# Framing NKOS Evaluation: HIVE's Approach

ASIST Annual Conference, Pittsburgh, PA
October 24, 2010

HOLLIE WHITE
JANE GREENBERG

# Overview

- Interdisciplinary needs: early evaluation
- HIVE—Helping Interdisciplinary Vocabulary Engineering
  >>Goals, status, and design
- HIVE evaluation
- Applicability to NKOS

# Purpose of HIVE

Address CV (controlled vocabulary) cost, interoperability, and usability constraints for interdisciplinary collections

- COST: Expensive to create, maintain, and use
- INTEROPERABILITY: Developed in silos (structurally and intellectually)
- USABILITY: Interface design and functionality limitations have been well documented

# Background Research: Interdisciplinary NKOS needs

Vocabulary analysis

- 600 keywords, Dryad partner journals

  Facets: taxon, geographic name, time period, topic, research method, genotype, phenotype

  Vocabularies: NBII Thesaurus, LCSH, the Getty's TGN, ERIC Thesaurus, Gene Ontology, IT IS (10 vocabularies)

Results

431 topical terms, exact matches

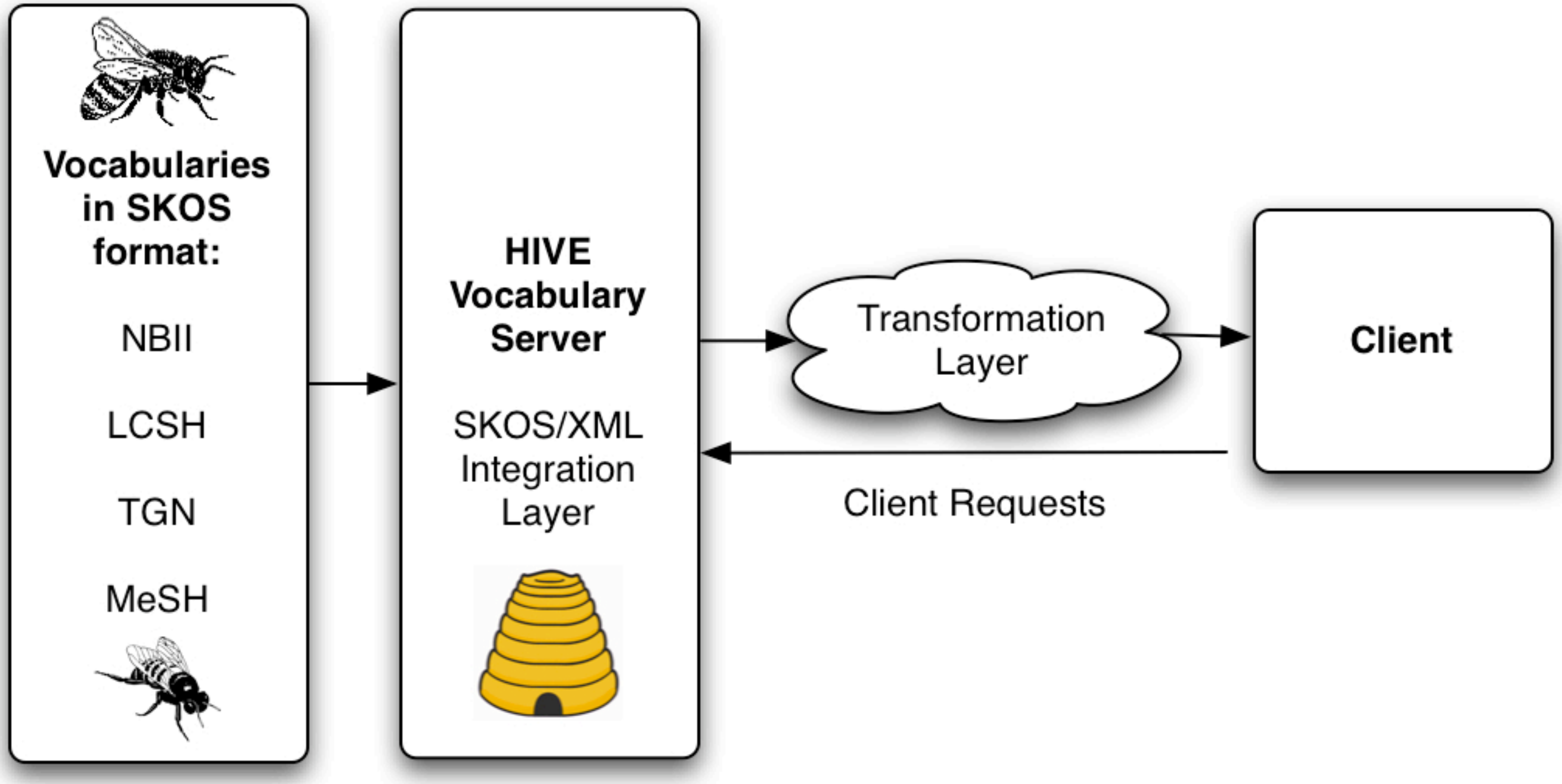
NBII Thesaurus, 25%; MeSH, 18%

531 terms (research method and taxon)

LCSH, 22% found exact matches, 25% partial

Conclusion: Interdisciplinary collections need multiple vocabularies

# HIVE model



- <AMG> approach for integrating discipline CVs
- Model addressing CV cost, interoperability, and usability constraints (interdisciplinary environment)

