In: Proceedings of the 6th International Conference on Asia-Pacific Library and Information Education and Practice, Manila, Philippines, October 28-30, 2015. Quezon City: University of the Philippines, School of Library and information Studies.

MODELLING QUESTIONNAIRE SURVEY DATA TO SUPPORT DATA CURATION

GUANGYUAN SUN
CHRISTOPHER S.G. KHOO

Wee Kim Wee School of Communication and Information
Nanyang Technological University, Singapore
E-mail: gsun003@e.ntu.edu.sg,chriskhoo@pmail.ntu.edu.sg

Abstract

Data curation is an important activity in e-Science and e-Social Science that seeks to derive new insights by integrating existing data from multiple sources and applyingdata-intensive computational modelling and analysis to the data.Data curation initiatives have so far focused on scientific communities, especially in the bioscience area. Data curation of social science research data is in a nascent stage, and it is not clear what information in the research needs to be captured, how the data should be represented, and the issues in data integration and reuse. The paper reports an ongoing study of the information requirements for metadata and knowledge representation of social science quantitative datasets for data curation—to support data integration and reuse. Five types of metadata are proposed: contextual information, information about the sample as a whole, the structure of the dataset, provenance of the dataset (including relations between datasets), and miscellaneous information needed to use the dataset correctly. The paper reports an analysis of a sample of 14 social science survey questionnaires to identify the types of information related to the structure and semantics of datasets that need to be represented in the metadata. Beginning with metadata information that is needed for statistical analysis, the paper examined the semantic and hierarchical relations between attributes as well as between attribute values in datasets. Some issues involved in representing a few common demographic attributes are discussed, as well as the types of relations between datasets that need to be described in the metadata.

Keywords: data curation; reuse and integration; social science; quantitative data; survey questionnaires; metadata; knowledge representation

Introduction

Data curation is an important activity in e-Science and e-Social Science communities. With advancements in information technology, scientific research has evolved from a paradigm focused on empirical observations to e-Science that derives knowledge and insights from data-intensive computational modelling and simulation (National Science Board, 2005). "Curation enhances the long-term value of existing data by making it available for further high quality research ("What is digital curation?", 2014). Reuse and integration of research data from multiple sources and domains will allow researchers to use past measurements to generate new knowledge in a timely and efficient manner (Zimmerman, 2008).

In the past, data curation initiatives have focused on scientific communities, especially in the bioscience area. Data curation in social science is in a nascent stage, and it is still not clear what kinds of data need to be curated, how they should be represented and stored, and how the curated data can be integrated and reused. Several data curation process models have been proposed (e.g., Crowston & Qin, 2011; Higgins, 2008; UK Data Archive, 2015), specifying the stages and steps in the curation of data of a research project, and the guidelines for each step. The DCC Curation Lifecycle Model (Higgins, 2008) specifies four levels of full-lifecycle actions (*Description and Representation Information, Preservation Planning, Community Watch and Participation,* and *Curate and Preserve*), followed by eight sequential actions (*Conceptualise*,

Create and Receive, Appraise and Select, Ingest, Preservation Action, Store, Access Use and Reuse, and *Transform*)and three occasional actions (*Dispose, Reappraise,* and *Migrate*).This study focuses on the full-lifecycle action of*Description and RepresentationInformation*and the sequential action of *Use and Reuse*.

Different kinds of data present different representation challenges, because of the nature and characteristics of the data as well as how the data is used. This study will focus on quantitative social science data which is defined as data that is related to people, organization and society which can be stored in a tabular format (e.g., Microsoft Excel file, CSV file, SPSS file, etc.). As social scientists also make use of statistical data collected and published by government and international agencies (e.g., census data), we extend the scope of the study to such statistical data that is related to people and people groups.

This paper reports a preliminary study of the metadata and knowledge representation requirements in curating social science quantitative research datasets, especially datasets from questionnaire surveys. In particular, we seek to identify the types of information relating to the datasets that need to be described, modelled, represented and stored—to support future data integration and data reuse.

Questionnaire survey data and government statistical data are often stored in a tabular form for easy statistical analysis. The rows in the table usually represent individual cases (representing the unit of analysis), but can also represent people groups, organizations and places. The columns usually represent attributes (fields or variables) and attribute values related to each case. The datasets are expected to provide the following types of information: demographic, social, economic, health, government, geographic, trade, business, etc. Even though the table is flat, sometimes there is implied hierarchical structure (e.g., a set of fields may represent a complex attribute).

Good metadata and knowledge representation is needed to support data integration and reuse:

- Other social scientists who were not involved in collecting the dataset need to understand and interpret the data quickly and so that they can reanalyze the data from a different perspective or using a different conceptual framework. A consistent data and metadata representation language with clear semantics (i.e. vocabulary control and meaning assigned using an ontology) is required to support understanding and interpretation.
- Social scientists need to understand the data to integrate multiple datasets (or to integrate a curated dataset with their own dataset) and analyze patterns across multiple domains or contexts to synthesize new insights not obtainable from the individual datasets. Thus, datasets need to be represented with clear semantics to enable users to identify common attributes in two datasets that can be used for linking the records.
- A good knowledge representation scheme with clear semantics can also support automatic integration of data sets to enable linked data applications and automatic big data analysis.

