DEMONSTRATING HIVE AND HIVE-ES: SUPPORTING TERM BROWSING AND AUTOMATIC TEXT INDEXING WITH LINKED OPEN VOCABULARIES

UC3M: David Rodríguez, Gema Bueno, Liliana Melgar, Nancy Gómez, Eva Méndez

UNC: Jane Greenberg, Craig Willis, Joan Boone

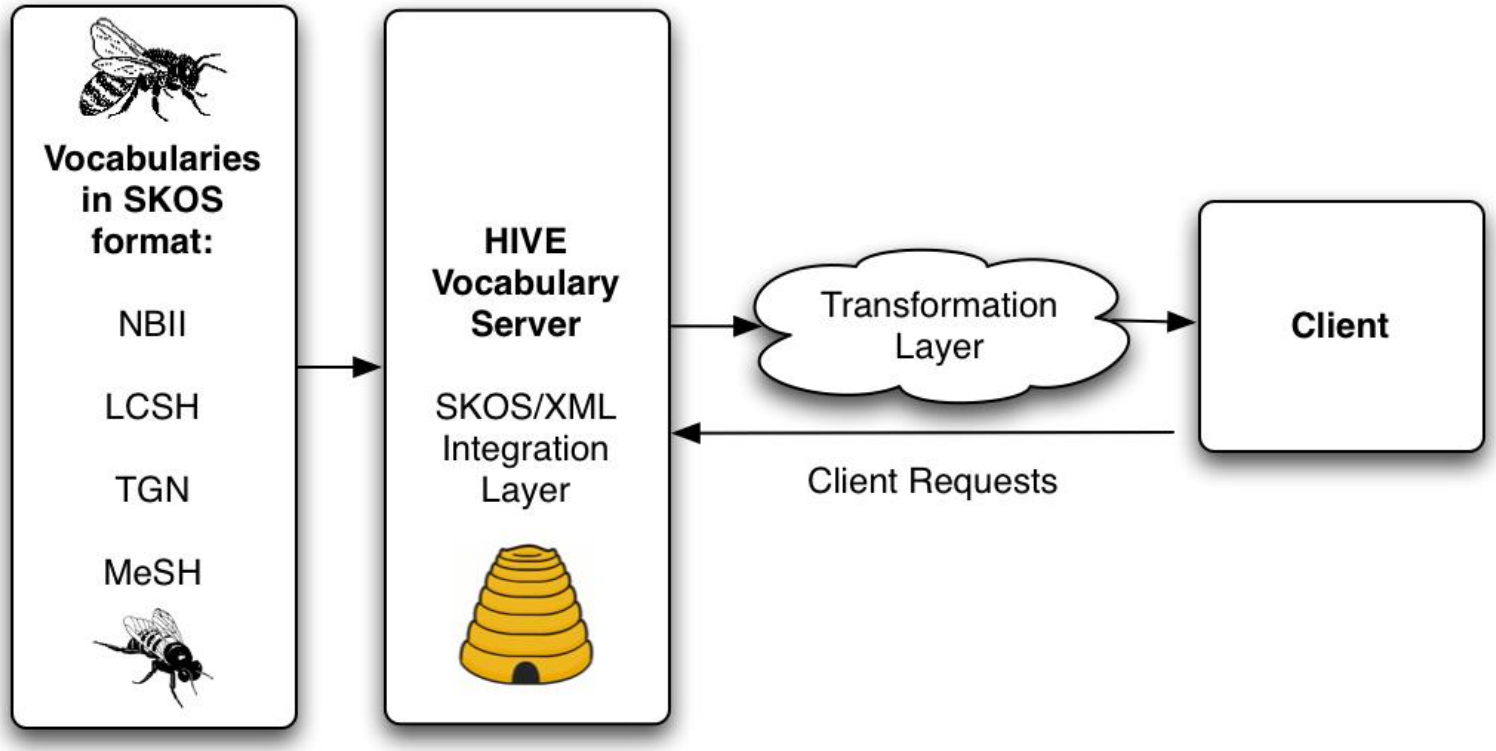
The 11th European Networked Knowledge Organization Systems (NKOS) Workshop

TPDL Conference 2012, Paphos, 27th September 2012

Contents

1. Introduction to HIVE and HIVE-ES
2. HIVE architecture: Technical Overview
3. Information retrieval in HIVE: KEA++/MAUI
4. HIVE in the real world: implementations, analysis/studies, challenges and future developments

What is HIVE?



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

What is HIVE?

HIVE Goals

- Provide efficient, affordable, interoperable, and user friendly **access to multiple vocabularies** during metadata creation activities
- Present a **model** and an **approach** that can be replicated
  —> not necessarily a service

