AGROVOC-CAT
Mapping project

# The FAO Multilingual Thesaurus (AGROVOC) and the Chinese Agricultural Thesaurus (CAT) **Mapping** Project

4th NKOS Workshop

September 22, 2005 Vienna

## 4th European Networked Knowledge Organization Systems (NKOS) Workshop

## EDCL2005 September 22nd, Vienna

# Outline

AGROVOC-CAT
Mapping project

4th NKOS Workshop
September 22, 2005 Vienna

- Objective
- Benefits
- Application
- AGROVOC and CAT
- Definitions
- Mapping relationships
- Pre-processing
- Work procedure
- Tools
- Using Protégé
- Implications
- Conclusions

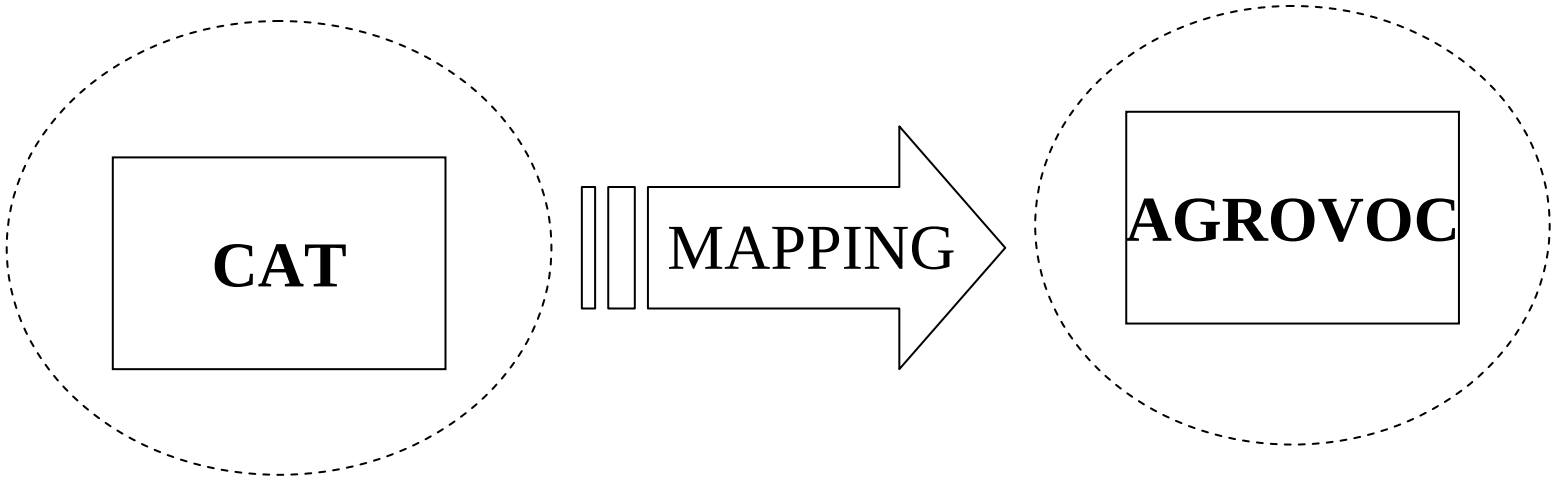
AGROVOC-CAT
Mapping project

# Objective

**To map** two multilingual agricultural terminologies

4th NKOS Workshop

September 22, 2005 Vienna



# Benefits

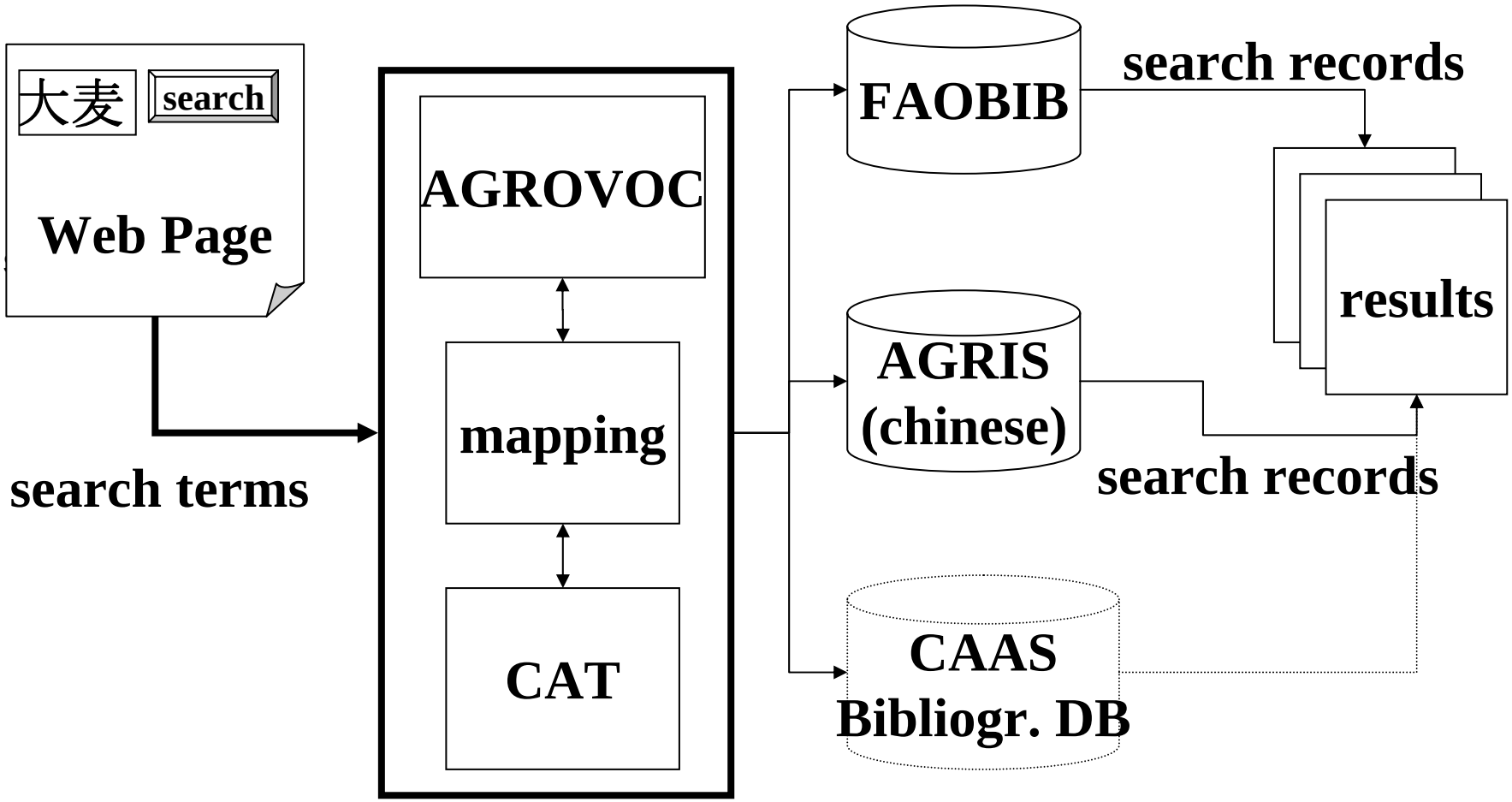




- **Multilinguality**: Enrich Chinese-English vocabulary in the agricultural domain
- **Domain coverage**: Expand and deepen coverage of the agricultural domain
- **Interoperability**: Extend access to one vocabulary system via another.

# Application: Terminology Brokering

**AGROVOC-CAT**
**Mapping project**



# AGROVOC and CAT



- **AGROVOC**:
  - –**27736** English terms: 16769 descriptors, 10967 non descriptors
  - –**25060** Chinese terms: 16628 descriptors, 8432 non descriptors
  - –**1240** top terms
  - –organized in **130** categories (AGRIS/CARIS)
  - –includes biological taxonomy and geographical names
- **CAT**:
  - –**64638** Chinese terms: 51614 descriptors, 13024 non-descriptors
  - –**51400** descriptors has at least one translation
  - –**2332** top terms
  - –organized in **40** categories (e.g. crops, etc.)
  - –includes biological taxonomy and geographical names