The main challenge of data representation and storage arises from the fact that data are collected from different sources in different contexts, represented with different data structures and stored in different data repositories in different file formats. The challenges to reuse and integration of multiple sources of data are:

- Data from multiple sources are sometimes inconsistent with each other. Different people use different variable names for the same concept, and use different symbols to represent the same value. For example, *gender* and *sex* refer to the same concept, and the values can be coded as “male/female”, “0/1”, “1/0” or “M/F”. So we need to develop an authority list of concepts and corresponding variables names, as well as the semantics for the possible values.
- The scale used for variable values can be different (e.g., temperature in Fahrenheit versus Celsius, and 5-point versus 7-point Likert scale).

- The insufficiency of context information which is crucial for users to determine the relevancy of the data.

Literature Review

According to Ball (2010), data curation is "the process of selecting, normalizing, annotating and integrating data from journals, reports or third-party databases into a database on a given topic, in order to keep it up-to-date and relevant" (p.5). Data curation is also defined as "the active and ongoing management of data through its lifecycle of interest and usefulness to scholarship, science, and education, which includes appraisal and selection, representation and organization of these data for access and use over time" (Shreeves & Gragin, 2009, p. 5). Lord and Macdonald (2003) further pointed out that "For dynamic datasets this [data curation] may mean continuous enrichment or updating to keep it fit for purpose" (p.18).

The concept of data curation is widely embraced across the scientific communities, and the bioscience community has developed the most mature practice (Ball, 2010). The main technique used to represent knowledge in the bioscience data curationcollaborations is to build ontologies (i.e. controlled and structured vocabularies) to describe and link the biological data (Bult et al., 2008). The knowledge representation system is developed in 3 stages (Orchard et al., 2012).

- The first stage is to describe data. Individual researchers maintain research data separately. When submitting data to a data repository, a common file format for representing data is required. That is to say, a stipulated list of information is required to be supplied to describe data. This enables user to download, combine, visualize and analyse data in a single format from multiple sources.
- The second stage is to coordinate curation. Synchronization of curation strategies is addressed in order to avoid redundant work on the same data. The curation strategies refer to the rules and controlled vocabularies that are used to curate biological data. They need to be standardized and synchronized across different data repositories within according consortiums.
- The third stage is quality control. Based on the ontology, released XML files are checked to ensure that their use of controlled vocabularies and assigned relations are syntactically and semantically correct.

The three stage structure can be used for reference when building knowledge representation system in social science fields. However, it is conceivable that it would be more difficult to describe the data and develop controlled vocabularies and rules since social science research data is more arbitrary and less structured.

Research studies on reusing curated social science data is in a nascent stage.Many studies have discussedconcerns about sharing data by social scientists(e.g.,Tenopir et al., 2011; Zenk-Möltgen & Lepthien, 2014). Others focused on the data selection and preservation activity. For example, Gutmann, Schürer, Donakowski and Beedham (2004) reviewed the selection, appraisal and retention of social science data in two archives: the Inter-university Consortium for Political and Social Research (ICPSR) from the US and the UK Data Archive (UKDA). They found that the primary appraisal guidelines of these two archives were the degree of significance of the research, the uniqueness of the data, and the degree of usability of data. Dehnhard, Weichselgartner and Krampen (2013) studied German psychological researchers' practice to deposit quantitative data in the data archive PsychData, developed by Leibniz Institute for Psychology Information. They found that the minimum selection criterion in PsychData was the existence of peer-reviewed publications based on the data. Beyond this criterion, many other criteria have been adopted dependent on specific cases, but all fulfilling the underlying principle of "PsychData should mainly preserve psychological data sets of unique value for the psychological research community" (p. 174).

Analysis Approach

Quantitative datasets are typically represented and stored in a tabular form with rows and columns of data:

- Each row represents a *case* or *record* of a unit from the population of interest. The *unit of analysis* is often individual persons in social science research, but can be groups of people defined in some way. More generally, statistical datasets from public and private organizations can describe units that are physical or abstract objects that are instances of any class of resources. The population of interest is defined by the researcher for the purpose of the researcher's study, and can be all the instances of the class (e.g., every living human in the world) or a subset of the class instances (e.g. every resident of Singapore). For statistical datasets from organizations, the population is defined by the organization that collected the data, or the process that collected the data.
- Each column representsa *variable* or *attribute*of the units of analysis (e.g. people's gender). In social science research, the attributes usually refer to different aspects of individuals or people groups.
- Each cell is a combination of a row (case) and column (attribute), and contains a datum that represents a value for a case's attribute (e.g. a person's gender)
- A tabular dataset may have one or more *key* or *identifier* attributes that have a unique value for each case. The values of an identifier (ID) attribute can be used to identify each case unambiguously, and can be used to integrate or join two tabular datasets on the ID attribute.

