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.

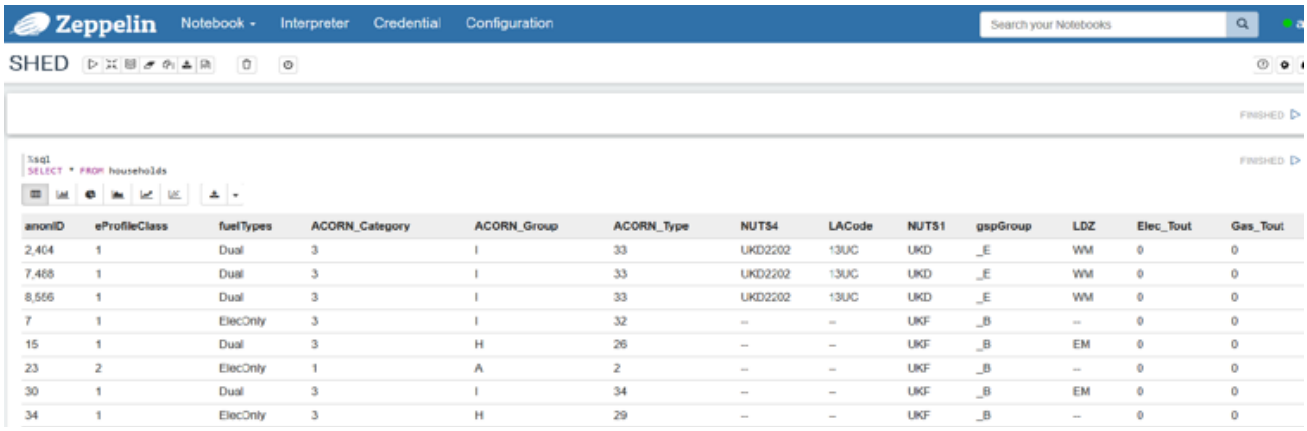
Sample Use Case: Exploratory Data Analysis with Household Energy Data

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.

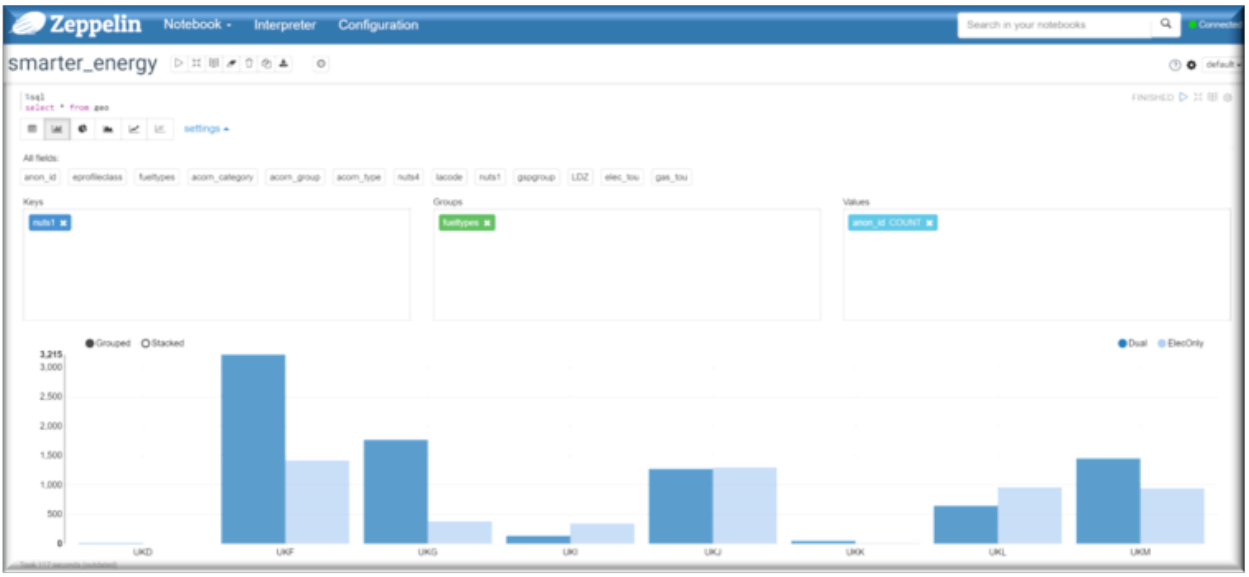


Figure 2 Zeppelin cross tabulation comparing electricity only and dual use households

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.

