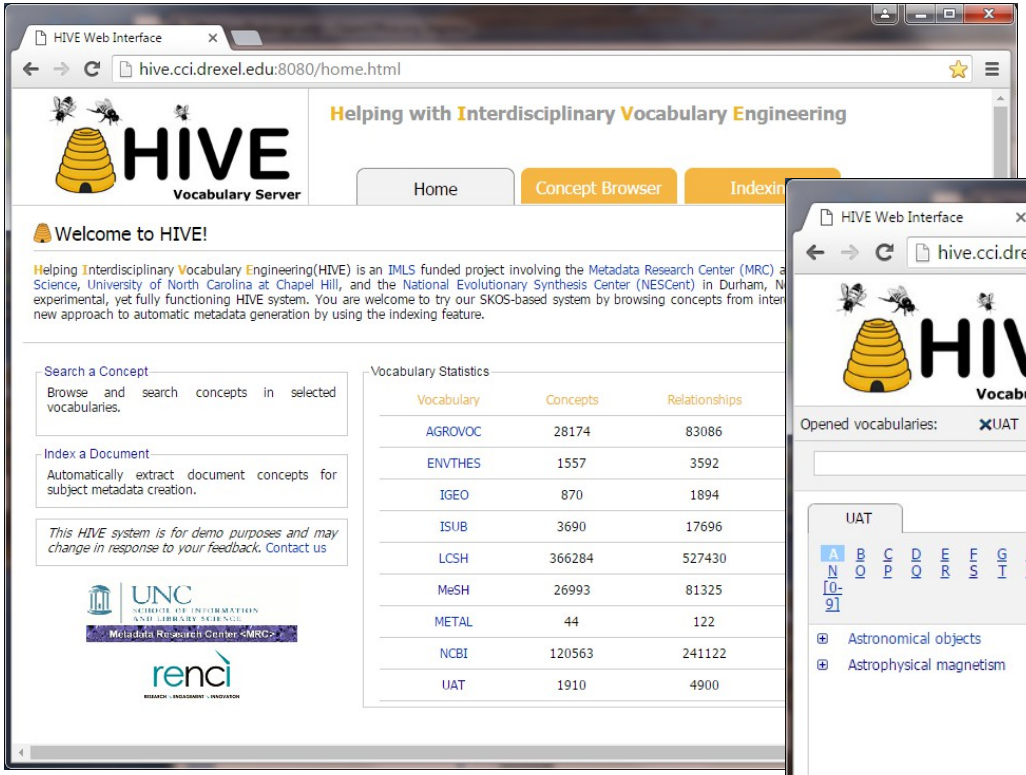
HIVE Update

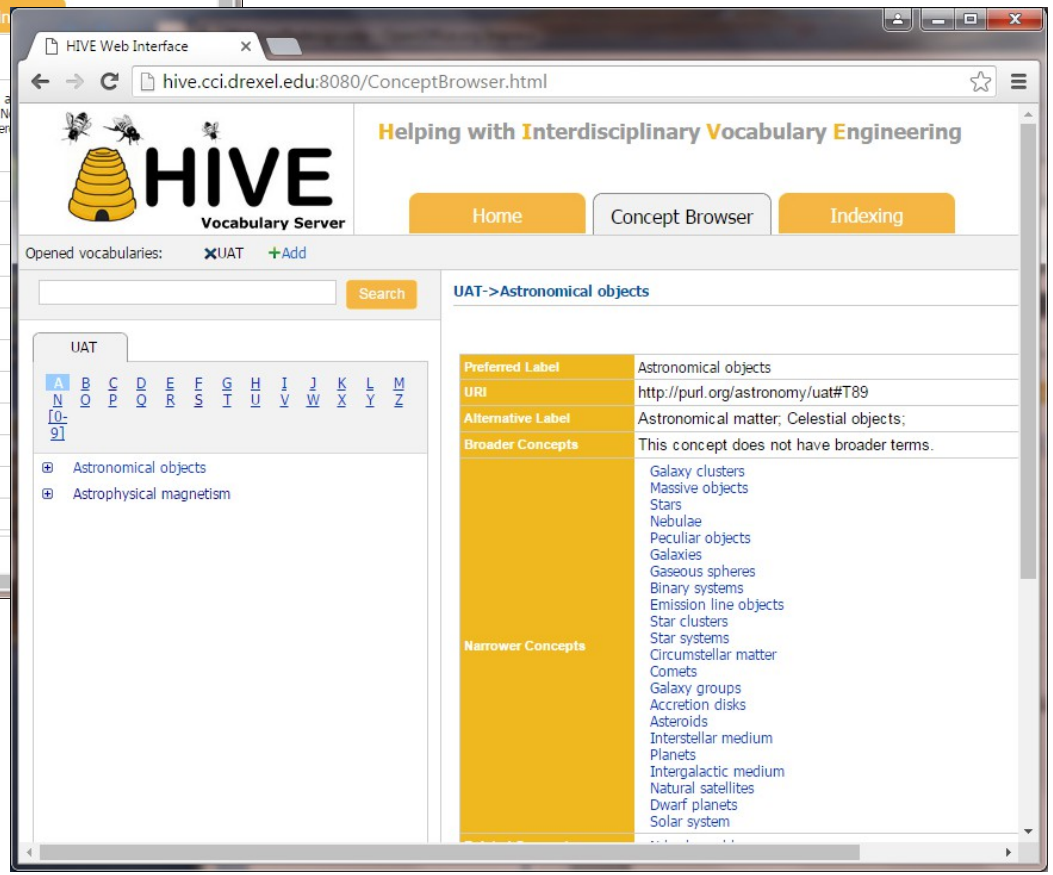
Joan Boone
jpboone@email.unc.edu

HIVE 1.0

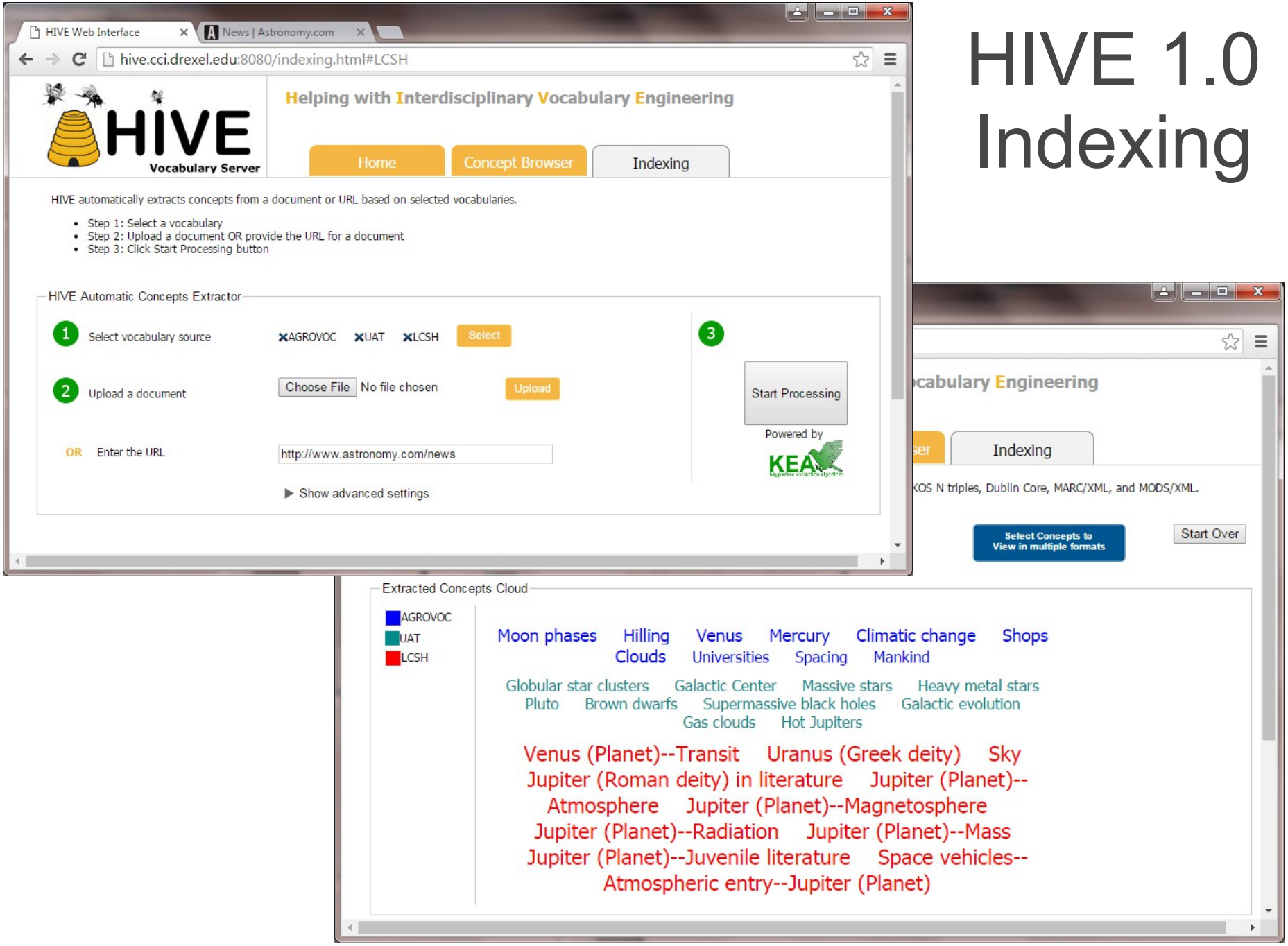
Available Vocabularies



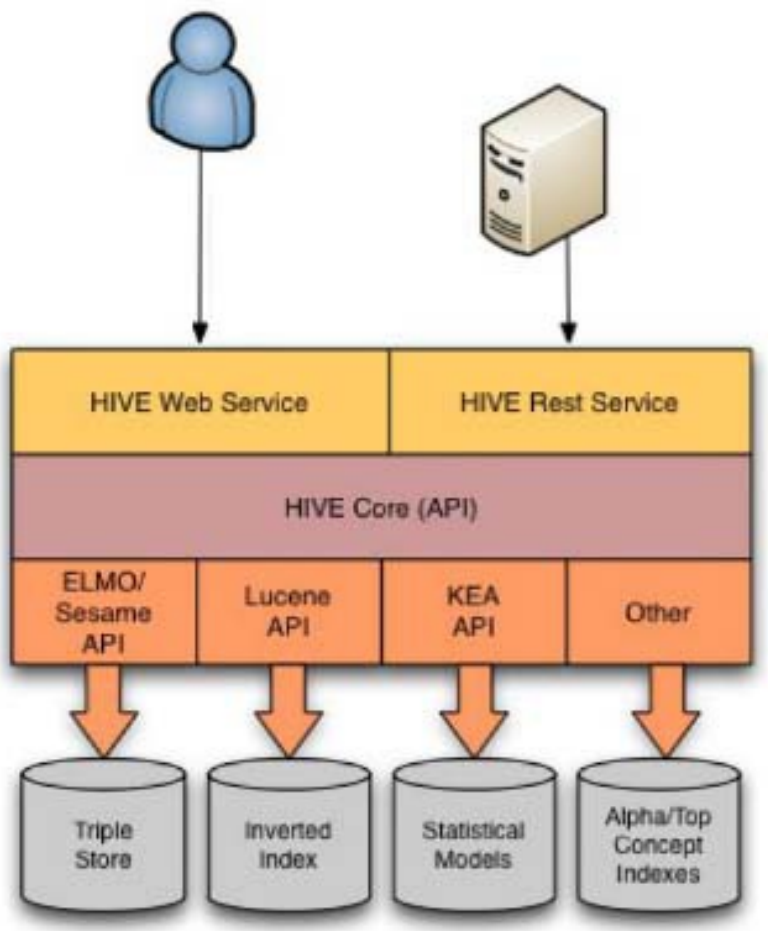
