New Tricks? Old Tools to Solve New Problems: A Pilot Project Using UDC at the UK Data Archive

Suzanne Barbalet, CESSDA Thesaurus Development & Bibliographic Services Officer
Nathan Cunningham, Functional Director for Big Data Network Support, UK Data Service

An important characteristic of classification schemes is that they are stable structures so it might seem something of an anomaly to consider the introduction of a classification scheme into a data archive that is preparing to meet the challenges of new and novel forms of data (NNfD), particularly when such a system of collection organisation has not been used before. However, the very strength of a classification scheme is to "act as a model and a map of the domain" (Broughton, 2016: 339) and in this role it appears to be a useful tool to assist with organising and browsing new subject domains that NNfD may introduce us to and to complement keyword control of content of collection materials by thesauri. While generally favoured as a 'mark and park' tool classification schemes have been used online to organise topics for gateways such as Intute[1] or for the organisation of self-deposit documents in the Research Repository schemes[2] so it is not a path untrodden.

'Future-proofing' Access to NNfD

At the UK Data Service we prefer to refer to new and novel forms of data (NNfD) and 'smart data' rather than 'big data'. Smart data is wide data (high variety), not necessarily deep data (high volume) and it is made up of "feature-rich content and context (time, location, associations, links, interdependencies, etc) that enable intelligent and even autonomous data-driven processes, discoveries, decisions, and applications (Borne, 2016).