In this preliminary study, we analyzed 14 questionnaires collected from the following sources:

- 6 from articles published in Journal of the Association for Information Science and Technology
- 4 from the UK Data Archive Survey Question Banks (http://surveynet.ac.uk/index/search.aspx?collectionid=1099)
- 1 from a General Social Survey in the US (http://www3.norc.org/GSS+Website/Publications/GSS+Questionnaires/)
- 2 questionnaires from faculty members of our school—one for a Singapore Internet use survey, and one for a social networking sites (SNS) use survey
- 1 from a PhD thesis taken from ProQuest Dissertations and Theses database

We propose that the following types of information at different levels have to be modelled and represented for data curation:

1. Contextual information about the dataset, including the research objectives, hypotheses, and research framework.
2. Information about the sample in the dataset, including sampling method, and the attributes that apply to all the individuals in the sample, especially demographic attributes.
3. The structure and semantics of individual tables (datasets) and table columns (attributes)
    a. Unit of analysis (i.e. what kind of entities/cases do the rows represent)
    b. Attributes (columns)
    c. Attribute values.
4. Provenance of the dataset (including relations to other datasets that it was derived from), and the operations performed on the source dataset, including cleaning, rescaling, enrichment and modelling.
5. Other issues that other researchers/users should be aware of.

This paper focuses on part 3—representingthe structure and semantics (meaning) of the dataset.

As a knowledge representation system cannot represent an infinite number of concepts, our approach is to identify recurrent patterns that can be used as basic building blocks, lists of concepts that have a closed membership (limited number of items), and underlying dimensions or facets (with a closed list of categories or values).In designing the metadata and knowledge representation of social science research

datasets, we make use of concepts in the areas of metadata schema, resource description framework (RDF), ontology and linked data.

Dataset Representation

General Issues

Quantitative datasets are typically stored in a spreadsheet (e.g., Microsoft Excel spreadsheet) or a database table. In a spreadsheet or database, a datatype is specified for each column—usuallyinteger, real number (floating point number), double (double-precision floating point number), character, character string (of a certain maximum length) or Boolean (true|false). We shall refer to this datatype as a *mathematical datatype*.

Statistical analysis is often applied to quantitative datasets. Each statistical analysis package as its own data file representation (e.g.,IBM SPSS .sav format). Figure 1 shows an SPSS variable view screen that defines variables in the dataset.

Most statistical data files will represent, for each attribute, the type of measure or*statistical datatype*:

1. *categorical* (or nominal), which may be subdivided into *dichotomous* (binary valued) or *polytomous* (more than 2 categories)
2. *ordinal*, which may be subdivided into *rank* (1st, 2nd, 3rd, etc.) and *ordered categories* (e.g., 5-point Likert scale)
3. *scale*, which may be subdivided into *interval* and *ratio* scale.

We shall refer to this as the *statistical datatype*. These statistical datatypes determine what statistical analysis techniques are appropriate for analyzing the attributes and how they should be prepared for analysis.

As shown in the IBM SPSS variable view screen (Figure 1), other types of information are specified for each attribute:

- Attribute code ("Name" in the Figure)
- Type (i.e. mathematical datatype)
- Label (user-friendly descriptive label)
- Valid values (for categorical variables), and a user-friendly label for each value
- Missing values (values used for indicating a missing value). 2 or more values can be used to indicate different reasons for the missing value

The other columns in Figure 1 indicate formatting and presentation preferences.

The dataset metadata information described above is well-known to social science researchers. They are part of the metadata that need to be stored together with the dataset values, to support statistical analysis and the interpretation of the statistical analysis results. This basic metadata elements for social science quantitative datasets is summarized in Table 1. Each attribute in a dataset needs to have a URI to map it to a concept in an ontology, and thus assign it meaning.

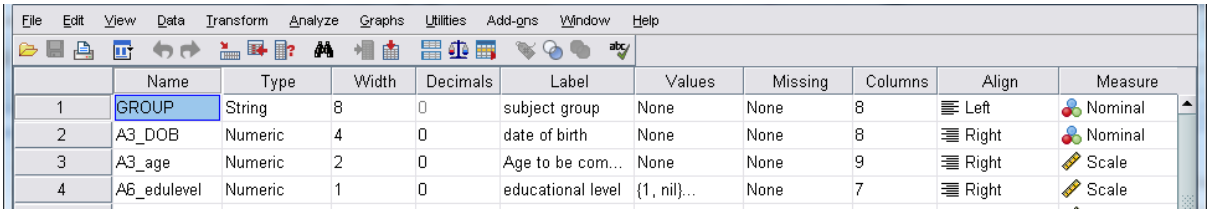

Figure 1. IBM SPSS variable view screen

Table 1

Basic Metadata Elements for a Social Science Quantitative Dataset

Metadata for dataset

element	***label***	***value semantics***	***notes***
unit	Unit of analysis	person | group | place | organization	