Phases

1. Building HIVE

- Vocabulary preparation
- Server development

2. Sharing HIVE

- Continuing education

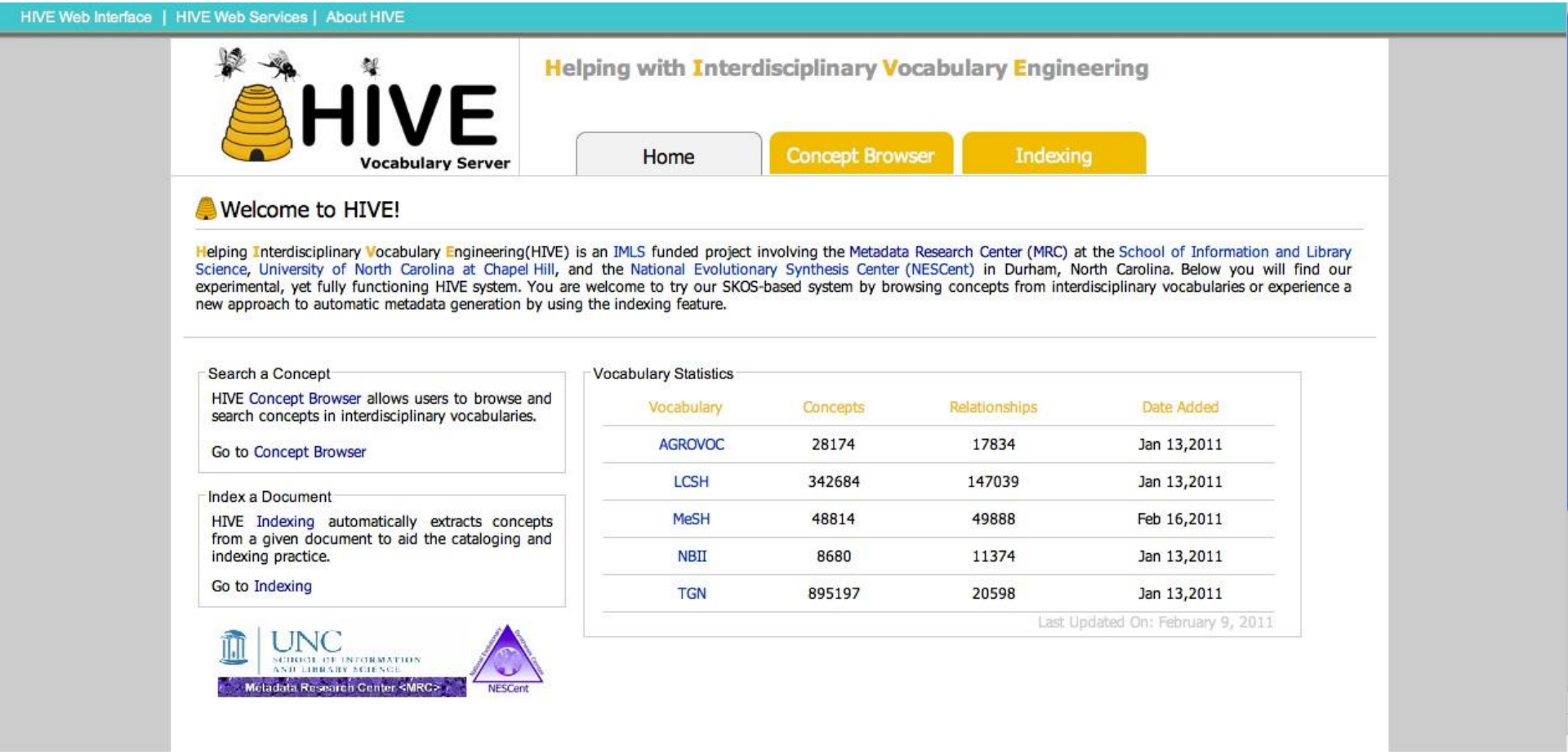
3. Evaluating HIVE

- Examining HIVE in Dryad reposit.
- Automatic indexing performance

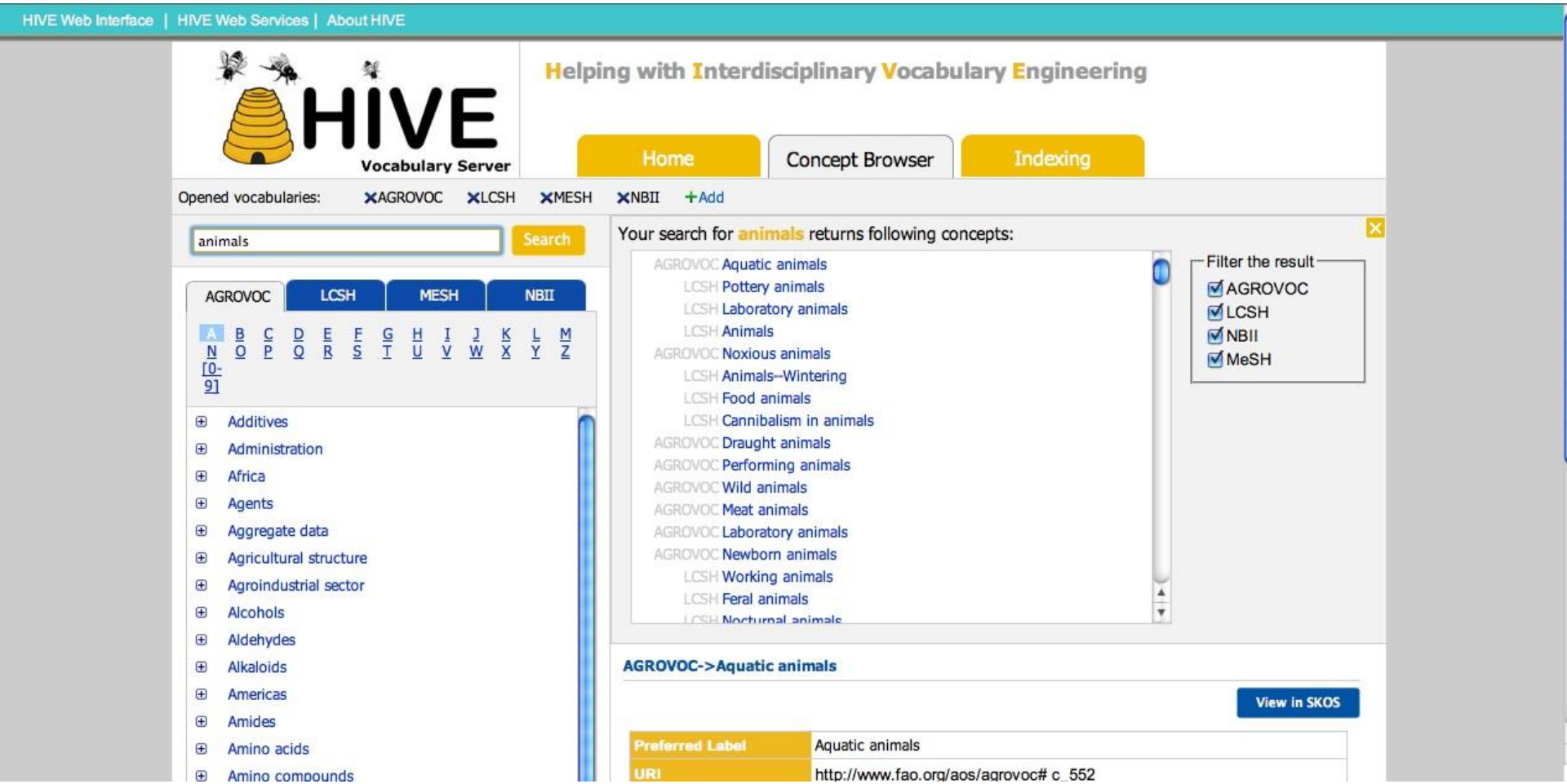
4. Expanding HIVE

HIVE-ES, HIVE-EU…

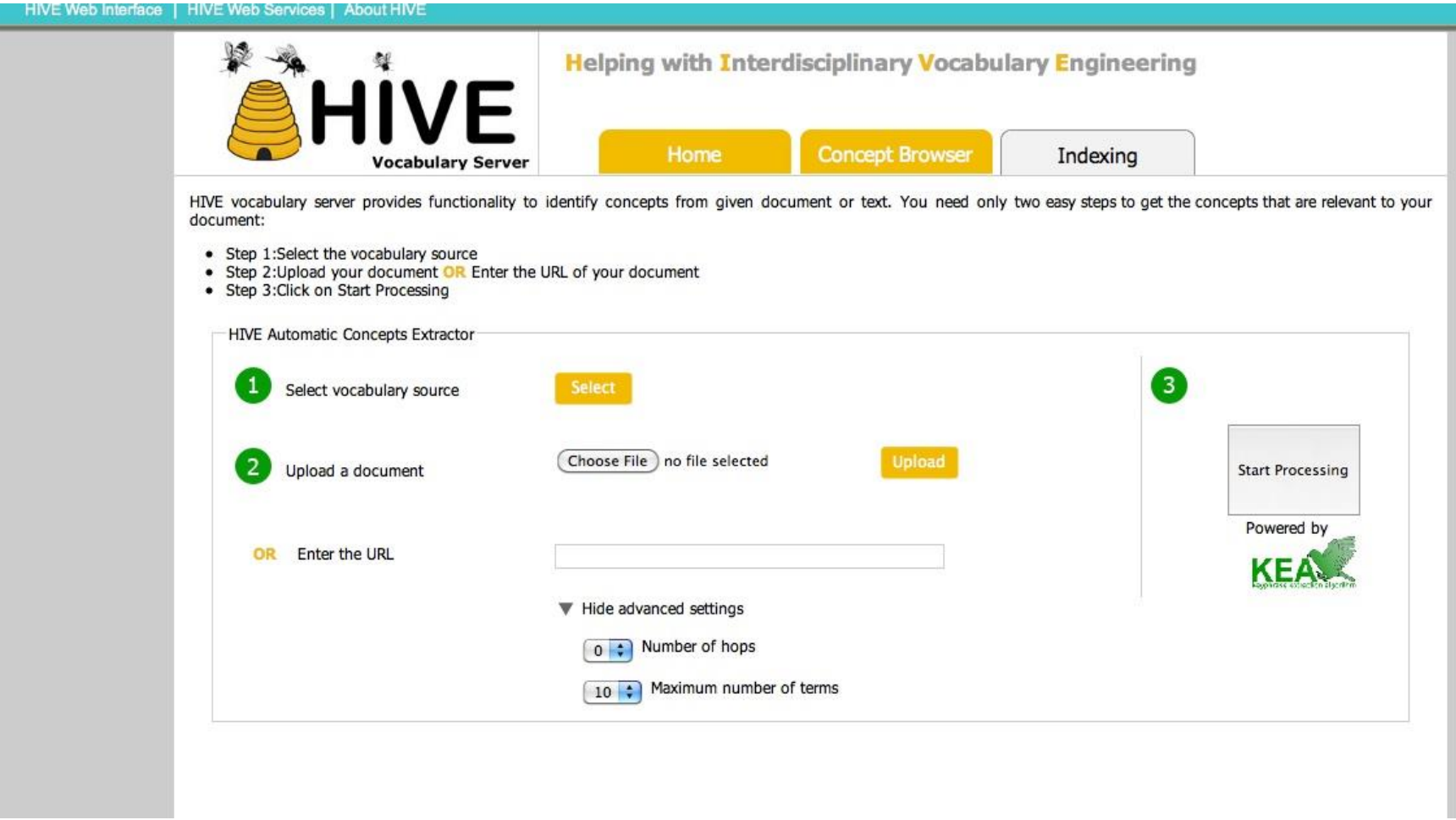
HIVE Demo Home Page



HIVE Demo Concept Browser



HIVE Demo Indexing



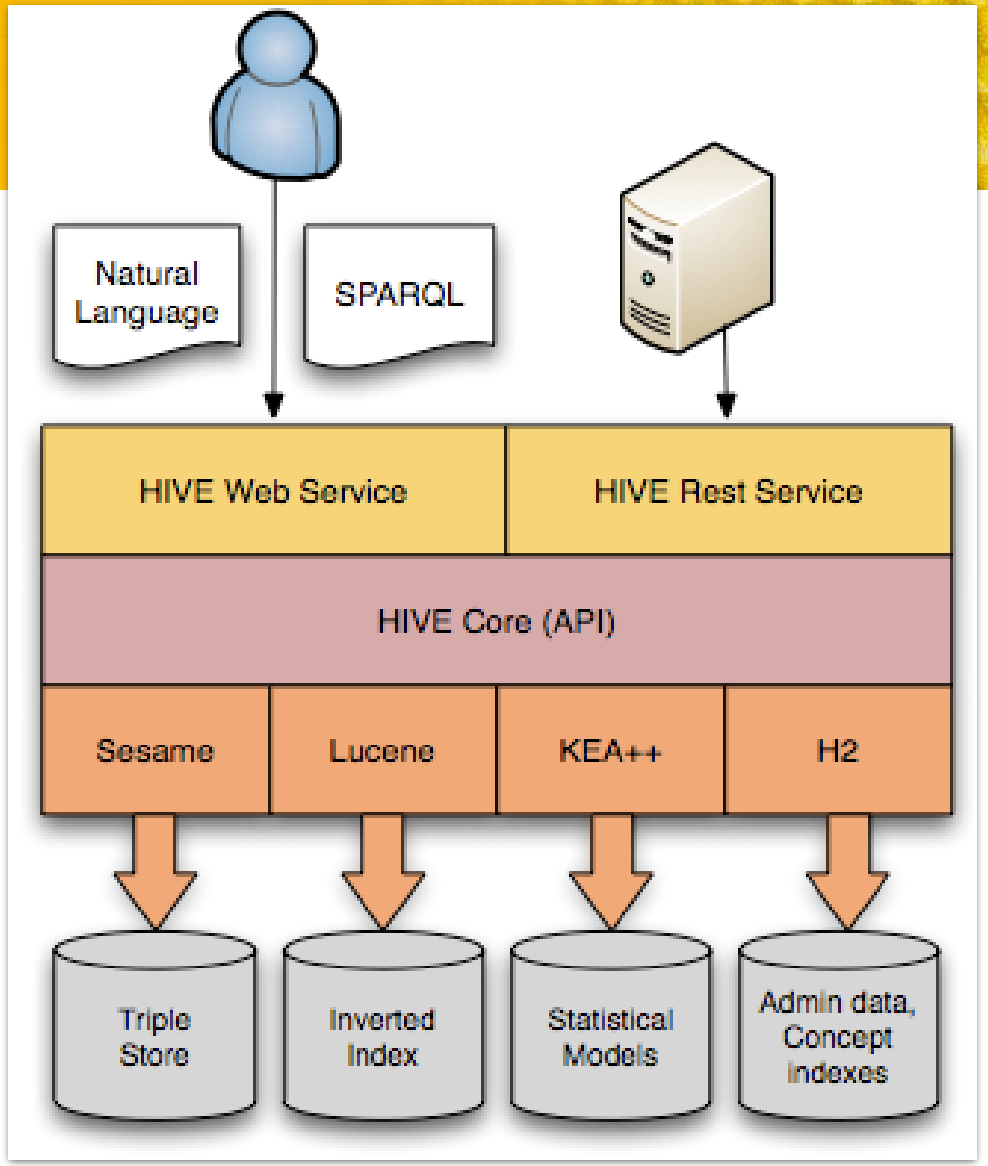
What is HIVE-ES

- **HIVE-ES** or HIVE-Español (Spanish), is an application of the HIVE project (Helping Interdisciplinary Vocabulary Engineering) for exploring and using methods and systems to publish widely used Spanish controlled vocabularies in SKOS.
- HIVE-ES chief vocabulary partner is the National Library of Spain (BNE): skosification of **EMBNE** (BNE Subject Headings)
- Establishing alliances for vocabularies skosification: BNCS (DeCS), CSIC IEDCYT (several thesauri).
- HIVE-ES wiki: http://klingon.uc3m.es/hive-es/wiki/
- HIVE-ES demo server: http://klingon.uc3m.es/hive-es
- HIVE-ES demo server at nescent: http://hive-test.nescent.org/

HIVE ARCHITECTURE: TECHNICAL OVERVIEW

HIVE Technical Overview

- HIVE combines several open-source technologies to provide a framework for vocabulary services.
- Java-based web services
- Open-source Google Code http://code.google.com/p/hive-mrc
- Source code, pre-compiled releases, documentation, mailing lists



HIVE Components

- **HIVE Core API**

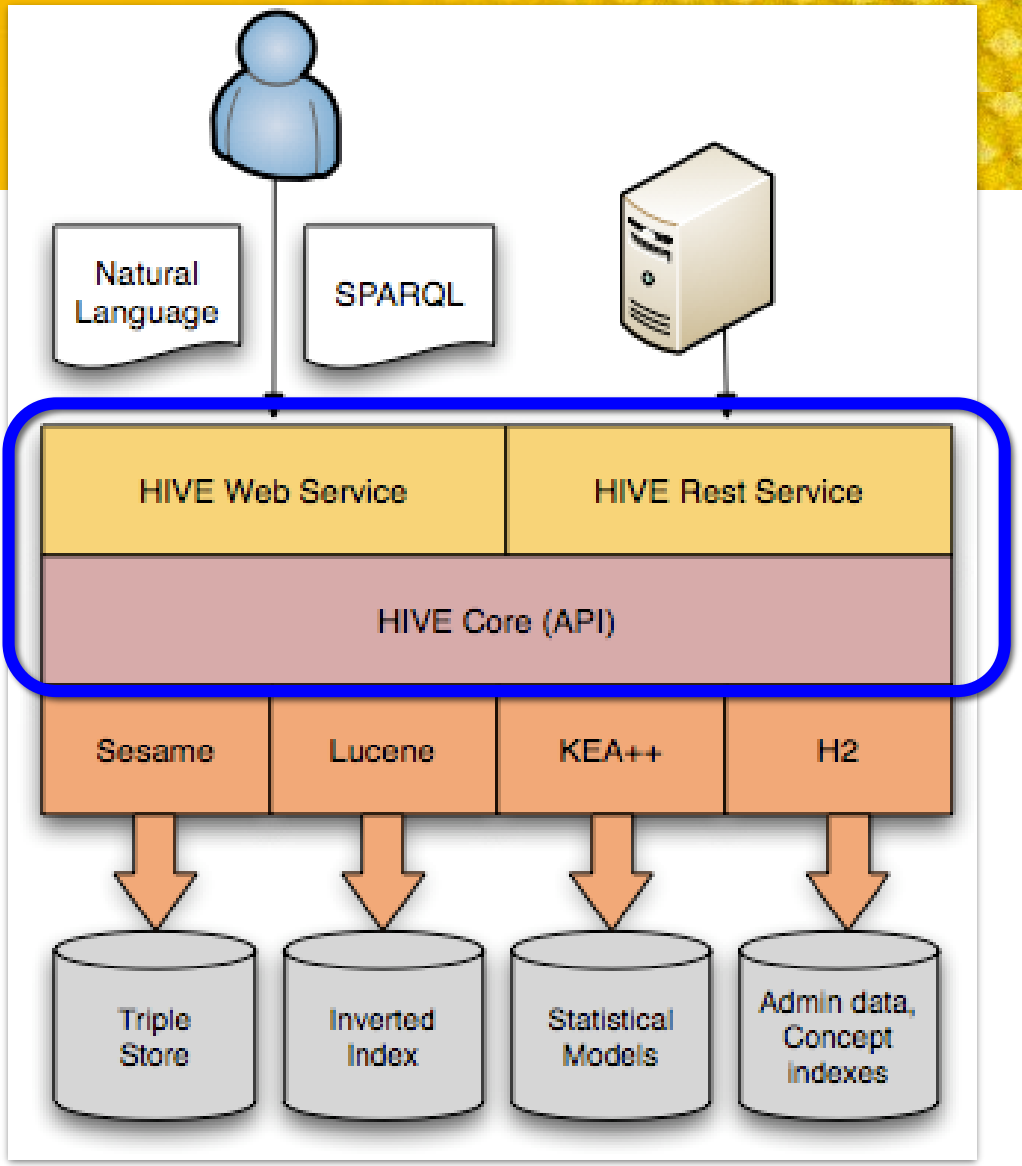
  Java API for vocabularies management

- **HIVE Web Service**

  Google Web Toolkit (GWT) based interface (Concept Browser and Indexer)

- **HIVE REST API**

  RESTful API



HIVE Supporting Technologies

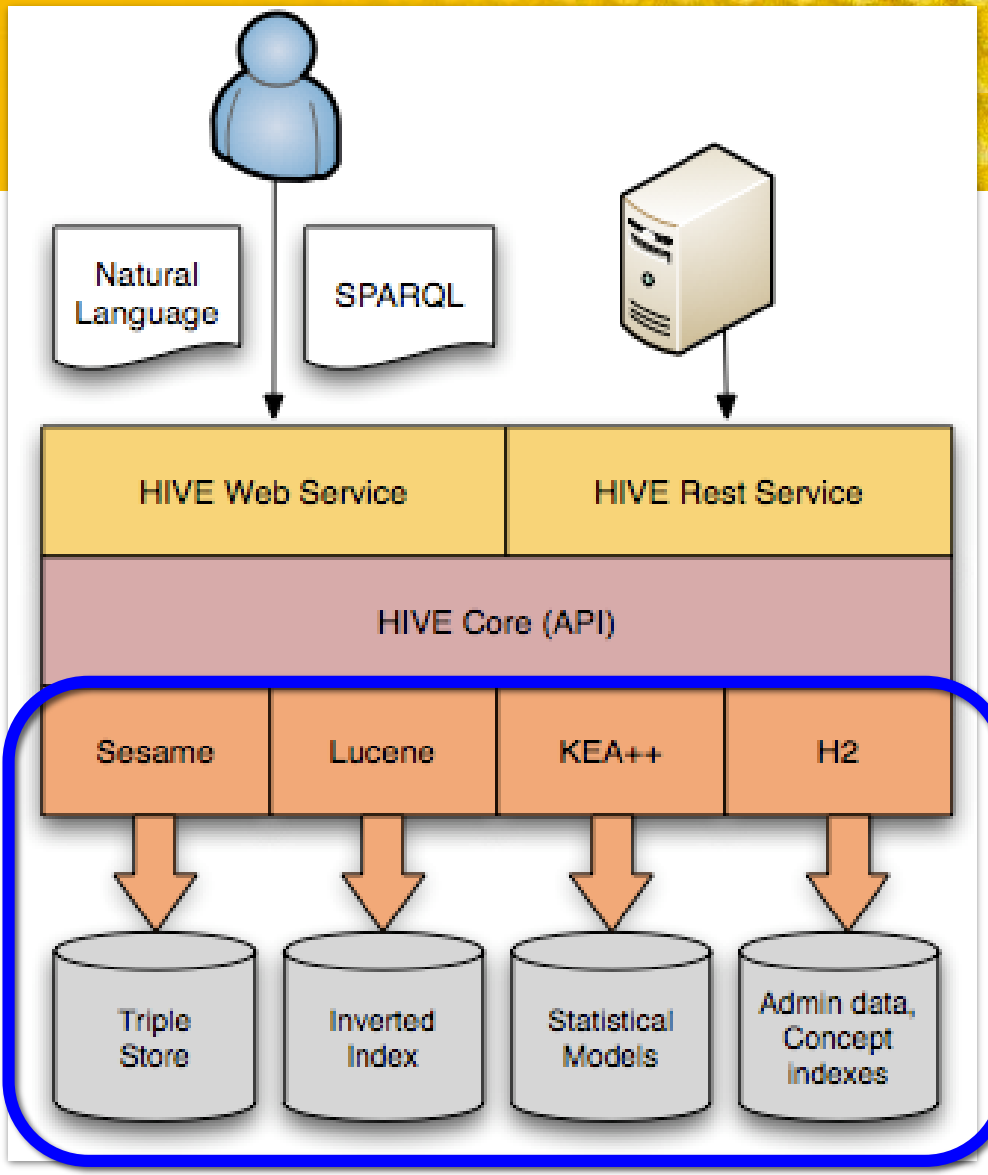
Sesame (OpenRDF): Open-source triple store and framework for storing and querying RDF data

Used for primary storage, structured queries

Lucene: Java-based full-text search engine

Used for keyword searching, autocomplete (version 2.0)

KEA++/Maui: Algorithms and Java API for automatic indexing



edu.unc.ils.hive.api

SKOSServer:

Provides access to one or more vocabularies

SKOSSearcher:

