Packaging and Distributing Data Collections for the Web

Tyng–Ruey Chuang[1,2,⋆]

[1] Institute of Information Science, and
[2] Research Center for Information Technology Innovation
Academia Sinica
Nangang 115, Taipei, Taiwan

What make data open and free to all to reuse? We take the position that for data to be free on the web, it shall be packaged and distributed like free software. To make data collections (*i.e.*, datasets) easy to use and useful to many, we need to consider issues related to their usage both *on and off* the web. These issues necessarily involve software support and tool development. Tools made for and lessons learned from free software development, hence, shall apply when data collections are to made free.

In addition, many a data collection is content collection whose elements can be datasets, documents, programs, and other kinds of works. The documents may describe the datasets, for example, and the programs validate and visualize them. A collection may contain as well creative works based on and derived from datasets and other elements (from the current collection or other collections). The constituting elements of a data collection can be in textual forms. To release, revise, and redistribute data collections is to work on them as they are software packages.

We take the view that for a data collection to be open, they shall be freely downloaded, adapted, mixed with others, and rehosted for other services. Being available and accessible on the Web by itself is not sufficient. A data collection must be easily ported to other computer systems, either on or off the Web, for it to be called open. Starting from these considerations, we list below some of the main issues and offer our viewpoints.

⋆ The views expressed here are those of the author, and are not necessarily those of Academia Sinica. Academia Sinica is a member of W3C.

Identifiers and references for/to elements in a data collection shall be relative and local. Use absolute URLs (*i.e.*, "permanent links") as IDs does not help data portability.

Packaging and depositing tools and practices for source code management can be used to help release data collections. It shall not matter where a data collection comes from as long as it can be properly authenticated.

Validation and revision tools shall be used. Datasets shall be accompanied by programs that validate their structure and integrity. Source code for programs that revise, validate, package data collections shall be made available. It must be free to modify and redistribute these utility programs.

Co–publication of datasets and documents shall be considered at the same time when the data is being generated. While the datasets are machine–processable, the documents aim for human readers. The two shall cross reference each other (*e.g.*, embedding semantic data in HTML documents with RDFa, as well as providing triple endpoints to access dataset documents). Tools shall be developed and used to extract datasets from documents, as well as to generate natural language text from datasets.

Independent services built from the same data collection are encouraged. These include meta services such as data catalogs and repositories. It shall be made easy to fork data collections and run new services based on their derivatives.

Rights and norms could be barriers to wide dissemination of some data collections but they need not be so. One way to get flexibility for maximal reuse is to treat all elements in a data collection homogeneously, for example, by declaring them to be in the public domain.

We shall illustrate these issues by the use of an exemplary public domain image collection on the Web.