Social science research data curation: issues of reuse

Guangyuan Sun, gsun003@e.ntu.edu.sg
Christopher S.G. Khoo, chriskhoo@pmail.ntu.edu.sg
Wee Kim Wee School of Communication & Information, Nanyang Technological University, Singapore

Libellarium, IX, 2 (2016): 59 – 80.
UDK: 005.92:004.63=111
DOI: http://dx.doi.org/10.15291/libellarium.v9i2.291
Conceptual paper

Abstract

Data curation is attracting a growing interest in the library and information science community. The main purpose of data curation is to support data reuse. This paper discusses the issues of reusing quantitative social science data from three perspectives of searching and browsing for datasets, evaluating the reusability of datasets (including evaluating topical relevance, utility and data quality), and integrating datasets, by comparing dataset searching with online database searching. The paper also discusses using knowledge representation techniques of metadata and ontology, and a graphical visualization interface to support users in browsing, assessing and integrating datasets.

KEYWORDS: data curation, data reuse, search, browse, reusability assessment, data integration, knowledge representation, graphical visualization

Introduction

Data curation is attracting a substantial interest in the e-Science and e-Social Science communities. *Data curation* refers to the management of datasets (usually research data) to make them available for use by other researchers beyond the lifespan and purpose of the project for which the data were collected. The main purpose of data curation is to support data reuse by other users. Rusbridge et al. (2005) argued that "curation … [places] emphasis on publishing data in ways that ease reuse and [promote] accountability and integration" (p. 32). The U.K. Digital Curation Centre (2016) states on its website that research data can be reused later in other research projects: "Curation enhances the long-term value of existing data by making it available for further high quality research." Reuse not only saves cost from repeated or overlapping data collection, but also offers an alternative avenue to knowledge creation. Zimmerman (2008) found that integration and reuse of research data from multiple sources allowed ecology researchers to generate new knowledge in a timely and efficient manner.

Most previous studies in data curation have focused on data sharing issues (e.g., Tenopir et al. 2011, Zenk-Möltgen and Lepthien 2014). Others have focused on data selection and preservation criteria in specific data repositories (Gutmann, Schürer, Donakowski and Beedham 2004, Dehnhard, Weichselgartner and Krampen 2013). Very few papers have discussed data reuse issues, especially for social science data.

Currently, social science research data are seldom reused. There are several barriers to data reuse. Law (2005) discussed the ethical concerns held within the research community related with data reuse, including the potential harm to the original researchers, to individual research subjects and to a vulnerable population in case of data confidentiality breach. Other researchers have discussed statistical difficulties in analyzing secondary data (e.g., Bulmer, Sturgis and Allum 2009, Hakim 1982, Hyman 1972). Some other researchers have pointed out that an insufficiency of context information is a prominent barrier to data reuse (e.g., Baker and Yarmey 2009, Birnholtz and Bietz 2003, Carlson and Anderson 2007, Markus 2001). Rusbridge et al. (2005) argued that the provision of context information would promote data reuse.

In the data curation community, the definition and scope of data reuse is not well-explored and a consensus has not been reached. The words *use* and *reuse* appear to be used interchangeably in the data curation literature. This paper takes the position that as long as data is used for purposes other than that for which they were originally collected, data are *reused*. In other communities, the term "secondary use of research data" is employed to refer to this activity. However, this expression is seldom used in the data curation literature.

Curated social science research data are expected to be reused in two ways:

Other social scientists who were not involved in collecting the dataset can analyze the data from a different perspective or using a different conceptual framework;

Social scientists can integrate multiple datasets (or 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.

We characterize the distinction between *use* and *reuse* as the significant conceptual gap that can be expected between the meaning of terms used in the original data material (including dataset, data collection instrument, and research report), and the concepts and context of the user seeking to reuse the data. Knowledge representation techniques are expected to help the user bridge this conceptual gap, as the user attempts to relate the terms and concepts in the curated data with the user's context and task.

So, an important part of managing data for reuse is to design the data representation format to support reuse by other researchers. Shreeves and Cragin (2008) emphasized that *data curation* included "representation … of these

data for access and use over time" (p. 93). Knowledge representation techniques, especially metadata schema and ontology, can be used to represent the structure of data and specify the meaning of data elements, as well as represent context information needed for interpretation and reuse by other researchers.

Few researchers have examined the knowledge representation issues related to social science data utilization. Hox and Boeije (2005) described the knowledge representation requirements for searching, retrieving and evaluating the usability of secondary quantitative datasets based on their experience as social science researchers reusing data for multilevel analysis. They pointed out that to support reuse a good description of the data is needed on data collection methods and procedure, study purpose, operationalization, entities being studied, sampling criteria, and any known biases. They proposed three "characteristic problems" of data reuse in general:

1. researchers must locate data sources;
2. researchers must be able to retrieve the relevant data; and
3. researchers must evaluate the data quality to meet their own research requirements and methodological quality criteria.

Bibb (2007) also summarized three limitations of reusing secondary population health data:

1. difficulty in locating required data;
2. incongruity of primary and secondary research objectives; and
3. assessment of data quality.

In addition to searching and browsing for data and evaluating the relevance and reusability of data and data quality, we propose that data integration issues are also important and that support for data reuse should include support for data integration—to link research data from multiple sources. Data integration is extensively used in business environments, especially in the context of business intelligence, online analytical processing and data mining to support business decision making (Dayal, Castellanos, Simitsis and Wilkinson 2009) and customer relationship management (Rygielski, Wang and Yen 2002).

