

A tree full of leaves: description logic and data documentation

Phil Edwards, Judith Aldridge, Karen Clarke

Uniformity: impossible

The documentation of datasets on a conceptual level, so as to make multiple datasets simultaneously accessible and comprehensible, is a key element of the e-social science project and a prerequisite for the construction of a social science 'Knowledge Grid'. However, this is perhaps more problematic in the social sciences than in other fields. Achieving the goal of a single overarching conceptual vocabulary may be difficult in the physical sciences; in the social sciences the goal itself is deeply problematic, for three reasons. Firstly, the concepts on which the data of the social sciences is constructed are often **imprecise**. In a survey question relating to driving a car while under the influence of alcohol, no two users' definitions of being 'under the influence' are likely to be alike; the data produced by the survey is not directly comparable with the results of roadside breath tests, for example. Nevertheless, concepts which are inherently subjective and not amenable to precise definition may remain valid in their own right, by virtue of the social practices which they represent. There may be no precise definition of 'anti-social behaviour', for example, but incidents of 'anti-social behaviour' can be (and are) reported, enumerated and analysed.

Secondly, as this example suggests, many social science concepts are **contested**: they are defined differently by different groups of users, both among professionals and among researchers. 'Anti-social behaviour' is one example where an explicit definition exists (as given in legislation) but interpretations in practice differ widely. In other cases the definition itself may vary from one group to another. One survey might define 'drug' as any dedicated psychoactive substance, including alcohol and tobacco; another might define it as any socially problematic psychoactive substance, excluding alcohol and tobacco but including solvents; administrative data from police sources might take a narrower view, confining the 'drug' rubric to substances controlled and classified under the terms of the Misuse of Drugs Act. Even assuming an agreed definition of 'drug', sub-classifications such as 'soft drug' and 'hard drug' remain problematic.

Thirdly, social science concepts are **mutable**. Statistical series covering the 2003-4 period will show a steep drop in offences relating to Class B drugs and a correspondingly sharp rise in Class C-related offences, corresponding to the reclassification of cannabis. Nor is this simply a question of changes in higher-level conceptual groupings. The concepts to be found in administrative records change over time as legislative and reporting requirements change; the conceptual vocabulary of surveys also changes over time, as the wording of questions changes in response to perceived changes in society. Until 1980, the standard definition of a 'household' in government surveys required household members to be catered for, for at least one meal a day, by the same person. Even when the concept of shared cooking arrangements had filtered through, it took until 2001 for the concept of 'Head of Household' to be replaced by the less loaded 'Household Reference Person'.

The propensity of social survey vocabularies to change over time is worth lingering over, as it highlights and clarifies the broader issues we have identified. It is easy to identify ways in which older vocabularies fail to adequately capture social reality as we understand it now, but the point here is not that any one conceptual vocabulary is 'wrong': what is interesting and significant here is the continuing process of change itself. Retroactively applying our current conceptual vocabulary to earlier datasets would be an error akin to inferring major changes in drug use from the recent changes in Class B- and Class C-related activity.

It is worth noting, also, that the shifting vocabularies of social surveys introduce two distinct difficulties in comparing successive survey datasets. On one hand, different concepts correspond to different social realities: what is true of a 'Household Reference Person' would not necessarily have been true of a 'Head of Household'. On the other hand, it should be borne in mind that survey vocabularies are changed in response to changes which are perceived to have taken place in society. In other words, the fact that 'Head of Household' was deemed inappropriate for the 2001 Census suggests - if we assume relatively slow processes of social change - that it was already becoming inappropriate at the time of the 1991 Census. The problem, then, is twofold. To some extent 'Household Reference Person'

captures a social phenomenon which did not exist when the term applied was 'Head of Household'. On the other hand, to some extent the old 'Head of Household' concept already captured the emerging 'Household Reference Person' reality, but did so imperfectly; the 2001 change casts a retrospective shadow over the 1991 concept, suggesting that it was not entirely comparable with earlier uses of the same term. This twofold process of social and analytical change cannot but continue: there is no taxonomic fix. In practice, when a major survey release is published, the accompanying metadata often includes not only a definition of key terms, but discussion of how and why the definitions have changed since the previous release. This information is valuable for the social scientist, both as a framework for understanding statistical data and as a body of data in its own right. To document successive survey releases using a single conceptual vocabulary would severely distort the data being documented.

These considerations suggest a broader point about the nature of social science concepts. Our argument is that they are imprecise, their meanings are contested between different groups and that they change over time. However, this does not imply that they are imperfect, that one group's vocabulary is objectively more 'correct' than another's or that change over time equates to progress. Social science concepts grow out of debates as to the nature and structure of social reality, and inevitably bear the traces of those debates. Rather than imposing 'scientific' rigour on the source data, thorough and consistent documentation of social science concepts requires us to reflect the fluid and contested contexts in which they exist.

