Experiences in Mapping Multiple Vocabularies in Agriculture

Lori Finch
National Agricultural Library
lfinch@nal.usda.gov

Presented at the joint CENDI / NKOS event at the World Bank
“New Dimensions in Knowledge Organization Systems”
September 11, 2008

On the Record: Recommendations

“On the Record: Report of The Library of Congress Working Group on the Future of Bibliographic Control” - January 9, 2008

Findings and Recommendations

- 4.3 Optimize LCSH for Use and ReUse
- 4.3.3. Encourage Application of, and Cross-Referencing with, Other Controlled Subject Vocabularies
- 4.3.3.3 Explore mechanisms to exploit cross-vocabulary linkages to enhance retrieval, without limiting to the headings explicitly provided in individual bibliographic records

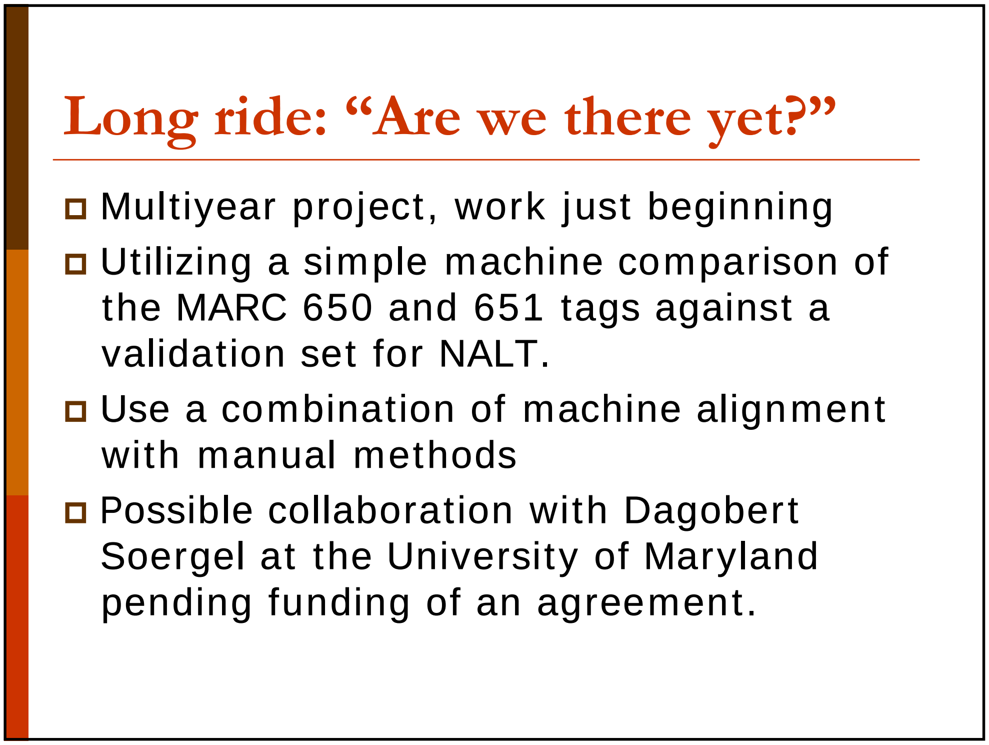
Controlled Vocabulary at NAL

- AGRICOLA – 5 million bib records
- LCSH used for cataloging of books and other media
- NALT used for indexing of articles, signed chapters, etc.

It would be beneficial to the Library's patrons if they could conduct simultaneous subject searches in both the "Books" and "Articles" catalogs no matter which thesaurus is used.

