Cognisearch Business: a business information search service on the Web

Armel Fotsoh Tawofaing

Laboratoire LIUPPA, BP-1155, 64013 PAU Université Cedex, France
aftawofaing@univ-pau.fr
armel@cogniteev.com

January 19, 2016

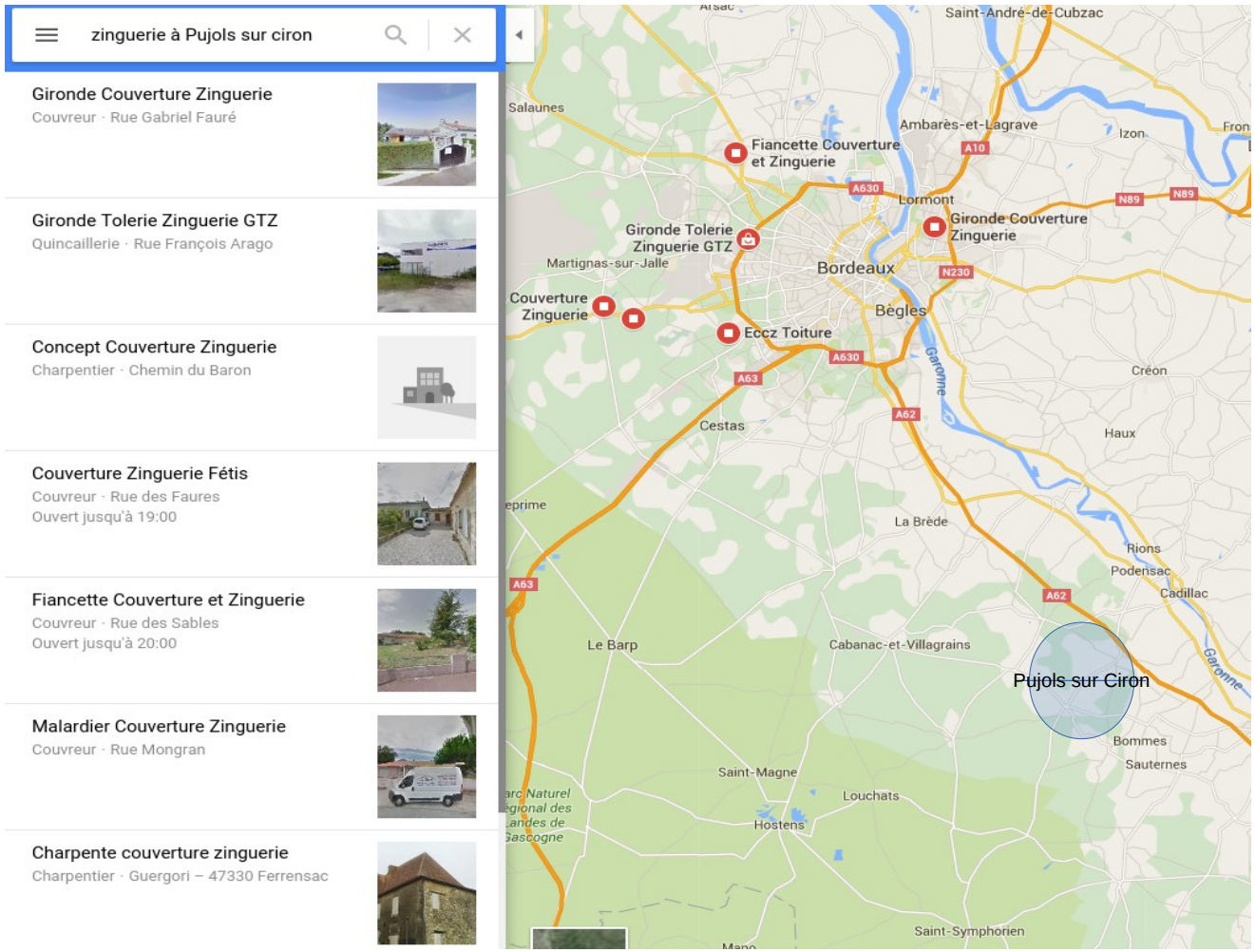
Plan

- Introduction
- Related work
- Proposal
  - Business Entity Model
  - General architecture of the service
  - Address Extraction
  - Resources Enrichment
- Prototype
  - Process Flow Implementation
  - Experiment query
- Conclusion