This paper discusses issues of reusing social science research data in the context of data curation, and how knowledge representation can address some of the issues. The perspective taken in this paper is that a research dataset must be curated together with the data collection instrument (e.g., questionnaires) as well as a report detailing context information such as the research objectives and hypotheses of the study. Furthermore, we think the data collection instrument and reports must be actively curated, meaning the information in these materials must be carefully selected, organized, represented, and linked to relevant datasets, rather than been passively stored as PDF documents with the datasets. We discuss data reuse issues in relation to the following activities

that a researcher seeking to reuse data is expected to engage in: searching and browsing for data, evaluating the reusability of data, and integrating and analyzing the data. Social science research can be divided broadly into qualitative and quantitative research (Neuman 2005). Our discussion focuses on quantitative research data—numerical data amenable to statistical analysis.

Dataset searching framework

Users need to interact with data repository systems to search for and evaluate datasets. The architecture of data repository systems is similar to that of information retrieval systems and digital library systems. Indeed digital library systems are often repurposed into data repository systems, using software platforms such as DSpace, EPrints and Fedora which are widely used for implementing digital libraries (Rice 2009). Some university libraries have used their institutional repositories that store faculty publications to double up as data repositories (Witt 2009, Walters 2009).

A model of user online database searching (also referred to as *online searching* or *online information retrieval*) provides a useful starting point and framework for a systematical analysis of user dataset searching in data repository systems.

User interaction with information retrieval systems have been modeled from different perspectives by different researchers. We outline a model in Table 1 based on Ellis' (2005) information seeking model, Bell's (2015) and Walker and Janes' (1999) online database searching models, and Saracevic's (1997) and Pian, Khoo and Chang's (2016) models of relevance judgment. The model identifies the main steps in online database searching and user interaction with an information retrieval system. The second column of Table 1 suggests how the same steps may apply to dataset searching. The steps in the database searching model correspond to the typical interface architecture of information retrieval systems, which is generally comprised of a query interface, a result interface containing a collection of document surrogates, and an interface providing access to the document. Changes to the interface architecture will affect user interaction with the system.

In online database searching, the user searches for documents containing mainly textual information. The search target is information in the document content that is relevant or useful in satisfying the user's information need. The user's information need is very varied even in the context of academic research. The target information can be the research results reported, the methodology used, the framework used, etc. In dataset searching, the information need is more constrained, and we can assume the target of the search to be a subset of a research dataset to be reused in two ways—to analyze from a different perspective, or to integrate with another dataset to synthesize new insights.

Table 1. A model of online database searching and online dataset searching

Online database searching	Online dataset searching
Purpose: to retrieve documents that contain relevant or useful information in the context of the user's research study. Stages:	Purpose: to retrieve datasets that can be reused in the context of the user's research study. Stages:
Select an appropriate database to search (or search a directory of databases).	Select an appropriate data repository (or search a directory of data repositories).
In the query screen: select appropriate search terms, and construct a query statement related to the user's information need.	Select appropriate concepts or terms, and construct a query statement related to the user's research objectives and concepts.
In the index display screen: browse the alphabetic list of index terms, thesaurus display or hierarchical classification, and select terms or concepts.	Browse the alphabetic, hierarchical or graphical display of concepts or terms (including variable names), and select concepts/terms.
In the summary search result screen: browse the summary search result (comprising a list of metadata records and abstracts), and select records that are likely to contain relevant information.	Browse the summary search result screen: Browse the dataset summary search result (comprising a list of metadata records describing datasets), and select datasets that are likely to be reusable; or Browse the variable summary search result (comprising a list of metadata records describing variables), and select variables that are likely to be reusable.
In the detailed record screen: read the detailed record or full-text document, and decide whether the content contains relevant or useful information.	Read the detailed metadata: In the detailed dataset record screen, read the detailed metadata record of a specific dataset, and decide whether a subset of the concepts or variables (representing a subset of the dataset) is relevant or reusable; or In the detailed variable record screen, read the detailed metadata record of a specific variable, and decide whether this variable and the dataset it is taken from are relevant or reusable.
Make use of the information found in the document content in decision making or in a research study.	Reuse a subset of the dataset in some way.

Online database systems typically allow the user to carry out fielded searching in author, title and subject fields, and keyword searching in the title, abstract and full-text. A browse search in an alphabetic display of index terms or a hierarchical taxonomy or classification scheme may also be supported. For dataset searching, searching in author, title, subject and other metadata fields is expected. However, as the user is looking for a dataset, we expect the user to search for variable names, attributes and concepts relating to people, people groups and social phenomena. The variable names and concepts will be related

to research questions, theories and frameworks. The demographic characteristics of the population and sample, and the data collection method and instrument will also be of interest. Clearly, metadata fields containing such information for searching and browsing will be useful.

Table 2. Summary of ICPSR and UK Data Archive's support for dataset and variable name searching

Search Entity	Type of Searching & Browsing	ICPSR	UK Data Archive
Dataset search	Fielded search	NA	Title Subject Depositor Data type (e.g., cohort and longitudinal studies, census data) Key data (e.g., Labour Force Survey) Country Kind of data (e.g., numeric, alpha-numeric) Spatial Unit (e.g., counties, districts) Analysis Unit (e.g., individual, family and household)
	Keyword search	Title Subject Abstract	Title Subject Abstract
	Browse search	By topic By series By geography By investigator	NA
Variable search	Keyword search	Variable name Variable label Question text Answer choices	Variable name Variable label Question text Answer choices
	Browse	By series By studies	By series By survey

Notes: The UK Data Archive allows the user to carry out fielded searching in not only title, subject, and depositor fields, but also in fields denoting unique information about the dataset (e.g., data type and key data), characteristics of the data (e.g., kind of data), and demographic characteristics of the sample (e.g., country or spatial unit).; ICPSR supports a variety of browse functions: users can browse datasets by a hierarchical classification of research topics, or by an alphabetical listing of study series, data collection geography, and names of investigators.; Both ICPSR and UK Data Archive have a variable and question bank, which enable the user to search for individual variables in datasets. Users are allowed to perform keyword searching in the variable name, label, question text, and answer choices. Users can also browse a summary result screen of variables by selecting from an alphabetic list of studies collecting datasets, or a list of series that the study belongs to.