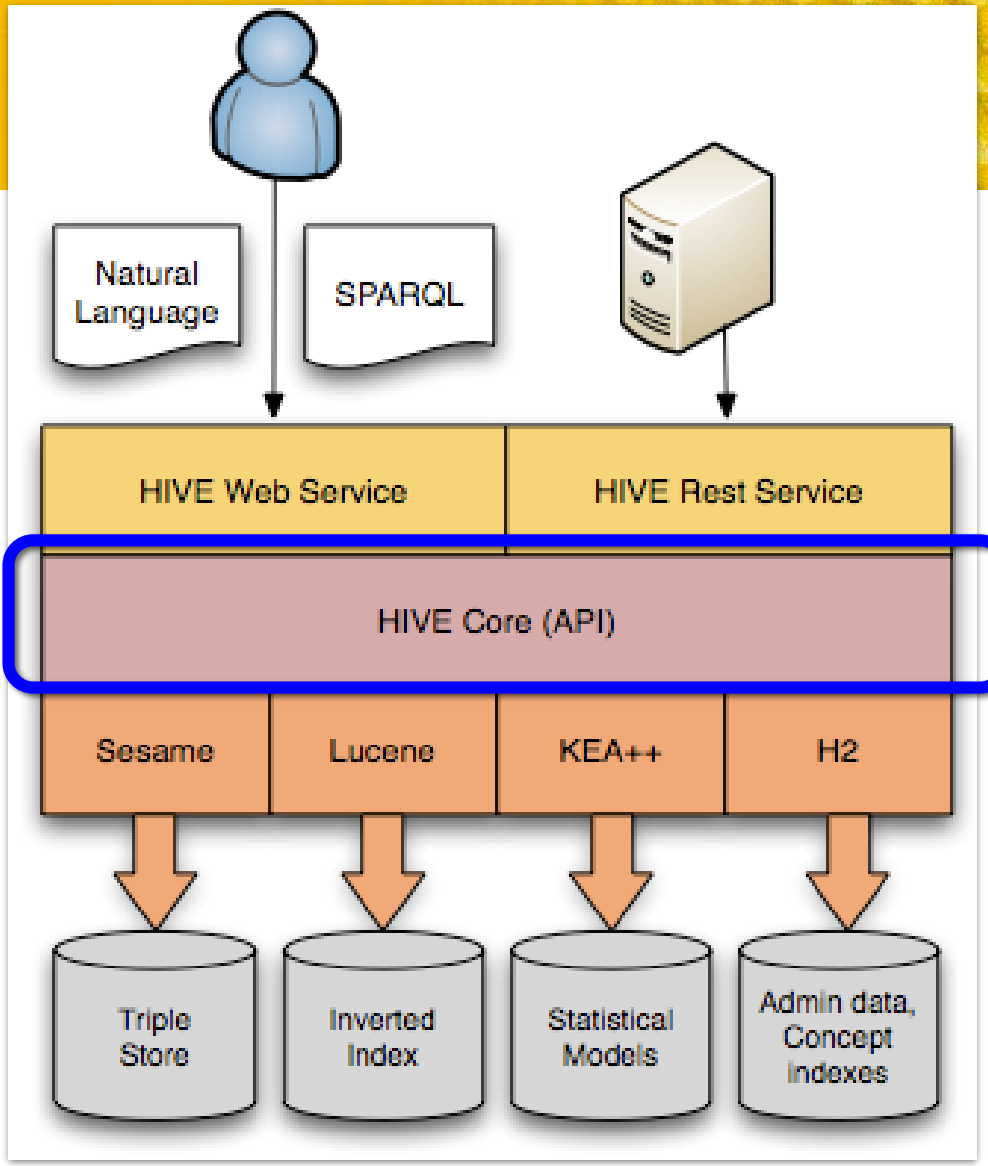
Supports searching across multiple vocabularies

SKOSTagger:

Supports tagging/keyphrase extraction across multiple vocabularies

SKOSScheme:

Represents an individual vocabulary (location of vocabulary on file system)



AUTOMATIC INDEXING IN HIVE

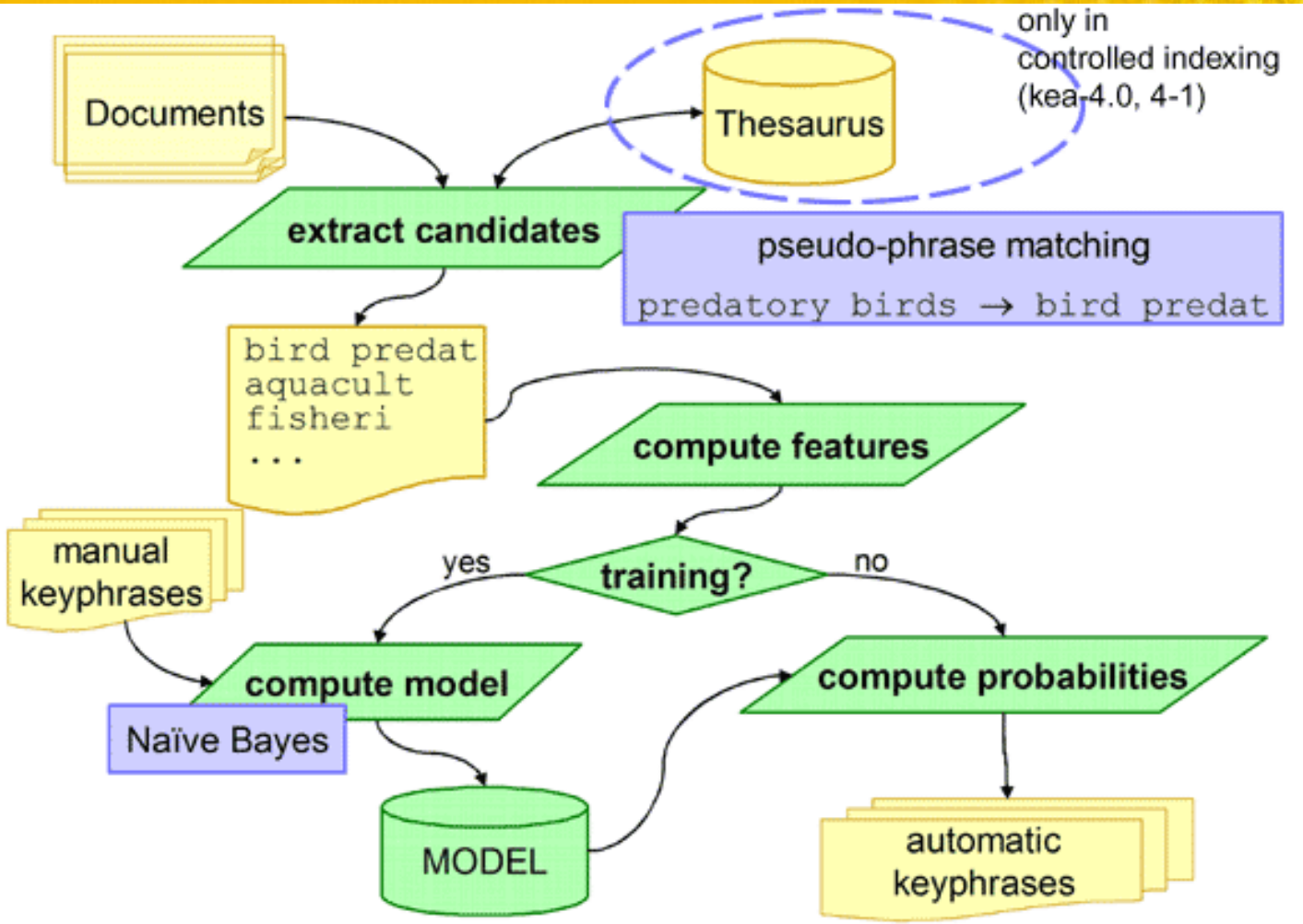
About KEA++ http://www.nzdl.org/Kea/

- Machine learning approach. http://code.google.com/p/hive-mrc/wiki/AboutKEA
- Domain-independent machine learning approach with minimal training set (~50 documents)....
- Leverages SKOS relationships and alternate/preferred labels
- Algorithm and open-source Java library for extracting keyphrases from documents using SKOS vocabularies.
- Developed by Alyona Medelyan (KEA++), based on earlier work by Ian Witten (KEA) from the Digital Libraries and Machine Learning Lab at the University of Waikato, New Zealand.

KEA
keyphrase extraction algorithm

Medelyan, O. and Whitten I.A. (2008). "Domain independent automatic keyphrase indexing with small training sets." *Journal of the American Society for Information Science and Technology*, (59) 7: 1026-1040).

KEA Model


Documents
Thesaurus
only in
controlled indexing
(kea-4.0, 4-1)
extract candidates
pseudo-phrase matching
predatory birds → bird predat
bird predat
aquacult
fisheri
...
compute features
manual
keyphrases
yes
training?
no
compute model
Naïve Bayes
compute probabilities
MODEL
automatic
keyphrases

KEA++ at a Glance

- Machine learning approach to keyphrase extraction
- Two stages:
  - **Candidate identification**: find terms that relate to the document's content
    - Parse the text into tokens based on whitespace and punctuation
    - Create word n-grams based on longest term in CV
    - Remove all stopwords from the n-gram
    - Stem to grammatical root (Porter) (aka "pseudophrase")
    - Stem terms in vocabulary (Porter)
    - Replace non-descriptors with descriptors using CV relationships
    - Match stemmed n-grams to vocabulary terms
  - **Keyphrase selection**: uses a model to identify the most significant terms

KEA++ candidate identification

- Stemming is not perfect…

Original	Stemmed
"information organization"	"inform organ"
"organizing information"	"inform organ"
"informative organizations"	"inform organ"
"informal organization"	"inform organ"

KEA++: Feature definition

- **Term Frequency/Inverse Document Frequency:** Frequency of a phrase's occurrence in a document with frequency in general use.
- **Position of first occurrence:** Distance from the beginning of the document. Candidates with high/low values are more likely to be valid (introduction/conclusion)
- **Phrase length:** Analysis suggests that indexers prefer to assign two-word descriptors
- **Node degree:** Number of relationships between the term in the CV.

MAUI http://maui-indexer.googlecode.com

- **Maui, an algorithm for topic indexing**, which can be used for the same tasks as Kea, but offers additional features.
- MAUI features:
  - term assignment with a controlled vocabulary (or thesaurus)
  - subject indexing
  - topic indexing with terms from Wikipedia
  - keyphrase extraction
  - terminology extraction
  - automatic tagging

MAUI Feature definition

- **Frequency statistics**, such as term frequency, inverse document frequency, TFxIDF;
- **Occurrence positions** in the document text, e.g. beginning and end, spread of occurrences;
- **Keyphraseness**, computed based on topics assigned previously in the training data, or particular behaviour of terms in Wikipedia corpus;
- **Semantic relatedness**, computed using semantic relations encoded in provided thesauri, if applicable, or using statistics from the Wikipedia corpus;

Software inside MAUI

- **Kea** (Major parts of Kea became parts of Maui without modifications. Other parts, extended with new elements)
- **Weka** machine learning toolkit for creating the topic indexing model from documents with topics assigned by people and applying it to new documents. (Kea only containes a cut-down version of Weka (several classes), Maui includes the complete library.)
- **Jena library** for topic indexing with many kinds of controlled vocabularies. It reads RDF-formatted thesauri (specifically SKOS) and stores them in memory for a quick access.
- **Wikipedia Miner** for accessing Wikipedia data
  - Converts regular Wikipedia dumps into MySql database format and provides an object-oriented access to parts of Wikipedia like articles, disambiguation pages and hyperlinks.
  - Algorithm for computing semantic relatedness between articles, to disambiguate documents to Wikipedia articles and for computing semantic features.

HIVE IN THE REAL WORLD

Who's using HIVE?

HIVE is being evaluated by several institutions and organizations:

- **Long Term Ecological Research Network (LTER)**
  - Prototype for keyword suggestion for Ecological Markup Language (EML) documents.
- **Library of Congress Web Archives (Minerva)**
  - Evaluating HIVE for automatic LCSH subject heading suggestion for web archives.
- **Dryad Data Repository**
  - Evaluating HIVE for suggestion of controlled terms during the submission and curation process. (Scientific name, spatial coverage, temporal coverage, keywords).
  - Scientific names (IT IS), Spacial coverage (TGN, Alexandria Gazetteer), Keywords (NBII, MeSH, LCSH). http://www.datadryad.org
- Yale University, Smithsonian Institution Archives

Ecology Letters, (2009) **12**: 351–366 doi: 10.1111/j.1461-0248.2009.01285.x

REVIEW AND SYNTHESIS

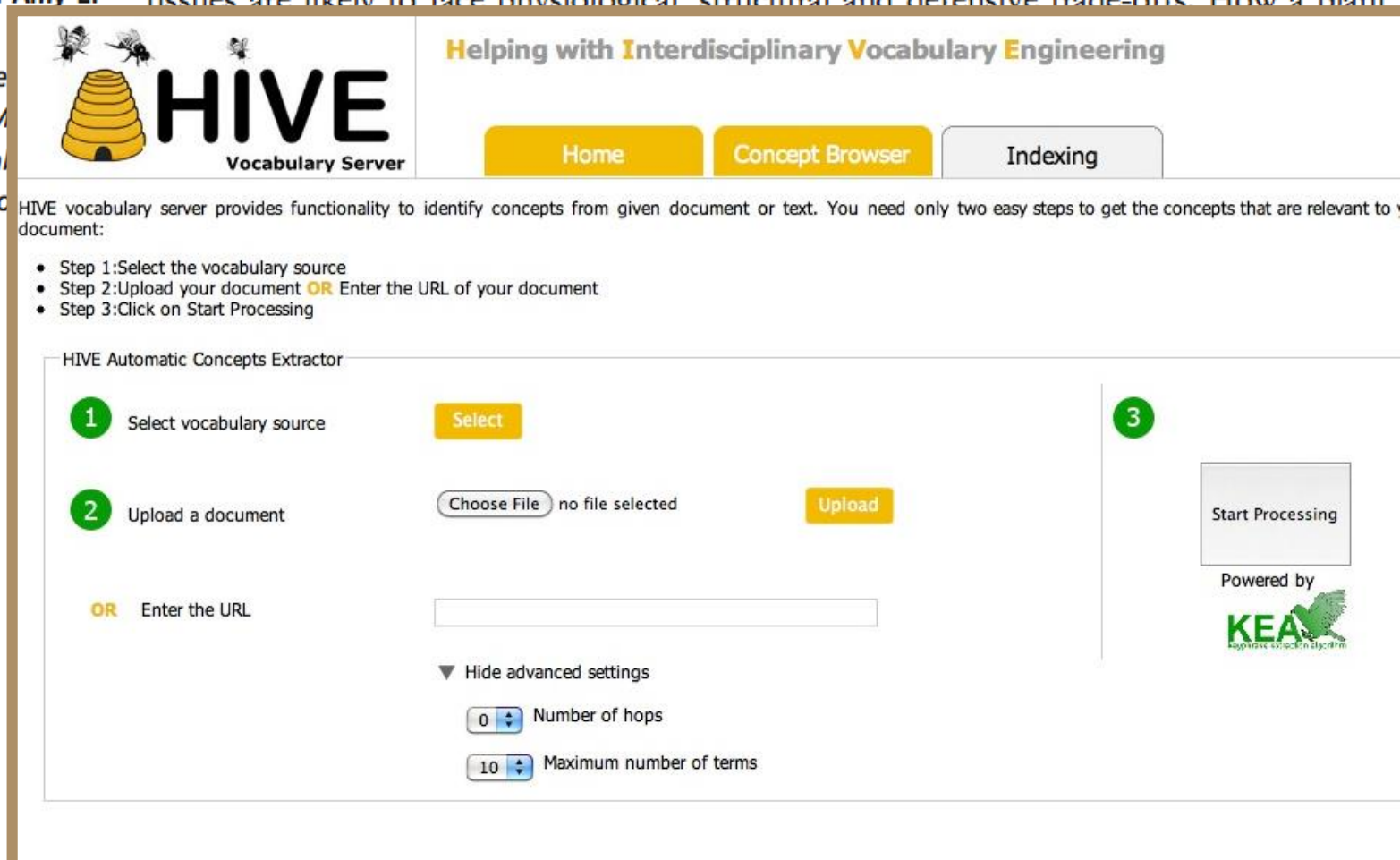
Towards a worldwide wood economics spectrum

