ILAIS 2006 Doctoral Consortium

A Model for Citing Simple Data

Yitzchak Miller
milleri@mail.biu.ac.il

Supervisor: Dr. Judit Bar-Ilan

Department of Information Science, Bar-Ilan University

Abstract

Citing data appearing in online databases in the Internet poses a challenge for authors since they cannot supply a "classical" reference as they do in the case of printed content. Supplying the URL together with the time of retrieval is not satisfactory, since the data retrieval process usually does not supply a static URL. Also, in the case of numerical databases such as statistical databases, revisions of data may result in readers' (and authors' alike) frustration when they encounter values that differ from those originally cited.

We propose a model for citing simple factual data taken from an online source that copes with the above mentioned problems by treating the single datum as an independent citable element.

The main principles of the proposed model are as follows:

1. The use of a structured definition for simple data, based on a set of identifiers for publisher, object, property, etc. which are taken from controlled lists.
2. The inclusion of "as of" time as an integral part of the datum's definition.
3. The ability to treat a single datum as an independent citable element while being able, at the same time, to relate it to a publishing framework that provides additional contextual information.
4. The inclusion of "time of publication" as a fundamental component in any transaction of adding new datum or updating an existing one by the data provider. Maintaining all previous versions of both the data values and their related metadata is a necessary condition for turning data into properly citable elements.

The proposed data citation model introduces a new approach to the notion of a reference to a data source. The reference is no longer bound to a certain textual passage located within a classic document, but rather, is defined logically as a proposition within a data model space of some domain of discourse. This approach is aligned with the more general approach of defining a document in terms of function rather than physical format as presented in (Buckland, 1998).

The reference according to the proposed model plays a dual role. It enables the retrieval of the cited datum from the source and at the same time provides a definition for it. This in turn, fits within the framework of the Semantic Web initiative (Berners-Lee et al., 2001) which strives for semantically well defined content in order to enable its automated processing.

We report on an implementation of a prototype that enables the retrieval and the direct citing of data independently of the structure of the source or the query used. It enables readers an easier

access to the original data without losing the ability to retrieve additional contextual information from the source. The prototype is based on data from two databases: the World Factbook of the CIA and WDI – World Development Indicators of the World Bank. The model was validated both by experts and by potential users of the system. The proposed model can be utilized for creating tools for checking data credibility, or for comparing data from different sources.

References

Buckland, M. (1998). What is a "Digital Document"? [Electronic version, preprint]. *Document Numérique*, *2*(2), 221-230. Retrieved October 8th, 2006 from http://www.ischool.berkeley.edu/~buckland/digdoc.html

Berners-Lee, T., Hendler, J., & Lassila, O. (2001, May). The Semantic Web. *Scientific American, 284,* 34-43.