Towards an open-source model for data and metadata standards

Ariel Rokem

Under construction

Please excuse our dust while we work on this report, which is currently under heavy construction.

Abstract

Recent progress in machine learning and artificial intelligence promises to advance research and understanding across a wide range of fields and activities. In tandem, an increased awareness of the importance of open data for reproducibility and scientific transparency is making inroads in fields that have not traditionally produced large publicly available datasets. Data sharing requirements from publishers and funders, as well as from other stakeholders, have also created pressure to make datasets with research and/or public interest value available through digital repositories. However, to make the best use of existing data, and facilitate the creation of useful future datasets, robust, interoperable and usable standards need to evolve and adapt over time. The open-source development model provides significant potential benefits to the process of standard creation and adaptation. In particular, development and adaptation of standards can use long-standing socio-technical processes that have been key to managing the development of software, and allow incorporating broad community input into the formulation of these standards. By adhering to open-source standards to formal descriptions (e.g., by implementing schemata for standard specification, and/or by implementing automated standard validation), processes such as automated testing and continuous integration, which have been important in the development of open-source software, can be adopted in defining data and metadata standards as well. Similarly, open-source governance provides a range of stakeholders a voice in the development of standards, potentially enabling use-cases and concerns that would not be taken into account in a top-down model of standards development. On the other hand, open-source models carry unique risks that need to be incorporated into the process.

Introduction

The Brain Imaging Data Structure BIDS (Gorgolewski et al. 2016) is an example of a data standard.

Recommendations

We make the following recommendations:

1. Training for data stewards and career paths that encourage this role.
2. Development of meta-standards or standards-of-standards. These are descriptions of cross-cutting best-practices. These can be used as a basis of the analysis or assessment of an existing standard, or as guidelines to develop new standards.

Gorgolewski, Krzysztof J, Tibor Auer, Vince D Calhoun, R Cameron Craddock, Samir Das, Eugene P Duff, Guillaume Flandin, et al. 2016. "The Brain Imaging Data Structure, a Format for Organizing and Describing Outputs of Neuroimaging Experiments." *Sci Data* 3 (June): 160044. https://www.nature.com/articles/sdata201644.