# Semantic Exploration of Native American Ethnobotany

*Submitted as required engineering project report*

**Indiana University ILS-Z636: Data Semantics, Fall 2014**

**Professor Ying Ding**

Stefan M. Furrer
Indiana University
Bloomington, IN
stfurrer@indiana.edu

Jeremy J. Yang
Indiana University
Bloomington, IN
jejyang@indiana.edu

## ABSTRACT
A semantic knowledge base and web interface was developed for the exploration of Native American Ethnobotany. We used the Jena API and toolkit and modern W3C semantic web standards including RDF, Turtle and Sparql to model the data. Public ontologies were used in compliance with Linked Open Data (LOD) principles. NCBI Taxonomy[1] and MeSH[2] controlled vocabulary were used as well.

## Categories and Subject Descriptors
H.4 [**Information Systems Applications**]: Miscellaneous

## Keywords
Semantic web, RDF, Sparql, ethnobotany, library science, informatics, data science, cultural heritage

## 1. INTRODUCTION
Ethnobotany studies both the plants and their relationship to a culture, in this case, indigenous people of North America. In addition to its value to social science, it is often the case that cultural heritage knowledge[9] can provide unique insights for the natural sciences, for example, traditional medicines. Native American Ethnobotany[3] has been studied and cataloged by various scholars[13] including for an online database at U. Michigan - Dearborn[4], which was first released as an electronic database in 1977. Cultural heritage knowlege in general, and ethnobotany in particular, is highly contextual. In other words, to correctly represent and apply knowledge across cultures requires sufficient understanding of those cultures, especially their languages and modes of communication. This can be considered metadata, but a deeper understanding of metadata leads us directly to concepts and methods of data semantics technologies. Learning from cultural heritage is a type of deep semantic translation. This project applies semantic methods and technologies to the ethnobotanical knowledge learned and propagated over the millenia by Native Americans. However, the breadth of this project is limited by the scope of this class project.

[1]http://www.ncbi.nlm.nih.gov/taxonomy
[2]http://www.ncbi.nlm.nih.gov/mesh
[3]https://en.wikipedia.org/wiki/Native_American_ethnobotany
[4]http://herb.umd.umich.edu/

## 2. THEORY AND BACKGROUND
This project and the Data Semantics course itself are based on several foundational disciplines including: library and computer science, logic, and linguistics[1]. This fusion is generally reflected in the new fields of "informatics" and "data science." The terminology in part reflects a change of emphasis, toward the data, information and knowledge and its representation, semantics and processing. In its highest aspiration, knowledge processing is synonymous with intelligence. With this very abstract background, the research question driven by this project is whether empirical knowledge of Native American ethnobotany can be represented and queried via semantic methods, and whether such methods can yield advantages. By definition, ethnobotany data is empirical, based on oral traditions and teachings, culturally propagated knowledge, accompanied by strong belief and contrasting sharply with modern laboratory results with scant data but mechanistic implications. A systematic evaluation and comparison of indigenous pharmacopoeias is of great interest for natural product discovery in the pharmaceutical industry as well as to contribute to improved health care in marginalized regions. The cultural and linguistic diversity represented in the numerous collections of ethnobotanical accounts worldwide presents an unique opportunity for the vision of the Semantic Web, where heterogenous data can be queried and different services are integrated [4].

### 2.1 Related Work
Medicinal plants are an important component of indigenous medicine systems since ancient time and often perceived as a cultural heritage. Major research efforts in collecting ethnobotanical reports have resulted in vast amount of information about medicinal plants and cultures [8]. With having often a cultural or regional focus, such ethnobotanical databases lack uniformity and interlinkability [16], both ma-