To obtain a better understanding of what dataset searching entails, we examined two of the most well-known social science data repositories, the Interuniversity Consortium for Political and Social Research (ICPSR - https://www.icpsr.umich.edu/icpsrweb/) and the UK Data Archive (http://www.data-archive.ac.uk/), and summarized their support for dataset searching in Table 2.

In the summary search result screen of an information retrieval system, the user browses a list of document surrogates, comprising selected metadata fields, to make *predictive* relevance judgments—to identify documents that are likely to contain relevant information (Pian, Khoo and Chang 2016). For these selected records, the user then reads the detailed document content (if the full text is available) and makes *evaluative* relevance judgments of whether some information in the document content is relevant.

In dataset searching, users are also expected to scan a summary list of dataset surrogates or variable surrogates to identify datasets that are likely to be relevant to the user's research, before examining the individual detailed surrogate records. Looking at the ICSPR and UK Data Archive repository systems, the metadata elements in the surrogate records are very different from online document databases:

1. *Dataset summary result screen* displays the title of the dataset or questionnaire survey, name of the data collector(s) and the institutional affiliation (see Figure 1).
2. *Detailed dataset record screen* displays the dataset series that the dataset belongs to, the principal investigators, the data collector institution, the sponsoring institution, grant numbers, subject categories, and an abstract introducing the dataset, dataset coverage and data collection methodology (see Figure 2).
3. *Variable summary result screen* displays the variable name, variable label, question text, the title of the dataset that the variable is taken from, and the year of the data collection (see Figure 3).
4. *Detailed variable record screen* displays the full question text, answer choices, number of responses to each answer, variable type, valid data values, and sampling information (e.g., sampling method, sample frame) (see Figure 4).

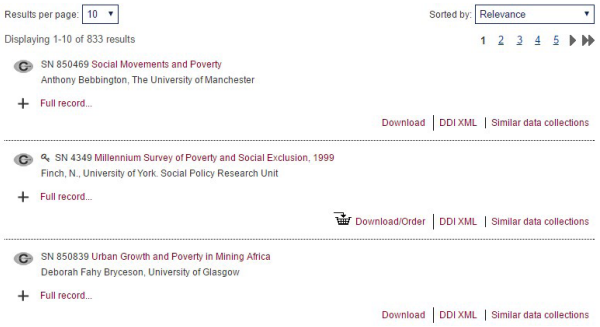

Figure 1. Dataset summary result screen in UK Data Archive

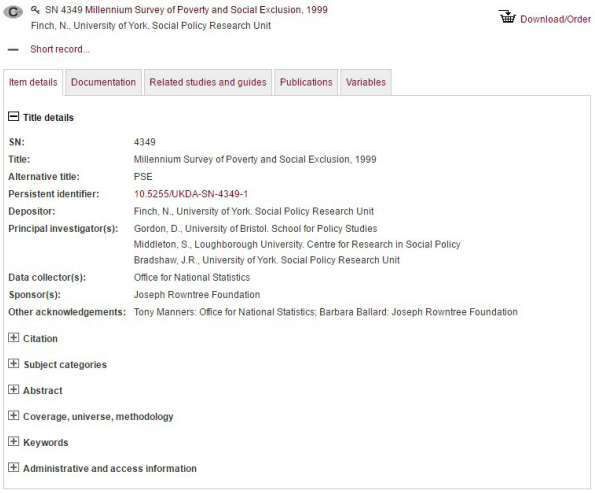

Figure 2. Detailed dataset record screen in UK Data Archive

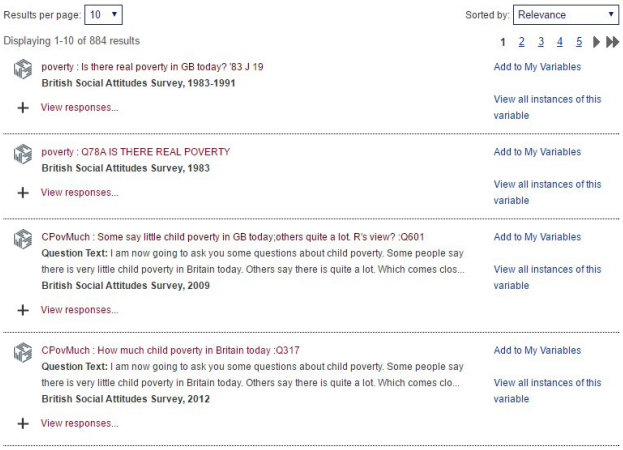

Figure 3. Variable summary result screen in UK Data Archive

VARIABLE DETAILS

Variable	CPovMuch		
Label	Some say little child poverty in GB today;others quite a lot. R's view? :Q601		
Question text	I am now going to ask you some questions about child poverty. Some people say there is very little child poverty in Britain today. Others say there is quite a lot. Which comes closest to your view?		
Responses	1	There is no child poverty in Britain today	84
	2	There is very little child poverty in Britain today	531
	3	There is some child poverty in Britain today	1458
	4	There is quite a lot of child poverty in Britain today	1262
	8	Don't know	86
	9	Refusal	0
Disclaimer	Please note that these frequencies are not weighted.		
Location	British Social Attitudes Survey, 2009		
Interviewer Instructions	CARD B8		
Universe	Adults (18 and over) living in private households in Great Britain (excluding the 'crofting counties' north of the Caledonian Canal). ;Adults;National		
Sampling	Multi-stage stratified random sample;Sampling is conducted in four stages; from 1993 the sample has been drawn from the Postcode Address File, whereas in previous years it was drawn from the Electoral Register.		
Study Type	Repeated cross-sectional study. The BSA survey is conducted annually.		

