An OpenRefine and XSL transformation workflow

Evaluation

Celian RINGWALD

THE HUMANITIES DATA LAB

Three Case Studies:

1. The CraikSiteIndex

University of Calgary

2. Devonshire Manuscript

University of Victoria

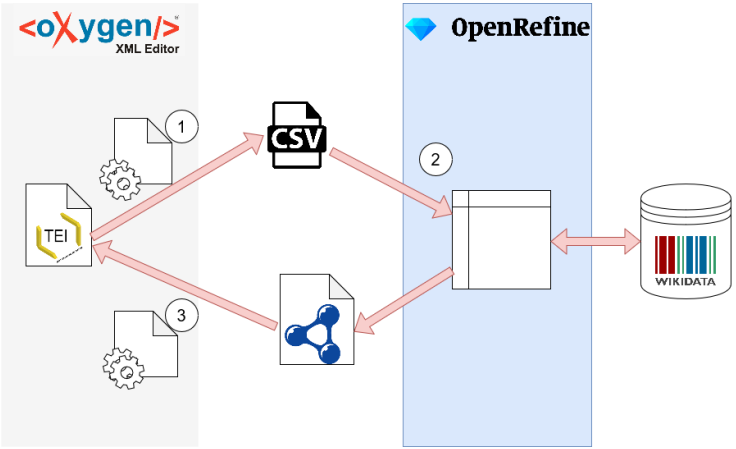
3. LGLC

LGLC

LESBIAN AND GAY LIBERATION IN CANADA

University of Ottawa

The global approach :
a mix of scripted and manual steps



1. We will export content we want to enrich from the TEI file as a CSV file [via XSLT]
2. We will use this file for getting via OpenRefine the interesting information that we want and then exporting it into a RDF file [manual]
3. Finally we will create a base XSLT containing a xsl map from the RDF file that we can use to inject data back into the TEI file [via XSLT]

But as every encoded project has different goals and each TEI file is different, we will have to adapt our process to each new TEI file!

Entities stats of our examples

Corpus					
CraikSiteIndex	273	306	55	31	5
Devonshire Manuscript	22	0	0	0	0
LGLC project	0	1216	103	1425	0

Person reconciliation

TEI Persons prosopograhic data

Wikidata ontology

Q5 : humans

```xml
<listPerson>
  <person xml:id="DMC" sex="2">
    <persName>
      <surname type="married">Craik</surname>
      <surname type="maiden">Mulock</surname>
      <forename>Dinah</forename>
      <forename>Maria</forename>
    </persName>
    <birth when="1826"/>
    <death when="1887"/>
    <occupation>Writer</occupation>
    <nationality>English</nationality>
    <event type="marriage" when="1865">
      <label>Marriage</label>
      <desc><persName ref="#DMC">Dinah Maria Mulock</persName> married <persName
          ref="#GeorgeCraik">George Lillie Craik.</persName></desc>
    </event>
  </person>
```

- P21: gender
- P725: given name
- P569 : birth date
- P570 : death date
- P106 : occupation
- P27 : citizen of

Place reconcialiation

Depending of the project, places could mean
a vaste variety of geographic appelation,
a place could be :

- Delimited area: Continent / Country / Region / Province / City ….
- An adress, a street, a building
- Could have different admited names

….
And could also be :

