UiO : University of Oslo Library

EDUG mapping workshop, Naples, April 15 2015
Are D. Gulbrandsen & Dan Michael O. Heggø

The mapping tool ccmapper (concept context mapper)

UiO **: University of Oslo Library**

Presentation overview

- Guidelines in ISO 25964-2 on computer assisted matching
- Overall goals for ccmapper
- System architecture
- Finding candidate mappings
- Interaction design demo

Guidelines in ISO 25964-2 on computer assisted matching

- Candidate mappings should be reviewed by an expert
  - In ranked order
  - View of full context for concepts
    - Source vocabulary
    - Target vocabulary
  - Support in selecting the appropriate type of mapping

Notes for previous slide

In ch 14 "Techniques for identifying candidate mappings"
the standard doesn't specifically recommend computer assisted matching
As it says that
"Traditionally the identification of mappings is an intellectual process." (on p 38)

On the other hand it provides guidelines in ch 14.2 "Computer assisted direct matching" (p 40)
"It is sometimes possible to automate the process, at least in part, using a matching algorithm."

Citations from ISO 25964-2 ch 14.2, p 40:
The candidate mappings identified by the matching processes described should be assembled for review by an expert.

For each concept in the source vocabulary, the expert should be able to view the complete record (including scope note, broader and narrower terms).

The viewing interface should make it easy to check the complete context of each concept identified in the target vocabulary.

It should also support the expert in selecting the appropriate type of mapping for the candidate(s) he approves.

Only after an expert has reviewed them should any of the candidate mappings be approved and established.

Source vocabulary	Candidate Mappings	Target Vocabulary
Concept and context		Concept and context

Notes for previous slide

Our interpretation of the ISO guidelines.

Concept context mapper (Ccmapper) will support mapping in two main ways:
- By suggesting the most relevant candidate mappings for a given concept
- By providing a good overview of the contexts associated with concepts in both the source and target vocabularies

We have chosen the user interface metaphor Dashboard when developing the tool.
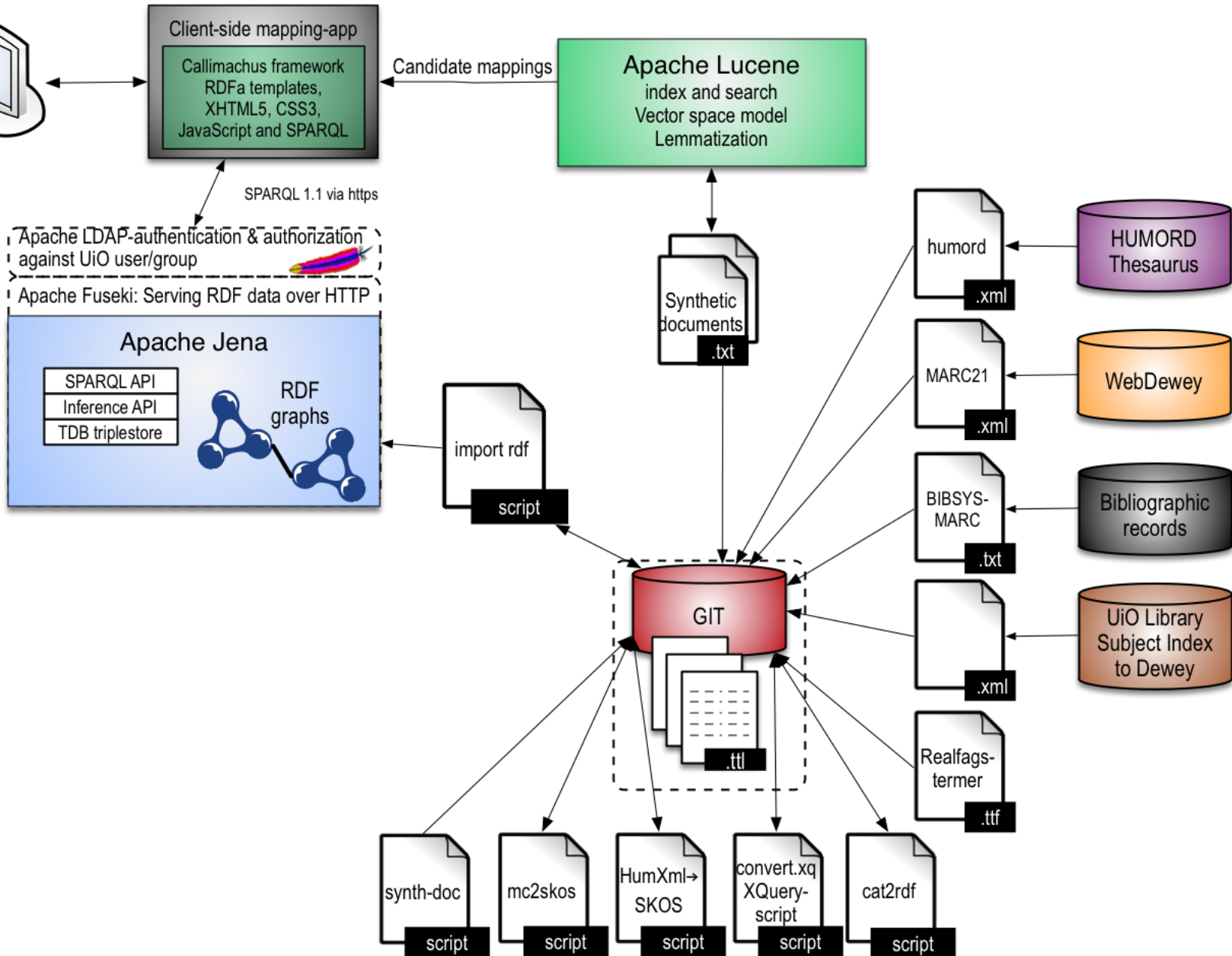The user interface will have to integrate information from several systems:
- Webdewey
- The Humord thesaurus
- Catalog data
- The UiO Library Subject Index to Dewey

