Faster and Better Metadata Authoring using CEDAR's Value Recommendations

Marcos Martinez-Romero, Martin J. O' Connor, Ravi D. Shankar, Maryam Panahiazar, Debra Willrett, Attila L. Egyedi, John Graybeal, and Mark A. Musen

Stanford University, Stanford, CA 94305, U.S.A.

CEDAR
CENTER for EXPANDED DATA ANNOTATION and RETRIEVAL

NIH
Big Data to Knowledge(BD2K)

CEDAR is developing novel methods and tools to simplify the process by which investigators annotate their experimental data with metadata

In biomedicine, **good metadata is crucial** to finding experimental datasets, to understand how experiments were performed, and to reuse data to conduct new analyses

The **CEDAR Workbench** (cedar.metadatacenter.net) is a set of Web-based tools for the acquisition, storage, search, and reuse of metadata acquisition forms (metadata templates). One of those tools is the **Metadata Editor**, which allows users to fill in metadata templates with metadata

Scientists

fill in templates →

Metadata Editor

templates filled in

We have enhanced the Metadata Editor with **value recommendation** capabilities

stored   indexed

CEDAR's Value Recommender identifies common patterns in the metadata repository, and generates **real-time suggestions** for filling out metadata templates

Metadata Repository
mongoDB JSON-LD

Search Engine
elasticsearch

real-time analysis

Value Recommender

suggest values

BioSample Human (Metadata Editor)

My Template
Fields
Sample N me
Organism
Organism Part
Sampling Location
Sampling Protocol

Sample Name*
056X

Organism
Homo sapiens

Tissue
brain

Sex
male

Isolate*
N/A

Age*
74 years

Biomaterial Provider*
Life Technologies

Optional Attribute 1

Name
disease

Value

- Parkinson's disease (DOID) (39%)
- central nervous system lymphoma (DOID) (27%)
- autistic disorder (DOID) (22%)
- melanoma (DOID) (5%)
- Edwards syndrome (DOID) (2%)
- schizophrenia (DOID) (1%)

The 'BioSample Human' template contains the fields used in BioSample for human samples

BioSample is a database from the National Center for Biotechnology Information (NCBI) that captures descriptive information about the biological samples investigated in projects

The suggestions provided are context-sensitive, meaning that the values predicted for a particular field are **generated and ranked based on previously entered values**

more examples →

The Value Recommender supports both **free-text values** and **controlled terms**. In this particular example, the system suggests terms from the Human Disease Ontology (DOID)

✔ **Decrease authoring time**

✔ **Increase metadata quality**

✔ **Reusable metadata**

CEDAR's value recommendation technology leverages **existing metadata** to make the authoring of **high-quality metadata** a **manageable task**

Figures above show the top 10 suggestions provided by the Value Recommender for the 'disease' field of the 'Biosample Human' template. These recommendations were generated based on over 35K instances of the 'BioSample Human' template

Chart 1: disease = ? (no context) — Recommendation score
hepatocellular carcinoma, lipid metabolism disorder, carcinoma, lung cancer, lymphoma, cancer, acute myeloid leukemia, asthma, breast cancer, rheumatoid arthritis

Chart 2: tissue = lung, disease = ? — Recommendation score
lung cancer, chronic obstructive pulmonary disease, lung squamous cell carcinoma, idiopathic pulmonary fibrosis, lung adenocarcinoma, adenocarcinoma, carcinoma, cancer, non-small cell lung carcinoma, pulmonary fibrosis

Chart 3: tissue = liver, disease = ? — Recommendation score
hepatocellular carcinoma, hepatitis C, liver cirrhosis, carcinoma, liver cancer, cancer, hepatocellular adenoma, hepatitis B, adenoma, alcohol abuse

Based on: Panahiazar et al. Context Aware Recommendation Engine for Metadata Submission. First Int Work Capturing Sci Knowl. 2015;3–7.

11/2016