Metadata for each dataset attribute

element	***label***	***value semantics***	***values***
URI			URI to map the attribute to an ontology
code	attribute code		
label			
mDatatype	mathematical datatype	integer|real|double| character|string| Boolean	
sDatatype	statistical datatype	categorical|dichotomous| polytomous|ordinal| rank|ordered|scale| interval|ratio	
validValues			This is a set of validValues in the column and the corresponding concept (URI) each value represents. Depends on the mDatatype and sDatatype.
missingValues			

For each attribute, the *validValues*need to be carefully represented to support interpretation of statistical analysis results. They may also be used for linking records across datasets, thus enabling data integration. The *validValues*depend on the mathematical datatype (e.g., real or string) and the statistical datatype (e.g., categorical or scale):

- *scale* values: can be any integer or real number. The values tend to fall within a range, and minimum and maximum values can often be specified. Scale values need to be assigned meaning by specifying the unit of measure (e.g., frequency, percentage, and years). The unit of measure can be complex, for example 3 times a year (freq per duration). Scale values can also have a complex structure with a few parts (e.g., hrs:min:sec and year:month:day).
- *rank* values: are integer values that start with 0 or 1, and increase until the sample size is reached.

- *ordered* categories: are integer ranges from 0 or 1 to probably not greater than 10. The meaning of the categories, and the direction of the ordering need to be indicated. For example, 1="strongly agree", 2="agree", 3="somewhat agree", etc. The scale can also be in reverse direction: 1="strongly disagree", etc. To complicate matters, each category may represent a range of numbers or interval (e.g., age groups 18-20yrs, 21-25yrs, 26-30yrs, etc.). That is, they can represent binned values from interval data, and be represented and analyzed as a scale attribute. This needs to be represented in the attribute metadata.
- *dichotomous* (binary) values: are usually 0|1, 1|2, a|b, A|B, true|false, T|F, yes|no, or y|n, but can also be character strings that reflect the attribute (e.g., male|female or M|F).
- *polytomous* categories: can be represented as integers, but are often represented by character strings that are unpredictable. Nevertheless, the values have meaning in the domain of the study and need to be mapped to concepts in an ontology.

Although a dataset is usually stored in a "flat" table for statistical analysis, there may be relations among the attributes that need to be represented. There are at least the following kinds of relations:

- Groups of related attributes: for example a set of demographic attributes, a set of Likert-scale questions, and a set of questions pertaining to religion.
- An attribute with sub-attributes in a hierarchical structure:for example, a question may offer a set of categories of which the respondent can choose more than one category. In this case, each category has to be represented by a separate column of Boolean values—to indicate whether the category is selected by the respondent. A related case is when the respondent can choose only one category, but the variable is stored as dummy variables (with the categories occupying separate columns) to facilitate regression analysis.
- An attribute that is contingent on another attribute:for example, a question may ask the respondent skip to a specific question, if the answer to the question is "yes". For example: *Do you have a full-time job? If yes, what is your annual gross income?* In other words, the attribute is valid only for a subset of the respondents, as determined by the value for another attribute.
- An attribute derived from one or more other attributes: for example, a set of dummy variables derived from a categorical variable, and an interaction variable derived by taking the product of two attributes. The mathematical formula or operations used to derive the variable may need to be represented.

Attribute categorical values can also have a hierarchical structure, or be grouped into a taxonomy. This is common for attributes with a large number of categories. Examples are the ethnic group categories (see Table 4) and occupation categories (see Table 5).

Socio-demographic Attributes

To support interpretation and reuse by other social scientists, the attributes in a dataset and the *validValues* for each attribute need to be assigned meaning by mapping them to concepts in a knowledge representation scheme. For this project, we propose to use an ontology, specifically OWL (Web Ontology Language) Level 2, as the knowledge representation scheme. The choice is obvious as OWL is used in the Internet environment to support semantic web and linked data applications.A comprehensive ontology needs to be constructed to support data curation of social science quantitative data.

It is difficult to construct an ontology to cover all the concepts used in social science research. However, there are some sets of attributes that are commonly used. Socio-demographic attributes can be found in almost every questionnaire survey dataset, though some attributes (e.g., gender and age group) are more common than others (e.g., income group). Because socio-demographic variables are common, they are often used to aggregate records into groups of people with the same values for a particular

socio-demographic variable, e.g. zip code, income group, age group, education level, and ethnic group. Table 2 lists the more common socio-demographic variables.