We think good interaction design will be an important success factor.

Overall goals for ccmapper

- System integration
  - One dashboard, not four separate systems
- Time saving for the mapping expert
- Consistency between mappers
- Linked Data
  - System integration
  - Share information
- Universal applicability
  - (with DDC as target vocabulary)

UiO : **University of Oslo Library**

Client-side mapping-app

Callimachus framework
RDFa templates,
XHTML5, CSS3,
JavaScript and SPARQL

Candidate mappings

Apache Lucene
index and search
Vector space model
Lemmatization

SPARQL 1.1 via https

Apache LDAP-authentication & authorization
against UiO user/group

Apache Fuseki: Serving RDF data over HTTP

Apache Jena

SPARQL API
Inference API
TDB triplestore

RDF
graphs

Synthetic
documents
.txt

humord
.xml

HUMORD
Thesaurus

import rdf
script

MARC21
.xml

WebDewey

BIBSYS-
MARC
.txt

Bibliographic
records

GIT

.ttl

UiO Library
Subject Index
to Dewey
.xml

Realfags-
termer
.ttf

synth-doc
script

mc2skos
script

HumXml→
SKOS
script

convert.xq
XQuery-
script
script

cat2rdf
script

Finding a mapping algorithm (1)

Source

- Preferred term

  – String equality
  – Meaning

Target

- Heading (caption)

Notes for previous slide

String comparison of preferred terms seems to be the most common strategy.
- Edit distance can give measure of string equality.

Common challenges:
- Synonyms (different terms used to name the same entity)
- homonyms (same terms used for different entities).

Basic string Normalisation:
- Case normalisation
- Diacritics suppression

- "Words do not have meanings; people have meanings for words" The linguist Nelson Francis (1967)
- Context give meaning to words.
Both statements are true.

ISO 25964-2 (p 32)
Mappings to or from a class or category in a monohierarchical scheme should treat the class/category as a pre-coordinated concept whose meaning can be established by inspecting all its superordinate and subordinate classes as well as any scope notes associated with it.
Inspection of the caption alone is inadequate.

We're mapping concepts, not terms.
Context is important.

Finding a mapping algorithm (2)

Source

- Preferred term
- Context
  - Non-preferred terms?
  - Parent concepts?
  - Definition?
  - Statistical mapping (analysis of bibliographic records)?

Target

- Class number
- Heading (caption)
- Context
  - Relative index terms?
  - Notes?
  - Parent classes
    - Heading
    - Notes?
    - Relative index terms?
  - UiO library subject index to Dewey

Notes for previous slide

Mappings to or from a class or category in a monohierarchical scheme should treat the class/category as a pre-coordinated concept whose meaning can be established by inspecting all its superordinate and subordinate classes as well as any scope notes associated with it.

Inspection of the caption alone is inadequate.

(ISO 25964-2 p 32)

Equality between two "Bag of words".

Each context can be seen as a separate component in deciding concept equivalence and should be individually weighted.

Our approach is the "Vector space model", used in most modern web search engines.