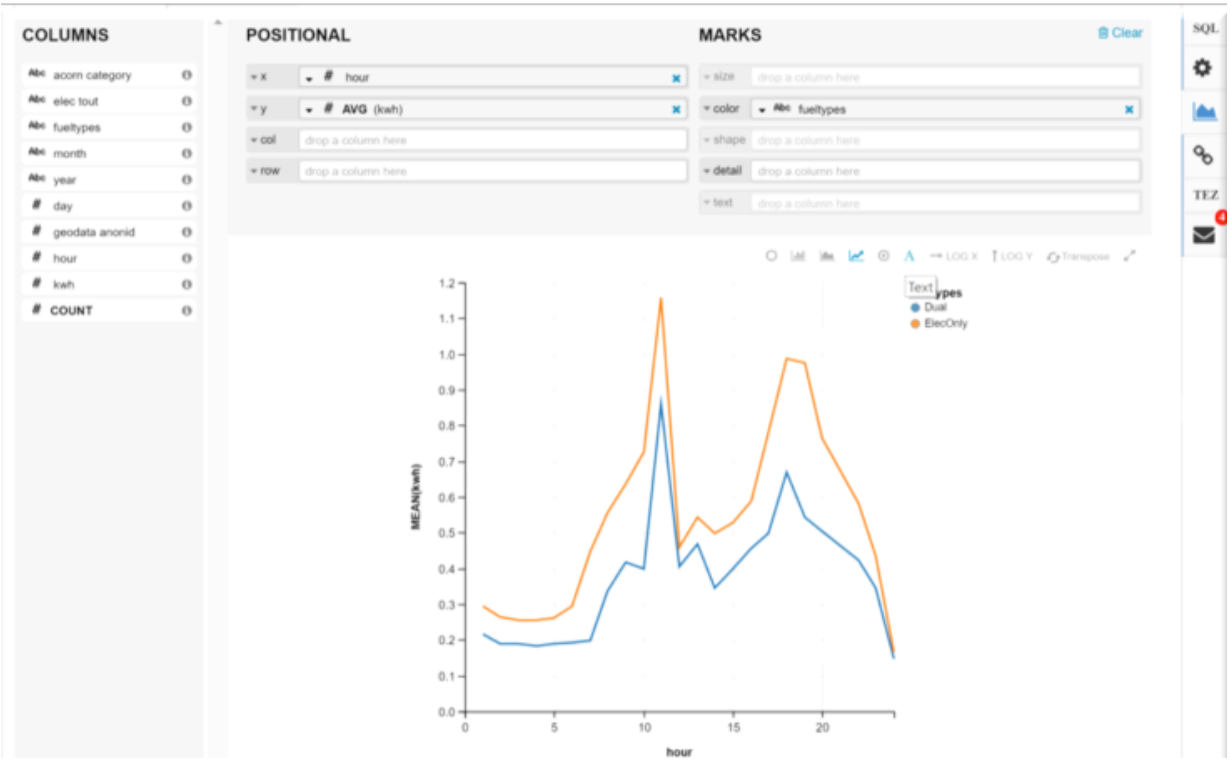


Figure 3 Comparing mean energy consumption, December 2009

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.

