SemJoconde

Toward a Semantic Representation
Of the
Joconde Database
Jean-claude.moissinac@telecom-paris.fr
Telecom Paris

Authors: Jean-Claude Moissinac, François Rouzé, Bastien Germain, Piyush Wadhera

October 2020
Presenter

• Jean-Claude Moissinac
• Associate Professor
• Telecom Paris

• Research lead for the Data&Musée project
  • Aiming to discover knowledge from data provening from museums and monuments enriched by the Linked Open Data
SemJoconde Goal

• Contribute to LOD with Knowledge Graphs about French Cultural Heritage
• Get an enriched semantic version of the Joconde database
• Part of the Data&Musée project
Data sources

• Joconde Database
  • Collected by the French Ministry of Culture
  • Data about 580035 creations, part of the French cultural heritage

• Wikidata

• Wikidata: Joconde project
  • https://www.wikidata.org/wiki/Wikidata:WikiProject_Vocabulaires_Joconde/
  • Expert linking between Joconde database vocabulary and Wikidata entities
    • 5/30/2020, validated links:
      • 2560 creators, 153 domains, 47 places, ...
Joconde database content

• For each creation
  • Ref: an unique string, internal ID of the work
  • Name of creators: anonymous, ”RODIN Auguste” ...
  • Title: "Tête d'homme barbu"
  • Domains: ”Néolithique”, “photographie”,
  • Techniques: ”bas relief”
  • Period: ”2e quart 20e siècle”
  • Dimensions: ”H. 27; l. 6.1 ; P. 4.2”
  • Location: ”Grenoble ; musée Stendhal”
## Source fields and ground truth

<table>
<thead>
<tr>
<th>Category (field)</th>
<th>Validated in ground truth</th>
<th>Source</th>
<th>% of validated links</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creators</td>
<td>2560</td>
<td>37828</td>
<td>6.7 %</td>
</tr>
<tr>
<td>Domains</td>
<td>153</td>
<td>153</td>
<td>100 %</td>
</tr>
<tr>
<td>Object types</td>
<td>77</td>
<td>5766</td>
<td>1.3 %</td>
</tr>
<tr>
<td>Places</td>
<td>35</td>
<td>3593</td>
<td>0.9 %</td>
</tr>
<tr>
<td>Epochs</td>
<td>500</td>
<td>831</td>
<td>60.1 %</td>
</tr>
<tr>
<td>Periods</td>
<td>60</td>
<td>346</td>
<td>17.3 %</td>
</tr>
</tbody>
</table>

Source= fields in the Joconde database  
Ground truth = validated links in the [Wikidata Joconde Project](https://wikidata.org)
Model

• Based on CIDOC-CRM
  • CIDOC = ICOM International Committee for Documentation
    • Developing standards for information management in museums
  • CRM = Conceptual Reference Model
  • Used by: Europeana, British Museum, Paris Musées Collections, ...

• With added properties like
  • DMP5_has_url, for a link to a significant web page
  • DMP19_has_period; ex.: 19th century
Artwork Model
Process

• Create three URIs for each work
• List the possible values for each field
• Clean values
  • Ex.: sculpture=Sculpture
• Count the number of works of each value
• Try to link each value with Wikidata
• Expert check of link for the most used values
• Check links against a ground truth to evaluate quality of the resulting representation
3 for each work

- **URI=Universal Resource Identifier**
- **Based on the REF field of the Joconde database**
  - Example: REF = AE037477
- **Creation event**
  - [http://datamusee.givingsense.eu/onto/event/creation/AE037477](http://datamusee.givingsense.eu/onto/event/creation/AE037477)
- **Physical work**
  - [http://datamusee.givingsense.eu/onto/work/physical/AE037477](http://datamusee.givingsense.eu/onto/work/physical/AE037477)
- **Conceptual work:**
  - [http://datamusee.givingsense.eu/onto/work/concept/AE037477](http://datamusee.givingsense.eu/onto/work/concept/AE037477)
Entity linking, a simple algorithm

