MANCHESTER 1824

The University of Manchester

Cathie Marsh Centre for Census and Survey Research

Variables, Datasets and Finding what you want: Developing Online Search Tools

CCSR Working Paper 2008-11
Anthony Rafferty and Sam Smith
s.smith@manchester.ac.uk

A common problem when searching repositories for secondary microdata is finding useful data to meet specific requirements. Variables are a fundamental building block of data analysis and usage. This paper covers the benefits to users from a search system that generates information and cross-references for variables in each file in the 650 (and growing) large-scale UK Government datasets supported by ESDS Government and the Samples of Anonymised Records. Use of broad but highly targeted search combined with the integration of a variety of sources of data, documentation, and metadata facilitates a powerful search platform.

Variables, Datasets, and Finding What You Want: Developing Online Search Tools

Anthony Rafferty*, Sam Smith†
Cathie Marsh Centre for Census and Survey Research‡
Samples of Anonymised Records & ESDS Government

IASSIST 2008

Abstract

A common problem when searching repositories for secondary microdata is finding useful data to meet specific requirements. Variables are a fundamental building block of data analysis and usage. This paper covers the benefits to users from a search system[1] that generates information and cross-references for variables in each file in the 650 (and growing) large-scale UK Government datasets supported by ESDS Government[2] and the Samples of Anonymised Records[3]. Use of broad but highly targeted search combined with the integration of a variety of sources of data, documentation, and metadata facilitates a powerful search platform.

*anthony.rafferty@manchester.ac.uk
†s.smith@manchester.ac.uk
‡www.ccsr.ac.uk
[1] `www.ccsr.ac.uk/esds/variables`
[2] `www.esds.ac.uk/government`
[3] `www.ccsr.ac.uk/sars/`

Contents

1 Introduction

This paper discusses the effects of a system built to display information about variables in a highly accessible manner - letting users find the information on data that they want, when they're looking for it. Users go to google[4] for what they're looking for. This paper covers one of the methods of how we help them find it.

The processes discussed here are conceptually simple, but are run on a large scale. Additionally, when we find an issue with one dataset, we run through all our data to find and correct similar issues elsewhere.

At the simplest level, what this system does is create a web page containing a univariate distribution of every variable in every dataset, and put them all on our website. We then build tools where those pages can be found by users doing research.

One of the biggest things that our users lend us for a little while is their attention. And they want it back pretty quickly. They come to our site to look for something, and want to get on with their research. Publishing metadata to the web in a usable fashion allows a wide range of benefits, internally from the ability to create enhanced services, and even more value from giving users the ability to use the strengths of the Internet ecosystems[5].

Some basic consideration has been given to extending the system to take advantage of emergent and novel ideas. Similarly, such a metadata platform lets you build things on top of it which are of direct value to users, and are impossible without the strong foundation.

Data analysis starts with a research question, "How has employment of Pakistani and Bangladeshi women in the UK changed over time?"'. From there, data is[6] found which provides the required information to examine that question. With a wide variety of data sources for UK academics, questions of quality, suitability and the real world issues of accessability come into play. There is a very great temptation on the part of the researcher to use the same dataset that they used for their last piece of research, with the immediate benefits of familiarity and understanding that come from experience, rather than spend significant time looking for a new dataset which may or may not exist.

While, as data custodians, we care about datasets, the research community cares about

[4] or another search engine, but pretty much google

[5] for those playing the bingo cards, this isn't Web2.0, it's barely even Web1.0

[6] or, at least, should be

issues. We have a survey containing y datasets; users wish to talk about *employment*, *health of elderly people*, or *Pakistani and Bangladeshi women's employment in England*. These issues use varying subsets of data based on the specific question, and cut across the lines.

Our primary tool allows researchers to put in the word they're looking for, and be shown all the datasets, surveys and their variables which contain that word[7], which is one of a number of services building on the existence of the pages themselves. Another allows users to put in 3 terms of interest, and find the datasets in which all those terms appear within the same dataset.

We can build on top of these tools, services such as email alerts[8] which notify users of the data they want to know about. These are conceptually simple but impossible to provide without those standard URLs of pages. This system lets users target the precise variable that they're interested in, and what they're looking for drops into their inbox, without having to give prior consideration to which dataset discusses it, or having to perform regular searches. When they get that email, they can click through to the variable in which they're interested.

