# From raw survey data to a structured register : evidence from the Census



Yves-Laurent Bénichou
Frédéric Comte
Julie Djiriguian
**Benjamin Sakarovitch**
Eric Sigaud

# *census is a boring party*

*identifying (in the register) the declared employer (in the census)*

1. the problem and the current state of affairs
2. the proposed solution
3. in the context of innovation in the NSI

# employer in the census

the *who do you work for* question

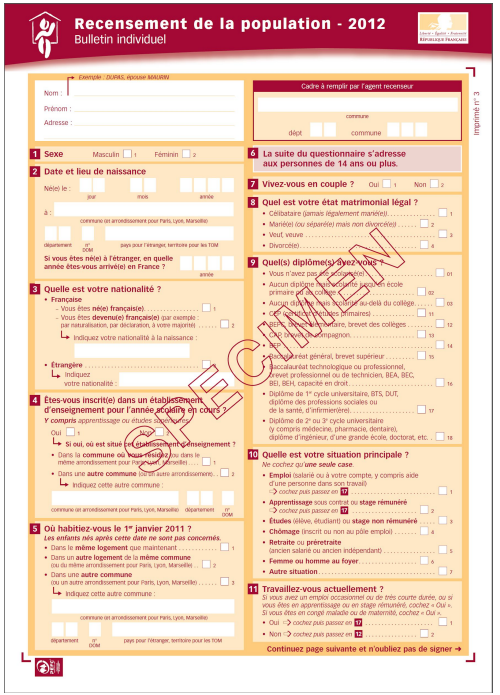
- what is it ?
- why asking it ?
- how is it processed ?

# declaring the employer in the census

- what is the **name of the company** that employs you ?
- what is the **address** of the place you mostly go to work ?
- what is the **activity** of the company that employs you ?

\+ what is your job ?

***⇒ then INSEE tries to match it to the exact company in its register***



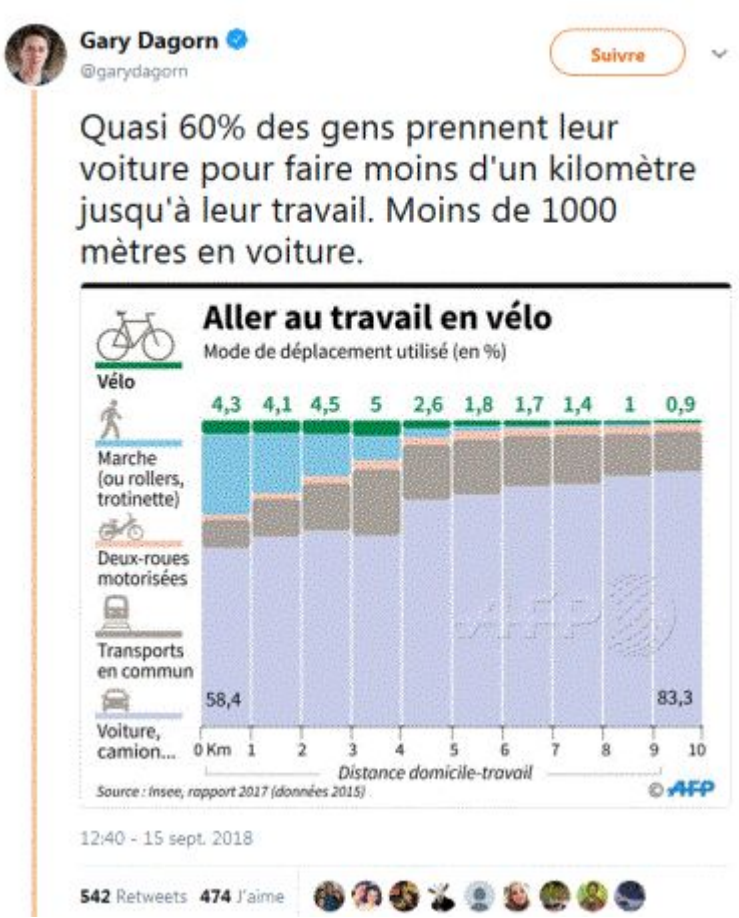
# what indicators does it fuel ?