It follows that the task of conceptually 'grid-enabling' major social science datasets cannot begin by ironing out inconsistencies and resolving ambiguities. Rather, documenting the datasets must include documenting the definitions of the conceptual framework on which the datasets are built, however imprecise or inappropriate these concepts might appear in retrospect. This will also involve preserving - and exposing - the variations between different sources, and between successive releases from a single source.

Taxonomy: irreducible

Conceptual metadata in social science data sources can, in principle, be documented by two main approaches, both of which offer to resolve the problem of conceptual variability. A 'top-down' approach is driven by a single controlled vocabulary of concepts, adopted by agreement between researchers and data users as a usable approximation of the concepts used in different sources. For ease of analysis, a vocabulary of this type can be organised into a taxonomic tree-structure: we might assume as a starting-point that 'drugs' divide into 'Class A', 'Class B' and 'Class C', for example, then go on to classify 'cannabis' as a type of 'Class B drug' and 'marijuana' as a type of 'cannabis'. The lowest-level concepts in the taxonomic hierarchy can then be cross-referenced to the data sources which contain data organised around that concept, or an equivalent concept.

As this example suggests, this approach has a number of hazards. Firstly, there is a risk that the imprecision, contestedness and variability of social science concepts will simply be enshrined in the uniform taxonomy: rather than a precise and unambiguous reference point against which source vocabularies can be measured, this may develop its own idiosyncrasies and imprecisions, effectively turning into one more source-level vocabulary. Secondly, the correspondence between the uniform taxonomy and the conceptual vocabulary of individual datasets may not be exact; indeed, the more rigorously the terms composing the uniform taxonomy are defined, the more inexact the correspondences are likely to be. (The concept of '3,4-methylenedioxymethamphetamine' is more precise than 'Ecstasy'; as a result, it corresponds less well with concepts used 'in the wild'.) Thirdly, a single conceptual hierarchy imposes divisions between closely-related concepts: 'arson' can be defined as a type of 'fire' or as a type of 'criminal damage', but not both. Lastly, the conceptual tree-structure imposes associations between 'parent' and 'sibling' concepts, colouring the user's perception of the field; like the terms themselves, these associations are inherently unlikely to correspond to conceptual associations in any one data source.

A 'top-down' approach is exemplified by the European Language Social Sciences Thesaurus (ELSST). The Madiera portal (MADIERA 2006) allows researchers to explore ELSST and access European survey data which has been linked to ELSST keywords. The limitations of the top-down approach can be gauged from ELSST's concepts relating to drug use. *Drug Abuse*, *Drug Addiction*, *Illegal Drugs* and

Drug Effects are all ‘leaf’ concepts - headings which have no subheadings under them. However, they are in different parts of the overall ELSST tree: for example, *Drug Abuse* is under *Social Problems->Abuse*, while *Drug Effects* is under *Biology->Pharmacology*. Although the hierarchy is augmented by a list of ‘related’ concepts, to some extent facilitating horizontal as well as vertical navigation, the hierarchy inevitably makes some types of enquiry easier than others. Anyone using the ELSST ‘tree’ will be visually reminded of the affinities identified by ELSST’s authors between *Pharmacology* and *Physiology*, or between *Drug Abuse* and *Child Abuse*. These problems follow from the initial design choice of a single conceptual hierarchy.

The alternative is a ‘bottom-up’ approach, retaining the conceptual vocabulary of individual sources and allowing users to run enquiries on this basis. This approach addresses the problem of conceptual variability by preserving and exposing it: the user engages directly with concepts as they are found ‘in the wild’ (*in vivo* concepts). As such, this approach has none of the problems associated with the first approach. However, the absence of an underlying taxonomic structure creates its own problems. A uniform taxonomy offers obvious ‘ways in’ to the data, suggesting (for example) that a search on ‘drug use’ could be narrowed down to ‘heroin use’ or broadened out to ‘consumption’; the ‘bottom-up’ approach cannot offer this. Secondly, with this approach the vocabularies of individual sources both limit and define the enquiries which can be conducted: a search for ‘drugs’ will not retrieve information classified under ‘controlled substances’, and vice versa. This problem can to some extent be alleviated by the definition of synonyms (e.g. ‘alcohol’ with ‘drink’ and ‘marijuana’ and ‘cannabis’); however, this is likely to create the opposite problem, obscuring the distinctions used within individual sources and returning too much of the wrong data. Lastly, this approach gives no indication of the associations which exist between concepts within data sources, other than by reproducing the structure of the original documents.