## HIVE Goals

- Automatic metadata generation approach that dynamically integrates discipline-specific controlled vocabularies encoded with the Simple Knowledge Organisation System (SKOS)

- *Provide efficient, affordable, interoperable, and user friendly access to multiple vocabularies during metadata creation activities*

- *A model that can be replicated*

## Three phases of HIVE:

1. *Building HIVE*

- *Vocabulary preparation*
- *Server development*

2. *Sharing HIVE*

- Continuing education (*empowering information professionals*)

3. *Evaluating HIVE*

- Examining HIVE
  - Primate Life Histories Working Group
  - Wood Anatomy and Wood Density Working Group

# HIVE Partners

## Vocabulary Partners

- Library of Congress:  LCSH
- the Getty Research Institute (GRI):  TGN (Thesaurus of Geographic Names )
- United States Geological Survey (USGS):  NBII Thesaurus, Integrated Taxonomic Information System (ITIS)
- Agrovoc Thesaurus

## Advisory Board

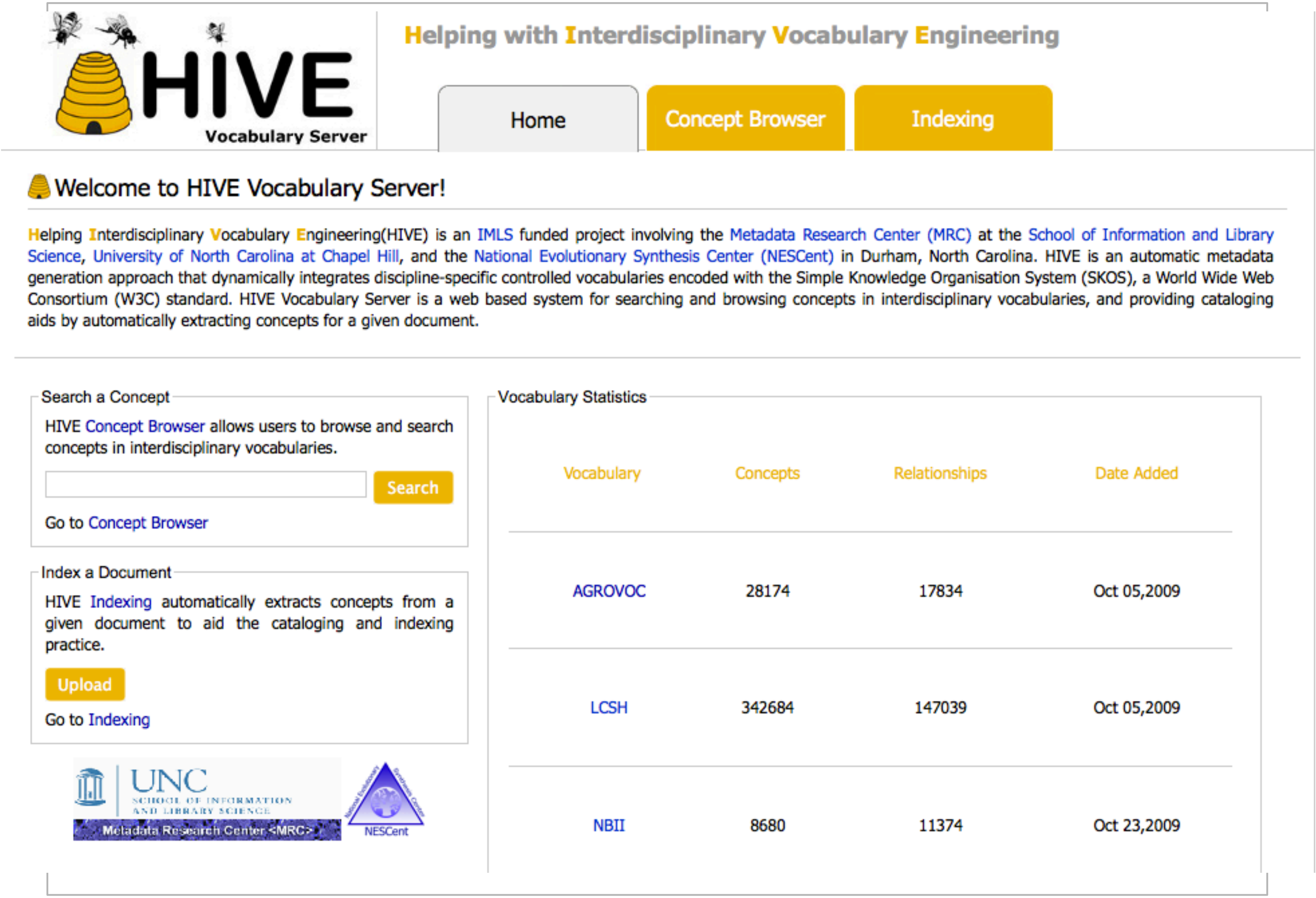
- Jim Balhoff, NESCent
- Libby Dechman, LCSH
- Mike Frame, USGS
- Alistair Miles, Oxford, UK
- William Moen, University of North Texas
- Eva Méndez Rodríguez, University Carlos III of Madrid
- Joseph Shubitowski, Getty Research Institute
- Ed Summers, LCSH
- Barbara Tillett, Library of Congress
- Kathy Wisser, Simmons
- Lisa Zolly, USGS

