Bottom-up curation of terminology for experimental variables: the Ontology of Experimental Variables and Values (OoEVV)

Gully APC Burns[1] and Jessica Turner[2]

[1] Intelligent Systems Division, ISI / USC; [2] MIND Research Network, UNM.

Introduction

The selection, definition and use of experimental variables is possibly the closest that a bench scientist comes to working with ontological concepts directly. When working with their own data, they understand the importance of standardizing their vocabulary, of defining exactly what they are measuring and how they measured it. Here, we describe work that empowers experimental scientists to define the experimental variables that they are using in a simple, bridging ontological framework (expressed as an 'ontology design pattern', ODP) that can then make those definitions available as ontologically defined terms. We emphasize a minimal ontological commitment and tool building that uses widely-used data-entry software (Microsoft Excel) to promote understandability and ease of use. We also incorprate mechanisms for interoperability with other ontologies and terminologies such as EFO, OBI, the NINDS Common Data Elements (CDE), and efforts like dbGap (scientist-driven repositories of variable definitions). As the Knowledge Engineering Working Group of the Biomedical Informatics Research Network (BIRN), we provide terminology support for the mediation technology development in several domains (neuroimaging, NHP HIV Vaccine development, immunology, radiation oncology, *etc*).

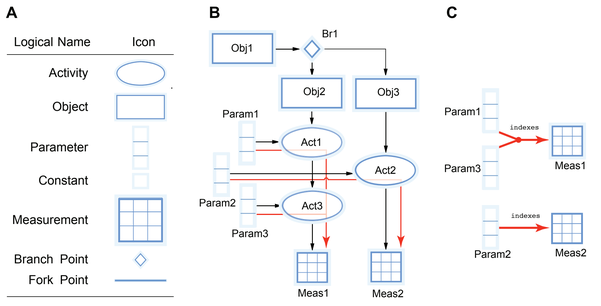

Figure 1: The underlying premise of the KEfED model

Knowledge Engineering from Experimental Design (KEfED)

KEfED is a knowledge representation of 'experimental observational assertions', based on the statistical relations between variables. KEfED elements (see Fig. 1) allow a curator to build data structures based on the dependencies between parameters, constants and measurements that can be derived from a flow diagram of an experimental protocol. Each measurement is indexed by parameters and constants by tracing a path through the protocol back to its starting point, and any parameter or constant falling on this path is used as an index for the measurement. This simple idea provides the motivating need to develop and lightweight, expressive standard terminology of elements to be used in these models.

Ontolology Design Patterns *vs.* Formal Ontologies

Formal ontologies are necessary for developing reasoning systems using formal logic. Ontology design patterns (ODPs) are a complementary approach, focusing less on the decidability and completeness of the representation, but more on providing an expressive, simple, reusable conceptual model that could ease the creation of standardized data elements. An ODP is designed to be reused in multiple contexts by groups without adherence to a centralized authority, and must be simple, non-restrictive and clear enough to promote reuse by others. We define an ODP for variables used in scientific experiments called the 'Ontology of Experimental Variables and Values' (OoEVV) to provide a practical curation process for domain experts .

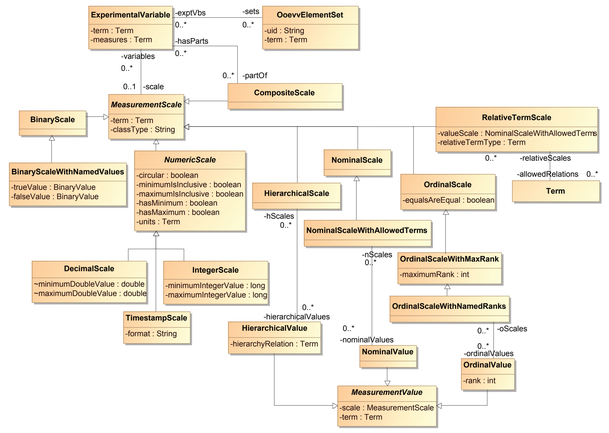

Figure 2: UML class diagram of basic structure of OoEVV showing detailed representation of the specification of measurement scales

Ontology of Experimental Variables and Values (OoEVV)

In an illustrative example, human subjects with or without schizophrenia participated in an functional Magnetic Resonance Imaging (fMRI) scan while performing a particular auditory oddball task (Ford *et al.* 2009, Schizophren. Bull. 35:58-66). The variables in this experiment include subject characteristics (diagnostic group, age, gender, performance on the task), as well as the experimental conditions of the oddball task (listening to the oddball or standard stimulus), and variations in the data collection methods (which fMRI scanner was used). Each variable is defined with its own mathematical characteristics for this study: 'Age' and the 'BOLD signal' are continuous numeric variables. 'Diagnostic category' or 'gender', have no units and cannot be added or subtracted meaningfully. OoEVV captures this usually implicit information.