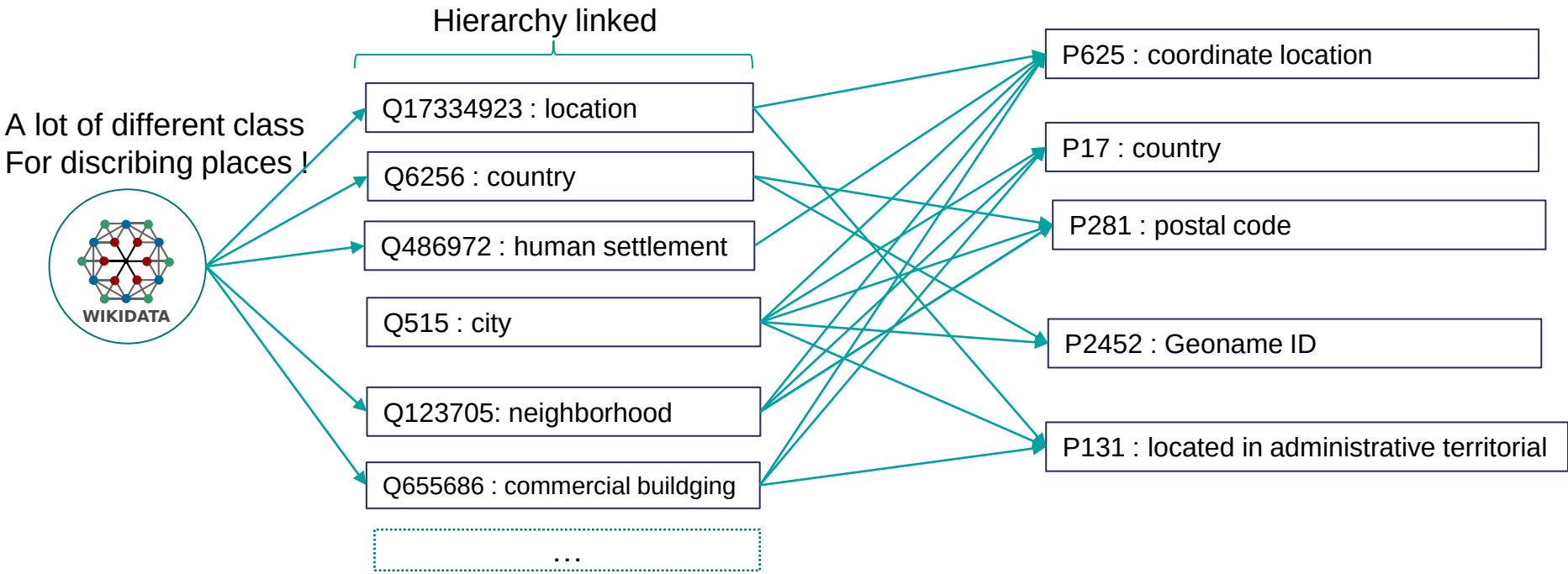
- An overlapping place, not officially limited
- A lapsed place

Theses entities are generally
Ambiguous !

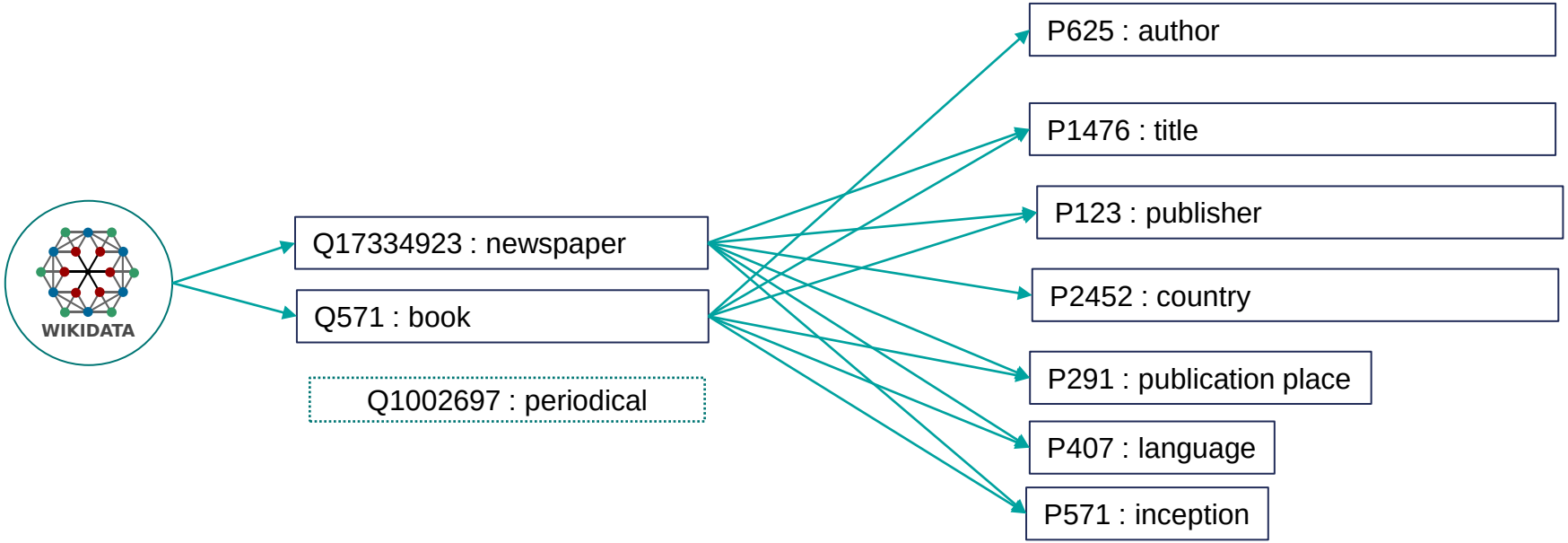
Wikidata offer an ontology for describing them, we will generally
need to discrimine places refered in a TEI
project by working on one by one type of entity.

