



européana

Sur des approches
d'alignement semi auto-
matique

Antoine Isaac

Atelier : Données liées et données à lier : quels outils pour
quels alignements ?

Mardi 10 juillet 2018, BnF

Quadruplettes à Buffalo
Agence de presse Meurisse
1909, National Library of France
France, Public Domain



Co-financed by the European Union
Connecting Europe Facility

Pourquoi suis-je là ?

- Alignment semi-automatique de vocabulaires
 - STITCH
 - TELplus
 - EuropeanaConnect
- SKOS et implementations
 - RAMEAU, LCSH etc.
- Library Linked Data
- Europeana



Insects and Fruit
Jan van Kessel
1660 - 1665, Rijksmuseum
Netherlands, Public Domain

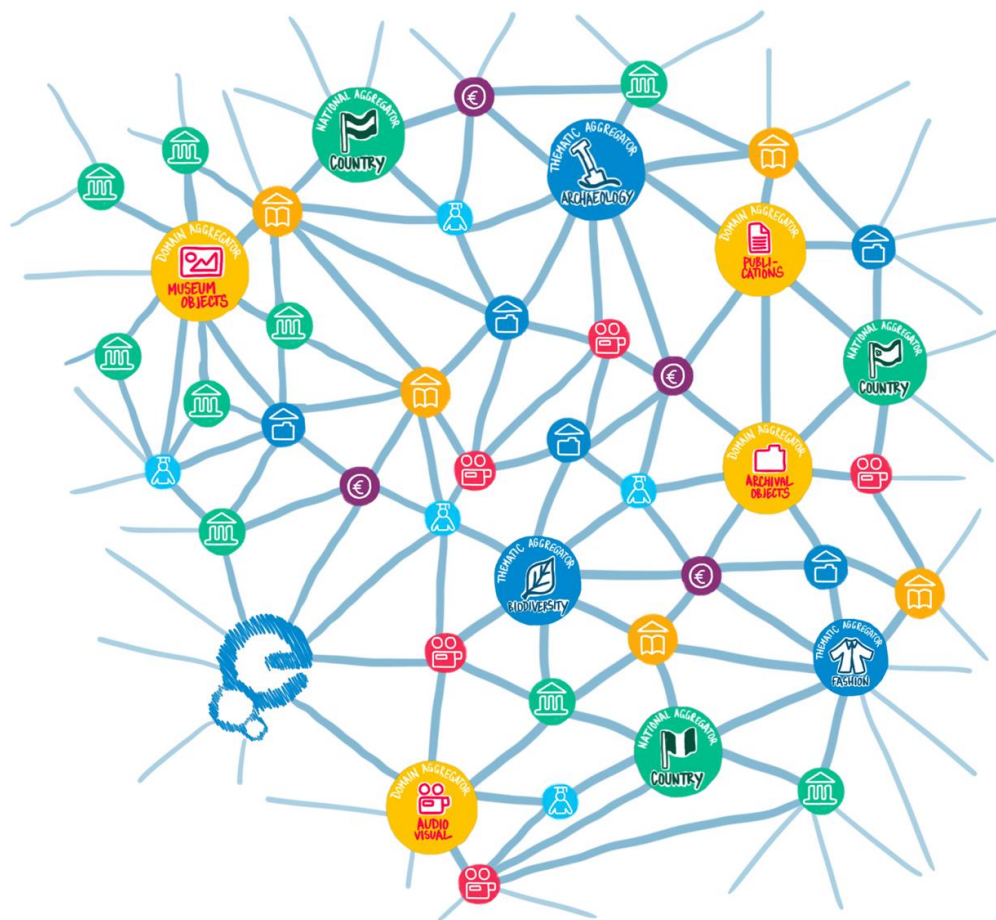
europeana

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 | **Centre de Recherche en Ethnomusicologie**
Ariane Néroulidis, Véronique Ginouvès | **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**

J.V. KESSEL. 5

Europeana?



