IDCC | *Practice Paper*

Frictionless Data: Making Research Data Quality Visible

Dan Fowler
Open Knowledge International

There is significant friction in the acquisition, sharing, and reuse of research data. It is estimated that eighty percent of data analysis is invested in the cleaning and mapping of data [1]. This friction hampers researchers not well versed in data processing techniques from reusing an ever-increasing amount of research data available on the web and within scientific data repositories. Frictionless Data is an ongoing project at Open Knowledge International focused on removing the friction in working with data. We are doing this by developing a set of tools, specifications, and best practices for describing, publishing, and validating data. The heart of this project is "Data Package", a containerization format for data based on existing practices for publishing open-source software. This Practice Paper will report on current progress toward that goal.

Tabular Data Packages

Data Package is a format for storing useful metadata alongside a given dataset in a simple JSON file called "datapackage.json". Our current efforts focus on packaging the very common "tabular" type of data, for example, data naturally stored in CSV files. This is a clear area for improvement well illustrated by data guidelines recently issued by Wellcome Open Research. The guidelines mandated the following:

> Spreadsheets should be submitted in CSV or TAB format; EXCEPT if the spreadsheet contains variable labels, code labels, or defined missing values, as these should be submitted in SAV, SAS or POR format, with the variable defined in English [3].

Typically, data management plans mandate that researchers submit data in non-proprietary formats, like CSV, to ensure their long-term accessibility. SPSS, SAS, and other proprietary data analysis platforms are accepted as file formats because they provide features that, until recently, haven't been supported by a standard, non-proprietary analog. We believe a decentralized, open standard for publishing tabular research data based on an existing formats like CSV is critical to high fidelity data transport and preservation, and our experiences so far point to an unmet need for exactly this kind of approach.

Draft from 20th October 2016

Correspondence should be addressed to Dan Fowler <daniel.fowler@okfn.org>

The International Digital Curation Conference takes place on [TBC] in [TBC]. URL: http://www.dcc.ac.uk/events/international-digital-curation-conference-idcc

Our Data Package specifications further provide an extensible way to create and assign functional "schemas" for tabular data which define expected types, constraints (e.g. maximum and minimum values for columns), and relations between columns in an standard, open format. Importantly, these specifications require few or no changes to existing data. In this way, our approach enables further benefits: by describing data with type and constraint information in a machine-readable manner, datasets can be automatically validated for adherence to the simple rules defined by the researcher or repository.

Uptake

Having started work on this concept through the process of developing CKAN, OpenSpending, and other data-intensive civic technology projects, we are currently in the middle of a a project to trial this approach across various research domains. Over the last year, we have noticed a very positive reaction driven, we believe, by a relentless focus on keeping the standards as simple as possible to make it easy to develop useful tools.

In some disciplines, Data Packages can provide the foundation for a general framework for data publishing where none previously existed. While many research disciplines have existing standards for sharing data, some do not. For instance, we are working the "Open Archaeology" working group to define some basic standards for describing the type of data—a mix of tabular and geospatial data—typically generated in that discipline. In other disciplines, we have seen interest even where existing standards already exist. For instance, ecologists will typically share data using the Ecological Metadata Language [2]. However, we have identified a software project for distributing ecological datasets that is using the tabular data package format due in large part to the simplicity of implementation. We will be interviewing the team to further elaborate on their motivations for adoption.

Driving Data Quality

Through this approach, we expect broad-based improvements in data quality as well as increased re-use of data. Significant time and energy is currently lost to cleaning data by early career researchers, many of whom may be more interested in generating novel insights than the sometimes tedious mechanics of data "wrangling". By providing an enabling environment for tools to create and consume "well-packaged" data, we can empower these researchers to do more with less. In particular, these specifications allow for the integration of modular, automated data import and validation services into research data repositories. We suggest that data quality can thereby be made "visible" by enabling better quality control and providing standardized visualization options.

References

[1] Tamraparni Dasu and Theodore Johnson. *Exploratory data mining and data cleaning*. Vol. 479. John Wiley & Sons, 2003.

[2] *EML - Ecological Metadata Language*. Accessed: 20 October 2016. 2016. URL: https://knb.ecoinformatics.org/#tools/eml.

[3] *How to Publish - Data Guidelines*. Accessed: 20 October 2016. 2016. URL: https://wellcomeopenresearch.org/for-authors/data-guidelines.