The basic components of OoEVV are shown in Fig. 2 as a UML class diagram. An ***OoEVVElementSet*** instance denotes a collection containing all variables relevant to a given domain, such as fMRI. An ***ExperimentalVariable*** instance *measures* a 'quality' (a ***Term*** instance denoting a reference to the external characteristic within the world that the variable measures). In our example, 'age in years' and 'experimental condition' are two example variables so that the 'age in years' variable measures the age of the subject at the time of the experiment in years, which could be linked to the relevant term from the Phenotype, Attribute, and Trait Ontology (PATO, *PATO:0000011)*. The 'experimental condition' variable indicates whether the data were collected during the 'oddball' or 'standard tone' conditions of the auditory oddball task, and links to the Cognitive Paradigm Ontology (CogPO, *CogPOver1:COGPO_00110)*.

Each variable links to a ***MeasurementScale*** instance that delimits the types of computation that may be performed on a given variable and the range of possible values for a variable. The 'age in years' variable uses a ***IntegerScale*** (a specialization of ***NumericScale***), while the 'experimental condition' uses a ***NominalScale*** (denoting values that may only be compared to see if they are same). Other scale types also include ***OrdinalScale*** (denoting values that may only be ranked), ***BinaryScale*** (denoting variables that take only 'true' or 'false' values), ***RelativeScale*** (denoting values that take can only defined by their relation to other objects), and ***HierarchicalScale*** (with values organized in a hierarchical structure, such as organismal taxonomy). Since OoEVV is only a specification for experimental variable definitions, we use ***MeasurementValue*** instances to assist with the specification of each ***MeasurementScale*** rather than representing data (at this stage).

It is crucial to note that *this formulation allows us to define multiple variables that measure the same underlying quality with different mathematical scales*.

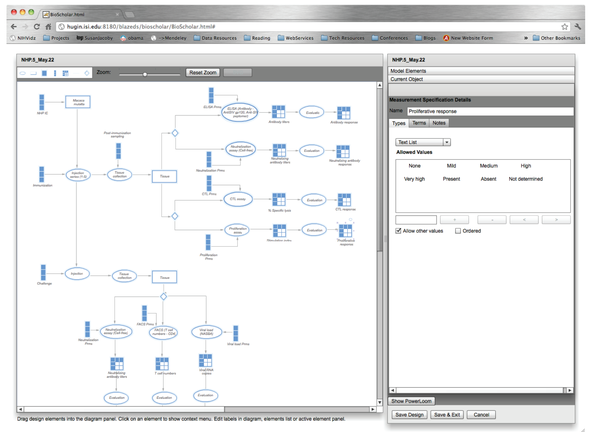

Figure 3: Screencap of the current BioScholar KEfED editor system

KEfED Curation

Fig. 3 shows 'BioScholar', a KEfED-enabled curation tool. This allows an researcher to draw a protocol in a graphical interface showing entities, processes and variables (constants, parameters and measurements) within an experiment. The system automatically builds data tables from the protocol design that could be used as the basis for a data repository. We have developed OoEVV to provide definitions of these elements as an ontology that can also support links to related terms in formal ontologies.

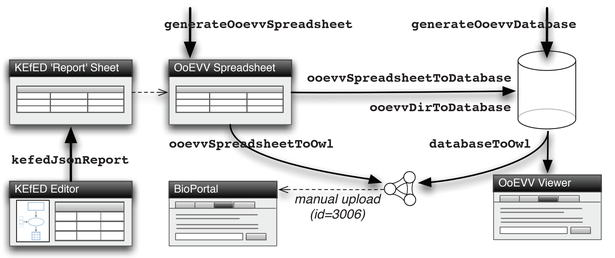

Figure 4: Organization of commands, components and data flow in the current KefedAdmin system.

