IASSIST Quarterly

VOLUME 40 – Number 4– 2016

More Data, Less Process? The Applicability of MPLP to Research Data

MMRepo - Storing qualitative and quantitative data

Servicing New and Novel Forms of Data

IN THIS ISSUE

Digital and Huge

Online at: *iassistdata.org/iq*

COLOPHON

IASSIST Quarterly

The **IASSIST Quarterly** represents an international cooperative effort on the part of individuals managing, operating, or using machine-readable data archives, data libraries, and data services. The **QUARTERLY** reports on activities related to the production, acquisition, preservation, processing, distribution, and use of machine-readable data carried out by its members and others in the international social science community. Your contributions and suggestions for topics of interest are welcomed. The views set forth by authors of articles contained in this publication are not necessarily those of **IASSIST**.

Information for Authors

The Quarterly is normally published four times per year. Authors are encouraged to submit papers as word processing files (for further information see: *http://www.iassistdata.org/iq/instructions*-authors). Manuscripts should be sent to Editor: Karsten Boye Rasmussen.: Email: *kbr@sam.sdu.dk*

.Announcements of conferences, training sessions, or the like are welcomed and should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event

Editor

Karsten Boye Rasmussen
Department of Marketing & Management
University of Southern Denmark, SDU
Campusvej 55, DK-5230
Odense M, Denmark
Phone: +45 6550 2115
Email: *kbr@sam.sdu.dk*

Deputy editor

Walter Piovesan
Simon Fraser University

Website editor

Michele Hayslett
University of North Carolina at Chapel Hill

In this issue

Title: Newsletter - International Association for Social Science Information Service and Technology
ISSN - United States: 0739-1137

Editor's Notes

When things get digital and huge. Doing the things right and doing the right things.

Welcome to the fourth issue of Volume 40 of the IASSIST Quarterly (IQ 40:4, 2016).

There is a lot of management involved in the data management carried out at data archives and with data collections. The phrase 'Doing the things right and doing the right things' belongs to fathers of modern management and is used to distinguish management vs. leadership, efficiency vs. effectiveness, and tactics vs. strategy. The winning authors of the 2016 IASSIST paper competition used the article 'More Product, Less Process: Revamping Traditional Archival Processing' (Mark A. Greene and Dennis Meissner, 2005) as their starting point for investigating the 'More Product, Less Process' (MPLP) approach for digital data. The winning paper 'More Data, Less Process? The Applicability of MPLP to Research Data' is written by Sophia Lafferty-Hess and Thu-Mai Christian. The authors work at the Odum Institute for Research in Social Science, University of North Carolina at Chapel Hill as Research Data Manager and Assistant Director of Archives. The paper was presented in the session 'Data Management Archiving/Curation Platforms' at the IASSIST 2016 conference in Bergen.

In their paper Lafferty-Hess and Christian set out to apply the principles and concepts formulated in MPLP to the archiving of digital research data. They discuss data quality, usability, preservation and access, leading to the question: What is the 'golden minimum' for archiving digital data? In terms of data archiving, spending too much effort on doing the things right may bring the trade-off problem that the resources are not sufficient to do all the things. Users in the digital world retrieve and consume lots of information by themselves, but digital data comes in forms that are seldom directly consumable without additional processing. When the authors also relate the phrase 'golden minimum' to the phrase 'good enough', management is again brought into the discussion. In my view, the short formulation of Herbert Simon's 'satisficing' concept in his theory of bounded rationality is 'good enough is best'. Lafferty-Hess and Christian are aware that shifting responsibility for certain data curation tasks from the data archive to the data producer and to the data user can present problems. Their best advice and hope for the future is that additional 'future research will help us build better understanding of the connection between user needs and data curation processes'.

The following paper 'MMRepo - Storing qualitative and quantitative data into one big data repository' is authored by Ingo Barkow, Catharina Wasner and Fabian Odoni, working at University of Applied Sciences Eastern Switzerland HTW Chur where Barkow is Associate Professor and Wasner and Odoni are research associates. They describe a prototype of their MMRepo project that addresses the problem of storing qualitative large binary objects with regular quantitative data in order to achieve the advantage of storing mixed mode data in the same infrastructure, whereby only one system needs to be provided and maintained. Linking to the first paper they are looking into the efficiency problem of doing the things right. When you are efficient you can do more things right. The project is trying to achieve this by combining CERN's Invenio portal with a Hadoop 2.0 cluster and DDI 3.3. The prototype was successful and the project continues. The paper was presented at the IASSIST 2016 conference in the session 'Technical Data Infrastructure Frameworks'.

Aidan Condron works with the Big Data Network Support team at the UK Data Service. At the IASSIST 2016 conference he presented 'Data Science: The Future of Social Science?' at the session 'Big Data, Big Science', and has submitted this presentation as the paper 'Servicing New and Novel Forms of Data: Opportunities for Social Science'. These 'new and novel' forms are, for example, social media data that present potential resources for researchers but also pose challenges for access provision and analysis. The paper introduces Data Service as a Platform (DSaaP), which is a project to establish technological infrastructure support. As with the MMRepo project, the DSaaP project will include both familiar and new and novel forms of data. The novel forms of data are often huge, and 'Hadoop' solutions are also at play here using a data lake built through use of open source software. The article also gives several demonstrations through graphs of energy consumption based on 3.7 billion datapoints. After the presentation and the paper, the Big Data Network Support team will standardise and generalise the procedures developed from their DSaaP project.

Submissions of papers for the IASSIST Quarterly are always very welcome. We welcome input from IASSIST conferences or other conferences and workshops, from local presentations or papers especially written for the IQ. When you are preparing a presentation, give a thought to turning your one-time presentation into a lasting contribution. We permit authors 'deep links' into the IQ as well as deposition of the paper in your local repository. Chairing a conference session with the purpose of aggregating and integrating papers for a special issue IQ is also much appreciated as the information reaches many more people than the session participants, and will be readily available on the IASSIST website at http://www.iassistdata.org.

Authors are very welcome to take a look at the instructions and layout:
http://iassistdata.org/iq/instructions-authors

Authors can also contact me via e-mail: *kbr@sam.sdu.dk*. Should you be interested in compiling a special issue for the IQ as guest editor(s) I will also be delighted to hear from you.

Karsten Boye Rasmussen
July 2017
Editor

More Data, Less Process? The Applicability of MPLP to Research Data

by Sophia Lafferty Hess[1], Thu-Mai Christian[2]

Abstract

In their seminal piece, "More Product, Less Process: Revamping Traditional Archival Processing," Greene and Meissner (2005) ask archivists to reconsider the amount of processing devoted to collections and instead commit to the More Product, Less Process (MPLP) 'golden minimum.' However, the article does not specifically consider the application of the MPLP approach to digital data. Data repositories often apply standardized workflows and procedures when ingesting data to ensure that the data are discoverable, accessible, and usable over the long-term; however, such pipeline processes can be time consuming and costly. In this paper, we will apply the principles and concepts outlined in MPLP to the archiving of digital research data. MPLP provides a useful lens to discuss questions related to data quality, usability, preservation, and access: What is the 'golden minimum' for archiving digital data? What unique properties of data affect the ideal level of processing? What level of processing is necessary to serve our patrons most effectively? These queries will contribute to the discussion surrounding how data repositories can develop sustainable service models that support the increasing data management needs of the research community while also ensuring data remain discoverable and useable for the long-term..

Keywords

data curation, data reuse, data quality, data service

Introduction

While Meissner and Greene's (2005) seminal article, More Product, Less Process: Revamping Traditional Archiving Processing, was written with traditional archives in mind, the authors' appeal for a critical assessment and recalibration of archival processing is no less relevant to digital data archives. Soon after the article was published, the new MPLP doctrine became the cause célèbre for much of the archival community, which the authors took to task for its proclivities toward all-or-nothing archival processing that had exacerbated growing backlogs of unprocessed and therefore inaccessible materials. MPLP reinstated user access as the highest of archive priorities, which absolved archivists of the minutiae of item-level arrangement, description, and preservation.

> What constitutes data quality has much to do with users' needs and preferences for discovering, accessing, interpreting, and using data.

For data archives, however, user access is very much tied to the minutiae. The usability of data is inextricable from its specific context: the research question the data were intended to answer, the instruments used to collect the

data, the software programs executed to manipulate and analyze the data, and the methods employed for data collection and analysis (Borgman, 2015). Data are complex, and the archival processing—or what we equate to data curation—required to make them available and usable for researchers has been informed by uncompromising standards of quality. Achieving these quality standards requires data archives to complete a laundry list of skill-intensive, labor-intensive, and time-intensive data curation tasks including normalizing file formats, mitigating confidentiality risks, checking for and correcting data errors, generating and enhancing descriptive metadata, assembling contextual documents, recording checksums, defining undefined variable and value codes, reconciling discrepancies between datasets and codebooks, and so on (Peer, Green, & Stephenson, 2014). Not so different from the backlog situation in traditional archives, compromises in data curation processes are inevitable given the nature of tightly resourced environments in which many data archives operate.

With increasing demand from funding agencies, journal publishers, and research communities for public access to quality data, data archives (and institutional repositories that are becoming ad hoc data archives) must examine the economies of curating archival data collections to the highest degree of quality at scale while keeping user needs at the forefront of curation approaches. The application of MPLP to data curation raises several essential questions about what compromises are allowable, if not inevitable, that will enable data archives to remain solvent while continuing to serve the needs of the user community.

In this paper, we discuss the Odum Institute Data Archive's application of the principal concepts of MPLP to data curation as part of an exercise to assess the scalability of data archive services as demand for them increases. MPLP offers a methodology that enabled us to not only reconsider, but also reaffirm the parameters of the standard data curation processes we use to provide access to quality data.

Data Quality Standards

One of Meissner and Greene's main criticisms of archival processing that obliged them to formulate their MPLP approach is "...the persistent failure of archivists to agree in any broad way on the important components of records processing and the labor inputs necessary to achieve them" (p. 209). This criticism cannot be wholly directed at data archivists, who have achieved some consensus on the important components of data curation as demonstrated in documented best practices that have received wide acceptance among data archives (ICPSR, 2012; Digital Curation Centre, n.d.). These best practices prescribe specific data curation actions that support data quality standards.

What constitutes data quality has much to do with users' needs and preferences for discovering, accessing, interpreting, and using data. Based on results from a study of users' perceptions of data quality, Wang and Strong (1996) identified four dimensions of data quality: 1) intrinsic, referring to the accuracy and credibility of the data; 2) contextual, or the relevancy of the data to the user's goals; 3) representational, relating to the ability to interpret and use the data; and 4) accessibility, or the ability to obtain the data. A more recent study conducted by Faniel, Kriesberg, and Yakel (2015) to determine the factors that elicit social science researchers' satisfaction with data reuse found that users associate data quality with attributes that align with Wang and Strong's quality dimensions. They include completeness (contextual), accessibility (accessibility), ease of operation (representational), and credibility (intrinsic) of the data. These aspects of data quality are often summed up in the notion of data being 'independently understandable' to their intended users (CSSDS, 2012; King, 1995; Lee, 2010; Peer, Green, & Stephenson, 2014).

