Web

Relational databases also facilitate the interaction of organizations and users through the web. Server-side scripts allow a website to display a customized view based on a user's request – the website retrieves information from an underlying database and then returns it to the user. Many online retailers also use the web to allow customers to interact with their databases through OnLine Transaction

[16] Microsoft SQL Server: http://www.microsoft.com/en-gb/server-cloud/products/sql-server
[17] Oracle: https://www.oracle.com/uk/index.html
[18] Amazon Relational Database Service: https://aws.amazon.com/rds
[19] DynamoDB: https://aws.amazon.com/dynamodb
[20] https://en.wikipedia.org/wiki/Cloud_database
[21] DB-Engines: http://db-engines.com/de/blog_post/23
[22] DB-Engines: http://db-engines.com/en/ranking

Processing (OLTP).[23] This type of processing allows a database to respond to users immediately. Applications that support OLTP enable update-intensive database management and allow for hundreds of concurrent users.[24] For this reason, it is useful for online retail or banking.

Relational databases are particularly useful for financial transactions, including on the web, because they can be tested against ACID properties. ACID properties help ensure that individual user interactions with a database do not create contradictory or incomplete records (Lake and Crowther, 2013). ACID is an acronym that stands for Atomicity, Consistency, Isolation, and Durability. Atomicity ensures that a transaction must be complete before it will commit to the database: if one part of the transaction fails, the entire transaction fails. Consistency ensures that no new transactions can have an adverse effect on pre-existing records. Durability ensures that the database system will hold onto any completed updates (new complete records) even if the system fails before they can be written to the disk. Isolation prevents one transaction from affecting another – if two transactions occur simultaneously, they will be committed consecutively, rather than over the top of each other.[25]

Although relational databases provide a reliable method for facilitating transactions on the web, they are not necessarily the best solution for all types of web interaction. As mentioned, unless specifically designed to do so, for instance through the use of OLTP, relational databases are not always well-suited to scaling, particularly in a way that supports simultaneous access by many users. In the 2000s, a shift began to occur as new types of databases were developed to do what relational ones could not. Lake and Crowther describe this dawning trend:

> 'However, as web-driven systems began to expand, particularly when mass-usage systems such as Facebook and Twitter began to take off, it became clear that the relational model is not good at everything. It is said in some quarters, for example, that Relational does not scale well. And the fact that these new mass-usage systems are global and data is typically spread across many nodes in many countries, is seen by some as something that relational does not cope with well' (2013).

Many online systems use relational databases to run websites to support user interaction. However, certain types of big data, even some transactional data, are better supported – more easily accessed and analysed – through new types of database models not based on relational logic.

5.4. Non-relational (SQL) Databases

The growth in use of the web and resulting boom in large web-based platforms, such as Amazon, has pushed the limits of relational database solutions. As a result, in the early 2000s non-relational (or non-tabular) databases began to gain increased attention. NoSQL databases are a method of storing, managing and retrieving data using an organizing principle other than tabular relations. Relational databases, while very good at facilitating transactions, do not traditionally scale up or handle distributed (horizontal) processing across multiple machines (Lake and Crowther, 2013). The term 'NoSQL' refers to those types of databases designed to store documents, not relational data, and to allow quick access through scalability and distributed processing. There is some debate regarding whether this term means No SQL, as in does not allow the use of the SQL query language, or if it means Not Only SQL, as in allows both SQL and other languages (Fowler, 2012). Ultimately, however, the move to NoSQL-model databases is a move away from ACID properties as an underlying set of rules. Because NoSQL databases are designed to allow relationships between objects that are not the 'same', with distributed rather than uniform updates, they often contradict the underlying principles of ACID properties. Rather, NoSQL databases preference availability, access to data by all users through redundant copies (or nodes), and 'partition tolerance'. Partition tolerance 'refers to the ability of the database to find alternate routes

[23] OLTP on Wikipedia: https://en.wikipedia.org/wiki/Online_transaction_processing
[24] Oracle Database VLDB and Partitioning Guide: http://docs.oracle.com/cd/E11882_01/server.112/e25523/part_oltp.htm
[25] Tutorial's Point, DBMS Tutorial: http://www.tutorialspoint.com/dbms/dbms_transaction.htm

through the network to get at data from various nodes should there be breaks in communications', or in other words, favours quick search results over consistent transactions (Lake and Crowther, 2013).

For these reasons, NoSQL database models have become a popular database model for both big web-based platforms (such as those developed by Google—MapReduce—and Facebook—Apollo) as well as for institutions interested in the capture and processing of big data because they increase data retrieval speed (Celko, 2014).[26] MongoDB[27] and Apache Cassandra[28] are two examples of types of NoSQL database models that have grown in popularity. (For a longer list, please see http://nosql-database.org) However, the relational model remains a very useful one for storing and analysing big data, as evidenced by its continued popularity, particularly when the data must reflect consistent, durable transactions. Relational databases simply store large amounts of data differently from non-relational databases:

> 'Relational databases can, and do, store and manipulate very large datasets. In practice these are often vertically scaled systems … NoSQL databases tend to be horizontally scaled; that is the data is stored on one of many individual stand-alone systems which run relatively simple processes' (Lake and Crowther, 2013).

Therefore strategies for managing and preserving big transactional data – data resulting from interaction with a database – must adapt to both SQL and NoSQL environments.

Different types of NoSQL database models provide different benefits. As with the choice between relational and non-relational models, the choice of NoSQL model should be based on the nature of the data and how it will be used. The four main types of NoSQL database models organize data in different ways. These four main types, Key-Value databases, Document databases, Column Family stores, and Graph databases, allow for the organisation of data in ways largely prevented by relational databases (Sadalage, 2014). **Table *1*** below compares some of the attributes of these NoSQL database types.

[26] DB-Engines: http://db-engines.com/en/ranking
[27] MongoDB: https://www.mongodb.org
[28] Apache Cassandra: http://cassandra.apache.org

Name	Description	Benefits & Examples
Key-value	Stores and retrieves 'associative arrays', also known as dictionaries or hashes. Each associative array (or dictionary) contains objects, or records. Each object may contain many different fields, each field containing data. These objects are stored and retrieved using a unique identifier or a 'key'.	Fast search results, excellent performance, scales up e.g. Redis[29] and Oracle NoSQL Database[30]
Document	A sub-class of key-value databases; stores and retrieves documents, including XML and JSON. In a document database, or 'store', all of the information about a document, or 'object', is stored as a single instance. Document stores do not require all of the documents to be exactly the same.	Useful for programming web applications subject to frequent change, provides a rich query language, allows easier migration from relational databases e.g. MongoDB,[31] CouchDB,[32] Terrastore,[33] and OrientDB[34]
Column Family	Stores and retrieves data as columns of related data. Individual objects in a column family database consist of 'tuples' containing a key-value pair. The key is mapped to a set of columns (the 'value') that contain different types of information. Like a relational database, a set of columns is like a 'table'. Each key is like a 'row'. The tuple (or 'triplet') contains a column name, a value, and a timestamp.	Individual rows do not have to have identical columns and columns can be added to any row at any time without having to add it to other rows e.g. Cassandra[35], HBase[36], and Hypertable[37]
Graph	Stores and retrieves entities and relationships between entities. Entities, also called 'nodes', contain properties. Relationships, referred to as 'edges', can themselves have properties. The nodes in a graph database are arranged by relationships. This allows the discovery of interesting trends among the nodes. Data in a graph database can be stored once then interpreted in many ways.	Traversing relationships is very fast, information about relationships can also be stored to add intelligence e.g. Neo4J[38] and Infinite Graph[39]

Table 1: Comparison of different types of NoSQL database models, or 'stores'. More information about each of these types of NoSQL database models can be found in Sandalage, 2014.

[29] Redis: http://redis.io
[30] Oracle NoSQL Database: http://www.oracle.com/technetwork/database/database-technologies/nosqldb/overview/index.html
[31] MongoDB: https://www.mongodb.org
[32] CouchDB: http://couchdb.apache.org
[33] Terrastore: http://www.nosqldatabases.com/main/2010/10/7/terrastore-a-document-database-for-developers.html
[34] OrientDB: http://orientdb.com/orientdb
[35] Apache Cassandra: http://cassandra.apache.org
[36] Apache HBase: https://hbase.apache.org
[37] Hypertable: http://www.hypertable.com
[38] Neo4J: http://neo4j.com
[39] Infinite Graph: http://www.objectivity.com/products/infinitegraph

6. Issues for Long-term Preservation

The number of organizations that collect or hold forms of transactional data is increasing as many switch to web-based interactions with users or implement electronic methods of tracking everyday activities. New methods for archiving and long-term preservation will enhance the value and usability of these data. As the OECD Global Science Forum articulate in their report *New Data for Understanding the Human Condition*:

> 'Formats of digital data are much more complicated and diverse than the traditional format of a data set ... . Communications and transactions involve two or more units, and they can be represented by networks or more complicated structures. This leads to the necessity of using non-traditional methods for data management and data analysis' (2013).

Many databases designed for the capture of web-based or other digital transactions can be large and volatile – changed or updated frequently. For some database systems, such as those underlying retail interactions or automated tracking, the data are never 'finished' but open-ended. In these instances, any snapshot or derivative dataset might quickly become incomplete or irrelevant. In some cases, the meaning of raw data depends heavily on a wider structure – or environment. If data are embedded in a complex structure of applications, they can often be fragmented and may require documentation and re-deployment of underlying processes or data models to ensure meaningful access. In the commercial sector, these issues have been approached using data warehousing—a strategy developed for comparing static data from several sources in order to gain business intelligence. Data warehousing refers to the use of a relational database designed for query and analysis rather than for transaction processing (Lane, 2002). This technique has potential as a digital preservation solution. The E-ARK Project[40] is exploring the use of data warehousing in archives. Complexity and cost may be a barrier to the adoption of this technique in archives, but the commercial sector is beginning to address these challenges (Hughes, 2016).