Figure 4. Detailed variable record screen in UK Data Archive

Unlike in online database searching where the user can read the content of the retrieved document to assess relevance, the user is unable to assess the relevance of a dataset by reading its numerical content! Relevance is assessed by examining the information in the detailed metadata record, but also by performing a simple statistical analysis to generate summary statistics to assess data quality.

A fundamental concept in information searching is the concept of *information need,* and the mental activity of translating an information need into search terms and a query statement appropriate to the search system. Borgman (1996) noted that "the success of a query is a function of the ability to translate the intended meaning into a set of search terms that are contained in the bibliographic records in the catalogue and that convey the intended meaning." (p. 496). In dataset searching, query formulation involves translating the user's research objectives and research concepts into dataset attributes and variable names,

as well as terms in dataset titles. Contextual constraints such as data collection method and geolocation may also need to be translated into variable names and keywords to search in the metadata fields.

Taylor (1962) characterized four levels of information needs. Initially, the need is felt but unexpressed and vague (visceral need). A conscious need may be expressed, possibly to another person. A clearer, rational and formalized statement of the need may be later formulated. Finally, a compromised statement of the need meeting the system requirements may be submitted as a query to an information retrieval system.

From our informal conversations with social scientists, researchers have different levels of familiarity with using secondary data and with using data archives:

1. Some researchers have used secondary data extensively. For the data archive that they regularly visit, they usually know the exact variable names they want. However, they do sometimes browse unfamiliar data archives to look for datasets of interest.
2. Many researchers are not familiar with using secondary data. They may have a rough idea of the variables they want, but cannot anticipate which types of variables would be available in a data archive, nor the likely variable names used.

We conjecture that both groups of researchers have information needs that are more constrained and less varied than for online database searchers in general, and can probably provide formal statements of their need. Relevance judgment will be based on whether the retrieved dataset contains variables of interest. Researchers with a conscious but vaguely defined research purpose may browse datasets by topic, author, geolocation or type of population. However, to reuse a dataset for a particular research purpose, a well-defined formal research purpose is expected. Unless a dataset contains the needed variables, is collected using a specific method, and is of acceptable data quality that can address the user's research need, the dataset is likely to be considered useless. This is unlike typical situations in online database searching where the relevance criteria are vaguer and unexpressed. After all, information expressed in text can be used in many ways, whereas quantitative data are used in statistical analyses to address specified questions.

Issues of data reuse

Searching and browsing for data

Library and information professionals and researchers have traditionally distinguished between two main types of searching (e.g., Baker and Lancaster 1991, Walker and Jane 1999, Wildemuth and O'Neill 1995):

1. Known item search: searching for a particular document for which the author or title is known;

2. Subject search: searching for any documents dealing with a particular subject or to answer a particular question.

Researchers who are experienced with secondary data reuse have a clear idea which data archive to go to when they need data, and may also be looking for a known dataset series. They are expected to perform a *known item search*. When using unfamiliar data archives, they may have a rough idea of the types of dataset that can possibly contain the desired variables, and may carry out some kind of subject search. To locate a relevant dataset, they may need to *browse* the dataset repository, and read metadata information and descriptions of each data collection program.

Users who are less familiar with secondary data reuse are expected to do more *browsing* in a data repository to gain an overview idea of what kinds of datasets are available, what kinds of metadata information is provided, and the system functions (including searching) available. They are expected to perform more iterations of interaction with the system, and are likely to take more time to browse dataset metadata, data dictionaries or codebooks giving variable definitions, and questionnaires giving the actual questions and answer choices. As Hox and Boeije (2005) put it, promising data should "appear to cover the topic, and described in a language that the researcher can read" (p. 596).

An examination of well-known social science data archives, ICPSR, UK Data Archive and the General Social Survey website, suggests a major knowledge representation issue: there are not enough of the right types of metadata information, and too much of the wrong types of information to help users to quickly identify reusable data. In current social science data archives, the metadata elements used to describe a dataset are designed from the perspective of librarians and archivists for comprehensive documentation. For users (i.e. social scientists), these metadata elements are insufficient to help them to quickly grasp key information about a dataset to make a relevance judgment. At the same time, there is an overload of archival information irrelevant to researchers' needs. Thus, the current situation may be that only determined users are likely to access these data archives: social scientists are unlikely to casually browse the data archives to explore the possibility of reusing the datasets. The UK Data Archive has, for example, at least 39 types of metadata elements in a detailed dataset record (see Figure 2). This raises the questions: 1) What kinds of metadata will support users to quickly review datasets for various types of reuse? 2) How should the metadata information be organized and displayed to the user? Studies are needed to answer these questions.

Issues of evaluating the reusability of data

Saracevic (1997) defined relevance as "the relation between the state of knowledge and cognitive information need of a user, and texts retrieved, or in the file of a system, or even in existence", and distinguished five manifestations

of relevance: system or algorithmic relevance, topical or subject relevance, cognitive relevance, situational relevance or utility, and motivational or affective relevance (p. 320). Schamber, Eisenberg, and Nilan (1990) synthesized three views of relevance after reviewing the extant literature, and presented one view of relevance as "a dynamic concept that depends on users' judgments of quality of the relationship between information and information need at a certain point in time" (p. 774).