The method will also be able to integrate statistically based mapping.

Statistical co-occurrence mapping

- A subject heading and a DDC number occurring on the same bibliographic records form a pair
- If the pair occurs relatively frequently compared to how often the subject heading and the DDC number themselves occur, we might suggest a mapping candidate.
  - (In practice we might use some statistical measure such as the association ratio or the log-likelihood ratio to find the association)

Vector space model and search

- Synthetic documents
  - Synthesize a text document representing every concept in source vocabulary
    - Concept = term and context
    - All components weighted individually
  - Synthesize a text document representing every concept in target vocabulary
    - Concept = term and context
    - All components weighted individually
- Linguistic normalization

Notes for previous slide

Standard search technology:

- Documents and queries as vectors:
  http://nlp.stanford.edu/IR-book/html/htmledition/queries-as-vectors-1.html
- Tf-idf weighting:
  http://nlp.stanford.edu/IR-book/html/htmledition/tf-idf-weighting-1.html

In combination with Linguistic normalization (morphological normalization of terms):

- Elimination of stop words
- Lemmatization
- Compound splitting (splitting of compound words

The article Monolingual document retrieval for European languages (Hollink, Kamps, Monz, & De Rijke, 2004) shows that compound splitting for Swedish increases mean average precision by 25 % compared to standard lemmatization. We are, however, unsure as to whether compound splitting can be achieved in this project.

We must use a flexible approach:
The method must handle mapping a concept with only one preferred term and no more context (science terms)
The method must handle concepts with a wide variety of contexts:

- Non-preferred terms
- Hierarchical context (parent concepts and child concepts)
- Definition
- Notes

We should be able to combine the method with statistical based mapping

Ccmapper interaction design demo

Following images replaces demo.

UiO : University of Oslo Library

Humord Thesaurus	WebDewey	

Horses   **Next**

Library records with horses in subject heading (opens in new window)

Horses	Search

UiO Library Subject Index to Dewey

Horses < Zoology
599.6655

Horses < Animal husbandry
636.1

Historic treatment < Horses < Animal husbandry
636.1009

Pedigrees < Horses < Animal husbandry
636.10822

Studbooks < Horses < Animal husbandry
636.10822

Harnesses < Horses < Animal husbandry
636.10837

Mapping Candidates   Search WebDevey

T3C--36296655 Horses—arts

Arts and literature dealing with specific themes and subjects > Natural and physical phenomena; mathematics > Animals > Mammals--arts
Horses—arts.

=EQ	~EQ	BM	NM	RM	Reject

Comments

599.6655 *Equus caballus (horse)

Animals (zoology) > *Mammalia (Mammals) > *Ungulates > *Perissodactyla (Odd-toed ungulates) > *Equidae
Equus caballus, **Horses**--zoology, Mustang--zoology, Przewalski's horse, Wild horse.

=EQ	**~EQ**	BM	NM	RM	Reject

Should this be =EQ or ~EQ?

636.1 Horses

Agriculture > Animal husbandry > Specific kinds of domestic animals
Horses, Horses--animal husbandry.

=EQ	~EQ	BM	NM	RM	Reject

Comments

005.84 *Malware

Computer programming, programs, data > Data security
Computer viruses, Malware, Spyware
Trojan **horses** (Computer security), Viruses (Computer security), Worms (Computer security).

=EQ	~EQ	BM	NM	RM	**Reject**

Comments

Save

599.6655 *Equus caballus (horse)

Relative Index Terms:
Equus caballus, Horses--zoology, Mustang--zoology, Przewalski's horse, Wild horse

Class here: Mustang, Przewalski's horse, wild horse

Class interdisciplinary works on horses in 636.1

Add To Mapping Candidates

Inherited Notes   Subordinate Classes

Notes from parent classes:

599.665 *Equidae
Including: asses
Class here: Equus
Class interdisciplinary works on asses in 636.18

599.66 *Perissodactyla (Odd-toed ungulates)
Including: Tapiridae (tapirs)

599.6 *Ungulates
Class here: hoofed mammals, comprehensive works on big game animals
Class Sirenia in 599.55
Class big game hunting in 799.26
For a specific kind of nonungulate big game animal, see the kind, e.g., bears 599.78

599 *Mammalia (Mammals)
Class here: warm-blooded vertebrates, Eutheria (placental mammals)
Class interdisciplinary works on species of domestic mammals in 636
For Aves, see 598
See Manual at 599
History:
Interdisciplinary works on species of domestic mammals relocated to 636 1996, Edition 21.