• Giving a label
• Produce variants: lowercase, uppercase, title case, push the first word in the last place
• Use the Wikidata search service with each variant
  • Within a sparql query
  • With a target for types of entities (see sample in the next slide)
• If only one link found, get it
• If several links found
  • If the label for one link matches the searched label in lowercase, get the link
  • Else get nothing
Example: link about a creator

• Leonard de Vinci
  • Variants: leonard de vinci, LEONARD DE VINCI, Leonard De Vinci, De Vinci Leonard

• Possible types:
  • Wikidata entities
  • Painter (Q1028181), sculptor (Q1281618), drawer (Q15296811), artist (Q483501), photographer (Q33231)… (see article)
Techniques

• 9697 terms used
• Hundred most used covers 88% of works
• Useful Wikidata types:
  • Artistic technique Q11177771
  • Painting material Q3300034
• Exact match between these types and terms: 45
• Coverage of these 45 terms: 43%
## Creators

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage or Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>No known creator</td>
<td>16.83 % of works</td>
</tr>
<tr>
<td>Creator known as «Anonymous »</td>
<td>10.72 % of works</td>
</tr>
<tr>
<td>Identified creators</td>
<td>426,885 works</td>
</tr>
<tr>
<td>Creators producing more than 10 works</td>
<td>5,217 works</td>
</tr>
<tr>
<td>Validated links in the ground truth</td>
<td>2,560 links</td>
</tr>
<tr>
<td>Productive creators in the ground truth</td>
<td>1,325 (covering 25.04 % of works)</td>
</tr>
<tr>
<td>Links found by our algorithm</td>
<td>1,180</td>
</tr>
<tr>
<td>Exact links found</td>
<td>1,177</td>
</tr>
</tbody>
</table>
**Evaluation about creators**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of creators validated by the Wikidata Joconde project and targets of the evaluation</td>
<td>1315</td>
</tr>
<tr>
<td>Number of creators for which our algorithm finds a Wikidata entity</td>
<td>1180</td>
</tr>
<tr>
<td>Number of exact links found</td>
<td>1177</td>
</tr>
<tr>
<td>Precision relative to creators</td>
<td>99.74</td>
</tr>
<tr>
<td>Recall relative to creators</td>
<td>88.83</td>
</tr>
<tr>
<td>F1 score</td>
<td>93.97</td>
</tr>
</tbody>
</table>
# Museums and other sites

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sites in the Joconde database</td>
<td>3593</td>
</tr>
<tr>
<td>Sites linked in the ground truth</td>
<td>0</td>
</tr>
<tr>
<td>«Richest» sites (with more than 100 works)</td>
<td>304</td>
</tr>
<tr>
<td>Coverage of the richest sites</td>
<td>98.43 % (580035 works)</td>
</tr>
<tr>
<td>Found links by our algorithm</td>
<td>155</td>
</tr>
<tr>
<td>Manually checked links (first sites)</td>
<td>76</td>
</tr>
<tr>
<td>Exact links in the checked links</td>
<td>100 % (76)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Searched links</td>
<td>76</td>
</tr>
<tr>
<td>Found links</td>
<td>42</td>
</tr>
<tr>
<td>Exact links</td>
<td>42</td>
</tr>
<tr>
<td>Precision</td>
<td>100</td>
</tr>
<tr>
<td>Recall</td>
<td>55.26</td>
</tr>
<tr>
<td>F1</td>
<td>71.18</td>
</tr>
</tbody>
</table>
Conclusion

• Closed to 600,000 works described by triples
• 191,8315 links to wikidata
• Publicly available: https://github.com/datamusee/semjoconde
• Asserted level of quality
• Possibly useful for future works
  • Machine learning
  • French cultural heritage vocabulary
Questions

- Jean-claude.moissinac@telecom-paris.fr