In dataset searching, we expect users to make use of at least the following types of relevance or relevance criteria to assess data reusability:

1. Topicality or topical relevance: users need to assess whether the dataset is relevant to their research objectives, contains the desired variable(s), collected in an appropriate way, and collected from the desired population;
2. Data quality: whether the data collection method, data preparation and dataset characteristics (e.g., amount of missing data) meet the user's research quality requirements;
3. Utility: whether the dataset can be used to derive new results or insights, or otherwise meet the user's research needs.

Evaluating topical relevance

Making a relevance judgment requires the user to make use of his or her domain knowledge to interpret the dataset metadata information, and relate the information to the user's research needs. However, compared with online database searchers, we suggest that the needs of dataset searchers are more specific, and the criteria for good social science research are well-known.

In assessing topical relevance, as Bibb (2007) pointed out, one can expect some conceptual incongruity between the research objectives for which the dataset was constructed, and the intended reuse objectives of the user. Some users formulate their research questions to fit existing data: they develop their research questions based on available and accessible data. In this case, the incongruity issue is less of a problem as long as the data quality is good enough to meet the users' research quality requirement. This is often the case in research fields such as cross-national comparative studies where there may be only one data source for a particular topic, as the cost of data collection is high. In fact, if there is only a single available data source, the researcher may be prepared to compromise on data quality.

In situations where there are alternative sources of data for a topic, researchers are expected to look for the dataset that best fits their research needs, including the research objectives. The users may have more stringent requirements for the dataset, and thus the conceptual incongruity issue may be important for the user to grapple with. Hox and Boeije (2005) pointed out that small "differences

in definitions, classifications, and concepts among studies may have an impact on the reusability of the data". In practice, such differences always exist because data are collected from different sources in different contexts. Users need to understand and interpret the secondary data quickly, which includes understanding the meaning of variable names and variable values, and the relationships between the variables.

When evaluating the relevance of datasets for reuse, users need to relate the meaning of terms used in the original data material (including dataset, data collection instrument, and research report) and the concepts and context of the user's research study. Knowledge representation using ontologies can help to bridge this conceptual gap. Researchers in various fields have long been using ontologies to solve semantic incongruity issues (e.g., Wache et al. 2001, Noy 2004; Alexiev, Breu and Bruijn 2005). An ontology can explicitly describe the semantics of data, using a controlled vocabulary to unambiguously denote concepts, and semantically identify and associate similar variables. Different types of relationships between concepts can be used to specify how related variables are similar or different. Especially, they can be used to relate variables representing concepts at different levels of granularity or abstraction. This can help users to quickly grasp the underlying similarities behind apparent incongruities.

Evaluating data quality

Data quality affects the accuracy and validity of research results. After assessing topical relevance, users need to assess the methodological quality of the dataset, based on information on the data collection method (including the survey procedure and sampling design), data preparation activities performed on the raw dataset, and the algorithm used to compute any derived variable. Other types of information that can support data quality assessment include the level of representativeness of the sample and the number of respondents for the question that the variable is derived from.

To better support data reuse, it is desirable to store both the raw dataset and the cleaned up dataset after data preparation activities have been performed on it. This is so that users can clean the data in a way that is appropriate to the research questions they are investigating and meet their methodological criteria. For example, if some variables are indexes based on a few raw variables, users need to know how the indexes are computed, and be able to recalculate the indexes differently, consistent with their own research study.

Pollack (1999) discussed the quality of data in terms of its reliability and validity. Reliability assessment evaluates whether the data were collected and coded in an accurate and consistent manner that is replicable, and validity assessment evaluates whether the variables in the dataset can indeed address the research questions. Users are likely to take the authoritativeness of the data collection

agency into consideration when determining data quality, and assign datasets from trustworthy agencies a higher usability level.

Among the different types of information, it can be expected that some are more important than others in supporting users to make reuse decisions. For example, the statistical data type of a variable value constrains the type of statistical analysis researchers can use. It is desirable for the data type of a variable to be listed as one type of metadata information, so users can evaluate data reusability before downloading the dataset. Research is needed to identify what kinds of information relating to data quality are important to social scientists' assessment of reusability.

Evaluating utility

A dataset, however relevant and of high quality, will prove its worth only in actual use—in yielding useful results and providing new insights in the context of a research study. The utility of a dataset can only be assessed after the user has performed statistical analysis on it. Thus, basic statistical analysis functions should be available in data repository systems to support preliminary analysis of the data, in order to assess its likely utility.

Thus, dataset evaluation includes some number crunching to identify characteristics that indicate likely utility. In contrast, relevance judgment in online database searching involves reading and interpreting textual content. It can be expected that assessing the utility of a dataset will be more time consuming and effortful as it involves using statistical analysis tools, rather than just human mental processing.

Issues of data integration

There are two types of data integration, reflecting different research objectives. Researchers may integrate secondary data collected from multiple sources, or they may integrate their own data with secondary data. Operationally, datasets can be combined by appending one dataset to another if the variables (i.e. columns) are the same. If one dataset has *n* rows and the other dataset has *m* rows, then the combined dataset has *n*+*m* rows. Alternatively, two datasets can be merged or joined on one or more key variables—variables that the two datasets have in common. In this case, the merged dataset has more variables (columns).