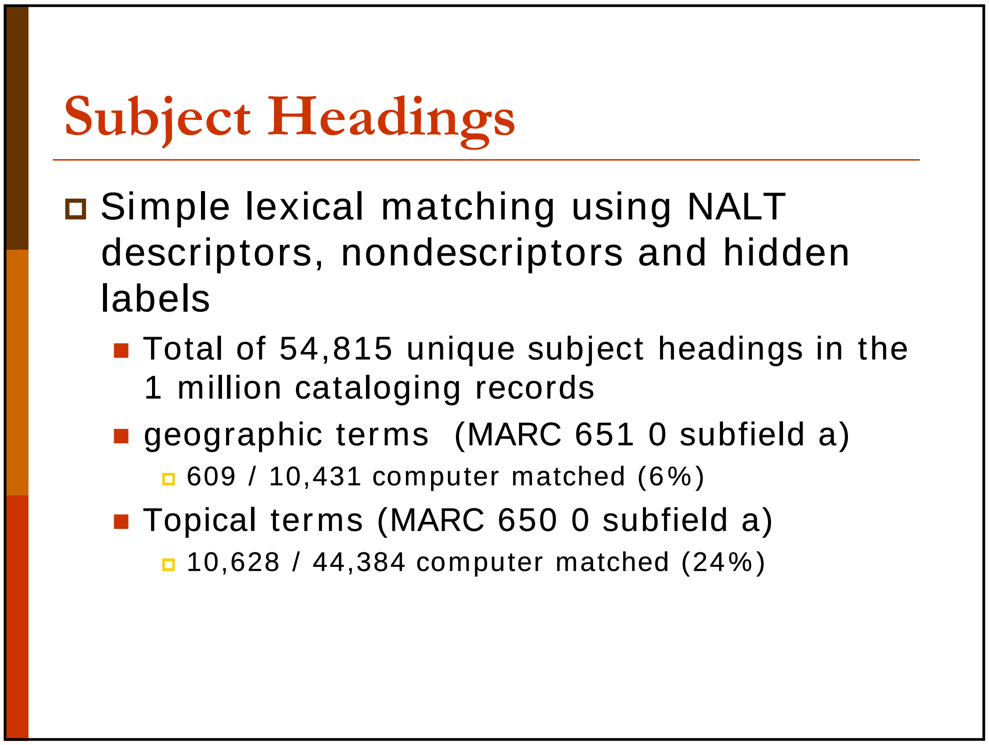
LCSH to NALT mapping project

- Objectives:
  - Alignment of LCSH to NALT so that there is an automated assignment of NALT to existing and new cataloging records.
  - Creation of NALT MARC authority records with links to LCSH and make this file available on the thesaurus website.
  - Creation of SKOS file with the LCSH-NALT alignment using SKOS mapping properties



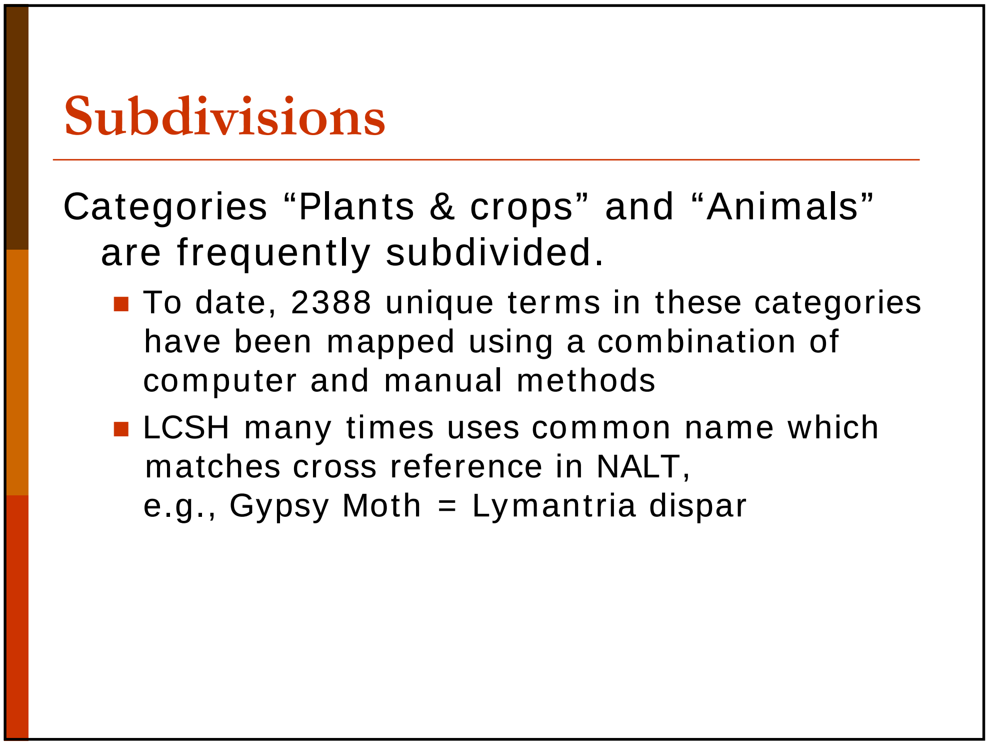
Long ride: "Are we there yet?"

- Multiyear project, work just beginning
- Utilizing a simple machine comparison of the MARC 650 and 651 tags against a validation set for NALT.
- Use a combination of machine alignment with manual methods
- Possible collaboration with Dagobert Soergel at the University of Maryland pending funding of an agreement.