Table 2

Common Socio-Demographic Variables

Gender	
Ethnic group	
Occupation	
Age	Age group
	People group by age (e.g. adults, young adults, teens, children)
Education level	
Income	
Social class	
Citizenship	
Marital status	

Gender

Gender is a dichotomous variable, with two categories of*male* and *female*.*Male*is often listed first in survey questionnaires. Though the categories are well-known and unambiguous, the attribute values used in each dataset vary: 1 or 2, A or B, a or b, Male or Female, and M or F. Table 3 lists the variations found in the questionnaires we analyzed.

Table 3

Example Gender Values

1 Male	a. Male	A. Male	Male
2 Female	b. Female	B. Female	Female

Ethnic group/Race

Ethnic group questions are normally in multiple choice formats. The value and number of categories vary from country to country. For example, compared withSingapore,surveys in the UK and the US cover a wider ethnicity range and are categorized in a more detailed way. Table 4 compares ethnic group categories from four survey questionnaires: the Singapore Internet use survey, two surveys from UK Data Archive Survey Question Banks, and one General Social Survey in the US.

Table 4

Ethnic Group/Race Categories

Singapore Internet survey	**General Household Survey 2000 (from UK Data Archive Survey Question Banks)**
1Chinese	1White
2Malay	2Black - Caribbean
3Indian	3Black - African
4Others	4Black - Other Black groups
	5Indian
	6Pakistani
	7Bangladeshi
	8Chinese
	9None of these
National Statistics Opinions Survey 2009 (from UK Data Archive Survey Question Banks)	**General Social Survey 2014 (US)**
1.00 White British	Indicate one or more races that you consider yourself to be.
2.00 Any other White background	1. White
3.00 Mixed - White and Black Caribbean	2. Black or African American
4.00 Mixed - White and Black African	3. American Indian or Alaska Native
5.00 Mixed - White and Asian	4. Asian Indian
6.00 Any other Mixed background	5. Chinese
7.00 Asian or Asian British - Indian	6. Filipino
8.00 Asian or Asian British - Pakistani	7. Japanese
9.00 Asian or Asian British - Bangladeshi	8. Korean
10.00 Asian or Asian British - Any other Asian background	9. Vietnamese
	10. Other Asian
11.00 Black or Black British - Black Caribbean	11. Native Hawaiian
	12. Guamanian or Chamorro
12.00 Black or Black British - Black African	13. Samoan
13.00 Black or Black British - Any other Black background	14. Other Pacific Islander
	15. Some other race
14.00 Chinese	NO MORE MENTIONED
15.00 Any Other	DON'T KNOW
98.00 Refusal	REFUSED

Comparing Singapore Internet use survey question with the three questions from the UK and the US, *Malay* is a unique category only for Singapore, whereas the categories of *Indian* and *Chinese* appear in all the four questions. The semantics of *Chinese* appears to be comparable. However, the meaning of *Indian* seems to vary across these three countries. This will cause confusion when linking ethnic groups across the surveys. The two surveys from UK Data Archive Survey Question Banksmake distinctions between *Indian*, *Pakistani* and *Bangladeshi*. When mapping them to Singapore Internet use survey, there is an issue of determining whether the*Indian* race in Singapore includes *Pakistani* and *Bangladeshi*. In the General Social Survey in the US, the Indian race is represented by two categories of *American Indian or Alaska Native* and *Asian Indian*. Presumably, *Asian Indian*is the same as the*Indian* category in the Singapore Internet use survey.

The two surveys from UK Data Archive Survey Question Banks divide ethnic groups into four big categories of*white*, *black*, *Asian*, and *mixed*. Each category is further subdivided into a hierarchical structure. The National Statistics Opinions Survey in 2009 (the lower left cell of Table 4)distinguishes *White British* from *Any other White background.* This is different from the General Household Survey in 2000 (the upper right cell of Table 4), which doesn't make this distinction. Thus, if the two questions are linked, the*White*category in theGeneral Household Survey in 2000 should include *White British* plus *Any other White background* in theNational Statistics Opinions Survey in 2009.

The National Statistics Opinions Survey in 2009 also forms a hierarchy by combining the ethnic group attribute with the nationality. The suffix of *British*is added to ethnic groups of *white*, *black* and *Asian*,to indicate British citizens of different ethnic categories. However, the nationality tag is missing in the mixed race groups, for example, *Mixed - White and Asian*. There is potential confusion or overlap between some categories, for example *Chinese* versus*Asian or Asian British - Any other Asian background.*

The General Social Survey in the US shows a completely different categorization of ethnic groups.The category structure does not follow a clear pattern. The US questionnaire is less interestedin distinctions within the white and within the black ethnic groups, i.e. no distinctions are made between British Americans, German Americans and Italian Americans, and between blacks of different origins.The UK questionnairedistinguishes between*Black Caribbean* and *Black African*, whereas the US questionnairelump them under *Black or African American*. However, the US questionnairehas categories of *Native Hawaiian, Guamanian or Chamorro, and Samoan*.