This is the quality standard to which research data are being held, particularly those subject to data management and sharing mandates that have grown in popularity among funders and journals (e.g., National Endowment of the Humanities, 2012; National Institutes of Health, 2003; National Science Foundation, 2010; Nature, 2015; PLOS, 2014; Science, n.d.). To some degree, this standard also acknowledges the recent scrutiny of published scientific studies, a concerning number of which were reported to have failed to meet the reproducibility benchmark of scientific integrity (Chang & Li, 2015; Freedman et al., 2015; Open Science Collaboration, 2015). In response, the scientific community has called for greater research transparency, which carries the presumption that data underlying published findings are not only shared, but shared in professional data archives that have the expertise and infrastructure to ensure that data are independently understandable to the research community (DA-RT, 2015; Center for Open Science, 2015).

Data archives have long accepted the charge from the scientific community to meet data quality standards to support research transparency. For years, data archives have instituted baseline protocols for acquiring data submissions, preparing data materials for repository ingest, and providing access to usable dataset files based on archival standards for trustworthy repositories. The Reference Model for an Open Archival System (OAIS) is something of a magna carta of archival standards, informing the processing approaches of many data archives (Lee, 2010). OAIS provides a framework of high-level concepts for understanding the requirements for long-term preservation and access of materials. Fundamental to OAIS is the concept of the 'Designated Community,' which is defined as an "identified group of potential Consumers who should be able to understand a particular set of information" (CSSDS, 2012, p. 1-11). In accordance with OAIS, data archives are responsible for giving access to materials that meet the 'independently understandable' criterion for data quality. Meeting this criterion requires that data packages held in data archives include sufficient information for users to apprehend the content, context, and structure of the data, as well as information regarding the unique identity, original source, and allowable uses of the data. What this has meant in practice for the Odum Institute is that, even beyond the various automatic ingest processes executed by archival system technologies, the data archivist is responsible for performing an assortment of critical data curation tasks. Table 1 provides a complete illustration of the Odum Institute data curation pipeline.

This skill-, time-, and resource-intensive data curation is similar to Peer, Green, and Stephenson's (2014) data quality review adopted by Yale University's Institution for Social and Policy Studies (ISPS) and the data curation pipeline employed at the Inter-university Consortium

for Political and Social Research (ICPSR) (Vardigan, 2007). This high-level, or maximal, data curation approach involves an exhaustive list of processing actions that are possible to execute only in small-scale operations as in the case of ISPS, or for well-resourced operations such as ICPSR. For data archives for which neither category applies, maximal data curation may not be feasible and/or sustainable even though our users require it. Here we arrive at an impasse where we need to confront problems of expectation management and resource management as we weigh user requirements for data quality against data archive capabilities.

DATA CURATION PIPELINE

STANDARD CURATION	• Review the dataset file package to ensure all components necessary to describe and interpret the data are present (i.e., codebook, instruments, reports, etc.) • Build the document set (i.e., construct codebooks, locate external documents) • Review data for confidentiality risks • Review data for errors (i.e., wild or out-of-range codes, missing or inconsistent variables, undefined missing values) • Perform data cleaning operations to anonymize data, correct data errors and inconsistencies, and standardize missing values	• Assign a persistent identifier (i.e., DOI) • Apply standard vocabulary • Generate standard DDI metadata to include methodological information and links to associated publications • Add full variable and value label text to dataset	• Normalize files to non-proprietary, software-agnostic preservation formats • Generate derivative files for widely-used software platforms
REPLICATION VERFICATION	• Review the replication data materials for completeness (i.e., README file, code file, etc.) • Review the code for inclusion of commands and comments required for execution • Execute code and compare results to the tables and figures in the manuscript	• Link the replication dataset to the published article	

Table 1. Odum Institute Data Curation Pipeline

More Data, Less Process?

Resolving the problems of expectation management and resource management is at the core of MPLP, which "...can help archivists make decisions about balancing resources so as to accomplish their larger ends and achieve economies in doing so..." (Meissner & Greene, 2010, p. 176). MPLP petitions archivists to pursue the 'good enough,' or 'golden minimum,' in archival processing work, which gives permission to archivists to spend the minimum amount of effort necessary to serve users' needs. Anything beyond the minimum must have "clearly demonstrable business reasons" (p. 240). However, what is 'good enough' for traditional archives may not be 'good enough' for data archives.

To determine what level of processing is considered 'good enough,' MPLP directs archivists to examine three primary task areas: arrangement, description, and preservation. In traditional archives, arrangement refers to the organization of files into physical and intellectual collections in order to preserve the context of the files' creation as well as the order of the files as they were created. Description provides detailed information about the context, characteristics, and content of the materials to allow users to discover them and evaluate their relevance. Preservation deals with the long-term maintenance and protection of materials (Society of American Archivists, n.d.). Though the materials MPLP refers to differ from data objects, these archival processing activities do have their equivalents in data curation.

Arrangement

In MPLP, finding the 'good enough' in arrangement tends towards deliberations over re-labeling and re-foldering archival materials, and whether or not doing so for individual objects is necessary to fulfill the intended purpose of arrangement. According to Meissner and Greene, as well as other foundational texts on archival practices, arrangement is a way to organize materials both physically and intellectually in a way that preserves their context (Society of American Archivist, n.d.). MPLP disputes the meticulousness with which some archivists organize and apply labels to individual objects. Rather than impulsively engaging in such "overzealous housekeeping, writ large" (p. 241), MPLP insists that the archivists discharge themselves of such object-level physical arrangement, which contributes little to users' understanding of the context. Greene and Meissner wrote: "If a user is given an understanding of the whole and the structure and identity of its meaningful parts, then the vagaries that occur within a folder will not prove daunting, and probably not even confusing" (p. 241).

In applying MPLP recommendations to arrangement of data collections, primary focus is on the 'understanding of the whole,' which, for data, is an understanding of their context. This context is contained in codebooks that define each variable and value code; documented data collection instruments such as interview or survey protocols; methodology reports containing comprehensive information on data collection, cleaning and analysis procedures; links to related research products including publications citing the data; and the programming code used to execute data analysis. Arrangement of data is ensuring the presence of these materials and the sufficiency of the information contained in these materials so that the data are 'independently understandable.' Where any of these documents do not exist, we might construct them from scratch, a cumbersome practice of stitching together information from the data producer, related publications, or any other sources that offer useful clues about the context of the data. Data archivists might also insist on performing a meticulous variable-by-variable check of the dataset file to identify and correct errors and inconsistencies.

MPLP questions how much of this attention and diligence to contextual materials is necessary for users' understanding of the whole. Reviewing datasets and correcting coding errors is as, if not more, tedious as shuffling documents among folders. Assembling and copyediting supplemental documents for a dataset is not such a far cry from the meticulous practice of re-labeling folders. Instead, processing approaches for archival arrangement for data should keep focus on the goal of 'understanding of the whole' and determine what the fundamental requirements are for achieving that goal. What is most critical for understanding data is having the information necessary to decipher cryptic variable names and undefined value codes. Data archives may need to reconsider the benefits to users of providing additional and/or enhanced supplementary materials and performing variable-by-variable checks against the amount of resources the archive has to commit to these practices.

Description

As is the case for any type of archival material, the primary purpose of description is to assist users in discovering and accessing materials of interest. MPLP suggests that archivists provide enough information to afford users 'decent access' without expending extra effort on composing lengthy descriptions of an individual object or its context. MPLP discourages verbosity in description, which is considered gratuitous and does not necessarily lend itself to an increase in users' understanding of the materials or their location.

Description as it is performed in the data curation pipeline involves applying standard vocabularies, generating metadata, enhancing variable labels, and assigning a persistent identifier for the data. The generation of standardized Data Documentation Initiative (DDI) metadata for data discovery is extended to include both methodological and contextual details extracted from the document set. While this provides robust metadata for search and discovery, MPLP asks us to consider whether it is perhaps 'good enough' to provide basic discovery metadata without taking the time to incorporate these methodological details such as sample size, weighting procedures, and other contextual information that is available within supplementary documents. Greene and Meissner make the point that as archivists it is not our job to do the research for our patrons, and efficiencies could potentially be gained by minimizing the generation of metadata.

Generating metadata and assigning a persistent identifier also underlie the creation of a stable data citation. Data citation is an essential practice for not only ensuring data producers receive appropriate attribution but also providing persistent access to data, documentation, and code. The Joint Declaration of Data Citation Principles (2014) communicates the importance of data citation as a scholarly practice and provides information on the purpose, function, and attributes of data citations. These principles highlight the role data archives play in the creation of data citations by generating metadata and assigning persistent identifiers and reaffirm this as an essential curation practice.

The other key description task within our pipeline is the enhancement of variable level metadata. For social science survey data, this often takes the form of adding the complete question text to variables. This variable level description allows for much more detailed and comprehensive discovery, examination, and analysis of the data within the repository platform. However, the MPLP model would suggest against this 'item level' description as a processing benchmark and would instead suggest archivists focus on describing the materials as a whole. Understanding how researchers interact with and use repository metadata would help us understand what is 'good enough' for description. Repositories have employed usability testing to inform interface design and the expansion of platform functionalities (Gibbs et al., 2013), an extension of these types of studies could increase our understanding of what metadata fields are most useful to researchers and provide additional evidence for how best to serve users' access needs.

Preservation

Because our focus is on the work of the archivist, a discussion of archive systems technology that are required to effectively preserve data is beyond the scope of this paper. While much of archival preservation actions take place within technological systems, there are

some preservation tasks that archivists perform. Since digital materials are far more fragile than analog materials (Rothenberg, 1999), we concede that MPLP's 'good enough' does not apply as readily to preservation of digital data. However, a consideration of MPLP in our examination of activities in the data curation pipeline--file normalization and optimization--that support long-term preservation allows us to identify potential efficiencies.

Normalizing files into open or preferred file formats allows files to remain accessible and protects against obsolescence. Without normalizing files, data may become unreadable and therefore unusable. A paramount requirement when serving users is ensuring digital material remain accessible into the future; therefore, normalization can be seen as an essential processing practice. In some cases, multiple different derivative copies of a dataset may also be created to allow expanded access to the data. While this increases the dissemination of the data and facilitates reuse, one file normalized into a non-proprietary file format would serve basic user requirements. Although the researcher would then have to read the file into his or her preferred software and variable-level metadata stored within the software package would be lost, as long as that contextual information is available within accompanying documentation then researchers would still be able to fundamentally understand the data. The original file in the proprietary format may also be made available alongside the preservation copy. Perhaps 'good enough' is creating a single non-proprietary version of the data file.

Another possible option includes shifting the burden to the data depositor and only accepting certain file formats for inclusion within the repository. For instance, guideline two of the Data Seal of Approval (2013) states that "the data producer provides the data in formats recommended by the data repository" (p. 12). This guideline shows how a DSA-certified data repository at a minimum must provide recommendations for appropriate file formats but the onus may be placed upon the data producer to comply. Another more automated solution can be seen in certain repository software platforms, such as the Dataverse, that generate a derivative preservation copy for certain data file types upon ingest (Crosas, 2011). An expansion of these types of system functionalities could also lessen the processing burden.

Good Enough" For Data Curation

A reconsideration of primary data curation activities has helped to identify those activities that are essential to ensuring access to quality data. For each of the three processing task areas, there are some activities that, if not performed, will likely make it impossible for users to discover, interpret, and use the data. We offer this 'minimal curation' model as a point of reference for which to engage the data archives community in a discussion on the necessity of intensive data curation processes for supporting data quality and reuse, particularly for tightly resourced environments.