One successful ‘bottom-up’ approach is the framework for documenting survey data developed by the Data Documentation Initiative (DDI). The DDI standard makes it possible to search on keywords associated with surveys, sections of surveys and individual questions; the short text of individual questions is also searchable. Searches of DDI metadata can also be run from the Madiera portal: a search on ‘marijuana’, for instance, brings back short text items including the following:

- CONSUMED HASHISH,MARIJUANA
 - Health Behaviour in School-Aged Children (Switzerland, 1990)
- Smoking cannabis should be legal? Q2.31
 - Scottish Social Attitudes Survey (Scotland, 2001)
- Q92C DRUGS EV B OFFERED - MARIJUANA
 - Eurobarometer 37.0 (EU-wide, 1992)

This way in to the data makes it easy for a well-prepared researcher to track the use of particular *in vivo* concepts. However, this gain comes at the cost of some information. There is wide variation both in the terminology used in the surveys and in the concepts to which they refer. It is not obvious that the ‘bottom-up’ search will bring back everything relevant to the researcher, or that everything it does bring back will be relevant. This approach clearly limits the extent to which conceptual vocabularies can be documented, and consequently sets limits to the progress which can be made towards grid-enablement of social science datasets.

Mention should be made here of a variant on the ‘bottom-up’ approach, the ‘social bookmarking’ or ‘tagging’ approach. According to advocates of the ‘tagging’ approach, top-down taxonomies like the Dewey Decimal System - or ELSST - are an artificial imposition on the world of knowledge, which is better represented as a set of individual acts of labelling (Shirky 2005). Rather than implement a single authoritative set of labels, it is argued that the labelling task should be thrown open to the user community; this is referred to variously as ‘ethn classification’ and ‘folksonomy’. Given enough active taggers, tagging practices for an individual source will converge over time on a limited set of tags, and hence a limited conceptual vocabulary. Tagging can even produce structure: associations between tags can be established by virtue of the number of sources which are tagged with a particular pair of tags. The combination of a data source, a unique tag and an identifiable user also makes it possible to ‘triangulate’ (Vander Wal 2005), finding other sources which a trusted user has labelled with a

particular tag or finding other users who share one's own evaluation of a source. Advocates of tagging suggest that it will ultimately resolve all the drawbacks of existing 'bottom-up' methods, making it possible to replace the 'trees' of hierarchical taxonomies with a pile of 'leaves' (Weinberger 2006).

In the context of e-social science, the 'tagging' approach is suggestive. A 'tag cloud' approach (Speroni 2005) could be used to display keywords used in searches, creating a visualisation of the content of a data source as seen by users. However, for ethnoclassification to offer anything like an alternative to top-down taxonomies would require several users to carry out a large number of tagging operations on any given source. Given the relatively small numbers of researchers who are ever likely to be interested in any given social science dataset, tagging is best seen as an adjunct to structured taxonomies rather than a replacement.

Description: logical

Neither the 'top-down' nor the 'bottom-up' approach articulates the conceptual assumptions which underlie the construction of a dataset - assumptions expressed both in the definition of *in vivo* concepts and in relationships between them. For example, in one survey smoking cannabis might be a type of petty crime; in others it might figure as a type of leisure activity or a potential health risk. Rather than choosing one 'correct' answer (the ELSST approach) or leaving these associations undocumented (the DDI approach), we propose to offer a coherent hierarchy of *in vivo* concepts for each individual source, based on the definitions (explicit and implicit) used in each source itself. Comparing the *in vivo* conceptual hierarchies used in multiple datasets will enable researchers both to see where concepts are directly comparable and to see where - and how - their definitions diverge and overlap.

To document hierarchies of *in vivo* concepts, we shall use description logic (DL) and the Semantic Web language OWL-DL (Web Ontology Language - Description Logic), with the ontology tool Protege and the automated reasoner FaCT++. OWL-DL is a subset of the full version of Web ontology language OWL; it has been designed to support ontology building using description logic and automated reasoning procedures, and is "the maximal subset of OWL Full against which current research can assure that a decidable reasoning procedure can exist for an OWL reasoner." (Bechhofer et al 2004).

OWL-DL makes it possible to formulate a precise logical specification of concepts such as

- use of cannabis in the form of hash oil in the month prior to the survey
- use of either crystal meth or crack by the respondent, at any time
- seizures of Class A drugs by HM Customs in the financial year 2004/5

OWL-DL's logical structures also make it possible to navigate the conceptual vocabulary of sources like those indicated above, in three directions: downward (from 'Class A drug' to specific drugs seized in the period and having this classification); upward (from 'hash oil' to 'cannabis'); or across (from the *in vivo* concept of 'crack' in a survey dataset to the organising concept of 'cocaine' in a central ontology).

