Extracting Linked Data from statistic spreadsheets

Tien-Duc Cao tien-duc.cao@inria.fr

Ioana Manolescu ioana.manolescu@inria.fr

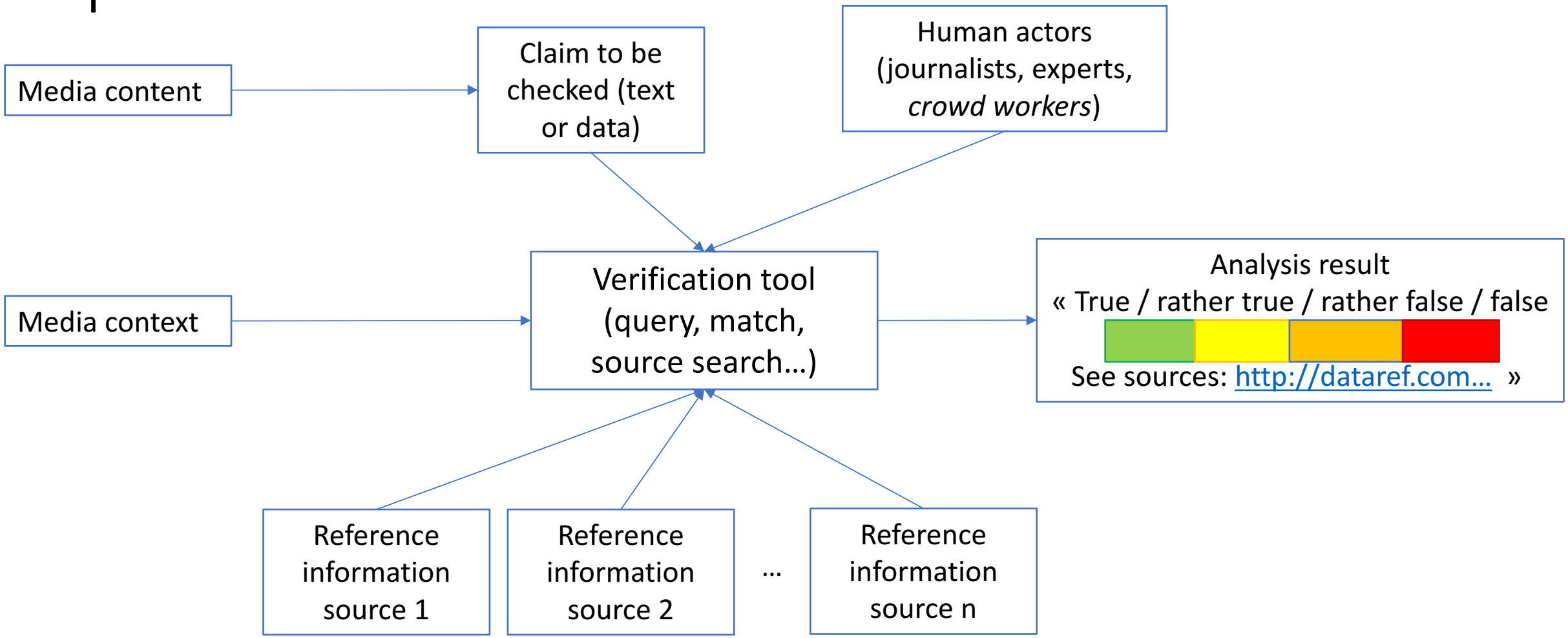
Xavier Tannier xtannier@limsi.fr

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

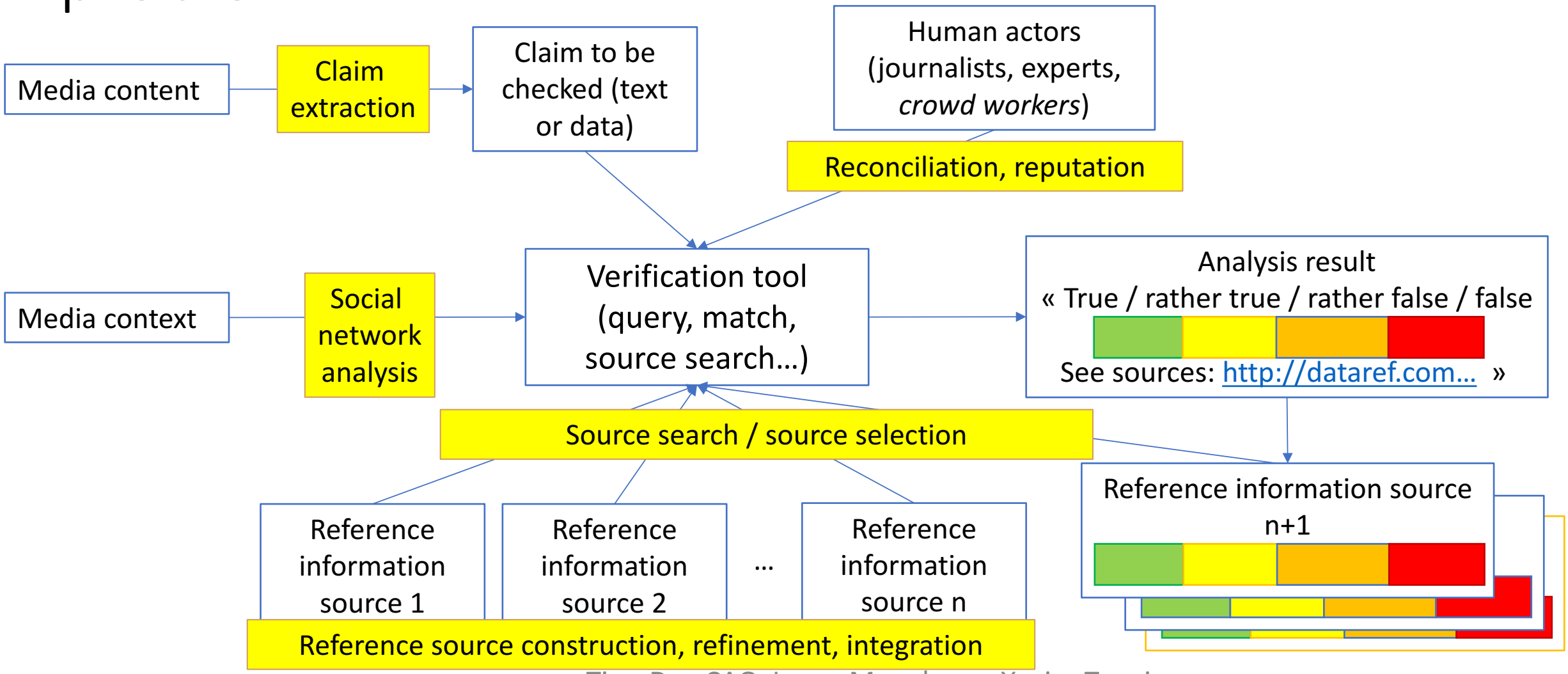
Agenda

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



1. Fact-checking is a content management problem



