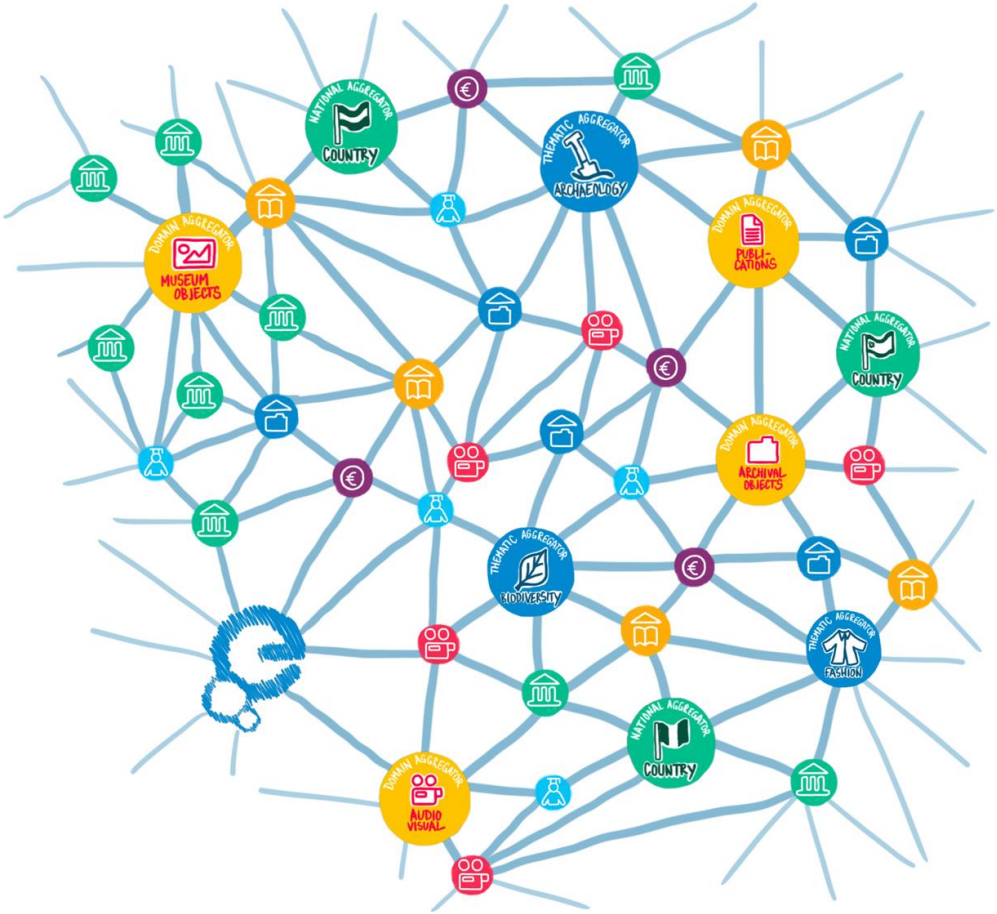
Linking subject labels in Cultural Heritage Metadata to MIMO vocabulary using CultuurLink

Hugo Manguinhas, Valentine Charles, Antoine Isaac| **Europeana Foundation**
Tom Miles| **The British Library**
Aude Lima| **The Centre de Recherche en Ethnomusicologie**
Ariane Néroulidis, Véronique Ginouvès| **The Maison Méditerranéenne des Sciences de l'Homme**
Dimitra Atsidis, Maarten Brinkerink| **Netherlands Institute for Sound and Vision**
Michiel Hildebrand| **Spinque B.V.**
Sergiu Gordea| **Austrian Institute of Technology**

Co-financed by the European Union
Connecting Europe Facility

What is Europeana?