WORKSHOPS HOSTS:  Columbia Univ.; Univ. of California, San Diego; Univ. of North Texas; Universidad Carlos III de Madrid, Madrid, Spain

# HIVE Construction

- HIVE stores millions of concepts from different vocabularies, and makes them available on the Web by a simple HTTP
  - Vocabularies are imported into HIVE using SKOS/RDF format
- HIVE is divided in two different modules:

1. HIVE Core
   - SKOS/RDF storage and management (SESAME/Elmo)
   - **SMART HIVE**: Automatic Metadata Extraction and Topic Detection (KEA ++ and MAUI)
   - Concept Retrieval (Lucene)
2. HIVE Web
   - Web user Interface (GWT—Google Web Toolkit)
   - Machine oriented interface (SOAP and REST)


HIVE
Vocabulary Server
Home
Concept Browser
Indexing
Search
Upload
UNC
SCHOOL OF INFORMATION AND LIBRARY SCIENCE
Metadata Research Center <MRC>
National Evolutionary Synthesis Center
NESCent

Helping with Interdisciplinary Vocabulary Engineering

## Welcome to HIVE Vocabulary Server!

Helping Interdisciplinary Vocabulary Engineering(HIVE) is an IMLS funded project involving the Metadata Research Center (MRC) at the School of Information and Library Science, University of North Carolina at Chapel Hill, and the National Evolutionary Synthesis Center (NESCent) in Durham, North Carolina. HIVE is an automatic metadata generation approach that dynamically integrates discipline-specific controlled vocabularies encoded with the Simple Knowledge Organisation System (SKOS), a World Wide Web Consortium (W3C) standard. HIVE Vocabulary Server is a web based system for searching and browsing concepts in interdisciplinary vocabularies, and providing cataloging aids by automatically extracting concepts for a given document.

Search a Concept

HIVE Concept Browser allows users to browse and search concepts in interdisciplinary vocabularies.