Subject Headings

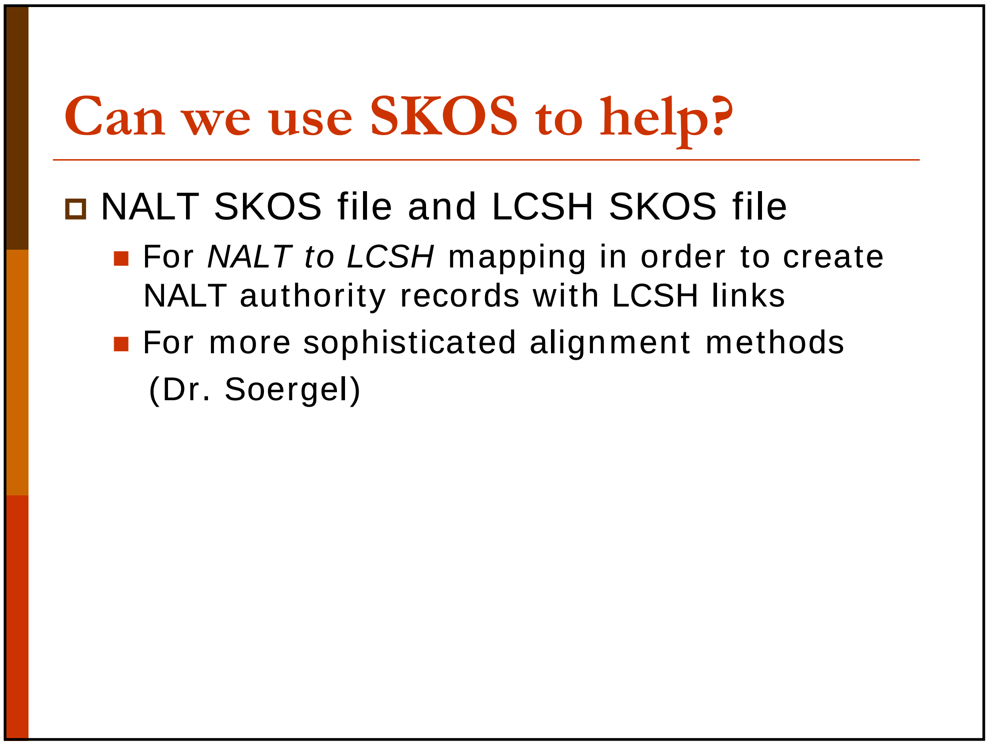
- Simple lexical matching using NALT descriptors, nondescriptors and hidden labels
  - Total of 54,815 unique subject headings in the 1 million cataloging records
  - geographic terms (MARC 651 0 subfield a)
    - 609 / 10,431 computer matched (6%)
  - Topical terms (MARC 650 0 subfield a)
    - 10,628 / 44,384 computer matched (24%)



Subdivisions

Categories “Plants & crops” and “Animals” are frequently subdivided.

- To date, 2388 unique terms in these categories have been mapped using a combination of computer and manual methods
- LCSH many times uses common name which matches cross reference in NALT, e.g., Gypsy Moth = Lymantria dispar

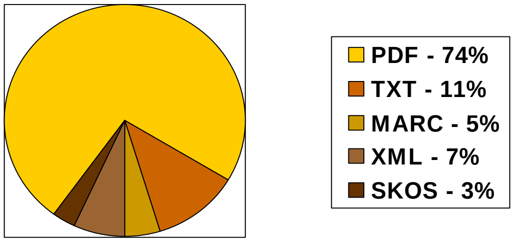


Can we use SKOS to help?

- NALT SKOS file and LCSH SKOS file
  - For *NALT to LCSH* mapping in order to create NALT authority records with LCSH links
  - For more sophisticated alignment methods (Dr. Soergel)

Popularity of data formats

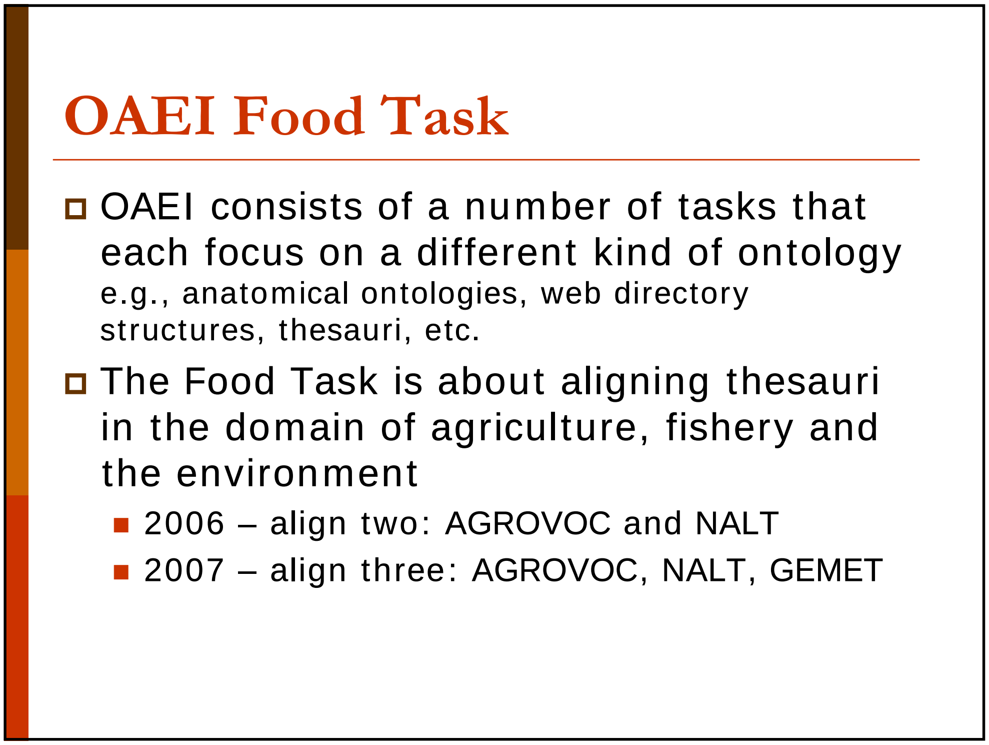
- NALT available for downloading in PDF, TXT, US MARC, XML and … SKOS (in 2006) http://agclass.nal.usda.gov/download.shtml
- 8,000 downloads per year