In the MPLP-based 'minimal curation' model (see Table 2), arrangement is reduced to the single task of data file package review, description requires only metadata generation and persistent identification, and preservation is limited to file normalization.

"MINIMAL" DATA CURATION PIPELINE

ARRANGEMENT	DESCRIPTION	PRESERVATION
• Review the dataset file package to ensure all components necessary to describe and interpret the data are present (i.e., codebook, instruments, reports, etc.)	• Assign a persistent identifier (i.e., DOI) • Generate basic descriptive standard DDI metadata for discovery and access	• Normalize files to non-proprietary, software-agnostic preservation formats

Table 2. Proposed Minimal Data Curation Pipeline

Arrangement

Most important to arrangement is ensuring that necessary documentation is included within the data package so that users can understand the context of the data. Whether this documentation takes on the form of a codebook or survey instrument, or some other format, at a minimum documentation should define variables and values and provide some indication of the research methodology and process, for which links to external publications may be sufficient. No longer part of arrangement in the minimal data curation scheme is the variable-by-variable review of the data to identify and remedy errors, discrepancies, and/or sensitive information in the data. By scaling back on the comprehensive data review to this degree, the archive may no longer be able to guarantee the quality of the minutiae of every dataset. In some cases, variables and values may be left undefined, missing values inconsistently or incorrectly coded, and sensitive variables in the dataset might be awaiting unauthorized disclosure.

Certainly, these compromises have the potential to impact overall usability; however, the goal of 'minimal curation' is to ensure that enough information is present for users to understand and interpret the data as a whole. The users still have contextual information available to them so as to assess the overall credibility of the data and to determine whether or not the data are relevant to them. The presence of variable and value definitions in codebooks enables users to make necessary corrections in the data. It is also

not unreasonable to set policies that make data producers responsible for removing sensitive information in their data files and users responsible for reporting the presence of sensitive data. Should the data archiving and research community determine that comprehensive variable level review is the 'golden minimum' for data, then we must also provide "clearly demonstrable business reasons" that dictate this additional task and take into account the additional resources that will be required as a result.

Description

The minimal curation model reduces the amount of metadata generated for a given dataset. Instead of generating extensive DDI metadata and enhancing the variable labels for question-text search queries, the archive would simply generate enough descriptive metadata to allow users to discover and access the data and understand the general scope and topic of the data. This descriptive metadata would also include a persistent identifier and all the information required for a standard data citation.

While this strategy may compromise some of the discovery and online analysis potential for a dataset, it would still serve basic discovery and accessibility requirements. Description to support understanding of the content and context of the data would be left to information contained in supplementary documentation.

Preservation

In regard to preservation, the archive would continue to normalize files into a non-proprietary, system-agnostic file format, as we believe this is necessary to ensure that users are able to properly render the data into the future even as hardware and software systems become obsolete. Rather than producing several different file derivatives, normalization is limited to a single file format that users may convert for use in various software platforms.

Although we did not discuss other preservation activities such as generating and recording checksums, performing fixity checks, and migrating digital content, these are archival processes that are essential for long-term access and reusability and therefore cannot be compromised. However, these preservation activities are performed by archival platform systems and have little effect on archivist-led data curation processes.

What we have identified as a minimal data curation pipeline is neither an endorsement of a new standard of data curation, nor of MPLP itself. 'Minimal curation' is not suggested as an alternative to maximal curation. Maximal curation supports the sharing of the highest quality data that gives greater assurances that data will be reusable into the foreseeable future. 'Minimal curation' is presented to address conflicting priorities in a search for efficiency gains.

Discussion

The outcome of the MPLP exercise of reconsidering data curation processes is a recognition and greater appreciation of our commitment to providing access to quality data for our user community. In our examination of each of our current data curation activities, we were able to reaffirm the value of our practices to our users and their specific needs and expectations for quality data. As MPLP predicted, this exercise reminded us that "choices can be uncomfortable" (p. 233) when attempting to find efficiencies in our current practices, all of which we deem indispensable to our users. But we have little choice but to do so as we anticipate an increase in demand for data curation services. MPLP forced us to think about how each task in our data curation pipeline contributes to our goals. In doing so, we also reaffirmed the necessity and non-negotiable nature of some tasks that must be performed regardless of their intensity.

The search for 'good enough' for data has again left us in a quandary since in many ways meeting the requirements for reuse requires labor-intensive data quality review processes. Several data repositories have implemented a variety of resource management strategies for addressing the challenge of providing high quality data access with limited support. For example, the UK Data Archive has developed different levels of data curation to most effectively respond to users' varying needs (UK Data Archive, 2013). A key aspect of this is clear communication of the data curation tasks that will (and will not) be performed as part of a program of expectation management that distinguishes roles, rights, and responsibilities of the data producer, data user, and data archive.

Shifting responsibility for certain data curation tasks from the data archive to the data producer and data user assumes that data producers and users have an understanding of data quality requirements and the tasks required to meet those requirements, which, unfortunately, is not always the case. To address these challenges, information professionals have produced a proliferation of educational materials and programs to teach researchers strategies for effectively managing their data with eventual data archiving and sharing in mind. While online education programs (such as MANTRA and the Research Data Management and Sharing MOOC) have the potential to reach researchers worldwide, the impact of such education programs is not immediate and does not necessarily guarantee that data meet the standard of being 'independently understandable.'

Tools that facilitate and provide additional functionalities to streamline data curation processes also present opportunities for efficiency gains. Likewise, tools, such as the Open Science Framework, that help moderate and structure research workflows with an end goal of archiving and sharing data have the potential to assist researchers in creating data packages that meet data reuse requirements. However, even with these tools, certain tasks will continue to require manual data curation processes.

Ultimately, determining which data curation processes are essential for archiving and sharing of data that meet certain quality standards requires further research. This research will provide the empirical evidence and rationale for data archives' roles in curating data for

reuse in accordance with the needs of our designated community. Although previous studies have already clearly demonstrated the importance of contextual information (Faniel & Jacobsen 2010; Faniel et al., 2012), additional research is needed to investigate: 1) the designated community's expectations of the archive's role in providing quality data; 2) how variable level reviews affect reuse; 3) how the presentation of contextual information affects use; 4) how users interact with contextual and variable level metadata; and 5) how specific data curation tasks performed by archives directly impact the data quality and satisfaction criteria discussed within the literature.

By expanding our knowledge on the connections between user needs and data curation processes, we will be better equipped to determine what is 'good enough' for data. Likewise, we will be able to substantiate the necessity of data curation, whether it be maximal or minimal, for informed reuse and make clear that simply making data available does not automatically equate to data that are useable. We will then be able to expand and build upon initiatives advocating for the development of sustainable funding models for data archives (ICPSR, 2013).

Conclusion

Performing the conceptual exercise of applying MPLP in many ways raised more questions than it answered. MPLP reaffirmed our belief that a certain amount of processing is necessary to adequately meet users' needs. MPLP also brought to light some gaps in our knowledge about data use that prevents us from truly determining the minimum amount of processing needed. Future research will help us build better understanding of the connection between user needs and data curation processes. In many ways, the exercise suggests that 'good enough' for data still sets the bar pretty high, and building sustainable models to fund data curation will require the data archiving community to articulate the amount of skills, time, and labor that are non-negotiable when a high level of data quality is expected.

Essentially, this exercise boils down to a quotation from Clifford Lynch: "it is clear that an enormous imbalance exists between the resources currently available to fund these efforts and the potentially almost infinite demands of a fully realized data stewardship program; a key strategy in managing this imbalance is the effective use of the specific policy goals, such as data reuse, as shaping and prioritizing mechanisms in shaping an overall stewardship effort" (Lynch, 2013, p. 408). With the growth of data sharing mandates and the increasing focus on research transparency, data archives will play an essential role. However, questions still remain as to how we can best support these needs in a sustainable way that results in data that meet the requirements for reuse.

REFERENCES

Borgman, C.L. (2015) Big data, little data, no data: scholarship in the networked world. Cambridge, Massachusetts: The MIT Press.

Chang, A.C., Li, P. (2015) Is economics research replicable? Sixty published papers from thirteen journals say "usually not." Finance and Economics Discussion Series 2015-083. Washington DC: Board of Governors of the Federal Reserve System. doi:10.17016/FEDS.2015.083

Center for Open Science. (2015) Transparency and Openness Promotion (TOP) Guidelines. Available from: https://cos.io/top/#signatories

Consultative Committee for Space Data Systems. (2012) Reference model for an open archival information system (OAIS) (Magenta Book No. 650.0-M-2). Washington, DC: National Aeronautics Space Agency.

Crosas, M. (2011) The Dataverse Network®: An open-source application for sharing, discovering and preserving data. D-Lib Magazine 17 (1-2). doi:10.1045/january2011-crosas

DA-RT. (2015) The Journal Editors' Transparency Statement (JETS). Available from: http://www.dartstatement.org/#!blank/c22sl

Data Citation Synthesis Group. (2014) Joint Declaration of Data Citation Principles. Martone M. (ed.) FORCE 11, San Diego CA. Available from: /datacitation

Data Seal of Approval. (2013) Data Seal of Approval Guidelines (v.2). Available from: http://datasealofapproval.org/media/filer_public/2013/09/27/guidelines_2014-2015.pdf

Digital Curation Centre (DCC). (n.d) Curation reference manual. Available from: http://www.dcc.ac.uk/resources/curation-reference-manual

Faniel, I.M., Jacobsen, T.E. (2010) Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. Computer Supported Cooperative Work (CSCW), 19 (3-4), p. 355–375. doi:10.1007/s10606-010-9117-8

Faniel, I.M., Kriesberg, A., Yakel, E. (2012) Data reuse and sensemaking among novice social scientists. Proceedings of the American Society for Information Science and Technology, 49 (1), p. 1–10. doi:10.1002/meet.14504901068

Faniel, I.M., Kriesberg, A., Yakel, E. (2015) Social scientists' satisfaction with data reuse. Journal of the Association for Information Science and Technology. doi:10.1002/asi.23480

Freedman, L.P., Cockburn, I.M., Simcoe, T.S. (2015) The Economics of Reproducibility in Preclinical Research. PLoS Biol 13 (6): e1002165. doi:10.1371/journal.pbio.1002165

Gibbs, E., Lin, L., Quigley, E. (2013) Dataverse usability evaluation: Final report. Available from: http://dataverse.org/files/dataverseorg/files/dataverse_usability_report-participant_omitted.pdf?m=1458571553

Greene, M.A., Meissner, D. (2005) More product, less process: Revamping traditional archival processing. The American Archivist, 68 (2), p. 208-263.

Inter-university Consortium for Political and Social Research (ICPSR). (2012). Guide to social science data preparation and archiving (5th ed.). Ann Arbor, MI: ICPSR. Available from: http://www.icpsr.umich.edu/files/deposit/dataprep.pdf

Inter-university Consortium for Political and Social Research (ICPSR). (2013, June 24-25) Sustaining domain repositories for digital data: A call for change from an interdisciplinary working group of domain repositories. Available from: http://www.icpsr.umich.edu/files/ICPSR/pdf/DomainRepositoriesCTA16Sep2013.pdf

King, G. (1995) Replication, replication. PS: Political Science & Politics, 28 (3), p. 444–452. doi:10.2307/420301

Lee, C.A. (2010) Open Archival Information System (OAIS) reference model. In Encyclopedia of Library and Information Sciences. Taylor & Francis, p. 4020–4030.