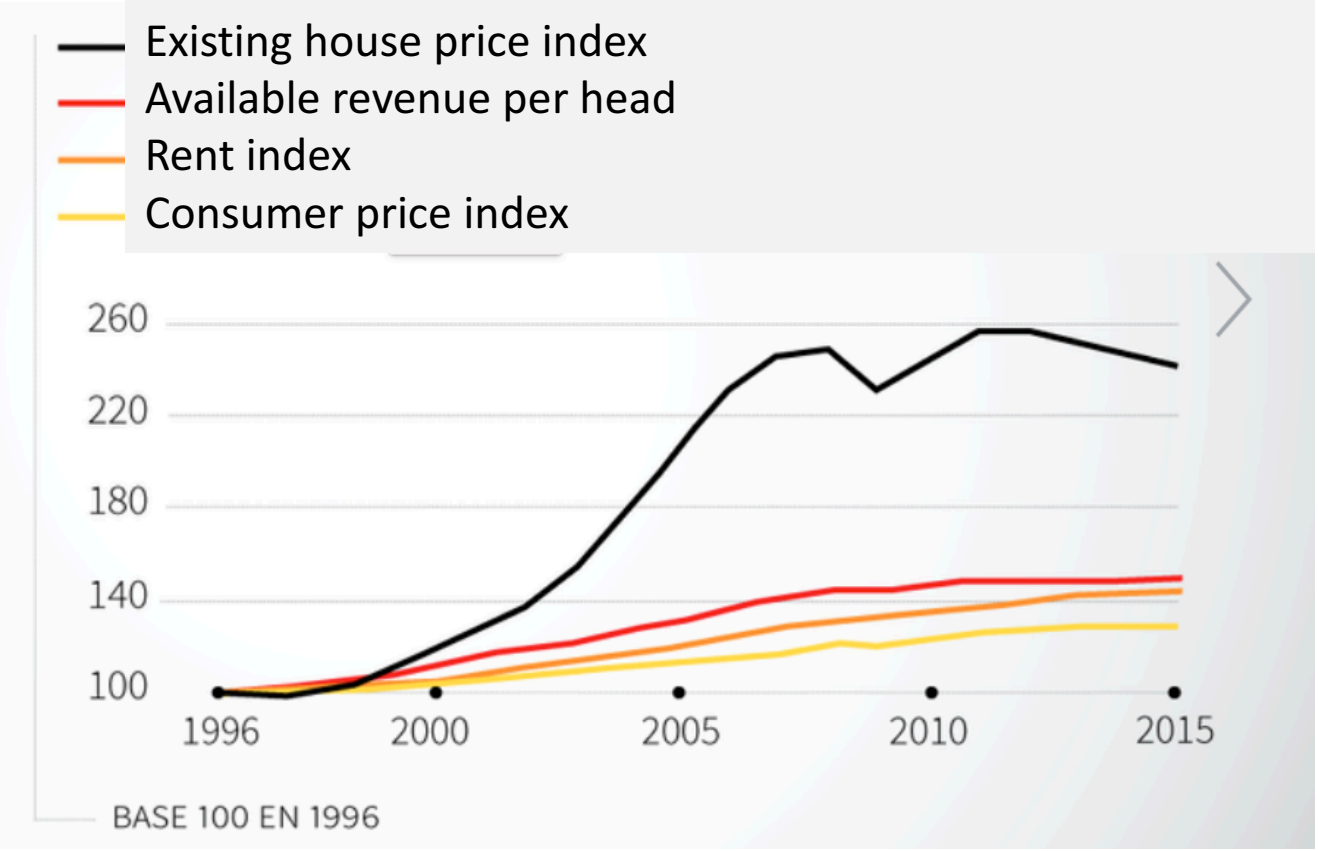
1. Context

- 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 [1], France's economic and societal statistics institute are often valuable data providers



[1] https://insee.fr/

http://abonnes.lemonde.fr/les-decodeurs/portfolio/2017/04/18/les-fractures-francaises-1-5-le-logement-les-raisons-de-la-crise_5112859_4355770.html