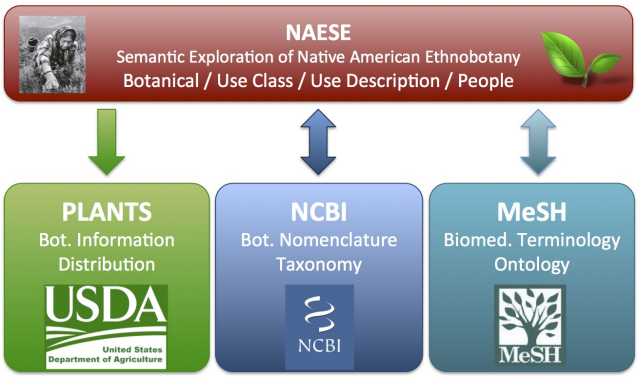


**Figure 1: Naese data sources and connections.**

jor efforts of the Linked Open Data[5] initiative. The integration of ontologies, Semantic Web languages (XML, RDF and OWL) or taxonomic standards that would enable an automated data exchange and search federation, has been discussed for particular cultural collections [10], [15], [20], but is not consistently implemented on a global scale. One of the few examples of Semantic Web efforts in the area of traditional medicine is RDF-TCM[6], containing information on over 800 herbs used in Traditional Chinese Medicine (TCM) and links to DBpedia[7] and DailyMed[8] among others [19].

## 3. DATASETS

The database of Native American Ethnobotany focuses on the traditional usage of plants by Native American people, as medicines but also other purposes, such as food [12],[13]. The data was accessed from the Native American Ethnobotany database on September 13, 2014 through a series of genera specific searches and the data processed into a tabular format. For the purpose of this project, only information on ethnomedicinal plant knowledge was considered, or 25'311 different usages for 3'074 kinds of plants (species, subspecies, varieties), ranging from cold remedies and analgesics to witchcraft uses.

### 3.1 Data Standards and Annotation

The heterogeneity of ethnobotanical data, amplified by cultural and linguistic differences, together with a lack of standard terminology are adding complexity to the integration and interlinking with other information resources [7]. The utilization of a common terminology to describe a medicinal plant and its usage are prerequisites for developing a system that allows automated data exchange. In addition to plant usage categories assigned in Native American Ethnobotany, the particular bits of information about the utilization of plants were indexed using Medical Subject Headings (MeSH) vocabulary [5]. The NLM Medical Text Indexer (MTI)[9] was utilized to efficiently and consistently annotate the detailed plant usage indications with an ontology based biomedical terminology [14]. Out of a total of 19'382 unique medicinal plant use indications, 727 or 4% resulted in errors and could not be annotated with corresponding MeSH terms. A manual evaluation of the selected records indicated an excellent and correct indexing, although in some cases word sense ambiguity lead to contextually incorrect annotation, such as the tagging of "(plant) gum" with "Gingiva" (MeSH: D005881) [18].

**Table 1: Datasets used and number of unique entities represented in project**

| Name | Publisher | Entities |
|---|---|---|
| Native American Ethnobotany | U. Michigan | 25'311 |
| Taxonomy | NCBI | 2'692 |
| MeSH | NLM | 1'442 |
| PLANTS | USDA | 2'921 |

Only few databases emphasize the application of a standardized identification and taxonomy for botanicals and synonyms, superfluous, ambiguous or misspelled taxon names result in mismatched records. The Taxonomic Name Resolution Service (TNRS)[10] [6] provides a botanical name parsing and fuzzy matching against multiple reference taxonomies, including the NCBI taxonomy[11] [17],[3] and USDA PLANTS[12] database. Careful manual curation of the taxonomy matching results was performed to enable a correct and accurate alignment with the taxonomies if possible. Due to different taxonomic coverages, an assignment of to the exact subspecies or species was not possible in every case. Missing botanical information resulted in assigning to a higher taxonomic level. For the 3'074 different botanicals, the taxonomic alignment resulted in 114 different subspecies/variety, 2'230 species and 348 unique genus.