Jerome Chave,[1]* David Coomes,[2] Steven Jansen,[3] Simon L. Lewis,[4] Nathan G. Swenson[5] and Amy E. Zanne[6,7]

[1]*Laboratoire Evolution e*
Diversité Biologique, UM
CNRS/Université Paul Sa
Bâtiment 4R3 F-31062 To
France

Abstract

Wood performs several essential functions in plants, including mechanically supporting aboveground tissue, storing water and other resources, and transporting sap. Woody tissues are likely to face physiological, structural and defensive trade-offs. How a plant



Ecology Letters, (2009) **12**: 351–366 doi: 10.1111/j.1461-0248.2009.01285.x

REVIEW AND SYNTHESIS

Towards a worldwide wood economics spectrum

Jerome Chave,[1]* David Coomes,[2] Steven Jansen,[3] Simon L. Lewis,[4] Nathan G. Swenson[5] and Amy E. Zanne[6,7]

[1]Laboratoire Evolution et Diversité Biologique, UMR 5174, CNRS/Université Paul Sabatier Bâtiment 4R3 F-31062 Toulouse, France

Abstract

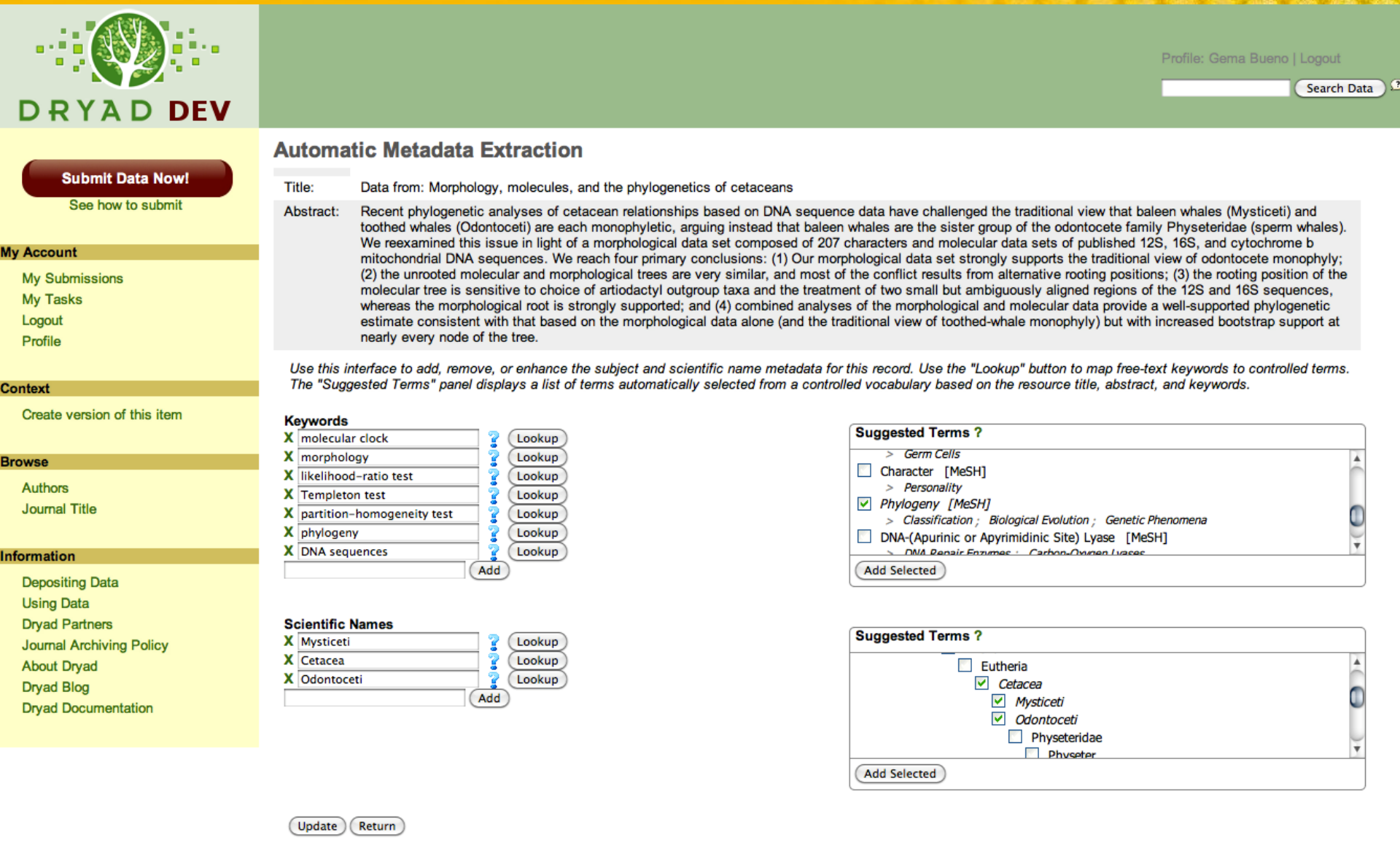
Wood performs several essential functions in plants, including mechanically supporting aboveground tissue, storing water and other resources, and transporting sap. Woody tissues are likely to face physiological, structural and defensive trade-offs. How a plant optimizes among these competing functions can have major ecological implications, which have been under-appreciated by ecologists compared to the focus they have given to leaf function. To draw together our current understanding of wood function, we identify and collate data on the major wood functional traits, including the largest wood density database to date (8412 taxa), mechanical strength measures and anatomical

Extracted Concepts Cloud

AGROVOC
LCSH
NBII

Reaction wood Wood--Figure Wood--Discoloration Calavicci, Al (Fictitious character) Lāt, al- (Arabian deity) Murphy, Al (Fictitious character) Density Soils--Density Population density Recessive traits Traits (genetics) Dominant traits Associated species Species diversity Numbers of species Plant anatomy Plant litter Plant condition Leaf spots Leaf prints Leaf blowers Brushes, Carbon Electrodes, Carbon Carbon taxes Growth Fetus--Growth Growth (Plants) Infiltration water Water--Color Drinking water

Automatic metadata extraction in Dryad



Automatic Indexing with HIVE: pilot studies

- Different types of studies:
  - Usability studies (Huang 2010).
  - Comparison of performance with indexing systems (Sherman, 2010)
  - Improving Consistency via Automatic Indexing (White, Willis and Greenberg 2012)
  - Systematic analysis of HIVE indexing performance (HIVE-ES Project Members)

Usability tests

(Huang 2010)

- **Search A Concept:**
  - Average time: librarians 4.66 m., scientists, 3.55 m.
  - Average errors: librarians 1.5; scientists 1.75.
- **Automatic indexing:**
  - Average time: librarians 1.96 m., scientists 2.,1 m.
  - Average errors: librarians 0.83; scientists 1.00.
- **Safisfaction rating:**
  - SUS (System Usability Scale): librarians 74.5; scientists 79.38.
- **Enjoyment and concentration** (Ghani's Flow metrics)
  - Enjoyment: librarians 17, scientists 15.25.
  - Concentration: librarians 15.83, scientists 16.75.

Automatic metadata generation: comparison of annotators (HIVE / NCBO BioPortal)

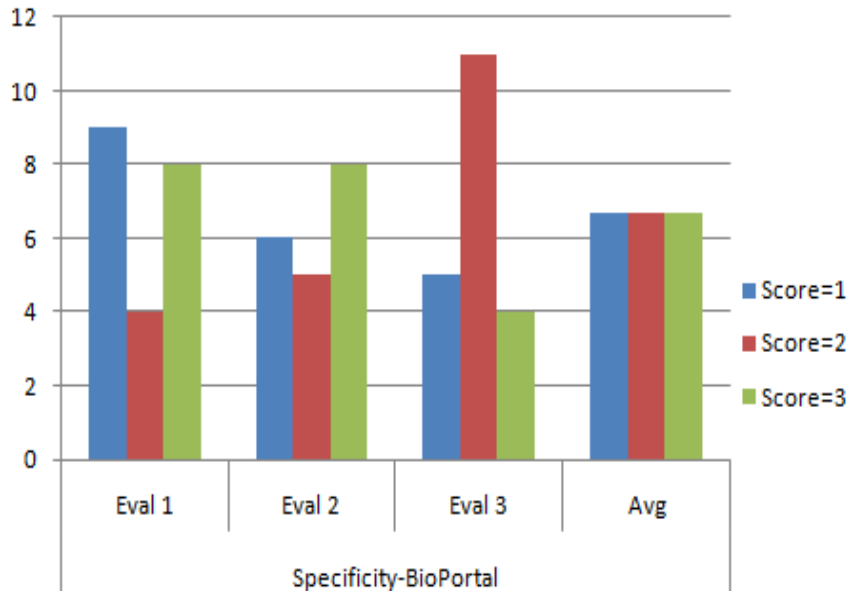
(Sherman 2010)

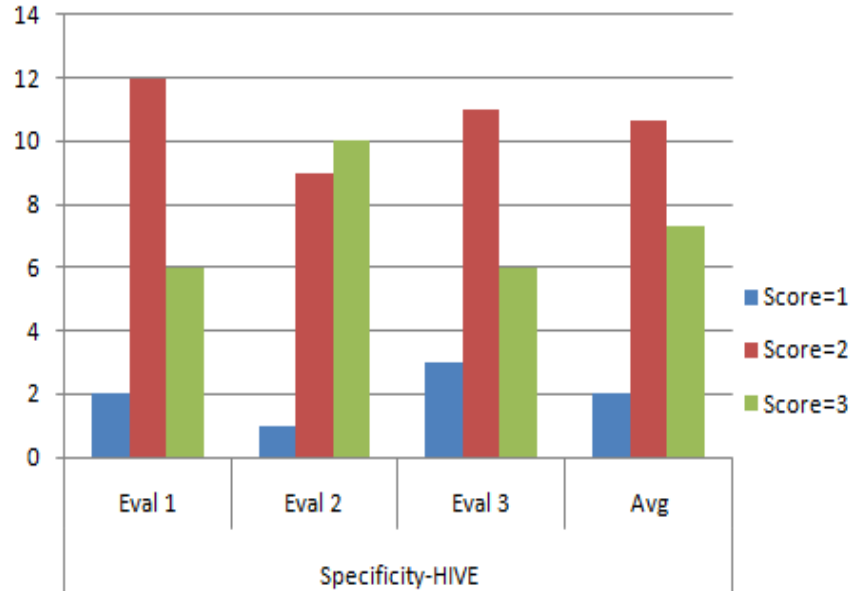
- BioPortal: term matching. Vs. HIVE: machine learning.
- Document set: Dryad repository article abstracts (random selection): 12 journals, 2 articles journal = 24
- Results: HIVE annotator:
  - 10 percent higher specificity.
  - 17 percent higher exhaustivity.
  - 19.4 percent higher precision.

Automatic metadata generation: comparison of annotators (HIVE / NCBO BioPortal)

(Sherman 2010)

Specificity



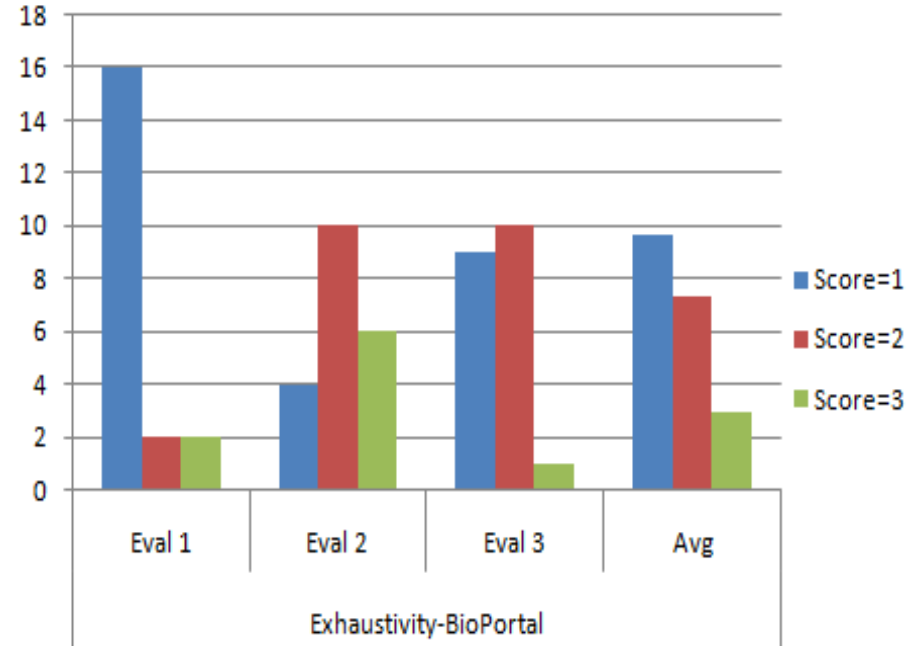


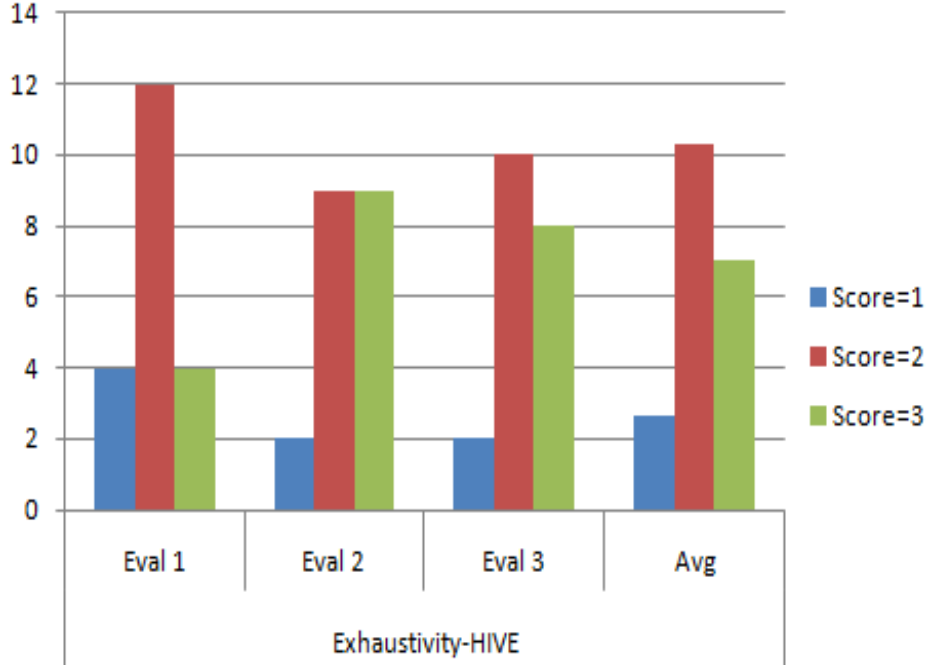
Figures 2 &3. Specificity (by evaluator)

