A Standards-Based Model for Metadata Exchange

Martin O'Connor, Stanford University; *Marcos Martínez-Romero, Stanford University; Attila L. Egyedi, Stanford University; Debra Willrett, Stanford University; John Graybeal, Stanford University; Mark A. Musen, Stanford University*

There has been a dramatic increase in the availability of biomedical data sets derived from scientific experiments. High-quality descriptive metadata is seen as essential to facilitate the discovery and interpretation of these data sets. The biomedical community has developed templates to describe metadata for a variety of experiment types, providing a strong foundation for the development of a large number of public metadata repositories.

Unfortunately, these templates rarely share common structure or semantics. Moreover, biomedical repositories usually require proprietary submission formats that are often loosely connected to underlying template specifications. Crucially, these formats typically lack standard mechanisms for semantically annotating the metadata in templates. These difficulties combine to ensure that most metadata submissions have weak semantic content.

A key shortcoming is the absence of an interoperable format for metadata exchange. Driven by the goals of the Center for Expanded Data Annotation and Retrieval (CEDAR) (metadatacenter.org), we have developed such a format. We created a lightweight standards-based template model that provides both a structural specification of metadata and principled interoperation with controlled terminologies and Linked Open Data.

In addition to semantically marking up templates themselves, the model supports ontology-based value constraints to ensure that metadata conforming to these templates are linked to controlled terminologies. The model also provides mechanisms to support template composition, with the aim of increasing reuse of metadata fragments across templates.

We developed an implementation of this model using the Web-centric JSON format. The associated JSON Schema (json-schema.org) and JSON-LD (json-ld.org) specifications provide standard technologies to represent the structural aspects of CEDAR's template model and the linkage to semantic technologies.

We created a Web-based ecosystem driven by the model to provide an end-to-end workflow for metadata acquisition and management. We released a public alpha version of the system in September 2016 (cedar.metadatacenter.net).