### 3.2 Knowledge Representation: Modeling the Data

The ethnobotanical data and its biomedical and taxonomic annotation was organized and converted into a knowledge representation of the domain. The raw data files were processed, combined, and converted to RDF, serialized as RDF-XML and TTL formats, using OpenRefine[13]. An RDF data skeleton was defined according to domain concepts and predetermined search scenarios. Throughout the development of the data skeleton, care was taken to identify concepts (or classes) of prominent dictionaries, such as Dublin Core Metadata Intiative (DCMI)[14] that can be reused, such as dc:title. The knowledge base skeleton was structured for a botanical to be identified by its botanical and common name, as well as taxonomy links. Each botanical is a member of a certain use description that references the corresponding biomedical terminology, usage category as well as the Native People for which the use was reported. The NCBI taxonomy and MeSH terminology were integrated through interlinking to the URI's of the RDF representations of the

[5]http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
[6]http://www.open-biomed.org.uk/rdf-tcm/ (SPARQL endpoints not operational 12/12/2014)
[7]http://dbpedia.org/
[8]http://dailymed.nlm.nih.gov/dailymed/
[9]http://skr.nlm.nih.gov/batch-mode/mti.shtml
[10]http://tnrs.iplantcollaborative.org
[11]http://www.ncbi.nlm.nih.gov/taxonomy/
[12]http://plants.usda.gov
[13]http://openrefine.org
[14]http://dublincore.org/documents/dcmi-terms/

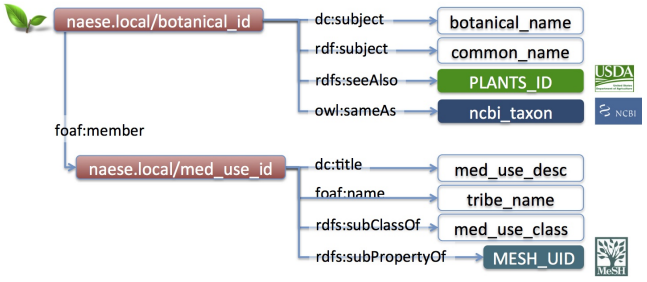


**Figure 2: Naese knowledge base skeleton.**

corresponding datasets in bio2rdf[15] [2]. Since the USDA PLANTS taxonomy is not available in RDF format, the knowledge base was populated with the corresponding ID instead of referencing these resources (using URI).

# 4. ENGINEERING PROJECT: WEB APPLICATION NAESE

The purpose of the web app in this project is to demonstrate the semantic data methods which are the central aim. An effective web app not only provides end users with a usable interface, but can accelerate and improve development by providing a convenient testing tool. Web apps can be conveniently deployed over heterogeneous networks for distributed audiences. This was an important capability for our distributed team of distance students, located in Ohio and New Mexico, for a course based in Indiana.

## 4.1 Methods and Tools

Naese was implemented as a Java Enterprise web app and deployed via WAR file using Tomcat. Naese was built using Jena 2.12.0[16]. Jena is described as "A free and open source Java framework for building Semantic Web and Linked Data applications". In this project, several Jena methods and tools were employed, including the Java API and command-line applications.

With Jena there are several options for dataset storage. To implement a self-contained WAR, for convenient and robust web app deployment, the RDF data file (RDF-XML) was included and an in-memory datastore is initialized on web app deployment. This improves performance at runtime by avoiding overhead, by pre-loading data. Jena can also be used to deploy a Sparql endpoint (service) using Fuseki[17]. Also Jena TDB[18] can be used to implement persistent storage from an input file, and support read/write access. Sparql endpoint deployment also supports LOD standards and applications from generic Sparql clients, including federated queries.

## 4.2 Case Study: Example of Pain Remedies

The benefits of using standardized terminology and ontologies is highlighted for instance, on pain remedies utilized by Native American people. Querying the dataset for botanicals associated with the MeSH term "Pain" (mesh: D010146), does recruit botanicals from many different use classification, including analgesic, dermatological and gynecological uses, among others. A close look at the use descriptions of these botanicals does highlight the value of the controlled terminology of MeSH and the MTI tagger, as not only "pain", but also synonyms of the term, such as "aches" or "sore" are classified alike. The ranking of these botanicals by their most frequent use, in terms of number of tribes associated with the botanical and "pain" MeSH term, products a significance ranking or cross-cultural validation, biased by a botanicals geographic distribution. *Acorus calamus*, also called Sweet Flag or Calamus, is one of the top ranked species and the analgesic effect of ethanolic root extracts has been studied in mice and significant activity has been observed [11].