**1/ Coding the socio-professional status**

*exact company in the register ⇒ activity code at the finest level ⇒ useful to code the socio-pro status*

**2/ Home-to-work commuting distance**

*currently it is computed as the distance between the 2 town halls…*



# the current (costly) procedure

2 **coding machines** (from the 80s) with decision rules

- from company name and address : identification in the register
- in case of doubt : coding the activity to decide
- if unsuccessful : goes to employees doing it manually

44% of declared employers are coded automatically (from a total of 1.7M)

70 persons do the 56% left for 5 months

*even by hand they find the exact company in only 70% of the cases (rest is activity-coded or blanked)*

# why is automatisation a challenge ?

***imprecisions in the declaration :***

- incoherent address (street or municipality)
- declared activity does not correspond to the one registered
- vague description of the employer : eg ministry of education without further indication (middle school…) and/or address

***registration of companies :***

name not corresponding to the usual denomination

- management contracts
- acronyms
- persons names

DECATHLON

Carrefour

# how to assess performance ?

*No ground truth*

Special process to assess accuracy of handmade identification :

- every five years or so - last from 2014
- on 32000 census sheets (30000 manual / 2000 auto)
- double additional manual check + third for decision in case of disagreement

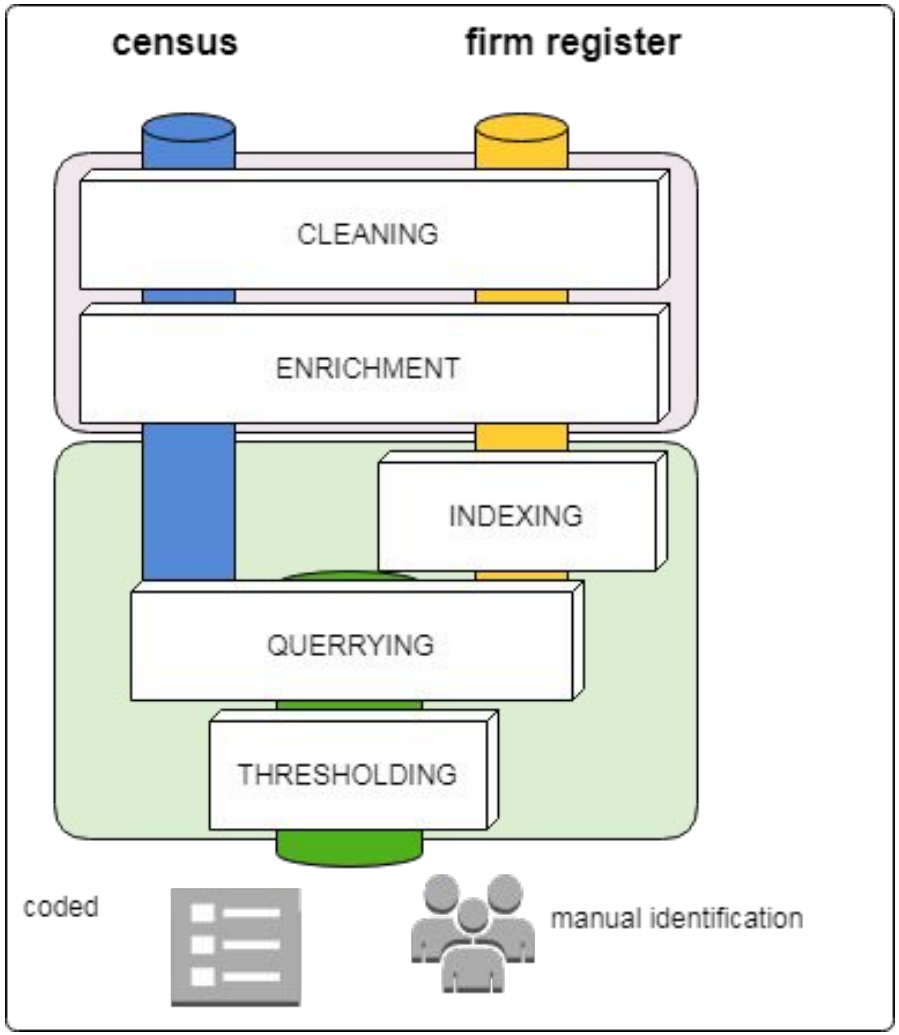
⇒ on the 2000 automatically coded : final coding coherent in 86% of cases
⇒ on the 30000 manually coded : final coding coherent in 66% of cases

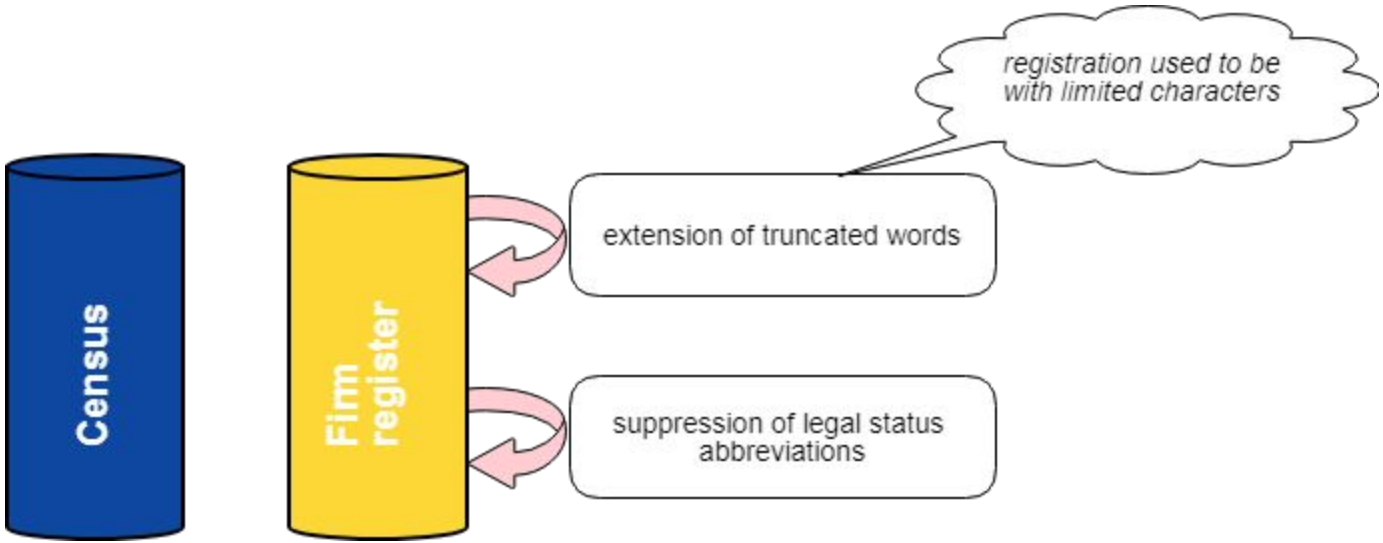
*only metrics are : **coverage** (number of employers automatically coded) and **consistency** (with current process of coding)*

# Developing a prototype

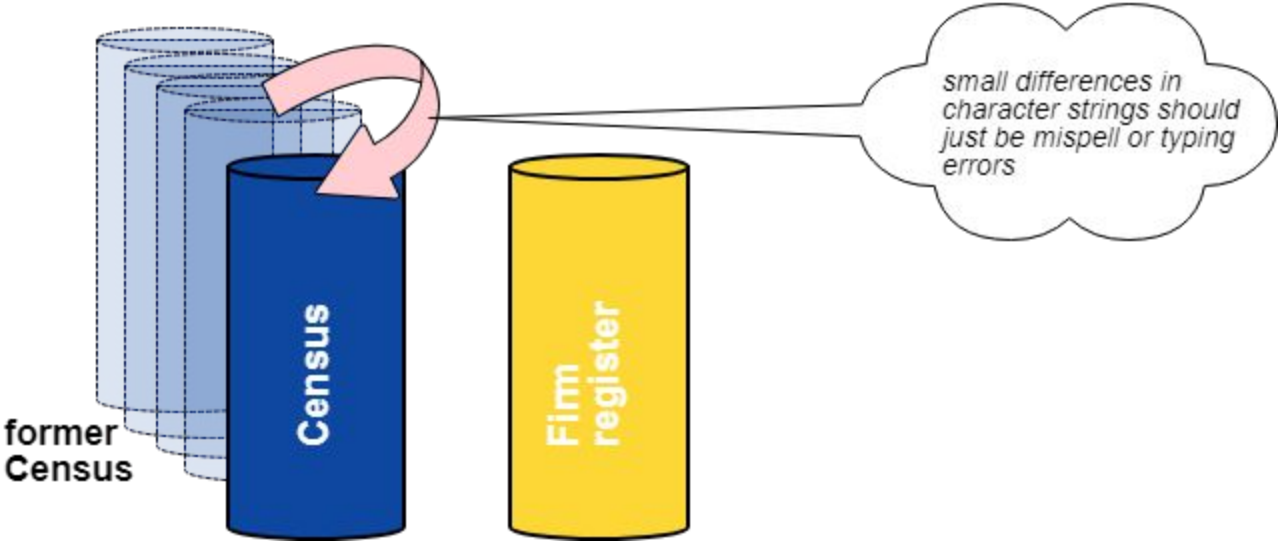
a search engine to solve the scalability challenge

elasticsearch

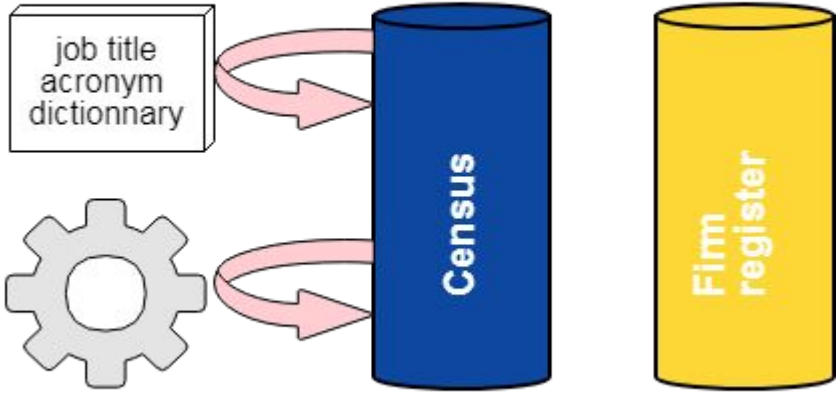

census
firm register
CLEANING
ENRICHMENT
INDEXING
QUERRYING
THRESHOLDING
coded
manual identification


Census
Firm register
extension of truncated words
suppression of legal status abbreviations
registration used to be with limited characters

## clusterizing similar observation (per field) : uniformisation to most common declaration

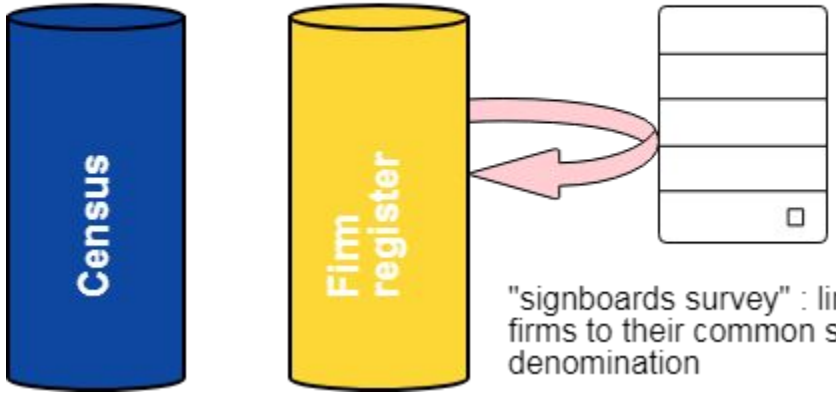


## enriching the declared activity of the employer



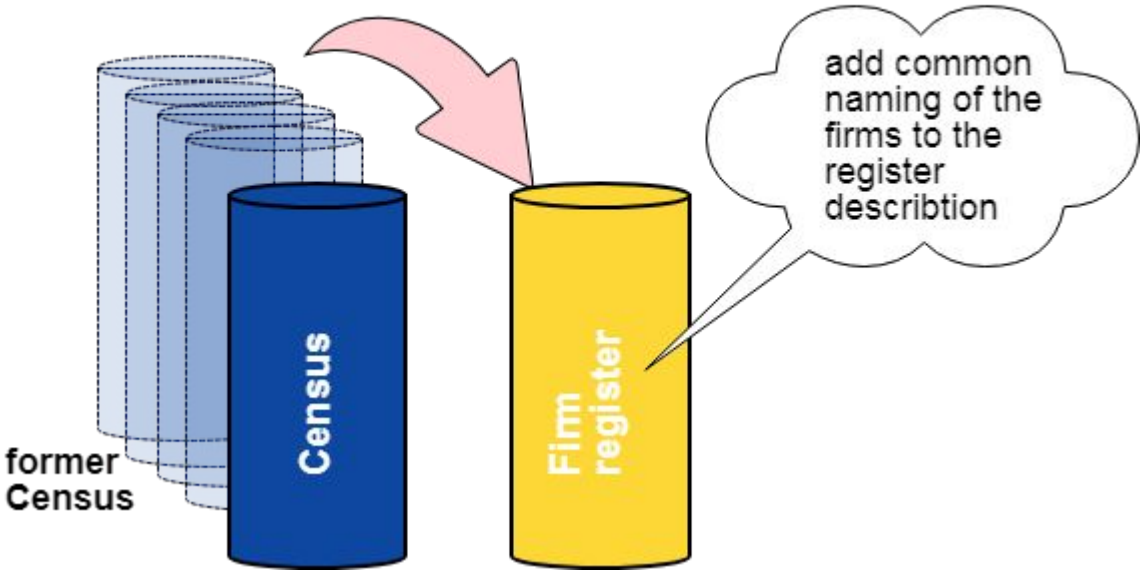

coding declared activity to "NACE"

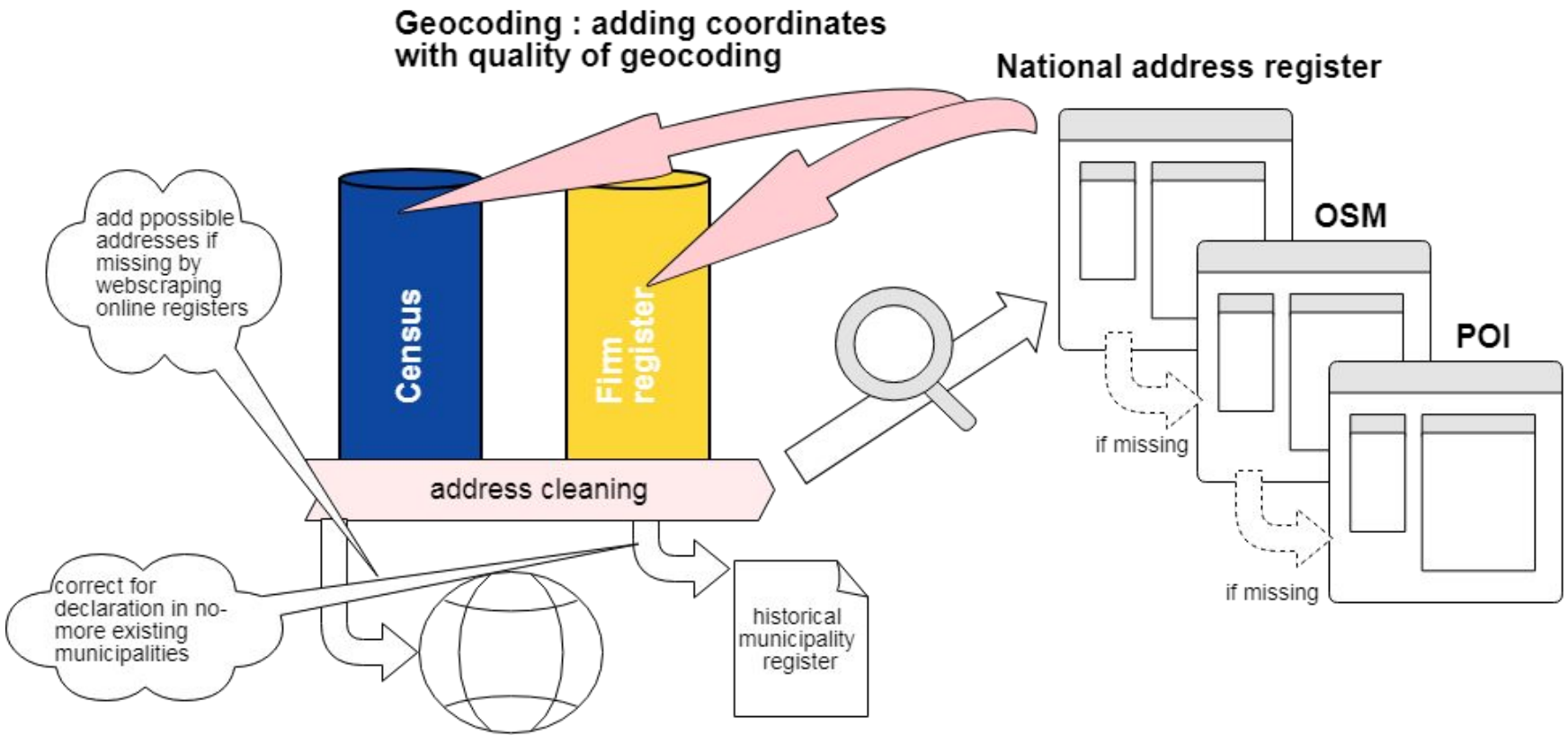
## enriching firms describtion with survey from the NSI

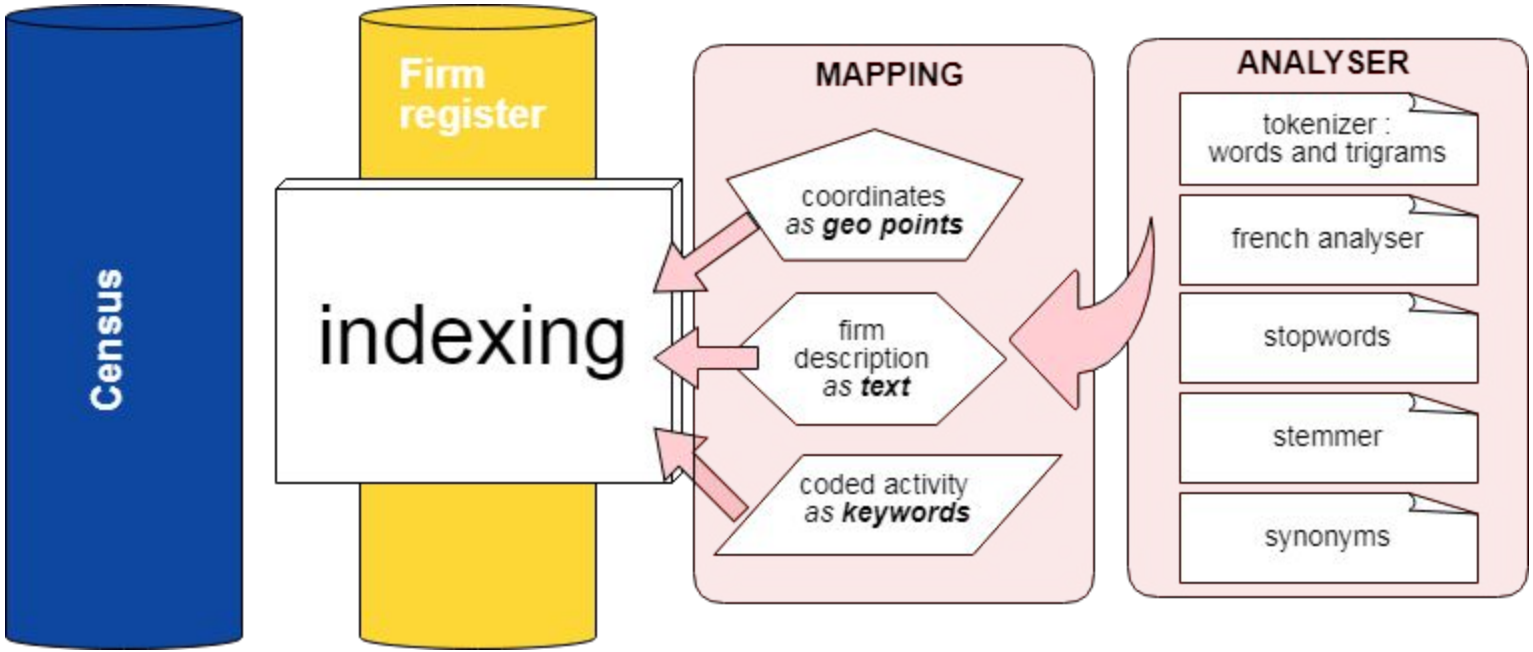


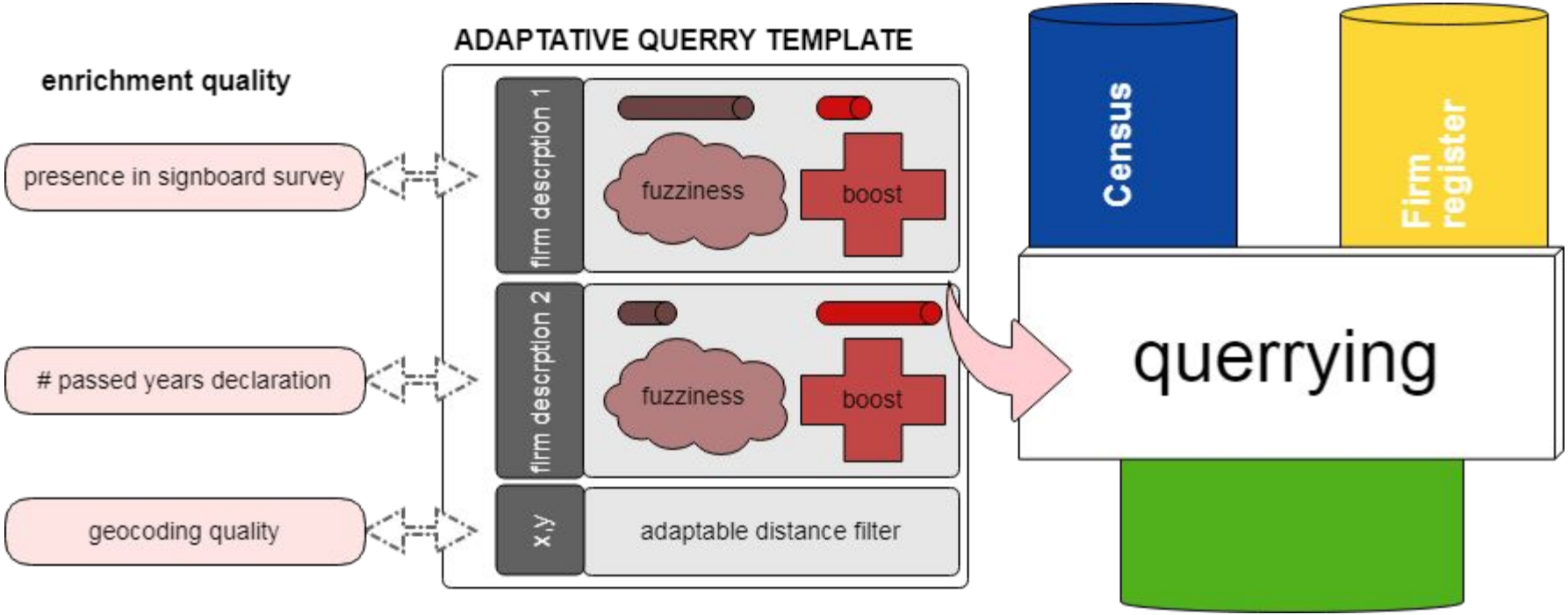

"signboards survey" : linking firms to their common signboard denomination

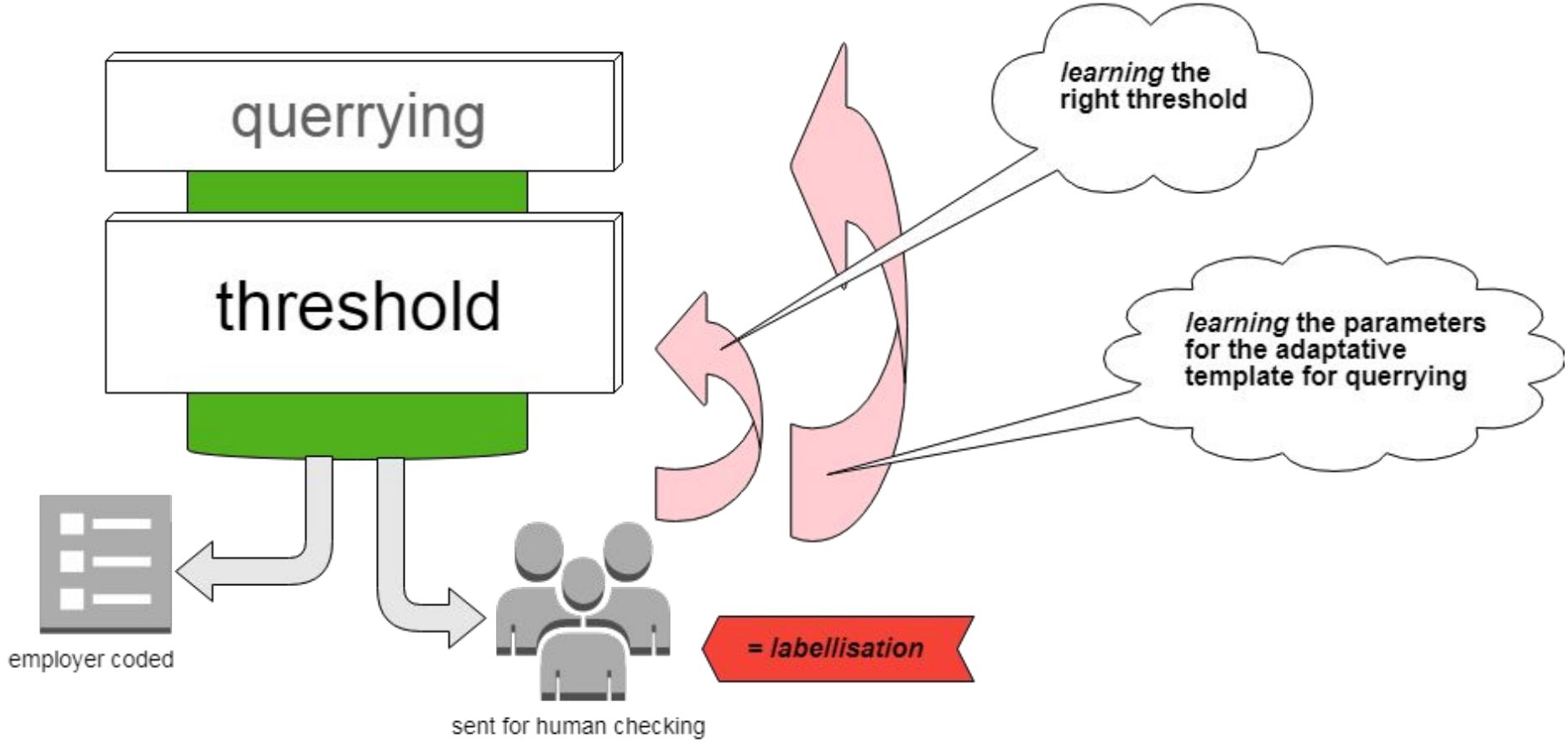
## crowd sourcing from the past : enriching the register from passed years identifications




Geocoding : adding coordinates
with quality of geocoding
National address register
add ppossible addresses if missing by webscraping online registers
Census
Firm register
address cleaning
OSM
POI
if missing
if missing
correct for declaration in no-more existing municipalities
historical municipality register


Census
Firm register
indexing
MAPPING
coordinates as geo points
firm description as text
coded activity as keywords
ANALYSER
tokenizer : words and trigrams
french analyser
stopwords
stemmer
synonyms


enrichment quality
presence in signboard survey
# passed years declaration
geocoding quality
ADAPTATIVE QUERRY TEMPLATE
firm descrption 1
fuzziness
boost
firm descrption 2
fuzziness
boost
x,y
adaptable distance filter
Census
Firm register
querrying


querrying
threshold
learning the right threshold
learning the parameters for the adaptative template for querrying
employer coded
sent for human checking
= labellisation

# conclusions

final word is yet to come

- actual solution leaves room for improvement
- enriching the data with many additional sources
- indexing for scalability
- identification in the firm register has many other applications
- identification in a register is a common issue

# a general process of modernisation

the brand new “SSP lab” to foster innovation for official statistics in France

- hackathon last winter
- promoting data science in the NSI and around
- on the long run create frameworks that allow to capitalize on the work produced by gestionnaires