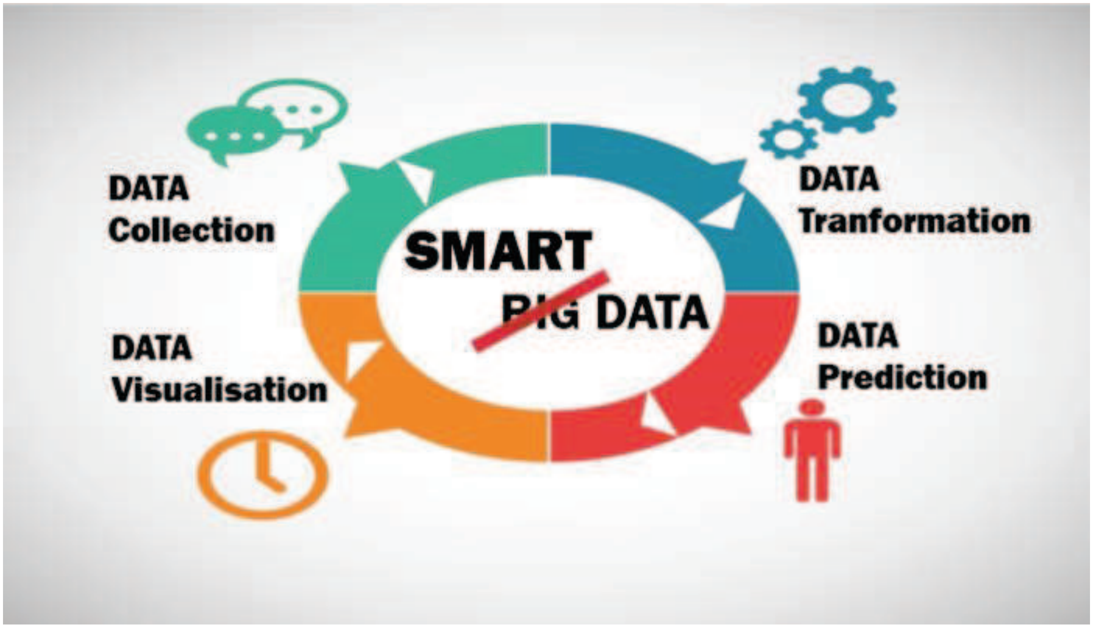


Fig. 1: Smart Data

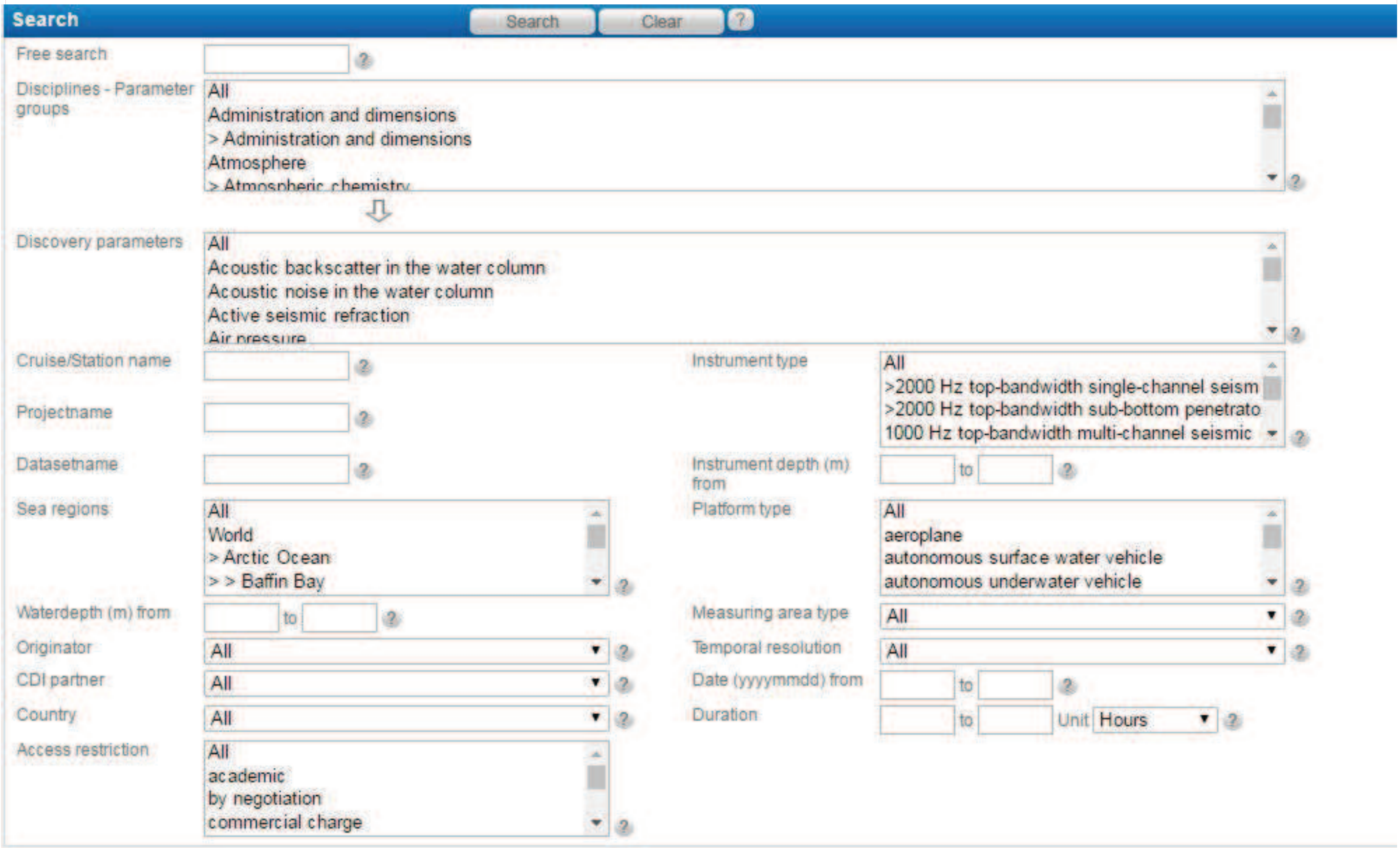
[1] https://digital.humanities.ox.ac.uk/project/intute

[2] For example http://repository.essex.ac.uk/

While ‘big data’ has been used by commercial organisations for market research for many years the use of such data for social sciences and humanities research is relatively new. As we acquire NNfD our role as a data service is becoming more proactive. The UK Data Service and the South African data service DataFirst have jointly undertaken a research project[3] that provides a model for social science research. This research on fuel poverty analyses government data in two countries and is a particularly good example of data services exploring the potential of smart data. In turn this work provides some insight into the metadata that best supports data discovery systems in the future.

Our staff have been encouraged to acquire additional skills. They now have the skills to enable them to identify and scale the data for access, linking and analysis and to ensure that data conforms to Research Data Management (RDM) principles which underpin all UK Data Service archiving and curation work.