Automatic metadata generation: comparison of annotators (HIVE / NCBO BioPortal)

(Sherman 2010)

Exhaustivity





Figures 4 & 5. Exhaustivity (by evaluator)

Improving Consistency via Automatic Indexing

(White, Willis & Greenberg 2012)

- **Aim**: Comparison indexing with and without HIVE aids.
- **Document set**: Scientific abstracts.
- **Vocabularies**: LCSH, NBII, TGN
- **Participants**: 31 (librarians, technologists, programmers, and library consultants.)

Table 1. Average inter-indexer consistency within-subjects with and without an automatic indexing aid

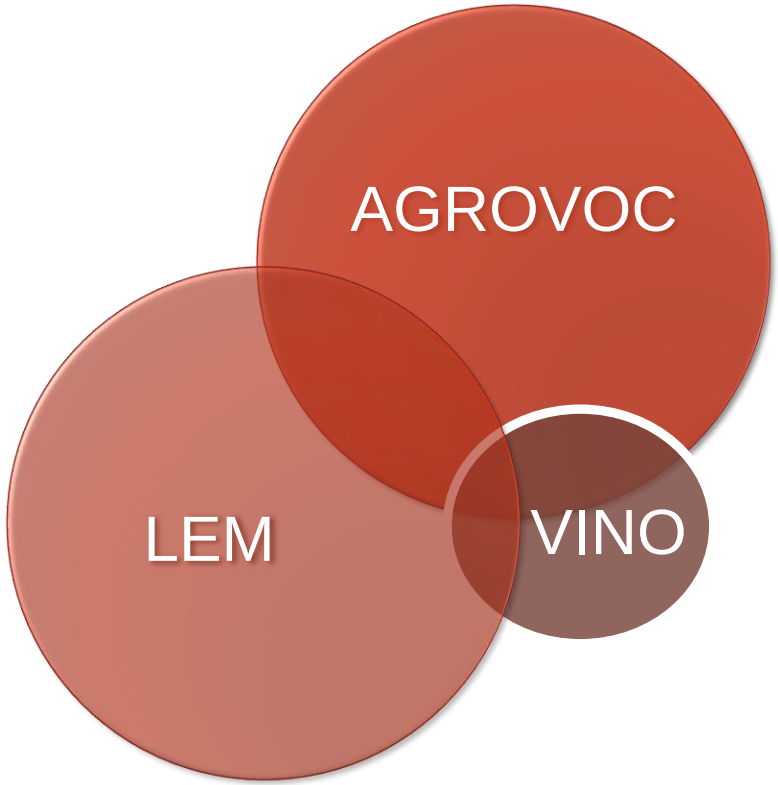
Task	Inter-indexer consistency	
	R (Mean)	H (Mean)
Free-text keywords	28.64%	18.29%
HIVE - Relevant	54.10%	24.61%
HIVE - Not Relevant	35.81%	24.61%

Systematic analysis of HIVE indexing performance: Initial research questions

- What is the best algorithm for automatic term suggestion for Spanish vocabularies, KEA or Maui?
- Do different algorithms perform better for a particular vocabulary?
- Does the number of extracted concepts represent significant differences of precision?
- Does the minimum number of term occurrence determines the results?
- Are the term weights assigned by HIVE consistent with the human assessment?

Systematic analysis of HIVE indexing performance: Pilot study

- **Vocabularies**: LEM (Spanish Public Libraries Subject Headings); VINO (own-developed thesaurus about wine); AGROVOC.
- **Document set**: Articles on enology, both in Spanish and English.



Systematic analysis of HIVE indexing performance: Pilot study

- **Variables**:
  1. Vocabulary: LEM, AGROVOC, VINO.
  2. Document language: ENG / SPA.
  3. Algorithm: KEA, MAUI.
  4. Nº of minimum ocurrences: 1, 2.
  5. Number of indexing terms. 5, 10, 15, 20.

  16 tests per document/vocabulary

  - Other parameters and variables for next experiments:
    - Document type, format and length (nº of words).
    - Number of training documents per vocabulary.
- **Data**: concept probability/ Relevance N/Y / Precision (1-4).
- **Participants**: project members / indexing experts.

HIVE-ES
Servidor de vocabularios

Helping with Interdisciplinary Vocabulary Engineering

Home | Concept Browser | Indexing

Welcome to HIVE!

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. Below you will find our experimental, yet fully functioning HIVE system. You are welcome to try our SKOS-based system by browsing concepts from interdisciplinary vocabularies or experience a new approach to automatic metadata generation by using the indexing feature.

Search a Concept

Browse and search concepts in selected vocabularies.

Index a Document

Automatically extract document concepts for subject metadata creation.

This HIVE system is for demo purposes and may change in response to your feedback. Contact us

UNC School of Information and Library Science
Metadata Research Center <MRC>
NESCent

Vocabulary Statistics

Vocabulary	Concepts	Relationships	Last Updated
AGROVOC	28174	83086	jun 12, 2011
embne	30992	68497	jul 6, 2012
embne2	351901	80163	jul 27, 2012
lem	17323	35980	jul 6, 2012
lem2	17323	35980	sep 19, 2012
vino	920	2612	sep 23, 2012
vino3	460	1306	sep 23, 2012

TermDoc Tecnologías aplicadas a la Información y la Documentación

HIVE-ES
Servidor de vocabularios

Helping with Interdisciplinary Vocabulary Engineering

Home | Concept Browser | Indexing

HIVE automatically extracts concepts from a document or URL based on selected vocabularies.

- Step 1: Select a vocabulary
- Step 2: Upload a document OR provide the URL for a document
- Step 3: Click Start Processing button

HIVE Automatic Concepts Extractor

1 Select vocabulary source — Select

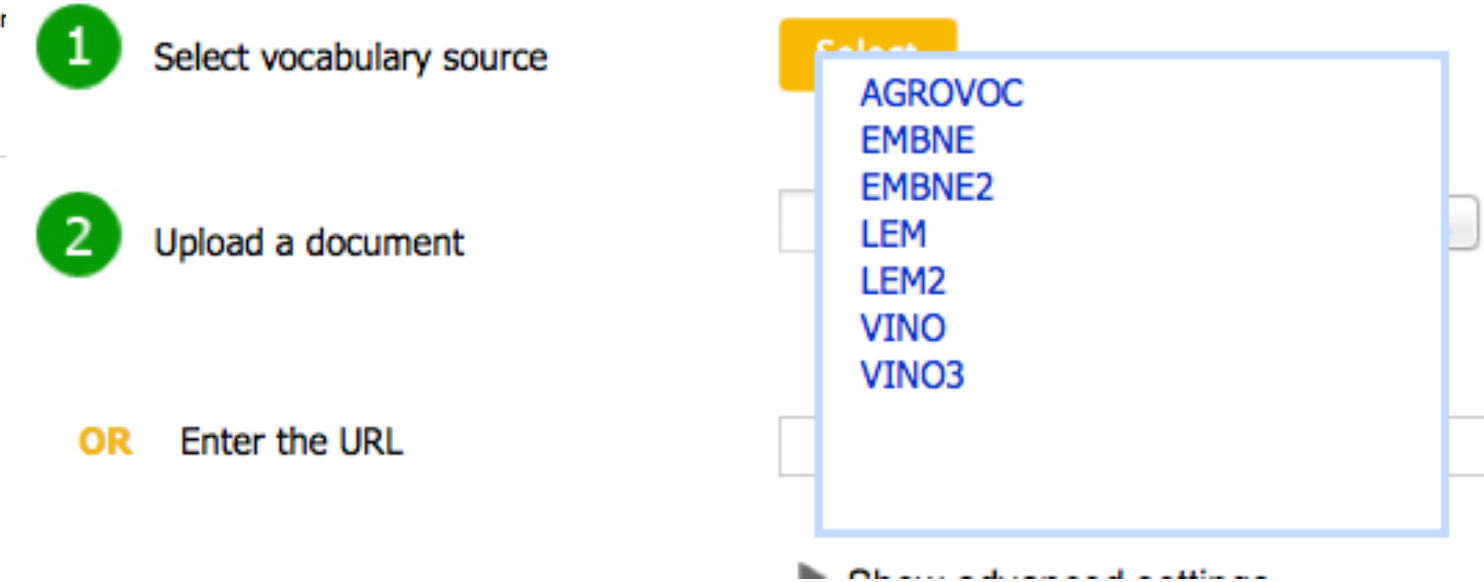
2 Upload a document — Examinar... — Upload

OR Enter the URL

3

Start Processing

d by



Helping with Interdisciplinary Vocabulary Engineering

Home | Concept Browser | Indexing

HIVE automatically extracts concepts from a document or URL based on selected vocabularies.

- Step 1: Select a vocabulary
- Step 2: Upload a document OR provide the URL for a document
- Step 3: Click Start Processing button