# Definitions (1/2)





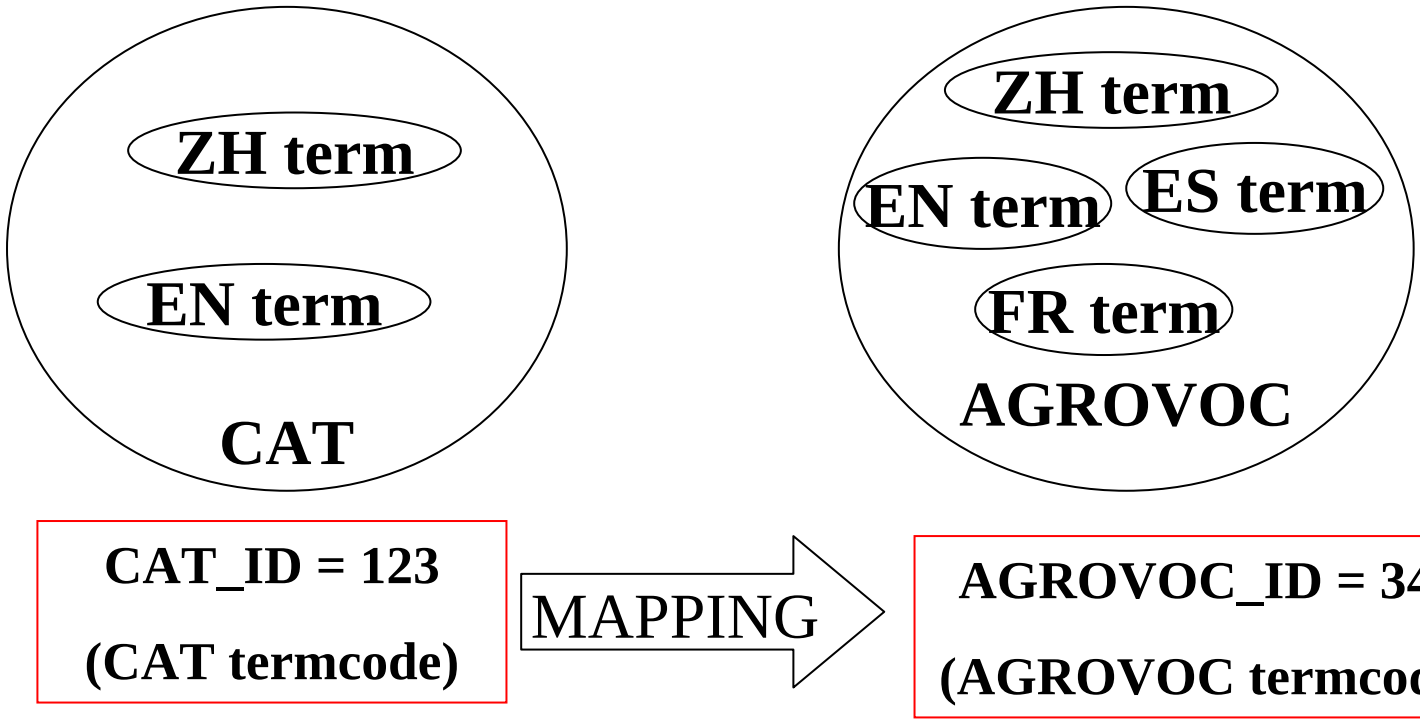
- The **source** vocabulary is CAT.
  The **target** vocabulary is AGROVOC.
- A **term** is a lexical representation of a concept.
- An **entry** in CAT consists of the Chinese term and any English translation(s) along with its relations to other entries.
  An **entry** in AGROVOC consists of at least one English or Chinese term along with their translations as well as its relations to other entries.
- **Mapping** means linking an entry in the source vocabulary to an entry in the target vocabulary.

# Definitions (2/2)

AGROVOC-CAT
Mapping project

4th NKOS Workshop

September 22, 2005 Vienna



# Mapping relationships

- Exact match
  - SKOS: exactMatch
  - OWL: equivalentTo, sameAs
- Broader/Narrower match
  - SKOS: broadMatch, narrowMatch
  - OWL: subClassOf
- OR, AND, NOT operators
  - SKOS: OR, AND, NOT
  - OWL unionOf, intersectionOf, complementOf
- Partial equivalences
  - SKOS:  minorMatch, majorMatch

# Pre-processing: Automatic identification of candidate exact matches





## Candidate exact matches

1. **Automatic** processing for matching terms in CAT and AGROVOC for both Chinese and English terms: **exact** match.
2. **Semi-automatic** processing to identify where either only the English or the Chinese terms match: **candidate exact** match + manual review for synonyms in the non-matching language.