OoEVV Tools and Curation

A goal of OoEVV is to provide a framework that domain experts can easily use. Fig. 4 shows the functional organization of a command-line application that uses spreadsheets to curate terminology (using standard file-sharing tools such as DropBox, Google Docs or Subversion to manage the files). Each separate Excel workbook corresponds to a separate ***OoEVVElementSet***. This permits us to provide detailed examples and instructions for handling exception cases in a way that we may adjust as the project progresses. The user can create a formatted spreadsheet (*generateOoevvSpreadsheet*) that may be filled out according to our curation manual (see http://www.isi.edu/projects/ooevv/curation). The user may add the contents of this file to an OWL file (permitting users to run a command (*ooevvSpreadsheetToOwl*) repeatedly over a set of spreadsheets to build an extended representation). A user may aggregate multiple spread-sheets into a MySQL database (*ooevvDirToDatabase / ooevvSpreadsheetToDatabase*) which then may be examined in a web-viewer application (Fig. 5). This example shows an antibody (typically used as a parameter in an experiment), and links to the EFO definition of an antibody. Finally, to provide a centralized set of definitions, a curator may run the *databaseToOwl* function that generates an OWL file to check that the model generated by the process is classifiable. This file may then be uploaded to the National Center of Biomedical Ontology's bioportal system to provide a centralized, versioned representation of OoEVV (http://bioportal.bioontology.org/ontologies/3006).

BIRN Applications and Users

A primary capability of our work within BIRN is to provide a simple methodology for us to construct ontologies for end-users that are appropriate for their needs. Given the large overhead incurred by building ontologies in various domains, we developed OoEVV to identify sets of sub-elements needed for their experimental work. Within BIRN, this was typically based on support of the BIRN mediator system (Ashish et al., 2010, Front. Neuroinform. 4:118). As an ODP, we anticipate that OoEVV tools may be used as a support system for other ontologies as our implementation improves. We currently are focussed on supporting numerous experimental domains including (a) neuroanatomical tract tracing experiments, (b) fMRI, (c) genetic childhood neurodevelopmental disorders, (d) radiation oncology studies, (e) stroke studies, (f) drug infusion studies, and (g) vaccine protection studies. Our development work within BIRN focusses on 'capabilities': http://www.birncommunity.org/capabilities/current-capabilities

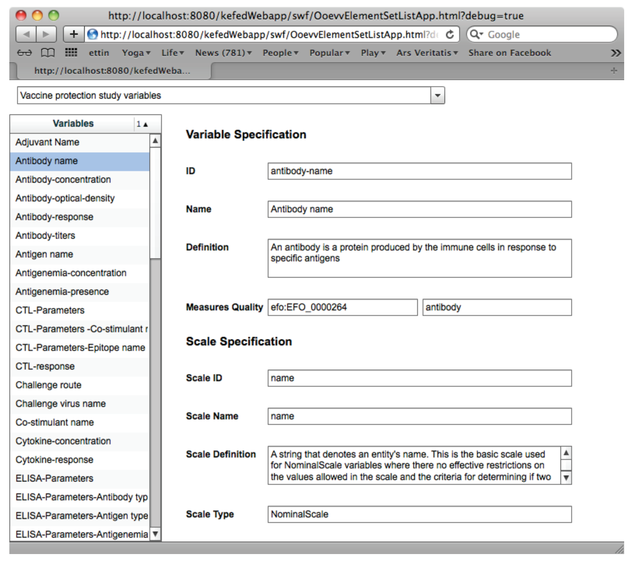

Figure 5: Ooevv Viewer Prototype

Related Work and Discussion

We have developped a lightweight system, appropriate to our needs but consistent with and interoperable with other efforts such as OBI and EFO. Some features of our representation are original and potentially important. The definition of multiple variables that measure the same underlying quality in different ways allows us to model accurately how different experimentalists gather data. Our detailed representation of different measurement scales provides an extensible 'type system' for each variable. We use UML as the base for our represention, restricting our use of OWL-based formal reasoning but enabling easier contruction of tools and curation processes. This is consistent with our focus on ODPs and tool construction but should be viewed as complementary and supportive of a more formal approach. Our OWL ontology is currently very simple, and future work will center on linking this to OBI and other efforts, further developing it's use in KEfED and other contexts in BIRN and information integration applications. The work that we describe here is supported by open-source tools available via http://bmkeg.isi.edu.

Acknowledgements

This work was supported by NIH with FBIRN (RR021992); Biomedical Informatics Research Network (RR025736); CogPO (MH084812); and BioScholar (GM083871). We thank Tom Russ, Swati Raina and Karthik Narasandra Manju-natha, Jose Luis Ambite, Maria Muslea, Naveen Ashish, Alex Paciorski, Ona Wu, and Vitali Moiseenko