Lynch, C. (2014) The next generation of challenges in the curation of scholarly data. In J. M. Ray (Ed.), Research data management: Practical strategies for information professionals. West Lafayette, Indiana: Purdue University Press. Available from: http://www.cni.org/wp-content/uploads/2013/10/Research-Data-Mgt-Ch19-Lynch-Oct-29-2013.pdf

Meissner, D., Greene, M.A. (2010) More application while less appreciation: The adopters and antagonists of MPLP. Journal of Archival Organization, 8 (3-4), p. 174–226. doi:10.1080/15332748.2010.554069

National Endowment for the Humanities (NEH). (2012) Data management plans for NEH Office of Digital Humanitites proposals and awards. Washington DC: National Endowment for the Humanities. Available from: http://www.neh.gov/files/grants/data_management_plans_2015.pdf

National Institutes of Health (NIH). (2003) Final NIH statement on sharing research data (No. NOT-OD-03-032). Bethesda, MD: National Institutes of Health.

National Science Foundation (NSF). (2010) Dissemination and sharing of research results. Arlington, VA: National Science Foundation. Available from: https://www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp#1

Nature. (2013) Availability of data, material and methods policy. Available from: http://www.nature.com/authors/policies/availability.html

Open Science Collaboration. (2015) Estimating the reproducibility of psychological science. Science, 349 (6251), aac4716. doi:10.1126/science.aac4716

Peer, L., Green, A., Stephenson, E. (2014) Committing to data quality review. International Journal of Digital Curation, 9 (1), p. 263–291. doi:10.2218/ijdc.v9i1.317

PLOS. (2014) Data availability policy. Available from: http://journals.plos.org/plosone/s/data-availability

Rothenberg, J. (1999) Ensuring the longevity of digital information. Washington, DC: Council on Library and Information Resources.

Science. (n.d.) Editorial policies: Data deposition. Available from: http://www.sciencemag.org/authors/science-editorial-policies

Society of American Archivists. (n.d.) Glossary of archival and records terminology: Preservation. Available from: http://www2.archivists.org/glossary/terms/p/preservation#.VxKnkfkrJ9M

Society of American Archivists. (n.d.) Glossary of archival and records terminology: Arrangement. Available from: http://www2.archivists.org/glossary/terms/a/arrangement#.VxKn2_krJ9M

UK Data Archive. (2015) Data ingest processing standards. Available from: http://www.data-archive.ac.uk/media/54782/cd079-dataingestprocessingstandards_08_00w.pdf

Wang, R.Y., Strong, D.M. (1996) Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 12, p. 5–33. doi:10.1080/07421222.1996.11518099

Notes

1. Sophia Lafferty-Hess is the Research Data Manager at the Odum Institute for Research in Social Science (228 Davis Library, CB# 3355, University of North Carolina at Chapel Hill), slaffer@email.unc.edu
2. Thu-Mai Christian is the Assistant Director of Archives at the Odum Institute for Research in Social Science (228 Davis Library, CB# 3355, University of North Carolina at Chapel Hill), thumai@email.unc.edu

MMRepo - Storing qualitative and quantitative data into one big data repository

by Ingo Barkow[1] Catharina Wasner[2] Fabian Odoni[3]

Abstract
In recent years, the storage of qualitative data has been a challenge to data archives using repositories that are based on relational databases, as large files cannot really be represented well in these structures. Most of the time, two or more structures have to be in place e.g. a fileserver that includes versioning for large files and a relational database for the tabular information. These structures necessitate the handling of multiple systems at the same time. With the arrival of Hadoop and other big data technologies, qualitative data and quantitative data can now be stored as mixed mode data in the same structures. This paper will discuss our findings in developing an early prototype version of MMRepo at the University of Applied Sciences Eastern Switzerland HTW Chur. Our prototype of MMRepo is a combination of the Invenio portal solution from CERN with a Hadoop 2.0 cluster using the DDI 3.3 beta metadata scheme for data documentation.

Keywords
Mixed mode, qualitative, quantitative, big data, repository

Introduction

Storing different kinds of data from different domains and enhancing them with metadata has been the main workflow of research data centers, data archives or scientific repositories for many years. Nevertheless, the usage of different file formats, metadata standards or IT infrastructures is challenging for data managers and IT managers in those facilities. Mixed mode data – meaning data derived from qualitative research (e.g. open interview formats, ethnographic studies using video and audio) and quantitative research (e.g. questionnaires, cognitive tests) – which arise from the trend of combining different research designs (Punch, 2009), poses a particularly significant challenge. Data from qualitative research differ in size and structure very much from data derived from a quantitative design. This can be illustrated by the following example: An observation of a classroom full of students doing a computer-based test by filming high definition videos from multiple angles and recording different audio tracks. This is actually a mixed mode design as it combines a qualitative ethnographic study with a quantitative design (the computer-based test). The qualitative design will have as a result several gigabytes of video and audio files leading to processes like transcription for scientific processing. In contrast, the computer-based tests result in datasets that are often stored in formats for statistical packages (e.g. SPSS, Stata, SAS, R) and that contain variables, including variable and value labels. Both types of data will be documented with different metadata standards with hardly any overlap between them due to the difference in domains.

Storage of qualitative data and quantitative data in repositories

From an IT perspective, the interesting question is this: Can different research data types be stored within the same technical infrastructure (e.g. file servers, relational databases, data warehouses) to enable search functionalities for users within the same frontend across all different data types? The next chapter therefore looks at the current processes of storing these kind of data in selected repositories, gives examples and explanations on why different systems exist in the same organization, and explains the objectives and design considerations of a big data driven approach.

The current state of the art in storing mixed mode data

As mixed mode data vary considerably in size, documentation and type, storing quantitative data and qualitative data in one structure is a challenge. In most data archives, these are the most common ways to handle mixed mode data.

1.) Storing both types in a relational database

If a relational database is used as the data storage mechanism, the quantitative data can be ingested in its tabular format (e.g. by importing an Excel table or SPSS file into a database table). The associated metadata could be stored in database tables as well using table joins or referential integrity to connect metadata and data thus allowing for variable shopping baskets or personal extracts (see Amin et al., 2011). This means by using these features of some repository systems (e.g. Questasy from CentERdata) the user does not have to download a Scientific Use File (SUF) and clean it from unnecessary variables but can choose in the portal which variables should be exported from the system. In many of the organizations connected to IASSIST or the DDI community (e.g. GESIS, DIPF, IAB, CentERdata) storing of tabular data in relational databases is therefore still a preferred way to document and store quantitative research.
However, while a relational database is advantageous for strong quantitative data, it does not work well for qualitative data. Typically, qualitative data would be stored in a relational database as a binary large object (BLOB) or an object of similar type directly in the table, which increases table size dramatically. The database would then be linked to the content of a file server. Sometimes, hybrid technologies like file streams would be used (e.g. SQL Server 2014; see Mistry and Misner, 2014). All these technologies do not combine very well. Relational database systems are not very good at handling BLOBs. There are limitations in size per single cell (usually 2GB – see SQL Server BLOB varbinary(max) datatype; Mistry and Misner, 2014) and the inflation of size normally leads to performance issues as database servers are optimized for handling small atomic data like short strings or numbers. The external linkage between relational database and file server also does not work very well as the two systems are separated. If users perform changes on the file server, the database server uses the information where the files have been stored, usually leading to dead links. File streams as a hybrid technology try to avoid this problem by letting the database server handle the file server automatically. This technology is only available in enterprise database servers like SQL Server or Oracle. Unfortunately, file streams have some technical disadvantages like e.g. the data outside of the table will not be part of the backup environment of the database server anymore. Furthermore, the outsourced data on the external file server is not a part of a database transaction (meaning the database server will not be able to roll back a transaction in case of a technical problem or break-off from the user side). In summary, filestreams fix the problem of broken links between file servers and database servers, but do not include the features relational databases are known and used for.

2.) Storing all data as files on a file server

The other option would be to handle quantitative data and qualitative data in their file-based state on a file server. Metadata would be provided one of three ways: by an external relational database, be attached as attributes of files, or by adding additional files that contain the metadata information. While this is a good way to handle qualitative data, the advantages of processing structured tabular information of quantitative data within a relational database are lost. In particular, quantitative data is simply processed in its file form so users can download it, while advanced features like variable shopping basket, personal extracts, and search functionalities for variables or basic tabulation from a portal solution would be heavily limited.

Advantages of storing mixed mode research data in the same technical infrastructure

When talking about the advantages of storing mixed mode research data in one infrastructure first the question has to be answered why it can be problematic to have different storage structures for qualitative and quantitative data.
From an IT perspective two different kinds of repositories also mean handling two times a complete set of hardware and software infrastructure. This means the IT administration, IT support and software development have a much higher effort, as this can be completely separate systems (different servers, different operating systems, different repository software, different frontend). If the user is supposed to be provided with one portal solution to search through two or even more repositories (basing on different kinds of data) a meta search solution has to be provided. This means the user types in a search request and the meta search solution divides this into different search requests running on the different back ends, collecting the results and collating them into one common result. Consolidating a multitude of different systems is a huge effort and a simpler one stop solution in the backend would lead to huge advantages as only one system has to be catered from hardware and software side.

Examples from data centers

To get a clearer picture of how data repositories handle different kinds of data, two organizations from Germany were selected as examples – the German Institute for International Educational Research (DIPF)[4] and the Leibniz Institute for Social Sciences (GESIS)[5]. Both institutions archive and distribute research data and publications and run one or more research data centers accredited by the German Data Council (RatSWD)[6]

DIPF is currently running three different repositories for different purposes (see Bambey et. al. 2013 and Bambey et. al. 2012):

- Qualitative data (e.g. school observations in video) are stored in the Medienarchiv[7]
- Questionnaires and answer schemes are stored in the Database for Quality of Schools (DaQS)[8]
- Data documentation regarding the framework program for educational research in Germany (over 300 projects) is stored in the metadata database of the Verbund Forschungsdaten Bildung[9]

Those three separate repository systems are derived from former projects and are run by three different organizations – DIPF, GESIS and IQB.[10] From a technological point of view they are completely different and optimized for their respective content: qualitative data, quantitative data, and metadata on quantitative and qualitative data.

A similar approach can be seen at GESIS where the following structures can be found:

- Quantitative data, questionnaires and study documentation are stored in the Data Catalogue (DBK)[11]
- Variable-level information is stored in the ZACAT portal[12]
- Metadata on microdata are stored in the MISSY system[13]
- Full-text social science documents are stored in the Social Science Open Access Repository (SSOAR)[14]
- Historical studies and time series are stored in HISTAT[15]
- Time series data from social indicators are stored in the online information system SIMon[16]

The diversity of GESIS systems is due to independent developments in different departments but as well to the requirements of different target groups and diverse digital resources. An integrated search function across all sources is under development.[17]

The diversity of repository systems in DIPF and GESIS developed organically with the organizations over the years. This pattern can be seen in many similarly-sized institutions around the world. The development of integrated repository systems was constrained by the previously mentioned limitations of relational database and file server data storage systems. The separation of storage based on data types is therefore valid and has to be seen as a product of the times in which they were developed. Many of the repository systems have been running for several years or in some cases even decades.