The Platform for Europe's Digital Cultural Heritage



We aggregate metadata:

- *From all EU countries*
- *~3,500 galleries, libraries, archives and museums*
- *More than 52M objects*
- *In about 50 languages*
- *Huge amount of references to places, agents, concepts, time*

[Europeana aggregation infrastructure]
Europeana| CC BY-SA

The Europeana Sounds project

Europeana Sounds aims to increase the amount of audio content available via Europeana

- *also improving geographical and thematic coverage*

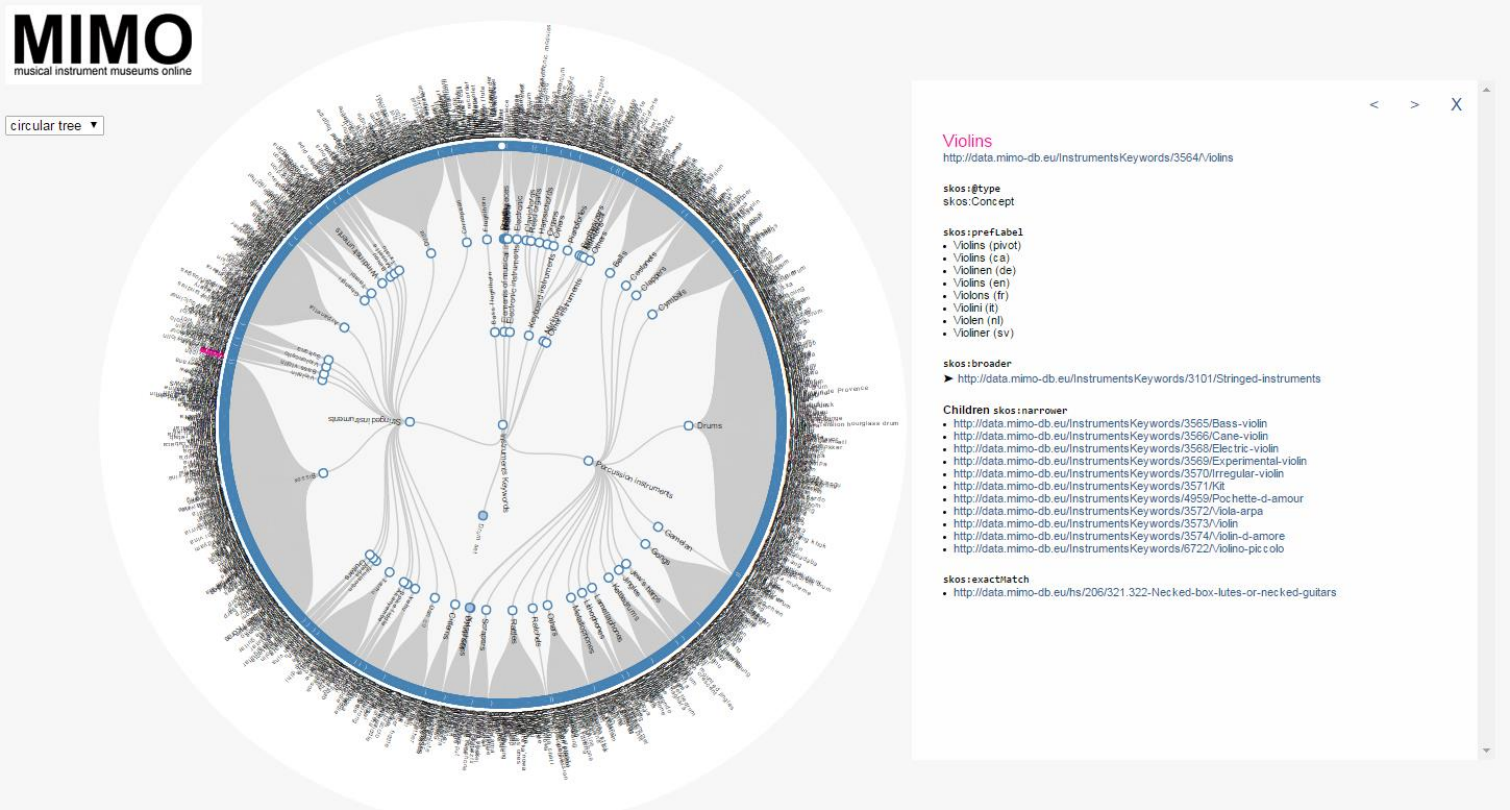
Apart from aggregation, it improves discovery and use of audio content, by enriching metadata through innovative methods