HIVE Automatic Concepts Extractor

1 Select vocabulary source ✖AGROVOC Select

2 Upload a document Examinar... Upload

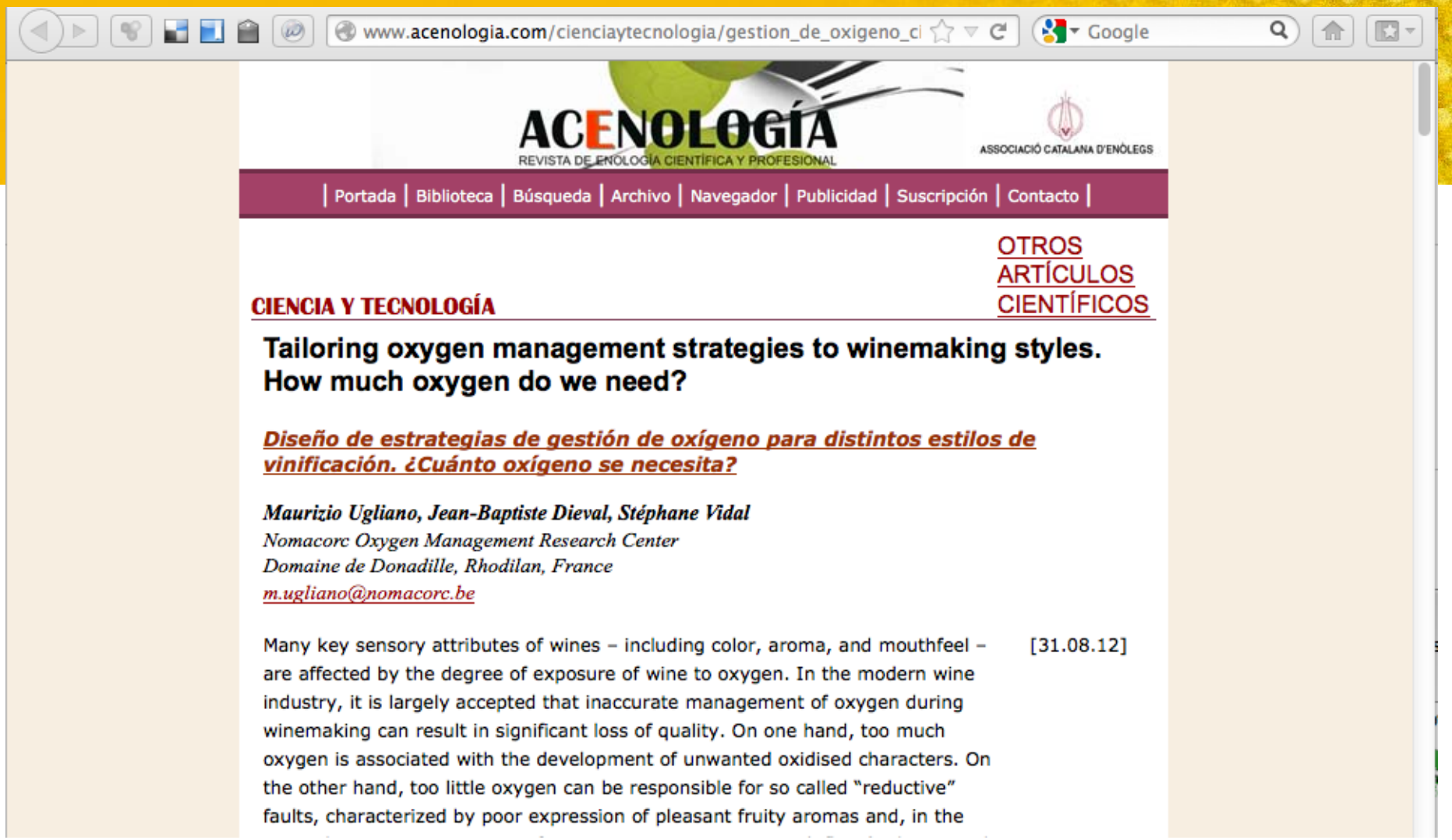
OR Enter the URL

▶ Show advanced settings

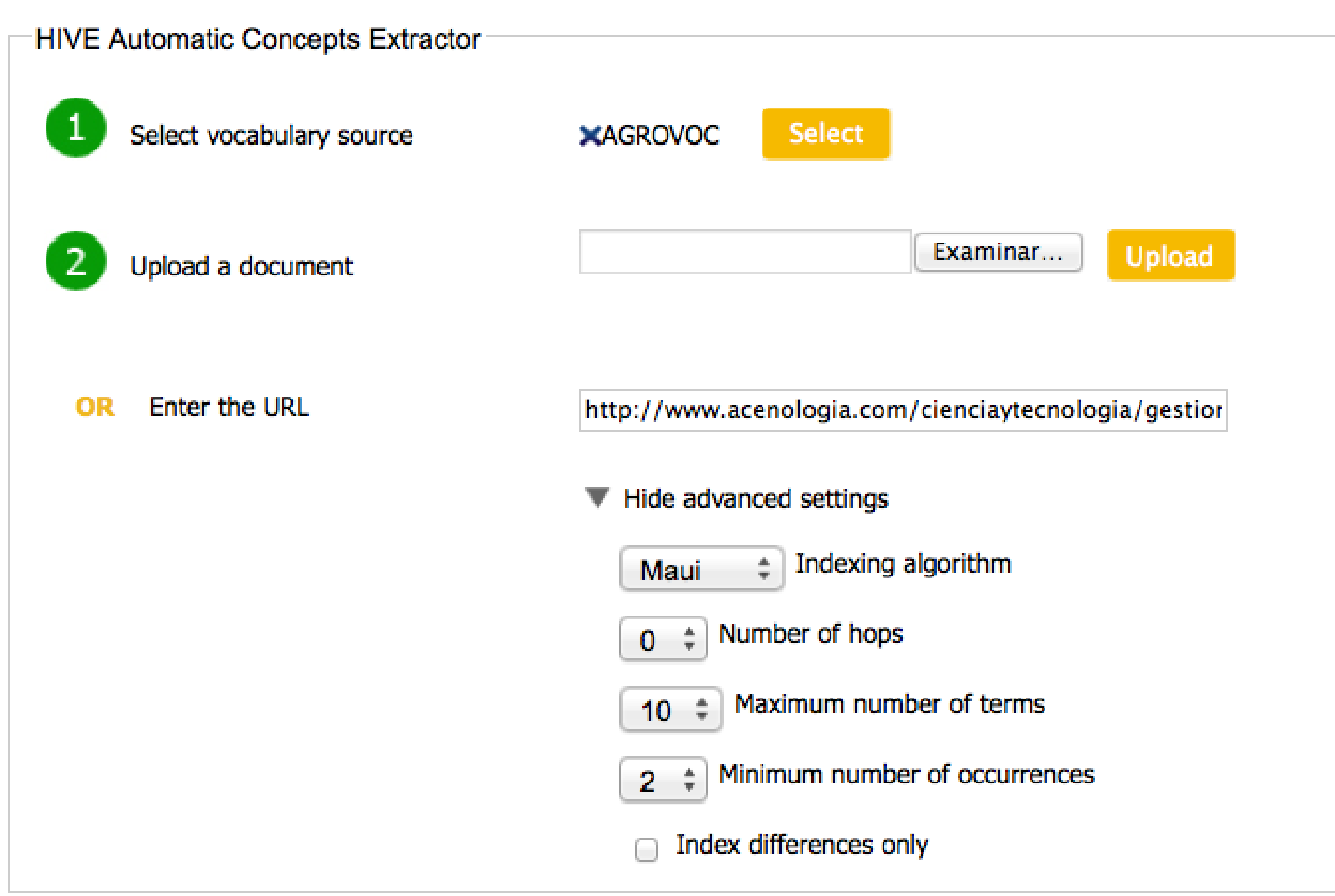
3

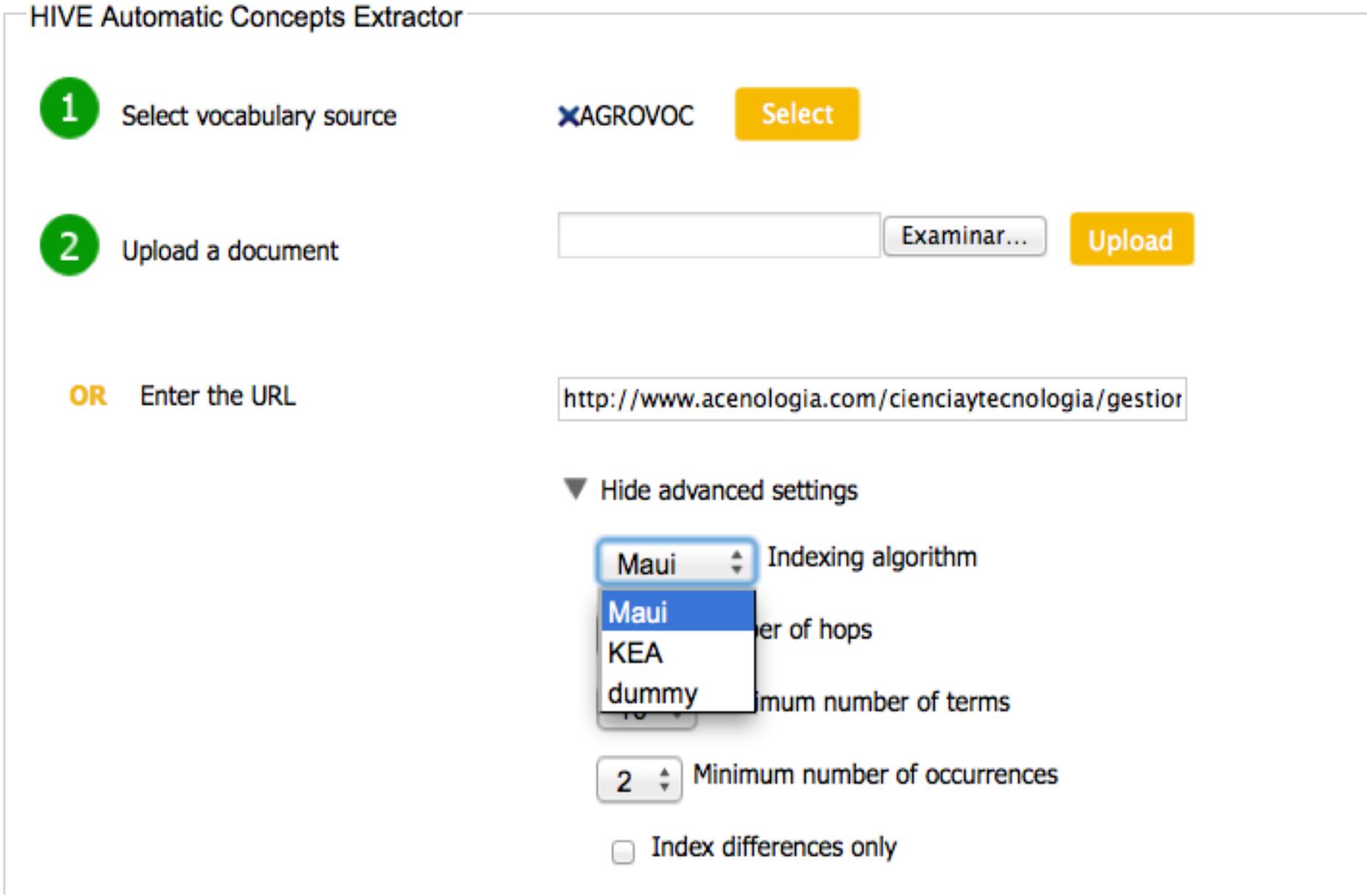
Start Processing

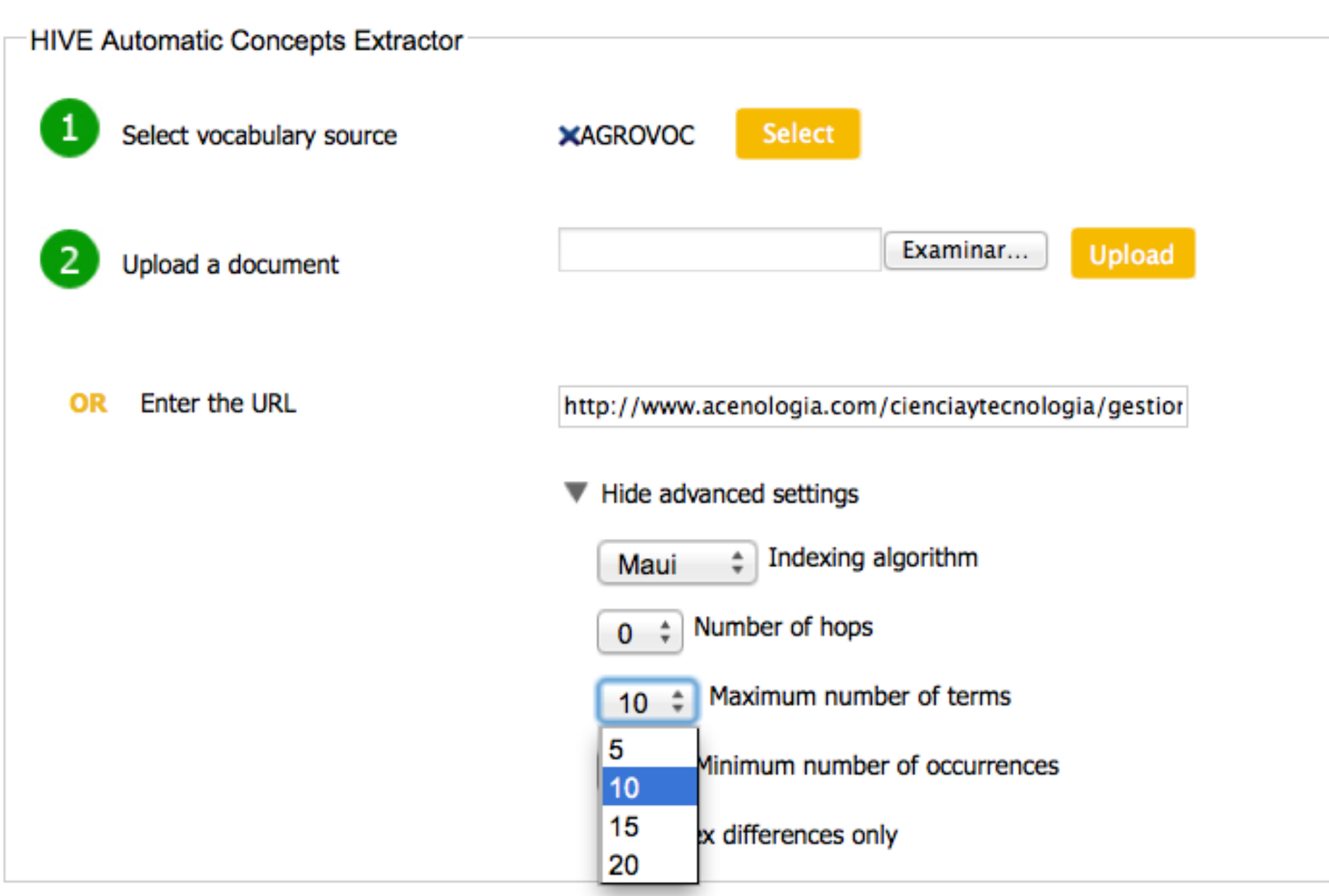
Powered by KEA

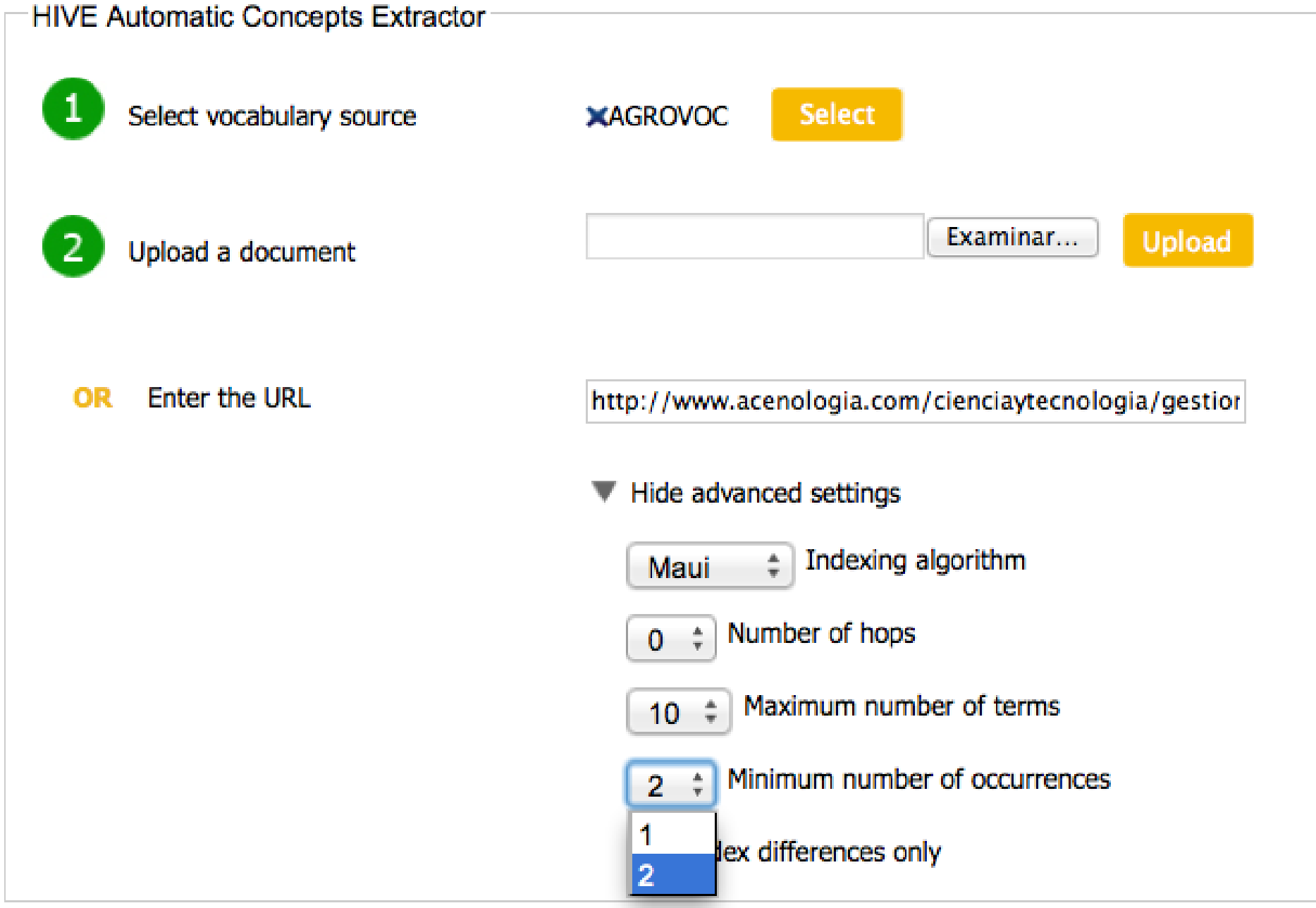


www.acenologia.com/cienciaytecnologia/gestion_de_oxigeno_cienc0812_eng.htm


HIVE Automatic Concepts Extractor
1
Select vocabulary source
AGROVOC
Select
2
Upload a document
Examinar...
Upload
OR
Enter the URL
http://www.acenologia.com/cienciaytecnologia/gestior
Hide advanced settings
Maui
Indexing algorithm
0
Number of hops
10
Maximum number of terms
2
Minimum number of occurrences
Index differences only