Occupation

Questions about occupation come in three styles:
1. Free text answer
2. A closed set of choices, which are quite comprehensive (see column A of Table 5)
3. Face-to-face semi-structured interview to probe for details.

An example of the semi-structured interview is found in the Health and Medicine Survey in 2009 that retrieved from UK Data Archive Survey Question Banks, where respondents are asked:

- What was your [LAST] (main) job [IN THE WEEK ENDING DATE]?
- What did you mainly do in your job?

Interviewers are then instructed to "probe manufacturing or processing or distributing [information] and main goods produced, materials used, wholesale or retail [information]" in order to form a "precise and detailed description of job and industry [that] avoid one word responses".

Table 5 gives three examples of occupation categories. Column A is a close set choices from a question in an article published in Journal of the Association for Information Science and Technology (Zhitomirsky-Geffet & Bratspiess, 2014). Column B is a list of 40 occupation categoriescoded by researchers from face-to-face semi-structured interviews in the National Statistics Opinions Survey 2009. Column C is an example of derived variables from column B by combiningthe 40 categoriesinto 8 categories for convenience in statistical analysis.

Table 5

Examples of Occupation Categories

A. Zhitomirsky-Geffet & Bratspiess (2014) paper published in Journal of the Association for Information Science and Technology	**B. National Statistics Opinions Survey 2009 (from UK Data Archive Survey Question Banks)**	**C. National Statistics Opinions Survey 2009 (from UK Data Archive Survey Question Banks)**
a) Heavy industry	1.0 Employers in large organisations	1.10 Large employers and higher managerial occupations
b) Hi-tech	2.0 Higher managerial occupations	1.20 Higher professional occupations
c) Law	3.1 Higher professional (traditional) - employees	2.00 Lower managerial and professional occupations
d) Medicine	3.2 Higher professional (new) - employees	3.00 Intermediate occupations
e) Accountancy	3.3 Higher professional (traditional) - self-employed	4.00 Small employers and own account workers
f) Economics	3.4 Higher professional (new) - self-employed	5.00 Lower supervisory & technical occupations
g) Teaching	4.1 Lower prof & higher tech (traditional) - employees	6.00 Semi-routine Occupations
h) Skilled craftsman/woman	4.2 Lower prof & higher tech (new) - employees	7.00 Routine occupations
j) Insurance	4.3 Lower prof & higher tech (traditional) - self-employed	8.00 Not classified
k) Unskilled labor	4.4 Lower prof & higher tech (new) - self-employed	
l) Customer service	5.0 Lower managerial occupations	
m) Clerical work	6.0 Higher supervisory occupations	
n) Marketing Management Sales	7.1 Intermediate - clerical and administrative	
o) Engineering	7.2 Intermediate - sales and service	
p) Biotechnology	7.3 Intermediate - technical and auxiliary	
q) Physics	7.4 Intermediate - engineering	
r) Chemistry	8.1 Employers in small organisations (non-professional)	
s) Nursing and paramedical professions	8.2 Employers in small organisations (agriculture)	
t) Pharmacy	9.1 Own account workers (non-professional)	
u) Social work	9.2 Own account workers (agriculture)	
v) Communications	10.0 Lower supervisory occupations	
x) Political science	11.1 Lower technical craft	
w) Psychology	…	
y) Information science	16.0 Occupations not stated or inadequately described	
z) Other	17.0 Not classifiable for other reasons	
	9998 Refusal	

Age

We observed the following types of age values in the sample questionnaires:

1) Age groups, listed in Table 6. The age ranges are different in different questionnaires, so they can only be linked approximately.
2) Exact age in years.
3) Date or year of birth. The surveys collected from UK Data Archive Survey Questions Bank provide instructions for handling incomplete data: "for day not given, enter 15 for day; for month not given, enter 6 for month."

Table 6

Example of Age Group Categories

Social network site use survey	**Singapore Internet use survey**	**Yuan & Belkin (2010) paper published in Journal of the Association for Information Science and Technology**	**Ho, Bieber, Song & Zhang (2013) paper published in Journal of the Association for Information Science and Technology**
17 - 20|	18-24|	16-25	Under 18
21 - 25|	25-34|	26-35	18–26
26 - 30|	35-44|	36-45	27–35
31 - 35|	45-54|	46-55	36–44
36 - 40|	55-64|	56-65	45–53
41 - 45|	65-74|	65+	54 and Over
46 - 50|	75+		
51 - 55|			
56 and above			

Income

Income level categories contain subtle variations: before tax (i.e. gross income) or after tax, household income or personal income, main job salary or income from all sources, and annual income or monthly income. Thus, additional information needs to be represented—to be processed when linking datasets. Table 7 shows example income categories from the sample questionnaires.