Vocabulary
Browse and Search



HIVE 1.0 Indexing



HIVE 1.0 Architecture



What's in the Dockerized HIVE

- Tomcat Server
- HIVE Web Services
- HIVE Core API
- Data stores and indexes
- Two vocabularies (AGROVOC, UAT)

Link to HIVE Docker files on GitHub

HIVE 1.0 Services

Vocabulary info

- List of available vocabularies
- Number of concepts and relationships for a vocabulary
- Alphabetic and top concept indexes for a vocabulary

Browsing and search

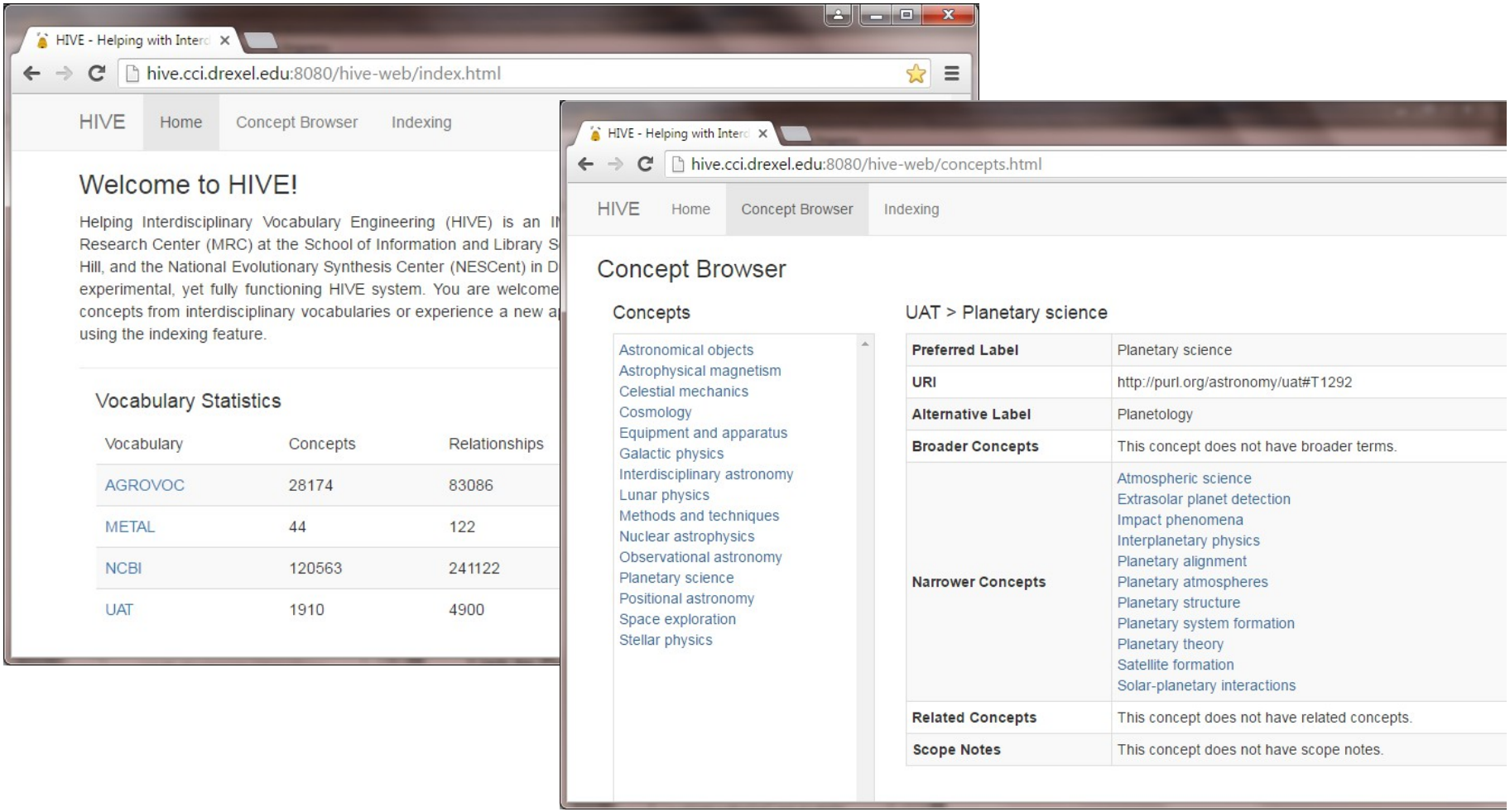
- Browse the vocabulary
- Browse the details for a specific concept (preflabel, altlabel, broaders, narrowers, relateds, SKOS format)
- Search concepts by keyword across multiple vocabularies