# Pre-processing: Automatic identification of candidate exact matches

AGROVOC-CAT
Mapping project

4th
Wo
Septe
200

| | Num. | Taxon. | Geogr. | Total | Action |
|---|---|---|---|---|---|
| Match English only (NB: Chinese may or may not match) | 4013 | 2619 | 192 | 6826 | Min 8000 Exact match |
| Match Chinese only (NB: English may or may not match) | 5767 | 1952 | 331 | 8050 | |
| Match English and Chinese | 2470 | 1547 | 143 | 4160 | Exact match |
| Match English but different Chinese | 624 | 546 | 15 | 1187 | Match not ensured |
| Match Chinese but different English | 3297 | 405 | 188 | 3890 | Tentative exact match |

# Other preparatory steps



- Convert thesauri to RDF-based format (OWL, SKOS)
  - USE/UF relations
    - AGROVOC → rdfs:subClassOf or rdfs:synonym
    - CAT → rdfs:synonym
  - BT/NT
    - AGROVOC and CAT → rdfs:subClassOf
  - RT
    - AGROVOC → owl:RTag
    - CAT → owl:RTca

# Work Procedure (1/3)





- Split CAT into rough **domains** based on top terms (split in several files) extract corresponding sub-trees
- Start from automatically identified exact and candidate exact matches and proceed to those to be manually processed

# Work Procedure (2/3)





- For manually processed mappings,
  - these should ideally be between conceptually identical entities, where CAT:ZH ~AG:ZH are synonymous terms, *and* CAT:EN ~ AG:EN are synonymous terms.
  - otherwise, CAT:ZH and AG:EN are the languages in each of the respective thesauri determine the mappings.

# Work Procedure (3/3): Fix inconsistencies

AGROVOC-CAT
Mapping project

4th NKOS Workshop
September 22, 2005 Vienna

- In agrovoc:
  - Oryza (Chinese translation '稻属')
  - NT Oryza sativa (Chinese translation '稻')
  - RT Rice (Chinese translation '稻米' )
- In CAT
  - 稻 (English translation 'Oryza sativa')
  - NT 水稻 (English translation 'Rice')

# Tools (1/3)





- **Custom solution (Excel sheet / RDBMS)**
  - ☺ easy to design template;
  - ☹ needs scripts to redefine the mapping in useable format;
  - ☹ requires separate access to vocabularies;
  - ☹ slow process and cumbersome;