The scope of the experiment

- Evaluate the use of a semi-automatic tool like CultuurLink for a concrete vocabulary alignment case, and
- Assess the coverage of the MIMO vocabulary for enriching Europeana Sounds datasets

About the MIMO vocabulary

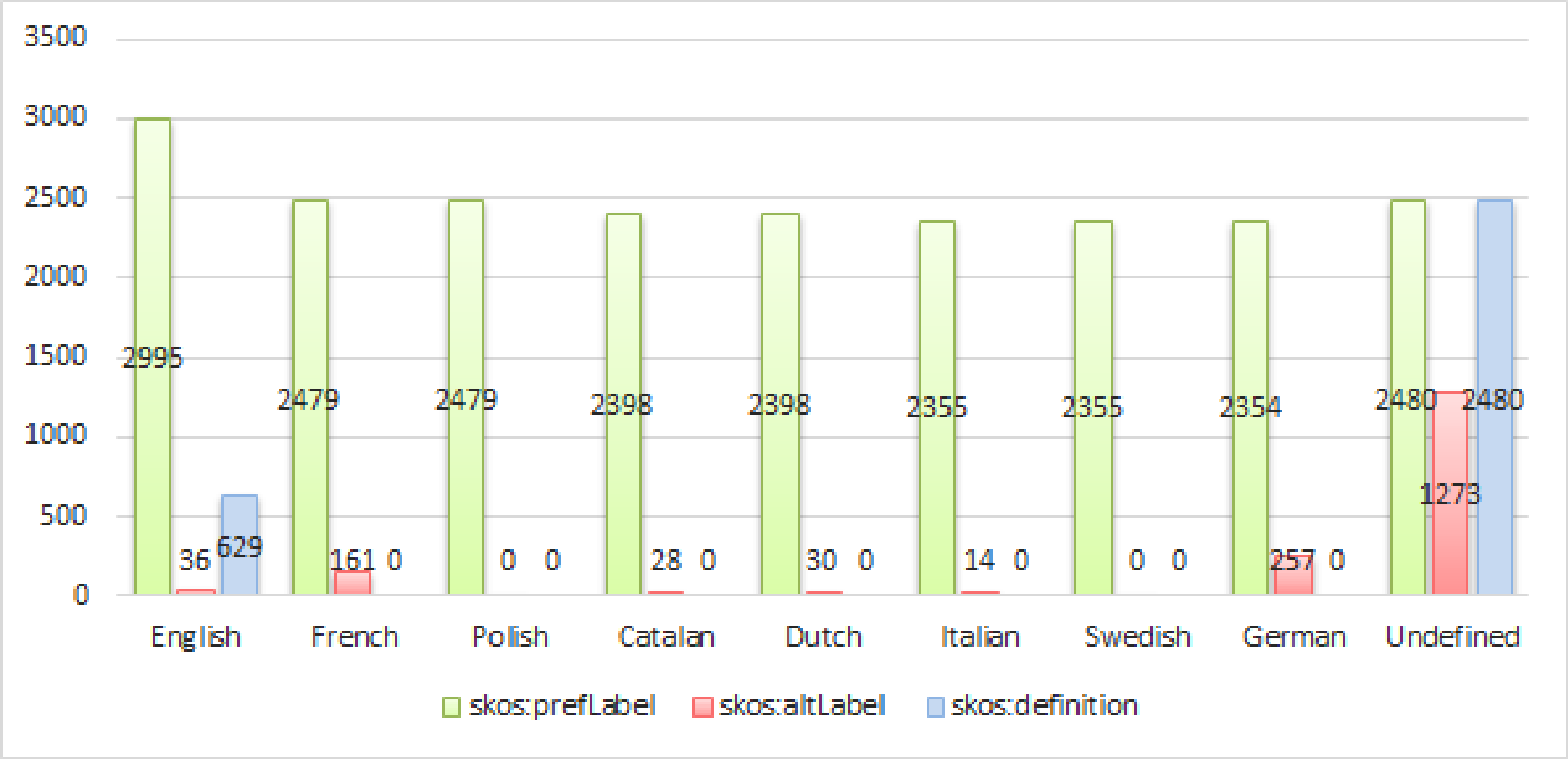
- A multilingual controlled vocabulary of musical instruments
- Developed within the Musical Instruments Museums Online project which gathered some of Europe's most important musical instruments museums



Why MIMO?

