Knowledge representation of social science research data for data curation and reuse

Sun, Guangyuan; Khoo, Christopher S. G.

2015

Sun, G., & Khoo, C. S. G. (2015). Knowledge Representation of Social Science Research Data for Data Curation and Reuse. Lecture Notes in Computer Science, 9469, 358-359.

<https://hdl.handle.net/10356/80318>

<https://doi.org/10.1007/978-3-319-27974-9>

© 2015 Springer International Publishing Switzerland. This is the author created version of a work that has been peer reviewed and accepted for publication by Lecture Notes in Computer Science, Springer. It incorporates referee's comments but changes resulting from the publishing process, such as copyediting, structural formatting, may not be reflected in this document. The published version is available at:
[<http://dx.doi.org/10.1007/978-3-319-27974-9>].

Downloaded on 26 Apr 2023 21:53:12 SGT

Knowledge Representation of Social Science Research Data for Data Curation and Reuse

Guangyuan Sun¹ & Christopher Khoo Soo Guan²,

Wee Kim Wee School Of Communication and Information,
Nanyang Technological University, 637718 Singapore

¹ gsun003@e.ntu.edu.sg

² chriskhoo@pmail.ntu.edu.sg

Abstract. The proposed study is in the area of data curation and knowledge representation. It focuses on the issues and methods of curating quantitative social science data, to advance the methodology of data curation, and data reuse and integration in the social sciences. Socio-economic data will be the object of the study. They are defined as data that is related to people, organization and society which can be stored in a tabular format.

Keywords: data curation, data reuse, data integration, knowledge representation, quantitative data, social science.

1 Introduction

Data curation is attracting 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 was collected. The main purpose of data curation is to support reuse by other users and integration with other datasets. The U.K. Digital Curation Centre pointed out that research datasets 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.”¹

Part of managing data for integration and reuse is to design the data representation format that supports data reuse by other researchers, and the integration of the curated data with other datasets for further analysis. Knowledge representation techniques, especially metadata schema and ontology, can be used to represent the structure of curated data and specify the meaning of data elements, as well as represent context information needed for interpretation and reuse by other researchers. The proposed study focuses on the knowledge representation issues of data curation.

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. It is still not clear what kinds of data need to be curated, how they should be

¹ <http://www.dcc.ac.uk/digital-curation/what-digital-curation>

represented and stored, and how the curated data can be integrated and reused. The proposed study focuses on the issues and methods of curating social science data, to advance the methodology of data curation and data reuse in the social sciences.

2 Research Objectives

This study will focus on the curation of quantitative social science data, which are numerical data related to people, organizations and society that can be stored in a tabular format (e.g., Microsoft Excel file, CSV file, SPSS file, etc.). The study is further limited to questionnaire survey data. However, as social scientists also make use of statistical data collected and published by government and international agencies, the scope of the study is extended to such official statistical data.

The objectives of the study are: 1) to develop a knowledge representation system for the data curation of quantitative social science research data; 2) to identify issues and solutions in the reuse and integration of curated data by social science researchers.

3 Research Method

Social science is a broad area that includes several disciplines. The study will focus on the discipline of sociology, which is arguably a core discipline in social science. However, the study will also investigate the extensibility of the knowledge representation system to two other social science disciplines—communication studies and information studies.

The knowledge representation system developed for representing sociology quantitative data has to satisfy certain requirements. It has to represent the following types of information accurately:

1. The meaning (semantics) of the data and dataset;
2. The research concepts that are operationalized in the questionnaire survey;
3. Various types of context information about the dataset;
4. Provenance of the dataset and relations with other datasets;
5. Other types of information needed to support researchers to find and evaluate the dataset for possible reuse, to integrate two or more datasets.

The first four types of information will be identified through content analysis of a sample of sociology questionnaires and the associated research reports. The last type will be identified through user (i.e. researchers in this case) studies. The requirements identified through content analysis will later be confirmed in the user study, to find out how useful these kinds of information are in supporting data reuse and integration, as well as to identify additional information needed.

Questionnaires will be collected from appendices of Master's theses and Doctoral dissertations in the ProQuest Dissertations and Theses database. Official statistical data will be collected from the online statistical services from government of public bodies. A group of 25 social science researchers in Singapore will be recruited as participants in the user study.