Collaborative Building of Controlled Vocabularies Crosswalks

Mateusz Kaczmarek, Sebastian R. Kruk, Adam Gzella
Digital Enterprise Research Institute
National University of Ireland, Galway

mateusz.kaczmarek@deri.org
sebastian.kruk@deri.org
adam.gzella@deri.org

Outline

- Motivations & problem statement
- JOnto framework overview
- Crosswalks algorithm

Controlled vocabularies nowadays

- Used for annotating libraries resources with subject headings and thesauri
- Amibguity reduced as each concept is described by one term
- Relations such as ‘narrower’, ‘broader’ and ‘related to’ adopted in many controlled vocabularies (but – independently)

Problems

- No explicit relations between concepts from different taxonomies
- As taxonomies grew larger and larger – it became pretty hard to do it manually
- Equivalent terms on different levels of hierarchy or differently spelled

↓

As a result – efficient information exchange between different entities is difficult

How could that be resolved?

- Find relations manually?
- Rebuild all taxonomies?
- Find it automatically using users community’s support!

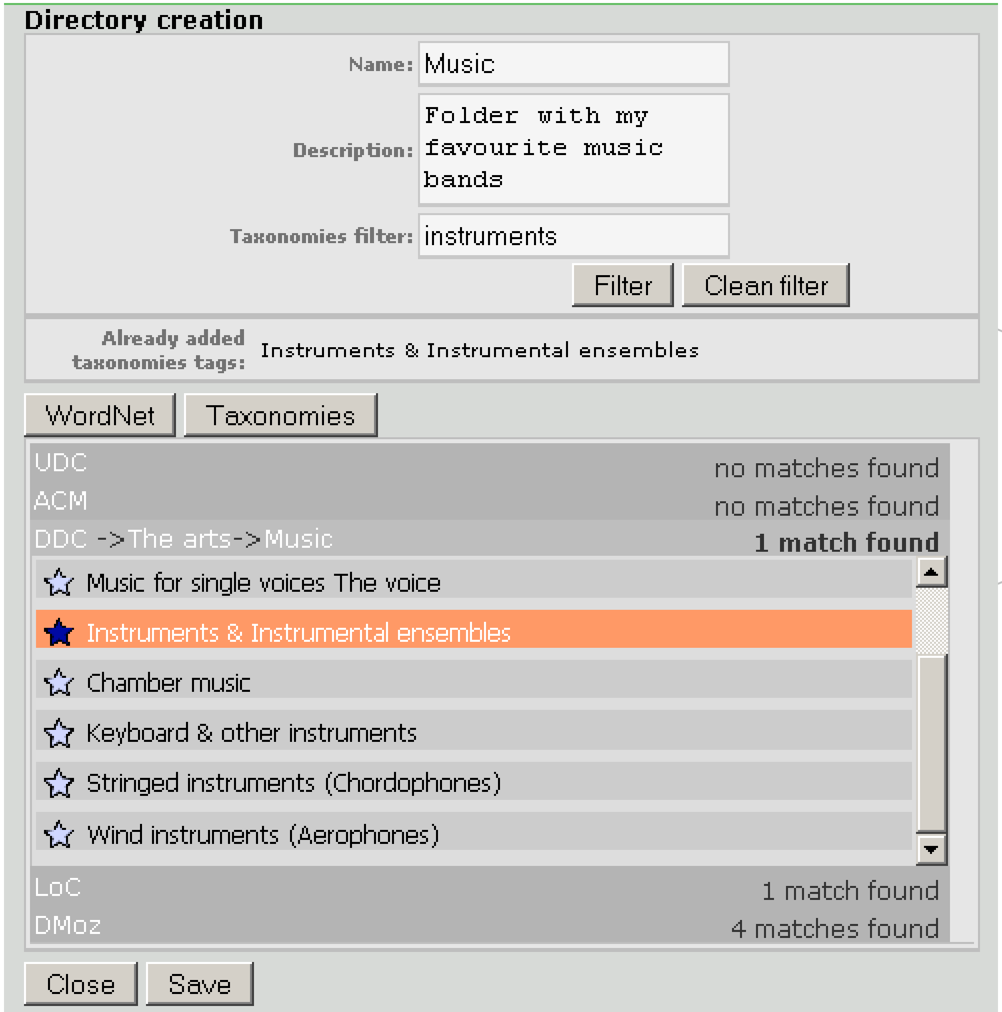
JOnto overview

- Tool for making annotations to resources
- Many taxonomies to choose from
- Java API for collecting RDF data
- Clear AJAX interface