There are implications for backend processes - we have the ability to take a zipfile of data and extract all the metadata. We can then rerun a similar process on newly deposited datasets (either another cross-section in a series, or an replacement deposit) and compare the metadata we get from that process to the existing dataset. This lets the data processor have a definitive list of changes between the two datasets based on the data itself.

Most services discussed here are freely available over the web, although the datasets they point to may not be available to non-UK academics. While we will provide URLs to specific pages where their features are discussed, you may wish to spend a few minutes having a look round to get a more rounded view. To make this paper clearer, we recommend that you look at the URLs referenced in the text before reading on them.

This paper describes the system built by the UK's ESRC Census of Population Programme[9] Samples of Anonymised Records (SARs) project[10] and extended by the Government sub-service[11] of the ESRC funded Economic and Social Data Service. While the SARs and ESDS Government are separate projects, they share the majority of staff and have an extremely close relationship, with ideas and innovations from one being shared. While most examples in this paper use ESDS Government, this is simply due to

[7] see section X.Y
[8] and RSS
[9] `www.census.ac.uk`
[10] `www.ccsr.ac.uk/sars`
[11] `www.esds.ac.uk/government`

it's larger size and more diverse and eclectic ecosystem of services.

This paper is the overview of what we do, what it makes possible and why it is useful and important. A companion working paper (designed to be read by those who have already read this one) contains significant low level practical detail . We start here with the high level view of the whole system and how it is used, in the second paper, go back through the processes in detail, benefiting from the knowledge of the big picture. This is vital as decisions at early stages have direct relevance to later outcomes. However the second paper need not be read all all in order to understand this one.

1.1 Acknowledgements

We must thank all those at ICPSR[12] for inspiring and detailed conversations when we visited them in 2006. Special thanks must go to Peggy Overcashier and Cole Whitman, who provided clear insight into what was possible given what they were doing, which was far in advance of what we had considered previously. We should also thank Kevin Schurer of the UK Data Archive, and especially Tanvi Desai of the LSE. As ever, Angela Dale, Gillian Meadows, Jo Wathan and Vanessa Higgins of ESDS Government have provided expert and vital feedback and ideas, sometimes they weren't even much work. ESDS Government must also thank the SARs[13] project, as part of the UK ESRC Census Programme[14] for the infrastructure, designs and time.

1.2 Terminology

ESDS Government supports 25 surveys, each of which is a time series of cross-sectional surveys dating back as far as the 1970s. A dataset is one cross-section of one survey, and there can be any number of datasets per year, depending on the survey frequency, and these surveys change significantly over time potentially without any name change in variables. ONS policy (oft imitated) means that any change in a variable generally leads to its name being changed, at least for the more recent surveys, which gives us some level of confidence that two variables in different cross-sections of the same survey are likely to be somewhat comparable. That does not necessarily hold across different datasets. The SARs are similarly designed datasets consisting of anonymised microdata from the 1991 and 2001 censuses.

A univariate distribution gives the range and spread of the values for a variable e.g. by

[12] `www.icpsr.umich.edu`
[13] `www.ccsr.ac.uk/sars`
[14] `www.census.ac.uk`

listing frequencies of each value.

2 650 Datasets across 25 surveys and 2 projects in 1 system

Our users are generally researchers academics and others looking to answer research questions. The system described here does not make data available, nor does it necessarily make any additional metadata available above that which could be made available by other means (although it may make such metadata much more useful).

What we have created is a webpage for each variable in every dataset, for every survey we support - e.g. for variable *actwkdy*2 from Quarterly Labour Force Survey Household Dataset, March - May, 2006, is `http://www.ccsr.ac.uk/esds/variables/lfs/lfs5441/actwkdy2/`

This page includes a simple univariate distribution of the variable[15], links to where it gets referenced in the documentation, links to other services' pages about that survey, and those links are direct to variable level information where they have it and make it available.

We also build variable indices at the dataset level, and variable and dataset indices at the survey level to allow browsing. These indices, operating in a tree like structure, make it easy for us to build additional services based on predictable URLs, and, more importantly, allow search engines to find those pages. We also link to the pages in other datasets for variables with that name[16].

