Using the Semantic Web to Improve Knowledge of Translations

Karen Smith-Yoshimura
OCLC Research
Challenges

- Different writing systems, different transliterations
- Metadata not always good enough
- Linked data opportunities

Why focus on translations?

https://www.nasa.gov/image-feature/africa-and-europe-from-a-million-miles-away
Why focus on translations?

For Developers:
Present information in the preferred language & script of the user

For Academics:
Understand information sharing across cultures
Translations

Leo Tolstoy: 97 languages
Rabindranath Tagore: 93 languages
Homer: 84 languages
Mahatma Gandhi: 52 languages
Isaac Bashevis Singer: 52 languages
Najīb Maḥfūẓ: 47 languages
Cao Xueqin: 27 languages
Murasaki Shikabu: 21 languages

HangingTogether, 2013-11-12
[By January 2017, person clusters have increased to 45 million]
Title: *Sein und Zeit*
Language: German
Author: Martin Heidegger
Created: 1927
WorldCat today

• Resources in nearly all languages
• More than 2.5 billion holdings contributed by libraries worldwide
• More than half the database is for works not in English

Languages

April 2017
Language of cataloging

- English: 160,000,000
- German: 60,000,000
- French: 30,000,000
- Dutch: 20,000,000
- Spanish: 10,000,000
- Danish: 5,000,000
- Swedish: 5,000,000
- Italian: 5,000,000
- Chinese: 5,000,000
- Japanese: 5,000,000

OCLC WorldCat
Language of cataloging and subject headings

*Sein und Zeit* by Martin Heidegger

Filosofía alemana [@es, Spanish]
Fundamentalontology. Ontologie. [@de, German]

哲学思想 [@zh, Chinese]
Leveraging language of cataloging

Language of cataloguing = fre

Language of cataloguing = eng
Après avoir vécu à Paris, Gaud, une belle femme, revient vivre en Bretagne et tombe amoureuse de Yann, marin sur le bateau "la Marie" qui navigue dans la mer du Nord, connue pour ses dangers, vers l'Islande. Yann n'aime que la mer... Ce classique est à la fois un récit de voyage et un roman d'amour.

"Iceland Fisherman (1888) was Loti's seventh novel and first considerable success. An impressionist writer of pervasive melancholy and great descriptive power, he served for many years in the French Navy, where he learned to know the sea in all its moods and to observe at first hand the scenic contrasts which are an outstanding feature of his books. In this tale, the hero, Yann, sturdy but wayward, and the girl Gaud, simple and complex by turns, reflect the ever-changing atmosphere of the element by which they live. There is also light, color and sound in Loti's magnificent pictures of the Icelandic seas, pictures which Laurence Bisson has described as 'a system of suggestion and evocation not without resemblance to the Symbolist technique.'"
Many languages in WorldCat written in non-Latin scripts

<table>
<thead>
<tr>
<th>Language</th>
<th>% with Original Script</th>
<th>% with Transliteration Only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>92%</td>
<td>8%</td>
</tr>
<tr>
<td>Japanese</td>
<td>88%</td>
<td>12%</td>
</tr>
<tr>
<td>Russian</td>
<td>12%</td>
<td>88%</td>
</tr>
<tr>
<td>Hebrew</td>
<td>77%</td>
<td>23%</td>
</tr>
<tr>
<td>Arabic</td>
<td>54%</td>
<td>46%</td>
</tr>
<tr>
<td>Thai</td>
<td>67%</td>
<td>33%</td>
</tr>
<tr>
<td>Korean</td>
<td>76%</td>
<td>24%</td>
</tr>
<tr>
<td>Greek</td>
<td>19%</td>
<td>81%</td>
</tr>
<tr>
<td>Persian</td>
<td>31%</td>
<td>69%</td>
</tr>
<tr>
<td>Ukrainian</td>
<td>8%</td>
<td>92%</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>25%</td>
<td>75%</td>
</tr>
<tr>
<td>Hindi</td>
<td>1%</td>
<td>99%</td>
</tr>
<tr>
<td>Yiddish</td>
<td>61%</td>
<td>39%</td>
</tr>
<tr>
<td>Tamil</td>
<td>25%</td>
<td>75%</td>
</tr>
<tr>
<td>Urdu</td>
<td>11%</td>
<td>89%</td>
</tr>
</tbody>
</table>

홍길동전

Ιλιάδα

源氏物語

Война и миръ

พระอภัยมณี
Wikipedia languages

As of 2017-05-31
Multilingual Linked Data Dataset

*The Grand Design* by Stephen Hawking and Leonard Mlodinow
  81 translations in 24 languages