HIVE Automatic Concepts Extractor
1
Select vocabulary source
AGROVOC
Select
2
Upload a document
Examinar...
Upload
OR
Enter the URL
http://www.acenologia.com/cienciaytecnologia/gestior
Hide advanced settings
Maui
Indexing algorithm
Maui
KEA
dummy
er of hops
imum number of terms
2
Minimum number of occurrences
Index differences only


HIVE Automatic Concepts Extractor
1
Select vocabulary source
AGROVOC
Select
2
Upload a document
Examinar...
Upload
OR
Enter the URL
http://www.acenologia.com/cienciaytecnologia/gestior
Hide advanced settings
Maui
Indexing algorithm
0
Number of hops
10
Maximum number of terms
5
10
15
20
Minimum number of occurrences
differences only


HIVE Automatic Concepts Extractor
1
Select vocabulary source
AGROVOC
Select
2
Upload a document
Examinar...
Upload
OR
Enter the URL
http://www.acenologia.com/cienciaytecnologia/gestior
Hide advanced settings
Maui
Indexing algorithm
0
Number of hops
10
Maximum number of terms
2
Minimum number of occurrences
1
2
dex differences only

HIVE Automatic Concepts Extractor

1 Select vocabulary source ✖AGROVOC Select

2 Upload a document Examinar... Upload

OR Enter the URL http://www.acenologia.com/cienciaytecnologia/gestior

▼ Hide advanced settings

Maui Indexing algorithm

0 Number of hops

10 Maximum number of terms

2 Minimum number of occurrences

☐ Index differences only

3

Start Processing

Powered by KEA

HIVE-ES Servidor de vocabularios

Helping with Interdisciplinary Vocabulary Engineering

Home | Concept Browser | Indexing

You can select multiple concepts from the cloud and view in the following formats: SKOS RDF/XML, SKOS N triples, Dublin Core, MARC/XML, and MODS/XML.

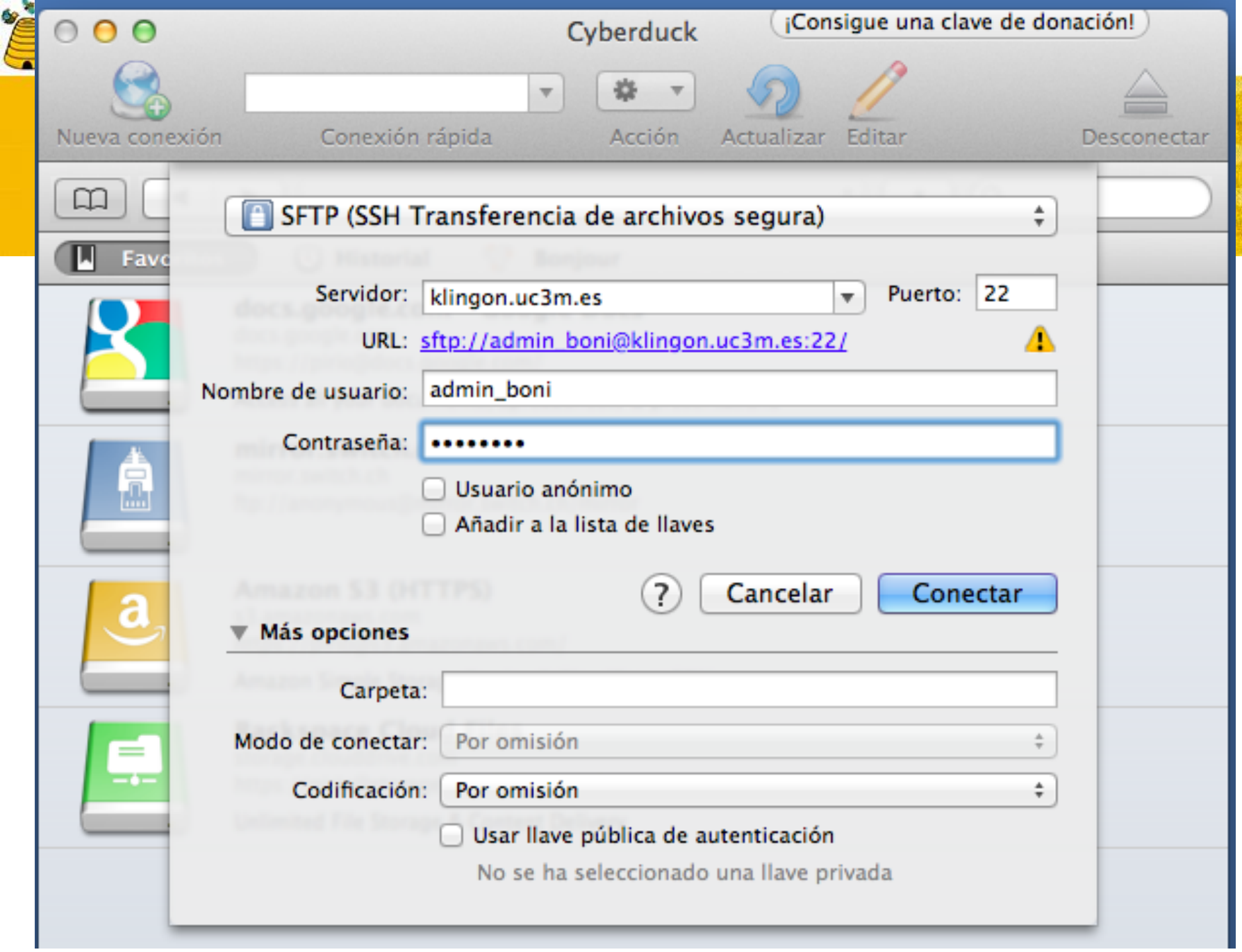
Select Concepts to

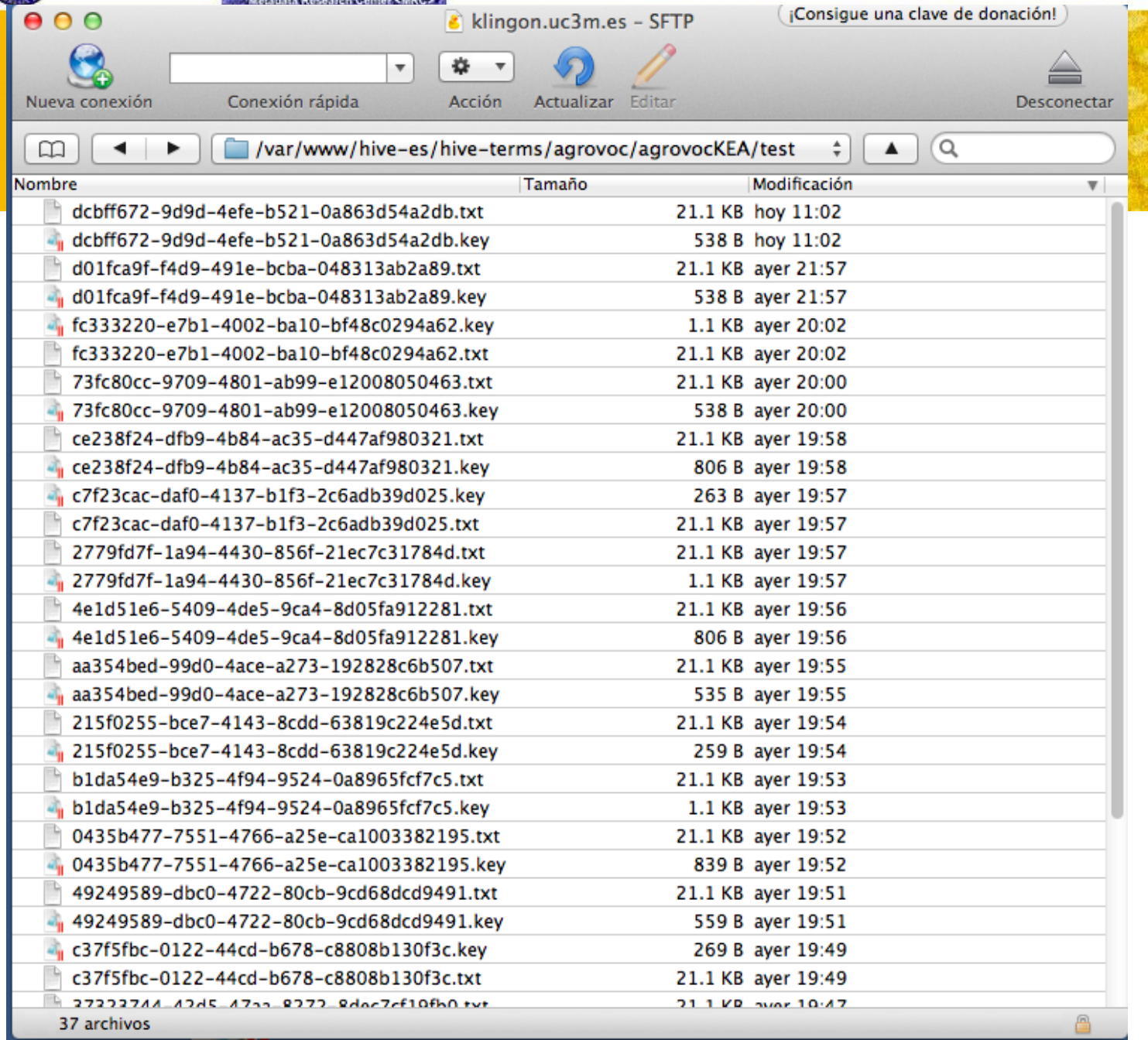
Start Over

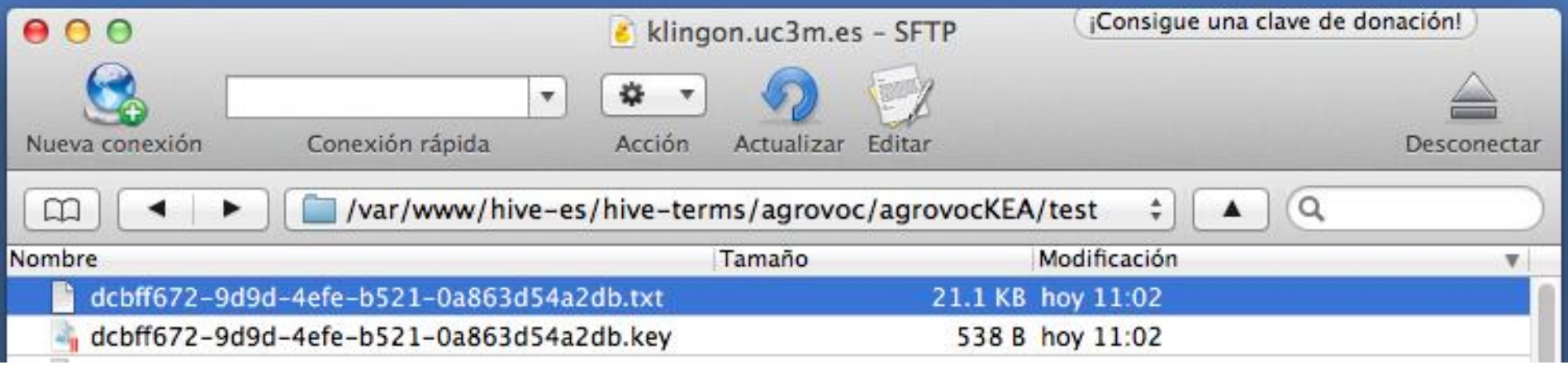
Extracted Concepts Cloud

AGROVOC

Wines Flavour Oxygen Winemaking Aging Bottling Bottles Glutathione Oxidation Oxidants