2.1 Google It!

While the most fastidious of data curators who are leading experts within their fields will add significant value for metadata, it is undoubtably the case that what novice users search for may not be added by those who think in far more advanced and nuanced terms.

[15] for overview purposes, we pretend that all variables have a univariate distribution, see later in the paper for how we deal with reality

[16] this is a dramatic oversimplification of what we actually do, which uses both the obvious enhancements and more

When you put the vast majority of our variable names[17], into Google[18], again using the example *actwkdy2*, the top result is the page that we discussed in the previous section. 'Googling' is the natural way for many people to now find their information, and now all our datasets are part of that natural workflow. If they want to find more information about it, it's right there, the same way you would find something else. The cross linking of variable to all relevant resources, irrespective of the organisation which produces them, aims to make the page as useful as possible and this benefits the user.

One of the main challenges of data support staff is to get information about data to where the users are. We can have the greatest services in the world, but if users need to go out of their way to find them, we will be spending far more time on outreach than building things and could never hope for services to be mainstream within our community. However, if we can easily find a way to integrate what we do into the normal workflow of users - e.g. google - that makes outreach far easier.

2.2 Attention!

One of the biggest things that our users lend us for a little while is their attention. And they want it back pretty quickly.

They come to our site to look for something, and are reading a specific set of pages, which gives us a very good indicator of what they're interested in. If they're looking at the Welsh Health Survey (WHS) pages[19], that clearly indicates that they're interested in the WHS. It is therefore useful to both users and to our aims to give users more information about the WHS while they're there. Whether it be a link on how to *Cite this data*[20], or notification about the latest dataset published for that survey. This is an extension of ESDS Government practice of putting *Latest News*[21] on the right of pages[22], and cross-linking to other Events, Data Releases and Publications.

One wall that we seem to be perpetually banging our heads against is getting users to cite the data that they use. This has many benefits[23] and needs to be made as easy as possible. Having all the datasets in a database makes it very easy to find the earliest and latest years available for any survey. A regular automated task then updates the

[17] those which aren't real words and make sense in context

[18] or your other favourite search engine

[19] `www.esds.ac.uk/government/whs`

[20] `e.g.www.esds.ac.uk/government/lfs/cite`

[21] `www.esds.ac.uk/government` second box on the right

[22] excluding those where we need the space for content

[23] for us, less so for the researchers, which is why it's hard. For what we do with citations, see `www.esds.ac.uk/government/citations`

web pages to always include the correct dates. Users can then be encouraged to cut and paste the online citation changing the dates[24]. If you can think of any ways of making this any easier for users, please let us know.

Detail is stored in the webserver logs of what pages people looked at, which were popular search engine queries that led them there etc. While this is not information which we have looked into at the time of writing, we expect that it will provide a useful view of what users are looking for on their own, and may provide novel insights when compared to what they look for when provided with an equally simple box to search the default catalogue[25].

While metadata is good, and more metadata is better, there is a limit to the amount that can be feasibly created by any organisation; sharing and cross linking benefits everyone. Build things to be reused into the future, and to be findable by the same processes our users

2.3 Building everything

The search engine takes metadata from a few different systems, and then produces our pages, where possible linking back to the source locations to find out more and see context. At a high level there is a large black box, in one end goes a set of SPSS files for a dataset, and out the other end comes a lot of webpages. Rather than being a black box, it is, a data pipeline, as introduced in papers from Cole whitman at ICPSR [26]. Data goes in at one end, and information comes out at the other. With all intermediate stages being driven by the input.

While this paper discussed the output, there has been no discussion so far of how it is produced. While here we discuss the concept of the data pipeline and how it works, the overview we present in the following few paragraphs is necessarily simplified, the technical detail is covered in the companion CCSR working paper. However, if you are only superficially interested in the mechanisms and care more about the output, just skip the companion paper[27].

It is worth pointing out that, at this point, the whole system runs in a fully automated fashion. The system picks up new datasets and pushes them through the pipeline. All stages are fully automated and a dataset can conceivably go from publication by

[24] if they remember

[25] `http://www.esds.ac.uk/search/searchStart.asp`

[26] `www.icpsr.umich.edu/ICPSR/org/publications/staff/ProcessMapping.pdf` and related papers

[27] However, if you think something in this section can't *possibly* work exactly as described, you may well be right.