We aggregate metadata:

- *Over 50M objects*
- *From 3,500 libraries, archives, museums*
- *From all EU countries*
- *In about 50 languages*
- *Huge amount of references to places, agents, concepts*

DENKSCHETS.NL

[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

Our experiment

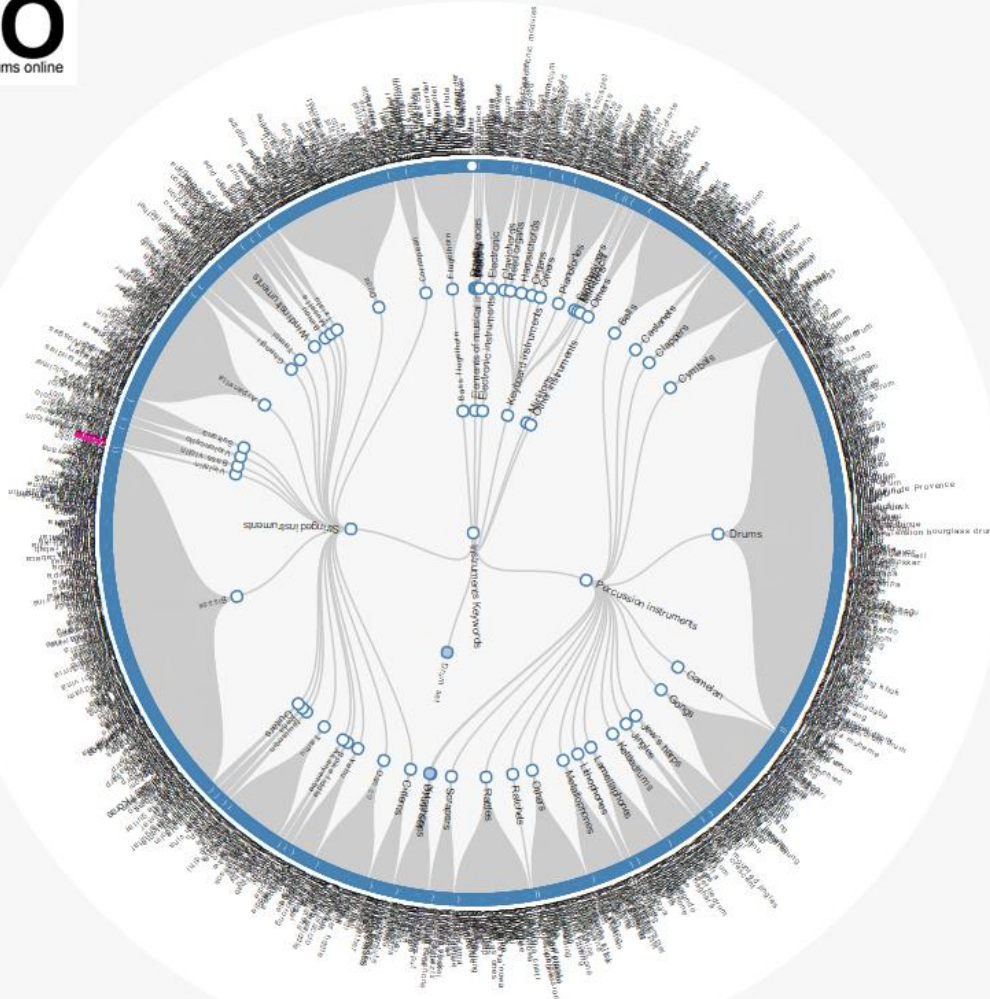
- Evaluate the use of a semi-automatic tool for a concrete vocabulary alignment case
- Assess the coverage of the MIMO vocabulary for enriching Europeana Sounds datasets

The MIMO Vocabulary

- A multilingual controlled vocabulary of musical instruments
- Developed by the [Musical Instruments Museums Online](http://www.mimo-db.eu/) project that gathers some of Europe's most important musical instruments museums