SKOS data used for evaluation of automated mapping of vocabularies

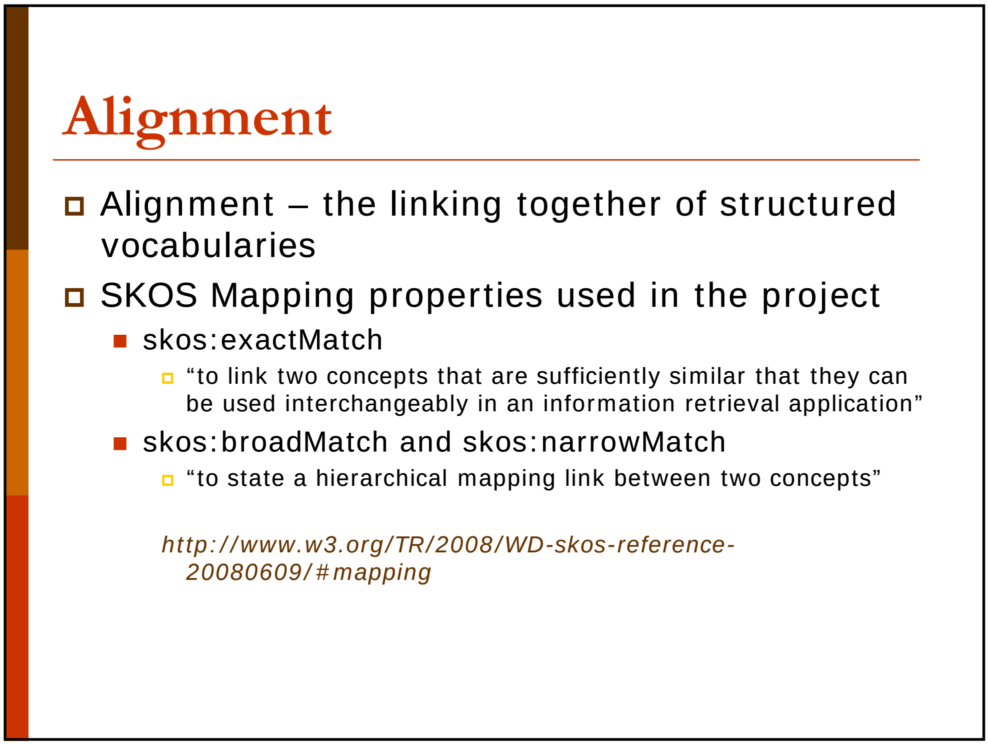
- Willem Robert van Hage, Vrije Universiteit, Amsterdam
- Ontology Alignment Evaluation Initiative oaei.ontologymatching.org

  OAEI organized to guide the development of automatic alignment systems and to evaluate the performance of these systems.



OAEI Food Task

- OAEI consists of a number of tasks that each focus on a different kind of ontology e.g., anatomical ontologies, web directory structures, thesauri, etc.
- The Food Task is about aligning thesauri in the domain of agriculture, fishery and the environment
  - 2006 – align two: AGROVOC and NALT
  - 2007 – align three: AGROVOC, NALT, GEMET



Alignment

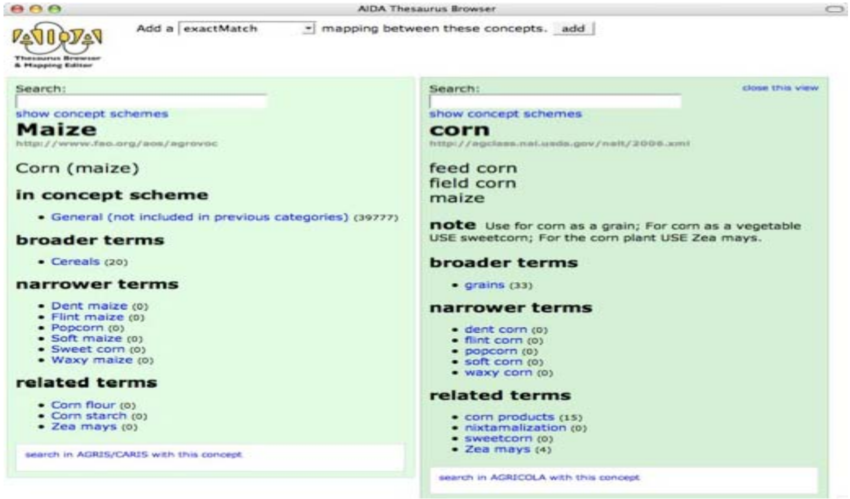
- Alignment – the linking together of structured vocabularies
- SKOS Mapping properties used in the project
  - skos:exactMatch
    - “to link two concepts that are sufficiently similar that they can be used interchangeably in an information retrieval application”
  - skos:broadMatch and skos:narrowMatch
    - “to state a hierarchical mapping link between two concepts”