To merge two or more datasets, the user needs to identify common variables in the datasets that can be used for linking the records. This is challenging because of the conceptual gap mentioned earlier, as the data were collected in different research contexts, from different subjects or entities, and using different methods. They may have different data structures and file formats. In attempting to reuse and integrate quantitative data from multiple sources, users are likely to encounter the followings issues (Sun and Khoo 2015):

1. Inconsistent attribute names: different people use different variable names for the same concept. For example, gender and sex refer to the same concept.
2. Inconsistent attribute values for categorical variables: different people use different symbols to represent the same attribute value. For example, the attribute values for gender can be coded as "male/female", "0/1", "1/0" or "M/F". For attributes with more than 2 categorical values (e.g., occupational categories and ethnic categories), the categories may be more or less detailed (i.e. small number of broad categories versus large number of narrow categories), or may have categories that do not match exactly.
3. Inconsistent scale: 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 values may also have been transformed in some way (e.g., with log transformation).

Key variables are needed to link records across datasets. A *key* is a variable (i.e. column) that exists in all the datasets to be merged, and has identical meaning and value semantics in all the datasets. Some key variables are defined by international, national or industry standards. For example, the Singapore Standard Occupation Classification (SSOC) is a national standard for classifying occupations, which is defined by the Singapore Department of Statistics (2015). It is designed for data collection in population census, household surveys and administrative data collection, to collect comparable data consistently over the years to identify meaningful trends. Users are able to integrate datasets based on the SSOC. Moreover, the SSOC adopts the basic framework of the International Standard Classification of Occupations 2008 (ISCO-08) developed by the International Labour Office, which has been the basis for many national occupation classifications. This makes possible the integration of datasets across national boundaries. Other commonly used *keys* are zip code and country names.

When two datasets do not share a key, then a pair of variables that are conceptually related in the two datasets can be prepared to form approximate keys. If one of the variables (e.g., zip code) is a specialization of the other (e.g., state or province), then the former can be generalized by aggregating its values into broader categories (e.g., converting zip code into state values). If the two variables are at the same level of granularity (e.g., street addresses), a common generalization (e.g., zip code) can perhaps be found to act as a key. Many Geographic Information Systems provide a function to merge datasets based on such operations on geolocations. To support data integration, such mapping and generalization functions need to be actively presented to alert the user to potential ways of merging datasets.

The bioscience research communities have been actively building ontologies to provide shared conceptualizations of biological domains, resolve naming

confusions, and specify semantic representation of data to support reuse and integration of biological data (Antezana et al. 2009, e.g., Bodenreider 2008, Costa, Lima, Sarraipa and Jardim-Gonçalves 2013). Likewise, in the social sciences, ontologies can be used to support social science knowledge representation, sharing and reuse. However, as social science researchers are investigating abstract social concepts rather than concepts of physical manifestations, it is expected to be more difficult to match variables and concepts arising from different frameworks, theories and schools.

Although it is difficult to construct an ontology to comprehensively cover concepts in social science, socio-demographic variables are common in social science datasets, and can be used by researchers to link records across datasets. We are constructing an ontology of socio-demographic concepts found in questionnaire surveys. Preliminary results and challenges of integrating demographic variables were discussed in Sun and Khoo (2015).

Knowledge representation to support dataset reuse

We propose five types of information to be described in the dataset metadata to support users in searching and browsing datasets, and assessing them for reuse:

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 (variables) 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 kinds of entities/cases do the rows represent)

b. Variables (columns)

c. Variable values. The variables should be linked to concepts in an ontology.

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 modeling.

We propose that the dataset metadata be displayed in a graphical visualization interface supported by knowledge representation techniques, to better support data reuse. This is in line with the perspective of data curation we have taken. We think the active curation should not only apply to social science datasets, but also include data collection instruments, and research reports related with

the datasets. This means the information in these text documents need to be examined, selected, and most importantly, linked to datasets. We also think that the relationship between datasets should be identified and represented, to support users in discovering reusable datasets. Thus, it is expected that totality of curated data will intrinsically have a network structure. Thus, we propose a graphical interface to better display the linkages between data collection instruments, research report and datasets, as well as to display the relationships between datasets. It will support users to navigate in the network, discover possible reusable datasets, and explore methods of data reuse.

It is expected that users will access and manipulate curated data through this interface. Given a particular social science concept that the user is researching, for example "social networking sites", the system can display a network of research concepts that have been studied in relation to "social networking sites" (see screen mockup in Figure 5). The graph shows specific research concepts that are found in the datasets, as well as higher-level, more generic concepts that they belong to (purple bubble in Figure 5). The research concepts in purple can be clicked on to display even broader-level concepts (blue bubble in Figure 5).

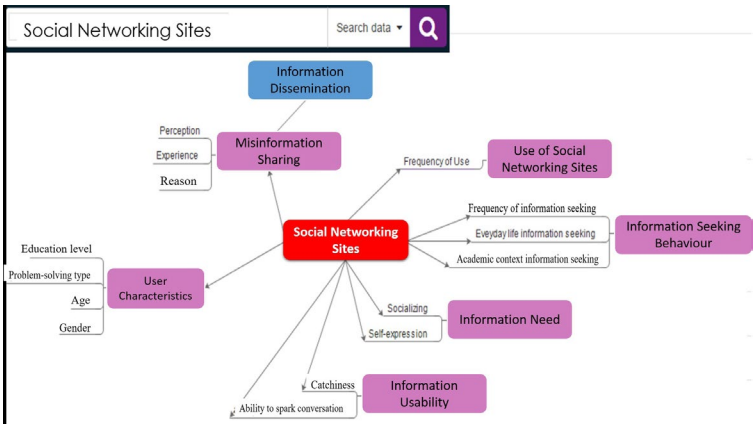

Figure 5. Mockup of an ontology graphical interface for a social science data portal system

By clicking on a specific concept, for example "socializing" (under the purple concept "Information Need" in Figure 5), and then selecting a specific record from the retrieved result list, users will be displayed with a graphical representation of the metadata for a dataset. The metadata are organized into different types of materials including questionnaire, datasets, and research reports. Figure 6 suggests how a dataset metadata may be visualized, with nodes representing the dataset (green node), the questionnaire (pink node), and the derived research reports (blue nodes). The pop-up text on the right represents the detailed metadata record describing the dataset.