Table 7

ExampleIncome Categories

Singapore Internet use survey	**Zhitomirsky-Geffet & Bratspiess (2014) paper published in Journal of the Association for Information Science and Technology**	**National Statistics Opinions Survey 2009 (from UK Data Archive Survey Question Banks)**	**General Social Survey 2014 (US)**
$2000 or less	0–3000	Annual Gross Income	In which of these groups did your
$2001-$3000	3001–4500	2.00 £520 up to £1,039	total family income, from all sources,
$3001-$4000	4501–6000	3.00 £1,040 up to £1,559	fall last year -- 2013 -- before taxes,

Singapore Internet use survey	Zhitomirsky-Geffet & Bratspiess (2014) paper published in Journal of the Association for Information Science and Technology	National Statistics Opinions Survey 2009 (from UK Data Archive Survey Question Banks)	General Social Survey 2014 (US)
$4001-$5000 $5001-$6000 $6001-$7000 Above $7000	6001–7500 7501–9000 9001–10500 10501–12000 12001–15000 15001–20000 20001–25000 25001+	4.00 £1,560 up to £2,079 5.00 £2,080 up to £2,599 6.00 £2,600 up to £3,119 7.00 £3,120 up to £3,639 8.00 £3,640 up to £4,159 9.00 £4,160 up to £4,679 10.00 £4,680 up to £5,199 11.00 £5,200 up to £6,239 12.00 £6,240 up to £7,279 13.00 £7,280 up to £8,319 … 38.00 £52,000 or more 96.00 Not enough information provided 97.00 No source of income 98.00 Refused 99.00 Don't know	that is. Just tell me the letter. Total income includes interest or dividends, rent, Social Security, other pensions, alimony or child support, unemployment compensation, public aid (welfare), armed forces or veteran's allotment. A. UNDER $1,000 B. $1,000 to 2,999 C. $3,000 to 3,999 D. $4,000 to 4,999 … Y. $150,000 or over DON'T KNOW REFUSED

Other Attributes

Other attributes in a dataset depend on the domain, the research objectives of the study, and the theoretical framework adopted in the study. On the surface, it does not seem feasible to construct an ontology to cover all the concepts that a social science researcher might study. Domain thesauri with controlled vocabulary may be needed to conflate different variable names that researchers might use. Nevertheless, there may be a limited set of common concepts and attributes of people that are often studied in social science research. For example, we have found that many questions in social science questionnaires seek to find out the following aspects about people:

- Their perception about something
- Their opinion about some issue
- Their attitude towards some issue
- Their behavior (what they do)
- Their knowledge/understanding about something
- Their possession (what they have)

Content analysis of a bigger set of questionnaires is ongoing to identify common concepts and dimensions in questionnaire questions.

Derived and AggregatedData

An important kind of dataset reuse is in integrating two or more datasets to form a merged dataset, to find new patterns not obtainable from the individual datasets. To join two datasets, they must have a common attribute with the same meaning, which must have compatible attribute values that can be matched.

For raw data where the units of analysis are individuals, the records should be matched on individual IDs. As research datasets from different authors are unlikely to describe the same individuals, aggregated datasets where records are aggregated by a socio-demographic, geographic or institution variables are more likely to be integrated with other datasets and reused. This suggests that particular attention must be paid to the representation of socio-demographic, geographic and institution variables, together with information about the unit of analysis of the dataset and what operations were used to aggregate the data.

What constitutes raw data depends on the research area and even the particular study. Luckily for the social science domain, individual people are probably the smallest unit of analysis. Thus, we use the term *raw data*to refer to datasets where the unit of analysis is individuals.An aggregated dataset refers to data where the unit of analysis are groups of people defined in some way. The aggregated data may be derived from the raw data, by grouping individuals by the values of one or more variables (socio-demographic, geographic or institutional variable).

Relations between the aggregated datasets and the raw (source) dataset need to be modelled and represented. Indeed, relations between different versions of a dataset, and between any kind of derived dataset and the source dataset need to be modelled. Relations between datasets include the following types:

1. Different versions of essentially the same dataset, with the same number of rows and columns. The new version may be derived from the source version through error corrections and different kinds of data cleaning. The new version retains all the information of the source version.
2. A derived dataset with additional columns derived from the source attributes. The derived dataset is a superset of the source dataset, i.e. all the rows and columns of the source is retained, plus additional derived columns.
3. A derived dataset with a subset of the attributes of the earlier dataset, i.e. with a few columns dropped from the source dataset.
4. A derived dataset with a subsample of the records of the source dataset.
5. An aggregated dataset with the records grouped according to the values of one or more variables. A summary measure needs to be applied to each attribute in the dataset, for example mean for scale variables, and frequencies for the categories of categorical variables.
6. Enhanced dataset, with additional attributes added through linking with other datasets.
7. Summary dataset, with only summary measure for each attribute, for example mean, standard deviation, and maximum and minimum values for scale variables, and frequency counts of categories for categorical variables.