- Developed in DERI since 2006

Where is it currently used

- JeromeDL – a social semantic digital library
- notitio.us – a tool for semantic information discovery, browsing and sharing



Directory creation

Name: Music

Description: Folder with my favourite music bands

Taxonomies filter: instruments

Already added taxonomies tags: Instruments & Instrumental ensembles

UDC no matches found

ACM no matches found

DDC ->The arts->Music **1 match found**

- ☆ Music for single voices The voice
- ★ Instruments & Instrumental ensembles
- ☆ Chamber music
- ☆ Keyboard & other instruments
- ☆ Stringed instruments (Chordophones)
- ☆ Wind instruments (Aerophones)

LoC 1 match found

DMoz 4 matches found

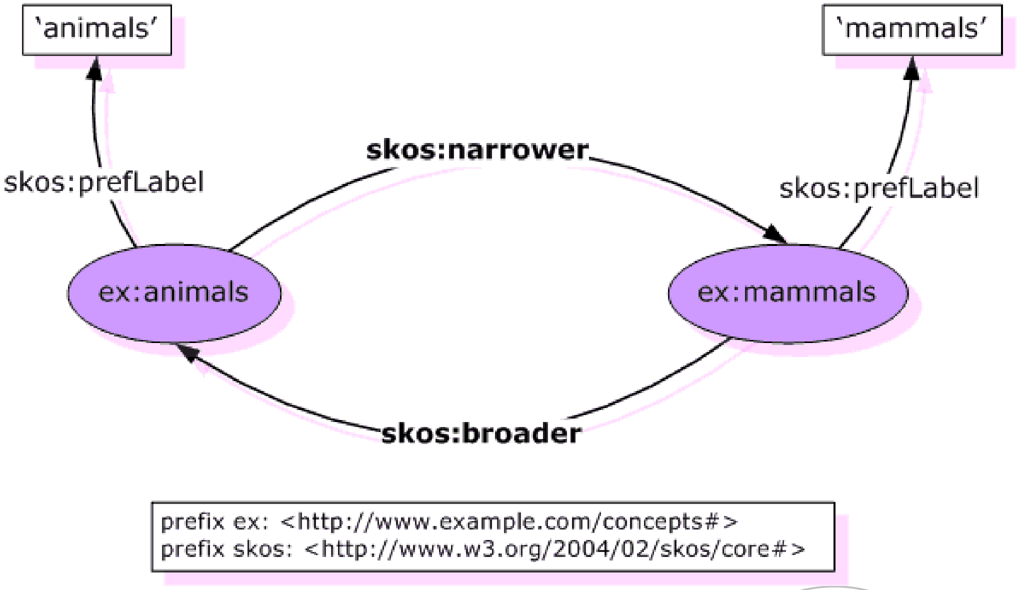
JOnto – implemented taxonomies

- DMoz (Open Directory Project Taxonomy – three levels)
- DDC (Dewey Decimal Classification)
- ACM (Association for Computing Machinery Classification)
- UDC (Universal Decimal Classification – outline)
- LCC (Library of Congress Classification – outline)

- WordNet 2.1

JOnto – implementation details

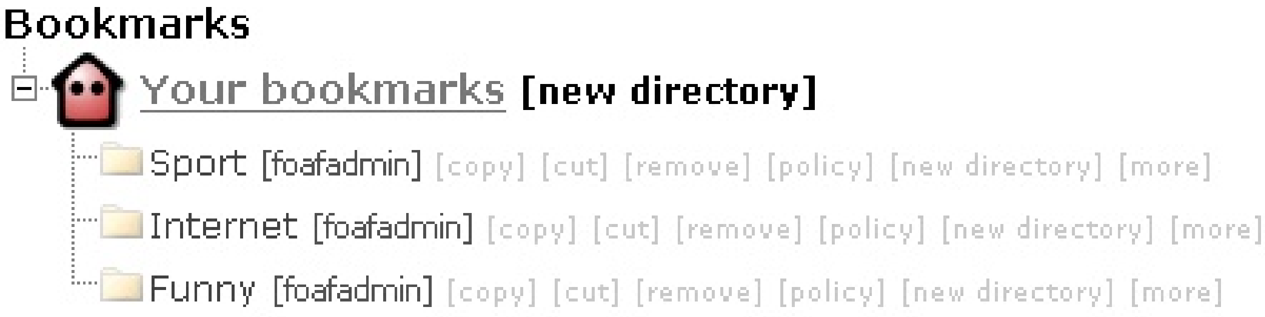
- All taxonomies are stored in Sesame RDF database
- Support for SKOS specification was implemented



```
prefix ex: <http://www.example.com/concepts#>
prefix skos: <http://www.w3.org/2004/02/skos/core#>
```