We are aware that our traditional clientele of academic researchers, policy makers, educators and students will broaden in the future and that their research needs may not be limited to what can be discovered in an online catalogue. It will be important to create user profiles to explore what aspects of NNfD users from outside our traditional community may require and consider the best way we might meet their future needs. For example, the Natural Environment Research Council (NERC) engages a similar multidisciplinary research network, makes accessible cutting edge quantitative research and its vocabulary service needs to be flexible enough to respond to new policy issues. Here topic access has proved useful as an initial point of entry for a cross-disciplinary search.



Free search

Disciplines - Parameter groups: All / Administration and dimensions / > Administration and dimensions / Atmosphere / > Atmospheric chemistry

Discovery parameters: All / Acoustic backscatter in the water column / Acoustic noise in the water column / Active seismic refraction / Air pressure

Cruise/Station name

Projectname

Datasetname

Sea regions: All / World / > Arctic Ocean / > > Baffin Bay

Waterdepth (m) from … to …

Originator: All

CDI partner: All

Country: All

Access restriction: All / academic / by negotiation / commercial charge

Instrument type: All / >2000 Hz top-bandwidth single-channel seism / >2000 Hz top-bandwidth sub-bottom penetrato / 1000 Hz top-bandwidth multi-channel seismic

Instrument depth (m) from … to …

Platform type: All / aeroplane / autonomous surface water vehicle / autonomous underwater vehicle

Measuring area type: All

Temporal resolution: All

Date (yyyymmdd) from … to …

Duration … to … Unit Hours

Fig. 2: Example of a Topic Search

[3] https://ukdataservicesmartenergydata.wordpress.com/ Accessed 17/02/2017

It was the need for such flexibility that led us to consider a proposal to 'future proof' our subject category organisation by introducing a classification scheme to reduce legacy work involved in updating a flat list of subject categories or topics.

Why has subject or topic access been an important discovery tool for the UK Data Service? Data archived for the purpose of secondary data analysis have a particular set of discoverability issues that is independent of changes in the range and variety of data available for social science and humanities research. For example to ensure the validity and reliability of the survey methodology the concept attached to survey questions and scales in social science research in particular will not always be apparent. A high proportion of survey questionnaires embed standard scales to develop a measurement of a variable in the study. Scales may not be described in the documentation nor easily identified in the questionnaire. Even if identified the variables themselves may not appear relevant to the 'study' topic until the research design is described in the resulting publications. The range of indicators of ethnicity is one example. The UK Office for National Statistics recommends 'country of birth', 'nationality', 'language spoken at home', 'skin colour', 'national/geographical origin', and 'religion' as indicators of ethnicity. Thus we index our studies[4] to variable level applying keywords from our thesaurus Humanities and Social Science Electronic Thesaurus (HASSET) and supplement subject access by also allocating subject category terms from a separate controlled vocabulary.

An acknowledgement of these discoverability issues can be found in the Data Documentation Initiative (DDI), the international standard tailored to describe data. DDI provides both a keyword field, which we populate with HASSET keywords, and an optional subject field.

Our subject controlled vocabulary is a flat list which requires periodic review to ensure that topics reflect new areas of funded research and current policy issues. If we managed it by using a classification scheme to introduce structure into this controlled vocabulary then editing the list would be a simple task. A pilot study was undertaken in 2014 - 2016 to assess the feasibility of classifying the whole data collection before the size of collection made this task uneconomic.

Research data in the digital age is growing exponentially (Corti et al 2014:1). It appears that the application of exploratory analysis using NNfD will become an integral part of the process of curating data for social science research and will impact on the way we manage our resources and on the services we provide. The UK Data Service has developed expertise in automatic indexing and improved the quality of our thesauri. The UK Data Service platform that is nearing completion will provide data analysis features and harness existing knowledge organisation resources to create a vocabulary service. It will not be an exclusively open data platform. The 'data lake' will be a secure repository capable of storing raw data in any format and will facilitate variable searches to the lowest possible level of granularity. Support will be provided for users with expertise ranging from the student to researchers who want to develop their own bespoke data tools. All will require efficient and reliable access tools, not only to locate variables but also sources of open raw data.