Vision and objectives of a big data driven repository

Nevertheless, the question remains whether all research data can be unified in one system by using more modern approaches, not least because the IT administration of multiple existing systems alone takes up many resources. One candidate for this is big data technology. Possible benefits of using big data technology to store different research data types include

- The use of cluster based file systems. Big data file systems like HDFS2 automatically split files and workload across multiple servers including redundant copies. This means it has inbuilt fail save and processing capabilities already from the software side (no need to use expensive hardware).
- Access to robust semantic search systems like SOLR and Elasticsearch. As big data was originally developed for handling large unstructured data within search engines it comes with sophisticated search capabilities which go beyond a simple string matching, but also features understanding of underlying concepts in e.g. a full-text search to improve the results for the users.
- Applicability of text mining or natural language processing. As large quantities of unstructured data can be stored and processed in parallel big data offers the possibility to analyze these data with advanced syntactical or semantic methods.

Furthermore, big data technologies can manage unstructured data like qualitative data. Having all data types in one system would be less costly and resource intensive because no meta search platform as described before has to be set up and only one system has to be developed and maintained.

Another advantage from user perspective is new methods of data analytics and data science can be used e.g. for analyzing data across different data types with the possibilities like text mining or natural language processing offered within the big data solution. These can be combined with classical statistical analysis from the sphere of social sciences and offer a scientific value add.

Design considerations of using big data as a unified repository

Big data solutions like Hadoop[18] were originally developed as search engine companies like Yahoo or Google were not able to store the masses of data needed to offer their services. While relational databases or data warehouses rely heavily on clear data structures, the design paradigms of big data technology are different. Instead of having structured data on expensive cluster hardware, big data technology was designed to allow parallel processing of unstructured data on inexpensive hardware. This basis was extended over the years from a more file systems based approach like Hadoop 1.0 to a multi-layer platform as can be seen from figure 1.

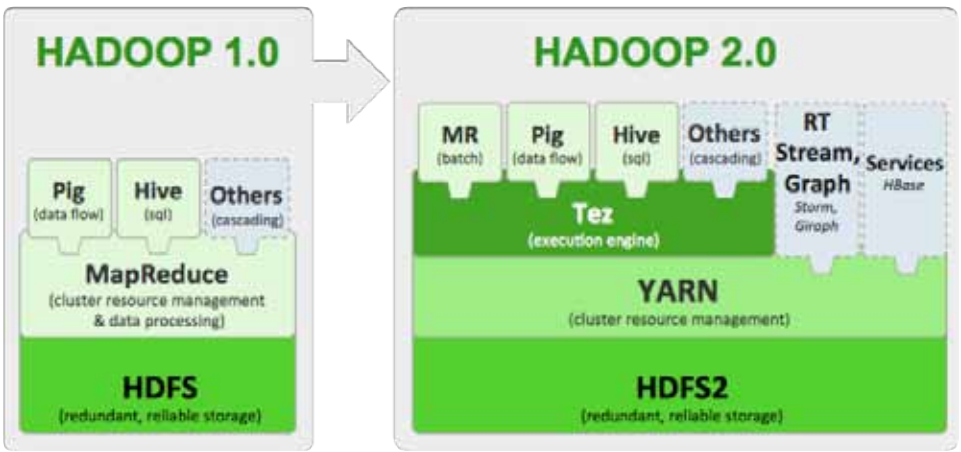


Figure 1 – The development from Hadoop 1.0 to Hadoop 2.0 (Hortonworks 2013)

The addition of services is especially interesting in big data developments like Hadoop 2.0. One additional service in Hadoop 2.0 is Hbase, a non-relational (NoSQL) and column-oriented database modeled after Google's BigTable[19]. It runs on

top of HDFS2, the native file system of Hadoop designed to store data among multiple computers, a cluster, by breaking up the data into blocks and distributing them throughout the cluster[20]. Another additional service is Hive which enables the use of SQL-like queries on the cluster.

Hadoop 2.0 could be the technical basis of a unified repository, whereby tabular content like metadata or qualitative data can be stored within Hive and large quantitative datafiles within HDFS2. This is the design idea which led to the MMRepo prototype project at the University of Applied Sciences Eastern Switzerland HTW Chur.

MMRepo as a prototype of a unified repository

As described in the chapter before, MMRepo was started as a prototype project to experiment with metadata, qualitative data and quantitative data within a single big data-based repository infrastructure. At the time of writing, MMRepo is a small project meant to test the performance and feasibility of the big data approach and must be considered a work in progress. The project started on January 1st, 2016 and its first phase ended on September 30th, 2016. The biggest obstacle in this respect was Invenio 3.0 final is as of today not released yet (March 2017). This means a large part of the final frontend testing had to be postponed to a later phase of the project. The current plan of Invenio specifies the release of the final version of 3.0 for summer 2017. The frontend testing will therefore be performed at a later date in the follow-up project called LifeCycleLab and not be part of this paper.

Structural design of MMRepo

The following test scenario has been set up to test the feasibility of the project. As the system's backend, a Hadoop 2.0 cluster is used with Hbase and HDFS2 as services. This backend will be combined with an Invenio[21] 3.0 beta frontend in the second project phase as the project is too small to develop portal functionalities by itself. The advantage of Invenio in this context is that it offers a modular framework for repositories where the data storage can be exchanged with something else (in our case a big data cluster). Furthermore, Invenio offers advanced features like semantic search or versioning, which benefit especially qualitative data. The structural design can be seen in the following figure. 2

The quantitative data and qualitative data used in the project are test data from previous studies at HTW Chur plus sample data from DIPF and the Research Data Center (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB)[22] . Therefore, a variety of test data has been imported into Hbase 2.0 and HDFS2. As the metadata schema for the quantitative data, Data Documentation Initiative (DDI)[23] Lifecycle 3.3 beta is used. For documenting the qualitative data, several internal groups at HTW are in favor of adopting the Metadata Encoding and Transmission Standard (METS)[24] . However, the final decision has not been made yet.

Figure 2 – Software structure of MMRepo

Hardware layout of the MMRepo prototype testing

As MMRepo uses Hadoop 2.0 as its backend, for proper testing, it is necessary to have a cluster environment. Only then, the advantages of parallel processing like the MapReduce algorithm can be exploited. Unfortunately, the project is much too small to employ even the least expensive servers. To simulate a multitude of nodes, the decision was made to set up the system on small third generation Raspberry Pi microcomputers.

Raspberry Pi 3 offers a quadcore 1.2 GHz ARM processor with 1GB of RAM plus SDHC card support and is an excellent less-costly solution for prototyping. A huge advantage for the Hadoop solution is additionally the support of LAN and Wireless LAN meaning two separated networks. For the current test setup the cluster network (LAN) and the user access (WLAN) can be separated as mixing the network traffic

Figure 3 – Raspberry Pi 3 (taken from www.raspberrypi.org)

from users and between server nodes might influence the results by slowing down reaction times.

From the hardware layout of big data solutions it does not make sense to test the capabilities by setting up e.g. two large powerful servers with inbuilt redudant hardware (e.g. multicore, multiple redudandant hard disks, fail safe networking) as the Hadoop software layout rather demands a high number of cheap servers with inexpensive hardware which can operate in parallel. The redudancy and the parallel processing is provided by the big data software itself. We also decided against employing virtualization (e.g. VMware vSphere, Citrix XenServer). If we tested the setup e.g. by setting up eight virtual servers in reality all virtual machines might end up on the same physical machine or might be shifted within several physical servers with different hardware layouts during the test (e.g. VMware – vMotion between nodes) thus influencing our results.

A prototype setup using the cheapest possible hardware is therefore closer to the specifications Hadoop was originally intended for. As we want to test primarily the search capabilities using an array of multiple servers we chose this setup for the prototype. In later phases of this project this prototype will be replaced by an array of inexpensive servers.

Results from the prototype

In late autumn 2016 we started to test the backend of the prototype after setting up the hardware with a limited number of Raspberry Pi nodes (one master, seven slaves) after it was clear Invenio 3.0 will not be released before the end of this project phase. We therefore decided on testing qualitative data and quantitative data on the backend of the cluster. Installing Hadoop on the inexpensive hardware was uneventful as there are in the meantime multiple How-Tos to be found (e.g. from IBM[25]). Some settings had to be modified manually in the environment variables, as Raspberry Pi is an unfamiliar hardware for Hadoop, but the How-Tos provided enough support.
To test the setup and set the correct block size for Hadoop the following datasets were implemented:

- US Census Data (2013_ACSSF_All_States_All_Tables)
  - Size: 469 MB
  - Format: Tables (*.csv)
  - Usage in Hadoop: Tables stored in Hbase (not relationally connected)

- Wikipedia – US Version (11/2016)
  - Size: 49 GB
  - Format: SQL dump
  - Usage in Hadoop: Relational database in Hbase

- Wikipedia – US Version (06/2008)
  - Size: 7.2 GB
  - Format: Static HTML dump
  - Usage in Hadoop: Files in Hbase

The selection of example datasets shows a mix of qualitative and quantitative data, but also the first problem in the prototypical setup. The amount of data used does not qualify as big data in a classical sense (4 Vs – Volume, Velocity, Veracity, Value – see e.g. Meyer, 2013) as the Volume is only several gigabytes while in big data repositories we rather aim at hundreds of terabytes or exabytes in the near future. Nevertheless, as the cluster is far from powerful from a systems' performance perspective the amount of data should be sufficient.

To see if the Hadoop cluster functions properly we started by implementing simple word count operations. The Raspberry Pi based cluster worked according to specifications, but ran very slowly. Part of the performance could be optimized by changing the block sizes. Also, the Hadoop software is not optimized to the Raspian operating system running underneath it, so the full performance of CPU and chipset is not used. As a first result it can be said the Raspberry setup is interesting to explore the possibilities of Hadoop on a real cluster especially for teaching purposes in an university environment, but it is not sufficient for performance testing or even productive setup.

From a conceptual perspective the results were much more promising. The ideas of storing files into the cluster-based filesystem HDFS and the tables into the NoSQL database Hbase worked fine. NoSQL database in this respect means "Not only SQL" a non-relational database layout which offers more flexibility in database layout while losing some transaction capabilities. The MySQL database dump from Wikipedia was imported into Hbase by using Sqoop. To see if the overall setup works we created search requests for HDFS2 and Hbase using SOLR which is embedded into Hadoop. As Hbase is built on top of HDFS2 as a column-oriented non-relational database

system it worked well with SOLR so essentially we were able to use one hardware platform, one software platform and one search engine for the purpose of searching through qualitative and quantitative data as a first step.

Nevertheless, there is a limitation. Although SOLR searches through files and tables it does not mean the results are meaningful per se. In the end we get text extracts from documents and rows of tables as result sets which can be considered an intermediate step from a presentation point of view. A real frontend search solution needs adaptations in the presentation layer to have a user-friendly version of the results. Currently the whole setup from the backend to the portal is very crude and can be considered a proof of concept for further funding, but not an out-of-the-box usable solution. Nevertheless, the basic functionality is already available and therefore our project can continue. We are currently certain to be able to exchange our currently separated repositories into one big data solution, although the real development work on much better hardware has to start first.

References

Amin, A., Barkow, I., Kramer, S., Schiller, D. & Williams, J. (2011). Representing and Utilizing DDI in Relational Databases. Minnesota : DDI Working Paper Series [DOI:http://dx.doi.org/10.3886/DDIOtherTopics02].