# Custom (manual) solution: Excel

AGROVOC-CAT
Mapping project

4th NKOS Workshop

September 22, 2005 Vienna

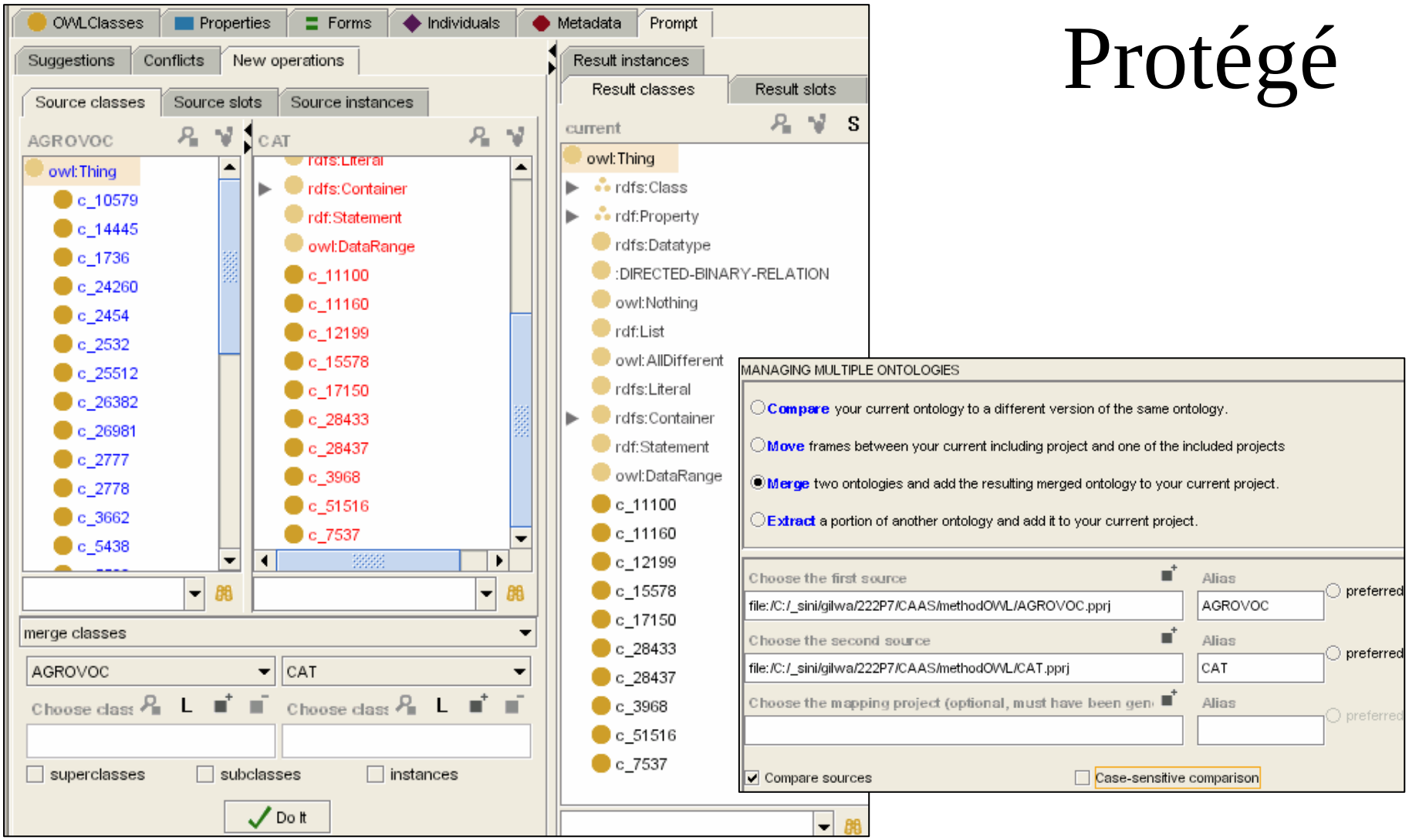
| | B | C | D | E | F | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|
| 1 | CAT-ID | CAT-ZH | CAT-EN | Relation | C-preferential | AG-ID | AG-EN | AG-ID | AG-EN |
| 2 | 17150 | 禾谷类作物 | | Exact | Descriptor | 25512 | Cereal crops | | |
| 3 | 15578 | 谷类作物 | Insect pests o | Exact | Non-Descriptor | 25512 | Cereal crops | | |
| 4 | 3968 | 穆子 | Eleusine coraca | Exact | Descriptor | 2532 | Eleusine coracana | | |
| 5 | 28433 | 龙爪稷 | Sophora japoni | Exact | Non-Descriptor | 10579 | Eleusine abyssinica | | |
| 6 | 28437 | 龙爪粟 | Dactyloctenium | Exact | Non-Descriptor | 2532 | Eleusine coracana | | |
| 7 | 51516 | 鸭脚粟 | Opisthorchis a | Exact | Non-Descriptor | 2532 | Eleusine coracana | | |
| 8 | 12199 | 非洲稷 | Genlisea afric | Exact | Non-Descriptor | 2532 | Eleusine coracana | | |
| 9 | 7537 | 大麦 | | Exact-OR | Descriptor | 3662 | Hordeum vulgar | 823 | Barley |
| 10 | 11100 | 二行大麦 | Two-row barley | NT1 | Descriptor | 3662 | Hordeum vulgare | | |
| 11 | 11160 | 二棱大麦 | Two-shear shee | NT1 | Non-Descriptor | 3662 | Hordeum vulgare | | |

# Tools (2/3)



- **Protégé / Prompt**
  - ☺ creates an owl mapping file;
  - ☺ automatic suggestions;
  - ☹ performance;
  - ☹ needs modifications;