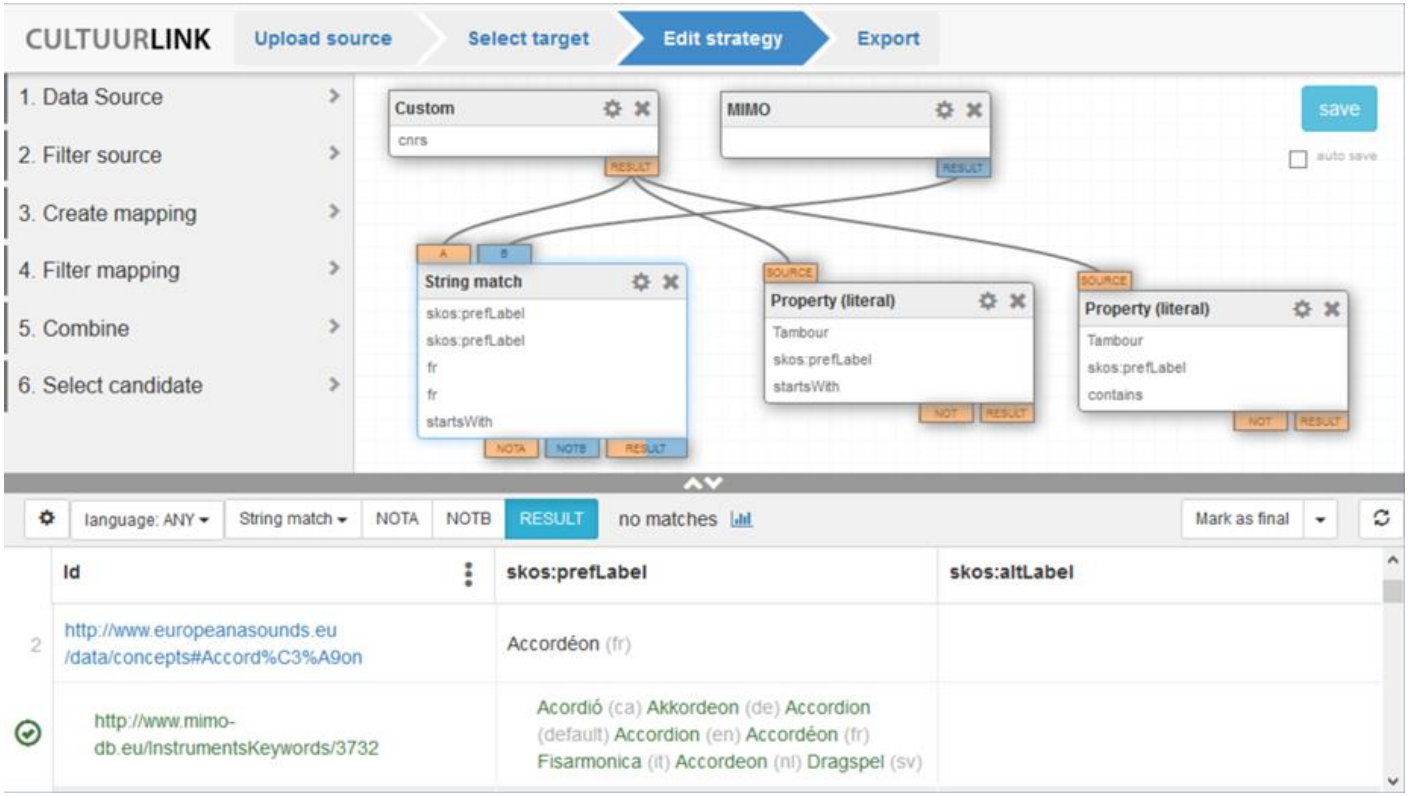
- A significant part of the subjects present in Europeana Sounds collections refer to musical instruments
- Good coverage
  - gathers a total of 3121 musical instruments used by professionals such as Hornbostel-Sachs (641)
  - contains terms in 8 different languages (English, French, Polish, Catalan, Dutch, Italian, Swedish, German)
- Technically available on the Web
  - Follows the Linked Data best practices and recipes (RDF, SKOS, content negotiation)
- Openly available (CCO)

Overview of MIMO language coverage



What is CultuurLink?