Summary of Significant Challenges

Volume and capacity: data can be big to start with, multiple snapshots over time may swamp archival storage and/or workflow processing capability

Volatility: data change rapidly and might draw on multiple changing sources

Multiple entry routes: different sources of data creation can lead to data quality issues

Context: understanding the context and how the data were created may be critical in preserving the meaning behind the data

Data purpose: preservation planning is critical in order to make preservation actions fit for purpose while keeping preservation cost and complexity to a minimum

Legalities: many forms of transactional data, particularly those of interest for re-use, contain personal information about individuals, making them subject to Data Protection regulations. They may also pose issues of Intellectual Property Rights legislation and/or restrictive licensing arrangements. The legal complexities attached to these data make it important to retain metadata documenting relevant conditions and risks

[40] E-ARK Project: http://www.eark-project.com

How data are collected or created can have an impact on long-term preservation, particularly when database systems have multiple entry points, leading to inconsistency and variable data quality. Documentation that describes the possible values, or content, of particular data fields may only partially answer this challenge. In some instances, a database or system of databases can be viewed from multiple access points. In these cases, point-of-view and personalization may also need to be preserved in order to make the data meaningful. Preserving a user's experience with a database through a web transaction could be achieved in a number of ways. For instance, it could be captured through screen casting or web harvesting and bundled along with metadata or linked with other forms of archived database content. Software preservation techniques could even be applied to preserve the underlying system and the method of interaction with the data, but this would obviously introduce considerable complexity to the preservation process.

Verifying the authenticity of records within a database also poses a challenge. While ACID properties can support the development of database systems that ensure the completeness and integrity of records, the attendant contextual information needed to authenticate records will be unlikely to remain associated with individual items of data. Relational databases are designed to break data down into separate parts so that they can be re-arranged or combined in different ways to answer different questions. However, breaking data down into smaller components means that the context around individual records is often lost. For instance, a database that supports an online retailer will capture the data generated by individual transactions, such as payment information or inventory levels. These data will not necessarily represent the version of the website at the time of sale or the scripts used to interact with the underlying database. Records of transactional data are much more useful for analysis if they can be reliably supported by contextual information, such as the technology underpinning the data or the programming used to interact with the database.

Transactional data often impose information security requirements and liabilities for the archives that hold or distribute them. Common legal and ethical issues stemming from transactional data include confidentiality, privacy, data protection, and copyright. These legal and ethical issues pose particular problems in cases where data are collected for purposes other than analysis or research. While relevant legislation does allow for the preservation of these data, deciding legality often entails navigating different types of exceptions applicable only in specific circumstances. Beyond legal issues, transactional data challenge current approaches to ethics as well. While de-identification and anonymization techniques may help reduce the risk of accidental disclosure in isolated datasets, the application of computational processing that combines data from multiple sources dramatically increases the likelihood of re-identifying individuals represented in the data. Furthermore, gaining consent from individuals to re-use data collected primarily for another purpose could be incredibly difficult or impossible, especially if a dataset contains hundreds or even thousands of data subjects.

Storage capacity poses an increasing challenge as these data grow and users find new ways of re-using them. For instance, if an organization preserves an active database by taking periodic snapshots, the size of the archival copies may quickly take up all available storage space. The funding and resources required for frequent increases in storage make it difficult for most organizations to sustain this level of growth.

Approaches to the preservation of transactional data, for organizational analysis or for research by other institutions, depends heavily on how and why those data will be used in the future. Many organizations that collect these data may have legal or regulatory reasons for preserving transactional data, for example, as evidence of transactions. They may want to preserve data in a way that also facilitates computational processing for consumer analysis or evaluation of business activities. These organizations may want to preserve entire databases and may also want to preserve derivative datasets created for particular types of analysis.

None of these challenges listed poses an unsurmountable obstacle to robust preservation of transactional data. As new implementations of database technology – both SQL and NoSQL – are created to store, manage and retrieve data, preservation techniques will need to adapt. Some forms of NoSQL database models, for instance, may require more documentation of the applications used to interact with the data, as programmers increasingly rely on information stored at the application level (Sadalage, 2014). In some cases, SQL databases also pose challenges to current preservation techniques (explored in

more detail in Section 9). As the capacity to link data with other data sources increases, for instance through the Semantic Web or open data repositories, the demand for re-using archived data may put strain on current preservation practices. Again, more information from the application layer, the software used to interact with the data, may be required to make the data more usable. As database technologies change and more and better solutions arise, it will be ideal for organizations to strategize for the preservation of databases, data, and applications. Ideally, this strategizing will take place at an early stage, for instance when choosing a database model that best fits an institution's requirements. A long-term preservation strategy should be a consideration when choosing a database model. Early preservation planning will avoid losing data later in the process. The PLATO tool, developed through the EU-funded SCAPE project, offers support in implementing preservation planning.[41]

7. Approaches to Curatorial and Organizational Challenges

Using data for a purpose other than the one for which it was originally collected creates a number of legal and ethical concerns.

Transactional data, as collected by government and other organizations, 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 data. Issues of ownership and intellectual property rights may pose issues, but more often, transactional data contain personal data and, in the UK, must comply with the Data Protection Act (1998) (DPA) and ethical standards for protecting data subjects. In many cases, the legal and ethical issues can be resolved, for instance through de-identification, but organizational mechanisms or institutional culture may prevent the use of these data. The adaptation of these data for research often also poses quality issues, such as incomplete data or datasets too large for most archival repositories to handle. This section discusses the issues of preserving transactional data that can be addressed through curatorial strategies and organizational policies.

The use cases presented later 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 they were originally collected creates a number of legal and ethical concerns. If the original data contain personal information about individuals, re-using or sharing those data could be prevented by the DPA. Furthermore, the ability to merge datasets by matching variable points from two or more (sometimes many more) datasets increases the likelihood of accidental disclosure. For instance, a dataset that holds an entry for a doctor living in the Greater London area in 2016 with a particular income may safely conceal the data subject's identity. However, if combined with a dataset that adds information, for instance, that the doctor living in the Greater London area also has 12 children, the doctor's identity may be easy to identify due to the highly unusual circumstances of having 12 children in the year 2016. Therefore, to ensure compliance, actions must be taken to prevent accidental disclosure of individual identities, such as using a trusted third party to replace personal identifiers (ADT, 2012). Once data controllers, or data owners, have assessed that data can be shared legally, 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. An assessment of preservation risks should be carried out at the point of sharing or merging data. Also at this early stage, it is crucial to gain the necessary permissions to preserve data and derivative datasets for the necessary amount of time.

7.1. Data Protection

In the UK, the Data Protection Act (DPA) (1998)[42], partly an implementation of the EU Directive 95/46/EC (adopted in 1995), is the principal source of legislation governing the management and re-use of data. The DPA aims to ensure that the data of private individuals are not used for purposes other than those

[41] Plato: http://www.ifs.tuwien.ac.at/dp/plato/intro
[42] UK Data Protection Act (1998): http://www.legislation.gov.uk/ukpga/1998/29/contents

agreed and understood by the individual who provided the data. It also aims to prevent the accidental disclosure of individuals' identities to the public or unintended audiences. The Act distinguishes between personal data and sensitive data. Sensitive data include information such as racial or ethnic origin, political opinions, physical or mental health or condition, sexual life, and the commission or alleged commission of any offence.[43] Though the accidental disclosure of any type of personal data could potentially put an individual at risk, sensitive data have the potential to make individuals vulnerable to discrimination or harm. The risk of accidental disclosure due to the sharing or re-use of data, however, is negligible (PHRDF, 2015). According to a 2015 report to The Wellcome Trust, most data controllers base their security policies on an 'intruder' scenario, which is a worst-case scenario and statistically unlikely to occur (*ibid*.). The legal framework that protects the personal data of individuals regulates what data can be shared and how. To comply with this framework, sophisticated techniques and methods for de-identification and anonymization have been developed that reduce the risk of accidental disclosure.

Many organizations are obliged to enforce measures for data security, retention schedules, and de-identification in order to comply with the DPA. The issues of data protection, and the sharing of government data (or 'administrative' data) in particular, have been widely covered by the Administrative Data Research Network (ADRN), particularly in *The UK Administrative Data Research Network: Improving Access for Research and Policy*, a report from the Administrative Data Taskforce from December 2012. The third section of the report, 'Legal and Ethical Issues', provides a concise overview of the challenges facing the re-use of government data for non-commercial research. Different types of government organizations have different powers governing how they can share data. The development of legal gateways, provided under statutes, do allow specific organizations to share data (ADT, 2012). Overall, however, allowances for re-use in non-commercial research do not apply universally or uniformly to government data in the UK (*ibid*.). Graham Laurie and Leslie Stevens at the Administrative Data Research Centre-Scotland have developed guidance for government organizations and researchers on the legal and ethical restrictions to sharing administrative data (2014). **Figure *2*** below illustrates a straightforward framework for implementing this guidance. Their report includes decision-making rubrics for approaching the task of sharing data. They discourage unnecessarily cautious practices in favour of using legal allowances to share data for research that could contribute to the public good (Laurie and Stevens, 2014).

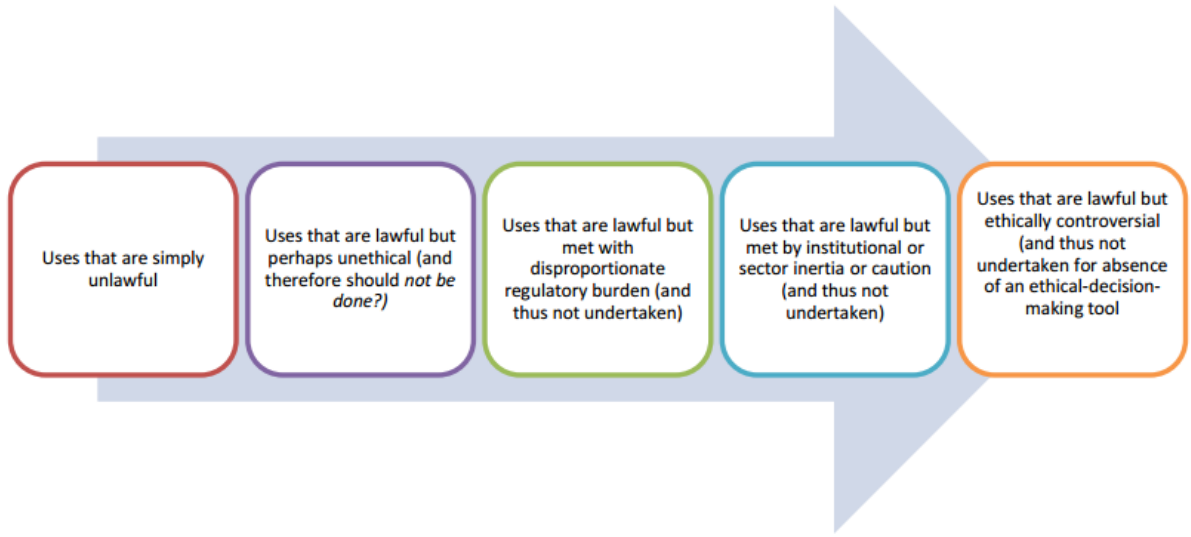

Figure 2: Administrative Data Decision-making Matrix

Image from 'The Administrative Data Research Centre Scotland: A scoping report on the legal & ethical issues arising from access & linkage of administrative data' by Graham Laurie and Leslie Stevens (see References for full citation). Printed with permission.