MIMO
musical instrument museums online

circular tree ▾



Violins

<http://data.mimo-db.eu/InstrumentsKeywords/3564/Violins>

skos:type
skos:Concept

skos:prefLabel

- Violins (pivot)
- Violins (ca)
- Violinen (de)
- Violins (en)
- Violons (fr)
- Violini (it)
- Violen (nl)
- Violiner (sv)

skos:broader

► <http://data.mimo-db.eu/InstrumentsKeywords/3101/Stringed-instruments>

Children skos:narrower

- <http://data.mimo-db.eu/InstrumentsKeywords/3565/Bass-violin>
- <http://data.mimo-db.eu/InstrumentsKeywords/3566/Cane-violin>
- <http://data.mimo-db.eu/InstrumentsKeywords/3568/Electric-violin>
- <http://data.mimo-db.eu/InstrumentsKeywords/3569/Experimental-violin>
- <http://data.mimo-db.eu/InstrumentsKeywords/3570/Irregular-violin>
- <http://data.mimo-db.eu/InstrumentsKeywords/3571/Kit>
- <http://data.mimo-db.eu/InstrumentsKeywords/4959/Pochette-d-amour>
- <http://data.mimo-db.eu/InstrumentsKeywords/3572/Viola-arpa>
- <http://data.mimo-db.eu/InstrumentsKeywords/3573/Violin>
- <http://data.mimo-db.eu/InstrumentsKeywords/3574/Violin-d-amore>
- <http://data.mimo-db.eu/InstrumentsKeywords/6722/Violino-piccolo>

skos:exactMatch

- <http://data.mimo-db.eu/hs/206/321.322-Necked-box-lutes-or-necked-guitars>

Why MIMO?

- A significant part of Europeana Sounds collections refer to musical instruments and MIMO has good coverage of them
 - Gathers a total of 3121 musical instruments
 - Contains terms in 8 different languages (English, French, Polish, Catalan, Dutch, Italian, Swedish, German)
 - Based on established classification (Hornbostel-Sachs)
- Technically available on the Web
 - Follows the Linked Data best practices and recipes (RDF, SKOS, content negotiation)
- Openly available (CCO)
- Used in the DOREMUS project

What is CultuurLINK?

- Semi-automatic vocabulary alignment tool
- Based on a prototype from EuropeanaConnect
- Online service freely available

The screenshot shows the CultuurLINK interface in the 'Edit strategy' step. The sidebar on the left lists the workflow steps: 1. Data Source, 2. Filter source, 3. Create mapping, 4. Filter mapping, 5. Combine, and 6. Select candidate. The main workspace contains several strategy boxes: 'Custom' (cnrs) and 'MIMO' (both with 'RESULT' buttons), a 'String match' box (with 'NOTA', 'NOTB', and 'RESULT' buttons), and two 'Property (literal)' boxes (one with 'NOT' and 'RESULT' buttons, the other with 'NOT' and 'RESULT' buttons). A 'save' button and an 'auto save' checkbox are in the top right. The bottom section shows a table of results with columns for 'Id', 'skos:prefLabel', and 'skos:altLabel'.

Id	skos:prefLabel	skos:altLabel
2	http://www.europeanasounds.eu/data/concepts#Accord%C3%A9on	Accordéon (fr)
	http://www.mimo-db.eu/instrumentsKeywords/3732	Acordió (ca) Akkordeon (de) Accordion (default) Accordion (en) Accordéon (fr) Fisarmonica (it) Accordeon (nl) Dragspel (sv)

<http://cultuurlink.beeldengeluid.nl>

Participants and their collections

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

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 session on CultuurLINK
- asked participants to perform the alignments
- collected and assessed the alignments and feedback from the participants