Indexing

- Index a URL or text document using multiple vocabularies

List of available HIVE Services and examples at Drexel

Redesign of Web Interface using HIVE 1.0 Services



Developed with basic front-end web technologies (HTML, CSS, JavaScript) using HIVE 1.0 services.

Link to HIVE 1.0 Web interface redesign at Drexel

HIVE 2.0 Redesign Objectives

- Implement with Python to address skills availability for developing and supporting HIVE
- Simplify vocabulary import and data model
- Improve web services interface for indexing and search functions
- Leverage existing Python tools and libraries
  - CherryPy for web framework
  - Natural Language Toolkit (NLTK 3.0) for text analysis: parsing, stemming, filtering, classification
  - RDFLib to read, parse, and transform RDF vocabularies
  - SQLite for relational database access

HIVE 2.0 Work in progress

- Import of RDF vocabularies using RDFLib, with conversion to a SQLite database
- Text processing using NLTK to tokenize, stem, and filter document content
- Installed CherryPy web server at Drexel
- Deployed test app that generates top concept hierarchy from UAT database

Converting RDF vocabularies to SQL database

RDF file snippet

```
<rdf:Description rdf:about="http://www.altbibl.io/astronomy/uat/xl_en_15xytrsw">
    <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
    <hasStatus xmlns="http://art.uniroma2.it/ontologies/vocbench#">Published</hasStatus>
    <dct:created  rdf:datatype="xsd:http://www.w3.org/2001/XMLSchema#dateTime">2014-09-10T19:51:06Z</dct:created>
    <dct:modified rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2014-09-10T19:54:46Z</dct:modified>
    <skos:broader rdf:resource="http://www.altbibl.io/astronomy/uat/xl_en_zk4ftwhn"/>
    <skos:inScheme rdf:resource="http://www.altbibl.io/astronomy/uat"/>
    <skosxl:altLabel rdf:resource="http://www.altbibl.io/astronomy/uat/xl_en_3527dfde"/>
    <skosxl:prefLabel rdf:resource="http://www.altbibl.io/astronomy/uat/xl_en_10ba7619"/>
</rdf:Description>
```

- Used RDFLib to parse the vocabulary in RDF format, and create a graph database of triples
- Graph database is large (~90MB)
  - Slow to access and search
  - Contains 30K+ triples (for a vocabulary with ~1800 terms)
- SQLite (relational) version is fast and small (~1.1MB)
  - Removed triples not used by HIVE
  - 4K+ 'triples' are defined as relations

HIVE 2.0 Test App:
Top Concept Hierarchy for UAT