Go to Concept Browser

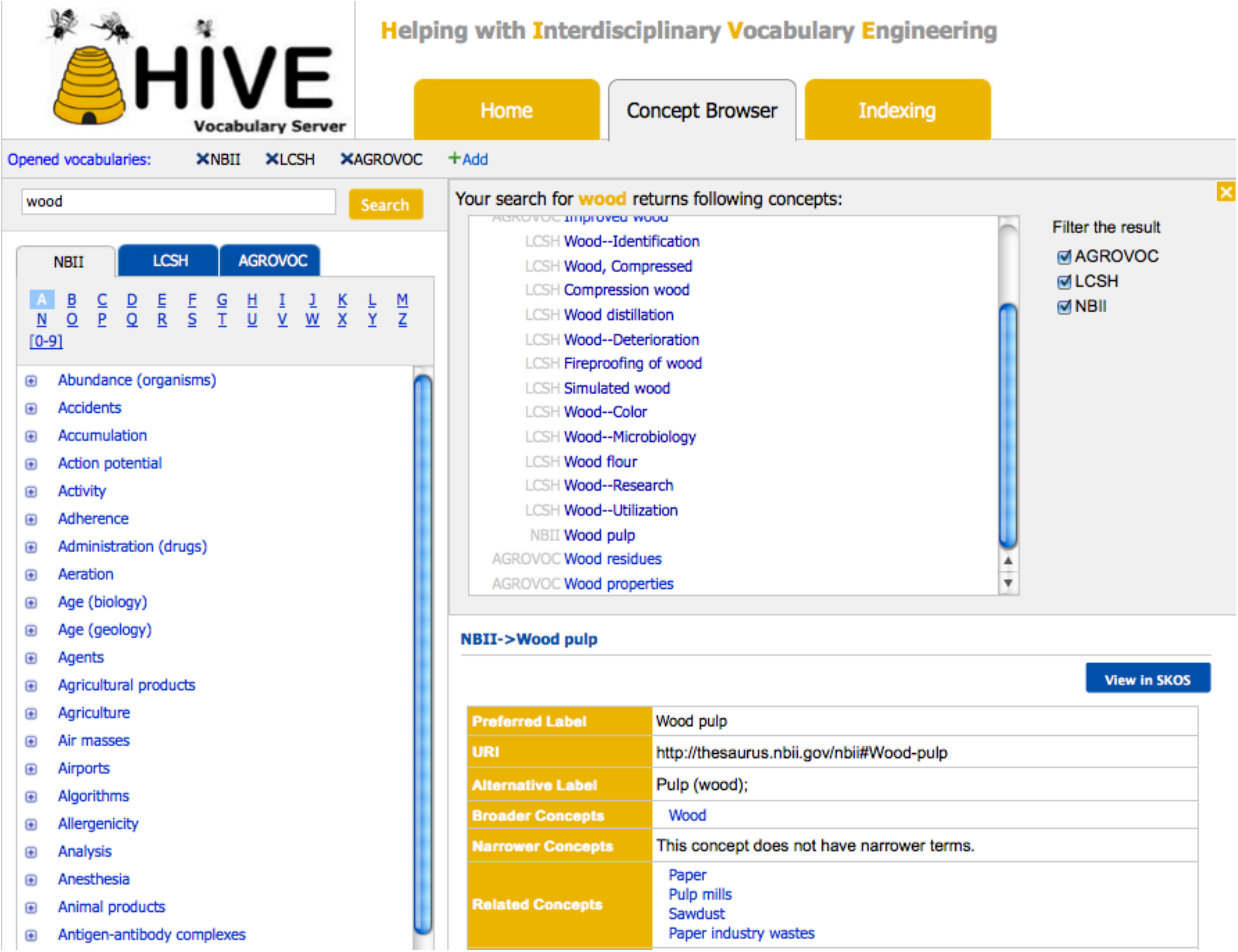
Index a Document

HIVE Indexing automatically extracts concepts from a given document to aid the cataloging and indexing practice.