UKDA to being included in our system without any manual intervention. A significant advantage of this full automation is that, as more features are added, it is easy to enhance the entire system by simply passing all datasets back through it. They are then reprocessed to get the latest new features as for all new datasets While this takes time, it is fully automated and so can run in the background without any significant staff time to do it.

We have a script which watches the UKDA website for new datasets that are not present in our database. When it sees one appear, it adds it and flags it with the current time[28], and whether it's new or an update to an existing dataset.

This dataset is then picked up by another regular process, which downloads the SPSS edition of the dataset to the local system for processing. It is then unzipped, and a list of SPSS data files within the extracted archive created for processing based on their filename. While UKDA generally also makes available a Stata dataset (amongst potentially others), we use SPSS as it is a somewhat consistent standard for all our datasets. What we discuss below is possible with other data formats, however we picked SPSS for convenience. The only time we use the SPSS software, rather than just opening SPSS files, is when we need to work around errors in the data formats.

To obtain the values and labels for the data from the SPSS files, we use scripts to provide the variable level information about labelling. SPSS portable files are either converted to SPSS .sav format[29], or run through the *xlabels* script written by Frank Stetzer of UW Milwaukee[30]. SPSS .sav files are processed by the *spssread.pl* script by Scott Czepiel[31] which produces a tab seperated file of information about the variables or values. These two outputs are merged to provide the univariate distribution information, along with variable and value labels. That is the extent of the information that is stored in most SPSS files. The distributions of each variable are obtained using R[32].

The internal format that is produced is substantially similar to the format used by Survey Documentation and Analysis System (SDA) at Berkeley[33] with decisions made for the same reasons, although, as they were made independently, there are negligible differences. While DDI is an excellent interchange format for sharing data between organisations and for forward archiving of final metadata, it is both hierarchical and complicated in many ways that the univariate distributions aren't. As a result, while we can both import from and export to DDI without a problem, internally, we use

[28] this provides the hook which makes our email alerting service works - `www.esds.ac.uk/government/shes/datanotify` - another easy attention grabbing system

[29] for why this happens see the detailed process

[30] and available from `ftp.uwm.edu/pub/stetzer`

[31] `http://czep.net/data/spssread/`

[32] see `www.R-project.org` for more details

[33] `http://sda.berkeley.edu` - a great system that we don't use

something that's only as complicated as we need.

The UKDA publish a number of PDFs at the dataset level containing a huge amount of information such as user guides, variable codebooks and derivation or collection information. For the variable page we used as an example above[34], there are 10 PDFs plus two text files of documentation, with the PDFs totaling 2244 pages, into which users should not be expected to dive into without some reasonable indication that what they're looking for is in there somewhere. Additionally, novice users are not going to know where to start (page 1 of which file?) and will probably just ignore it until they can't.

The univariate distributions we have previously created contains a list of variable names, such as *actwkdy2*, the vast majority of which have an interesting names. What knowledgeable users would do is simply look through the PDF for that word, and ignore everything else, less experienced users will search a PDF for a keyword which may or may not be there. Converting the PDF content to text[35] allows us to completely automate that process for each variable, and produce a list of pdf files and page numbers in which each variable name is mentioned. We can then link to the right page in the PDF. While there are a number of false positives for some variables, and we often pick up references to variables where they have been used to derive others, this is a highly useful and extremely easy process, and one user's false positive may be another's serendipity.

Cross linking to compatible external services requires predictability of URLs, which is something that the majority of the UK academic support community does extremely well. This allows us to know where content is, and point to it without having to actually go out and find it.

The newer versions of Nesstar[36] make the DDI metadata behind the interface easily available. Combining that with the list of datasets published to the server, we can easily import the DDI published to Nesstar into the system[37]. This also provides the ability to link to specific variables within Nesstar where they are present. Additionally, when new datasets are published into Nesstar, we automatically identify them for reuse.

Once the variety of metadata sources have been collected and processed, it is all loaded into a relatively simple database from which the web pages are created. The use of the database allows for the searches required to be relatively simple[38] and also allows for additional services to be built on top.