JOnto use case in notitio.us

1. User wants to create a new directory for his bookmarks

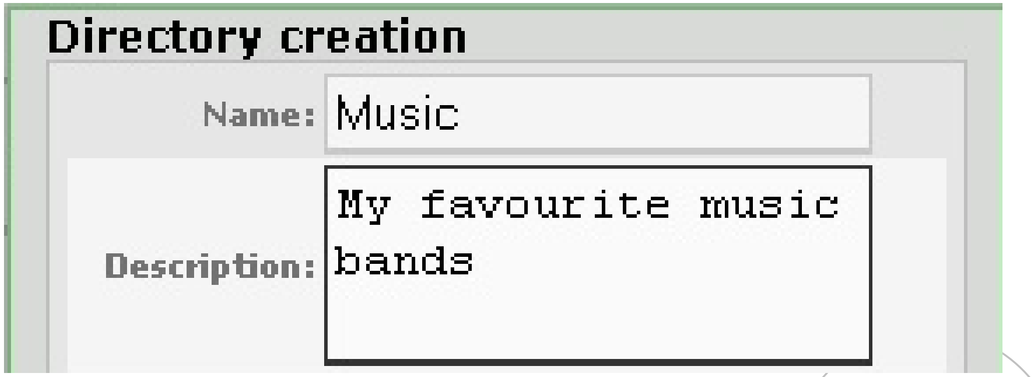


Bookmarks

Your bookmarks **[new directory]**

- Sport [foafadmin] [copy] [cut] [remove] [policy] [new directory] [more]
- Internet [foafadmin] [copy] [cut] [remove] [policy] [new directory] [more]
- Funny [foafadmin] [copy] [cut] [remove] [policy] [new directory] [more]