Bambey, Doris; Rittberger, Marc (2013): Das Forschungsdatenzentrum (FDZ) Bildung des DIPF: Qualitative Daten der empirischen Bildungsforschung im Kontext. Standards und disziplinspezifische Lösungen. In: Huschka, Denis; Knoblauch, Hubert; Oellers, Claudia; Solga, Heike (2013) (Hrsg.): Forschungsinfrastrukturen für die Qualitative Sozialforschung. Berlin: Scivero-Verlag, S. 63-71. URL: http://ratswd.de/dl/downloads/forschungsinfrastrukturen_qualitative_ sozialforschung.pdf (20.03.2015).

Bambey, Doris; Reinhold, Anke; Rittberger, Marc (2012): Pädagogik und Erziehungswissenschaft. In: Neuroth, Heike; Strathmann, Stefan; Oßwald, Achim; Scheffel, Regine; Klump, Jens; Ludwig, Jens (Hrsg.): In: Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme. Boizenburg: Hülsbusch, S. 111-135.

Hortonworks (2013). Apache Hadoop Patterns of Use.

Andreas Meier (2013). Relationale und postrelationale Datenbanken – Leitfaden für die Praxis, Springer-Verlag.

Mistry, Ross and Stacia Misner (2014). Introducing Microsoft SQL Server 2014.

Punch, K. F. (2009). Introduction to research methods in education. London: Sage.

Notes

1. Ingo Barkow is an Associate Professor for Data Management at the University of Applied Sciences Eastern Switzerland HTW Chur and can be reached by email: ingo.barkow@htwchur.ch
2. Catharina Wasner is a research associate at the University of Applied Sciences Eastern Switzerland HTW Chur
3. Fabian Odoni is a research associate at the University of Applied Sciences Eastern Switzerland HTW Chur
4. http://www.dipf.de
5. http://www.gesis.org
6. http://www.ratswd.de
7. http://www.fachportal-paedagogik.de/forschungsdaten_bildung/medien.php?la=de
8. http://daqs.fachportal-paedagogik.de
9. http://www.forschungsdaten-bildung.de
10. https://www.iqb.hu-berlin.de
11. https://dbk.gesis.org
12. http://zacat.gesis.org
13. http://www.gesis.org/missy
14. http://www.ssoar.info
15. http://www.gesis.org/histat
16. http://gesis-simon.de
17. http://www.gesis.org/en/research/applied-computer-and-information-science/information-retrieval/
18. http://hadoop.apache.org/
19. http://www-01.ibm.com/software/data/infosphere/hadoop/hbase
20. http://www-01.ibm.com/software/data/infosphere/hadoop/hdfs
21 http://invenio-software.org
22 http://fdz.iab.de
23. http://www.ddialliance.org
24. http://www.loc.gov/standards/mets
25. Alan Verdugo - Building a Hadoop Cluster with Raspberry Pi - https://developer.ibm.com/recipes/tutorials/building-a-hadoop-cluster-with-raspberry-pi/#r_overview

Servicing New and Novel Forms of Data:

Opportunities for Social Science

by Aidan Condron[1]

Abstract
For social and economic researchers, many useful but previously unavailable sources of data have become at least potentially accessible in recent years. These 'new and novel' forms of data (NNfD), such as social media data or smart meter data, represent potentially invaluable resources for researchers, but pose challenges for access provision and analysis. This short article introduces Data Service as a Platform (DSaaP), a project currently underway at the UK Data Service to establish a technological infrastructure supporting data archivists and social and economic researchers in managing and analysing both familiar and new and novel forms of data. It presents an overview of NNfD in social science contexts, introduces the DSaaP system, and sketches short examples of DSaaP capabilities in analysing NNfD, drawn from an associated UKDS project Smarter Household Energy Data: infrastructure for policy and planning, before concluding with some reflections on the potential value added to social scientific research by Data Service as a Platform.

Keywords
Big Data, Data Science, Social Science, Data Archiving, Hadoop

Introduction

The UK Data Service[2] (UKDS) Big Data Network Support[3] (BDNS) team is currently engaged in a major project to develop Data Service as a Platform (DSaaP), a technological infrastructure supporting data archivists and social and economic researchers in managing and analysing both familiar and new and novel forms of data (NNfD). NNfD encompasses very large datasets, sometimes referred to as 'big' data, but not all, or all aspects of NNfD are necessarily 'big' in this way.

This short paper presents an overview of NNfD in social science contexts, introduces the DSaaP system, and sketches short examples of DSaaP capabilities in analysing NNfD, drawn from an associated UKDS project Smarter Household Energy Data: infrastructure for policy and planning,[4] before concluding with some reflections on the potential value added to social scientific research by data science techniques and Data Service as a Platform.

An accompanying video demonstration can be viewed here.[5]

New and Novel Forms of Data

Although the term 'big data' has been popularised in recent years, NNfD are not just about the size of the files, but about a broader view of what, how, when, and where data can be collected, stored, linked and analysed to further research[6] (OECD, 2013). Many new data sources on social and economic activity have emerged such as social media, smart meter and other household consumption data, internet usage, sensor and footfall readings, and many others. While much, if not all, of this this data is not collected specifically for the purposes of social and economic research, there many exciting possibilities for reuse by researchers, presenting a potentially invaluable resource. Even if some of these data have been collected for some time now, they represent new and novel forms of data in a social science context.

Some of these datasets are very large, but not all, or all aspects are necessarily 'big' in this way. In any case, the most advantageous approach to 'big data' is often to downscale or reduce to manageable sizes, whether by aggregation or mining for the more valuable elements (Siems & Wolf, 2007). Smart meter data, for example, contains a huge number of readings, but these are more useful in context of associated geodemographic data on much smaller numbers of households from which the readings are drawn. Analysis and findings are enriched by linking to other data sources such as meteorological data, which does tend to be voluminous, and housing stock and energy performance certificates, which are more compact. In making a case for analysis of NNfD as a progressive factor in social science, and in thinking big about data, it's not just a matter of scale, but of innovation in assessing what data sources can be harnessed to answer research questions, and how linking and triangulation can enhance analyses and findings. A big data approach doesn't always mean using massive files and processing power!

Data Service as a Platform - DSaaP

Our concept of DSaaP is a 'data lake' system capable of storing data of any kind, which will cater for both traditional data and NNfD drawn from current and future UKDS holdings. The data lake is a secure, format-agnostic repository providing a powerful, scalable, suite of tools for data processing. It is built on Hadoop,[7] a software framework that facilitates high-speed processing of datasets of any size by distributing data across networks (referred to as clusters) of linked computers (referred to as nodes).

KDNuggets, a leading data science website, offers the following working definition of a data lake.[8] A data lake is a storage repository that holds a [potentially] vast amount of raw data in its native format, including structured, semi-structured, and unstructured data. The data structure and requirements are not [necessarily] defined until the data is needed'. This differs from data warehouses, where data is strictly formatted and structured to meet specific, pre-defined reporting functions.

Developing DSaaP is a staged, medium-term enterprise, involving installing cloud-based and on-premises Hadoop infrastructure, establishing ingest pipelines to populate the data lake (which will recombine, store and tag datasets in a Resource Descriptive Framework (RDF) triplet[9] and key-value pair format) and providing access channels and user endpoints for researchers to work with data. Data stored within it are assigned randomly machine generated globally unique identifiers[10] (GUID)s at the lowest possible level of granularity, down to the field, record, or even data point level. By drawing on W3C standardised Vocabulary Services,[11] this tagging facilitates dynamic reassembly, interlinking, and querying of data according to user requirements.

Access channels and endpoints are designed to cater for user requirements, and are determined by engagement with the research community. The approach adopted in developing DSaaP is pro-active, moving from making data available through downloads bundles from catalogues to scoping and providing solutions for secure data access, management, linking, and analysis within the DSaaP environment. Approved researchers should be able to log into DSaaP work areas with access to the data they will work with, and as functionality develops, DSaaP will facilitate self-service for researchers and other users, unifying data tools and querying, linking and analysing data across a complex environment. The strategic impetus driving the project is to create a twenty-first century data solution for the ongoing data access and curation communities.

In developing the DSaaP, BDNS has adopted the philosophy of the Open Data Platform initiative (ODPi),[12] focusing on developing a standardised data working environment facilitated entirely through open-source software. In referring to an 'Open Data Platform', this does not mean that the platform support is limited to open data. Openness refers to the data lake's open source software build, and to the ongoing cross-community technological development of which the DSaaP itself is a part and an exemplar, and in which DSaaP developers and users are members.

Support will be provided for users across a wide range of technical expertise, catering for novices who needs to navigate and query data using point and click or drag and drop interfaces, researchers interested in applying traditional techniques such as linear regression or analysis of variance (ANOVA), analysts interested in employing machine learning algorithms such as random forest, nearest neighbours, or text mining techniques, up to technologists or developers who want to develop their own bespoke data tools. In keeping with the UKDS public service ethos and with the spirit of open science, it is hoped that researchers using the DSaaP will participate in knowledge sharing and functionality development by contributing to shared repositories of software code and analytical techniques.

While capable of scaling as necessary to accommodate NNfD, all DSaaP-based data storage and access provision will be governed by established Research Data Management (RDM) principles which underpin all UKDS archiving and curation work.[13] UKDS Data is protected by enterprise-grade security and governance, with access governed by the UKDS three-tier classification of open, safeguarded, and controlled data.[14] If sensitive data is ingested to DSaaP, research will be regulated within the 'five safes'[15] framework of Safe People, Safe Projects, Safe Settings, Safe Outputs, and Safe Data. As working with very large datasets or linking multiple datasets can pose risks to anonymity and privacy, rigorous machine actionable Statistical Disclosure Control (SDC) checks should be applied when data is ingested, to determine appropriate levels of access security, and to all potentially disclosive outputs.

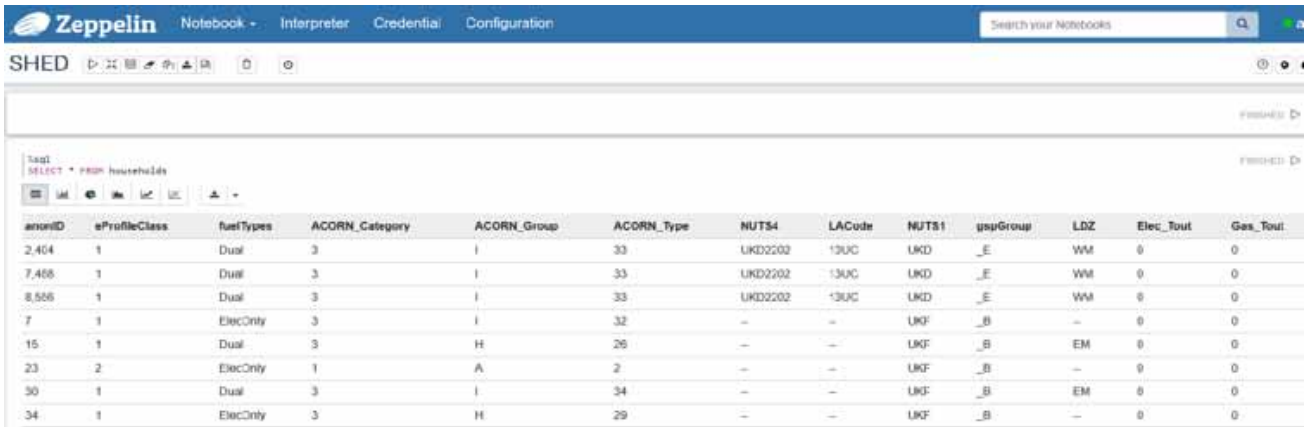
Sample Use Case: Exploratory Data Analysis with Household Energy Data

DSaaP capabilities in generating value from NNfD is illustrated by work on Smarter Household Energy Data: infrastructure for policy and planning[16] (SHED), an associated project in partnership with the University of Cape Town' DataFirst[17] and University College London's Energy Institute,[18] which 'focuses on data infrastructure and brings together data professionals, energy researchers and policymakers in SA and the UK'. DSaaP and SHED intersect, as the SHED infrastructure component will be provided by DSaaP, while SHED will provide DSaaP use cases, from ingest to endpoint, of data valuable in household energy consumption researchers.