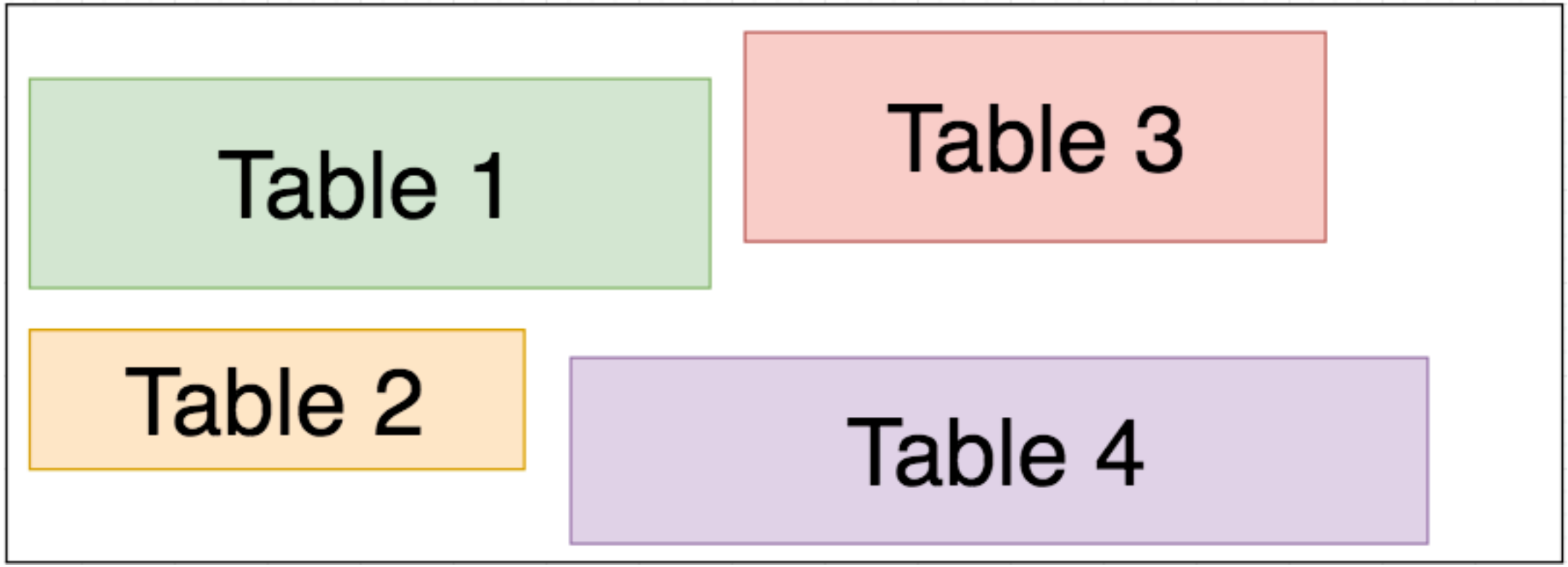
2. The road to high quality data…

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

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		…	…	…	…	…	…	…	…	
…	…	…	…	…	…	…	…	…	…	