[43] ICO Guide to Data Protection: https://ico.org.uk/for-organizations/guide-to-data-protection/key-definitions

Though commercial organizations may refuse to share data for business reasons – namely to maintain data assets that could reduce profits if released to competitors – the DPA is the main *legislative* restriction. Commercial organizations interested in sharing data for non-commercial research purposes must also ensure the privacy of their data subjects. Public and private sector organizations alike benefit from the ability to share data, for instance to improve services. The DPA recognizes these benefits and, despite regulation, provides for data sharing. The UK Information Commissioner's Office provides a 'Data sharing code of practice' that outlines general guidance and best practice for public and private organizations interested in sharing data. As the ICO guidance points out, however:

> 'There may well be other considerations such as specific statutory prohibitions on sharing, copyright restrictions or a duty of confidence that may affect your ability to share personal data. A duty of confidence may be stated, or it may be implied by the content of the information or because it was collected in circumstances where confidentiality is expected – medical or banking information, for example. You may need to seek your own legal advice on these issues' (2011).

As demonstrated by the plethora of guidance in this area, many data controllers, or institutions who hold transactional data, find the legal restrictions to sharing data daunting. How organizations approach legal and ethical requirements will have a huge impact on how, or if, data make it into the hands of data analysts or researchers.

7.2. New EU Legislation: the European General Data Protections Regulation

Many of the particular legal pathways open for sharing data for non-commercial research will change in coming years. Though the underlying principles of data protection laws and regulations will remain the same, new implementations and exceptions will soon be established. In December 2015, new EU regulations were approved that will come into effect in 2018 to replace the current Data Protection Directive 95/46/EC and the UK Data Protection Act.[44] 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.

7.3. Legal Protection of Databases: Copyright and Sui Generis

In some cases, the preservation of an entire database will be restricted by other legal considerations, particularly copyright and the economic rights (or sui generis) of the rights holder of the database. If an organization wants to preserve a database they did not themselves create, or for which they are not the exclusive rights holder, they will need to observe regulations that protect the rights of the database creator (or legal rights holder). In some cases, an organization will need to arrange permissions for actions required for preservation at the time they acquire the database. Often, a delay in acquiring permissions can make it more difficult to identify the rights holder and more time-consuming to arrange an agreement. European law regarding these protections for databases are informed by EU Directive 96/9/EC, though individual nations have different implementations.[45] For instance, in the UK, the organization who employs the creator of a database holds exclusive rights; in Germany, the exclusive rights remain with the individual person who created the database.

The Directive provides some general guidelines to the rights in databases and provides, to an extent, definitions of terms. Article 1(2) of the Database Directive defines a database as: 'a collection of independent works, data, or other materials arranged in a systematic or methodical way and individually

[44] 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 here: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52012PC0010

[45] About Directive 96/9/EC Legal Protection of Databases: http://ec.europa.eu/internal_market/copyright/prot-databases/index_en.htm

accessible by electronic or other means'.[46] The Directive deliberately abstains from requiring a specific form and (as a result, it applies to a broad variety of databases: electronic or non-electronic, dynamic or static, printed or digital (Hoeren, Kolany-Raiser, Yankova, Hecheltjen, and Hobel, 2013).

The Directive forbids the reproduction, adaptation, or alteration of any database, in any form, that enjoys copyright protection. This protection does not extend to the content of the database; the copyright in the content and in the database are exclusive and do not impact the copyright status of the other. The actions restricted by the Directive are required for preservation; for instance, to copy a database for back-up or migrate a database to a new environment requires reproduction, adaptation, and alteration. Therefore, for a non-owner to preserve a database, they must first gain permissions. The economic rights in the database apply if the rights holder has made a 'qualitatively and/or quantitatively substantial investment in either the obtaining, verification or presentation of the contents (Article 7 [1])'.[47] Economic rights do not extend to the content of the database but only to the structured database itself.

Some exceptions exist within the Directive that allow for teaching, scientific research, public security, and administrative or judicial procedures (*ibid*.). These exceptions extend only to the *use* of the database, not to preservation. No national or EU legislation provides exceptions for the creation of a back-up copy to prevent data loss during migration or other preservation activities (*ibid*.). The protections in the Directive, however, do not last forever. Copyright expires 70 years after the death of the creator or rights holder (or 70 years after the death of the last living creator) and sui generis (economic rights) expire 15 years after the completion of the database.

More guidance on the legalities of preserving databases can be found through the Legalities Lifecycle Management[48] materials developed through the TIMBUS Project, including training videos.[49]

7.4. Organizational Policy and Data Sharing

Unfortunately, the complexity of legal and ethical issues prevent data owners from sharing data, even legally.

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 of the reuse. 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 organizations 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 prevents data owners from sharing data, even legally. Laurie and Stevens quote from the *Data Sharing Review Report* to illustrate the situation at most organizations considering whether or not to share data:

> '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"'[50] (2014).

This confusion belies a predominant 'culture of caution' at organizations in a position to share valuable data. One remedy for this 'fog of confusion' is education. The Big Data Network, including the

[46] EU Directive 96/9/EC: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31996L0009:EN:HTML
[47] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31996L0009:EN:HTML
[48] Legalities Lifecycle Management: http://timbusproject.net/portal/domain-tools/72-portal/domain-tools/334-lehalities-lifecycle-management-tool
[49] Legalities of Database Preservation training videos: https://www.youtube.com/playlist?list=PL2fxQHETrFueJk7wIaJdlmaibQogcNSYm
[50] Richard Thomas and Mark Walport, 'Data Sharing Review Report', 11 July 2008, Foreword. http://systems.hscic.gov.uk/infogov/links/datasharingreview.pdf

Administrative Data Research Network, positioned as they are 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.

7.5. Challenges to Merging Data

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 ingestion into data repositories or onto researchers' machines. Many of these problems are caused by a lack of uniform approaches and standards used by organizations (or different government bodies) when collecting data. Some organizations are better funded than others or allocate more funding to the development of systems for collecting and storing data. As a result, many types of data that could be merged or compared are not compatible. As Moody highlighted in her definition of big data (see Introduction), these datasets are often simply larger and more complex than the datasets researchers or data managers are accustomed to handling (2015).

This challenge could be attributed to a problem of scale, meaning repositories face a growing issue of storage capacity as well as processing power. The lack of a consistent approach across related organizations also leads to broken or incomplete data. 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 service provided by the ESRC's Big Data Network 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.

7.6. Standards and Documentation

Organizations have only fairly recently begun to capture and re-use transactional data, and other forms of big data. Although commercial sector organisations have been storing and analysing data as part of Business Intelligence for several decades (Hughes, 2016), few archival standards or best practice guidelines have been developed to provide a benchmark for successful preservation of these data. The initiative researchobject.org is one example of work currently underway to develop best practice for re-using data in research and enhancing the ability to share data in machine-readable form.[51] For now, related guidance and standards, though not developed directly for preserving transactional data, provide a degree of stability and assurance to the practice of preserving these data.

Data Documentation Initiative (DDI)

[51] researchobject.org: http://www.researchobject.org

For re-use of statistical and social science data, the Data Documentation Initiative (DDI) provides a range of standards recognized internationally. They offer specifications for the Lifecycle management of research data and a Codebook that provides a compressed standard used to document simpler survey data. The Lifecycle specification, now on version 3.2, documents and manages data from the planning stages to publication to analysis. Both specifications are based on XML and the Lifecycle specification is modular and extensible. DDI also provides a controlled vocabulary that can be used with the specification or for other applications. They also provide DDI-RDF vocabularies that makes it possible to publish metadata about datasets in the Web of Linked Data.[52] The development of DDI takes into consideration other established international standards. The current version of DDI maps relationships to a number of standards including Dublin Core, MARC, METS, and PREMIS.[53] Mapping to PREMIS, in particular, could make it simpler for data creators to provide necessary preservation metadata at an early stage.

Commercial Sector Guidance

Both the commercial sector and the public sector can find guidance to best practice from a number of sources. The Association for Data-driven Marketing and Advertising (ADMA) have published a *Best Practice Guideline* for commercial organizations interested in exploiting big data to increase customer engagement and improve marketing strategies (2013). This guidance emphasizes the need to curate big data so that they can be compared with other sources. It also provides information about the privacy laws that affect how big data can be used in a commercial context. The Administrative Data Research Centre-Scotland at the University of Edinburgh have delivered best practice guidelines for the sharing of administrative data from the public sector (Laurie and Stevens, 2014). These guidelines provide a review of established standards and frameworks of governance as well as guidance on using historical data and forging commercial partnerships. Best practice and guidelines for curation and legal review help establish the long-term value of these data. As these practices develop, it will be necessary to address the actions needed to further ensure long-term access to commercial, public sector, and research data.

The technical actions required to preserve databases also employ different types of standards, standard formats, and languages. Those standards will be covered in the following section on Technical Solutions.

8. Case Studies

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, currently still face considerable obstacles to capture and sharing. The ADRN follows a 'create and destroy' model; in other words, all data are destroyed at the conclusion of individual research projects.[54] The ADRN centres, for administrative data, and the other Big Data Network research centres, for business and local government data, help negotiate the use of third-party data, but these centres often do not have ownership or even possession of those data.