- Semi-automatic Vocabulary Alignment Tool
- Successor to EuropeanaConnect's Amalgame



http://cultuurlink.beeldengeluid.nl

Why CultuurLink?

- Freely available
  - as an online open service that any user can use
- Users have the ability to design and experiment with different alignment strategies
  - helps the task of discovering new alignments between two vocabularies
  - users can define and combine strategies that apply different techniques or parameterizations
- Manual control
  - alignments are identified through an automatic means but strategies are designed by users
  - users can decide which alignments are correct and can assign a specific meaning (e.g. skos:exactMatch, skos:related, skos:broadMatch)
- User friendly
  - allows non-technical savvy users to easily perform fairly complex tasks

The participants and their collections

- The British Library (BL) participated with 3 collections:
  - A selection of Asian instruments (**1,099 records**) from the "Colin Huehns Asia Collection"
  - a selection from the “Peter Cooke Uganda Collection” (**1,312 records**)
  - and the “Keith Summers English Folk Music Collection” (**1,326 records**)
- The Centre de Recherche en Ethnomusicologie (CREM)
  - participated with a test collection of **36 records** published in the CD “Musical Instruments of the World”
- The Maison Méditerranéenne des Sciences de l'Homme (MMSH)
  - participated with a collection of **25 records** about folk music
- The Netherlands Institute of Sound and Vision (NISV)
  - participated with a collection of **6,608 records** containing commercial 78 rpm records (Handelsplaten) from different genres like light music, classical music and opera.

Alignment of vocabulary terms

We decided to focus on the vocabulary terms within the subject fields of the metadata as opposed to aligning the full vocabulary used by the providing institution, because:

- not available for use outside the organization and/or in a data structure that suits a vocabulary alignment tool
- we preferred to report on alignments for the subjects used in the source datasets and not on all possible subjects

What have we done?

For each collection we:

- Extracted a SKOS vocabulary out of the subject terms found in the object metadata
- Set-up a permanent session on CultuurLink
- Asked providers to perform the alignments
- Collected and assessed the alignments and feedback obtained from the Data Providers

Concept definition obtained from the MMSH dataset

```
<skos:ConceptScheme
      rdf:about="http://www.europeanasounds.eu/data/mmsh/concepts#ConceptScheme">
</skos:ConceptScheme>
<skos:Concept rdf:ID="grelot">
    <skos:inScheme rdf:resource="#ConceptScheme"/>
    <skos:prefLabel>grelot</skos:prefLabel>
    <skos:note rdf:resource="http://mint-
projects.image.ntua.gr/data/sounds/http://phonotheque.mmsh.huma-
num.fr/dyn/portal/index.seam?page=alo&amp;aloId=9800"/>
    <skos:note rdf:resource="http://mint-
projects.image.ntua.gr/data/sounds/http://phonotheque.mmsh.huma-
num.fr/dyn/portal/index.seam?page=alo&amp;aloId=9775"/>
    <skos:note rdf:resource="http://mint-
projects.image.ntua.gr/data/sounds/http://phonotheque.mmsh.huma-
num.fr/dyn/portal/index.seam?page=alo&amp;aloId=9801"/>
    <skos:note rdf:resource="http://mint-
projects.image.ntua.gr/data/sounds/http://phonotheque.mmsh.huma-
num.fr/dyn/portal/index.seam?page=alo&amp;aloId=9768"/>
    <skos:note rdf:resource="http://mint-
projects.image.ntua.gr/data/sounds/http://phonotheque.mmsh.huma-
num.fr/dyn/portal/index.seam?page=alo&amp;aloId=9798"/>
    <skos:note rdf:resource="http://mint-
projects.image.ntua.gr/data/sounds/http://phonotheque.mmsh.huma-
num.fr/dyn/portal/index.seam?page=alo&amp;aloId=9788"/>
</skos:Concept>
```