SHED project work has involved scoping the household energy field through reading and undertaking engagement with the research community, canvassing requirements through roundtable meetings[19] and collaboration with individual researchers working on household energy. While DSaaP is developing discipline-agnostic, generic systems and tools of value to a wide user base, this association with specific research and datasets facilitates pilot project initiation, data processing test cases, and analytical proofs of concept.
More general, generic data analysis work has also been carried out by UKDS staff on a large dataset collected during the Energy Demand Research Project (EDRP), a series of experimental trials involving smart meters on household energy consumption during 2008-2010 which was deposited by the Department of Energy and Climate Change (DECC) with the UKDS for curation in late 2014.[20] The EDRP data presented an initial technical challenge for assessment and curation, as it included files over 12 GB in size containing hundreds of millions of records, far too big to be opened with familiar desktop software, and indeed too big to be fully loaded into the memory of standard PCs, regardless of the software used. A DSaaP prototype facilitated loading, opening, and employing Exploratory Data Analysis (EDA) techniques to explore the data.[21] Pioneered by John Tukey in the late 1970s (Tukey, 1977) and now accepted as an important component in data science, EDA involves initial, non-hypothesis driven, investigation of data, developed by generating summary statistics, plotting variable distributions and time series, and transforming data, and is often used as a crucial first step towards understanding data, particularly large, unfamiliar datasets (Marsh, C & Elliott 2008, O'Neil and Schutt, 2013: 34-40).

Exploring the EDRP datasets, which previously seemed impenetrable, is relatively easy with DSaaP. Geodemographic information on the households included in the study consists of fifteen variables including a household anonymous identifier, data on the types of fuel available to the household (electricity or electricity and gas), energy consumption pricing structure (whether fixed rate or Time of Use Tariff(Tout)), geographic region, and socioeconomic status as defined by the ACORN classification system. Once data is loaded, variables

can be viewed as standard data tables, familiar to anyone who has worked with spreadsheets, SPSS, or any type of database software, as seen in figure 1 below.



anonID	eProfileClass	fuelTypes	ACORN_Category	ACORN_Group	ACORN_Type	NUTS4	LACode	NUTS1	gspGroup	LDZ	Elec_Tout	Gas_Tout
2,404	1	Dual	3	I	33	UKD2202	13UC	UKD	_E	WM	0	0
7,468	1	Dual	3	I	33	UKD2202	13UC	UKD	_E	WM	0	0
8,566	1	Dual	3	I	33	UKD2202	13UC	UKD	_E	WM	0	0
7	1	ElecOnly	3	I	32	–	–	UKF	_B	–	0	0
15	1	Dual	3	H	26	–	–	UKF	_B	EM	0	0
23	2	ElecOnly	1	A	2	–	–	UKF	_B	–	0	0
30	1	Dual	3	I	34	–	–	UKF	_B	EM	0	0
34	1	ElecOnly	3	H	29	–	–	UKF	_B	–	0	0

Figure 1 Data table viewed in Zeppelin

Apache Zeppelin[23] is an analytical tool integrated into DSaaP which supports multiple language backends, such as Python, Scala, and Structured Query Language (SQL), and provides powerful visualisation tools. The user interface is accessed through standard web browsers such as Google Chrome or Mozilla Firefox, meaning that users with DSaaP log in credentials need not install any additional software to work with data on the system. With non-controlled data, this can be done from any internet-connected location. The data table view is accompanied by a series of interactive graphic views, where variables can be selected and manipulated with clicks or drag and drop, with Zeppelin dynamically generating visualisations such as bar, line, or scatter plots. These types of functionality speed and ease EDA and other forms of analysis, particularly when working with NNfD. As a first EDA step with the EDRP data, univariate, bivariate and multivariate distributions of these variables were rapidly and dynamically visualised using 'out of the box' features. Figure 2 below shows a bivariate bar chart produced with drag and drop commands, graphing the study population of households by geographical region and types of fuel used.

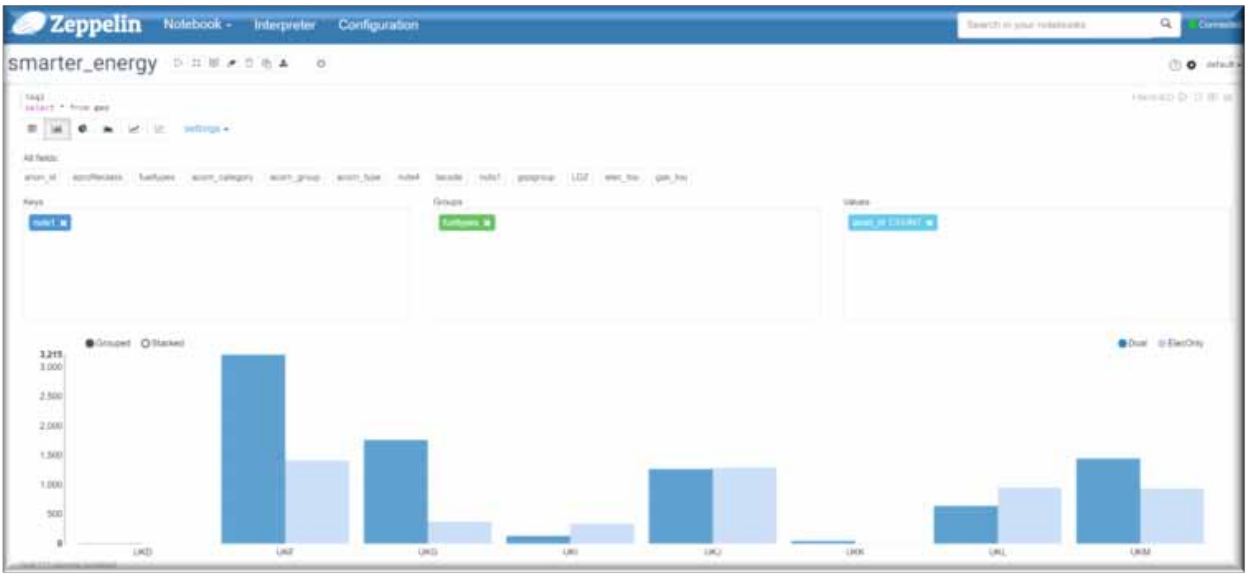


Figure 2 Zeppelin cross tabulation comparing electricity only and dual use households

While other software packages can certainly produce bar charts, DSaaP facilitates linking to and working with data on a much greater scale. While approximately 14,000 households are included in this study, data was collected at half-hourly intervals over a thirty month period, generating 413,000,000 records on electricity usage alone (a figure which was itself unknown before loading and counting in DSaaP). Apache Hive,[24] a data warehousing interface included with DSaaP is configured for aggregation and analysis of large data sets like this. Figure 3 below, a visualisation generated by Hive, graphs mean household electricity usage over an average twenty-four hour period in December 2009, demonstrates DSaaP capabilities in drawing meaning and value from the data. 'Dual' households with both gas and electricity installed are represented by the blue line, while 'ElecOnly' households with households relying solely on electricity for energy, including heating are represented by the orange line. As might be expected, and can be seen clearly from the graph, dual usage households consume noticeably less electricity on a December day than electricity only households.

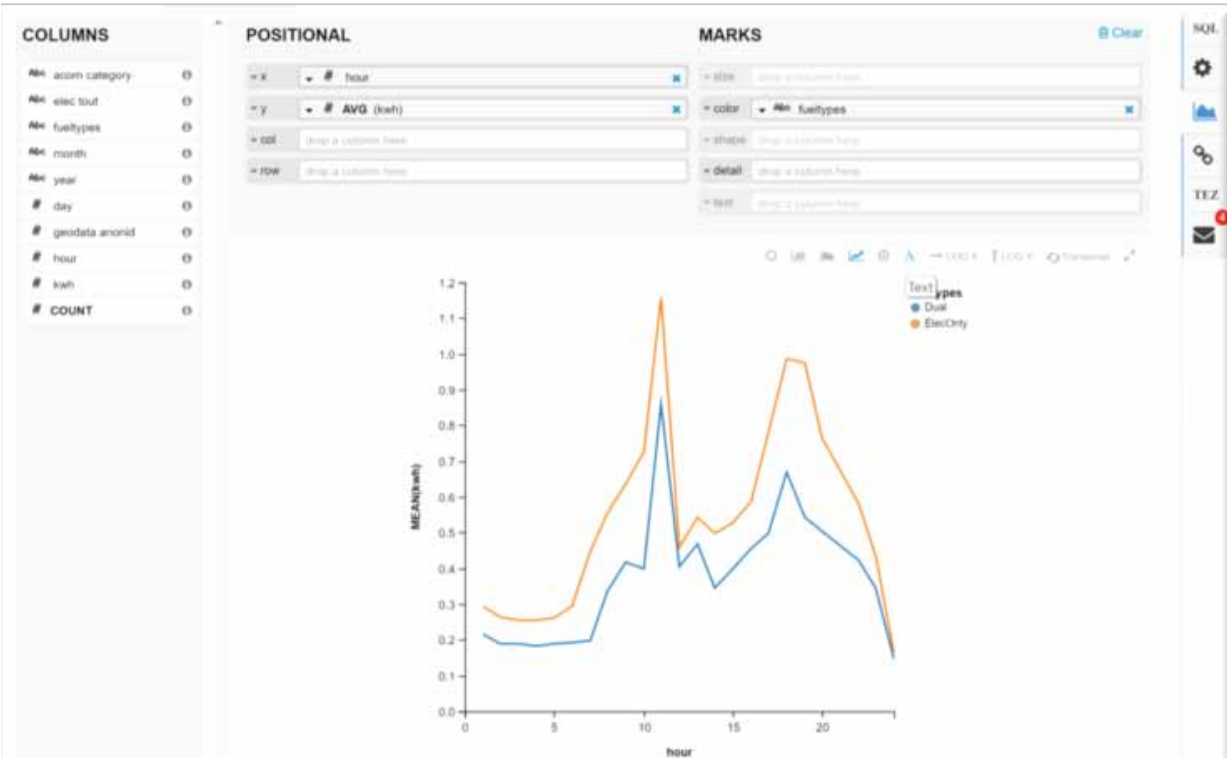


Figure 3 Comparing mean energy consumption, December 2009

This graph demonstrates some of the power of the DSaaP system. The graph represents a series of aggregations drawn from two linked data tables and hundreds of thousands of data points in a simple and easily readable output. Figure 4, below is a cross section of a 3 x 12 grid, extending this analysis across the twelve months for the years 2008-2010, an output drawing on over 3.7 billion data points.