The institutions which own the source data used by researchers 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. Commercial organizations do not often publish information about their data collection or preservation operations, though they are increasingly transparent regarding the treatment of personal data. Revealing company methods for collecting data could potentially jeopardize trade secrets and reduce profits. Commercial organizations, for these reasons, rarely share data, and when they do, they often insist on extraordinarily strict security and access measures. For that reason, this report does not present detailed information on any particular commercial dataset, but rather gives general descriptions of the types of data collected by commercial organizations. The use cases here are instances of how other, potentially similar, data can be re-used for research, and the accompanying management and preservation requirements. 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.

[52] More information about DDI standards: http://www.ddialliance.org
[53] DDI: http://www.ddialliance.org/standards/relationship-to-other-standards
[54] Administrative Data Research Network: http://adrn.ac.uk/protecting-privacy/secure-environment

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 Big Data Network Support 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 the time of selection and capture (or acquisition), in order to plan for transition to archival storage and access over time. Though other conditions may prevent the immediate preservation of these data, data curation benefits from long-term planning. Long-term planning, for example, gaining permissions for preservation actions such as duplicating data and capturing sufficient metadata, increases the usefulness of data for research. The following case studies present data at different stages in their lifecycles, but all demonstrate long-term value. They are written to stand alone and also provide context and practical examples of the issues described in this report.

8.1. Energy Demand Research Project: Early Smart Meter Trials at the UK Data Service (UKDS)

Background

Funded by the ESRC, the UK Data Service (UKDS) provides resources and support to researchers, teachers and policymakers who depend on high-quality social and economic data for research and analysis. UKDS collections curated at the UK Data Archive form Britain's largest collection of social and economic data, including key well-known national datasets such as the census, the National Household Survey, and the National Crime Survey.[55]

In this context, long-term preservation starts early in the lifecycle of these data.

Through its Big Data Network Support (BDNS) project and team, the UKDS is also extending its capacities to curate and facilitate management and analysis of new and novel forms of data, including very large datasets. BDNS plays a key role within the ESRC-funded Big Data Network,[56] coordinating and harmonizing workflows across three specialized research centres around the UK: the Urban Big Data Centre at the University of Glasgow, ESRC Business and Local Government Data Research Centre: University of Essex, and the Consumer Data Research Centre: University of Leeds and University College London.[57] The BDNS team are helping to increase access to these types of data through their own core services. One strategy for developing these services has been the curation of the Energy Demand Research Project smart meter datasets.

Example of Transactional Data

Energy Demand Research Project: Early Smart Meter trials (2007–2010) (EDRP)

Persistent Identifier (PID)

10.5255/UKDA-SN-7591-1

Description

The EDRP data derive 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. CSE also ensured de-identification of the portions of the data

[55] UK Data Service: https://www.ukdataservice.ac.uk
[56] Big Data Network: http://www.esrc.ac.uk/research/our-research/big-data-network/big-data-network-phase-2
[57] UK Data Service: https://www.ukdataservice.ac.uk/about-us/our-rd/big-data-network-support

sent to third parties for research, partially through clustering. 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 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.

Long-term Preservation Requirements

The EDRP datasets were collected from energy suppliers by CSE and deposited with UKDS by the Department of Energy and Climate Change. The UKDS now curates the EDRP data, providing authorized access through its Discover catalogue.[58] 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.

The BDNS team are using the EDRP as a test case for piloting a new system facilitating management, analysis, and visualization of big datasets. Re-storing the data on an Apache Hadoop cluster facilitates easier access and manipulation. Apache Hadoop is entirely open-source software that makes it possible to store large datasets on multiple machines in a way that allows them to be processed simultaneously – this is called 'distributed storage' and 'distributed processing'.[59] Hadoop clusters will allow the team at UKDS to create data visualizations, data products and to merge smart reader data with other data sources, such as weather data and fuel poverty data (Corti, Bolton, and Moody 2015).

In order to perform these actions, the EDRP datasets must have complete and accurate metadata in a format that can be compared with other datasets. If the metadata variables in the EDRP datasets have a different meaning than identical variables in other datasets, such as 'employment', they cannot be combined to make further connections or observations (*ibid*.: 'Metadata issues'). For big data, perhaps even more so than 'small' data, the preservation of associated metadata is vital for other researchers to understand what the data represents and how they can be used. In other words, large amounts of transactional data without sufficient metadata are unusable.

The UKDS has created a very useful model for archiving transactional data, from acquisition to curation for re-use in research to engagement with users. The UKDS team have not only ensured that the EDRP datasets have been curated and catalogued with documentation – including how the data were collected and why – but they have used the data to develop services to support and engage the research community. The infrastructure at the UKDS allows the team to use mature, tested processes for acquisition and preservation and also to provide training and services for researchers. This type of data requires new skills, or the adaptation of old ones, in order to perform analysis that will have any significant impact. UKDS help to increase the usability of these data and the quality of research outputs by broadening the community of scholars with the knowledge to work with this data effectively.

The UKDS is based at the UK Data Archive, a trusted digital repository that complies with (and helps to establish) standards of best practice. The Archive has been externally reviewed and approved for the Data Seal of Approval[60] and has extensive organizational policies for data management and preservation.[61] Part of the UKDS's role is to provide guidance for researchers on how to deposit data with the archive. Any data deposited at the UKDA, or any other trusted repository, enjoy the benefits of an established infrastructure and workflow for curating and preserving data. These benefits entail the expertise of specialists in the curation of big data, assurance that research will be reproducible, and the warranty of a trusted preservation framework. The model created by the UKDS, as exemplified by the EDRP datasets, provides support for re-use of these data, but also for their long-term preservation.

[58] UKDS Discover Catalogue: https://discover.ukdataservice.ac.uk/catalogue/?sn=7591&type=Data%20catalogue
[59] Apache Hadoop: https://hadoop.apache.org
[60] UK Data Archive: http://www.data-archive.ac.uk/curate/trusted-digital-repositories/standards-of-trust?index=1
[61] UK Data Archive: http://www.data-archive.ac.uk/about/publications

8.2. 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 (UCL), University of Liverpool, University of Oxford 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'.[62] The CDRC acts as a liaison between consumer-oriented organizations and trusted researchers in order to promote innovation in the use of data. Their partners include Acxiom, Appliances Online, CACI, Heart Research UK, and Shop Direct.[63] The CDRC offer a three-tier data service providing open, safeguarded and controlled data via a secure infrastructure for trusted researchers to access data. Researchers gain access in the case of open data through a process of simple registration, and to safeguarded and controlled data by application, reviewed by the data partner(s) concerned and an independent research approvals group. Upon approval, access to safeguarded data is made available via a secure download and to controlled data through a state-of-the-art secure lab. The CDRC provide an online catalogue of data available including metadata and support for researchers looking to access these data. Though many of the datasets in the catalogue are open – also available from other places on the web – the CDRC catalogue published on their open website provides added value as well as centralized discovery, making the datasets more searchable.

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 are one example of the several open data sources the CDRC has used to demonstrate the possibilities of analytics on these types of data. The OAC, LOAC, and TOAC 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 sharing similar population and built characteristics.[64] They are popularly used for a variety of product and service targeting in both the public and private sectors.

CDRC Long-term Preservation Requirements

The CDRC do hold some data, either temporarily or for the long term, enabling them to facilitate re-use by researchers and data analysts. Because of this primary function, citation and persistent identification are both high priorities in order to enable reproducibility. Reproducibility relies on proactive data management and robust preservation. The CDRC currently follows a number of rigorous management processes and relies on support from the UK Data Service for long-term preservation support where appropriate.

Different versions of datasets created by CDRC data analysts are separately deposited and catalogued. At deposit, CDRC staff provide metadata for the following attributes:

[62] Consumer Data Research Centre: https://www.cdrc.ac.uk/research/research
[63] Consumer Data Research Centre: https://data.cdrc.ac.uk
[64] Office for National Statistics: http://www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/ns-area-classifications/ns-2011-area-classifications/index.html

- Name of original dataset
- Version number
- Description of work done on it
- Data source (e.g. information about commercial partner)
- Contact at CDRC
- Restriction information from data licence

The ingest process begins when a CDRC researcher or data scientist uploads data or creates a new dataset from open data. The data are uploaded to a sandbox for review. An independent CDRC team member will check the data as part of quality assurance and also review and supplement the associated metadata. Finally, a project manager will do an overview of data licence information and record permissions in the metadata. A project manager then does a final quality assurance to ensure uniform data quality and a correctly formatted catalogue entry.

The role of the CDRC as a liaison between commercial organizations and researchers depends heavily on rights management. The lengthiest stage of acquiring data is often negotiations with commercial partners over the terms of data licences. This process, on average, takes eight to ten months and is supervised by university legal teams. The typical data licence, issued by one of the university partners on behalf of the CDRC, achieves a number of key allowances. Data licences ensure data sharing across all CDRC centres and establishes the deposit of outputs of research with the UK Data Service at the end of the project (February 2019). Usually, however, commercial partners do not grant permission to deposit original datasets with the UK Data Service. Most data licences also gain permission for the CDRC to publish limited metadata about the original dataset in CDRC's online catalogue. Some bespoke licence exceptions are granted for particular researchers (based on specific projects). Sometimes these bespoke arrangements generate aggregated outputs that can be deposited with the CDRC under a separate agreement.

All data held through negotiation with commercial partners is held for duration of the project on CDRC local servers within the secure labs. Any derivative datasets for which the CDRC gain ownership, therefore, require careful preservation planning. This planning includes the assignation of persistent identifiers to support citation so that other researchers and organizations can build on the work done by CDRC research teams and data scientists. The UKDS provides data management support for all of the Big Data Network research centres in order to ensure that the process of data management and preservation planning happens uniformly across the centres. This process will entail transferring datasets, each assigned a DOI by UCL, to UKDS ownership. Because of the legal issues of ownership and privacy attached to these types of data, the CDRC maintain detailed records of the licence agreements with their data source partners. This information will need to remain associated with archived datasets as metadata to make access simpler for future users and to maintain provenance.

8.3. Higher Education Data at the Administrative Data Research Network (ADRN)

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 (ADT, 2012). 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 wellbeing. Work achieved through the ADRN will hopefully influence a culture of caution currently inhibiting the sharing of valuable data (Laurie, 2014).