[34] `http://www.ccsr.ac.uk/esds/variables/lfs/lfs5441/actwkdy2/`
[35] actually XML using pdftohtml `pdftohtml.sourceforge.net`
[36] version 3.5 or better
[37] This is where the various bits of *Universe* and *Literal Question* fields come from
[38] the use of XSLT to do this would be extremely complicated

The variable webpages visible to the user are then generated from the database and written to disk. While they could be built on request, the permanence of generated pages allows the database to reflect the current state of the data, without concern about variables being lost during data updates[39]. Once these pages are created, they exist as a platform upon which to build additional services at any or many levels from a survey time-series down to a particular variable in a particular file. As metadata and these features get used more heavily, it finds more errors and inconsistencies in the metadata. Not because they weren't there before, but because no one noticed. More use makes metadata better.

2.4 A sense of scale

In terms of numbers, for ESDS Government, there is currently 482Mb of internal metadata created, built from 6.8Gb of PDF documentation, 758Mb of DDI metadata downloaded from Nesstar, and 11Gb of zipped source data (which expands to 62.6Gb when unzipped). The core of the system, the parts which do the SPSS processing and data loading, plus some codebook generation, is shared between ESDS Government and the SARs and totals aroudn 3000 lines of perl and shell scripts.

The ESDS Government specific portions, the areas which deal with the download of data, metadata, config creation, parsing of custom metadata and other sources is another 3000 lines of perl and shell scripts.

The system was original developed for the SARs website, and extending the code to do the basics for ESDS Government took about 2 days. By comparison, to download the data for the biggest survey, the system covers[40] took 6 days (running quietly in the background). That initial system would be immediately recognisable to our users today, even if some additional novel features were missing.

Communication overheads are minimised by attempting to autodetect as much of the information from our partner organisations as possible. This means that we are not reliant on the resources of partner (or just friendly) organisations to do any part of the work - we simply make use of what they do anyway.

[39] see later in the paper for a detailed discussion of this

[40] The Labour Force Survey - `www.esds.ac.uk/government/lfs`

2.5 International Comparators

We have previously mentioned the huge debt that we owe to the ICPSR[41] for their ideas. They have a similar system, which, at time of writing operates on a subset of their datasets as a demonstrator, and may be extended based on their successes.

We use the web-based Nesstar system[42] within the SARs project for data exploration; ESDS Government's data exploration is also through Nesstar run by the UK Data Archive[43]. Nesstar allows easy browsing of data and analysis for authenticated users[44]. Due to Nesstar's heavy reliance on Javascript and magical appearances of dynamically built menus, Google is unable to index even the public metadata published via Nesstar. While Nesstar is extremely good at what it is designed to do (easy access to data for people), this system is designed to compliment our Nesstar services[45], by providing the links from the higher level information. If it's in Nesstar alone, it will not be found unless users are using your search engine to look for it, or know it's already there. Those aren't the people our service aims to help.

CESSDA[46] has a pan-European data catalogue based around Nesstar and a multi-lingual theasuarus. While the current implementation uses Nesstar, and the data interchange format is DDI, it should be possible to integrate the ESDS Government datasets that do not appear in Nesstar to allow them to be found by the CESDDA system without the investment in Nesstar infrastructures and the staff time to load each dataset from the back catalogue.

This may be an area for future work since a significant portion of our datasets are not within the CESSDA system due to not being present in UKDA's Nesstar. ESDS Government datasets in Nesstar likely to be recent than the older members of series, which provides a disincentive[47] to researchers looking across time.

One advantage of this system is that it uses only free components, and so may provide an effective solution to institutions and organisations with low levels of resources. Because of the automation, it is easily and cheaply scalable to add additional datasets or countries without any consideration of license fees which could be an issue in some environments. Especially where the online exploration possible through Nesstar isn't immediately feasible due to a lack of supporting infrastructurs - such as authentication

[41] but we'll mention it at least another time later on to make sure that you didn't miss it
[42] `http://nesstar.ccsr.ac.uk/webview`
[43] `http://nesstar.esds.ac.uk/webview`
[44] For more on Nesstar, see `www.nesstar.com`
[45] We use Nesstar, and expect to continue to do so for the forseeable future
[46] The Council of European Social Science Data Archives - `www.cessda.org`
[47] as if any more were needed

etc.