Concept definition obtained from the MMSH dataset

```
<skos:ConceptScheme
  rdf:about="http://www.europeanasons.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&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&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&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&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&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&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.europeanasons.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.europeanasons.eu/data/concepts#flûte">
    <skos:exactMatch rdf:resource="http://www.mimo-db.eu/InstrumentsKeywords/3955"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.europeanasons.eu/data/concepts#grelot">
    <skos:exactMatch rdf:resource="http://www.mimo-db.eu/InstrumentsKeywords/2873"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.europeanasons.eu/data/concepts#ban">
    <owl:differentFrom rdf:resource="http://www.mimo-db.eu/InstrumentsKeywords/2498"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.europeanasons.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

Quick demo

The screenshot shows the CultuurLINK interface during the 'Edit strategy' step. The sidebar on the left lists the following steps:

1. Data Source
2. Filter source
3. Create mapping
4. Filter mapping
5. Combine
6. Select candidate

The main workspace displays two source boxes: 'Custom' (containing 'cnrs') and 'MIMO'. These are connected to three strategy boxes: 'String match', 'Property (literal)' (with 'startsWith'), and another 'Property (literal)' (with 'contains'). The 'String match' strategy is currently selected and shows the following configuration:

- skos:prefLabel
- skos:prefLabel
- fr
- fr
- startsWith

The bottom status bar shows the current strategy is 'String match' and the result is 'no matches'. The table below displays the results of the mapping process:

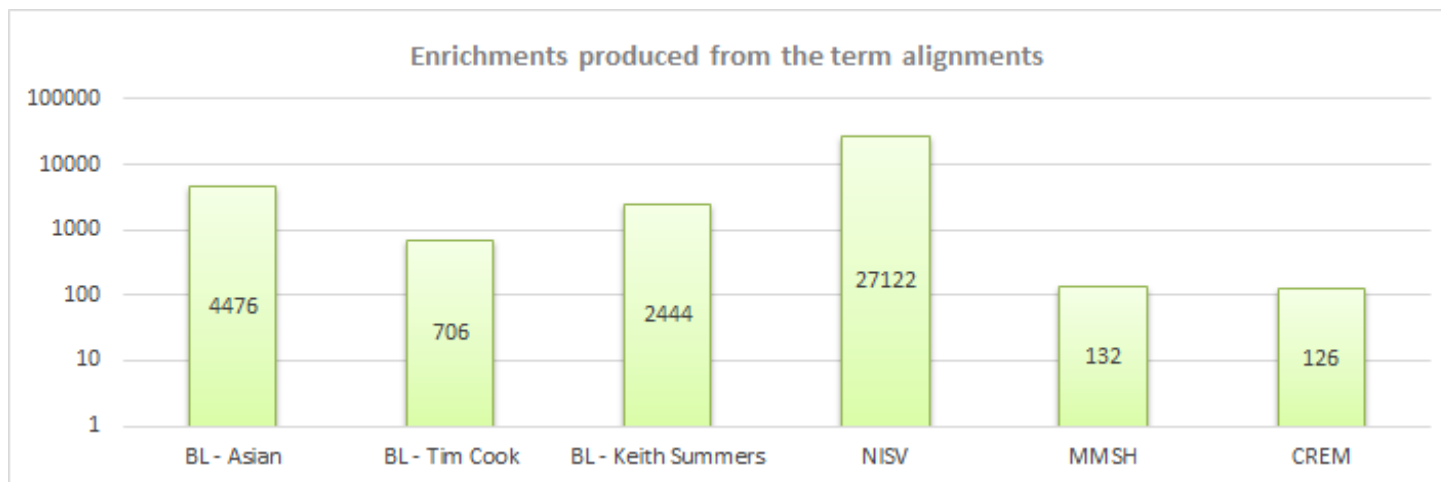
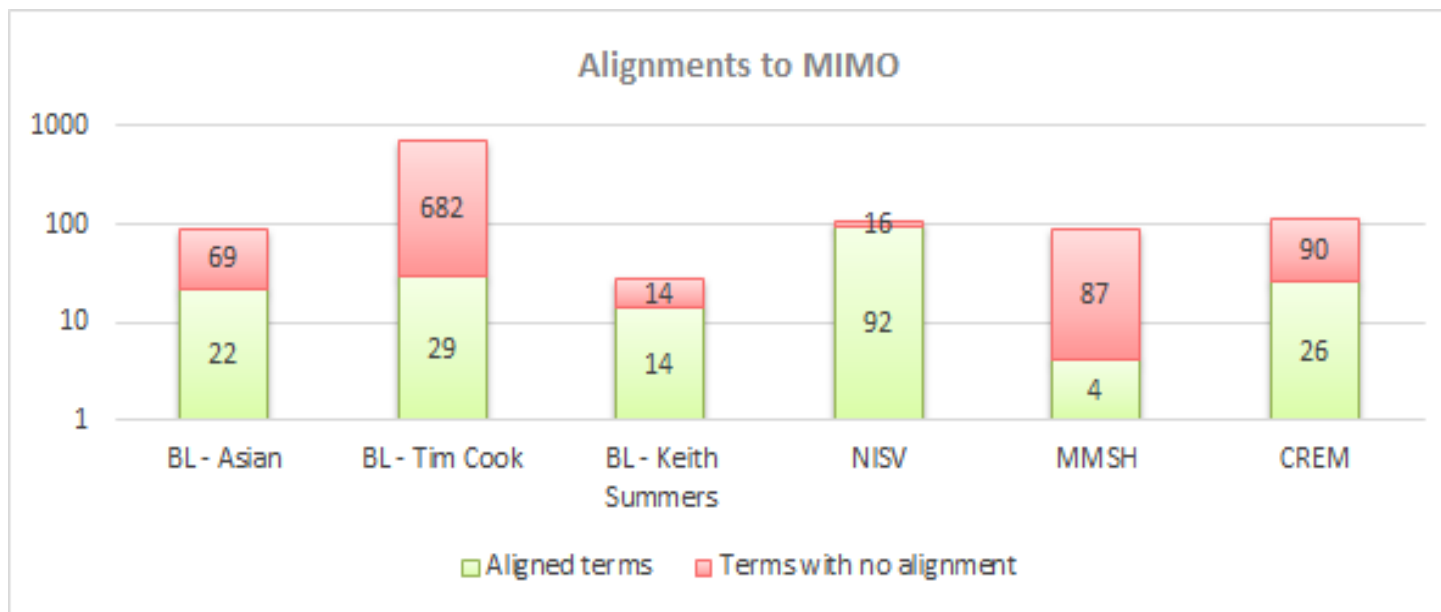
Id	skos:prefLabel	skos:altLabel
2	Accordéon (fr)	
	Acordió (ca) Akkordeon (de) Accordion (default) Accordion (en) Accordéon (fr) Fisarmonica (it) Accordeon (nl) Dragspel (sv)	

<http://cultuurlink.beeldengeluid.nl>

Why CultuurLINK?

- Users can play with different alignment strategies