2. The road to high quality data…

Sometimes there are more than 1 table per sheet



3. Extraction approach

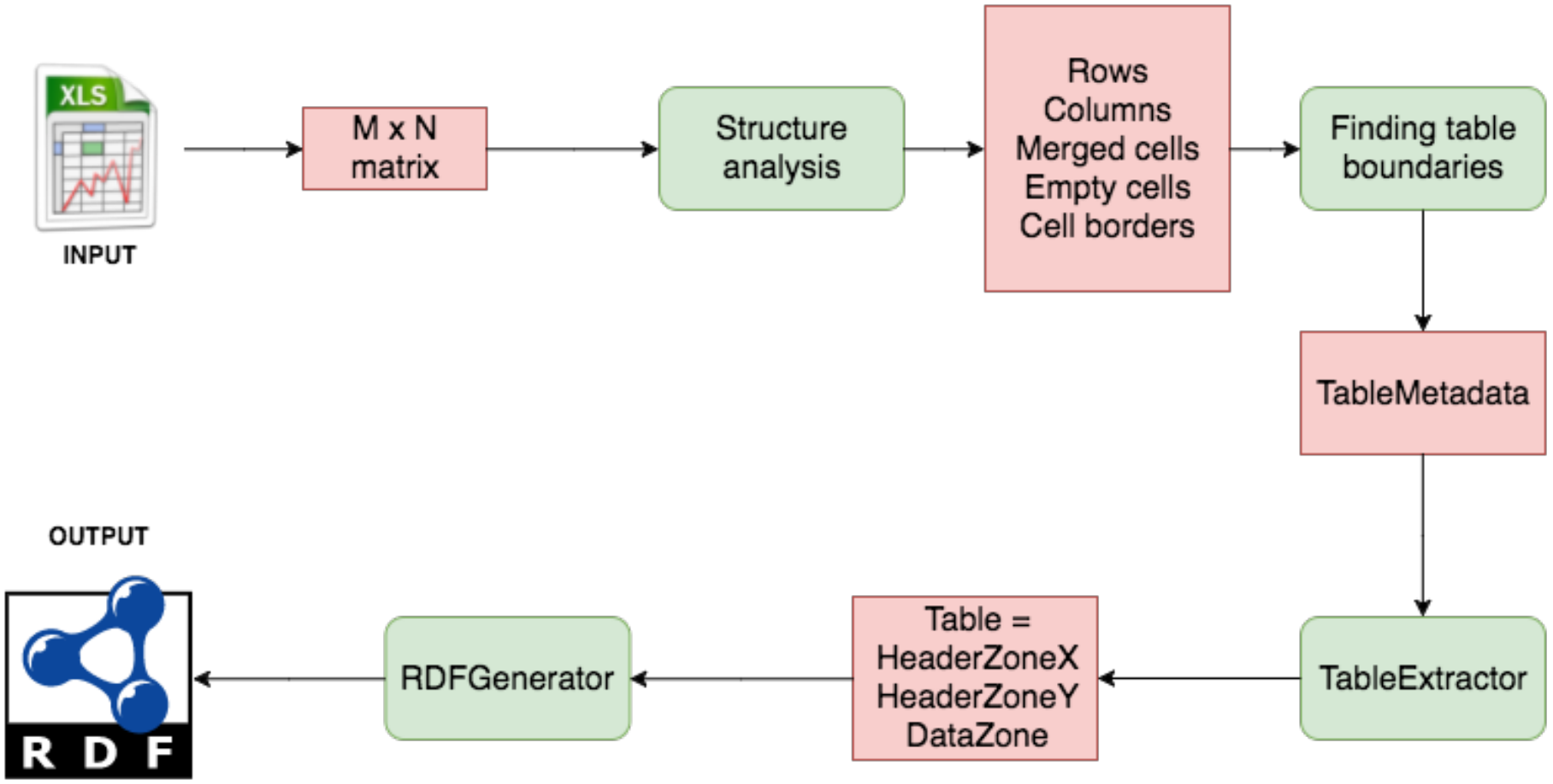


Image sources:
https://www.iconfinder.com/icons/7661/excel_microsoft_word_xls_icon#size=128
https://www.w3.org/RDF/icons/rdf_w3c.svg