http://www.w3.org/TR/2008/WD-skos-reference-20080609/#mapping

Participants

- 5 Computer applications in 2006
  - Falcon-AO – South East University
  - RiMOM – Tsinghua University
  - Prior – University of Pittsburgh
  - COMA++ - University of Leipzig
  - HMatch – University of Milan
- Human mapping
  - Manual mapping performed to evaluate performance of computer methods
  - Manually mapped a sample – 7% of terms
  - Participants used an online tool developed by Willem for the mapping

Manual mapping tool



Add a exactMatch mapping between these concepts.

Search:

show concept schemes

Maize

http://www.fao.org/aos/agrovoc

Corn (maize)

in concept scheme

- General (not included in previous categories) (39777)

broader terms

- Cereals (20)

narrower terms

- Dent maize (0)
- Flint maize (0)
- Popcorn (0)
- Soft maize (0)
- Sweet corn (0)
- Waxy maize (0)

related terms

- Corn flour (0)
- Corn starch (0)
- Zea mays (0)

search in AGRIS/CARIS with this concept

Search:

show concept schemes

corn

http://agclass.nal.usda.gov/nalt/2006.xml

feed corn
field corn
maize

note Use for corn as a grain; For corn as a vegetable USE sweetcorn; For the corn plant USE Zea mays.

broader terms

- grains (33)

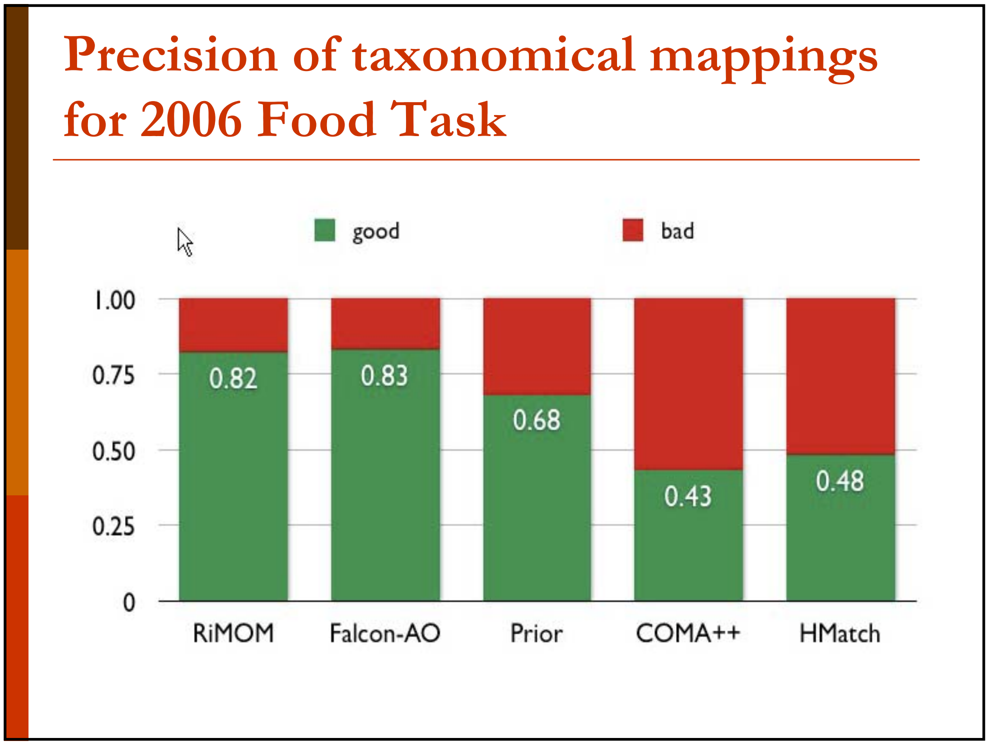
narrower terms

- dent corn (0)
- flint corn (0)
- popcorn (0)
- soft corn (0)
- waxy corn (0)

