Extracting Linked Data from statistic spreadsheets

Tien-Duc Cao tien-duc.cao@inria.fr
Ioana Manolescu ioana.manolescu@inria.fr
Xavier Tannier xtannier@limsi.fr

Semantic Big Data workshop, Chicago, May 19th, 2017
Agenda

1. Context: data journalism and journalistic fact-checking
2. Research problem: extracting linked open data from spreadsheets
3. Approach
4. Results
5. Future work
1. Fact-checking is a content management problem

- **Media content**
- **Claim to be checked (text or data)**
- **Human actors** (journalists, experts, crowd workers)

**Verification tool** (query, match, source search...)

- **Reference information source 1**
- **Reference information source 2**
- **...**
- **Reference information source n**

**Analysis result** « True / rather true / rather false / false


Tien-Duc CAO, Ioana Manolescu, Xavier Tannier
"Extracting linked data from statistic spreadsheets" 19/05/2017
1. Fact-checking is a content management problem

- **Claim extraction**
  - Claim to be checked (text or data)
  - Human actors (journalists, experts, crowd workers)
  - Reconciliation, reputation

- **Verification tool**
  - (query, match, source search...)

- **Source search / source selection**

- **Social network analysis**

- **Reference information source**
  - Reference information source 1
  - Reference information source 2
  - …
  - Reference information source n

- **Reference source construction, refinement, integration**

- Analysis result
  - « True / rather true / rather false / false 

---

Tien-Duc CAO, Ioana Manolescu, Xavier Tannier
"Extracting linked data from statistic spreadsheets"
1. Context

• Which **data source** can help us to fact-check a **statistical claim** from the media?
  
  • E.g: “The unemployment rate in France last year was 50%?”

• This work is a part of ContentCheck ¹ project

¹ https://team.inria.fr/cedar/contentcheck/
2. Research problem: high-quality reference data

- **National statistic institutes such as** INSEE, France’s economic and societal statistics institute are often valuable data providers.
2. The road to high quality data...

Unfortunately most of the data published by INSEE looks like this (our text coloring):

<table>
<thead>
<tr>
<th>l</th>
<th>c</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>The data reflects children born alive in 2015...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Region</td>
<td>Department</td>
<td>16-20</td>
<td>21-25</td>
<td>26-30</td>
<td>31-35</td>
<td>36-40</td>
<td>41-45</td>
<td>46-50</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Île-de-France</td>
<td>Essonne</td>
<td>215</td>
<td>1230</td>
<td>5643</td>
<td>4320</td>
<td>3120</td>
<td>1514</td>
<td>673</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Val-de-Marne</td>
<td>175</td>
<td>987</td>
<td>4325</td>
<td>3156</td>
<td>2989</td>
<td>1740</td>
<td>566</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Rhône-Alpes</td>
<td>Ain</td>
<td>76</td>
<td>1103</td>
<td>3677</td>
<td>2897</td>
<td>1976</td>
<td>1464</td>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>Ardèche</td>
<td>45</td>
<td>954</td>
<td>2865</td>
<td>2761</td>
<td>1752</td>
<td>1653</td>
<td>523</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mother's age at the time of the birth

Age below 30  Age above 31

Tien-Duc CAO, Ioana Manolescu, Xavier Tannier
"Extracting linked data from statistic spreadsheets" 19/05/2017
2. The road to high quality data...

Sometimes there are more than 1 table per sheet
3. Extraction approach

Tien-Duc CAO, Ioana Manolescu, Xavier Tannier
"Extracting linked data from statistic spreadsheets"
3. Extraction approach

Tien-Duc CAO, Ioana Manolescu, Xavier Tannier
"Extracting linked data from statistic spreadsheets"
3. Approach: finding table boundaries

1. Original sheet

Table 1
Table 2
Table 3
Table 4

2. Extract tables from column 0 to k1

Table 1
Table 2
Table 3
Table 4

3. Extract tables from column k2 to k3

Table 1
Table 2
Table 3
Table 4

Consider these cells as empty

Tien-Duc CAO, Ioana Manolescu, Xavier Tannier
"Extracting linked data from statistic spreadsheets"
3. Extraction approach

