Extracting Linked Data from statistic spreadsheets

Tien-Duc CAO, Ioana MANOLESCU, Xavier TANNIER http://team.inria.fr/cedar

Main idea: understanding and publishing INSEE data

Spreadsheets from INSEE

- French economic and societal data
- They publish PDF, HTML, Excel files

Extracting RDF data from spreadsheets

- Data is organized as table
  - header cells: text
  - data cells: number
- Heterogeneous

l \ c	1	2	3	4	5	6	7	8	9	10
1	The data reflects children born alive in 2015...									
2										
3			Mother's age at the time of the birth							
4			Age below 30			Age above 31				
5	**Region**	**Department**	**16-20**	**21-25**	**26-30**	**31-35**	**36-40**	**41-45**	**46-50**	
6	Île-de-France	Essonne	215	1230	5643	4320	3120	1514	673	
7		Val-de-Marne	175	987	4325	3156	2989	1740	566	
8		…	…	…	…	…	…	…	…	…
9	Rhône-Alpes	Ain	76	1103	3677	2897	1976	1464		
10		Ardèche	45	954	2865	2761	1752	1653	523	
11		…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…

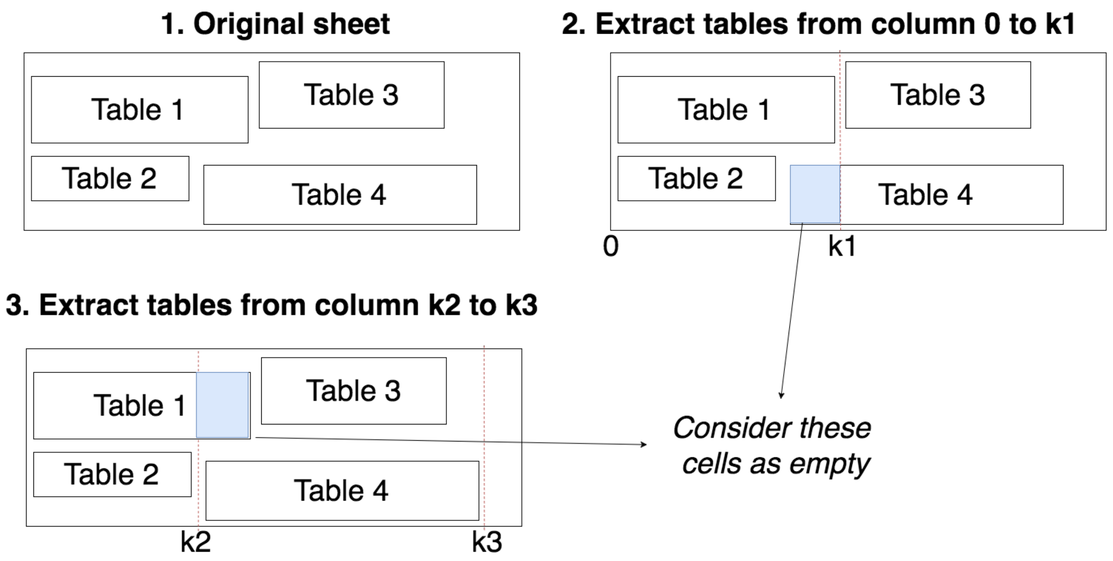
Applications: Data journalism and fact-checking