3. Extraction approach

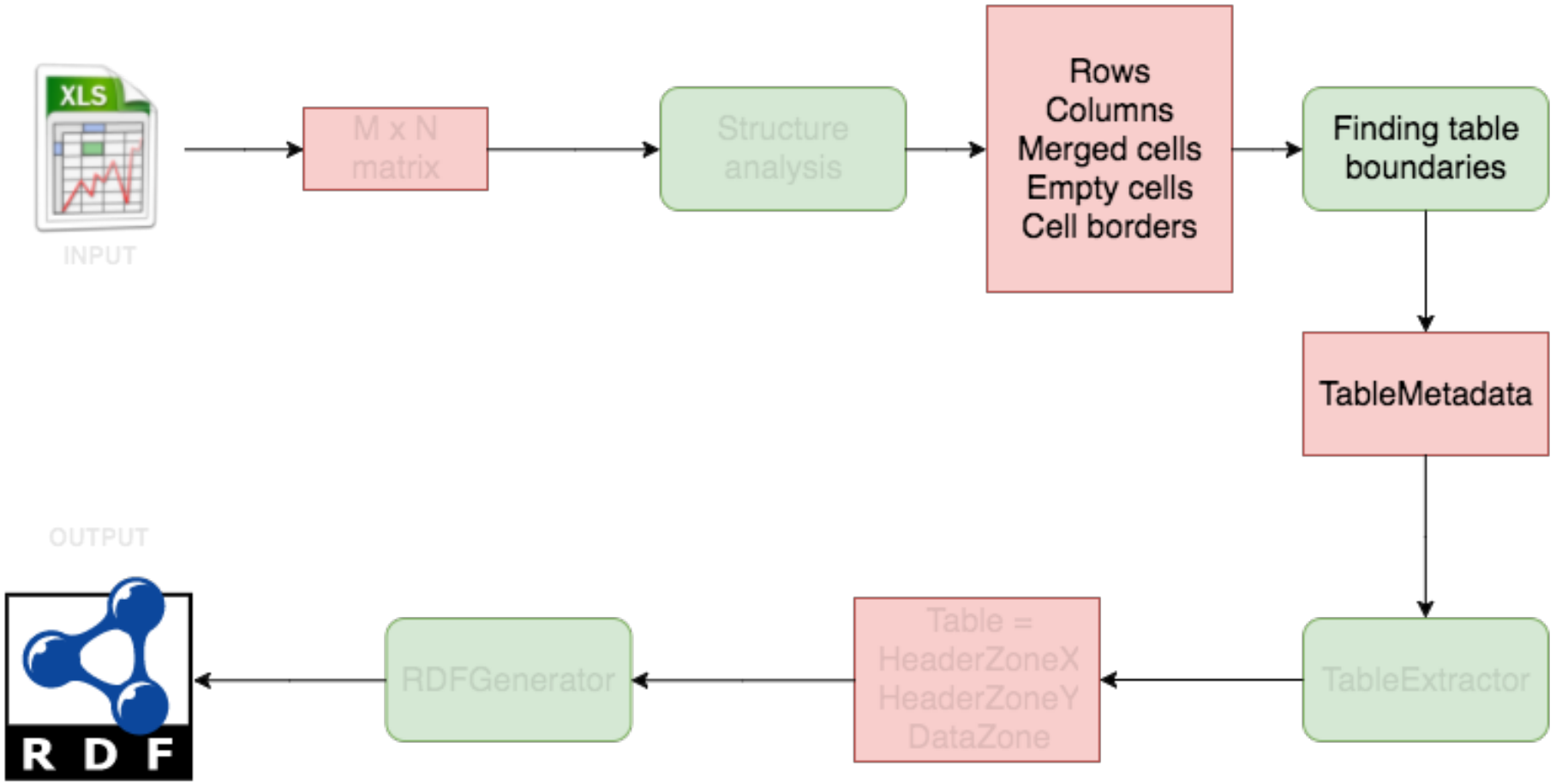
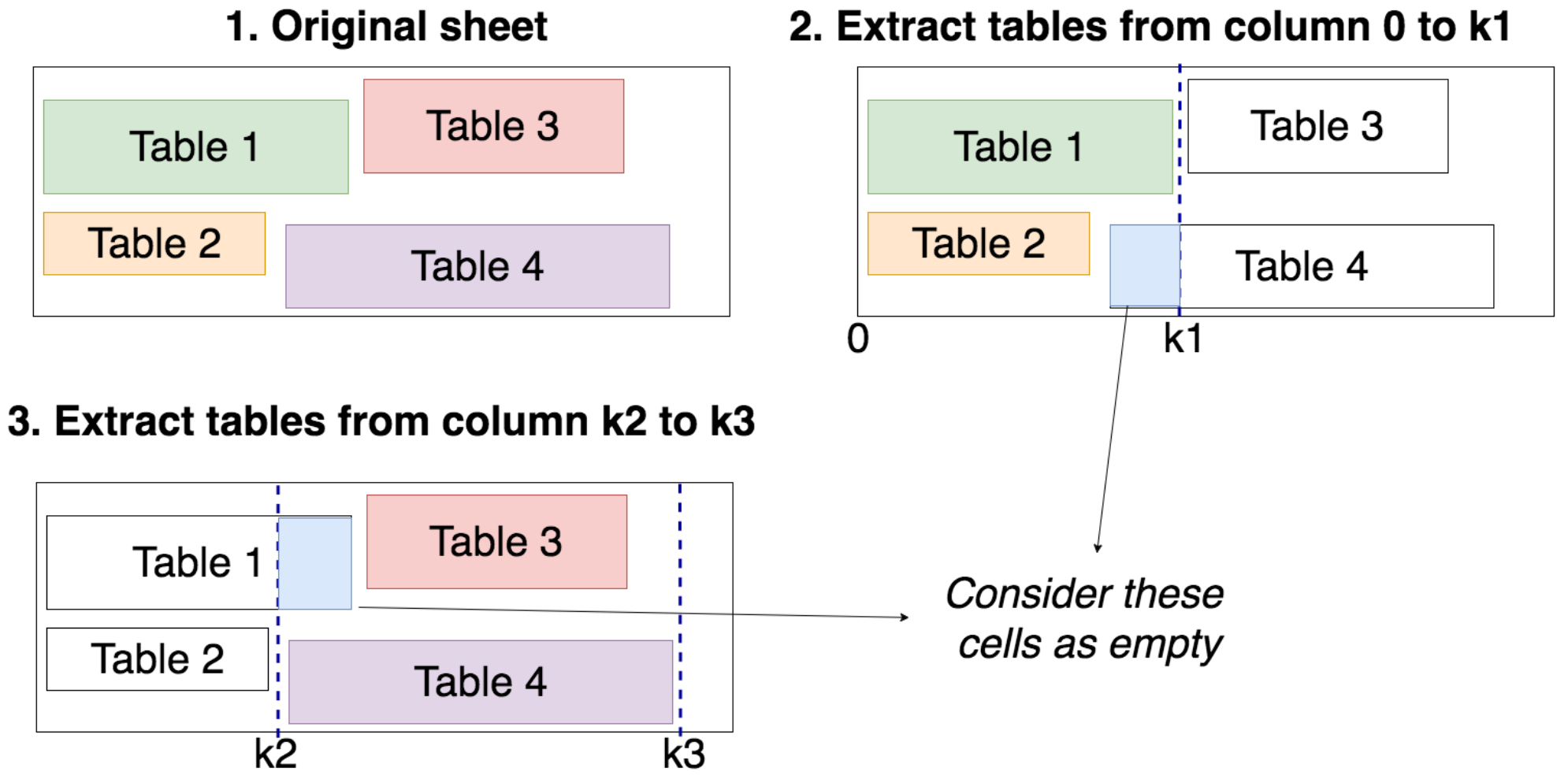


Image sources:
https://www.iconfinder.com/icons/7661/excel_microsoft_word_xls_icon#size=128
https://www.w3.org/RDF/icons/rdf_w3c.svg