Figure 6. Visualization of dataset, questionnaire, and research reports, and metadata records

Conclusion

It is generally acknowledged that the value of data curation lies in value addition to existing research data to support researchers to reuse the data and create new knowledge both now and in the future. The aim of data reuse raises many issues in practice. We have attempted to identify the main issues of reusing quantitative social science data from three perspectives of searching and browsing for datasets, evaluating the reusability of datasets (including evaluating topical relevance, utility and data quality), and integrating datasets, by comparing dataset searching with online database searching.

Researchers have different levels of familiarity with data reuse. When searching and browsing for data in a data repository, experienced researchers are expected to perform known-item searching, both of datasets and variable names. Researchers who are less familiar with data reuse are expected to carry out more browsing of metadata provided by the system to understand the curated datasets, and decide whether to reuse the data. In assessing a dataset for reusability, researchers need to overcome the conceptual incongruity between the concepts in their own research and the concepts represented by variables in the curated dataset as well as related contextual information (including the research objectives). We have proposed using knowledge representation techniques of metadata and ontology, and a graphical visualization interface to help users bridge the conceptual incongruity when browsing, assessing and integrating datasets.

The issue with current social science data archives seems to be an overload of information that does not support researchers in assessing the relevance, quality and utility of the datasets. More research is needed to identify the kinds of metadata and interface design that help researchers find relevant datasets and assess reusability. We have made some proposals based on an examination of the interfaces of existing data portals, an ongoing analysis of questionnaires and dataset variables and variable values, informal dialog with social science researchers in our university, and our own experience with social science research.

References

Alexiev, V., M. Breu, and J. Bruijn. 2005. Information integration with ontologies: Experiences from an industrial showcase. Chichester: John Wiley %26 Sons.

Antezana, E., M. Kuiper, and V. Mironov. 2009. "Biological knowledge management: The emerging role of the Semantic Web technologies." Briefings in Bioinformatics 10: 392-407. doi: 10.1093/bib/bbp024 https://doi.org/10.1093/bib/bbp024

Baker, K. S., and L. Yarmey. 2009. "Data stewardship: Environmental data curation and a Web-of-Repositories." International Journal of Digital Curation 4: 12-27. doi: 10.2218/ijdc.v4i2.90 https://doi.org/10.2218/ijdc.v4i2.90

Baker, S. L., and F. W. Lancaster. 1991. The measurement and evaluation of library services. 2nd ed. Arlington, VA: Information Resources Press.

Bell, S.S. 2015. Librarian's guide to online searching: Cultivating database skills for research and instruction. 4th ed. Santa Barbara, CA: Libraries Unlimited.

Bibb, S. C. G. 2007. "Issues associated with secondary analysis of population health data." Applied Nursing Research 20, 2: 94-99. https://doi.org/10.1016/j.apnr.2006.02.003

Birnholtz, J. P., and M. J. Bietz. 2003. "Data at work: Supporting sharing in science and engineering." Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work, 339-348. New York. doi: 10.1145/958160.958215 https://doi.org/10.1145/958160.958215

Bodenreider, O. 2008. "Biomedical ontologies in action: Role in knowledge management, data integration and decision support." Yearbook of medical informatics, 67-69. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2592252/

Borgman, C. L. 1996. "Why are online catalogs still hard to use?" Journal of the American Society for Information Science (1986-1998) 47, 7: 493-503.

Bulmer, M., P. J. Sturgis, and N. Allum. 2009. The secondary analysis of survey data. Los Angeles: SAGE. https://doi.org/10.4135/9781446263372

Carlson, S., and B. Anderson. 2007. "What are data? The many kinds of data and their implications for data reuse." Journal of Computer-Mediated Communication 12: 635-651. doi:10.1111/j.1083-6101.2007.00342.x https://doi.org/10.1111/j.1083-6101.2007.00342.x

Costa, R., C. Lima, J. Sarraipa, and R. Jardim-Gonçalves. 2013. "Facilitating knowledge sharing and reuse in building and construction domain: An ontology-based approach." Journal of Intelligent Manufacturing 27: 1-20. doi: 10.1007/s10845-013-0856-5 https://doi.org/10.1007/s10845-013-0856-5

Dayal, U., M. Castellanos, A. Simitsis, and K. Wilkinson. 2009. "Data integration flows for business intelligence." Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, 1-11. St. Petersburg, Russia. doi: 10.1145/1516360.1516362 https://doi.org/10.1145/1516360.1516362

Dehnhard, I., E. Weichselgartner, and G. Krampen. 2013. "Researcher's willingness to submit data for data sharing: A case study on a data archive for psychology." Data Science Journal 12: 172-180.doi: http://doi.org/10.2481/dsj.12-037 https://doi.org/10.2481/dsj.12-037

Digital Curation Centre. 2016. What is digital curation? (website page). http://www.dcc.ac.uk/digital-curation/what-digital-curation

Elis, D. 2005. "Ellis's model of information-seeking behavior." In Fisher et al. (Eds.), Theories of information behaviour, 138-142). Medford, NJ: Information Today.