 - users can define and combine strategies that apply different techniques or parameters of one technique
 - the tool facilitates experimentation to discover new alignments between two vocabularies
- Manual control
 - alignments are identified automatically 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)
- (Relatively) user-friendly
 - allows non-technical savvy users to easily perform fairly complex tasks

Quantitative results



Findings (1/2)

- Applying exact string matching of preferred labels is sufficient to align 50% of subjects
- Polysemy hurts, as usual, leading to incorrect alignments
 - e.g. “ban” or “zang” which means singing or song matching the instrument “zang”, a sort of cymbals or clapper bells
- Match labels across languages turned out be successful for finding matches based on vernacular terms
 - but it also increased the number of irrelevant alignments

Findings (1/2)

- More elaborate strategies were useful to discover more alignments:
 - less restrictive string matching like “contains”, “startsWith” or fuzzy matching both with distance 1 or 2 can surface broader/narrower relations
 - stemming enables aligning e.g. “Trompet” with “Trompetten” and “Accordeon” with “Accordeons” (in Dutch)
 - 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 doubtful terms

What about MIMO?

Participants found that MIMO had great features:

- good coverage of musical instruments and good language coverage compared to their local vocabulary
- simple hierarchy, practical for non musicologists
- updated families treating both electronic instruments and tools that are presented in contemporary music
- helpful concept definitions

It also has weak points:

- centred on occidental classical music structure
- lacks concepts to describe voice (texture, mechanism, etc.)

Pourquoi est-ce intéressant?

L'alignement semi-automatique tel que supporté par CultuurLINK permet d'envisager :

- La considération d'une expertise de domaine, à l'intérieur ou à l'extérieur des institutions (nichesourcing)
- Le passage à l'échelle
- Une flexibilité en termes des techniques d'alignement employées
- La vérification du contenu et de la pertinence des vocabulaires à aligner

Wikidata Mix'n'Match



Concours de cycles nautiques sur le lac
d'Enghien : Berregent piloté par Austerling

Agence de presse Meurisse

1914, National Library of France