3. Approach: finding table boundaries



3. Extraction approach

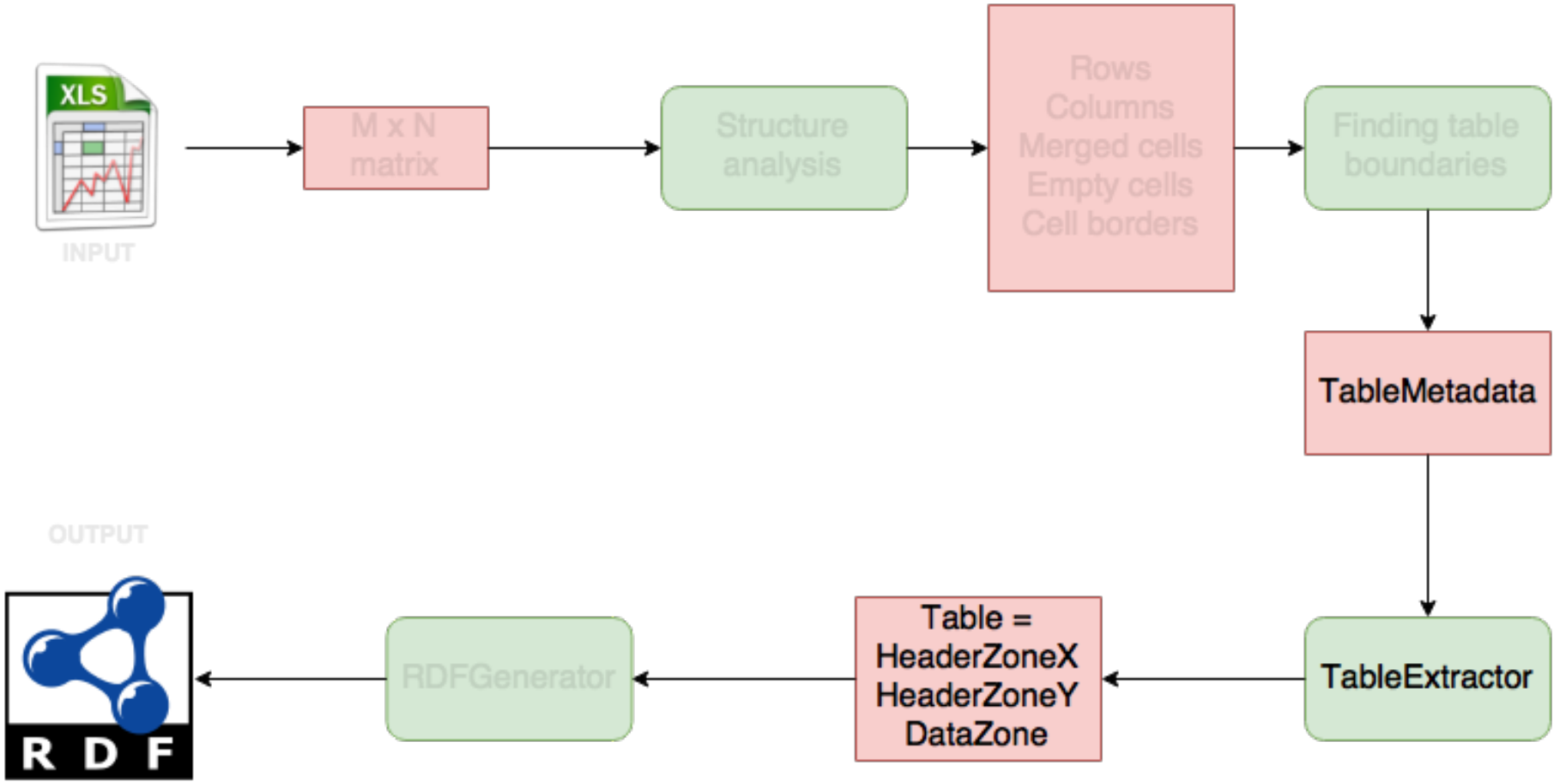


Image sources:
https://www.iconfinder.com/icons/7661/excel_microsoft_word_xls_icon#size=128
https://www.w3.org/RDF/icons/rdf_w3c.svg

3. Approach: table extractor

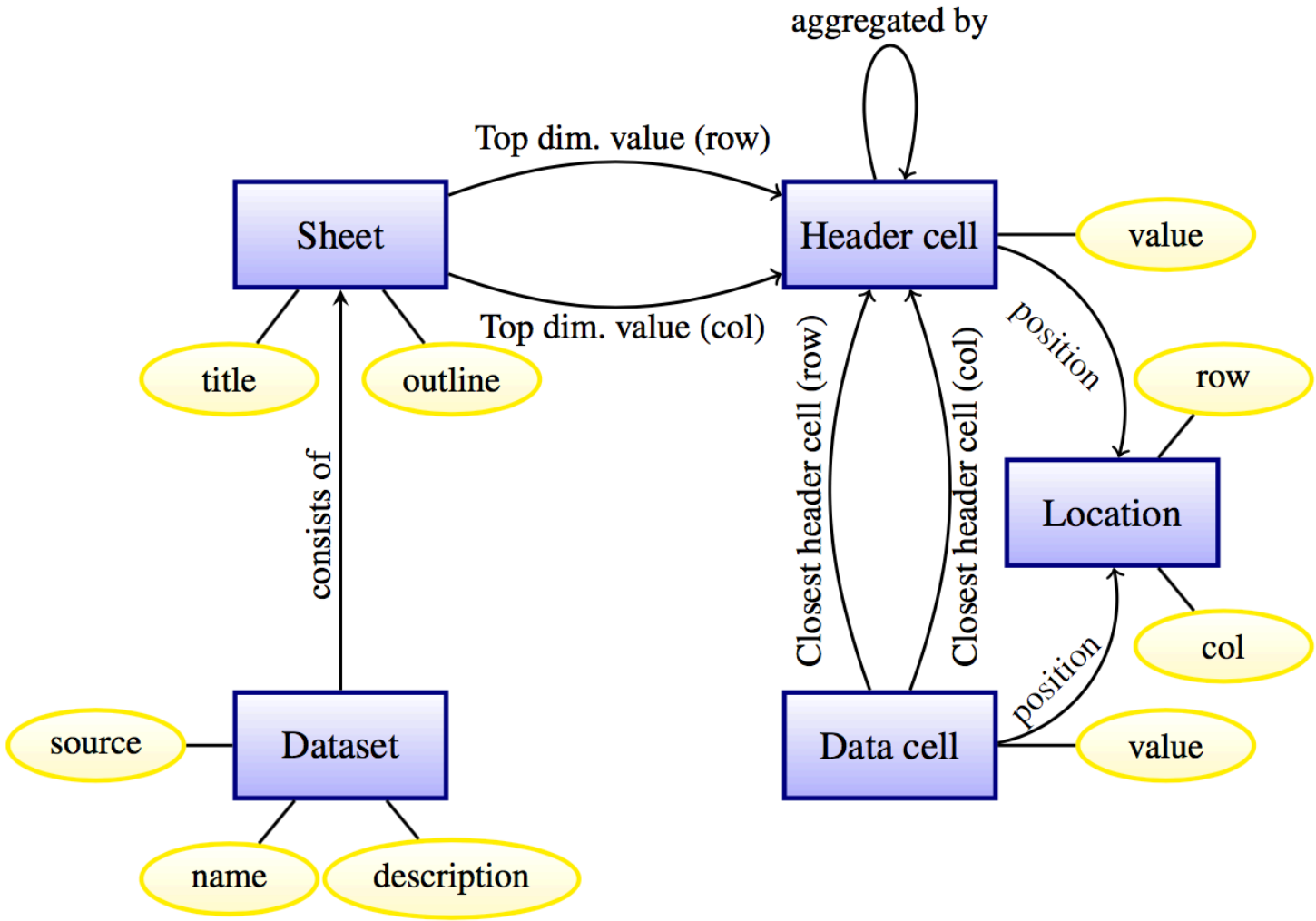
- Header cells *mostly* contain texts
- Their positions are at:
  - the top (header rows) of table
  - the left (header columns) of table
- Having more than 1 header rows/columns indicates **data aggregation**
- Data cells *mostly* contain numeric values