Given that a simple keyword search box is the way most searches are undertaken, and with increasingly large recall, attention has been drawn to the problem of subject searches. The question is posed "are we too obsessed with the notion of providing access to everything at the expense of the quality of the results?" (Tay, 2016: 113). Can we enrich such searches with tailored metadata?

[4]. We allocate a study number to each set of data and its documentation as it is processed. It may be either quantitative, qualitative or historical research. It is a consecutive number that retrieves all datasets and documentation relating to the particular research project.

The Pilot Study

We chose Universal Decimal Classification (UDC) for the pilot study as it is the scheme of choice in the UK for large special media libraries, but most importantly because it is an analytico-synthetic classification, that is its constituent parts can be analysed or parsed in the process of information retrieval. Where a specific new topic has not been included in the main tables it can be covered with the addition of an appropriate extension from the auxiliary tables. In addition, the notation allows the linking together of numbers from the main class.
We arranged ‘studies’ for processing by the main subject category allocated to each study in our current system. Series and virtual ‘studies’ were excluded from this pilot study sample, as they are easier to classify, so we selected non-series ‘studies’ for classification in order to gain the best estimate of the cost and time the task would take. Thus, in the main, efficiencies were made in the range of classification numbers required.

It was decided to make the codes as specific as possible until we knew how many auxiliary numbers would be required. We found the option to build numbers to include extra subjects particularly useful. When required we can retrieve these secondary subjects and allocate the data set accordingly to the extra or new category while still retaining all the required information in the full classification number attached to the study.

We included ‘form’ auxiliaries to explore the use of the notation to link data to resources that may be useful for training and outreach purposes; for example case studies, illustrations/photos or bibliographical references. We also added auxiliaries of ‘place’ and ‘time’. The latter, we thought, would be useful for easy identification of our historical data.

We created guides for the most popular UDC numbers and found that by the end of fourteen weeks a part-time employee, working 15 hours a week, completed the classification of 4,289 ‘studies’ Now, with access to UDC Online, the results are even more promising. Reviewing the content of one existing subject category via the code label allows for the identification of significant sub-categories which could form a new category in their own right. Within the subject category of ‘Society and Culture’ below we can see that ‘Leisure, Tourism and Sport’ is emerging as a significant new category.

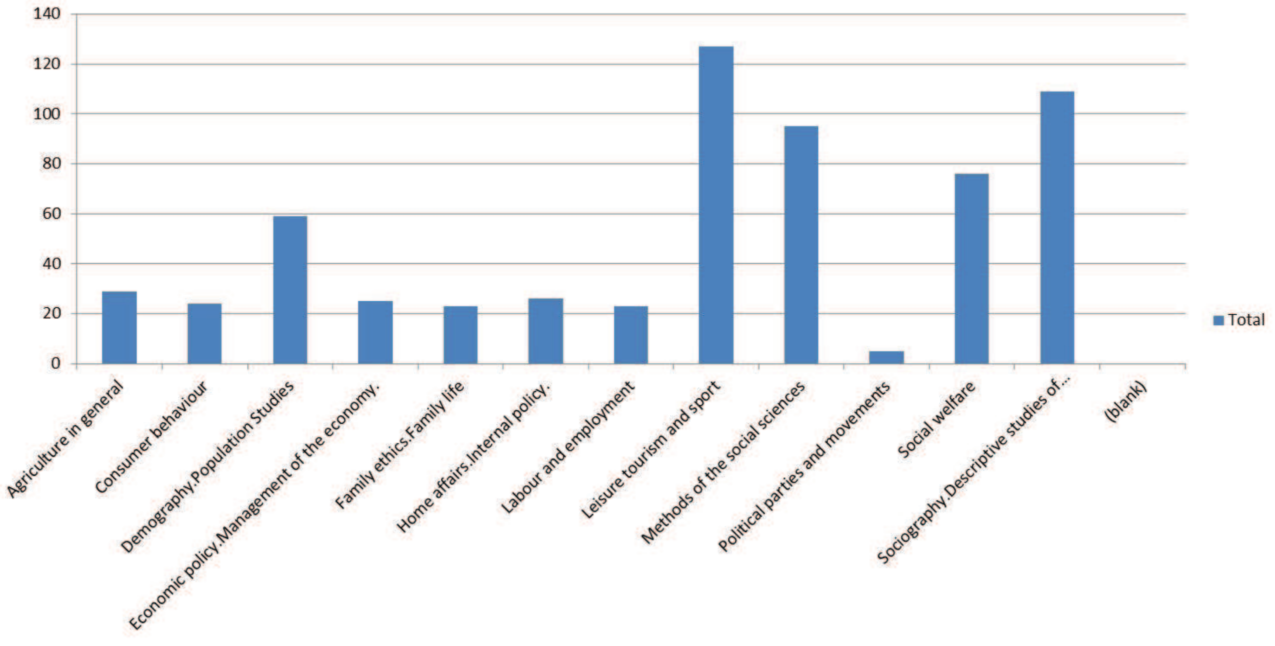
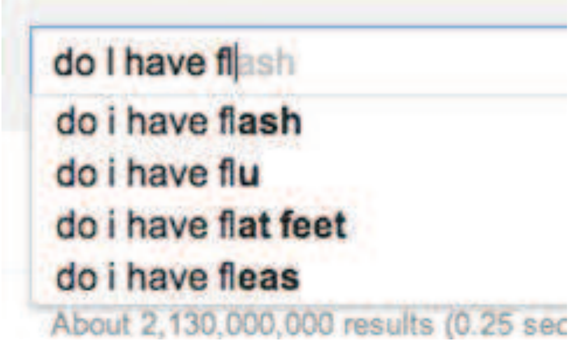