2. He writes a directory’s name and general description



Directory creation

Name: Music

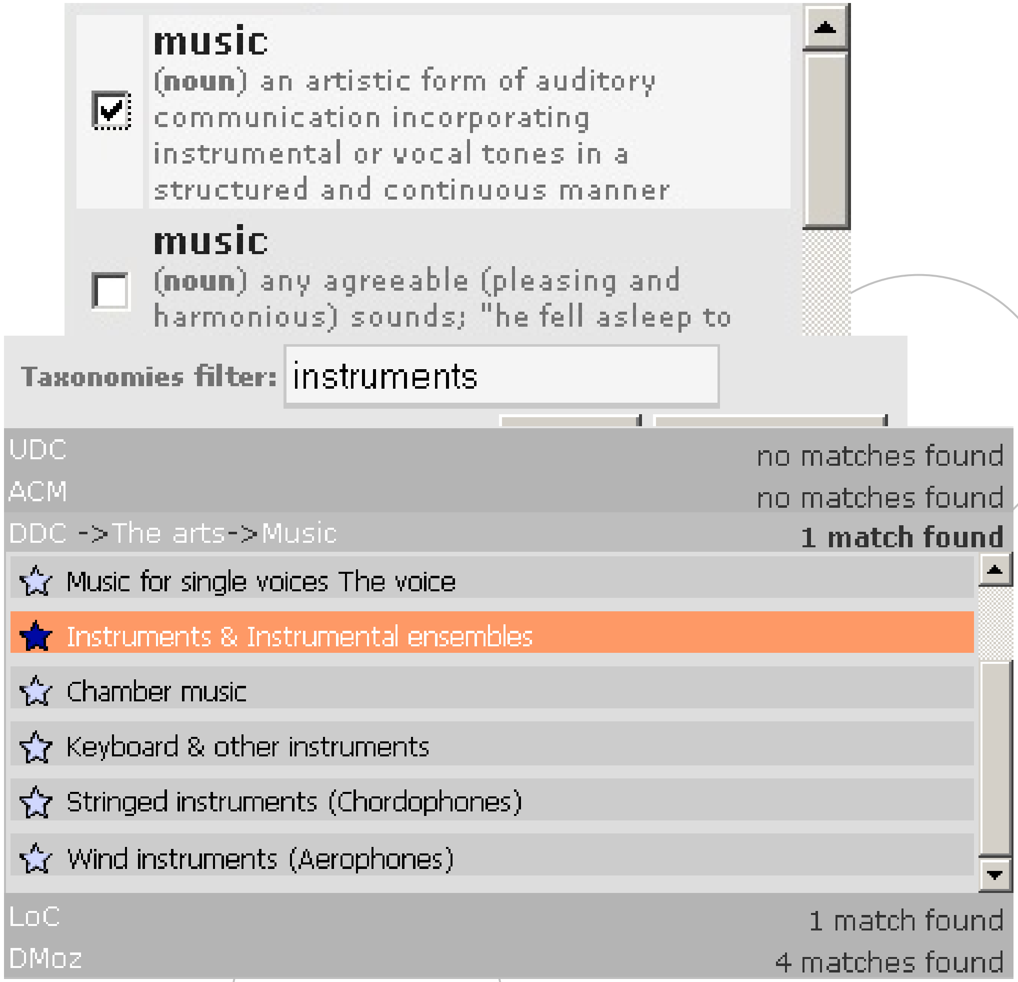
Description: My favourite music bands

JOnto use case in notitio.us

3. The name’s meaning is chosen from WordNet

4. Taxonomy categories may be filtered

5. Tags annotating the directory are chosen from different taxonomies



☑ **music**
(**noun**) an artistic form of auditory communication incorporating instrumental or vocal tones in a structured and continuous manner

□ **music**
(**noun**) any agreeable (pleasing and harmonious) sounds; "he fell asleep to

Taxonomies filter: instruments

UDC no matches found
ACM no matches found
DDC ->The arts->Music **1 match found**

☆ Music for single voices The voice
★ Instruments & Instrumental ensembles
☆ Chamber music
☆ Keyboard & other instruments
☆ Stringed instruments (Chordophones)
☆ Wind instruments (Aerophones)

LoC 1 match found
DMoz 4 matches found

Crosswalks algorithm – direct motivations

- Users should tag resources using concepts from different taxonomies

BUT

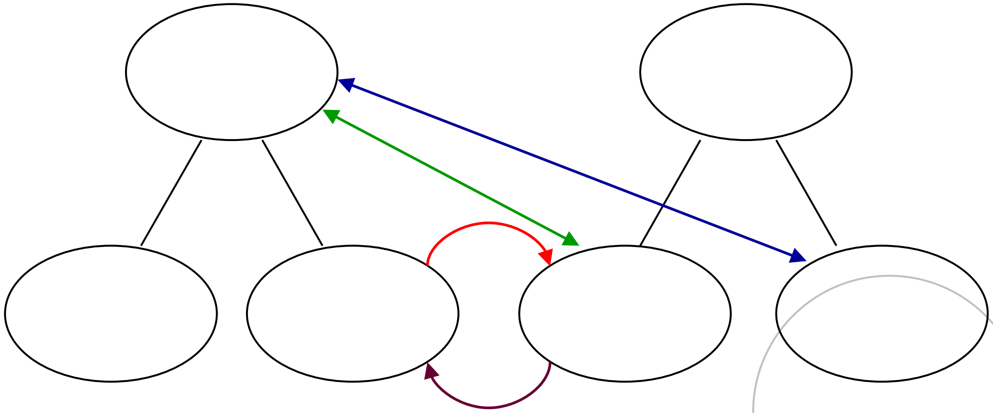
- Users are lazy – they don’t want to browse taxonomies
- Taxonomies are pretty large (e.g. DMoz – 8308 entries)
- Concepts in different taxonomies are labelled differently

THAT’S WHY

- Users should get suggestions on related concepts automatically

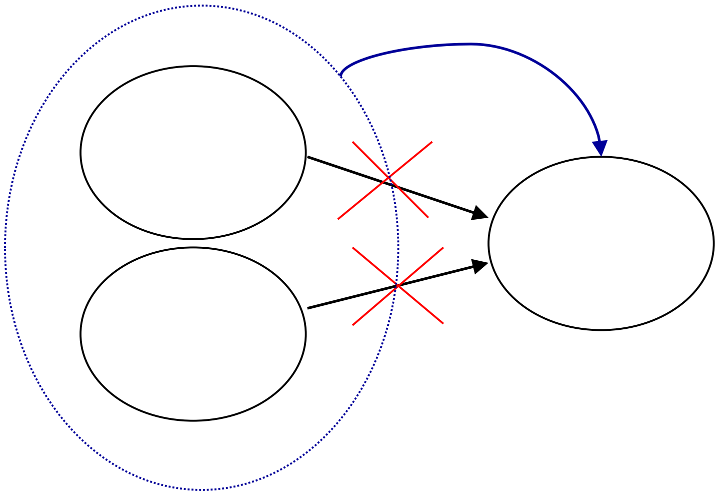
Crosswalks algorithm – types of relations

- Equivalent term
- Synonymous term
- Related term
- Narrower term
- Broader term
- Alternative language term (?)



