Under the umbrella of the Center for Enhanced Data Annotation and Retrieval (CEDAR) metadata enrichment effort, we are building the Semantic Annotation Pipeline (SAP) that automates the semantic annotation of biomedical data stored in community data repositories. The pipeline has two major segments: 1) transform the metadata stored in the community data repository to CEDAR JSON-LD format, and 2) add semantic annotations to the CEDAR-formatted metadata.

The Ingestor transforms the metadata in the community repository to instances of a JSON-encoded CEDAR community template. The template includes ontological information on value restrictions on specific metadata elements. The CEDAR Community Template encapsulates the metadata specific to that community.

The Annotator annotates metadata with ontology terms as specified in the template. The ontology terms are served by the NCBO BioPortal. At the core of the Annotator is a named-entity recognition tool such as NCBO Annotator.

The annotated metadata is stored in the CEDAR repository.

An Illustration of SAP: Semantic Annotation of GEO Sample Metadata

GEO Sample Metadata in GEO Repository

**SAMPLE = GSM1169960**
Sample_title = IPF23lung
Sample_type = RNA
Sample_source_name_ch1 = Homo sapiens
Sample_characteristics_ch1 = tissue: Lung
Sample_molecule_ch1 = total RNA
Sample_label_ch1 = biotin
Sample_protocol_ch1 = Biotinylated cRNA were prepared with the Ambion MessageAmp kit for Illumina arrays

GEO Sample Metadata in CEDAR JSON-LD format

```
{
  "sampleID": {
    "@value": "GSM1169960"
  },
  "description": {
    "@value": "IPF"
  },
  "tissue": {
    "@type": "http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C12801",
    "@valueLabel": "Lung Tissue",
    "@id": "http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C33024"
  },
  "disease": {
    "@type": "http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C7057",
    "@valueLabel": "Idiopathic Pulmonary Fibrosis",
    "@id": "http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C35716"
  }
}
```

Semantically Annotated GEO Sample Metadata in CEDAR JSON-LD format

```
{
  "sampleID": {
    "@value": "GSM1169960"
  },
  "description": {
    "@value": "IPF"
  },
  "tissue": {
    "@type": "http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C12801",
    "_valueLabel": "Lung Tissue",
    "@id": "http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C33024"
  },
  "disease": {
    "@type": "http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C7057",
    "_valueLabel": "Idiopathic Pulmonary Fibrosis",
    "@id": "http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C35716"
  }
}
```

GEO Sample Metadata in CEDAR JSON-LD format

```
{
  "sampleID": {
    "@value": "GSM1169960"
  },
  "description": {
    "@value": "IPF"
  },
  "tissue": {
    "@type": "http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C12801",
    "@value": "Lung"
  },
  "disease": {
    "@type": "http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C7057",
    "@value": "null"
  }
}
```

Semantically Annotated GEO Sample Metadata in CEDAR JSON-LD format

```
{
  "sampleID": {
    "@value": "GSM1169960"
  },
  "description": {
    "@value": "IPF"
  },
  "tissue": {
    "@type": "http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C12801",
    "@id": "http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C33024"
  },
  "disease": {
    "@type": "http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C7057",
    "@id": "http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C35716"
  }
}
```

An Illustration of SAP: Semantic Annotation of GEO Sample Metadata

CEDAR is supported by grant U54 AI117925 awarded by the National Institute of Allergy and Infectious Diseases through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov)