Searching for truth in a database of statistics

PhD student: Tien-Duc Cao

Supervisors: Ioana Manolescu and Xavier Tannier

Agenda

1. Problem statement
2. Overview
3. Data collection
4. Dataset search
5. Data cell search
6. Evaluation

1. Problem statement

- Given a **statistical claim** from the media
  - Which data sources are the most **relevant** to fact-check that claim?
    - E.g: the statistical claim *“unemployment rate of Île-de-France in 2016 was 20%”* can be fact-checked by looking at insee.fr’s Excel files of unemployment rate
  - What is the trusted known **figure** closest to the claim?

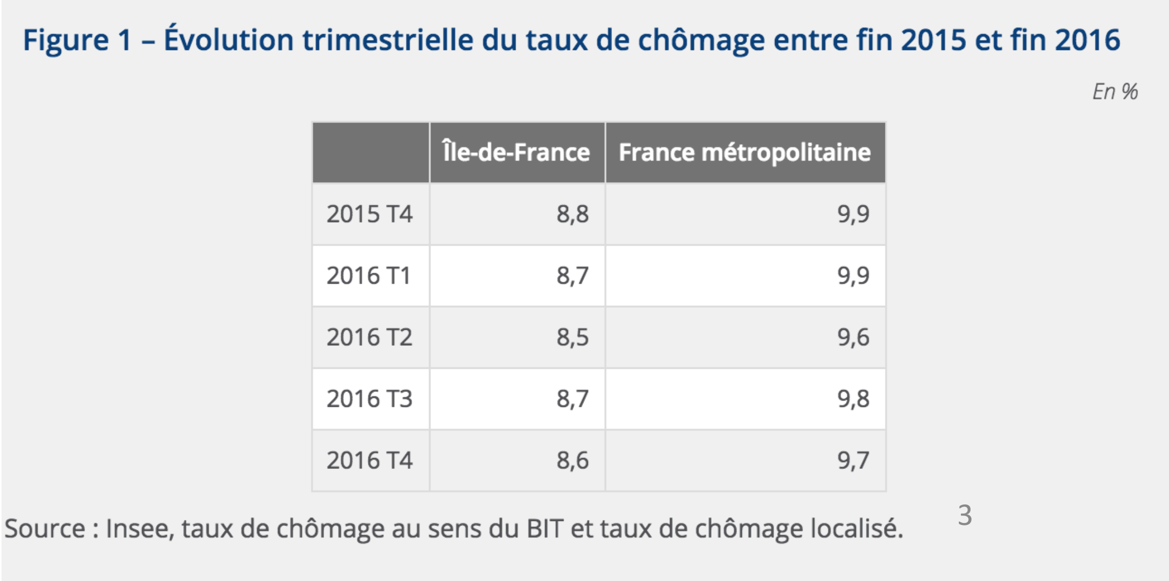


Figure 1 – Évolution trimestrielle du taux de chômage entre fin 2015 et fin 2016

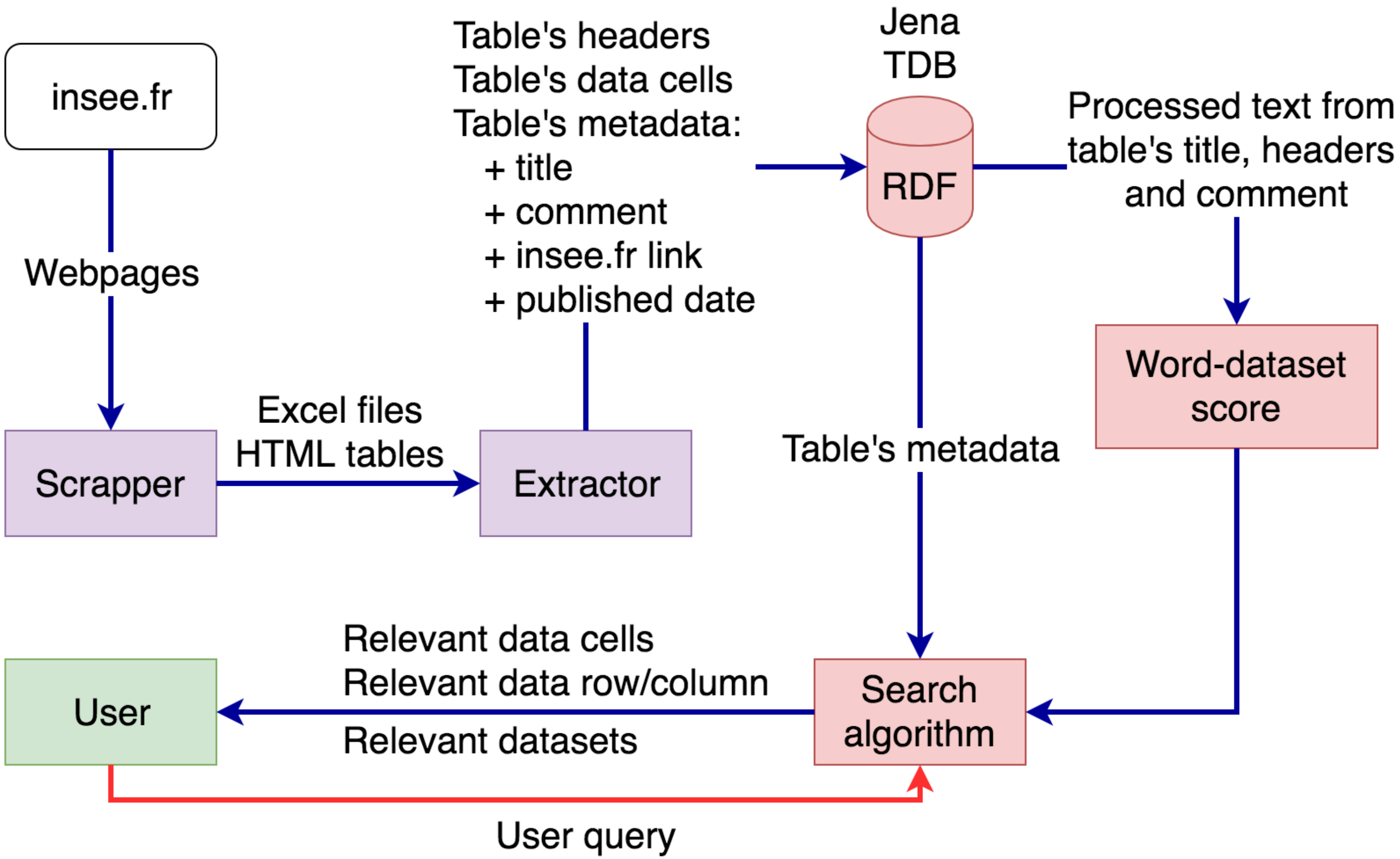
En %

	Île-de-France	France métropolitaine
2015 T4	8,8	9,9
2016 T1	8,7	9,9
2016 T2	8,5	9,6
2016 T3	8,7	9,8
2016 T4	8,6	9,7

Source : Insee, taux de chômage au sens du BIT et taux de chômage localisé.

https://www.insee.fr/fr/statistiques/2853194#tableau-Figure_2

2. Overview


insee.fr

Webpages

Scrapper

Excel files
HTML tables

Extractor

Table's headers
Table's data cells
Table's metadata:
+ title
+ comment
+ insee.fr link
+ published date

Jena
TDB

RDF

Processed text from
table's title, headers
and comment

Word-dataset
score

Table's metadata

Search
algorithm

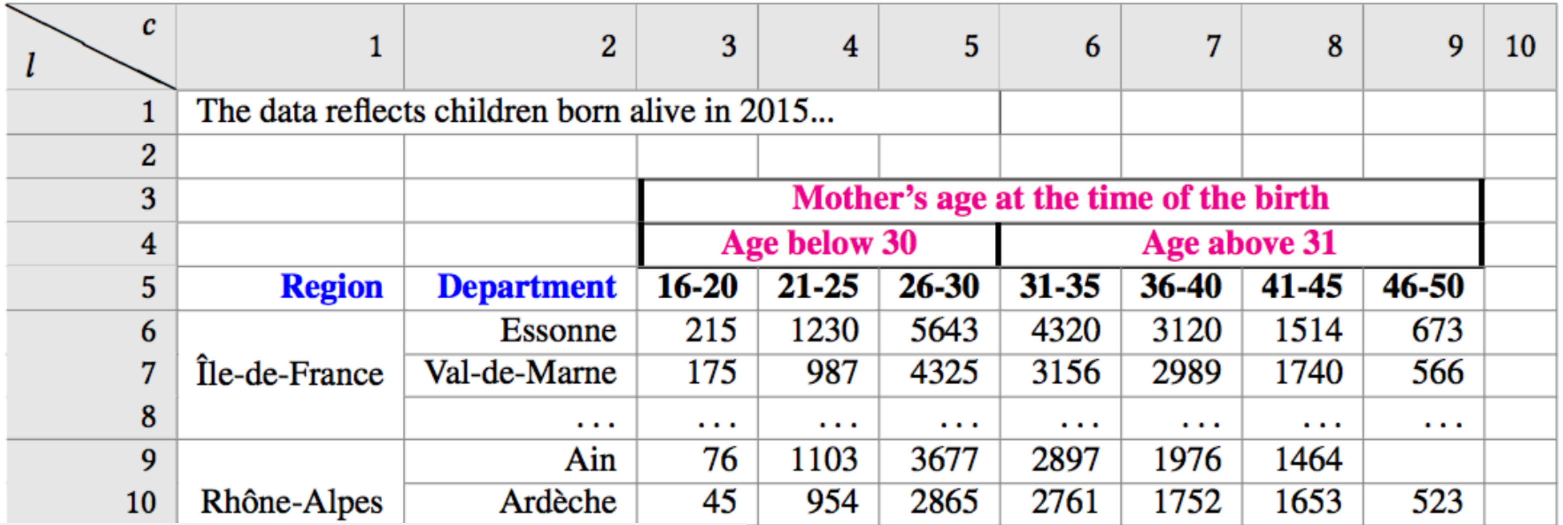
Relevant data cells
Relevant data row/column

Relevant datasets

User

User query

3. Data collection: INSEE tables



c l	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	
				...	...	...	...	...	...	
				...	...	...	...	...	...	

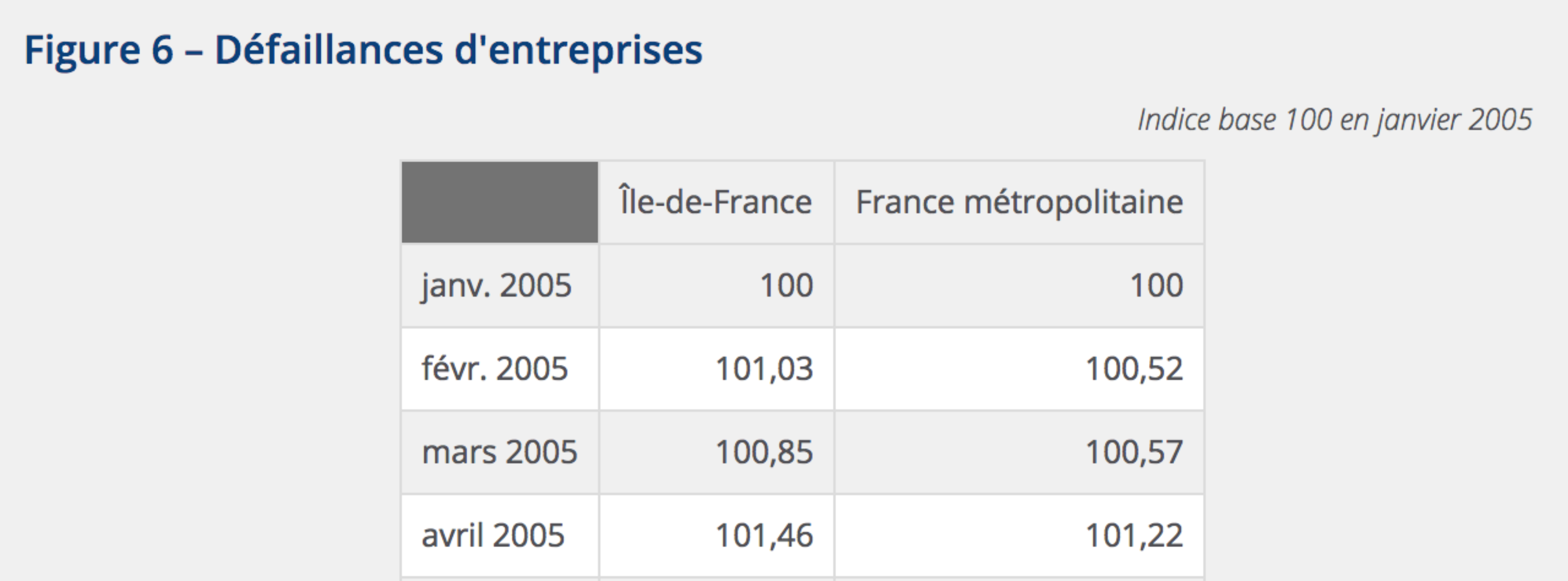


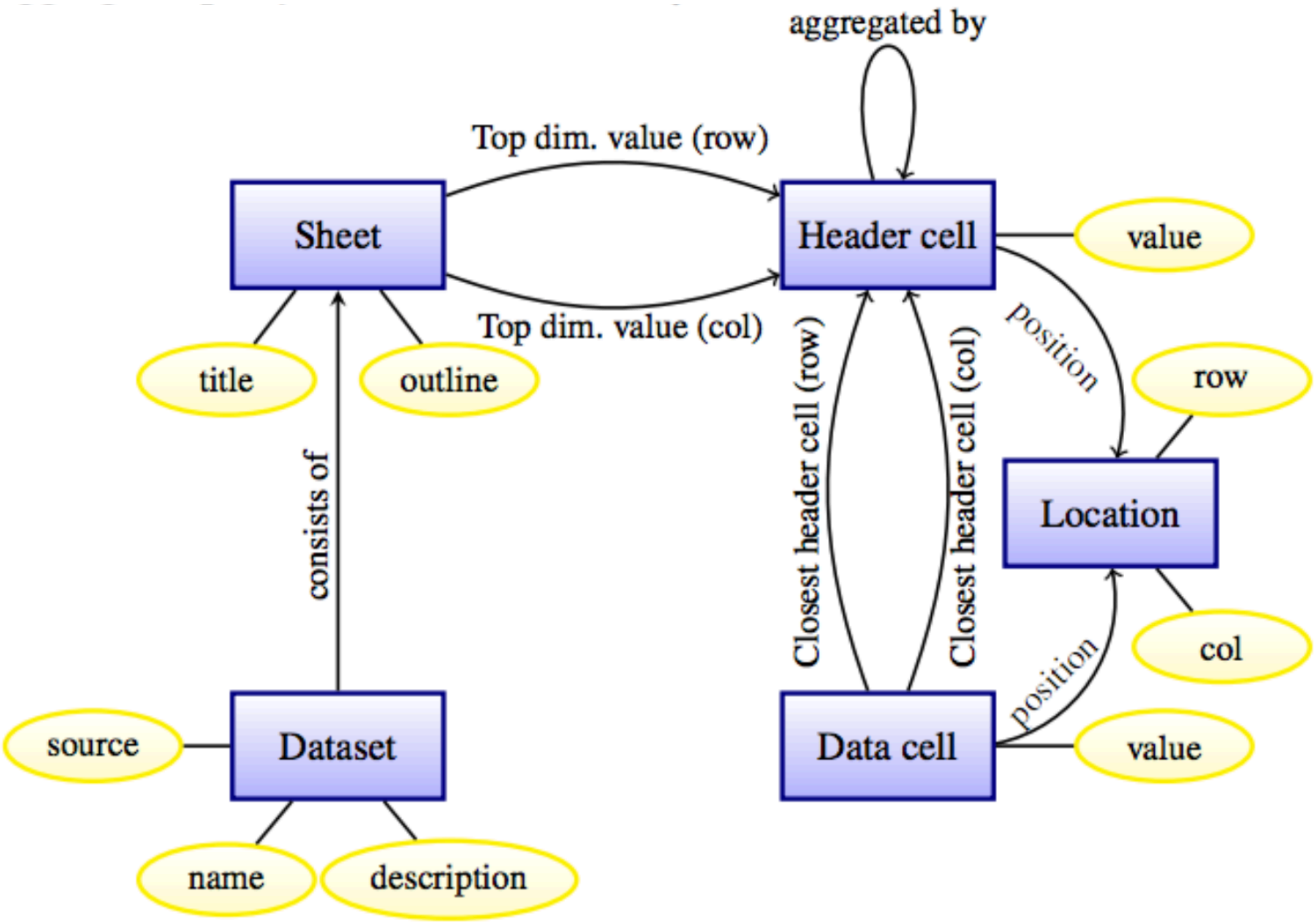
Figure 6 – Défaillances d'entreprises

Indice base 100 en janvier 2005

	Île-de-France	France métropolitaine
janv. 2005	100	100
févr. 2005	101,03	100,52
mars 2005	100,85	100,57
avril 2005	101,46	101,22

3. Data collection: extracted RDF

- Tien Duc Cao, Ioana Manolescu, Xavier Tannier ***Extracting Linked Data from statistic spreadsheets***. Semantic Big Data workshop (with SIGMOD 2017)
- Up to date, processed:
  - 22962 HTML tables
  - 91287 Excel tables



Conceptual data model

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

4. Dataset search

- A dataset is relevant if it contains the keyword or a similar word of that keyword
- Dataset's relevance score G1(t):
  - Query = W = {w1, …, wN}
    - Each w_i is a keyword
  - **G1(t) = score(w1, t) + … + score(wN, t)** where
    - t is a dataset
    - and score(w, t) is the word-dataset score of w in t

4. Dataset search: word-dataset score

- Text processing:
  - Collect texts from each table: title, comment, header rows and header columns
  - Tokenize text to words using KEA[1] tokenizer
  - Identify bigrams (e.g: “aide sociale”)
  - Identify top-50 similar words using French word2vec[2] model
- **score(w, D)**: word-dataset score for a keyword w in dataset D
  - 1.0 if w appears in d
  - s if w doesn’t appear in d but w’ is similar to d and the similarity score of w and w’ is s
  - computed from Geonames Hierarchy API[3]
    - E.g: 1.0 for *Paris*, 0.66 for *Île-de-France*, 0.33 for *France*

[1] https://github.com/boudinfl/kea

[2] Model was trained by Xavier from a big corpus of French articles

[3] http://www.geonames.org/export/place-hierarchy.html

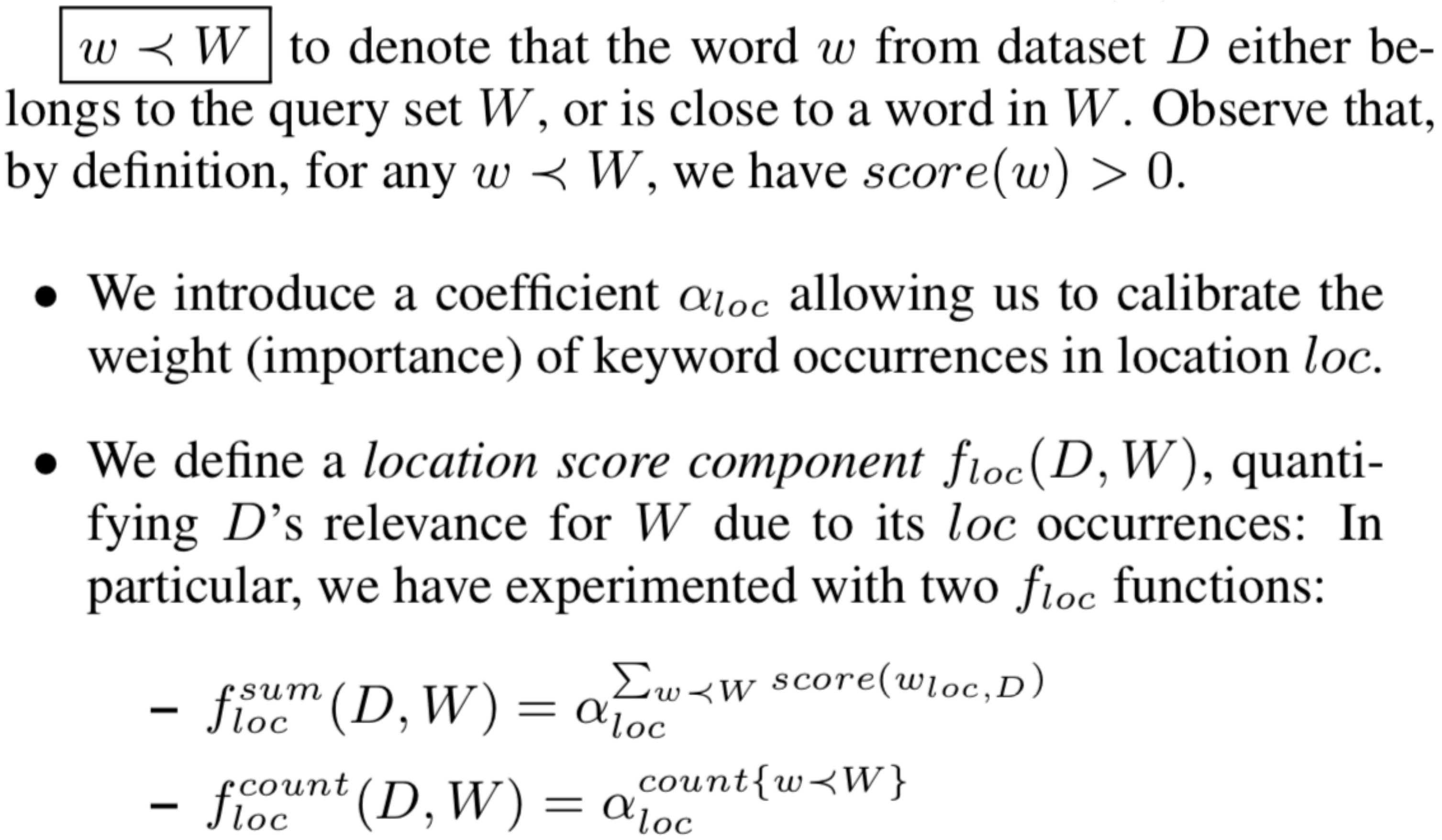
4. Dataset search

- Computing all relevance scores of 110k datasets and then selecting the top-k datasets (e.g: k = 20) is too slow
- Solution: adapted early-stop Fagin’s Threshold algorithm to our setting
    - n keywords
    - n sorted lists (descending order) of word-dataset scores (one for each keyword)
    - G = dataset’s relevance score as a **monotonous increasing** function
    - k = number of datasets with the highest value of G

4. Dataset search

- Positions of keywords are also important:
  - Match in title > match in comment
- Re-rank preselected datasets using score function **G2**:
  - To return k final answers, we identify the best 3k datasets using Fagin
  - Apply Fagin with G2 and 3k datasets, we obtain top-k datasets

4. Dataset search



$\boxed{w \prec W}$ to denote that the word w from dataset D either belongs to the query set W, or is close to a word in W. Observe that, by definition, for any $w \prec W$, we have $score(w) > 0$.

- We introduce a coefficient α_{loc} allowing us to calibrate the weight (importance) of keyword occurrences in location loc.
- We define a *location score component* $f_{loc}(D, W)$, quantifying D's relevance for W due to its loc occurrences: In particular, we have experimented with two f_{loc} functions:
  - $f_{loc}^{sum}(D, W) = \alpha_{loc}^{\sum_{w \prec W} score(w_{loc,D})}$
  - $f_{loc}^{count}(D, W) = \alpha_{loc}^{count\{w \prec W\}}$

4. Dataset search

$$g_2(D,W) = g_1(D,W)+\Sigma_{loc\in\{\mathtt{T},\mathtt{HR},\mathtt{HC},\mathtt{C}\}} f_{loc}(D,W)+f_{\mathtt{H}}(D,W)$$

- T: title
- HR: header row
- HC: header column
- C: comment
- H: header row or column

$$g_2^*(D,W) = \begin{cases} g_2(D,W), \text{if } f_{\mathtt{T}}(D,W) > 0 \\ 0, \text{otherwise} \end{cases}$$

5. Data cell search

- Answer can be a **data cell** or **set of data cells**
- A relevant data cell satisfies (by row and column headers) ideally all query keywords
    - Ideal: some keywords in row header and the others in column header

Créations d’entreprises dans la région Île-de-France

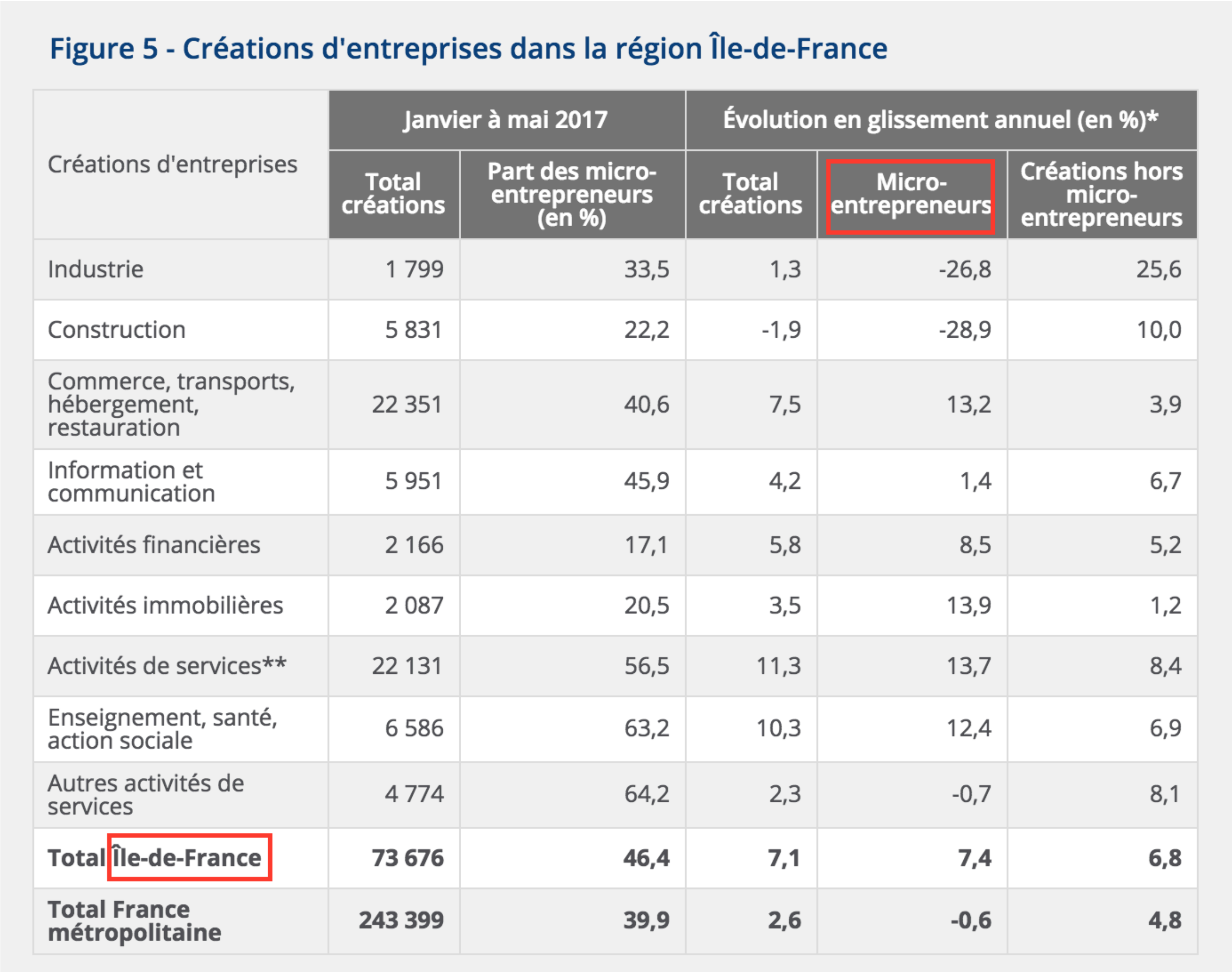


Figure 5 - Créations d'entreprises dans la région Île-de-France

Créations d'entreprises	Janvier à mai 2017		Évolution en glissement annuel (en %)*		
	Total créations	Part des micro-entrepreneurs (en %)	Total créations	Micro-entrepreneurs	Créations hors micro-entrepreneurs
Industrie	1 799	33,5	1,3	-26,8	25,6
Construction	5 831	22,2	-1,9	-28,9	10,0
Commerce, transports, hébergement, restauration	22 351	40,6	7,5	13,2	3,9
Information et communication	5 951	45,9	4,2	1,4	6,7
Activités financières	2 166	17,1	5,8	8,5	5,2
Activités immobilières	2 087	20,5	3,5	13,9	1,2
Activités de services**	22 131	56,5	11,3	13,7	8,4
Enseignement, santé, action sociale	6 586	63,2	10,3	12,4	6,9
Autres activités de services	4 774	64,2	2,3	-0,7	8,1
Total Île-de-France	**73 676**	**46,4**	**7,1**	**7,4**	**6,8**
Total France métropolitaine	**243 399**	**39,9**	**2,6**	**-0,6**	**4,8**

5. Data cell search

- Answer can be a **data cell** or **set of data cells**
- A relevant data cell satisfies (by row and column headers) ideally all query keywords
  - Ideal: some keywords in row header and the others in column header
  - Less than ideal: keywords only in row header --> return the whole row