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) has 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 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 Individualized 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.

Long-term Preservation of Original and Derivative Data

Although the ADRN can help facilitate access and discoverability of the HESA student record data, they do not hold this data. The ADRN does not have any archival or preservation accountability, but through its services and training for researchers, it can encourage uniform processes to access these data and support reliable data management during the research process. The HESA datasets, in particular, will require increasing documentation and robust preservation. In 2015, HESA launched a consultation on the future of higher education data.[65] This consultation underpins the new Data Futures programme that aims to modernize and improve the collection and delivery of higher education data.[66] These changes will have an impact on the ways the data are preserved by HESA but will also impact how research data based on these data can be created and preserved. The guidance offered by the ADRN can ensure that research data contain sufficient metadata to describe the underlying properties of the original data.

Current methods of data collection by HESA prevent longitudinal analysis because: 'complete data on the student population are gathered each year, but data about individual students are not gathered and cannot be linked between years'.[67] These data already required metadata and documentation to identify the fields of data collected for a relevant year. This information is important for future researchers to understand the context of the data for analysis and citation. As HESA transition to a new system, new issues arise that may impact both HESA and researchers.

The HESA Data Futures programme aims to improve information management by implementing data warehousing to provide infrastructure.[68] It also aims to implement new technology for both the collection of data and for facilitating access.[69] The transition to a new infrastructure for managing their data will require careful preservation to prevent data loss; for instance, they may choose to preserve legacy systems in order to maintain support for research and analysis using older datasets. It also

[65] HESA consultation on future of data: https://www.hesa.ac.uk/pr/3745-press-release-222 and analysis of responses: https://www.hesa.ac.uk/component/content/article?id=3741
[66] HESA Data Futures programme: https://www.hesa.ac.uk/component/content/article?id=3741
[67] ADRN HESA student record catalogue documentation: http://adrn.ac.uk/catalogue/cataloguepage?sn=888013#documentation
[68] HESA Proposals for Change: https://www.hesa.ac.uk/component/content/article?id=3740
[69] *Ibid.*

requires careful planning for the new preservation concerns introduced by the new systems. Standards for the citation of research data and reproducibility of data analysis in the academic sector are increasing every year. Increasingly, funders require that researchers publish the data underlying their written publications.[70] These requirements put greater importance on the sustainable preservation of both data held in retired systems as well as data collected and held in the new system once it is in place.

8.4. Summary

These examples demonstrate forms of transactional data – information generated through the interaction of individuals with third-party organizations – 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 have 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. The following section presents approaches to the types of challenges posed by transactional data to long-term preservation. It provides guidance on the legal, ethical, and organizational issues faced by many institutions looking to re-use these data, and points to a few solutions to alleviate these challenges. It also provides a section on the preservation of databases, a process that ensures long-term access to data held by institutions who collect data routinely. They also provide insight into the data management practices at these institutions for researchers and those interested in re-using data for analysis.

On their own, no one example from the cases above necessarily meets the general definition of 'big data' (such as the one quoted in 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 provides new opportunities for these approaches to research and analysis. As a result, data managers and archivists face increased challenges to curation and long-term preservation in order to maximize and build on these opportunities.

9. Technical Solutions: Preserving Databases

Current technical approaches to preserving transactional data primarily focus on the preservation of databases. Database preservation may not capture the complexities and rapid changes enabled by new technologies and processing methods; it does, however, provide an important foundation for further developing long-term preservation strategies for transactional data and other forms of big data, including SQL, document-oriented databases.

The preservation of transactional data in relational databases requires consideration for a wide range of data types captured by a variety of technologies, and owned and managed by an array of institutions, under varying legal and regulatory frameworks. A preservation strategy that encompasses this complex category of digital object has to break down data based on function at particular stages in the data lifecycle. Government organizations and commercial companies routinely collect and preserve data as evidence of transactions. Data analysts or marketing specialists may want the data for processing and analysis. The data re-used in research must be obtained from an external data source and curated or 'cleaned' to make it usable for researchers. For any of these stages in the life of the data to occur, the original source of the data must be preserved. While further action must be taken to capture the changes and evolution of these data through the interactions of users, preservation of the underlying databases offers an important solution. Datasets created through the analysis of data are strengthened by the

[70] DCC Funders' data policies: http://www.dcc.ac.uk/resources/policy-and-legal/funders-data-policies

ability to cite or point to an existing archival copy of the original data. This section outlines the strategies and approaches to performing the preservation of databases.

As with all preservation planning, the relevance of a specific approach depends on the organization's objectives.

Databases are a key technology to support the storage, organization, and interaction with digital information. In fact, most contemporary information systems cannot perform without them. Databases do not act solely as data containers; in fact, databases underlie a broad range of activities required for other applications and processes to function properly. They are responsible for enforcing important business constraints, performing data transformations and validations, managing access restrictions, and providing discovery and filtering. At the same time, they have to support the basic functions implemented by relational databases in real-time (also referred to as the CRUD operations: create, read, update, and delete). Because of this complexity, the long-term preservation of databases poses a considerable number of unique problems.

Different communities approach the archiving of databases in very different ways. IT personnel often view 'preservation' as backing up or copying data to a remote location. Archivists view preservation as a more holistic activity, requiring the content, information about the content (metadata), and information about any dependencies such as software or hardware. The sections below discuss the approaches considered robust by archivists, who must work closely with IT personnel and data creators within an institution.

9.1. Approaches

There are several approaches to preserving databases, each with its own strengths and weaknesses. As with all preservation planning, the relevance of a specific approach depends on the organization's objectives, the intended use by the designated community, and available resources.[71] The following sections outline the available approaches and their relative benefits and drawbacks. In some cases, the use of multiple approaches may provide the best solution. For instance, an organization may want to preserve the database in its original format as well as migrate it to a different format. This approach could prevent the preclusion of unexpected modes of re-use in the future. The decision to choose one of these approaches, or a combination of approaches, should be based on the evaluation of user needs and preservation requirements established during preservation planning.

Encapsulation

Encapsulation entails the collection of documentation about a database's technological environment. This documentation may include manuals of the Database Management System (DBMS), information about the end-user application, file format specifications, details about the operating system, and descriptions of hardware. Documentation may even include information about other applications that coexist in the same IT environment. One of the difficulties inherent to this approach lies in knowing the full extent of the documentation that will be needed to understand the content in the future. For this reason, encapsulation is rarely used by itself, but rather serves as a foundational activity or a supplement to other approaches, such as emulation or migration (Digital Preservation Testbed, 2001; Faria, 2015, Ferreira, 2006).

Emulation

Emulation entails the replacement of software and/or hardware components of the database technological stack with software that simulates the way these parts operate; for instance, the use of a virtual machine to imitate the hardware while keeping the rest of the technology stack intact (Faria, 2015; Ferreira, 2006). This strategy keeps the original environment of the database – one benefit of using

[71] For more support for preservation planning, see the *Digital Preservation Handbook*: http://www.dpconline.org/advice/preservationhandbook/organisational-activities/preservation-planning

emulation – but it could also be a disadvantage as both time and technological differences may make it difficult for the consumer to use the system in its original state (Lee, Slattery, Lu, Tang, and McCrary, 2002). Problems in the original technology may also be accidentally preserved (e.g. known security risks), and access restrictions may hinder access to consumers in the future (Thibodeau, 2002; Waugh, Wilkinson, Hills, and Dell'oro, 2000).

Emulation may also introduce problems regarding software licensing and intellectual property rights, as the operating system, the DBMS, as well as other co-existing applications, may have associated licences that deny the right to duplicate, access, or use the technology. Furthermore, emulation can be very complex to implement in contexts where databases are distributed over a network. In these instances, the whole networked setup must be preserved by the emulation environment (Thibodeau, 2002).

Migration/Normalization

The migration/normalization approach works by exporting certain properties[72] of the original database from its DBMS into another DBMS or file format more adequate for long-term preservation. This DBMS or file format must be carefully chosen to ensure the success of this strategy. The chosen DBMS or file format should be mature, open, widely adopted, well supported by the community, and transparent (Heslop and Wilson, 2002). Additionally, the chosen DBMS or file format should support the preservation requirements and future re-uses established during preservation planning (Dappert and Farquhar, 2009; Hockx-Yu and Knight, 2008). The main advantage of migration is the ability to disseminate database assets in a way that is easy for its future users to understand and re-use (Faria, 2015). The main disadvantage is the potential for data loss due to inadequate preservation formats, or less-than-perfect migration software (Ferreira, 2006).

Nonetheless, the normalization approach represents the current best practice when it comes to preserving databases. Preservation programmes for large-scale database archiving have been in place for over 10 years in countries such as Denmark and Sweden (The Danish National Archives, 2013). The scope of these programmes includes all public bodies, requiring them to notify The National Archives whenever a new information system is acquired or updated. In addition, the Archives take snapshots of running databases roughly every five years.

9.2. Standards, Best Practice, and Tools

Because database emulation and encapsulation are implemented essentially in the same way as with any other type of digital object, this report will focus on existing solutions to support migration. The following sections provide a list of preservation formats, tools, software, and services available to support the preservation of databases.

Practitioners of database preservation typically prefer simple text formats based on open standards. These include flat files, such as Comma Separated Value (CSV), annotated textual documents, such as Extended Markup Language (XML), and the international and open Structured Query Language (SQL) (ERPANET, 2003; Heuscher, Jaermann, Keller-Marxer, and Moehle, 2004). [73] These formats are not operational DBMSs, but are text-oriented containers to hold the data and other properties that have been carefully extracted from the original systems. The end-goal is to keep data in a transparent and vendor-neutral database technology and reintegrate these data into a live DBMS in the future for enhanced access (e.g. data mining).

Archival Data Description Markup Language (ADDML)

ADDML[74] is an XML-based format developed by the National Archives of Norway and Sweden to preserve databases. The format stores general metadata about the original database format, system, and

[72] As tables, columns, rows, views, users, permissions, triggers, stored procedures, etc.
[73] Standard Query Language as defined by the ISO/IEC Standard 9075
[74] http://xml.ra.se/addml/

database structure; it also holds content as plain-text files. The format can also contain and reference files in other formats (Geber, 2012).