## 4.3 Example Queries, Use Cases

One approach to data resource design is to begin with the desired queries, that is, the questions which should be answerable, and design the system to fulfill those design requirements. Whether this approach is used or not, typically, since development often iterates cyclically, the effect will be the same: databases are about answering questions, so good example questions are essential, and in many ways define the project. Output is excerpted below due to space limitations. Full queries and output is included in supplementary materials.

**Sparql headers:**

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf:
    <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX tax: <http://bio2rdf.org/taxonomy:>
PREFIX voc: <http://bio2rdf.org/taxonomy_vocabulary:>
PREFIX mesh: <http://bio2rdf.org/mesh:>
PREFIX usda:
    <http://plants.usda.gov/core/profile?symbol=>
PREFIX local: <http://naese.local/>
```

**Example query: Toothache remedies** Find all botanicals used for toothache. Note that toothache remedies are a subclass of analgesics. This illustrates a fundamental advantage of semantic methods, as the query could easily be broadened. Likewise, a species can be associated with related species via taxonomy subclasses.

```
SELECT DISTINCT ?id ?species ?tribe ?title ?subClassOf
WHERE {
  FILTER (regex(?subClassOf, 'Toothache', 'i')) .
  ?usage rdfs:subClassOf ?subClassOf .
  ?usage foaf:name ?tribe .
  ?usage dc:title ?title .
  ?id foaf:member ?usage .
  ?id dc:subject ?species .
}
ORDER BY ?species ?tribe
```

**Example query: Plant name** Search for botanical by official name (Latin). Show common name, tribe, usage, and external linking IDs.

[15]http://bio2rdf.org
[16]http://jena.apache.org/
[17]http://jena.apache.org/documentation/serving_data/
[18]https://jena.apache.org/documentation/tdb/


**Figure 3: Naese web app with example query loaded.**

```
SELECT ?common_name ?tribe ?title ?medical_class
    ?plants_uri ?ncbi_tax ?mesh_id
WHERE {
  ?id dc:subject 'Hibiscus tiliaceus L.' .
  ?id rdf:subject ?common_name .
  ?id rdfs:seeAlso ?ncbi_tax .
  ?id owl:sameAs ?plants_uri .
  ?id foaf:member ?usage .
  ?usage dc:title ?title .
  ?usage foaf:name ?tribe .
  ?usage rdfs:subClassOf ?medical_class .
  ?usage rdfs:subPropertyOf ?mesh_id .
}
```

**Example query: Federated query** Use SERVICE subquery for remote endpoint access, to include linked MeSH terms and descriptions. *(Not supported by Naese web app, due to Jena API limitations.)*

```
SELECT ?naese_id ?subject ?mesh_id ?mesh_term
    ?mesh_description
WHERE {
  ?naese_id dc:subject 'Hibiscus tiliaceus L.' .
  ?naese_id rdf:subject ?subject .
  ?naese_id rdfs:subPropertyOf ?mesh_id .
  ?mesh_id rdf:type local:mesh_UID
  SERVICE <http://mesh.bio2rdf.org/sparql>
  {
    SELECT ?mesh_id ?mesh_term ?mesh_description
    WHERE {
      ?mesh_id voc:mesh-heading ?mesh_term .
      ?mesh_id voc:mesh-scope-note ?mesh_description
    }
  }
}
```

**Example query: Bio2RDF MeSH endpoint** Using a generic Sparql client, fetch by MeSH ID. In effect this implements a manual federated search.