# Protégé


OWLClasses
Properties
Forms
Individuals
Metadata
Prompt
Suggestions
Conflicts
New operations
Source classes
Source slots
Source instances
AGROVOC
owl:Thing
c_10579
c_14445
c_1736
c_24260
c_2454
c_2532
c_25512
c_26382
c_26981
c_2777
c_2778
c_3662
c_5438
CAT
rdfs:Literal
rdfs:Container
rdf:Statement
owl:DataRange
c_11100
c_11160
c_12199
c_15578
c_17150
c_28433
c_28437
c_3968
c_51516
c_7537
merge classes
AGROVOC
CAT
Choose class
superclasses
subclasses
instances
Do It
Result instances
Result classes
Result slots
current
owl:Thing
rdfs:Class
rdf:Property
rdfs:Datatype
:DIRECTED-BINARY-RELATION
owl:Nothing
rdf:List
owl:AllDifferent
rdfs:Literal
rdfs:Container
rdf:Statement
owl:DataRange
c_11100
c_11160
c_12199
c_15578
c_17150
c_28433
c_28437
c_3968
c_51516
c_7537
MANAGING MULTIPLE ONTOLOGIES
Compare your current ontology to a different version of the same ontology.
Move frames between your current including project and one of the included projects
Merge two ontologies and add the resulting merged ontology to your current project.
Extract a portion of another ontology and add it to your current project.
Choose the first source
Alias
preferred
file:/C:/_sini/gilwa/222P7/CAAS/methodOWL/AGROVOC.pprj
AGROVOC
Choose the second source
Alias
preferred
file:/C:/_sini/gilwa/222P7/CAAS/methodOWL/CAT.pprj
CAT
Choose the mapping project (optional, must have been gen
Alias
preferred
Compare sources
Case-sensitive comparison

# Tools (3/3)



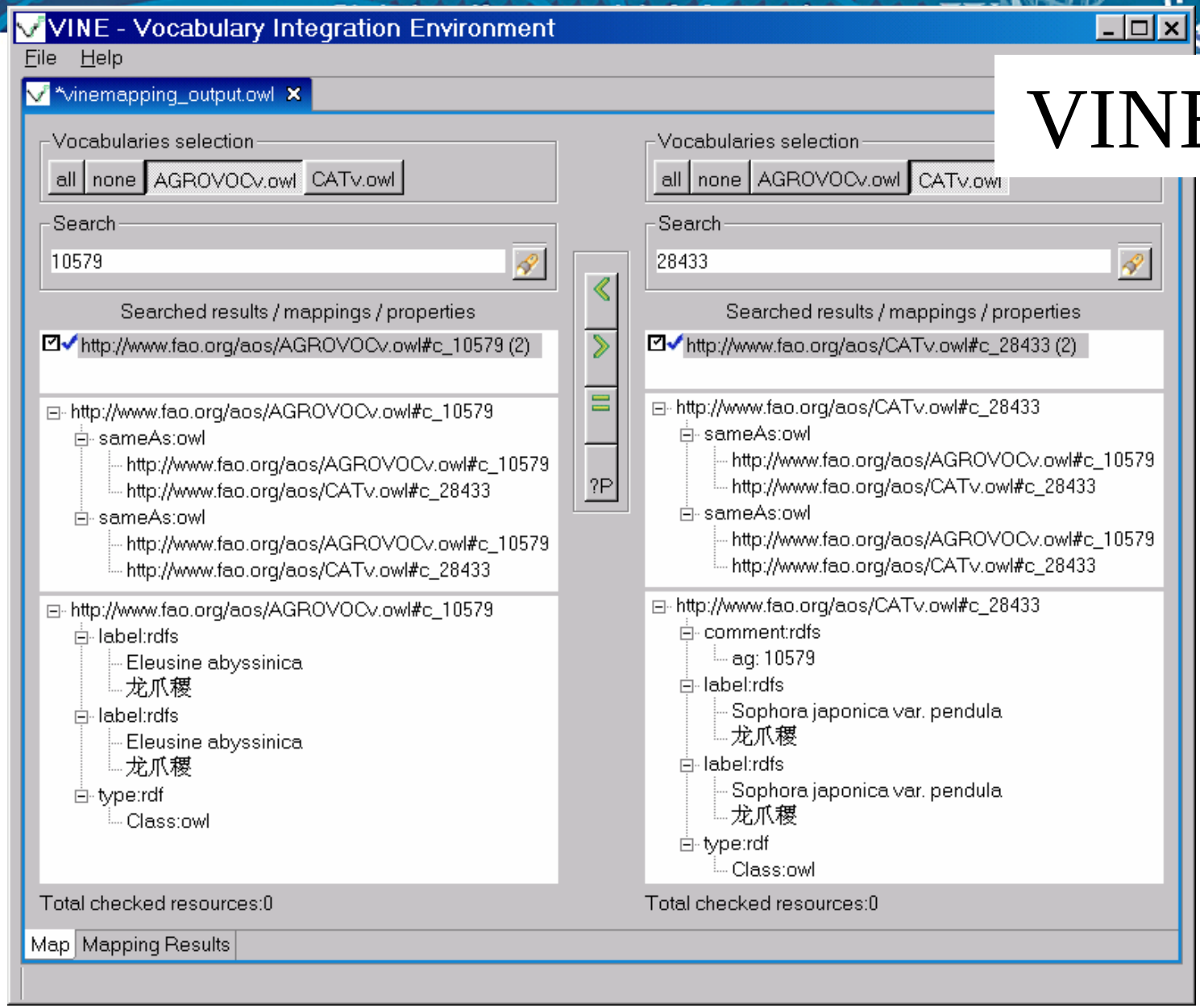
- **VINE** (Vocabulary Integration Environment)
  - ☺ creates an owl mapping file;
  - ☺ seems easy to make mapping (based on initial tests with small files);
  - ☹ user-defined relationships have to be re-keyed;
  - ☹ performance for large files reported to be poor;
  - ☹ problems reported in Windows environment, okay with Apple;
  - ☹ needs modifications;