Text found in dc:subject

URIs of the records are kept as skos:notes

The alignments obtained from CultuurLink

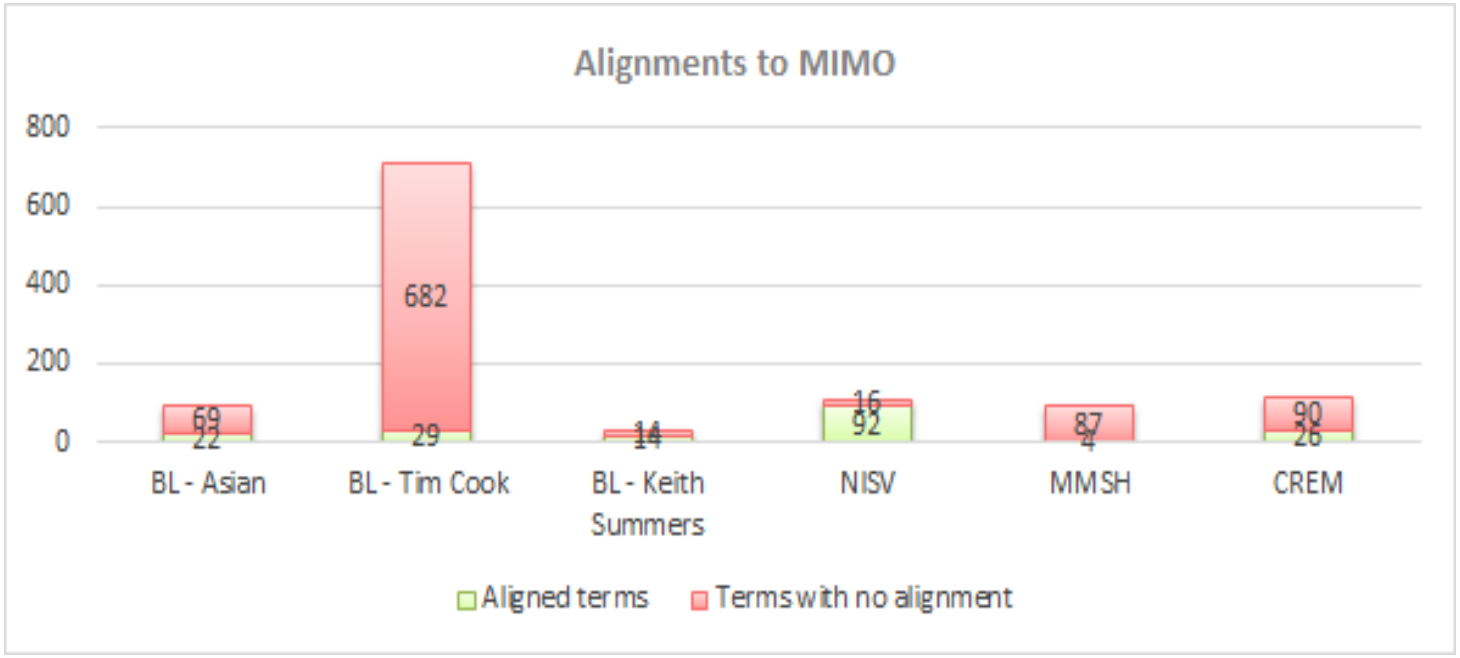
```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#" >
  <rdf:Description rdf:about="http://www.europeanasounds.eu/data/concepts#guitare">
    <skos:exactMatch rdf:resource="http://www.mimo-db.eu/InstrumentsKeywords/3237"/>
    <owl:differentFrom rdf:resource="http://www.mimo-db.eu/InstrumentsKeywords/5137"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.europeanasounds.eu/data/concepts#flûte">
    <skos:exactMatch rdf:resource="http://www.mimo-db.eu/InstrumentsKeywords/3955"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.europeanasounds.eu/data/concepts#grelot">
    <skos:exactMatch rdf:resource="http://www.mimo-db.eu/InstrumentsKeywords/2873"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.europeanasounds.eu/data/concepts#ban">
    <owl:differentFrom rdf:resource="http://www.mimo-db.eu/InstrumentsKeywords/2498"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.europeanasounds.eu/data/concepts#violon">
    <skos:exactMatch rdf:resource="http://www.mimo-db.eu/InstrumentsKeywords/3573"/>
  </rdf:Description>
</rdf:RDF>
```