```
$ sparql_query.py --url http://mesh.bio2rdf.org/sparql
    --rqfile mesh_remote.rq --fmt TTL
@prefix res: <http://www.w3.org/2005/sparql-results#> .
@prefix rdf:
    <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
_:_ a res:ResultSet .
_:_ res:resultVariable "MeSH_Term" , "MeSH_Description"
    .
_:_ res:solution [
res:binding [ res:variable "MeSH_Term" ; res:value
    "Plant Roots" ] ;
res:binding [ res:variable "MeSH_Description" ;
    res:value "The usually underground portions of a
    plant that serve as support, store food, and
    through which water and mineral nutrients enter
    the plant. (From American Heritage Dictionary,
    1982; Concise Dictionary of Biology, 1990)" ] ] .

sparql_query.py: input file: mesh_remote.rq
```

## 5. EVALUATION

The usability and functionality of the Naese web app was evaluated by the project team in the context of molecular discovery science. The simple textual input for Sparql does not provide GUI features for editing, but the example queries do provide multiple useful, editable templates. The output is simple text, and could be improved by formatting and alternate modes, especially downloadable files. Currently there is no support for federated queries (e.g. via SERVICE), so this could be a major enhancement, but would require re-architecture to employ a Sparql endpoint instead of embedded Jena TDB triple store. This would also enable rule based queries, utilizing utilizing the MeSH ontology and NCBI taxonomic classification. Further enrichment of the data set, for example through reconciliation with DBpedia, as well as integration of other ethnobotanical data sources, such as Plants for a Future[21] would allow searches across cultural and regional collections. A refinement of the

[21] http://www.pfaf.org

Table 2: Toothache query output (partial)

| species | tribe |
|---|---|
| Aesculus californica | Mendocino |
| Cephalanthus occidentalis L. | Cherokee |
| Cephalanthus occidentalis L. | Choctaw |
| Cephalanthus occidentalis L. | Kiowa |
| Cephalanthus occidentalis L. | Thompson |
| Juglans nigra L. | Cherokee |
| Juglans nigra L. | Choctaw |
| Juglans nigra L. | Thompson |
| Salix sp. | Cherokee |
| Salix sp. | Choctaw |
| Salix sp. | Kiowa |
| Salix sp. | Thompson |
| Sambucus racemosa L. | Okanagon |
| Sambucus racemosa L. | Thompson |
| Shepherdia rotundifolia Parry | Navajo, Kayenta |
| Zanthoxylum americanum P. Mill. | Iroquois |

Table 3: Plant name query output (partial)

| | |
|---|---|
| **common_name** | Sea Hibiscus |
| **tribe** | Hawaiian |
| **medical_class** | Pulmonary Aid |
| **title** | Bark and other plants crushed, water added, strained and resulting liquid taken for congested chest |

data model and further utilization of existing dictionaries, such as skos:altLabel for the "common name" and a local resource for the tribe (tribe_id).

## 6. CONCLUSIONS

A data model was developed to represent and process knowledge about Native American ethnobotany. This data model was implemented using Jena and deployed as a JEE web app for testing and usage. A range of queries which demonstrate a range of capabilities were developed and tested. The advantages of semantics are conferred and illustrated by subclass relationships such as toothache-remedies as a subclass of analgesics, and taxonomy subclass relationships to facilitate phylogenetic or chemotaxonomic similarity exploration.

## 7. ACKNOWLEDGMENTS

The authors would like to thank the instructor, Prof. Ying Ding, and assistant instructors of Data Semantics ILS-Z636, for their efforts and support, and especially for supporting the participation of distance learners through online education technologies, another very promising, rapidly evolving area.


Figure 4: USDA Plants page for Hibiscus tiliaceus L., symbol HITI[20].