*Les mots et les choses: Une archéologie des sciences humaines*
by Michel Foucault
  293 translations in 20 languages

*Pêcheur d'Islande* by Pierre Loti
  494 translations in 34 languages

*Principia philosophiae* by René Descartes
  889 translations in 14 languages

*Sein und Zeit* by Martin Heidegger
  570 translations in 33 languages
Multilingual linked data dataset

Data extracted from 4,073 WorldCat records

Enhanced by data from Wikidata

WHERE {
  ?wd wdt:P243 ?ocn;
  wdt:P364 ?originalLanguage;
  rdfs:label ?label.
  FILTER(LANG(?label) = ?originalLanguageCode)


  OPTIONAL {
    ?wc dct:mediator ?geoLocation
  }
}

<table>
<thead>
<tr>
<th>wd</th>
<th>label</th>
<th>translationCount</th>
<th>numTranslationLanguages</th>
<th>numGeolocations</th>
</tr>
</thead>
<tbody>
<tr>
<td>wd:Q93773</td>
<td>&quot;The Grand Design&quot; @en</td>
<td>&quot;80&quot;^^xsd:integer</td>
<td>&quot;24&quot;^^xsd:integer</td>
<td>&quot;5220&quot;^^xsd:integer</td>
</tr>
<tr>
<td>wd:Q3412135</td>
<td>&quot;Pêcheur d'Islande&quot; @fr</td>
<td>&quot;493&quot;^^xsd:integer</td>
<td>&quot;34&quot;^^xsd:integer</td>
<td>&quot;1592&quot;^^xsd:integer</td>
</tr>
<tr>
<td>wd:Q1195384</td>
<td>&quot;Principia philosophiae&quot; @la</td>
<td>&quot;888&quot;^^xsd:integer</td>
<td>&quot;14&quot;^^xsd:integer</td>
<td>&quot;2223&quot;^^xsd:integer</td>
</tr>
<tr>
<td>wd:Q404567</td>
<td>&quot;Sein und Zeit&quot; @de</td>
<td>&quot;566&quot;^^xsd:integer</td>
<td>&quot;33&quot;^^xsd:integer</td>
<td>&quot;2669&quot;^^xsd:integer</td>
</tr>
<tr>
<td>wd:Q1215306</td>
<td>&quot;Les Mots et les Choses&quot; @fr</td>
<td>&quot;293&quot;^^xsd:integer</td>
<td>&quot;20&quot;^^xsd:integer</td>
<td>&quot;1454&quot;^^xsd:integer</td>
</tr>
<tr>
<td>1</td>
<td>876191532</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>19970719s1997 it 000 0 ita d</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**No original language**

**No original title**

**Chiodi translator, not author**
All “translationOfWork” Label

Instead of WorldCat’s Einei kai chronos
# Original Work (in Chinese)
<http://worldcat.org/entity/work/id/1215997>
  a schema:CreativeWork;
  schema:creator <http://viaf.org/viaf/102266649> ; # "Gao, Xingjian"
  schema:inLanguage "zh";
  schema:name "靈山"@zh-hant.

# Translated Work (in English)
<http://worldcat.org/entity/work/id/145209748>
  a schema:CreativeWork;
  schema:creator <http://viaf.org/viaf/102266649> ; # "Gao, Xingjian"
  schema:translator <http://viaf.org/viaf/81663420> ; # "Lee, Mabel"
  schema:inLanguage "en";
  schema:name "Soul Mountain"@en ;
  schema:translationOfWork <http://worldcat.org/entity/work/id/1215997>
Work sets
- Series
- Editions
- Translations
- Publishers
- Subjects
- Classifications
- Materials
- Library holdings
- ...

Book instances
- Multilingual labels
- First publication date
- Original language and script of work
- First line
- Freebase ID
- MusicBrainz work ID
- ...

Multilingual labels
Other identifiers
Common descriptions
Original scripts

Work ID
Richer bibliographic descriptions
Translations
Library related identifiers

Courtesy of Shenghui Wang
Meeting the challenges

• WorldCat has more translations than any other resource.
• We can tag data elements from different languages of cataloging.
• We can ingest data from other linked data sources to present information in the preferred language and script of the user and associate translations to the original work.
Thank you!

Karen Smith-Yoshimura
smithyok@oclc.org
@KarenS_Y

Together we make breakthroughs possible.