UiO : University of Oslo Library

Humord Thesaurus

Natural Sciences > Biology > Zoology > Animals > Vertebrates > Mammals > Horses

Horses

[Next]

Library records with horses in subject heading (opens in new window)

| Horses | [Search] |

UiO Library Subject Index to Dewey

Horses < Zoology
599.6655

Horses < Animal husbandry
636.1

Historic treatment < Horses < Animal husbandry
636.1009

Pedigrees < Horses < Animal husbandry
636.10822

Studbooks < Horses < Animal husbandry
636.10822

Harnesses < Horses < Animal husbandry
636.10837

WebDewey

Mapping Candidates | Search WebDevey

| Horses | [Search] |

T3C--36296655 Horses—arts
Mammals--arts
Horses—arts.

005.84 *Malware
Data security
Computer viruses, Malware, Spyware
Trojan **horses** (Computer security), Viruses (Computer security), Worms (Computer security).

345.026286361 Horse stealing--law, . . .
Theft--law
Horse stealing--law, **Horses**--theft of--law.

357.2 Remount services
Mounted forces and warfare
Horses--remount services, Remount services.

364.16286361 Horse stealing, . . .
Theft of specific items
Horse stealing, **Horses**--theft of.

597.6798 *Sea horses
*Syngnathidae
Hippocampus, Sea **horses**, Seahorses.

599.6655 *Equus caballus (horse)
*Equidae
Equus caballus, **Horses**--zoology, Mustang--zoology, Przewalski's horse, Wild horse.

615.851581 Therapeutic use of horses
Therapeutic use of animals
Horsemanship--therapeutic use, **Horses**--therapeutic use.

636.1 Horses
Specific kinds of domestic animals
Horses, **Horses**--animal husbandry.

798.23 Riding

Agriculture > Animal husbandry > Specific kinds of domestic animals

636.1 Horses

Relative Index Terms:
Horses, Horses--animal husbandry

Class here: Class here equines

[Add To Mapping Candidates]

Inherited Notes | Subordinate Classes

Notes from parent classes:

636.1-636.8 Specific kinds of domestic animals
Except for modifications shown under specific entries, add to each subdivision identified by * as follows:
01   Philosophy and theory
[015765] Genetics
          Do not use; class in notation 2 from this table
02-06   Standard subdivisions
07   Education, research, related topics
079   Competitions, awards, financial support
          Do not use for animal shows and related awards; class in notation 1 from this table
08-09   Standard subdivisions
1   Showing
          Class here judging
2   Breeding
22   Breeding records
          Class here origin of the breed or breeds; herdbooks, pedigrees, studbooks
3   Care, feeding, training, veterinary medicine
35   Training
39   Veterinary medicine

Class comprehensive works in 636

See Manual at 636.1-636.8 vs. 636.088

636 Animal husbandry
Class here interdisciplinary works on species of domestic mammals

History

UiO : University of Oslo Library

Humord Thesaurus

Natural Sciences > Biology > Zoology > Animals > Vertebrates > Mammals

Horses   [Next]

Library records with horses in subject heading (opens in new window)

[Horses   ] [Search]

UiO Library Subject Index to Dewey

Horses < Zoology
599.6655

Horses < Animal husbandry
636.1

Historic treatment < Horses < Animal husbandry
636.1009

Pedigrees < Horses < Animal husbandry
636.10822

Studbooks < Horses < Animal husbandry
636.10822

Harnesses < Horses < Animal husbandry
636.10837

WebDewey

[**Mapping Candidates**] [**Search WebDevey**]

[Horses   ] [Search]

T3C--36296655 Horses—arts
Mammals--arts
Horses—arts.

005.84 *Malware
Data security
Computer viruses, Malware, Spyware
Trojan **horses** (Computer security), Viruses (Computer security), Worms (Computer security).

345.026286361 Horse stealing--law, . . .
Theft--law
Horse stealing--law, **Horses**--theft of--law.

357.2 Remount services
Mounted forces and warfare
Horses--remount services, Remount services.

364.16286361 Horse stealing, . . .
Theft of specific items
Horse stealing, **Horses**--theft of.

597.6798 *Sea horses
*Syngnathidae
Hippocampus, Sea **horses**, Seahorses.

599.6655 *Equus caballus (horse)
*Equidae
Equus caballus, **Horses**--zoology, Mustang--zoology, Przewalski's horse, Wild horse.

615.851581 Therapeutic use of horses
Therapeutic use of animals
Horsemanship--therapeutic use, **Horses**--therapeutic use.