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		...	...	...	...	...	...	...	...	
...	...	...	...	...	...	...	...	...	...	

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



4. Results

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

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

4. Sample extracted RDF

```
@prefix inseeXtr: <http://inseeXtr.excel/> .
@prefix rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

inseeXtr:YHierarchy rdfs:range inseeXtr:HeaderCell ;
    rdfs:domain inseeXtr:HeaderCell .
inseeXtr:XHierarchy rdfs:range inseeXtr:HeaderCell ;
    rdfs:domain inseeXtr:HeaderCell .
inseeXtr:closestYHeaderCell rdfs:range inseeXtr:DataCell ;
    rdfs:domain inseeXtr:HeaderCell .
inseeXtr:closestXHeaderCell rdfs:range inseeXtr:DataCell ;
    rdfs:domain inseeXtr:HeaderCell .
inseeXtr:belongsTo rdfs:range inseeXtr:Sheet ;
    rdfs:domain inseeXtr:Dataset .

<http://inseeXtr.excel/File:File_f448f9566734f29343b0a38801221197241dbe17aeb188a686b11c9f> inseeXtr:name "Auray.xls" ;
    rdf:type <http://inseeXtr.excel/:Dataset> ;
    inseeXtr:crawled_date "2017-05-09 01:33:24" ;
    inseeXtr:url "https://www.insee.fr/fr/statistiques/fichier/2386251/Tableaux_pays_bretagne.zip" ;
    inseeXtr:description "Les Pays de la région Bretagne" .
<http://inseeXtr.excel/Sheet:Sheet_6> inseeXtr:title "Postes salariés par secteur d'activité au 31 décembre 2013" ;
    rdf:type <http://inseeXtr.excel/:Sheet> ;
    inseeXtr:comment "Champ : ensemble des établissements hors défense et particuliers employeurs    Source : Insee, Clap" ;
    inseeXtr:belongsTo <http://inseeXtr.excel/File:File_f448f9566734f29343b0a38801221197241dbe17aeb188a686b11c9f> .
<http://inseeXtr.excel/HeaderCellY:HeaderCellY_1> inseeXtr:value "Pays" ;
    rdf:type <http://inseeXtr.excel/:HeaderCell> ;
    inseeXtr:YHierarchy <http://inseeXtr.excel/HeaderCellY:HeaderCellY_5> .
<http://inseeXtr.excel/HeaderCellY:HeaderCellY_2> inseeXtr:value "Poids Pays / Bretagne (en %)" ;
    rdf:type <http://inseeXtr.excel/:HeaderCell> ;
    inseeXtr:YHierarchy <http://inseeXtr.excel/HeaderCellY:HeaderCellY_5> .
<http://inseeXtr.excel/HeaderCellY:HeaderCellY_3> inseeXtr:value "Pays" ;
    rdf:type <http://inseeXtr.excel/:HeaderCell> ;
    inseeXtr:YHierarchy <http://inseeXtr.excel/HeaderCellY:HeaderCellY_6> .
<http://inseeXtr.excel/HeaderCellY:HeaderCellY_4> inseeXtr:value "Poids Pays / Bretagne (en %)" ;
    rdf:type <http://inseeXtr.excel/:HeaderCell> ;
    inseeXtr:YHierarchy <http://inseeXtr.excel/HeaderCellY:HeaderCellY_6> .
<http://inseeXtr.excel/HeaderCellY:HeaderCellY_5> inseeXtr:value "Nombre d'établissements actifs" ;
    rdf:type <http://inseeXtr.excel/:HeaderCell> .
<http://inseeXtr.excel/HeaderCellY:HeaderCellY_6> inseeXtr:value "Postes salariés" ;
    rdf:type <http://inseeXtr.excel/:HeaderCell> .
<http://inseeXtr.excel/HeaderCellX:HeaderCellX_1> inseeXtr:value "Agriculture" ;
    rdf:type <http://inseeXtr.excel/:HeaderCell> .
<http://inseeXtr.excel/HeaderCellX:HeaderCellX_2> inseeXtr:value "Industrie" ;
    rdf:type <http://inseeXtr.excel/:HeaderCell> .
<http://inseeXtr.excel/HeaderCellX:HeaderCellX_3> inseeXtr:value "Construction" ;
    rdf:type <http://inseeXtr.excel/:HeaderCell> .
<http://inseeXtr.excel/HeaderCellX:HeaderCellX_4> inseeXtr:value "Commerce" ;
    rdf:type <http://inseeXtr.excel/:HeaderCell> .
<http://inseeXtr.excel/HeaderCellX:HeaderCellX_5> inseeXtr:value "Services" ;
    rdf:type <http://inseeXtr.excel/:HeaderCell> .
<http://inseeXtr.excel/HeaderCellX:HeaderCellX_6> inseeXtr:value "Total" ;
    rdf:type <http://inseeXtr.excel/:HeaderCell> .
<http://inseeXtr.excel/DataCell:DataCell_0> inseeXtr:value 690.0 ;
    rdf:type <http://inseeXtr.excel/:DataCell> ;
    inseeXtr:posX 0 ;
    inseeXtr:posY 0 ;
    inseeXtr:closestXCell <http://inseeXtr.excel/HeaderCellX:HeaderCellX_1> ;
    inseeXtr:closestYCell <http://inseeXtr.excel/HeaderCellY:HeaderCellY_1> .
```

