Data Processing Across Domains Using Shared Libraries and Practices:

Recommendations

Useful data are both everywhere and nowhere at the same time. They are everywhere because organizations and programs around the world produce data, and they are nowhere because there is no systematic means of finding data sets fit for some specific purpose. This duality adds significant incentive to build the FAIR principles into data catalogs, libraries, archives, and the dissemination platforms of data producing organizations, such as national statistical offices.

An example of the need to obtain data from multiple sources at the same time was illustrated in this report. The methods through which to accomplish this are not so well-known nor understood. Here, we lay out a set of recommendations and best practices for groups to follow who plan to use data from many sources.

We will use the FAIR guiding principles to organize the recommendations, but metadata will play a central role in this discussion.

**Findable**:

Data cannot be evaluated and used if they cannot be found, and they cannot be found if they or a reference to them is not online. Digital catalogs, libraries, and archives are all possible avenues of access. These are necessary resources.

However, finding data sets is only part of the problem. The content of a data set, either the variables or cubes (multi-dimensional data structures) for structured, machine readable data and a detailed description of unstructured or semi-structured data (e.g., data, words, or tables, in a PDF file) needs to be available.

For example, the US Bureau of Labor Statics publishes 2 measures of total employment every month as part of its employment situation. One measure is from household survey data, and the other from establishment survey data. The numbers are methodologically different, are interpreted differently, and don’t agree. Depending on which measure is required, the required information must be provided.

**Accessible**:

Data that can be found are still not useful if they cannot also be obtained. The files that contain data must be listed in a catalog or other resource, as above, but if they cannot be downloaded for processing, they are not useful.

Some data must be used within a controlled environment, such as a research data center. For the purposes of this report, we assume data must be publicly and freely available. Formats for unstructured data are such that there is no automated navigation path to the data provided. Data held within a PDF file is such an example. Typically, these data can be read by humans, but they cannot be processed by computer.

Structured data, on the other hand, have the ability to be read by the computer and processed. Data in online databases accessible through an API are typical. Other formats include Comma Separated Variables (CSV), spreadsheets (Excel), and statistical package files (Stata, SAS, SPSS, R).

All of this information can be structured using metadata standards. For statistical data, standards such as DDI (Data Documentation Initiative), GSIM (Generic Statistical Information Model), or SDMX (Statistical Data and Metadata eXchange) are useful. Each of these has its strengths and weaknesses, but for the purposes of findability and accessibility, each can provide the necessary information.

**Interoperable**:

Interoperable data are those that a user other than the creator of the file can use without help. This turns out to be hard problem to solve completely. Not only do all the metadata have to be available, but permissions to use the data and the network connections have to be in place.

There are also several kinds of interoperability. Data interoperability often means data can be transferred and read from one machine to another, however the understanding of the data is not necessarily provided. Semantic interoperability is required t transfer understandings, and for this it is necessary to precisely adhere (conform) to a metadata specification. These are usually metadata standards, and some important ones for statistics were mentioned above.

Within each subject area domain, standards are often developed to ensure data and semantic interoperability for their data. In addition, there are domain independent standards that can be applied in many circumstances. The relational database query language SQL, the Dublin Core resource description set, and HTML (Hyper-Text Markup Language) for formatting Web pages are examples of these general-purpose standards. Many standards can be employed to help achieve some aspect of interoperability in applications. The choices as to which standards to use will depend on the particular situation. DDI, GSIM, and SDMX can all be used to interoperate with data that are of a statistical nature.

**Reusable**:

Reusable data are those that users can continue to find, access, and interoperate with over time. Even though data might be less relevant as time goes on, that does not mean they are still not needed. The requirements for time series analyses show that older data remain relevant for long periods of time.

To achieve this, all the requirements needed for data to be used once have to persist over time to support repeated use. For instance, the metadata needed for finding, accessing, and interoperating with data must be maintained in a persistent database. The metadata management community often refers to a metadata database as a repository. Therefore, metadata repositories have to be maintained.

This means the metadata, which of course are data, must be FAIR as well! In particular, a persistent metadata repository has to be findable for users to make use of it. Accessibility is necessary for obtaining the metadata to support all the uses of data. Interoperability, especially semantic interoperability, is necessary for users to understand the metadata, so they can understand the underlying data.