Go to Indexing

Vocabulary Statistics

| Vocabulary | Concepts | Relationships | Date Added |
|---|---|---|---|
| AGROVOC | 28174 | 17834 | Oct 05,2009 |
| LCSH | 342684 | 147039 | Oct 05,2009 |
| NBII | 8680 | 11374 | Oct 23,2009 |



HIVE
Vocabulary Server
Helping with Interdisciplinary Vocabulary Engineering
Home
Concept Browser
Indexing
Opened vocabularies: ✕NBII ✕LCSH ✕AGROVOC +Add
wood
Search
NBII
LCSH
AGROVOC
A B C D E F G H I J K L M
N O P Q R S T U V W X Y Z
[0-9]

- Abundance (organisms)
- Accidents
- Accumulation
- Action potential
- Activity
- Adherence
- Administration (drugs)
- Aeration
- Age (biology)
- Age (geology)
- Agents
- Agricultural products
- Agriculture
- Air masses
- Airports
- Algorithms
- Allergenicity
- Analysis
- Anesthesia
- Animal products
- Antigen-antibody complexes

Your search for wood returns following concepts:

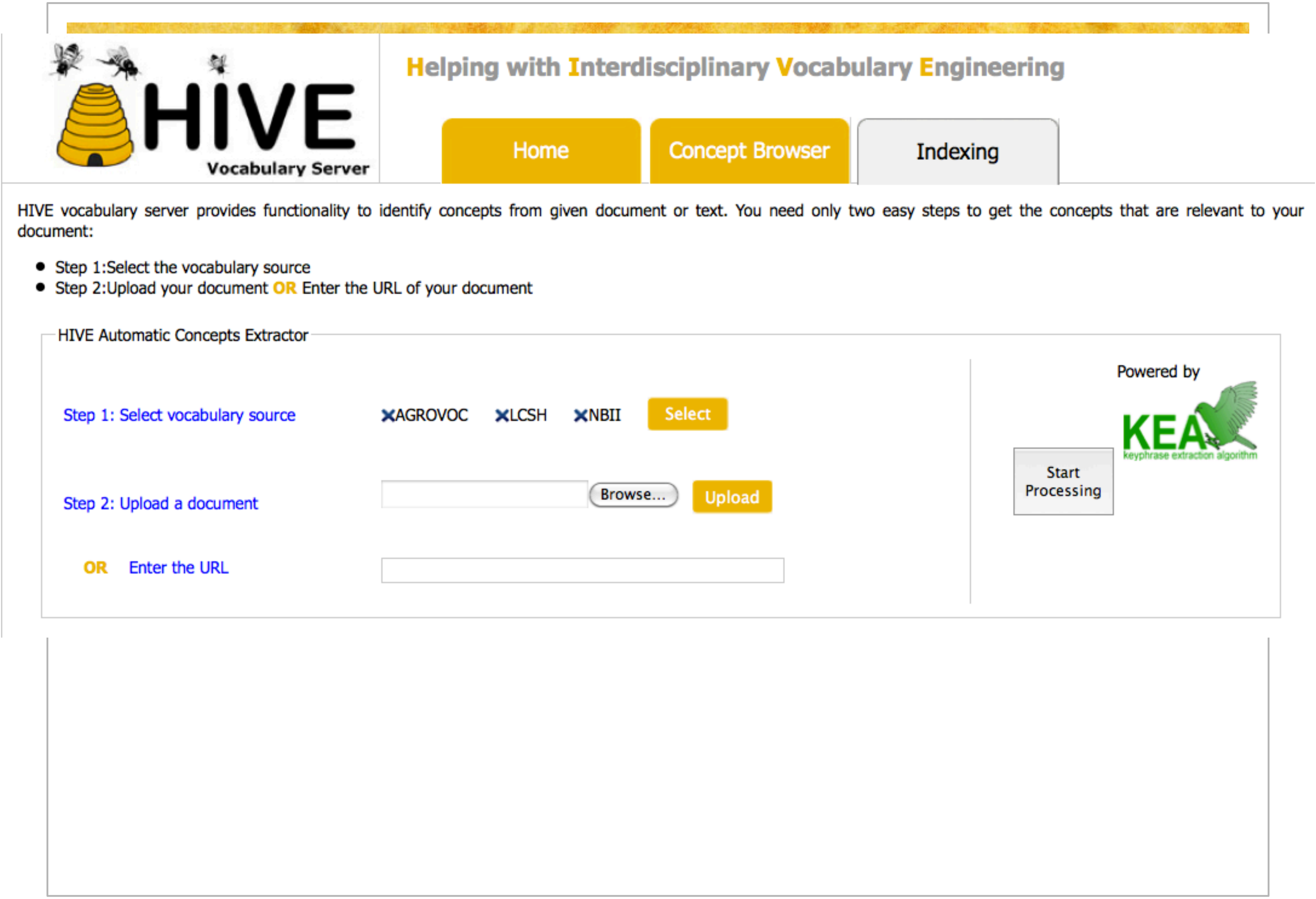
AGROVOC Improved wood
LCSH Wood--Identification
LCSH Wood, Compressed
LCSH Compression wood
LCSH Wood distillation
LCSH Wood--Deterioration
LCSH Fireproofing of wood
LCSH Simulated wood
LCSH Wood--Color
LCSH Wood--Microbiology
LCSH Wood flour
LCSH Wood--Research
LCSH Wood--Utilization
NBII Wood pulp
AGROVOC Wood residues
AGROVOC Wood properties