Introduction

Motivations

zinc-work in «Pujols sur Ciron»



Motivations

zinc-work in «Pujols sur Ciron»

zinguerie

Pujols sur Ciron (33210)

zinguerie à **Pujols sur Ciron** : **29** résultats

Trier par Pertinence

Localité | Horaires

1 **Ludovic Arnaud Charpente Couverture Tradit...** + d'infos

Ecrire un avis

Freylon, 33760 CANTOIS

17,32 km Ouvert

entreprises de couverture

Zinguerie

SITES INTERNET | PLAN | AFFICHER LE N°

2 **Moriceau (Sarl)** + d'infos

Ecrire un avis

3 r Jean Baptiste Perrin, 33320 EYSINES

42,69 km

zinguerie, entreprises de couverture

Prestations : Travaux de couverture, Travaux d'isolation, Travaux de zinguerie, Tr...

Zinguerie

E-MAIL | SITE INTERNET | PLAN | AFFICHER LE N°

Motivations

zinc-work in «Pujols sur Ciron»

Observations & Propositions

- **Observations:**
  - An increasing number of companies are present on Internet
  - These companies release their business and location information
    - Location information (Where): addresses
    - Thematic information (What): expertise fields, products, jobs
    - Contact information: phone numbers, fax, emails ...
  - Companies registration data is available in specific resources

- **Propositions:**
  - Gather companies' registration information from resources
  - Identify companies' websites
  - Extract company information from corporate websites
  - Merge extracted and registration data in a full business entity to supply a business local-based search service

Related work

State of the art

	Strong points	Limitations
Triou et al. (2007)	• Structuration of business information in an ontology organised by activity fields • Use of semantic to query constructed ontology	• Manually recorded data • companies' webistes are not analysed
Alhers (2013)	• Extraction of business information on the web • Extraction of German addresses with a good precision • Use "Pages Jaunes" data as input of the extraction process	• Analysed web pages come from DMOZ directory • Products and Jobs are not processed • Use of a Gazetteer containing all German street names for extraction for addresses

Proposal

Proposition

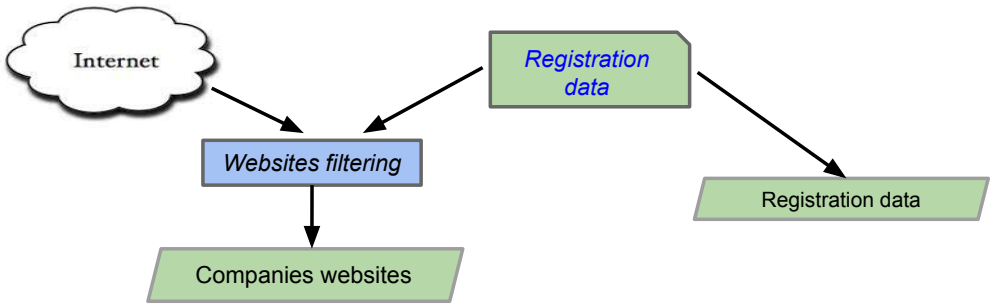
- We propose to build a business knowledge graph only from :
  - Business registration data
  - Companies websites
- Objectives :
  - Independence of manually recorded data
  - Exploitation of business data released and updated on websites.

Business model

Which information constitutes our entities?
Where do they come from?

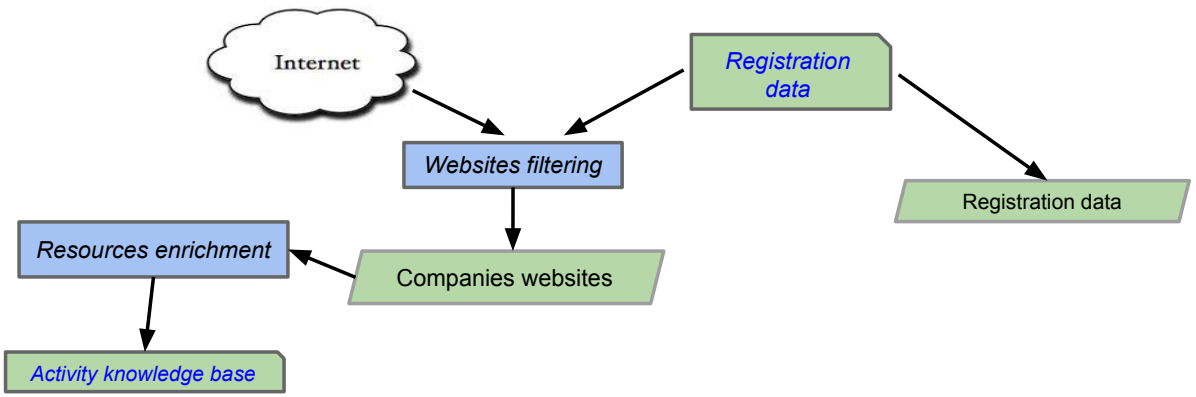
Business Entity Properties	**Representation Models**	**Data Sources**
Registration data	SIRENE (INSEE)	societe.com
Coordinates		
website	-	web
Address	Address (Etalab)	company website
Phone Numbers, emails, fax	-	company website
extended data		
Jobs	ROME (Pole Emploi)	company website
Activities	NAF (INSEE)	company website
Products	CPF (INSEE)	company website

