Development of the EDDA Study Design Terminology to Enhance Retrieval of Clinical and Bibliographic Records in Dispersed Repositories

Ashleigh Faith, School of Library and Information Sciences, University of Pittsburgh iSchool; Eugene Tseytlin, Department of Biomedical Informatics, University of Pittsburgh School of Medicine; Tanja Bekhuis, Department of Biomedical Informatics, University of Pittsburgh School of Medicine

Background:
Medical terminology varies across disciplines and reflects linguistic differences in communities of clinicians, researchers, and indexers. Inconsistency of terms for the same concepts and lack of machine-readable metadata impede discovery of information artifacts, such as records of clinical reports and scientific articles that reside in various repositories. To facilitate discovery, retrieval, and data sharing, the medical community maintains an assortment of terminologies, thesauri, and ontologies. Valuable resources include the US National Library of Medicine Medical Subject Headings (MeSH), Elsevier Life Science thesaurus (Emtree), and the National Cancer Institute Thesaurus (NCIT). It is increasingly important to identify medical investigations by their design features, as these have implications for evidence regarding research questions.

Purpose:
Recently, Bekhuis et al (2013) found that coverage of study designs was poor in MeSH and Emtree. Based on this work, the EDDA Group at the University of Pittsburgh is developing a terminology of study designs. In addition to randomized controlled trials, it covers observational or uncontrolled designs.

Methods:
Among the resources analyzed thus far, inconsistent entry points, semantic labels, synonyms, and definitions are common. The EDDA Study Design Terminology is freely available in the NCBO BioPortal (http://purl.bioontology.org/ontology/EDDA). Some of the preferred terms have several variants, definitions sometimes compete, as well as other concept identifiers useful for researchers. The beta version was developed using the Protégé ontology editor v.4.3 (http://protege.stanford.edu) and distributed as a Web Ontology Language (OWL) file. Dublin Core Metadata Initiative (DCMI) protocols are in place for recording overall terminology metadata and OWL annotations.

Results:
At this preliminary stage, the term matrix consists of 171 class axioms consisting of study design terms, related terms, and publication types. When possible, class axioms were annotated with definition(s), incompatibility status, legacy term(s), controlled vocabulary resource unique identification, semantic type, and variant term annotation properties. In addition, revision metadata was also captured with editor annotations consisting of the team member who modified the class axiom and the date of modification. The following process was used for axiom enhancement (Figure 1):
Through the annotation process, a total of 2,381 axiom annotations were recorded. This included 51 MeSH, 33 NCIT, and 27 Emtree exact match access points to EDDA Study Design terms. Both MeSH and NCIT access points enabled information to be recorded. However, 12 of 27 Emtree access points did not result in any information because of insufficient information. Because NCIT cross-references other controlled vocabularies, 33 Unified Medical Language System (UMLS) resources also contributed to axiom annotations. A total of 123 definitions, 14 instances of incompatibility, 33 legacy terms, 95 semantic types, and 1,349 term variations were recorded.

**Conclusions & Future Work:**

Identifying and retrieving reports of medical investigations by design features is increasingly possible, primarily through linking metadata. Further development entails adding definitions from other sources, mapping relationships among terms, and integrating terms from existing vocabularies, particularly the Information Artifact Ontology. A primary goal is to improve identification and retrieval of electronic records describing studies in dispersed data warehouses or electronic repositories.

**Acknowledgments**

This research was partially supported by the US National Library of Medicine (NLM), National Institutes of Health, grant number 5R00LM010943. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NLM.