636.1 Horses
Specific kinds of domestic animals
Horses, **Horses**--animal husbandry.

798.23 Riding

WebDewey (right panel)

Agriculture > Animal husbandry > Specific kinds of domestic animals

636.1 Horses

Relative Index Terms:
Horses, Horses--animal husbandry

Class here: Class here equines

[Add To Mapping Candidates]

[**Inherited Notes**] [**Subordinate Classes**]

636.1001-636.1009 Standard subdivisions
Notation from Table 1 as modified under 636.001-636.009.

636.101-636.108 Specific topics in husbandry of horses
Add to base number 636.10 the numbers following 636.0 in 636.01-636.08, e.g., housing horses 636.10831; however, for racehorses, see 636.12 for riding horses, see 636.13 for draft horses, see 636.15. **Class** training of riders and drivers, comprehensive works on training horses and their riders and drivers in 798.

636.109 Miniature horses
Regardless of breed.
History: Formerly located in 636.11-636.17 (1996, Edition 21)

636.11-636.17 Specific breeds and kinds of horses
Class miniature horses regardless of breed in 636.109.
Class comprehensive works in 636.1.
History: Miniature horses regardless of breed relocated to 636.19 (1996, Edition 21)

636.18 Other equines
Including zebras.
Relative Index Terms:
Asses, Asses--animal husbandry, Zebras--animal husbandry.

UiO : University of Oslo Library

Humord Thesaurus

Natural Sciences > Biology > Zoology > Animals > Vertebrates > Mammals > Horses

Horses   [Next]

[Library records with horses in subject heading (opens in new window)](#)

| Horses | [Search] |

UiO Library Subject Index to Dewey

Horses < Zoology
599.6655

Horses < Animal husbandry
636.1

Historic treatment < Horses < Animal husbandry
636.1009

Pedigrees < Horses < Animal husbandry
636.10822

Studbooks < Horses < Animal husbandry
636.10822

Harnesses < Horses < Animal husbandry
636.10837

WebDewey

| **Mapping Candidates** | **Search WebDevey** |

| Horses | [Search] |

T3C--36296655 Horses—arts
Mammals--arts
Horses—arts.

005.84 *Malware
Data security
Computer viruses, Malware, Spyware
Trojan **horses** (Computer security), Viruses (Computer security), Worms (Computer security).

345.026286361 Horse stealing--law, . . .
Theft--law
Horse stealing--law, **Horses**--theft of--law.

357.2 Remount services
Mounted forces and warfare
Horses--remount services, Remount services.

364.16286361 Horse stealing, . . .
Theft of specific items
Horse stealing, **Horses**--theft of.

597.6798 *Sea horses
*Syngnathidae
Hippocampus, Sea **horses**, Seahorses.

599.6655 *Equus caballus (horse)
*Equidae
Equus caballus, **Horses**--zoology, Mustang--zoology, Przewalski's horse, Wild horse.

615.851581 Therapeutic use of horses
Therapeutic use of animals
Horsemanship--therapeutic use, **Horses**--therapeutic use.

636.1 Horses
Specific kinds of domestic animals
Horses, **Horses**--animal husbandry.

798.23 Riding

WebDewey — detail

T3C--36296655 Horses—arts

Components: [T3C--362](#) + [599.6655](#)

Relative Index Terms:
Horses—arts

[Add To Mapping Candidates]

| **Inherited Notes** | **Subordinate Classes** |

Notes from parent classes:

T3C--362 Animals
Class here pets
Add to base number T3C--362 the numbers following 59 in 592-599, e.g., cats T3C--3629752
Class supernatural, mythological, legendary animals in T3C--374
See Manual at 800, T3C--362 vs. 398.245, 590, 636
History: expanded from T3C--36 1996, Edition 21

T3C--36 Natural and physical phenomena; mathematics
Standard subdivisions are added for natural and physical phenomena, mathematics together; for natural and physical phenomena together; for natural phenomena alone
Including fire, weather
Class here landscapes [formerly T3C--32], agriculture, nature, scientific themes, comprehensive works on scientific and technical themes
Class science fiction in literature in T3B--308762 in Table 3B
Class science fiction in the arts in T3C--15
Class times (e.g., seasons, times of day) in T3C--33
Class specific persons associated with natural and physical phenomena, with mathematics in T3C--351
Class animals or plants for a specific purpose with the purpose, e.g., medicinal plants T3C--3561, meat T3C--3564
For human anatomy, physiology, diseases, see