Crosswalks algorithm – additional features

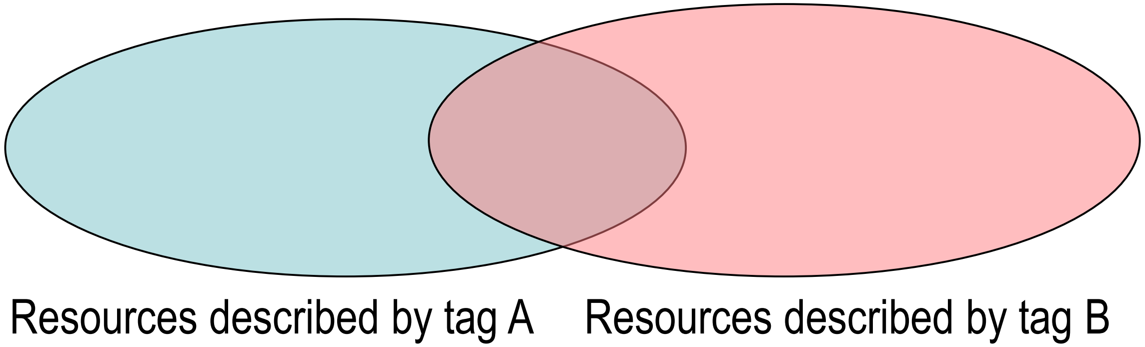
- „Context” notion



- Accuracy property

Crosswalks algorithm – first step

- Finding relations in general
- Determining common part of two sets:



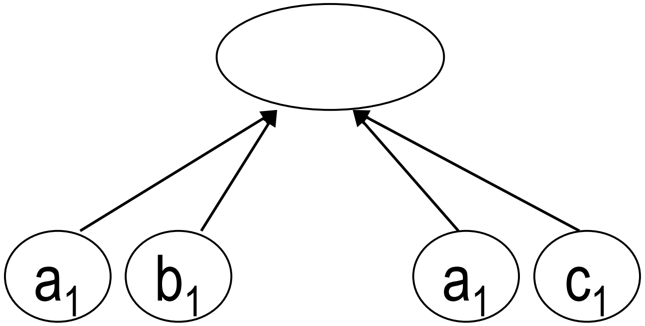
- If two tags exist together very often, they obviously are in one of forementioned relations

Crosswalks algorithm – first step (evaluation)

- Tests are run on database of del.icio.us tags (over 1,5 GB of RDF data) to determine correct wages for the measure:
  - How does the size of both sets affect the results?
  - Is it possible (in terms of computational complexity) to look for complex relations (i.e. including contexts) in this manner?
  - What conclusions can we make as far as specific relations are concerned? In particular – how users are tagging resources – do they use many synonimes together? Or do they rather use terms from many different domains to describe resource thoroughly?

Crosswalks algorithm – second step

- Determining specific relations
- Many constraints applied, making use of how resource was tagged by different users, e.g.:



- Making use of conclusions from the first step

Crosswalks algorithm – third step

- Utilizing users involvement!
- At first – large amount of data will be collected. Users will annotate some predefined resources with as many tags as they can.
- Two first steps of the algorithm could be made after that.
- As the automatic suggestions begin to work – the algorithm will get much feedback from users about how precise the suggestions are.

References

- Wiki pages - http://wiki.corrib.org/index.php/JOnto/
- JOnto website - http://jonto.sourceforge.net/
- notitio.us
- SKOS - http://www.w3.org/2004/02/skos/
- Taxonomies
  - ACM - http://www.acm.org/class/1998/
  - DDC - http://www.tnrdlib.bc.ca/dewey.html
  - DMoz - http://www.dmoz.org/
  - LCC - http://www.loc.gov/catdir/cpso/lcco/
  - UDC - http://www.udcc.org/index.htm

Summary

- Problems with usage of multiple controlled vocabularies
- JOnto – tool for annotating resources
- Crosswalks algorithm’s motivations and idea
- Crosswalks algorithm’s steps details

Questions?

Thank you for your attention!