Our engagement with OWL-DL is solely as users: although one of us has some background in commercial IT, we are not computer scientists and have no qualifications in pure mathematics. Consequently, we will not attempt a general description of OWL-DL, but will highlight three features which make it particularly well suited to our needs.

Firstly, as well as defining a logical hierarchy of concepts (implemented as classes and sub-classes), OWL-DL supports logical properties, which effectively relate classes to one another in non-hierarchical ways. For example, given two classes called 'Drug' and 'DrugUse' and a property called 'isUseOf', we could state - in machine-readable form - that

DrugUse - isUseOf - Drug

Subclasses inherit the restrictions placed on a class by its properties: in this example, any subclass of 'DrugUse' must be related to 'Drug' - or to a subclass of 'Drug' - by way of the property 'isUseOf'. Thus:

CannabisUse - isUseOf - Cannabis

...and so on. As banal as these examples seem, this is a major step forward from either the 'top-down' or the 'bottom-up' approaches described above: neither of these approaches supports this type of relationship in a logical (and hence machine-readable) form.

Secondly, OWL-DL supports inferred superclasses, and hence multiple superclasses for a given class. What this means is that a class can be defined in terms of its properties, such that any other class which has these properties will be considered a subclass of this class. In practical terms, this means that we can define 'Cannabis' as a 'PsychoactiveSubstance' and also define it as having a variety of properties - illegality, Class C classification, failure to induce physical dependency, etc. If we then define additional classes in terms of these properties - IllegalDrug, ClassCDrug, NonDependencyInducingDrug - the reasoner will infer that 'Cannabis' is a sub-class of each of these, as well as 'PsychoactiveSubstance', making it possible to reach the 'Cannabis' concept by way of any of these superclasses. Again, this is a major advance on what is offered by the ELSST and DDI approaches.

Thirdly, OWL-DL makes it possible to work with multiple ontologies alongside one another. This means that it is possible to produce both 'clean' ontologies, composed of precisely-defined organising concepts, alongside *in vivo* ontologies, faithfully representing the local assumptions and imprecise definitions of each given source. Crucially, it is also possible to implement correspondences between the two in logical and machine-readable form, by way of defined properties. We could, for instance, say that

[in vivo]:DopeSmoking - isATypeOf - [organising]:CannabisUse

or that

[in vivo]:Downers - hasType - [organising]:Benzodiazepine

This approach, which we are currently piloting against a set of drugs-related datasets, combines the source-based immediacy of the 'bottom-up' approach with the logical structure and clarity of the 'top-down' approach.

Conclusion

One of the defining characteristics of social science data is debate: social science concepts arise from debate and are indefinitely open to modification through debate. The inevitable result is a set of vocabularies characterised by imprecision, contestedness and mutability. In the context of the project of grid-enabling social science datasets, we believe that the challenge of these shifting vocabularies has been largely overlooked, with an implicit assumption that knowledge-management approaches which have been developed and tested in the physical sciences could be transferred to the social sciences. We believe that the result of applying such an approach to social science concepts would be imperfect at best, positively distorted at worst. As a more viable alternative, we suggest that the use of description logic at the level of individual datasets, together with an abstracted ontology of 'organising concepts', can preserve the conceptual universe of individual sources and present it in a comprehensible - and machine-readable - form.

We believe that a portal designed on these lines has the potential to make a major contribution to the grid-enablement of social science datasets, highlighting those datasets whose conceptual vocabularies cannot be equated as well as facilitating comparisons where vocabularies are directly comparable.

References

- Bechhofer, S. et al (2004), "OWL Web Ontology Language Reference", online at <<http://www.w3.org/TR/owl-ref/>>
- Multilingual Access to Data Infrastructures of the European Research Area (MADIERA) (2006), *Madiera Portal*, online at <<http://www.madiera.net>>
- Shirky, C. (2005), "Ontology is Overrated: Categories, Links, and Tags", online at <http://www.shirky.com/writings/ontology_ouerrated.html>
- Speroni, P. (2005) "On Tag Clouds, Metric, Tag Sets and Power Laws", online at <<http://blog.pietrosperoni.it/2005/05/25/tag-clouds-metric/>>
- Vander Wal, T. (2005), "Folksonomy Definition and Wikipedia", online at <<http://www.vanderwal.net/random/entrysel.php?blog=1750>>
- Weinberger, D. (2006), "Taxonomies and Tags: From Trees to Piles of Leaves", online at <http://www.hyperorg.com/blogger/misc/taxonomies_and_tags.html>