Tien-Duc CAO, Ioana Manolescu, Xavier Tannier
"Extracting linked data from statistic spreadsheets" 19/05/2017
3. Approach: table extractor

- Header cells *mostly* contain texts
- Their positions are at:
  - the top (header rows) of table
  - the left (header columns) of table
- Having more than 1 header rows/columns indicates **data aggregation**
- Data cells *mostly* contain numeric values

![Table Example]

---

Tien-Duc CAO, Ioana Manolescu, Xavier Tannier
"Extracting linked data from statistic spreadsheets" 19/05/2017
3. Approach: table extractor

1. We distinguish header/data row/columns using
   • data type of its cells (text, number, special value to indicate a missing value, null for empty cell)
   • formatting information of its cells: cell’s border, cells belong to merged cell
   • the types of its neighbor rows/columns

2. Based on these we identify the exact structure of each table
3. Conceptual data model

Tien-Duc CAO, Ioana Manolescu, Xavier Tannier
"Extracting linked data from statistic spreadsheets"
4. Results

- Collected **16011** Excel spreadsheets, extracted **74117** tables.
- Accuracy evaluation:
  - We selected randomly 100 Excel files $\rightarrow$ 2432 tables
  - We visually identified the header cells, data cells and header hierarchy and then compared with those obtained from our system.

<table>
<thead>
<tr>
<th>Category</th>
<th>Number</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tables correctly extracted</td>
<td>2214</td>
<td>91%</td>
</tr>
<tr>
<td>Tables incorrectly extracted</td>
<td>218</td>
<td>9%</td>
</tr>
</tbody>
</table>
4. Sample extracted RDF

<table>
<thead>
<tr>
<th>Postes salariés par secteur d’activité au 31 décembre 2013</th>
<th>Nombre d’établissements actifs</th>
<th>Postes salariés</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agriculture</td>
<td>690</td>
<td>583</td>
</tr>
<tr>
<td>Pays</td>
<td>2.1</td>
<td>2.6</td>
</tr>
<tr>
<td>Poids Pays / Bretagne (en %)</td>
<td>3,6</td>
<td>3,384</td>
</tr>
<tr>
<td>Industry</td>
<td>589</td>
<td>583</td>
</tr>
<tr>
<td>Pays</td>
<td>3,6</td>
<td>3,348</td>
</tr>
<tr>
<td>Poids Pays / Bretagne (en %)</td>
<td>3,6</td>
<td>3,348</td>
</tr>
<tr>
<td>Construction</td>
<td>1 051</td>
<td>1 812</td>
</tr>
<tr>
<td>Pays</td>
<td>3,8</td>
<td>2,5</td>
</tr>
<tr>
<td>Poids Pays / Bretagne (en %)</td>
<td>3,6</td>
<td>2,5</td>
</tr>
<tr>
<td>Commerce</td>
<td>1 558</td>
<td>3 385</td>
</tr>
<tr>
<td>Pays</td>
<td>3,7</td>
<td>2,4</td>
</tr>
<tr>
<td>Poids Pays / Bretagne (en %)</td>
<td>3,7</td>
<td>2,4</td>
</tr>
<tr>
<td>Services</td>
<td>8 415</td>
<td>12 078</td>
</tr>
<tr>
<td>Pays</td>
<td>3,7</td>
<td>1,9</td>
</tr>
<tr>
<td>Poids Pays / Bretagne (en %)</td>
<td>3,6</td>
<td>1,9</td>
</tr>
<tr>
<td>Total</td>
<td>10 303</td>
<td>21 184</td>
</tr>
</tbody>
</table>

Champ : ensemble des établissements hors défense et particuliers employeurs

Source : Insee, Clap
5. Future work

Verification tool (query, match, source search...)

Source search / source selection

Reference information source 1
Reference information source 2
Reference information source n

Reference source construction, refinement, integration
Thanks / questions?

Excel files and extracted RDF files
(10.5GB will be expired in May 29\textsuperscript{th} 2017)

https://goo.gl/4Y5Dtv

Source code: no expiration date :)

https://gitlab.inria.fr/cedar/insee-crawler
https://gitlab.inria.fr/cedar/excel-extractor