Gutmann, M., K. Schürer, D. Donakowski, and H. Beedham. 2004. "The selection, appraisal, and retention of social science data." Data Science Journal 3: 209-221.doi: http://doi.org/10.2481/dsj.3.209 https://doi.org/10.2481/dsj.3.209

Hakim, C. 1982. Secondary analysis in social research: a guide to data sources and methods with examples. London, UK: Allen and Unwin.

Hox, J. J., and H. R. Boeije. 2005. "Data collection, primary vs. secondary." Encyclopedia of social measurement 1: 593-599. https://doi.org/10.1016/b0-12-369398-5/00041-4

Hyman, H. H. 1972. Secondary analysis of sample surveys: Principles, procedures, and potentialities. New York: Wiley.

Law, M. 2005. "Reduce, reuse, recycle: Issues in the secondary use of research data." IASSIST Quarterly 29, 1: 5-10.

Markus, M. L. 2001. "Toward a theory of knowledge reuse: Types of knowledge reuse situations and factors in reuse success." Journal of Management Information Systems, 18, 57-94. doi:10.1080/07421222.2001.11045671

Neuman, W. L. 2005. Social research methods: Quantitative and qualitative approaches. London: Pearson.

Noy, N. F. 2004. "Semantic integration: A survey of ontology-based approaches." ACM SIGMOD Record 33: 65-70. doi: 10.1145/1041410.1041421 https://doi.org/10.1145/1041410.1041421

Pian, W., C. S. Khoo, and Y. K. Chang. 2016. "The criteria people use in relevance decisions on health information: An analysis of user eye movements when browsing a health Discussion Forum." Journal of Medical Internet Research 18, 6: e136. doi: 10.2196/jmir.5513 https://doi.org/10.2196/jmir.5513

Pollack, C. D. 1998. "Methodological considerations with secondary analyses." Outcomes management for nursing practice 3: 4, 147-152.

Rice, R. 2009. DISC-UK DataShare Project: Final Report. Joint Information Systems Committee. http://ie-repository.jisc.ac.uk/336/1/DataSharefinalreport.pdf

Rusbridge, C., P. Burnhill, S. Ross, P. Buneman, D. Giaretta, L. Lyon, and M. Atkinson. 2005. "The digital curation center: A vision for digital curation." Local to Global Data Interoperability: Challenges and Technologies, 31-41. doi: 10.1109/LGDI.2005.1612461 https://doi.org/10.1109/LGDI.2005.1612461

Rygielski, C., J. C. Wang, and D. C. Yen. 2002. "Data mining techniques for customer relationship management." Technology in society 24: 483-502. doi: 10.1016/S0160-791X(02)00038-6 https://doi.org/10.1016/S0160-791X(02)00038-6

Saracevic, T. 1997. "The stratified model of information retrieval interaction: Extension and applications." Proceedings of the American Society for Information Science 34: 313-327.

Schamber, L., M. B. Eisenberg, and M. S. Nilan. 1990. "A re-examination of relevance: Toward a dynamic, situational definition." Information Processing %26 Management 26: 755–775.

Shreeves, S. L., and M. H. Cragin. 2008. "Introduction: Institutional repositories: Current state and future." Library Trends 57: 89-97. doi: 10.1353/lib.0.0037 https://doi.org/10.1353/lib.0.0037

Singapore Department of Statistics. 2015. Singapore Standard Occupational Classification 2015. https://www.singstat.gov.sg/docs/default-source/default-document-library/methodologies_and_standards/standards_and_classifications/occupational_classification/ssoc2015-report.pdf

Sun, G. Y. and C. S. G. Khoo. 2015. "Modeling questionnaire survey data to support data curation." Proceedings of the 6th International Conference on Asia-Pacific Library and Information Education and Practice (A-LIEP 2015), 196-211. Manila, Philippines, October 28.

Taylor, R. S. 1962. "The process of asking questions." American Documentation 13, 4: 391-396. https://doi.org/10.1002/asi.5090130405

Tenopir, C., S. Allard, K. Douglass, A. Aydinoglu, L. Wu, E. Read, M. Manoff, and M. Frame. 2011. Data sharing by scientists: Practices and perceptions. Plos One, 6, e21101. doi 10.1371/journal.pone.0021101 https://doi.org/10.1371/journal.pone.0021101

Wache, H., T. Voegele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hübner. 2001. "Ontology-based integration of information: A survey of existing approaches." Proceedings of IJCAI-01 workshop: Ontologies and information sharing, 108-117. Seattle, WA. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.142.4390%26rep=rep1%26type=pdf

Walker, G., and J. Janes. 1999. Online retrieval: A dialogue of theory and practice. 2nd ed. Englewood, CO: Libraries Unlimited.

Walters, T.O. 2009. "Data curation program development in U.S. universities: The Georgia Institute of Technology example." International Journal of Digital Curation 3, 4: 83-92. https://doi.org/10.2218/ijdc.v4i3.116

Wildemuth, B. M., and A. L. O'Neill. 1995. "The 'known' in known-item searches: empirical support for user-centered design." College and Research Libraries 56, 3: 265-281. https://doi.org/10.5860/crl_56_03_265

Witt, M. 2008. "Institutional repositories and research data curation in a distributed environment." Library Trends 57, 2: 191-201. https://doi.org/10.1353/lib.0.0029

Zenk-Möltgen, W., and G. Lepthien. 2014. "Data sharing in sociology journals." Online Information Review 38: 709-722. doi: http://dx.doi.org/10.1108/OIR-05-2014-0119 https://doi.org/10.1108/OIR-05-2014-0119

Zimmerman, A. S. 2008. "New knowledge from old data: The role of standards in the sharing and reuse of ecological data." Science, Technology %26 Human Values 33: 631-652. doi: 10.1177/0162243907306704 https://doi.org/10.1177/0162243907306704