Fig. 3: UDC Notations Distribution for the Subject Category ‘Society and Culture’

‘Big Data’ and ‘Big Metadata’

While there is a popular view that ‘big data’ implies ‘big metadata’ currently there is no evidence to suggest that NNfD will fundamentally affect social science and humanities research methodology and that, in turn, the need for thesauri and controlled vocabularies will be replaced by Google style algorithms. The example of the Google Flu Trends research reflects the initial excitement.



do I have fl|ash

do i have flash

do i have flu

do i have flat feet

do i have fleas

About 2,130,000,000 results (0.25 sec

Fig. 4: Flu Trends (Source The Guardian 27 March 2014)[5]

Researchers warn of ‘big data hubris’ (Lazer, 2014). Google's own autosuggest feature may have driven more people to make flu-related searches and misled its Flu Trends forecasting system.

The future is in the making and we have yet to see exactly what discoverability issues will arise as researchers explore new and novel sources of data but the evidence to date suggests that just as the fundamental principles of social science methodology will hold in this new research environment so too will classic principles of archiving and managing collections of research data for secondary data analysis.

The core strength of classification schemes such as UDC is that a classification code crosses all borders. The code is not only instantly recognisable but also, wherever it is used to classify research data, the opportunity is there to share subject metadata.

References

Barbalet, S. (2015) Enhancing Subject Authority Control at the UK Data Archive: A Pilot Study Using UDC. In Slavic, A. & Cordeiro, M. I. *Classification and Authority Control. Proceedings of the International UDC Seminar.* Ergon-Verlag

Borne, K. (2016) Rocket-Powered Data Science: Data Reflections. http://rocketdatascience.org/ Accessed 24/02/2017.

Broughton, V. (2015) *Essential Cataloguing*. 2nd ed. London, Facet.

Corti L. et. al. (2014) *Managing and Sharing Research Data: A Guide to Good Practice*. Los Angeles: Sage.

Hjorland, B. (2017) Subject (of Documents) *Knowledge Organisation* 44 (1): 55-64.

Lazer, D. et. al. (2014) The Parable of Google Flu: Traps in Big Data Analysis. *Science*,14 Mar, 343 (Issue 6176): 1203-1205 http://science.sciencemag.org/content/343/6176/1203.full Accessed 27/02/2017

Slavic, A. (2006) UDC In Subject Gateways: Experiment or Opportunity? *Knowledge Organization* 33 (2): 67-85.

Tay, A. (2016) Managing Volume in Discovery Systems. In Spiteri, L ed. *Managing Metadata in Web-Scale Discovery Systems.* London, Facet.

[5] See https://www.theguardian.com/technology/2014/mar/27/google-flu-trends-predicting-flu