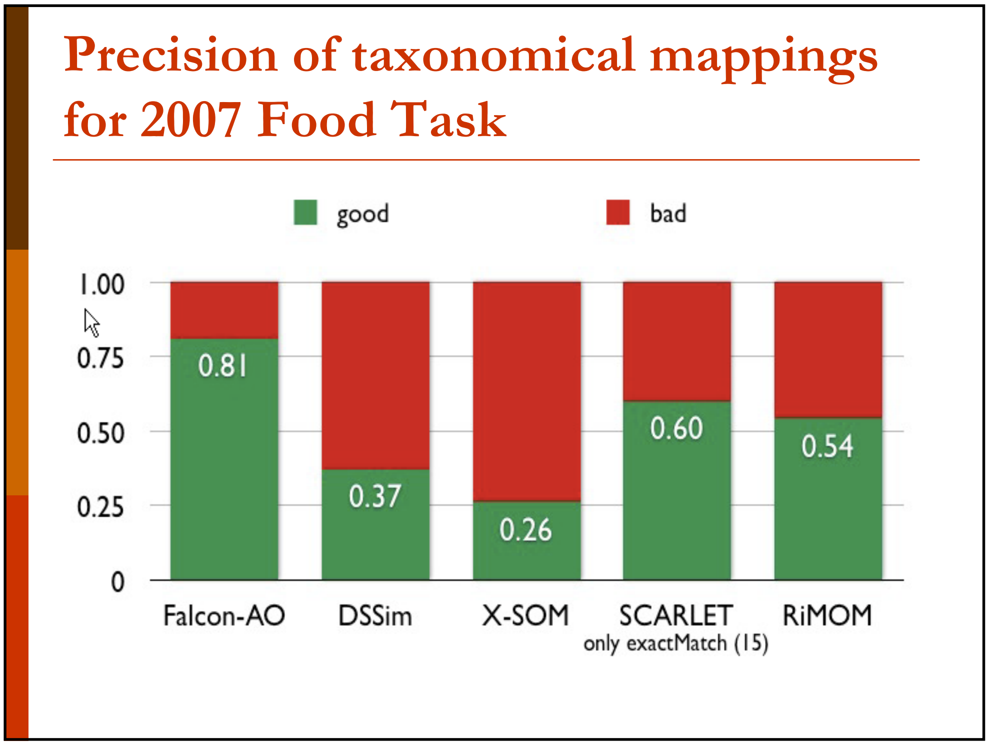
related terms

- corn products (15)
- nixtamalization (0)
- sweetcorn (0)
- Zea mays (4)