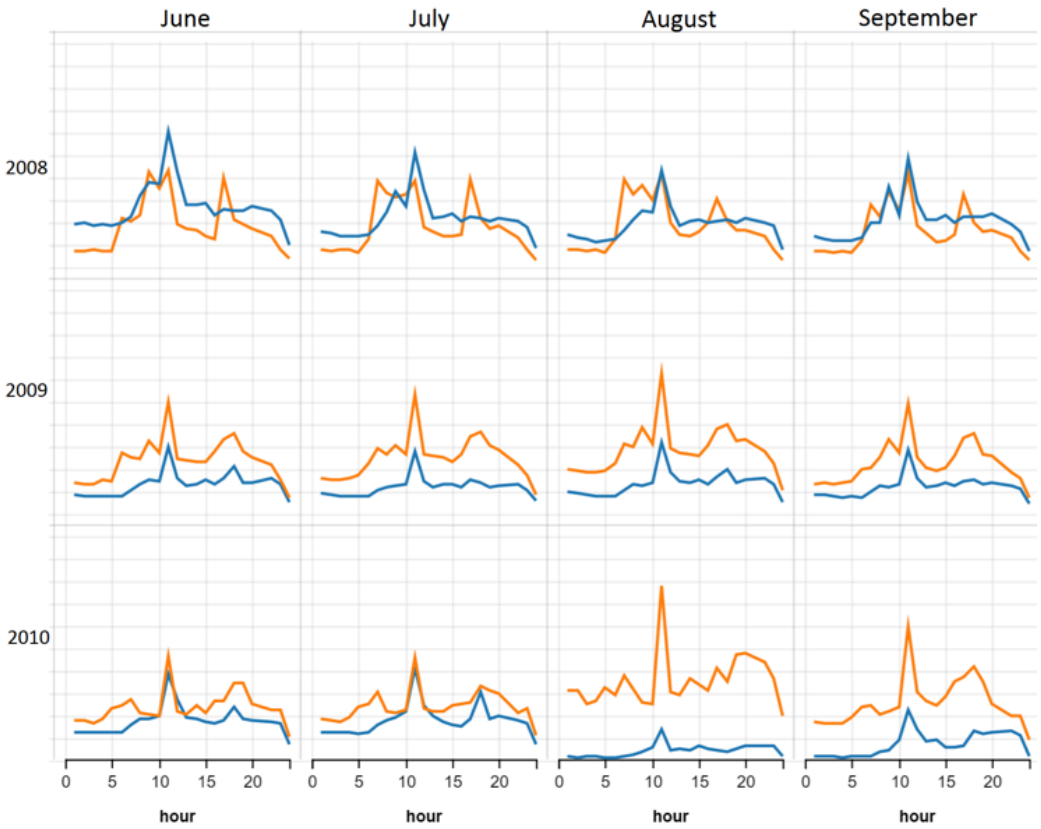


Figure 4 Energy Curves, June-September, 2008-2010

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.

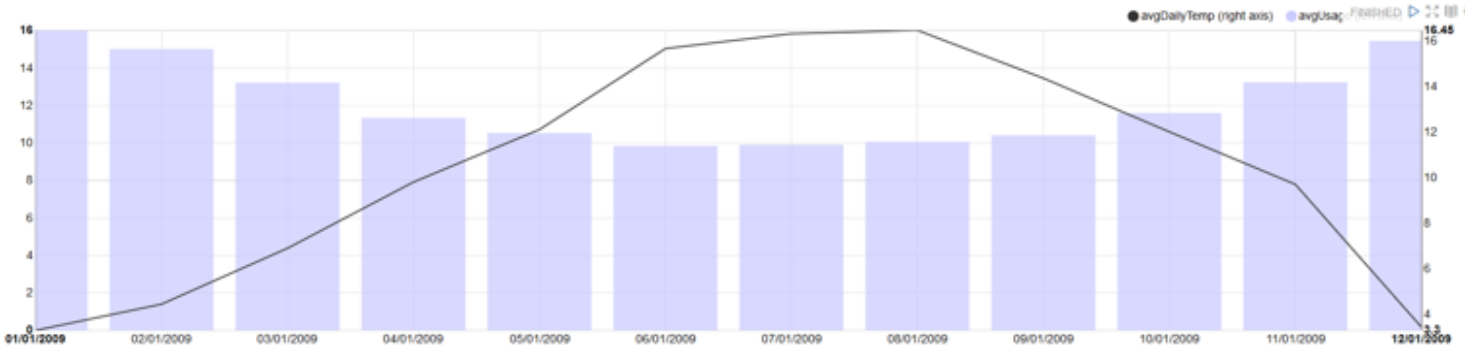


Figure 5 Line over bar monthly mean temperature and energy consumption

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