We will give some clues to map with the Wikidata ontology depending on the type of the entity.

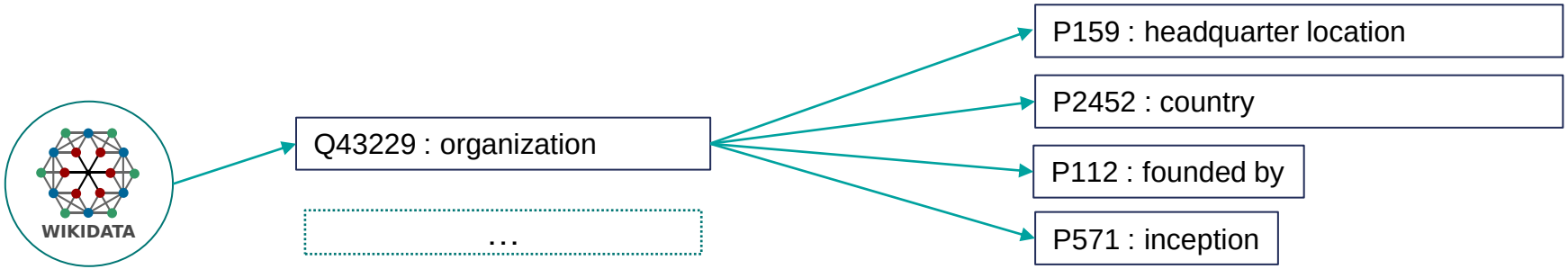
Place reconcialiation



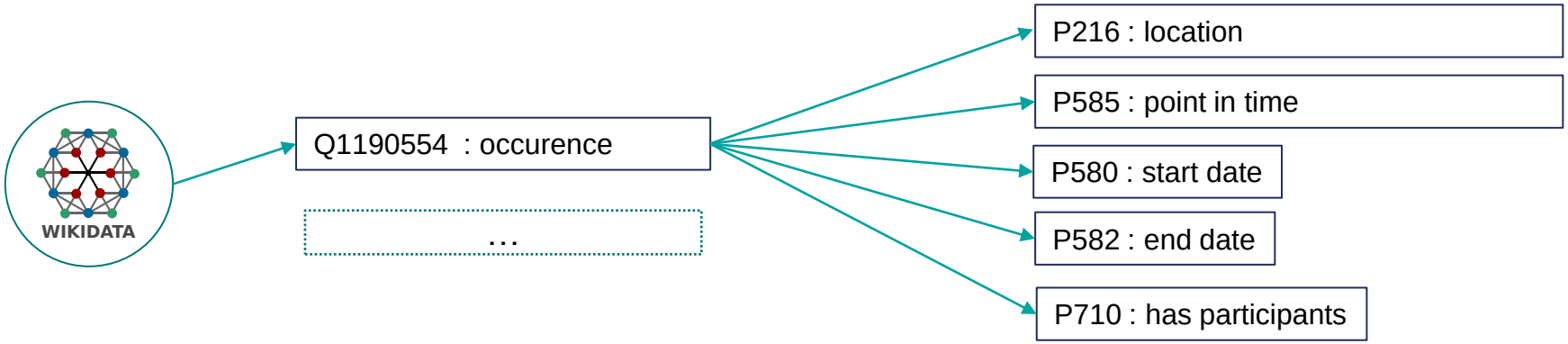
Document reconcialiation



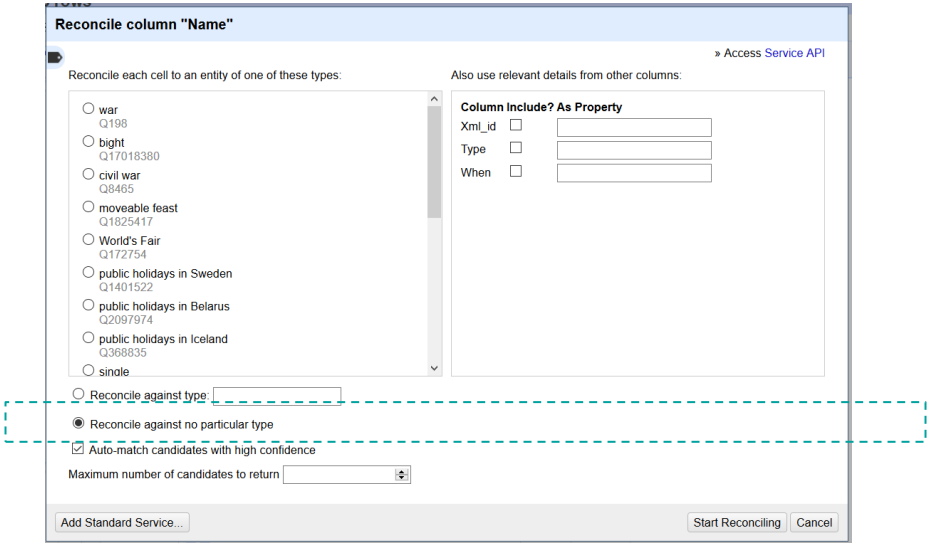
Organization reconcialiation



Event reconcialiation



What could we do if we havn't enough clues ?



We could be tempted to each time using this technique, but this one is very time consuming…

Results on our corpora : peoples

	Nb entities	authority ref average	Wikidata	VIAF	ISNI	GETTY
CraikIndex (manual and typed)	273	0,80	36 %	34 %	0%	10%
CraikIndex (only precise type unchecked)	273	0.46	15%	13%	12%	5%
CraikIndex (any type unchecked)	273	0.54	20%	16%	13%	4%
Devonshire (manual and typed)	22	1.8	100 %	86%	0%	19%
Devonshire (only precise type unchecked)	22	2	86%	63%	63%	13%
Devonshire (any type unchecked)	22	0.27	9%	9%	9%	0%

Results on our corpora : places

	Nb of entities	authority ref average	Wikidata	GEONAMES	GETTY
CraikIndex (checked)	306	0,87	34 %	34 %	10%
CraikIndex (only precise type unchecked)	306	0.15	9%	5%	6.5%
CraikIndex (any type unchecked)	306	0.08	4%	3%	0%
LGLC (checked)	1216	0,82	45%	35%	3%
LGLC (only precise type unchecked)	1216	0.04	3%	1%	0%
LGLC (any type unchecked)	1216	0.14	10%	34%	0%

Results on our corpora : document

	Nb of entities	authority ref average	Wikidata	VIAF	GETTY	ISNI
CraikIndex (checked)	55	0,07	4 %	4 %	0%	0%
CraikIndex (only precise type unchecked)	55	0.05	5%	0%	0%	0%
CraikIndex (any type unchecked)	55	0%	0%	0%	0%	0%