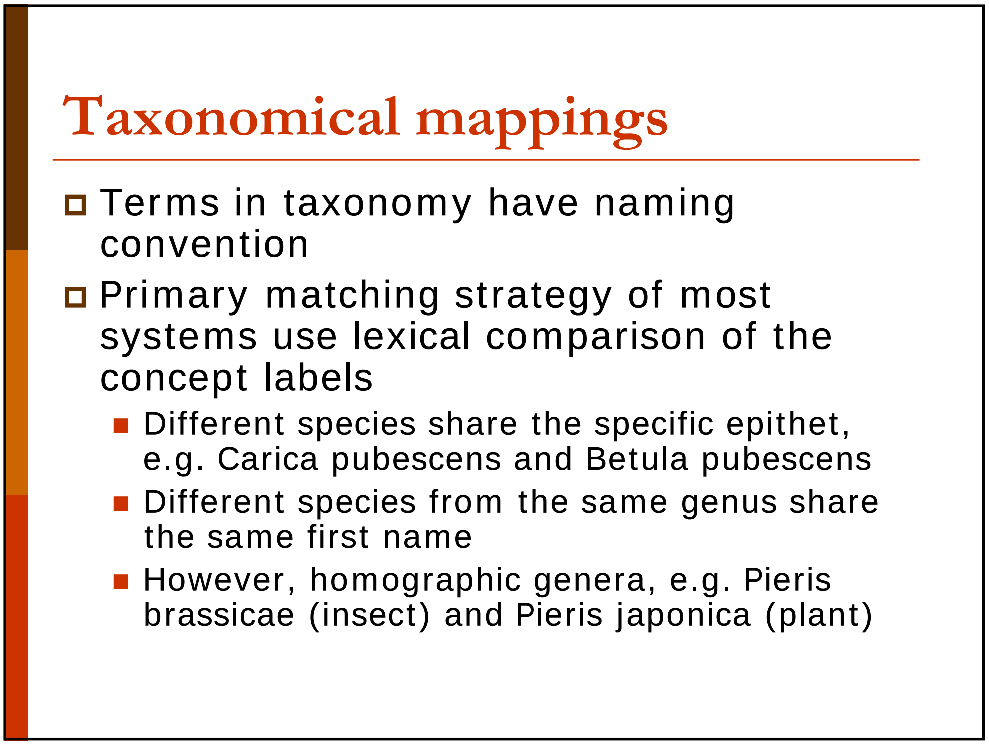
search in AGRICOLA with this concept



Precision of taxonomical mappings for 2006 Food Task

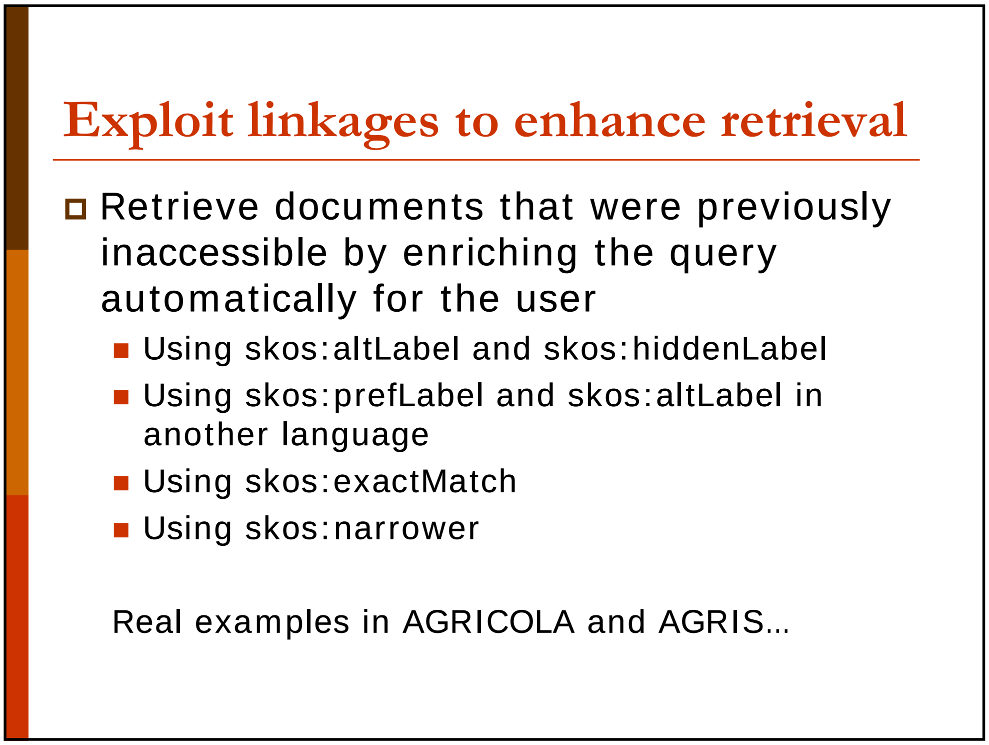


Precision of taxonomical mappings for 2007 Food Task



Taxonomical mappings

- Terms in taxonomy have naming convention
- Primary matching strategy of most systems use lexical comparison of the concept labels
  - Different species share the specific epithet, e.g. Carica pubescens and Betula pubescens
  - Different species from the same genus share the same first name
  - However, homographic genera, e.g. Pieris brassicae (insect) and Pieris japonica (plant)

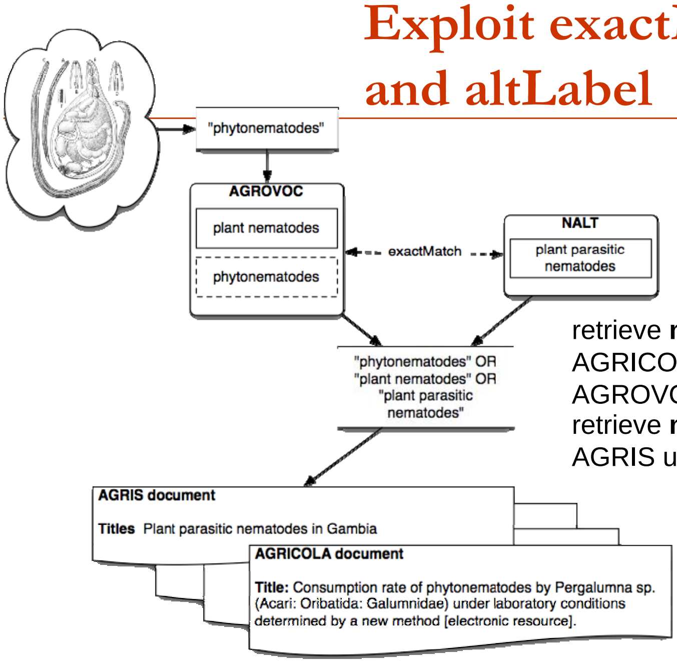


Exploit linkages to enhance retrieval

- Retrieve documents that were previously inaccessible by enriching the query automatically for the user
  - Using skos:altLabel and skos:hiddenLabel
  - Using skos:prefLabel and skos:altLabel in another language
  - Using skos:exactMatch
  - Using skos:narrower

Real examples in AGRICOLA and AGRIS…

Exploit exactMatch and altLabel



"phytonematodes"

AGROVOC

plant nematodes

phytonematodes

exactMatch

NALT

plant parasitic nematodes

"phytonematodes" OR "plant nematodes" OR "plant parasitic nematodes"