KRAM is a software application developed by the National Archives of Sweden that, among other digital preservation features, converts databases to the ADDML format and validates some aspects of the resulting ADDML file (e.g. it checks that the data are in agreement with the metadata). This software is used to convert outdated file formats from the 1970s and 1980s to ADDML for preservation.

RALF is also a software application developed by the National Archives of Sweden that essentially acts as a downsized version of KRAM. RALF checks metadata for correctness and verifies that the data files are in accordance with the technical requirements and correspond to the metadata. RALF supports ingest of an Excel file (that complies with a specific template) to facilitate the creation of an ADDML file containing all the relevant metadata.

Standard Data Format for Preservation (SDFP)

SDFP is an umbrella format developed as part of the Migration to Intermediate XML for Electronic Data (MIXED) project carried out at DANS.[75] It contains sets of XML schemas for various significant data types and builds on existing XML representations of file formats such as the Open Document Format (ODF). SDFP will expand as new data types are added (always remaining backwards compatible). Thus the format can be used as a device for containing and accumulating knowledge on the structure of file formats. One of the main functions of SDFP is to store tabular data contained in spreadsheets and databases.

The **MIXED tool**[76] (from the project of the same name) is a service that can convert files based on tabular data formats. It currently supports the formats Data Perfect (input only), Microsoft Access 2000 and 2002, dBase III and IV, and Microsoft Excel 2003. MIXED is an online service to which tabular data files can be uploaded and then downloaded as an SDFP. The service is extendable with plugins allowing it to support additional tabular formats.

Software Independent Archiving of Relational Databases (SIARD)

SIARD[77] is an open format developed by the Swiss Federal Archives. It has been a Swiss standard since 2013 (eCH-0165), designed for archiving relational databases in a vendor-neutral form. A SIARD archive consists of a ZIP-based package containing files based mostly on XML, SQL:1999, and Unicode. A SIARD file contains metadata about the database itself, the database content, and its structure, all written in a machine-readable format. SIARD also supports the preservation of structural constraints (such as keys and triggers) and is capable of saving large database objects (BLOBs and CLOBs) as files inside the SIARD archive (Heuscher *et al*., 2004; Swiss Federal Archives, 2008).

SIARD DK is an open format and a variation on the original SIARD format. It was created by the Danish National Archives to adapt SIARD to their specific needs. It differs from the original SIARD format in the following ways:

- a different hierarchy of folders and file placement specification within the package
- specification of normalization formats
- creation of the SIARD archive as a folder instead of a zip file

[75] Data Archiving and Networked Services (DANS): http://www.dans.knaw.nl
[76] MIXED tool: https://sites.google.com/a/datanetworkservice.nl/mixed/
[77] SIARD format: http://www.digitalpreservation.gov/formats/fdd/fdd000426.shtml

The creation of the SIARD archive as a folder rather than a zip file allows distribution of large databases across multiple storage devices. These changes were introduced to increase the flexibility of the format and efficiency of the tools that use it (Danish State Archives, 2010).

SIARD 2 is the most recent update to the SIARD format and is backward compatible with the original format. SIARD 2 was developed as part of the E-ARK project by providers of digital preservation services in collaboration with the Swiss Federal Archives and the Danish National Archives (Faria, Nielsen, Röthlisberger-Jourdan, Thomas, and Voss, 2015).[78]

The **Database Preservation Toolkit**[79] is an open-source tool that allows conversion between various database systems and the SIARD preservation format. The tool also enables the conversion of a preservation format into a live DBMS. Current version 2.0 supports the conversion to and from Microsoft SQL Server, MySQL, Oracle, PostgreSQL, SIARD 1, SIARD 2 and other JDBC-supported systems. It also supports Microsoft Access as an input format and SIARD DK as an output format (Ramalho, Faria, Silva, and Coutada, 2014).

The **SIARD Suite**[80] is a free Java software application developed by the Swiss Federal Archives to simplify the archiving of relational databases. The software supports Oracle, Microsoft SQL Server, MySQL, DB/2, Microsoft Access and SIARD 1 database formats. The suite also includes SIARD Edit which supports editing SIARD 1 metadata and allows basic viewing of database content.

KOST-Val[81] is an open-source tool created by KOST-CECO to validate files in SIARD 1 format. It can also validate other non-database formats such as JPEG, JP2, TIFF and PDF/A.

Other tools

CHRONOS[82] is a commercial software application that archives databases for long-term readability and usability. The software creates ZIP archives containing database data in comma separated value (CSV) files, allowing the data to be read without the original software. CHRONOS also saves metadata in XML format. This metadata includes: object names, data types, data ranges, commentaries and constraints. CHRONOS also supports partial and ongoing database archiving (Brandl and Keller-Marxer, 2007; CSP GmbH &Co. KG, 2015; Lindley, 2013).

DeepArc[83] was developed by the National Library of France with the XQuark Group (which no longer exists) to transform relational database content into XML for archiving purposes. It is part of the IIPC[84] tool suite for web archiving. DeepArc is an open-source graphical editor which allows users to map an existing relational data model to one or several target data models, specified as XML Schemas. The purpose of this tool is to migrate database structure and content to an open-source, structured format that will create or retain the link between the document and its information.

HPAIO (HP Application Information Optimizer) is a software application created by HP to relocate inactive data from production systems and legacy databases while preserving data integrity and access. It archives data as XML or CSV documents but embeds binary image files within the XML. Over time, this approach could hinder the monitoring of these embedded objects to prevent format obsolescence (Fitzgerald, 2013).

[78] E-ARK project: http://www.eark-project.com/
[79] Database Preservation Toolkit: http://www.database-preservation.com/
[80] SIARD Suite tool: https://www.bar.admin.ch/bar/en/home/archiving/tools/siard-suite.html
[81] KOST-Val tool: http://kost-ceco.ch/cms/index.php?kost_val_de
[82] CHRONOS tool: http://www.csp-sw.de/en/inhalt.php?kategorie=c271_Solutions_CHRONOS
[83] DeepArc tool: http://sourceforge.net/projects/deeparc
[84] International Internet Preservation Consortium (IIPC): http://netpreserve.org

This list represents a selection of tools demonstrated to support the preservation of databases. To further compare and assess appropriate tools, Gartner's 'Magic Quadrant for Structured Data Archiving and Application Retirement' provides some further guidance:
https://www.gartner.com/doc/reprints?id=1-2HYPOQ8&ct=150616

9.3. Best Practice for Future Usability

Database preservation entails more than the technical operations. Robust preservation of databases requires cooperation across multiple roles within an institution.

The first step to preserving a database is to choose the best possible format, either preserving the database in its original format or migrating to an alternative format. 'Best' means the most adequate format to meet the preservation goals of the organization that needs to archive the database. This necessitates choosing a format capable of holding data and all the necessary related objects as defined by a preservation policy. The format should be able to store data in a vendor- and technology-neutral format that is widely adopted and well supported (such as XML). The database should be converted to this format by a tool that can automate most of the process.

After a database is converted, encapsulation should be performed, by adding descriptive, technical, and other relevant documentation to understand the preserved data. This step could require the help of the content producers, the database manager, or application software developers. In order to ensure the capture of sufficient metadata and documentation, those with the most knowledge of the content and structure of the database should be consulted as database preservation entails more than the technical operations. Robust preservation of databases requires cooperation across multiple roles within an institution, and potentially between multiple institutions.

Databases rarely exist just on their own, but interact with other applications that allow input and output of data. Documenting these applications and how they relate to the database, for instance how information was introduced on the application and how it was then arranged on the database, may be key information required to understand the database content. If re-use of the original database system is a priority for the organization, steps should be taken to enable emulation, such as collecting and documenting the original DBMS software, related software, the operating system, storage, and hardware.

Lastly, all databases identified for long-term archiving should be submitted to a preservation environment that will curate it over time. Future access to the preserved data should be ensured through strategies such as simple discovery and location services, or advanced querying and transformation services. To ensure that data can still be used in advanced research contexts, the ability to ingest data into a live DBMS by means of migration or emulation is essential. This capability supports more advanced data analytics and processing.

9.4. Current Limitations and Future Research

Current strategies fail to preserve the semantics of the data when provided by the separate application that creates and uses the data.

Current approaches to database preservation have a number of limitations. There is an inherent difficulty in preserving the behaviour of a database. For instance, current strategies do not typically store procedures because they are often incompatible between DBMSs. For this reason, current approaches focus on the preservation of tabular data themselves and documentation of other database properties. Current strategies fail to preserve the semantics of the data when provided by the separate application that creates and uses the data. Also, incompatibilities between DBMSs may lead to data loss

or corruption. For instance, dates in different formats may not be converted correctly or data types that do not exist in both source and target DBMSs may be lost.

Additionally, the emergence of new database models, such as the NoSQL and big data stores, creates novel challenges. Currently, there are very few research activities or practical programmes available to support the preservation and curation of these types of database.

Current and future research forecasts an increase in possible uses of archived databases. Current tools and standards are not wellequipped to cope with analysis of preserved data, but data warehousing has potential to address this challenge. This topic forms part of the initiative of the E-ARK Project, a European Commission-funded project that will deliver new approaches to integrating techniques from data warehousing, Online Analytical Processing (OLAP), data mining, and semantic annotation (Delve, Schmidt, and Aas, 2014).

Advances in new database technology (both SQL and NoSQL) and corresponding advances in methods for archiving these databases, indicate an increasingly blurry boundary between data and applications. The field of database preservation will face this challenge on several fronts as it tackles new areas of growth in the use of archived databases. The need for information governance will increase as organizations grow to rely more heavily on archived database records as evidence. Database archiving is also beginning to respond to the trend toward big data analytics and petabyte scale archives, which require the ability to search and compare large amounts of archived and current data. All of these developments will rely on the capture of elements of both data and the applications used to interpret and create data. Whether a relational database or a graph data store, effective preservation will increasingly become about archiving more than just data in order to ensure that information can be reliably understood and re-used in the future.

9.4.1. Ongoing Research: the E-ARK Project

E-ARK is a European Commission-funded collaborative project, at the time of writing, currently undertaking research into the further development of preserving these types of data. The project partners, in co-operation with commercial systems providers, aim to create and pilot a pan-European methodology for electronic document archiving. They propose to achieve this by synthesizing existing national and international best practices, which will keep records and databases authentic and usable over time. The partners plan to implement their approach in several national contexts, using existing (near-to-market) tools, and services developed by during the project.