LGLC (checked)	103	0,8	9%	0%	0%	0%
LGLC (only precise type unchecked)	103	0	0%	0%	0%	0%
LGLC (any type unchecked)	103	0	0%	0%	0%	0%

Results on our corpora : organization

	NB of enties	authority ref average	Wikidata	VIAF	GETTY	ISNI
CraikIndex (checked)	31	0,32	16 %	16 %	0%	0%
CraikIndex (only precise type unchecked)	31	1	35%	32%	0%	32%
CraikIndex (any type unchecked)	31	0,6	29%	22%	0%	12%
LGLC (checked)	1425	0.08	4%	2%	0%	2%
LGLC (only precise type unchecked)	1425	0.1	4%	3%	0%	2%
LGLC (any type unchecked)	1425	0.28	16 %	7%	0%	4%

Results on our corpora : event

	Nb of entities	authority ref average	Wikidata	VIAF	GETTY	ISNI
CraikIndex (checked)	5	1.2	100 %	20 %	0%	0%
CraikIndex (only precise type unchecked)	5	0	0%	0 %	0 %	0 %
CraikIndex (any type unchecked)	5	0.2	20 %	0 %	0 %	0 %

Is Q1190554 type (occurrence) the good one to match ?

Conclusions

-> Could be very long to reconcielling via Wikidata API : need to think about local system for industrialization

-> Extremelly depending about Wikidata entity type choosen to link

-> Reasonable working for person / places

-> About organization : better without typing ?

-> Document : need we to precise more about what we wanna link ? We need to test the VIAF API.

-> A global processus to think, not only about openrefine part but also about the linkage optimisation :
in time and in accuracy.

Questions

> Thinking about optimized SPARQL queries depending on a entity typing prediction ?
> Global context seems to be catch when we try to reconciele could we go deeper in that ?