Enabling Precise Identification and Citability of Dynamic Data

Recommendations of the RDA Working Group on Data Citation

Andreas Rauber(✉)

Information and Software Engineering Group (IFS),
Institute of Software Technology and Interactive Systems (ISIS), Vienna
University of Technology, Vienna, Austria
rauber@ifs.tuwien.ac.at

1 Introduction

> *"Sound, reproducible scholarship rests upon a foundation of robust, accessible data. For this to be so in practice as well as theory, data must be accorded due importance in the practice of scholarship and in the enduring scholarly record. In other words, data should be considered legitimate, citable products of research. Data citation, like the citation of other evidence and sources, is good research practice and is part of the scholarly ecosystem supporting data reuse."* (Data Citation principles, [1])

While the importance of these Data Citation Principles is by now widely accepted, several challenges persist when it comes to actually providing the services needed to support precise identification and citation of data, particularly in dynamic environments. In order to repeat an earlier study, to apply data from an earlier study to a new model, we need to be able to precisely identify the very subset of data used. While verbal descriptions of how the subset was created (e.g. by providing selected attribute ranges and time intervals) are hardly precise enough and do not support automated handling, keeping redundant copies of the data in question does not scale up to the big data settings encountered in many disciplines today. Conventional approaches, such as assigning persistent identifiers to entire data sets or individual subsets or data items, are not sufficient to meet these requirements. This problem is further exacerbated if the data itself is dynamic, i.e. if new data keeps being added to a database, if errors are corrected or if data items are being deleted.

Starting from the Data Citation Principles we reviewed the challenges identified above and discussed the solutions and recommendations that have been elaborated within the context of a Working Group of the Research Data Alliance (RDA) on Data Citation: Making Dynamic Data Citeable. These approaches are based on versioned and time-stamped data sources, with persistent identifiers being assigned to the time-stamped queries/expressions that are used for creating the subset of data.

J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, pp. 663–664, 2017.
DOI: 10.1007/978-3-319-67008-9

We reviewed examples of how these can be implemented for different types of data, including SQL-style databases, comma-separated value files (CSV) and others, and took a look at operational implementations in a variety of data centers.

Reference

1. Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. In: Martone. M. (ed.) FORCE11. San Diego, CA (2014)