E-ARK's hope is to provide a single, scalable, robust approach capable of meeting the needs of diverse organizations, public and private, large and small, and able to support complex data types. The outputs of E-ARK have the potential to benefit public administrations, public agencies, public services, citizens and business by providing simple, efficient access to the workflows, and enabling re-use of information.

For more information about the project and its outputs, see their website http://www.eark-project.com

10. Conclusions

Digital preservation and the adoption of standards and best practice will make the difference between troves of crude data and curated collections of rich information.

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 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 ESRC-funded Big Data Network Support has created some useful services and infrastructure to help on both 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 study already using particular datasets, to show how they can be utilized. 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 BDNS by the ESRC demonstrates growing initiative to exploit transactional data for research to improve services and policies. The networks under BDNS provide a model for how this type of research might be facilitated and supported. In particular, the work undertaken by the research centres 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 that traditionally preserve digital content (e.g. repositories, libraries, archives) have long 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.[85] 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.[86] 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 BDNS and other institutions integrate data management and preservation into their processes.

Big data and the technologies used to generate, store, and analyse them significantly alter our understanding of the roles and uses of digital content. Data that reflect human interactions have become ubiquitous as the web transitions from a space for publishing content to a dynamic space of interaction underpinned by a system of networked platforms. Consumers increasingly rely on mobile devices and web services to perform everyday activities, from shopping to banking to managing utilities. The increased availability of these data alone, however, does not necessarily lead to a valuable new source of information. Without a coordinated effort to manage and curate these data, and understand their context, their value will be very short term. Lack of early preservation action would be short-sighted.

Though some forms of data are not suited to open access, transparent practices of data collection and preservation are paramount to cultivating a culture of trust between data owners and their community of users. Transparent practices in combination with consistent methods for curation and preservation could lead to more wide-spread data-driven services and programmes that can respond rapidly to new information. In addition to data curation, methods to capture the research process applied to these data could provide insights into possible re-uses. Understanding the actions performed on data helps reveal

[85] Two examples of these accreditation frameworks are TRAC (Trustworthy Repository Audit and Certification) developed by the US RLG and NARA and the Data Seal of Approval developed by the Dutch organization DANS.
[86] For a more detailed discussion of the issue of digital repositories and trust, see Yakel, E, Faniel, IM, Kriesberg, A, and Yoon, A (2013).

the meaning and significance of research and analysis. Research in the collection, curation, and analysis of transactional data, and other forms of big data, are still underway. Institutions in a position to lead development of new uses of novel forms of data must negotiate a complex terrain of legal, organizational, and technical challenges. As research continues and infrastructure becomes more established, digital preservation and the adoption of standards and best practice will make the difference between troves of crude data and curated collections of rich information.

11. Glossary

*Definitions with an asterisk derive from the OECD Ethics Glossary, or may originate from other sources (those sources are cited in the relevant definitions).

Access point: this refers to the application or interface used by someone to view or change a database. A database may have multiple access points.

ACID properties: an acronym for atomicity, consistency, isolation, and durability, the four desirable properties of a classic transaction processing system. Each transaction is an atomic (indivisible) unit of work that fails or succeeds as a unit. The database is always in a consistent state at the start and end of a transaction; no constraints are violated. Each transaction is isolated from all the other transactions against the database. Finally, the work done by a transaction is persisted in the database (durable) when a transaction succeeds (Celko, 2014).

Analytics (data analytics): the computational analysis of large sets of data, from one or many sources, to discover patterns or trends, such as about an event, phenomenon, or demographic. Discerns information within the data not visible without the hardware and software technology to manage, store, process, analyze, and synthesize very large sets of data.

Anonymization*: a process of ensuring that the risk of somebody being identified in the data is negligible. This invariably involves more than simple de-identification, but also requires that data be altered or masked in some way in order to prevent statistical linkage.

Apache Hadoop: a framework of open-source software and file systems that allow users to save big datasets across multiple computers or using cloud technology. Hadoop breaks large datasets into smaller parts in order to distribute them to multiple machines, but a user can still access the data as a whole from a single interface. Hadoop increases processing speed, making **data analytics** much faster.

Backward compatibility: an application or technology is backward compatible if it can read and process information created by an older application or technology, such as a legacy system. This feature helps prevent the obsolescence of formats.

Cloud computing: the practice of using a network of remote servers hosted on the Internet to store, manage, and process data, rather than a local server. Cloud computing used to refer to any distributed computing over a network (Celko, 2014).

Computational analysis or analytics: *see* Analytics.

Consent*: informed consent entails giving prospective participants sufficient information about the research and ensuring that there is no explicit or implicit coercion so that they can make an informed and free decision on their possible involvement. Information should be provided in a form that is comprehensible and accessible, typically in written form (or in a form that participants can access after the end of the research interaction), and time should be allowed for the participants to consider their choices and to discuss their decision with others, if appropriate. The consent forms should be signed off by the research participants to indicate consent. (Source: ESRC Framework for Research Ethics)

CRUD: an acronym for create, read, update, and delete. These are the basic functions implemented by relational databases for persistent storage.

CSV: A comma separated values (CSV) file stores tabular data (numbers and text) in plain text. Each line of the file is a data record consisting of one or more fields, separated by commas. There is no official standard for the CSV file format, but RFC 4180 provides a de facto standard for many aspects. CSV is supported by a very large number of tools, from spreadsheets such as Excel, OpenOffice and Google Docs to complex databases to almost all programming languages. (Sources: Open Knowledge http://data.okfn.org/doc/csv and Wikipedia https://en.wikipedia.org/wiki/Comma-separated_values)

Data controller (or data holder): the person who decides the purposes for which, and the manner in which, personal data is to be processed. This may be an individual or an organization registered with the Information Commissioner's Office (ADT, 2012).

Data model (or database model): a map or plan of all the data elements and how they relate to each other. A data model supports the creation of an information system, such as a database, by defining the definition and format of the data. The consistent implementation of a data model will ensure data are compatible across applications, allowing them to share data.

Data owner*: a legal entity which has the right to give permission for rights (intellectual property) to be used by others. A 'data owner' could be an individual or a corporation.

Data science: a cross-disciplinary research approach that uses large amounts of data for analysis; 'data science' is used by government, journalism, business, academic social science, computer science, and by the humanities.

Data subject: an individual who is the subject of personal data (ADT, 2012). The term refers to the individuals represented in a dataset.

Data warehousing: A data warehouse is a relational database that is designed for query and analysis rather than for transaction processing. It usually contains historical data derived from transaction data, but it can include data from other sources. It separates analysis workload from transaction workload and enables an organization to consolidate data from several sources. In addition to a relational database, a data warehouse environment includes an extraction, transportation, transformation, and loading (ETL) solution, an online analytical processing (OLAP) engine, client analysis tools, and other applications that manage the process of gathering data and delivering it to business users (Lane, 2002).

Database as a Service (DaaS): a method of storing and managing data using databases hosted via cloud computing and accessed remotely.

Database Management System (DBMS): a software application that allows end users to create, read, update, and delete data in a database. A DBMS provides an interface between a database and end users or other software applications and maintains consistency and allows access.

Dataset: a collection of data that usually corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question.

Designated community: an identified group of potential consumers who should be able to understand a particular set of information from an archive. These consumers may consist of multiple communities, are designated by the archive, and may change over time (OAIS term). (Source: Digital Preservation Handbook, http://handbook.dpconline.org/glossary#D)

Disclosure (accidental disclosure)*: disclosure relates to the inappropriate attribution of information to a **data subject**, i.e. an individual person or organization represented in a set of data. Disclosure has two components: identification and attribution. (Source: OECD Expert Group for International Collaboration on Microdata Access: *Final Report*)

Hadoop: *see* Apache Hadoop

Online transaction processing (OLTP): provides support for daily business applications. This is the niche that SQL has in the commercial market (Celko, 2014).

Open data*: data (datasets) that are: 1) accessible to anyone and everyone, ideally via the Internet; 2) in a digital machine-readable format that allows interoperation with other data; 3) available at reproduction cost or less; and 4) free from restrictions on use and re-use. (Source: OECD Expert Group for International Collaboration on Microdata Access)

Persistent identifier: a long-lasting reference to a digital resource. Typically it has two components: a unique identifier; and a service that locates the resource over time even when its location changes. (Source: Digital Preservation Handbook, http://www.dpconline.org/advice/preservationhandbook/technical-solutions-and-tools/persistent-identifiers)

Privacy*: someone's right to keep their personal matters and relationships secret, involving an obligation of the holder of information to the subject of the information to do so. (Source: UNECE, Principles and Guidelines on Confidentiality Aspects of Data Integration Undertaken for Statistical or Related Research Purposes, 2009)

Query language: a computer languages used to interact with a database or information system. Structured Query Language (SQL) is the most common language used with relational databases.

Re-deploy: the ability to run archived software or programming on preserved software or emulated software, such as a virtual machine.

Sandbox: a testing environment that isolates untested code changes. Sandboxing protects 'live' servers and their data from changes that could be damaging (regardless of the intent of the author of those changes) or which could simply be difficult to revert. (Source: Wikipedia, https://en.wikipedia.org/wiki/Sandbox_(software_development))

Screencast: a digital recording of computer screen output, also known as a video screen capture, often containing audio narration. The term screencast differs from the related term 'screenshot'; in that while a screenshot generates a single picture of a computer screen, a screencast is essentially a movie of the changes over time that a user sees on a computer screen, enhanced with audio narration. (Source: Wikipedia, https://en.wikipedia.org/wiki/Screencast)

Server-side scripting: a technique used in web development which involves employing scripts on a web server which produce a response customized for each user's (client's) request to the website. (Source: Wikipedia, https://en.wikipedia.org/wiki/Server-side_scripting)

Significant properties: characteristics of digital and intellectual objects that must be preserved over time in order to ensure the continued accessibility, usability and meaning of the objects and their capacity to be accepted as (evidence of) what they purport to be. http://www.significantproperties.org.uk. (Source: Digital Preservation Handbook, http://handbook.dpconline.org/glossary#S)

Stack: the different components required for running a database – the DBMS, an operating system, the database, and any other software applications used to operate the database.