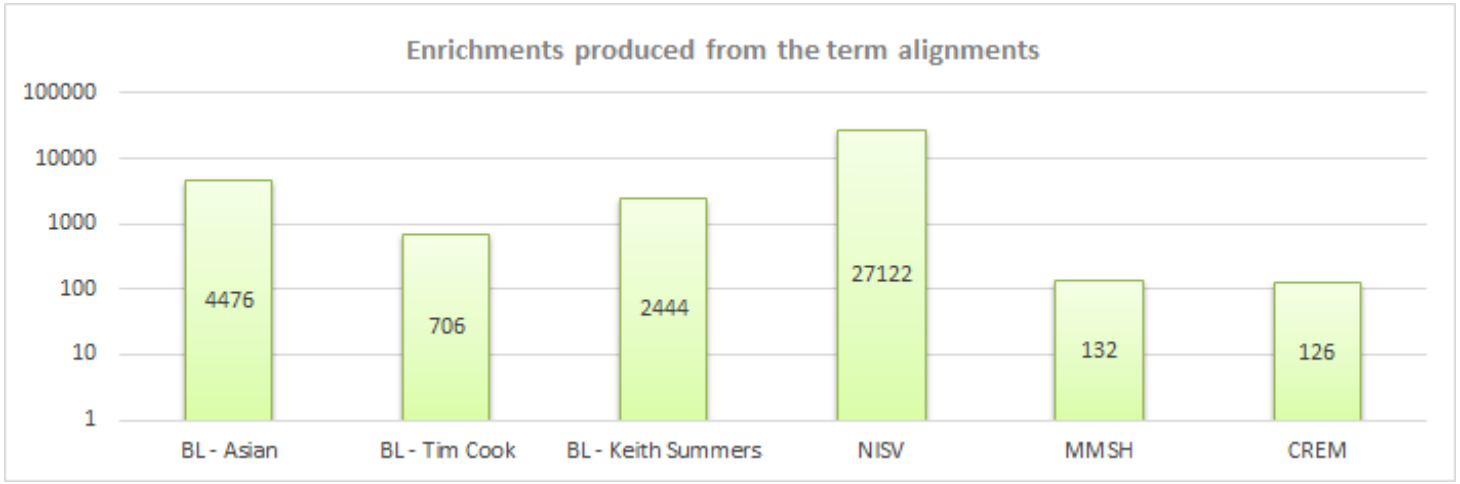
Subject term

MIMO concept

Alignments identified by the data provider for this subject

The quantitative results of the experiment





Findings identified when aligning with CultuurLink (1/2)

- Applying an exact string matching of preferred labels is sufficient to align ~50%
- Also incorrect alignments were identified due to polysemy reasons
  - e.g. “ban” or “zang” which means singing or song matching the instrument “zang”, a sort of cymbals or clapper bells
- Applying match against labels in any language turned out be very successful on finding matches based on vernacular terms
  - But also increased the number of irrelevant alignments

Findings identified when aligning with CultuurLink (2/2)

More elaborate strategies were found very helpful to discover more alignments:

- by using a **less restrictive string matching function** like “contains” or “startsWith”, to surface “broader” or “narrower” relations
- by activating **stemming**
  - e.g. “Trompet” was aligned with “Trompetten” and “Accordeon” with “Accordeons”, both in Dutch
- by applying **fuzzy matching** both with max distance of 1 and 2
- The “**NOT A**” functionality was found crucial to iteratively refine the strategy

Using such strategies also revealed some quality issues in the source metadata, such as:

- misspellings and unrecognizable/doubtful terms

What about MIMO?