Collaboration Inria - LIMSI

ANR ContentCheck project

Le Monde — LES DÉCODEURS VENONS-EN AUX FAITS

Algorithm outline

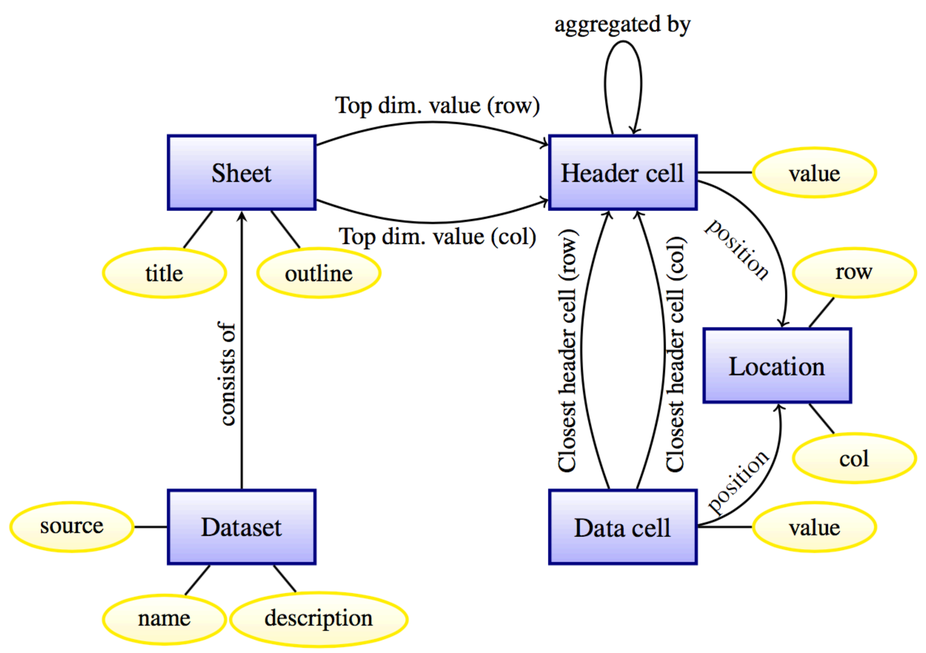


- Each sheet of spreadsheet is a matrix *S* of *M* rows and *N* columns. More than 1 table could appear in the same sheet.
- **Step 1**: identify a rectangle *R* (left = 0, right = *k*, top = 0, bottom = *N* – 1) where we can extract tables
- **Step 2**: identify horizontal boundaries which separate tables in *R* and then extract tables
- **Step 3**: identify a new rectangle *R* and come back to step 1

Table extraction algorithm

- Each item the matrix *S* represents a cell of spreadsheet.
- Based on cell's value (could be text, number, or just an empty cell) and cell's formatting information (whether this cell belongs to a merged cell, how many borders this cell has), we can infer the type of each row and column.
- From the type of row/column, we can identify the data cells and header cells.

Conceptual model



RESULTS

Extraction

- Crawled INSEE website and found 20k+ Excel files
- Extracted 70k+ tables
- Version 2 (able to extract multiple tables per sheet) development ongoing

Evaluation

- Version 1: assume that each sheet contains only 1 table
- We selected randomly 100 Excel files. They contained 2432 tables.
- For these 100 files, we visually identified the header cells, data cells and header hierarchy, which we compared with those obtained from our system.
- We consider a table is "correctly extracted" when all these are pairwise equal; otherwise, the table is "incorrectly extracted".

Category	Number	%
Tables correctly extracted	2214	91%
Tables incorrectly extracted	218	9%

Sample RDF output

```
@prefix insee: <http://insee.excel/> .

<http://insee.excel/File:File_education_scolarite>
    insee:name "Éducation-Scolarité.xls" ;
    insee:description "Effectifs scolarisés du second degré (*)" .
<http://insee.excel/Sheet:Sheet_0> insee:title "Éducation - Scolarité" ;
    insee:belongsTo <http://insee.excel/File:File_education_scolarite> .
<http://insee.excel/HeaderCellY:HeaderCellY_0> insee:value "2014-2015" ;
    insee:aggregate <http://insee.excel/HeaderCellY:HeaderCellY_1> ;
    insee:YHierarchy <http://insee.excel/Sheet:Sheet_0> .
<http://insee.excel/HeaderCellY:HeaderCellY_1> insee:value "Rentrée scolaire" ;
    insee:YHierarchy <http://insee.excel/Sheet:Sheet_0> .
<http://insee.excel/HeaderCellY:HeaderCellY_2> insee:value "2015-2016" ;
    insee:aggregate <http://insee.excel/HeaderCellY:HeaderCellY_1> ;
    insee:YHierarchy <http://insee.excel/Sheet:Sheet_0> .
<http://insee.excel/HeaderCellX:HeaderCellX_6> insee:value "Pays" ;
    insee:XHierarchy <http://insee.excel/Sheet:Sheet_0> .
<http://insee.excel/DataCell:DataCell_0> insee:value 6817 ;
    insee:posX 0 ;
    insee:posY 0 ;
    insee:closestXCell <http://insee.excel/HeaderCellX:HeaderCellX_6> ;
    insee:closestYCell <http://insee.excel/HeaderCellY:HeaderCellY_0> .
<http://insee.excel/DataCell:DataCell_1> insee:value 6951 ;
    insee:posX 0 ;
    insee:posY 1 ;
    insee:closestXCell <http://insee.excel/HeaderCellX:HeaderCellX_6> ;
    insee:closestYCell <http://insee.excel/HeaderCellY:HeaderCellY_2> .
```

Comparison with related work

Chen et at [1] built a system that could automatically extract relational data from spreadsheets. Their machine learning approach requires manual labeling 27,531 non-empty rows in spreadsheets as title, header, data or footnote. The model obtained a precision of 92.1% for top headers and 85.2% for left headers.

Reference

[1] Zhe Chen and Michael Cafarella. 2013. Automatic Web Spreadsheet Data Extraction. In *Proceedings of the 3rd International Workshop on Semantic Search Over the Web (Semantic Search)*. ACM, New York, NY, USA, Article 1, 8 pages. DOI: http://dx.doi.org/10.1145/2509908.2509909