Conclusion

We have reported an ongoing study of the requirements for metadata and knowledge representation of social science quantitative datasets for data curation—tosupport data integration and reuse. We proposed five types of information to be described in the dataset metadata: contextual information, information about the sample as a whole, the structure of the dataset, provenance of the dataset (including relations between datasets), and miscellaneous information needed to use the dataset correctly. We proposed to construct an ontology, as a knowledge representation scheme, to control the vocabulary, assign meaning to common concepts and specify common relations between concepts used in the metadata. We carried out an analysis of a sample of 14 social science survey questionnaires to identify in more detail the types of information related to the structure and semantics of datasets that need to be represented to support statistical analysis, interpretation of the data, and data integration and reuse. Beginning with metadata information that is needed for statistical analysis, we examined the semantic and hierarchical relations between attributes as well as between attribute values that need to be represented in the metadata. We then examined issues involved in representing a few common demographic attributes and their values, and issues involved in

matching attribute values for dataset integration. We then outlinedthe types of relations between datasets that need to be modelled.

We are planning to analyze a much bigger sample of social science questionnaires as well as statistical datasets from government agencies to develop a metadata application profile as well as an ontology to specify value semantics for dataset metadata. The metadata application profile and ontology will be evaluated in the following ways:

- Comprehensiveness—byapplying it to new survey questionnaires
- Usability—by working with our university library to create metadata for social science datasets submitted to the library's data repository
- Comprehensibility—byasking social scientists to review the metadata we will create, and suggest how the data may be reused and integrated with other datasets
- Computability—byapplying linked data technology to automatically link the datasets, using the metadata and ontology we shall construct.

References

Ball, A. (2010). Review of the state of the art of the digital curation of research data (version 1.1). *ERIM Project Document* erim1rep091103ab11. Bath, UK: University of Bath.

Bult, C. J., Eppig, J. T., Kadin, J. A., Richardson, J. E., & Blake, J. A. (2008). The Mouse Genome Database (MGD): Mouse biology and model systems. *Nucleic Acids Research*, 36(suppl 1), D724-D728.

Crowston, K., & Qin, J. (2011). A capability maturity model for scientific data management: Evidence from the literature. *Proceedings of the American Society for Information Science and Technology*, *48*(1), 1-9.

Dehnhard, I., Weichselgartner, E., & Krampen, G. (2013). Researcher's willingness to submit data for data sharing: Acase study on a data archive for psychology. *Data Science Journal*, *12*, 172-180.

Gutmann, M., Schürer, K., Donakowski, D., & Beedham, H. (2004). The selection, appraisal, and retention of social science data. *Data Science Journal*, *3*, 209-221.

Higgins, S. (2008). The DCC curation lifecycle model. *International Journal of Digital Curation*, *3*(1), 134-140.

Ho, S. M., Bieber, M., Song, M., & Zhang, X. (2013). Seeking beyond with IntegraL: A user study of sense-making enabled by anchor-based virtual integration of library systems. *Journal of the American Society for Information Science and Technology*, *64*(9), 1927-1945.

Lord, P., & Macdonald, A. (2003). *e-Science Curation Report: Data curation for e-Science in the UK: An audit to establish requirements for future curation and provision*. Digital Archiving Consultancy Limited.

National Science Board. (2005). Long-lived digital data collections: Enabling research and education in the 21st century. *National Science Board*. Retrieved July 15, 2015 fromhttp://www.nsf.gov/pubs/2005/nsb0540/

Orchard, S., Kerrien, S., Abbani, S., Aranda, B., Bhate, J., Bidwell, S., ... & Hermjakob, H. (2012). Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nature methods*,*9*(4), 345-350.

Shreeves, S. L., & Cragin, M. H. (2009). Introduction: Institutional repositories: Current state and future. *Library Trends*, *57*(2), 89-97.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A., Wu, L., Read, E., & ... Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS One*,*6*(6), e21101.

UK Data Archive. (2015). *Research Data Lifecycle*. Retrieved August 8, 2015, from http://www.data-archive.ac.uk/create-manage/life-cycle

What is digital curation? (2014). Retrieved July 20, 2015, from http://www.dcc.ac.uk/digital-curation/what-digital-curation

Yuan, X., & Belkin, N. J. (2010). Investigating information retrieval support techniques for different information-seeking strategies. *Journal of the American Society for Information Science and Technology*, *61*(8), 1543-1563.

Zenk-Möltgen, W., & Lepthien, G. (2014). Data sharing in sociology journals. *Online Information Review*, *38*(6), 709-722.

Zimmerman, A. S. (2008). New knowledge from old data: The role of standards in the sharing and reuse of ecological data. *Science, Technology & Human Values*, *33*(5), 631-652.

Zhitomirsky-Geffet, M., & Bratspiess, Y. (2014). Professional information disclosure on social networks: The case of Facebook and LinkedIn in Israel.*Journal of the Association for Information Science and Technology*. doi:10.1002/asi.23393