## 8. AUTHOR CONTRIBUTIONS

SF conceived the project, developed the data model, acquired and pre-processed the data using OpenRefine, producing RDF in RDF-XML and TTL formats, and developed use-cases via example Sparql queries. JY developed Jena API code, Jena TDB command-line workflows, and Jena Fuseki tools, to deploy the dataset via TDB, as a Sparql endpoint, and via Java Enterprise web app, deployed via WAR file. SF and JY cooperatively and equally authored the report, tested and revised Sparql queries and use cases.

## 9. REFERENCES

[1] D. Allemang and J. Hendler. *Semantic Web for the working ontologist, 2nd Edition*. Morgan and Kaufmann, 2011.

[2] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette. Bio2rdf: towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform*, 41(5):706–716, Oct 2008.

[3] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. Genbank. *Nucleic Acids Res*, 37(Database issue):D26–31, Jan 2009.

[4] C. Bizer, T. Heath, and T. Berners-Lee. Linked

data-the story so far. *International journal on semantic web and information systems*, 5(3):1–22, 2009.

[5] O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32::D267, 2004.

[6] B. Boyle, N. Hopkins, Z. Lu, J. A. Raygoza Garay, D. Mozzherin, T. Rees, N. Matasci, M. Narro, W. Piel, S. Mckay, S. Lowry, C. Freeland, R. Peet, and B. Enquist. The taxonomic name resolution service: an online tool for automated standardization of plant names. *BMC Bioinformatics*, 14(1):16, 2013.

[7] K. Cheung and H. Chen. Semantic web for data harmonization in chinese medicine. *Chinese Medicine*, 2010.

[8] M. Heinrich, J. Kufer, M. Leonti, and M. Pardo-de Santayana. Ethnobotany and ethnopharmacology–interdisciplinary links with the historical sciences. *J Ethnopharmacol*, 107(2):157–160, Sep 2006.

[9] E. Hyvonen. *Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Synthesis Lectures on the Semantic Web: Theory and Technology*, volume 1:3. Morgan and Claypool, 2012.

[10] B. Kamsu-Foguem, G. Diallo, and C. Foguem. Conceptual graph-based knowledge representation for supporting reasoning in african traditional medicine. *Engineering Applications of Artificial Intelligence*, 26(4):1348 – 1365, 2013.

[11] M. A. A. Khan and M. T. Islam. Analgesic and cytotoxic activity of acorus calamus l., kigelia pinnata l., mangifera indica l. and tabernaemontana divaricata l. *J Pharm Bioallied Sci*, 4(2):149–154, Apr 2012.

[12] D. E. Moerman. An analysis of the food plants and drug plants of native north america. *Journal of ethnopharmacology*, 52(1):1–22, 1996.

[13] D. E. Moerman. *Native american ethnobotany*, volume 879. Timber Press Portland, 1998.

[14] J. G. Mork, A. J. Jimeno-Yepes, and A. R. Aronson. The nlm medical text indexer system for indexing biomedical literature, 2013.

[15] S. Mustaffa, R. Ishak, and D. Lukose. Ontology model for herbal medicine knowledge repository. In D. Lukose, A. Ahmad, and A. Suliman, editors, *Knowledge Technology*, volume 295 of *Communications in Computer and Information Science*, pages 293–302. Springer Berlin Heidelberg, 2012.

[16] S. S. Ningthoujam, A. D. Talukdar, K. S. Potsangbam, and M. D. Choudhury. Challenges in developing medicinal plant databases for sharing ethnopharmacological knowledge. *J Ethnopharmacol*, 141(1):9–32, May 2012.

[17] E. W. Sayers, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, I. Mizrachi, J. Ostell, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, E. Yaschenko, and J. Ye. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 37(Database issue):D5–15, Jan 2009.

[18] M. Stevenson and Y. Guo. Disambiguation in the biomedical domain: The role of ambiguity type. *Journal of Biomedical Informatics*, 43(6):972–981, 2014.

[19] J. Zhao. Publishing chinese medicine knowledge as linked data on the web. *Chinese Medicine*, 5(1):1–12, 2010.

[20] X. Zhou, Z. Wu, A. Yin, L. Wu, W. Fan, and R. Zhang. Ontology development for unified traditional chinese medical language system. *Artif Intell Med*, 32(1):15–27, Sep 2004.