2.6 Technical considerations

The above overview, while hinting at the issues below, paints a simple and pretty picture of the dataset world. Pretending that it's all wonderful, datasets are tidy, and there aren't landmines in the middle of that nice picturesque pasture. Some of our data is more than 30 years old; some of our data makes little sense out of context, and some of it is as it was. Some of it has been converted from format to format a number of times over the years, and we're told nothing went wrong at any point in that process; and everything is just fine.

Reality has the landmines.

For details of the technical implementation, please see CCSR Working Paper 2008-11[48] which covers this in some detail.

2.7 History, SPSS and future implications

Datasets are often provided in one format, and then converted into the formats requested by users. While we currently get the majority of data deposited in SPSS format, historically this has not always been the case[49]. As a result, over time, the datasets have been run through a variety of processes to convert from one format into another. Possibly multiple times.

With the impending arrival of grid technologies, one of the things for archivists to be aware of is it is likely that their users will not be running SPSS. More generally, the software R[50] is becoming more widely used for statistical analysis, and R's input routines for SPSS software follow the exact specification for a SPSS Save or Portable file, and may not be able to open a file which is not what it claims. SPSS, on the other hand, is extremely liberal in what it will accept. If you have an old style SPSS export file, and simply rename the file to end *.por*, then SPSS will happily open it, and allow you to work on it and do the conversion to make it a portable file when it saves it.

[48] `http://www.ccsr.ac.uk/publications/working/`
[49] after all, what software were you using in 1975?
[50] `www.r-project.org`

As a result of this[51], it is possible, and in fact, easy to create a invalid dataset that SPSS will open correctly, but other software will not. Previously, due to the prevalence of SPSS, that has not been a significant issue. However, it is likely that such shortcuts will increasingly cause problems for users.

The simple and obvious solution to this was to run all the datasets through SPSS - opening and saving them one at a time. While this could be an incredibly tedious and error prone process to do manually, we created a script which ran through the datasets looking for files that SPSS would open, and created a SPSS Syntax file which opened every SPSS file, and saved them all our as *.sav* files. This has the side effect of making the processing simpler, as everything is the same format, although newer datafiles which are SPSS portable files may be processed directly. The format issues are much more focussed on historic data.

When processing more than 9000 SPSS files sequentially in a single syntax file, you can expect various things to go wrong. SPSS, for some reason, would very occassionally fail to open an SPSS file, and hence resave it's predecessor with the new name[52] so a checksum[53] of each processed file was created, and looked for two consecutive datafiles having the same checksum, with the second file being the resaved duplicate. These few problems were corrected by loading it into SPSS manually.

While there is very significant effort put into the manual creation of accurate and useful metadata, as some stages are a manual process, there is always the prospect of human error[54]. As part of the processing of the metadata we get from Nesstar, we have significant integrity checks against what we get out of the data itself. On a couple of occasions, there had been a minor error in a manual process, and 2 dataset identifiers had become switched. These errors generally don't affect users in their usage of the data, but which are things such as that can be detected are items like file identifiers in metadata being swapped, which affect large scale bulk processes. Where such errors are found generally by a metadata match being tried across sources and producing an error, it is useful to build at the previous stage a checking process which looks for those inconsistencies. It is then possible, and even desirable, to run the automated checking process across all the datasets in the system, which will find (and lead to being corrected) all errors of the kind, and while it may not prevent the recurrance of typos[55], it will lead to their discovery and correction.

Bulk downloads of data are likely to become more common. While variables are still not consistent over time, it is not inconceivable that the more Computer Science orientated

[51] which is in of itself a good thing
[52] This only happened on a couple of files, and was in no way reproducible
[53] we used MD5 for convenience
[54] although, we're fortunate in that this is very rare
[55] and if you know how to do this, please let me know

users, especially those with an interest in e-Social Science, will start looking to do more comparisons across time, rather than just run more complex models on one cross section. As e-Social Science becomes more mainstream, then the dominance of SPSS (and to a lesser extent, Stata) will begin to erode and a more heterogenous software enviroment may appear. In some ways, this is beginning to occur with the rise of R and M-Plus within the research community.

One thing that both projects have experienced is that as metadata and processes get used more heavily, more bugs or inconsistencies are found which no one has noticed before.