Filter the result

☑ AGROVOC
☑ LCSH
☑ NBII

**NBII->Wood pulp**

| | |
|---|---|
| Preferred Label | Wood pulp |
| URI | http://thesaurus.nbii.gov/nbii#Wood-pulp |
| Alternative Label | Pulp (wood); |
| Broader Concepts | Wood |
| Narrower Concepts | This concept does not have narrower terms. |
| Related Concepts | Paper<br>Pulp mills<br>Sawdust<br>Paper industry wastes |



HIVE vocabulary server provides functionality to identify concepts from given document or text. You need only two easy steps to get the concepts that are relevant to your document:

- Step 1:Select the vocabulary source
- Step 2:Upload your document **OR** Enter the URL of your document

HIVE Automatic Concepts Extractor

Step 1: Select vocabulary source ☒AGROVOC ☒LCSH ☒NBII

Step 2: Upload a document

**OR** Enter the URL

# HIVE Evaluation

Explanation:  Experimental solutions to evaluating the effectiveness of many vocabularies in one system

## Front-end:

usability

Ease of getting access to vocabularies

## Back-end:

performance

IR research in terms of relevancy (precision and recall)

# Front-end: Usability

LS and IS students (32 students)

- Understanding HIVE:  3.8 on 5 pt. scale
- Ease of navigation:  4.5
- Concept cloud a good idea:  3.3
- Indexing represent document accurately:
  2.0 (simple HIVE), 3.3 (*smart* HIVE)

Advisory board (10 members)

- Systems/technical folks want integration w/systems, Getty—EAD
- Librarians/KO folks, want to see term relationships
- Like tag cloud, want relevance percentages
- Color, placement of box, labels..

# Front-end: Usability

- Formal usability study 4 biologist, 5 information professionals

  ~ Tasks, usability ratings, satisfaction ranking

  - Average time to search a concept:

    Librarians: 6.53 minutes

    Scientists: 3.82 minutes

    ~ consistent w/research at NIEHS, 2 times as long

  - Average time for automatic indexing sequence

    Librarians: 1.91 minutes

    Scientists: 2.1 minutes

# Back-end: Performance

Vocabulary server comparison

- HIVE
  - SKOS
  - Some machine learning
- NCBO (National Center for Biomedical Ontologies) bioportal

  "ontologies that are actively used in biomedical communities" ( http://bioportal.bioontology.org/)

  194 ontologies; over one million terms

  - OWL/OBO
  - term matching

# Back-end: Performance

Method

- Experiment (quasi-experiment); content analysis
- 20 abstracts randomly selected from 10 Dryad partner journals
- 3 evaluators
  - 3 tier scale (good, fair, poor)
  - mean for each of 3 evaluators was averaged

**Specificity**

- NCBO Bioportal: 33.33% split evenly across
- Results inconsistent
- HIVE: good 10%, fair 53.33%, and poor 23.34%

**Exhaustivity**

- NCBO: good 48.33%, fair 36.67% , poor 15%
- HIVE: good 13.%, fair 51.67%, poor 35%

# NKOS

- Body of research literature in ILS for performance usability of information systems
  - Targeting NKOS is limited
- Scratch the surface w/NKOS
  - Practical system, needs of real people
- Much to learn from adapting and experimenting with these methods for NKOS

# Acknowledgments

- HIVE Development Team
- Dryad Repository Team
- Former SILS Masters students: Lina Huang and Jacquelynn Sherman

NSF National Science Foundation
WHERE DISCOVERIES BEGIN

INSTITUTE of Museum and Library SERVICES

DRYAD

UNC SCHOOL OF INFORMATION AND LIBRARY SCIENCE
Metadata Research Center <MRC>