France, Public Domain

Mix'n'Match

- Un outil de validation d'alignements de vocabulaires vers Wikidata
<https://tools.wmflabs.org/mix-n-match/>
- Les correspondances potentielles sont calculées par l'outil lors du chargement du vocabulaire
- Elles peuvent être validées par n'importe quel membre de la communauté Wikidata
- Nous encourageons nos partenaires à aligner leurs vocabulaires avec Wikidata en l'utilisant:

<https://pro.europeana.eu/page/get-your-vocabularies-in-wikidata>

(Sandra Fauconnier, Valentine Charles, Liam Wyatt)

Mix'n'Match et MIMO – les étapes (1/3)

- Convertir la hierarchie du vocabulaire MIMO en simple liste de termes
- Importer dans Mix'n'Match
- Définir une propriété Wikidata pour les résultats de l'alignement
- Valider manuellement les correspondances produites automatiquement (142)

# Ghichak	Broader term: Fiddles. Ghičak (de), Ghichak (en), Ghichak (fr), Ghichak (it), Ghichak (nl), Ghichak (sv), Ghichak (ca), Ghichak (pl).	<i>Automatically matched</i>
Ghaychak [Q1521728]	Musical instrument;	Confirm Remove

Mix'n'Match et MIMO – les étapes (2/3)

- Ajouter d'éventuelles correspondances manquantes

The screenshot shows the MIMO interface with a list of musical instruments and a dialog box for adding missing correspondences.

tools.wmflabs.org says:

Enter Q number of matching item

The list of instruments includes:

- # Alto dulcian
- # Alto fagotto
- # Bajon
- # Contrebasse à anche
- # Bassoon
- # Bassonore

The **Bassoon** entry is highlighted with a blue circle. The **Set Q** button for this entry is also circled in blue.

The **bassoon** (Q159998) entry is expanded, showing the following table:

Language	Label	Description	Also known as
English	bassoon	musical instrument	fagot
British English	No label defined	No description defined	
Dutch	fagot	No description defined	
French	basson	instrument à vent de la famille des bois	fagott

Mix'n'Match et MIMO – les étapes (3/3)

- Ajouter des précisions aux entités Wikidata correspondant aux concepts de MIMO, par exemple en rajoutant les liens hiérarchiques
- Créer de nouveaux (types d')instruments pour combler les lacunes de Wikidata

<https://tools.wmflabs.org/mix-n-match/#/catalog/391>

Flowers
Anonymous
1700-1799, Rijksmuseum
Netherlands, Public Domain

Thank you!



europaena

#AllezCulture



europaena
sounds