Structured Query Language (SQL): a standard language for accessing and manipulating databases. (Source: W3schools: http://www.w3schools.com/sql/sql_intro.asp)

Tabular data: data that are structured into rows, each of which contains information about something. Each row contains the same number of cells (although some of these cells may be empty), which provide values of properties of the thing described by the row. In tabular data, cells within the same column provide values for the same property of the things described by each row. This is what differentiates tabular data from other line-oriented formats. (Source: W3C, https://www.w3.org/TR/tabular-data-model/#model)

Virtual machine: an emulation of real or hypothetical computer hardware. Examples: VMware, VirtualBox.

Web harvesting (or crawling): the act of browsing the web automatically and methodically to index or download content and other data from the web. The software to do this is often called a web crawler. (Source: Digital Preservation Handbook, http://handbook.dpconline.org/glossary#C)

XML (eXtensible Mark-up Language): a widely used application-independent mark-up language for encoding data and metadata.

12. References

Association for Data-driven Marketing and Advertising (ADMA) 2013, *Best Practice Guideline: Big Data*, http://www.admaknowledgelab.com.au/compliance/compliance-help/general/data-and-privacy/codes-and-guides/best-practice-guideline-big-data

Administrative Data Taskforce 2012, 'The UK Administrative Data Research Network: Improving Access for Research and Policy', http://www.esrc.ac.uk/files/publications/themed-publications/improving-access-for-research-and-policy

Brandl, S, Keller-Marxer, P 2007, 'Long-term archiving of relational databases with Chronos', First International Workshop on Database Preservation (PresDB'07).

Celko, J 2014, *Joe Celko's Complete Guide to NoSQL: What Every SQL Professional Needs to Know about Non-Relational Databases*, MA: Morgan Kaufmann, DOI: 10.1016/B978-0-12-407192-6.09991-X.

Corti, L, Bolton, S, and Moody, V 2015, 'Effective data curation for big data', *Data Impact Blog*, 1 November 2015, http://blog.ukdataservice.ac.uk/696-2

CSP GmbH & Co. KG 2015, 'Declaration of Incorporation', CHRONOS Version 4.9.

Danish National Archives 2013, 'Strategy for archiving digital records at the Danish National Archives', https://www.sa.dk/en/wp-content/uploads/sites/2/2014/12/Strategy-for-archiving-digital-records-2013.pdf

Danish State Archives 2010, 'Bekendtgørelse om arkiveringsversioner' ('Executive Order on Submission Information Packages'), https://www.sa.dk/en/wp-content/uploads/sites/2/2014/12/Executive-Order-on-Submission-Information-Packages-Danish-national-standard.pdf

Dappert, A, Farquhar, A 2009, 'Significance is in the eye of the stakeholder', Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5714 LNCS, 297–308. DOI: http://doi.org/10.1007/978-3-642-04346-8_29

Delve, J, Schmidt, R, & Aas, K 2014, 'Long-term preservation of databases the meaningful way', DLM Forum Triennial Conference, Lisbon, Portugal, 10–14 November.

Digital Preservation Testbed 2001, 'Migration: Context and Current Status', The Hague http://kifri.fri.uniza.sk/~chochlik/diz_doc/sources/data_migr/Migration.pdf

Directorate-General for Research & Innovation 2016, 'Guidelines on Data Management in Horizon 2020', European Commission, http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

ERPANET 2003, 'The long-term preservation of databases', ERPANET Workshop Report, Bern.

Faria, L 2015, 'Automated Watch for Digital Preservation', University of Minho.

Faria, L, Nielsen, AB, Röthlisberger-Jourdan C, Thomas, H, and Voss, A 2015, eCH-0165 SIARD Format Specification 2.0 (draft).

Ferreira, M 2006, 'Introdução à Preservação Digital: Conceitos, estratégias e actuais consensos', https://repositorium.sdum.uminho.pt/bitstream/1822/5820/1/livro.pdf

Fitzgerald, N 2013, 'Using data archiving tools to preserve archival records in business systems – a case study',

http://purl.pt/24107/1/iPres2013_PDF/Using%20data%20archiving%20tools%20to%20preserve%20archival%20records%20in%20business%20systems%20%E2%80%93%20a%20case%20study.pdf

Fowler, Martin 2012, 'NosqlDefinition', *Martin Fowler bliki*, 9 January 2012, http://martinfowler.com/bliki/NosqlDefinition.html

Geber, M 2012, 'Transfers and Preservation of E-archives at the National Archives of Sweden', http://www.gosbook.ru/system/files/documents/2012/11/13/geber_m.pdf

Gorman, M 2005, 'Is SQL A Real Standard Anymore?', *The Data Administration Newsletter* (TDAN.com), Ed. Robert S. Seiner, http://www.wiscorp.com/is_sql_a_real_standard.pdf

Heslop, H, and Wilson, A 2002, 'An Approach to the Preservation of Digital Records', National Archives of Australia, http://www.naa.gov.au/Images/An-approach-Green-Paper_tcm16-47161.pdf

Heuscher, S, Jaermann, S, Keller-Marxer, P, and Moehle, F 2004, 'Providing Authentic Long-term Archival Access to Complex Relational Data', Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data, October, 18, https://arxiv.org/abs/cs/0408054

Hockx-Yu, H and Knight, G 2008, 'What to Preserve?: Significant Properties of Digital Objects', *International Journal of Digital Curation*, 3(1), 141–153, DOI: http://doi.org/10.2218/ijdc.v3i1.49

Hoeren, T, Kolany-Raiser, B, Yankova, S, Hecheltjen, M, Hobel, K 2013, *Legal Aspects of Digital Preservation*, Cheltenham: Edward Elgar Publishing Ltd. ISBN: 978 1 78254 665 8

Hughes, R 2016, Agile Data Warehousing for the Enterprise: A Guide for Solutions Architects and Project Leaders, DOI: 10.1016/B978-0-12-396464-9.00020-5

Information Commissioners Office (ICO) 2011, 'Data sharing code of practice', https://ico.org.uk/media/for-organizations/documents/1068/data_sharing_code_of_practice.pdf

Johnston, D and Henderson-Ross, J 2012, 'The New Data Values: Securing customer data as a renewable resource', Aimia Insights, http://www.aimia.com/content/dam/aimiawebsite/CaseStudiesWhitepapersResearch/english/WhitepaperUKDataValuesFINAL.pdf

Lake, P and Crowther, P 2013, 'Concise Guide to Databases: A Practical Introduction', London: Springer, DOI: 10.1007/978-1-4471-5601-7

Lane, P 2002, *Oracle9*i *Data Warehousing Guide*, Release 2 (9.2), https://docs.oracle.com/cd/B10500_01/server.920/a96520.pdf

Lappin, J 2011, 'The challenges of archiving databases – Podcast with Kevin Ashley', 9 July 2011, *Records Management Today* podcast series, hosted by Northumbria University, http://thinkingrecords.co.uk/2011/07/09/the-challenges-of-archiving-databases-podcast-with-kevin-ashley/

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

Lee, K, Slattery, O, Lu, R, Tang, X, and McCrary, V 2002, 'The State of the Art and Practice in Digital Preservation', *Journal of Research of the National Institute of Standards and Technology*, 107(1), 93–106, DOI: http://doi.org/10.6028/jres.107.010

Lindley, A 2013, 'Database Preservation Evaluation Report – SIARD vs. CHRONOS', Lisbon Technical University (IST), Portugal, 2–5 September,

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*, http://blog.ukdataservice.ac.uk/thinking-big-making-the-most-of-big-data-in-scientific-research

Müller, H 2009, 'Database Archiving', *DCC Briefing Papers: Introduction to Curation*, Handle: 1842/3346, http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation

Organization for Economic Co-operation and Development (OECD) February 2013, 'New Data for Understanding the Human Condition', *OECD Global Science Forum Report*, http://www.oecd.org/sti/sci-tech/new-data-for-understanding-the-human-condition.pdf

Public Health Research Data Forum (PHRDF) 2015, 'Enabling Data Linkage to Maximise the Value of Public Health Research Data: full report', http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtp059017.pdf

Ramalho, JC, Faria, L, Silva, H, and Coutada, M 2014, 'Database Preservation Toolkit : a flexible tool to normalize and give access to databases', http://purl.pt/26107/1/DLM2014_PDF/15%20-%20Database%20Preservation%20Toolkit.pdf

Rhind, D 2014, 'Drowning in Data: Who and what can we trust?', Advisory Panel on Public Sector Information, http://www.nationalarchives.gov.uk/documents/meetings/20140425-drowning-in-data.pdf

Sadalage, P 2014, 'NoSQL Databases: An Overview', ThoughtWorks Insights, https://www.thoughtworks.com/insights/blog/nosql-databases-overview

Schweers, S, Kinder-Kurlanda, K, Müller, S, and Siegers, P 2016 (forthcoming), '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, http://edpl.lexxion.eu/article/EDPL/2015/2/4

Swiss Federal Archives 2008, SIARD Format Description.

Thibodeau, K 2002, 'Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years', *The State of Digital Preservation: An International Perspective*, Council on Library and Information Resources.

UK Data Archive (UKDA) 2014a, 'Preservation Policy', http://www.data-archive.ac.uk/media/54776/ukda062-dps-preservationpolicy.pdf

UK Data Archive (UKDA) 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, http://doc.ukdataservice.ac.uk/doc/7591/mrdoc/pdf/7591_edrp_accompanying_documentation.pdf

Walkowiak, S and Moody, V 2012, 'The power of R: methods for processing big data', *Data Impact blog*, 9 January 2012, http://blog.ukdataservice.ac.uk/the-power-of-r-methods-for-processing-big-data

Waugh, A, Wilkinson, R, Hills, B, and Dell'oro, J 2000, 'Preserving digital information forever', 5th Conference on Digital Libraries, 175–184, DOI: http://doi.org/10.1145/336597.336659

Yakel, E, Faniel, IM, Kriesberg, A, and Yoon, A 2013, 'Trust in Digital Repositories', *International Journal of Digital Curation*, *8*:1, 143–156, DOI: 10.2218/ijdc.v8i1.251