# VINE

**AGROVOC-CAT Mapping project**

**4th NKOS Workshop**

**September 22, 2005 Vienna**



# Using Protégé (1/2)





- Set up Protégé with Prompt plug-in. Set journaling ON.
- Load headers file.
- Load CAT.pprj and AGROVOC.pprj. Set X as the preferred ontology.
- Don’t do multiple copy/merge operations at once.

# Using Protégé (2/2)





- Mappings will occur as follows in the resulting mapping ontology:
  - Mapping to single concepts in the target vocabulary
    - narrow match: subClassOf
    - broad match: super class
  - Mapping to multiple concepts requires the formulation of restrictions (anonymous classes)
    - CAT:A → AG:B AND AG:C: A subClassOf (B ∩ C)
    - CAT:A → AG:B OR AG:C: A subClassOf (B ∪ C)
    - CAT:A → AG:B AND NOT AG:C: A subClassOf (B ∩ ~C)

# Example Mapping

**AGROVOC-CAT**
**Mapping project**

| CAT-ID | CAT-ZH | CAT-EN | Map | AG-ID | AG-EN | AG-ID | AG-EN |
|---|---|---|---|---|---|---|---|
| 30854 | 糜子 | Senta flammea | Exact | 9748 | Cheena | | |
| 21596 | 稷 | Cneoranidea signatipes | Exact | 9453 | Broom corn (millet) | | |
| 50008 | 小麦×黑麦 | Mayetiola destructor | Exact-OR | 24260 | Triticale (gramineae) | 7949 | Triticales (product) |
| 49901 | 小麦 | | Exact-OR | 7950 | Triticum | 8373 | Wheats |

# Implications





- Training needed (Infrastructure set-up, methodology, etc.)
- Sources affected
- Quality can be monitored via queries
- Versioning:
  - Work should be done using the latest version of AGROVOC.
  - The mapping will be assigned a version.

# Conclusion

AGROVOC-CAT
Mapping project

- Valid idea and lots of benefits
- Finalization of the procedures
- Identify best tool to use

**AGROVOC-CAT**
**Mapping project**

Food and Agriculture Organization (FAO)
and
Chinese Academy of Agricultural Sciences (CAAS)

**4th NKOS Workshop**

**September 22, 2005 Vienna**

**anita.liang@fao.org**
**margherita.sini@fao.org**
**changc@mail.caas.net.cn**

# Thank you.