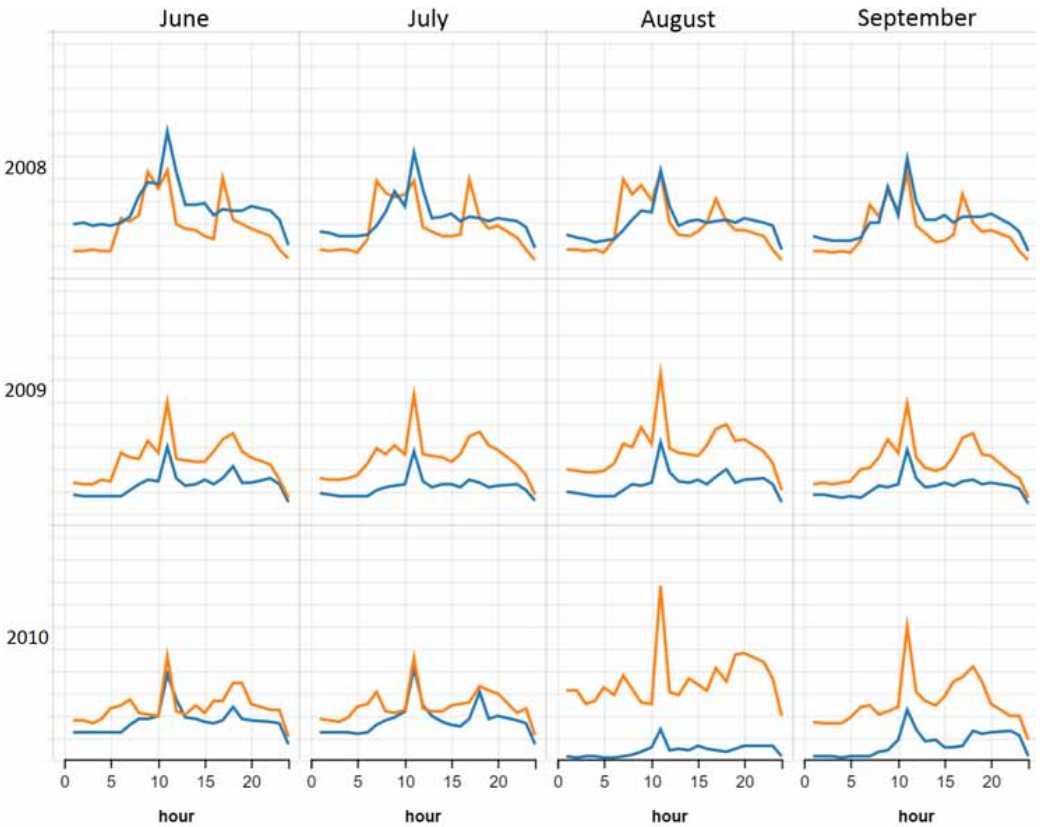


Figure 4 Energy Curves, June-September, 2008-2010

While household energy curves provide striking visualisations, which are recognised as useful analytical devices in the field (Palmer et al. 2014), it should be stressed again that every visualisation is based on a data table, which is produced by querying the underlying data. The graphs above are generated by selecting, subsetting, pivoting and plotting specific variables from EDRP, all standard Hive features, and display DSaaP's power to recombine and represent data at different levels of analysis, from high-level national and annual aggregations down to the fine granularity of single households and hourly intervals.

The energy curves are based on linking two data tables, one with data on household energy consumption, and some with data on the households themselves. Both tables were included with the EDRP dataset, and the common anonymised household identifier facilitated easy linkage. Linking to other data sources is also facilitated. Figure 5 below, a line over bar graph, plots mean electricity consumption against mean daily temperature in the East Midlands region of England, displaying a strong negative correlation, with household energy consumption rising as temperature falls.

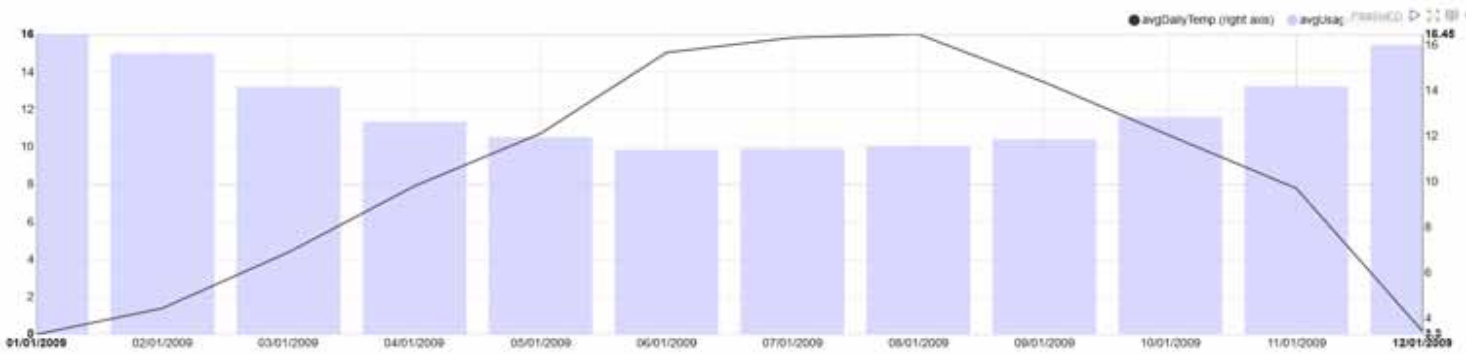


Figure 5 Line over bar monthly mean temperature and energy consumption

This graph is based on spatial and temporal aggregations of energy consumption (mean monthly household energy consumption across geographical region) linked to open data from an external source, the Meteorological Office. Despite being derived from a large number representing a modest level of complexity, it presents a clear, readily understandable visualisation of the information. Any of the outputs produced including derived data tables, descriptive statistics, and visualisations can be easily stored on DSaaP or, security permitting, downloaded for use or reproduction elsewhere.

Conclusion: new and novel forms of data and opportunities for social science

The examples above illustrate some basic DSaaP capabilities. Visually subsetting by categorical data, aggregating and pivoting large sets, displaying correlation, and linking to internal and external data sources are shown, but assuming availability of data, almost any imaginable analysis, visualisation, or derived data product required by social or economic researchers could be generated.
The usefulness of these data products is not determined by the technology, but by the research agenda and design. To return to the household energy consumption example, hourly data at the household level are required for answering questions about daily consumption patterns, but monthly aggregations at district level are more useful for exploring questions on seasonal or regional variations in energy consumption. The useful level of detail or granularity or is useful is determined by analysis performed and research questions asked, as are the appropriate tools to use. The examples above have shown Hive's utility in managing and analysing large numbers of datapoints, and some Zeppelin capabilities in dynamic plotting and graphing. Interoperability between these and other DSaaP features provides for a powerful analytical platform.

The key driver of this work is not just to understand these particular datasets, but to develop transferable understanding, expertise, and tooling, intended for wider use in working within the new data environment, from ingest to access and analysis. This is a research community oriented and involved project, scoping interest and requirements within the social science community, and working actively with researchers to develop the most useful systems. Community engagement is a two-way process. Collaboration with energy researchers has developed use cases based on actual research, while demonstrations of DSaaP features whether on video or conducted live have generated interest and ideas on how the systems can be used and how NNfD can be leveraged.

Moving forward, the Big Data Network Support team will standardise and generalise procedures developed from DSaaP work. For example, work on assessing, aggregating and visualising time series data developed with reference to the SHED project can be adapted and scaled to cover time series data from other areas. This institutional learning will be implemented through DSaaP and also disseminated through training, seminars, lectures, conference papers, and knowledge exchange programmes and partnership, all of which are already underway. The overall aim is to establish a general-purpose data services system for social and economic research, supporting both established and emerging analytical techniques, and both traditional and new and novel forms of data.

References

'New Data for Understanding the Human Condition: International Perspectives'
OECD Global Science Forum Report on Data and Research Infrastructure for the Social Sciences, Brussels, February 2013.

Marsh, C & Elliott, J, 'Exploring data: an introduction to data analysis for social scientists', Polity, Cambridge 2008.
O'Neil , Cathy & Rachel Schutt, 'Doing data science: Straight talk from the frontline', O'Reilly Media, Inc., New York 2013.

Palmer, et al., 'Further Analysis of the Household Electricity Survey Energy use at home: models, labels and unusual appliances', Cambridge Architectural Research Limited, Cambridge 2014.

Siems, K & Wolf, D, 'Burning the Hay to Find the Needle – Data Mining Strategies in Natural Product Dereplication' CHIMIA International Journal for Chemistry, Volume 61, Number 6, June 2007, pp. 339-345(7)

Tukey, John W, 'Exploratory data analysis', Pearson, London 1977.

Notes

1. Dr Aidan Condron | Senior Officer, Collections Development and Producer Relations | Big Data Network Support | UK Data Service | University of Essex | Wivenhoe Park | Colchester CO4 3SQ | T +44 (0) 1206 874254 | E acondron@essex.ac.uk
2. https://www.ukdataservice.ac.uk/
3. https://bigdata.ukdataservice.ac.uk/
4. https://www.ukdataservice.ac.uk/about-us/our-rd/smarter-household-energy-data
5. https://www.youtube.com/watch?v=0HbcAyUwWDY
6. https://www.oecd.org/sti/sci-tech/new-data-for-understanding-the-human-condition.pdf
7. http://hadoop.apache.org/
8. http://www.kdnuggets.com/2015/09/data-lake-vs-data-warehouse-key-differences.html
9. https://www.w3.org/TR/rdf11-concepts/
10. https://betterexplained.com/articles/the-quick-guide-to-guids/
11. https://www.w3.org/2013/04/vocabs/
12. https://www.odpi.org/
13. http://www.data-archive.ac.uk/curate
14. https://www.ukdataservice.ac.uk/get-data/data-access-policy
15. http://blog.ukdataservice.ac.uk/access-to-sensitive-data-for-research-the-5-safes/
16. https://www.ukdataservice.ac.uk/about-us/our-rd/smarter-household-energy-data
17. https://www.datafirst.uct.ac.za/
18. http://www.bartlett.ucl.ac.uk/energy/
19. https://ukdataservicesmartenergydata.wordpress.com/
20. https://discover.ukdataservice.ac.uk/doi?sn=7591#1
21. https://www.youtube.com/watch?v=0HbcAyUwWDY
22. http://acorn.caci.co.uk/
23. http://zeppelin.apache.org
24. https://hive.apache.org/
25. http://www.carltd.com/sites/carwebsite/files/Report%203_Models,%20labels%20and%20unusual%20appliances.pdf

IASSIST

INTERNATIONAL ASSOCIATION FOR SOCIAL SCIENCE INFORMATION SERVICE AND TECHNOLOGY

ASSOCIATION INTERNATIONALE POUR LES SERVICES ET TECHNIQUES D'INFORMATION EN SCIENCES SOCIALES

The **International Association for Social Science Information Service and Technology** (**IASSIST**) is an international association of individuals who are engaged in the acquistion, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, data base librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompases hard copy as well as machine readable data

Paid-up members enjoy voting rights benefit from reduced fees for attendance at regional and international conferences sponsored by **IASSIST**. Join today by filling in our online application:

http://www.iaassistdata.info/

Online Application

IASSIST Member ($50.00 (USD))
Subscription period: *1 year, on: July 1st*
Automatic renewal: *no*

Please fill in the information our Online Form

The application is in USD, however, we do accept Canadian Dollars, Euro, and British Pounds as well.

The membership rates in all currencies as well as the Regional Treasurers who manage them are listed on the Treasurers page