Total créations d’entreprises

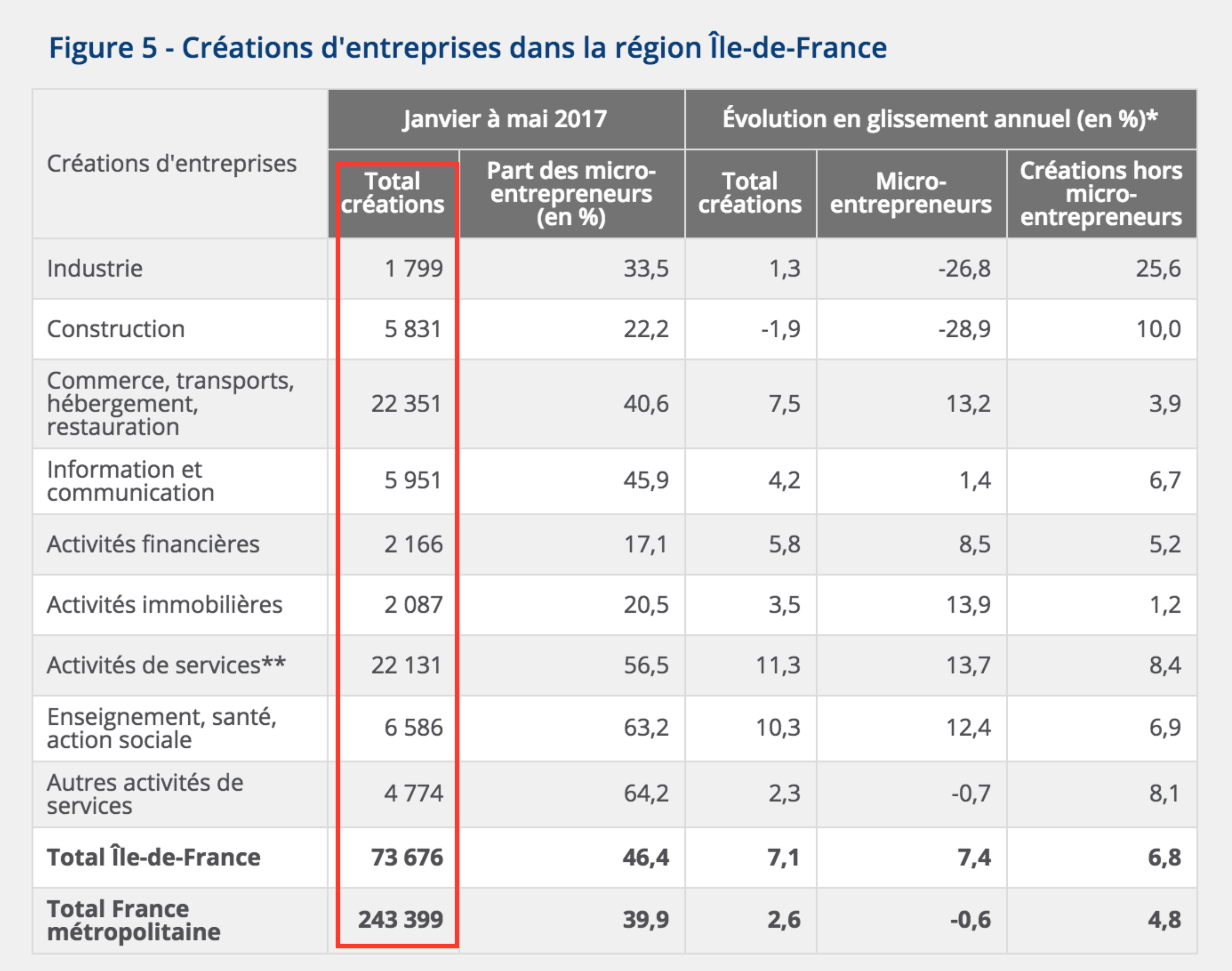


Figure 5 - Créations d'entreprises dans la région Île-de-France

Créations d'entreprises	Janvier à mai 2017		Évolution en glissement annuel (en %)*		
	Total créations	Part des micro-entrepreneurs (en %)	Total créations	Micro-entrepreneurs	Créations hors micro-entrepreneurs
Industrie	1 799	33,5	1,3	-26,8	25,6
Construction	5 831	22,2	-1,9	-28,9	10,0
Commerce, transports, hébergement, restauration	22 351	40,6	7,5	13,2	3,9
Information et communication	5 951	45,9	4,2	1,4	6,7
Activités financières	2 166	17,1	5,8	8,5	5,2
Activités immobilières	2 087	20,5	3,5	13,9	1,2
Activités de services**	22 131	56,5	11,3	13,7	8,4
Enseignement, santé, action sociale	6 586	63,2	10,3	12,4	6,9
Autres activités de services	4 774	64,2	2,3	-0,7	8,1
Total Île-de-France	**73 676**	**46,4**	**7,1**	**7,4**	**6,8**
Total France métropolitaine	**243 399**	**39,9**	**2,6**	**-0,6**	**4,8**

5. Data cell search

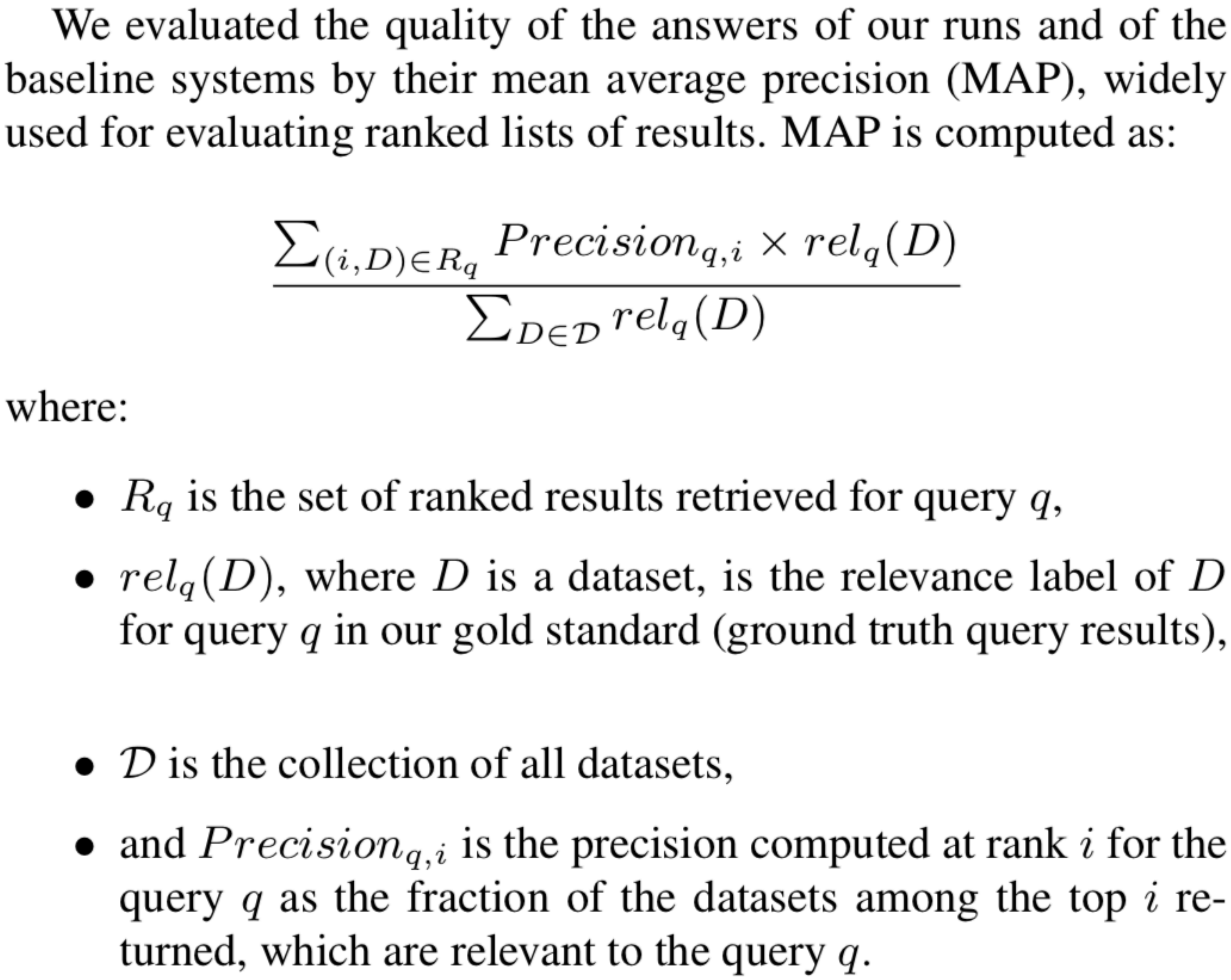
- Inputs: header rows and header columns in which at least 1 keyword appears
- Outputs: pair(s) of header row and header column that contain the maximum number of keywords.
  - If a parent header cell PHC of a cell HC contains some keywords, then we consider these keywords belong to HC, too
- We use a SPARQL query to ask for the data cells that belong to these pairs of header row and column