AGRIS document

Titles Plant parasitic nematodes in Gambia

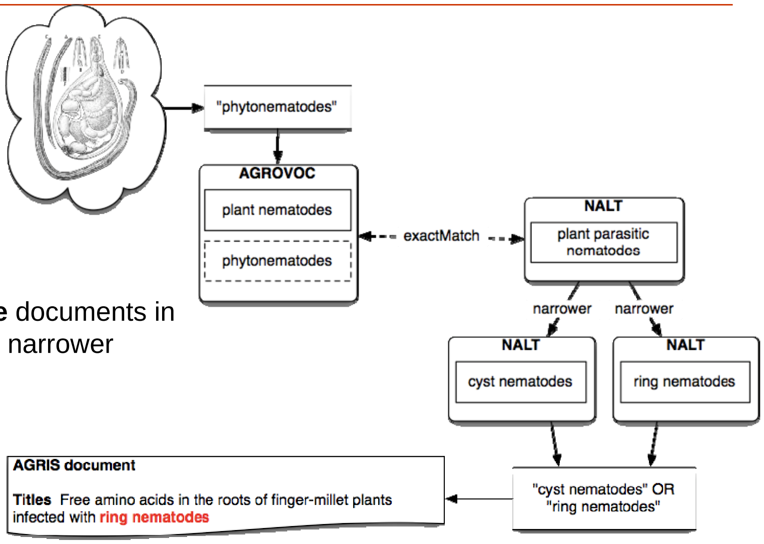
AGRICOLA document

Title: Consumption rate of phytonematodes by Pergalumna sp. (Acari: Oribatida: Galumnidae) under laboratory conditions determined by a new method [electronic resource].

retrieve **more** documents in AGRICOLA using AGROVOC

retrieve **more** documents in AGRIS using NALT

Exploit exactMatch and narrower



"phytonematodes"

AGROVOC

plant nematodes

phytonematodes

exactMatch

NALT

plant parasitic nematodes

narrower

narrower

NALT

cyst nematodes

NALT

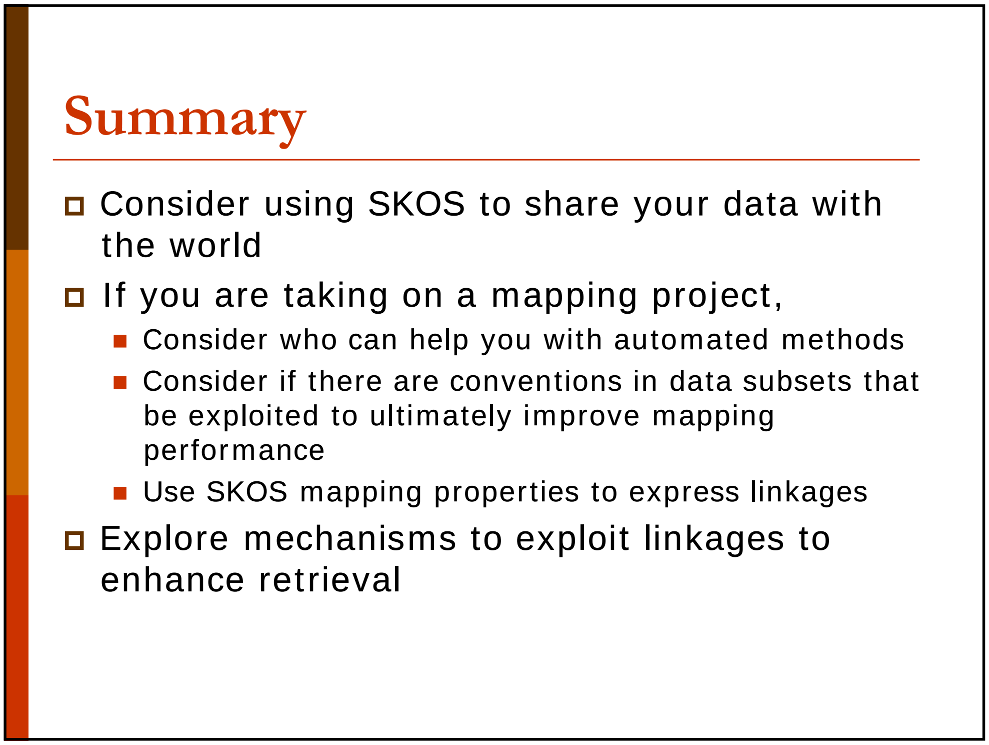
ring nematodes

"cyst nematodes" OR "ring nematodes"

AGRIS document

Titles Free amino acids in the roots of finger-millet plants infected with **ring nematodes**

retrieve **more** documents in AGRIS using narrower terms
from NALT



Summary

- Consider using SKOS to share your data with the world
- If you are taking on a mapping project,
  - Consider who can help you with automated methods
  - Consider if there are conventions in data subsets that be exploited to ultimately improve mapping performance
  - Use SKOS mapping properties to express linkages
- Explore mechanisms to exploit linkages to enhance retrieval