## A Model for Citing Simple Factual Data in the Internet Environment /

## Miller Yitzchak

## **Abstract**

Within the framework of scientific writing, a clear methodology has been developed that defines the way authors should cite content taken from external sources. The special case of citing numerical data, however, poses some challenges. Usually, authors do not regard the citing of data as an act of quoting and as a result do not provide details about the exact location of the data in the source publication as it is done in the case of a quotation. To avoid cumbersome documents, authors also tend to supply partial references to data sources, including, for example, only the name of the publisher without any further details concerning the specific publication. This is especially the case when a lot of data from several sources is included in the document. This situation makes it difficult for readers to track down the data in order to verify their correctness or in order to learn further about their context in the original publication.

The growing use of online databases on the Internet as a primary data source challenges citing data from these sources, even further. The data retrieved as a response to a user's query are not part of a document with numbered pages or tables, so the author cannot provide an "exact location" for them even if he or she wishes to do so. Using the Web address of the resulted data page isn't a solution either since in many cases this address is an ad hoc address created during the current retrieval operation. Authors and readers alike face another challenge concerning the handling of data revisions. Many numerical databases only publish the last, most current data. As a consequence, readers cannot retrieve the original value cited by an author, if this datum has been modified, since.

The various style guides for scientific writing do handle citation of content coming from databases on the Internet. However, these guidelines only refer to content in a traditional form such as articles, reports and data files with no explicit reference to a citation of individual data items. Ignoring the above mentioned

methodological questions and challenges, results in a shortage of clear and unified

guidelines for citing data, mainly from online databases.

Unified format of citation of references to data in a large collection of

publications can enable an automatic analysis of data citation patterns. For example,

as in the case of papers' citation analysis, one can infer the relative importance of a

certain data source compared to other sources, or can study the change in authors'

preferences of data sources over time. The absence of such citation guidelines

prevents exploiting this potential.

This thesis studies how data citation should be accommodated in order to cope

with the described challenges. We propose a conceptual model that describes both

how data should be published and how it should be cited. The core element of the

model is a simple datum defined by its subject, its property, possibly by its time

reference, and its publication time. The proposed model treats the datum as an

independently citable element which is uniquely defined using a formal definition.

The publication time component of the datum's identification is used to differentiate

between two revisions of the same measured phenomena, published in different times.

The model also defines an abstract element called a publication framework. A

publication framework bundles data together and provides them a context beyond the

definition and interpretation of each datum separately. In a concrete implementation, a

publication framework can be a single table, a multi-table digital document or an

entire database.

In order to demonstrate the feasibility of the proposed model and part of its validation

process, we present in this thesis an outline for building a system for data publication

and citation based on the proposed model. Using this outline we have implemented a

prototype for such a system that contains two numerical databases simulating two

different data publishers. We report the result of using this prototype in which the

users acted both as readers of texts containing data citations, as well as authors that

use the system to create such data citations

System No.

1131148