6. Evaluation

- We collected all the fact-checking articles published online by Les Décodeurs[1] between March 10th and August 2nd, 2014
- 75 articles that contain links to insee.fr → **55** queries
  - 29 queries for development set
  - 26 queries for test set
- We compute MAP@20 on test set for:
  - Our system
  - Google search
  - insee.fr's search

[1] http://www.lemonde.fr/les-decodeurs/

6. Evaluation



We evaluated the quality of the answers of our runs and of the baseline systems by their mean average precision (MAP), widely used for evaluating ranked lists of results. MAP is computed as:

$$\frac{\sum_{(i,D)\in R_q} Precision_{q,i} \times rel_q(D)}{\sum_{D\in\mathcal{D}} rel_q(D)}$$

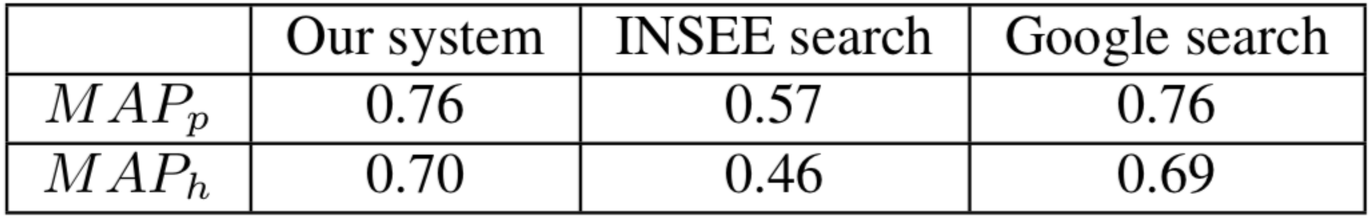
where:

- R_q is the set of ranked results retrieved for query q,
- $rel_q(D)$, where D is a dataset, is the relevance label of D for query q in our gold standard (ground truth query results),
- $\mathcal{D}$ is the collection of all datasets,
- and $Precision_{q,i}$ is the precision computed at rank i for the query q as the fraction of the datasets among the top i returned, which are relevant to the query q.

MAP$_h$ is the mean average precision where only highly relevant datasets are considered as relevant in rel$_q$(E) and Precision$_{q,i}$.

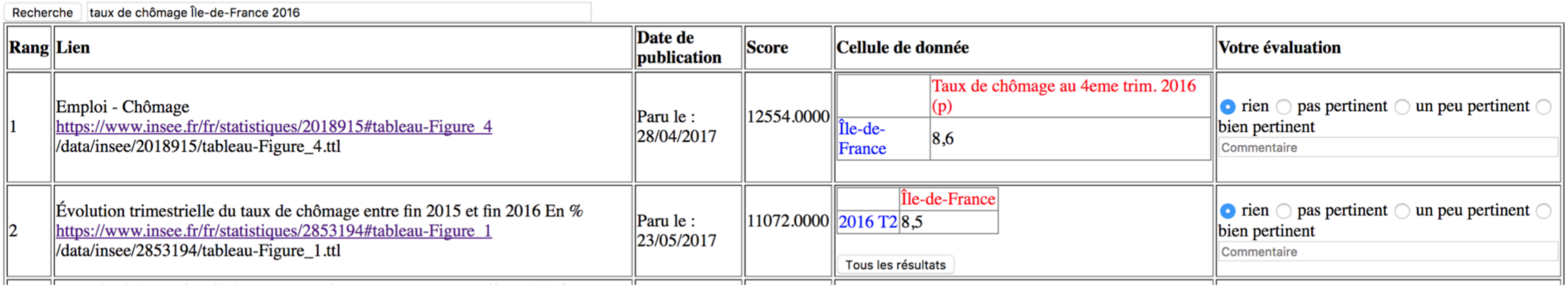
- **MAP$_p$** is the mean average precision where both partially and highly relevant datasets are considered relevant.

6. Evaluation



	Our system	INSEE search	Google search
MAP_p	0.76	0.57	0.76
MAP_h	0.70	0.46	0.69

Screenshot



taux de chômage Île-de-France 2016

<table>
<tr><th>Rang</th><th>Lien</th><th>Date de publication</th><th>Score</th><th>Cellule de donnée</th><th>Votre évaluation</th></tr>
<tr><td>1</td><td>Emploi - Chômage
https://www.insee.fr/fr/statistiques/2018915#tableau-Figure_4
/data/insee/2018915/tableau-Figure_4.ttl</td><td>Paru le :
28/04/2017</td><td>12554.0000</td><td><table><tr><td></td><td>Taux de chômage au 4eme trim. 2016 (p)</td></tr><tr><td>Île-de-France</td><td>8,6</td></tr></table></td><td>● rien ○ pas pertinent ○ un peu pertinent ○ bien pertinent
Commentaire</td></tr>
<tr><td>2</td><td>Évolution trimestrielle du taux de chômage entre fin 2015 et fin 2016 En %
https://www.insee.fr/fr/statistiques/2853194#tableau-Figure_1
/data/insee/2853194/tableau-Figure_1.ttl</td><td>Paru le :
23/05/2017</td><td>11072.0000</td><td><table><tr><td></td><td>Île-de-France</td></tr><tr><td>2016 T2</td><td>8,5</td></tr></table>Tous les résultats</td><td>● rien ○ pas pertinent ○ un peu pertinent ○ bien pertinent
Commentaire</td></tr>
</table>

Thank you very much

Merci beaucoup

どうもありがとうございました