Cyberduck
¡Consigue una clave de donación!
Nueva conexión
Conexión rápida
Acción
Actualizar
Editar
Desconectar
SFTP (SSH Transferencia de archivos segura)
Servidor:
klingon.uc3m.es
Puerto:
22
URL:
sftp://admin_boni@klingon.uc3m.es:22/
Nombre de usuario:
admin_boni
Contraseña:
Usuario anónimo
Añadir a la lista de llaves
Cancelar
Conectar
Más opciones
Carpeta:
Modo de conectar:
Por omisión
Codificación:
Por omisión
Usar llave pública de autenticación
No se ha seleccionado una llave privada


klingon.uc3m.es – SFTP
¡Consigue una clave de donación!
Nueva conexión
Conexión rápida
Acción
Actualizar
Editar
Desconectar
/var/www/hive-es/hive-terms/agrovoc/agrovocKEA/test
Nombre
Tamaño
Modificación
dcbff672-9d9d-4efe-b521-0a863d54a2db.txt 21.1 KB hoy 11:02
dcbff672-9d9d-4efe-b521-0a863d54a2db.key 538 B hoy 11:02
d01fca9f-f4d9-491e-bcba-048313ab2a89.txt 21.1 KB ayer 21:57
d01fca9f-f4d9-491e-bcba-048313ab2a89.key 538 B ayer 21:57
fc333220-e7b1-4002-ba10-bf48c0294a62.key 1.1 KB ayer 20:02
fc333220-e7b1-4002-ba10-bf48c0294a62.txt 21.1 KB ayer 20:02
73fc80cc-9709-4801-ab99-e12008050463.txt 21.1 KB ayer 20:00
73fc80cc-9709-4801-ab99-e12008050463.key 538 B ayer 20:00
ce238f24-dfb9-4b84-ac35-d447af980321.txt 21.1 KB ayer 19:58
ce238f24-dfb9-4b84-ac35-d447af980321.key 806 B ayer 19:58
c7f23cac-daf0-4137-b1f3-2c6adb39d025.key 263 B ayer 19:57
c7f23cac-daf0-4137-b1f3-2c6adb39d025.txt 21.1 KB ayer 19:57
2779fd7f-1a94-4430-856f-21ec7c31784d.txt 21.1 KB ayer 19:57
2779fd7f-1a94-4430-856f-21ec7c31784d.key 1.1 KB ayer 19:57
4e1d51e6-5409-4de5-9ca4-8d05fa912281.txt 21.1 KB ayer 19:56
4e1d51e6-5409-4de5-9ca4-8d05fa912281.key 806 B ayer 19:56
aa354bed-99d0-4ace-a273-192828c6b507.txt 21.1 KB ayer 19:55
aa354bed-99d0-4ace-a273-192828c6b507.key 535 B ayer 19:55
215f0255-bce7-4143-8cdd-63819c224e5d.txt 21.1 KB ayer 19:54
215f0255-bce7-4143-8cdd-63819c224e5d.key 259 B ayer 19:54
b1da54e9-b325-4f94-9524-0a8965fcf7c5.txt 21.1 KB ayer 19:53
b1da54e9-b325-4f94-9524-0a8965fcf7c5.key 1.1 KB ayer 19:53
0435b477-7551-4766-a25e-ca1003382195.txt 21.1 KB ayer 19:52
0435b477-7551-4766-a25e-ca1003382195.key 839 B ayer 19:52
49249589-dbc0-4722-80cb-9cd68dcd9491.txt 21.1 KB ayer 19:51
49249589-dbc0-4722-80cb-9cd68dcd9491.key 559 B ayer 19:51
c37f5fbc-0122-44cd-b678-c8808b130f3c.key 269 B ayer 19:49
c37f5fbc-0122-44cd-b678-c8808b130f3c.txt 21.1 KB ayer 19:49
37323744-42d5-47aa-8272-8dec7cf19fb0.txt 21.1 KB ayer 19:47
37 archivos



dcbff672-9d9d-4efe-b521-0a863d54a2db: Bloc de notas

Archivo Edición Formato Ver Ayuda

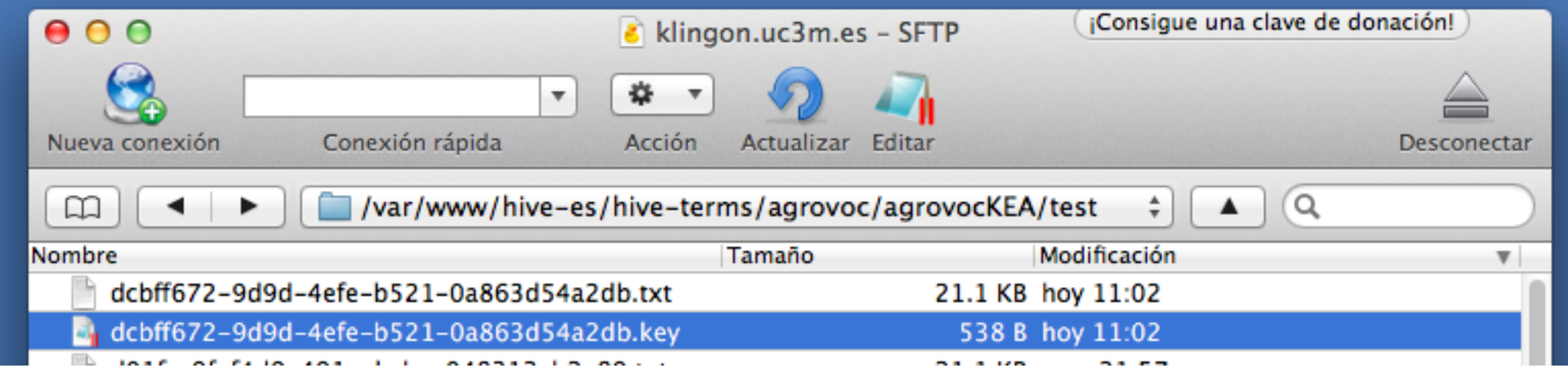
| Portada | Biblioteca | Búsqueda | Archivo | Navegador | Publicidad | Suscripción | Contacto |

CIENCIA Y TECNOLOGÃ•A OTROS ARTÃ•CULOS CIENTÃ•FICOS

Tailoring oxygen management strategies to winemaking styles. How much oxygen do we need? DiseÃ±o de estrategias de gestiÃ³n de oxÃ-geno para distintos estilos de vinificaciÃ³n. Â¿CuÃ¡nto oxÃ-geno se necesita?

Maurizio Ugliano, Jean-Baptiste Dieval, StÃ©phane Vidal Nomacorc
Oxygen Management Research Center Domaine de Donadille, Rhodilan, France
m.ugliano@nomacorc.be

Many key sensory attributes of wines Â- including color, aroma, and mouthfeel Â- are affected by the degree of exposure of wine to oxygen. In the modern wine industry, it is largely accepted that inaccurate management of oxygen during winemaking can result in significant loss of quality. On one hand, too much oxygen is associated with the development of unwanted oxidised characters. On the other hand, too little oxygen can be responsible for so called Â"reductiveÂ" faults, characterized by poor expression of pleasant fruity aromas and, in the most obvious cases, aromas of rotten egg, sewage, or struck flint (Ugliano et al 2009). In addition, the complex array of chemical reactions that contribute to softening tannin harshness and stabilizing colour during wine ageing are also closely connected to the oxidative processes that can potentially take place either in the cellar or in the bottle. Although these concepts have been long established in the wine industry, on a practical level it remains difficult to effectively assess the oxygen demand of a wine. In other words, in the vast space defined by too much to too little oxygen, the degree of oxygen exposure that will provide optimal sensory expression of a given wine is still hard to define. At a general level, it is accepted that wines obtained from certain grape varieties are particularly sensitive to oxygen, reflecting the fact that some of the chemical components key to their sensory attributes are strongly modulated by oxygen. Sauvignon blanc is a well documented example of an oxygen sensitive wine. In addition, anecdotal evidence Â- in some cases supported by scientific literature Â- suggests that moderate oxygen exposure is crucial to the development of certain key aroma attributes, as in the case of



dcbff672-9d9d-4efe-b521-0a863d54a2db.key

```
Wines   http://www.fao.org/aos/agrovoc#c_8406   0.5082
Flavour http://www.fao.org/aos/agrovoc#c_10893  0.4082
Oxygen  http://www.fao.org/aos/agrovoc#c_5477   0.3498
Winemaking  http://www.fao.org/aos/agrovoc#c_8405   0.3415
Aging   http://www.fao.org/aos/agrovoc#c_192    0.3415
Bottling    http://www.fao.org/aos/agrovoc#c_1032   0.3082
Bottles http://www.fao.org/aos/agrovoc#c_1031   0.3082
Glutathione http://www.fao.org/aos/agrovoc#c_11184  0.2498
Oxidation   http://www.fao.org/aos/agrovoc#c_5472   0.2415
Oxidants    http://www.fao.org/aos/agrovoc#c_24879  0.2415
```

Helping with Interdisciplinary Vocabulary Engineering

Home | Concept Browser | Indexing

You can select multiple concepts from the cloud and view in the following formats: SKOS RDF/XML, SKOS N triples, Dublin Core, MARC/XML, and MODS/XML.

Select Concepts to

Start Over

Extracted Concepts Cloud

AGROVOC

Wines Flavour Oxygen Winemaking Aging Bottling Bottles Glutathione Oxidation Oxidants

dcbff672-9d9d-4efe-b521-0a863d54a2db.key

```
Wines   http://www.fao.org/aos/agrovoc#c_8406   0.5082
Flavour http://www.fao.org/aos/agrovoc#c_10893  0.4082
Oxygen  http://www.fao.org/aos/agrovoc#c_5477   0.3498
Winemaking  http://www.fao.org/aos/agrovoc#c_8405   0.3415
Aging   http://www.fao.org/aos/agrovoc#c_192    0.3415
Bottling    http://www.fao.org/aos/agrovoc#c_1032   0.3082
Bottles http://www.fao.org/aos/agrovoc#c_1031   0.3082
Glutathione http://www.fao.org/aos/agrovoc#c_11184  0.2498
Oxidation   http://www.fao.org/aos/agrovoc#c_5472   0.2415
Oxidants    http://www.fao.org/aos/agrovoc#c_24879  0.2415
```

Systematic analysis of HIVE indexing performance: **Initial Results**

- The % of relevant extracted terms is higher in VINO (72-100%) and AGROVOC (≅80%) than in LEM (10-55%)→ More specific vocabularies offer more relevant results.
- A higher number of extracted concepts does not imply higher precision.
- A higher number of extracted concepts implies lower average probabilities.
- Probabilities are not always consistent with evaluators assessment of terms' precision.
- For VINO and AGROVOC, KEA always give the same probability to all the extracted terms. Maui offers variations.
- AGROVOC offers relevants results indexing documents both in English and Spanish (Agrovoc concepts in HIVE are in English).

LEM Vocabulary

Algorithm	Minim ocurrs.	N. max. of terms.	N. extracted terms	N. relevant terms	Precision	Average precision (human ass)	Average probability
KEA	1	5	5	2	40,00%	3,00	0,76924
KEA	1	10	10	2	20,00%	3,40	0,36195
KEA	1	15	15	6	40,00%	2,93	0,38091
KEA	1	20	20	11	55,00%	2,70	0,19683
KEA	2	5	5	2	40,00%	3,00	0,46836
KEA	2	10	10	3	30,00%	3,20	0,26720
KEA	2	15	15	6	40,00%	3,07	0,18331
KEA	2	20	20	8	40,00%	3,25	0,13799
Maui	1	5	5	1	20,00%	3,40	0,29956
Maui	1	10	10	1	10,00%	3,70	0,24965
Maui	1	15	15	4	26,67%	3,53	0,19738
Maui	1	20	20	5	25,00%	3,55	0,15245
Maui	2	5	5	1	20,00%	3,40	0,36346
Maui	2	10	10	1	10,00%	3,70	0,24965
Maui	2	15	15	4	26,67%	3,53	0,19738
Maui	2	20	20	5	25,00%	3,55	0,15245

VINO Vocabulary

Algorithm	Minim ocurrs.	N. max. of terms.	N. extracted terms	N. relevant terms	Precision	Average precision (human ass.1-4)	Average probability
KEA	1	5	5	5	100,00%	2,40	0,1689
KEA	1	10	10	9	90,00%	2,70	0,1689
KEA	1	15	15	14	93,33%	2,67	0,1689
KEA	1	20	16	12	75,00%	2,75	0,1689
KEA	2	5	5	5	100,00%	2,40	0,1689
KEA	2	10	10	9	90,00%	2,80	0,1689
KEA	2	15	11	9	81,82%	2,82	0,1689
KEA	2	20	10	9	90,00%	3,20	0,1689
Maui	1	5	5	3	60,00%	3,40	0,3105
Maui	1	10	10	8	80,00%	2,80	0,2084
Maui	1	15	15	11	73,33%	3,27	0,1274
Maui	2	5	5	4	80,00%	3,00	0,2146
Maui	2	10	10	9	90,00%	3,10	0,0371
Maui	2	15	11	8	72,73%	3,09	0,0338
Maui	2	20	11	9	81,82%	3,09	0,1313

Systematic analysis of HIVE indexing performance: Further research questions

- Integration and evaluation of alternative algorithms
  - What is the best algorithm for automatic term suggestion for Spanish vocabularies?
  - Do different algorithms perform better for title, abstract, full-text, data?
  - Does the extension/format of the input document influence the quality of results?
  - Which is the relationship between number of training documents and algorithm performance?
  - Do different algorithms perform better for a particular vocabulary/taxonomy/ontology?
  - Do different algorithms perform better for a particular subject domain?

Challenges

- Training of KEA++/MAUI models
  - General Subject Headings list vs. Thesaurus, number of indexing terms, number of training documents, specificity of documents.
- Combining many vocabularies during the indexing/term
  - matching phase is difficult, time consuming, inefficient.
  - NLP and machine learning offer promise
- Interoperability = dumbing down
  - ontologies

Limitations and future developments

- **Administration level:**
  - Administrator interface
  - Automatic SKOS vocabularies/ training document set uploading
  - Access to indexing results history through admin interface.
  - Vocabulary update and synchronization (→ integration of HIVE with LCSH Atom Feed http://id.loc.gov/authorities/feed)

- **Browsing/Search:**
  - Browsing multiple vocabularies simultaneously, through their mappings (*closeMatch*?)
  - Visual browsing of vocabularies' concepts.
  - Advanced search: limit types of terms, hierarchy depth, nº of terms.
  - Search results: ordering and filtering options, visualization options.

Limitations and future developments

- **Indexing:**
  - Indexing multiple documents at the same time.
  - Visualization options: cloud / list.
  - Ordering options: byconcept weights/ vocabulary, alphabetically, specificity (BT/NT).
  - Linking options: select and export SKOS concept, link it to document by RDF (give document an URI…)

- **Integration:**
  - Repositories and controlled vocabularies / author keywords.
  - Digital library systems.
  - Traditional library catalogs? Bound to disappear… RDA >> RDF bibliographic catalogs.

HIVE and HIVE-ES Teams

HIVE

HIVE-ES

Thank you!

- Metadata Research Center (UNC)
- NESCent (National Evolutionary Synthesis Center)
- Tecnodoc Group (UC3M)
- Duke University
- Long Term Ecological Research Network (LTER)
- Institute of Museum and Library Services
- National Science Foundation
- National Library of Spain

References

- Huang, L. (2010). *Usability Testing of the HIVE: A System for Dynamic Access to Multiple Controlled Vocabularies for Automatic Metadata Generation*. Master's Paper, M.S. in IS degree. SILS, UNC Chapel Hill.
- Medelyan, O. (2010). *Human-competitive automatic topic indexing*. Unpublished dissertation.
- Medelyan, O. and Whitten I.A. (2008). "Domain independent automatic keyphrase indexing with small training sets." *Journal of the American Society for Information Science and Technology*, (59) 7: 1026-1040).
- Moens, M.F. (2000). *Automatic Indexing and Abstracting Documents*. London: Kluwer.
- Shearer, J. R. (2004). A Practical Exercise in Building a Thesaurus, *Cataloging & Classification Quarterly*, 37:3-4, 35-56
- Sherman, J. K. (2010). *Automatic metadata generation: a comparison of two annotators*. Master's Paper, M.S. in IS degree. SILS, UNC Chapel Hill.
- Shiri, A. and Chase-Kruszewsky, S. (2009). Knowledge organization systems in North American digital library collections. Program: electronic library and information systems. 43 (2) pp. 121-139
- White, H.; Willlis, C.; Greenberg, J. (2012) The HIVE Impact: Contributing to Consistency via Automatic Indexing. *iConference 2012*, February 7-10, Toronto, Ontario, Canada ACM 978-1-4503-0782-6/12/02.