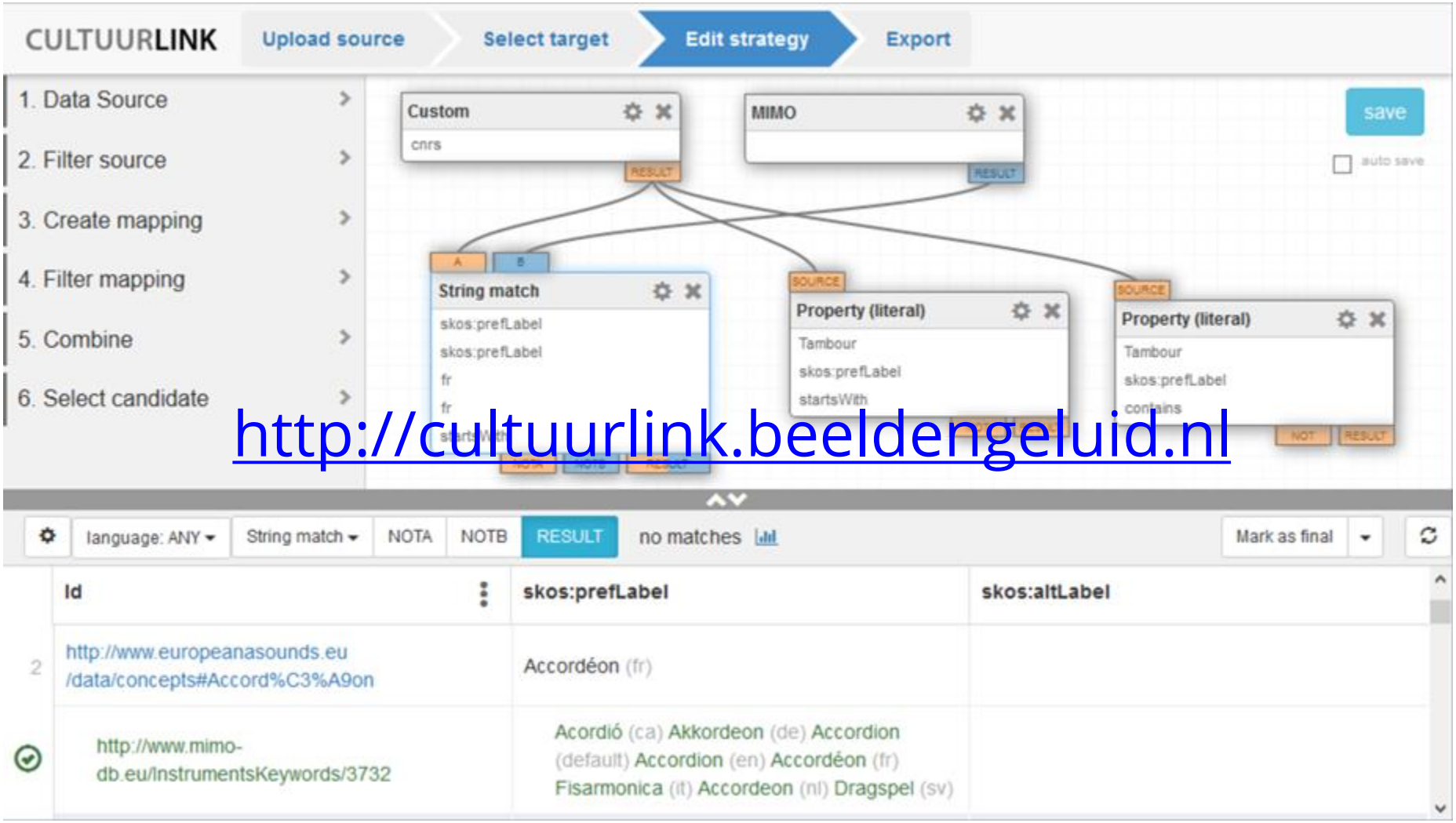
In general the Data Providers found MIMO:

- Good coverage of musical instruments
- Good language coverage comparing to their local vocabulary
- Simplified hierarchy allowing it to be understandable and practical for non musicologists
- Includes updated families treating both electronic instruments and tools that are presented in contemporary music
- Helpful concept definitions

However,

- Lacks concepts to describe voice (texture, mechanism, etc.)
  - But may be enriched by the DOREMUS project with vocal terms from the IAML mediums of performance thesaurus
- Centred on occidental classical music structure

Quick demo



Flowers
Anonymous
1700-1799, Rijksmuseum
Netherlands, Public Domain

Thank you!

europeana

#AllezCulture

europeana
sounds

europeana.eu @EuropeanaEU

Co-financed by the European Union
Connecting Europe Facility