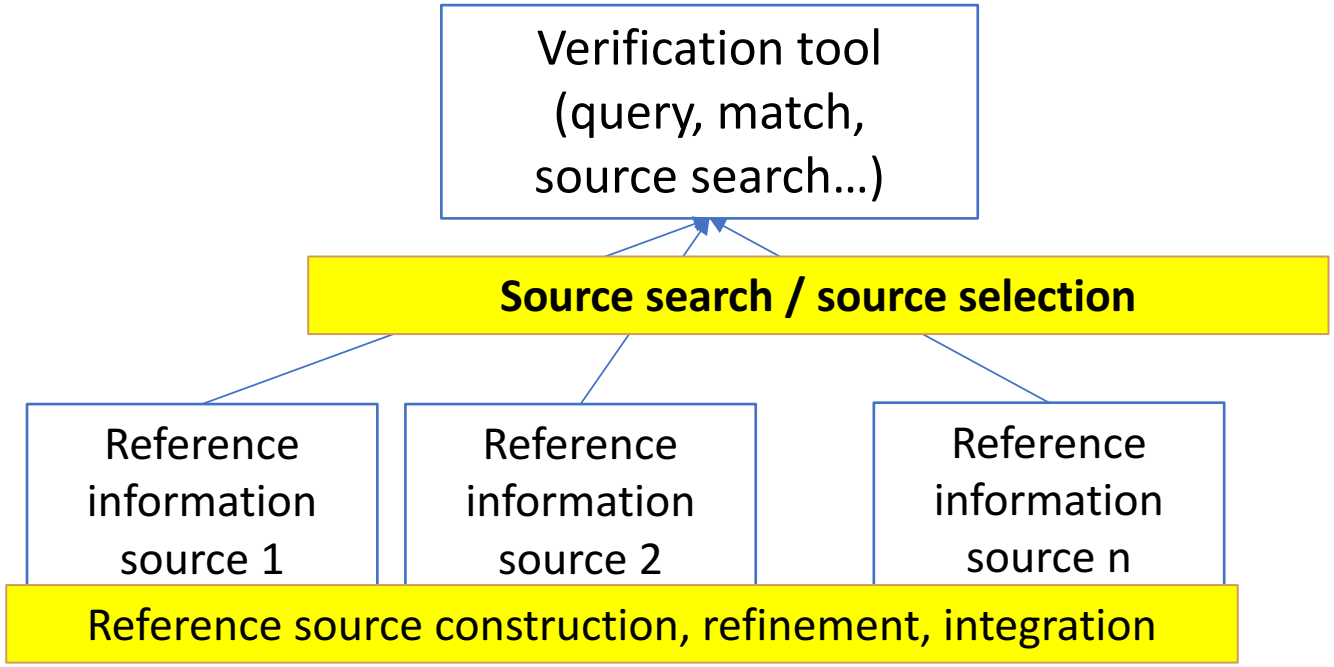
Postes salariés par secteur d'activité au 31 décembre 2013

	Nombre d'établissements actifs		Postes salariés	
	Pays	Poids Pays / Bretagne (en %)	Pays	Poids Pays / Bretagne (en %)
Agriculture	690	2,1	563	2,6
Industrie	589	3,6	3 348	2,0
Construction	1 051	3,8	1 812	2,5
Commerce	1 558	3,7	3 385	2,4
Services	6 415	3,7	12 076	1,9
Total	**10 303**	**3,6**	**21 184**	**2,0**

Champ : ensemble des établissements hors défense et particuliers employeurs

Source : Insee, Clap

5. Future work



Thanks / questions?

Excel files and extracted RDF files
(10.5GB will be expired in May 29th 2017)
https://goo.gl/4Y5Dtv

Source code: no expiration date :)
https://gitlab.inria.fr/cedar/insee-crawler
https://gitlab.inria.fr/cedar/excel-extractor