Website Filtering

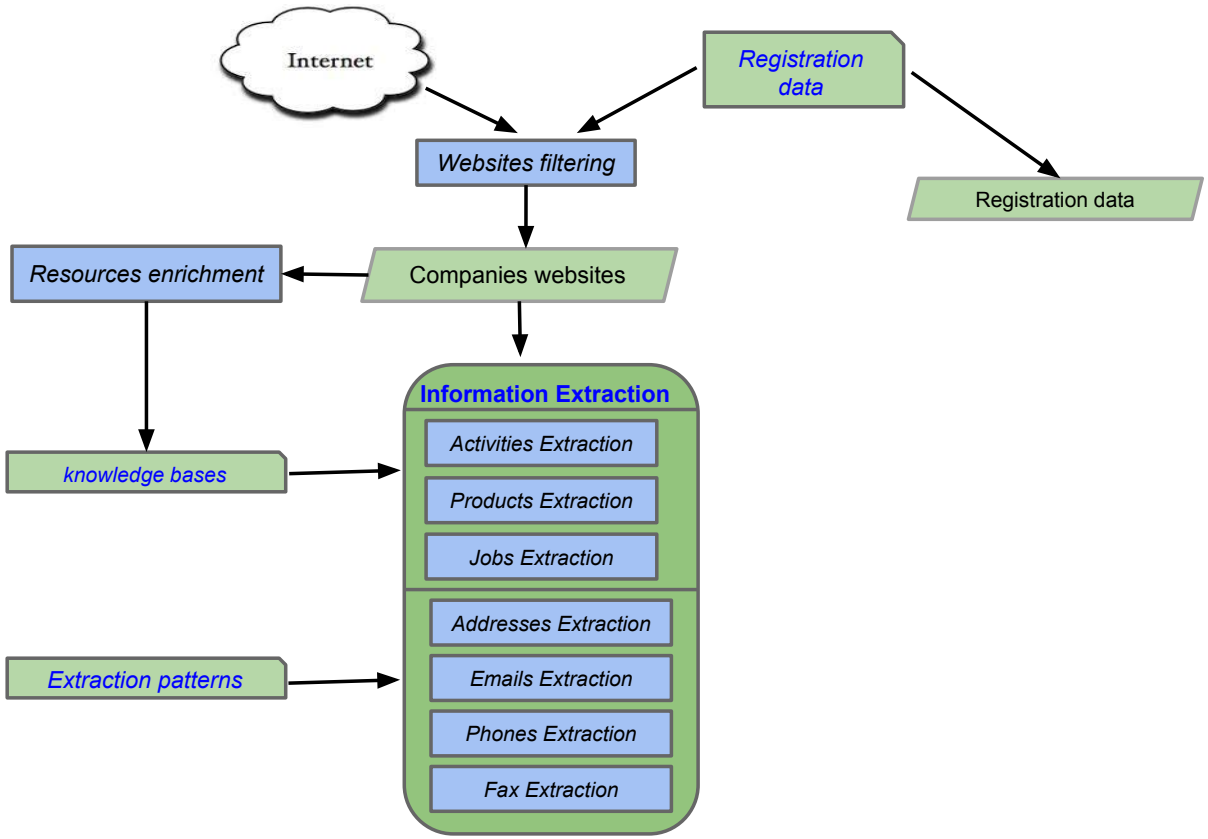


Resources Enrichment

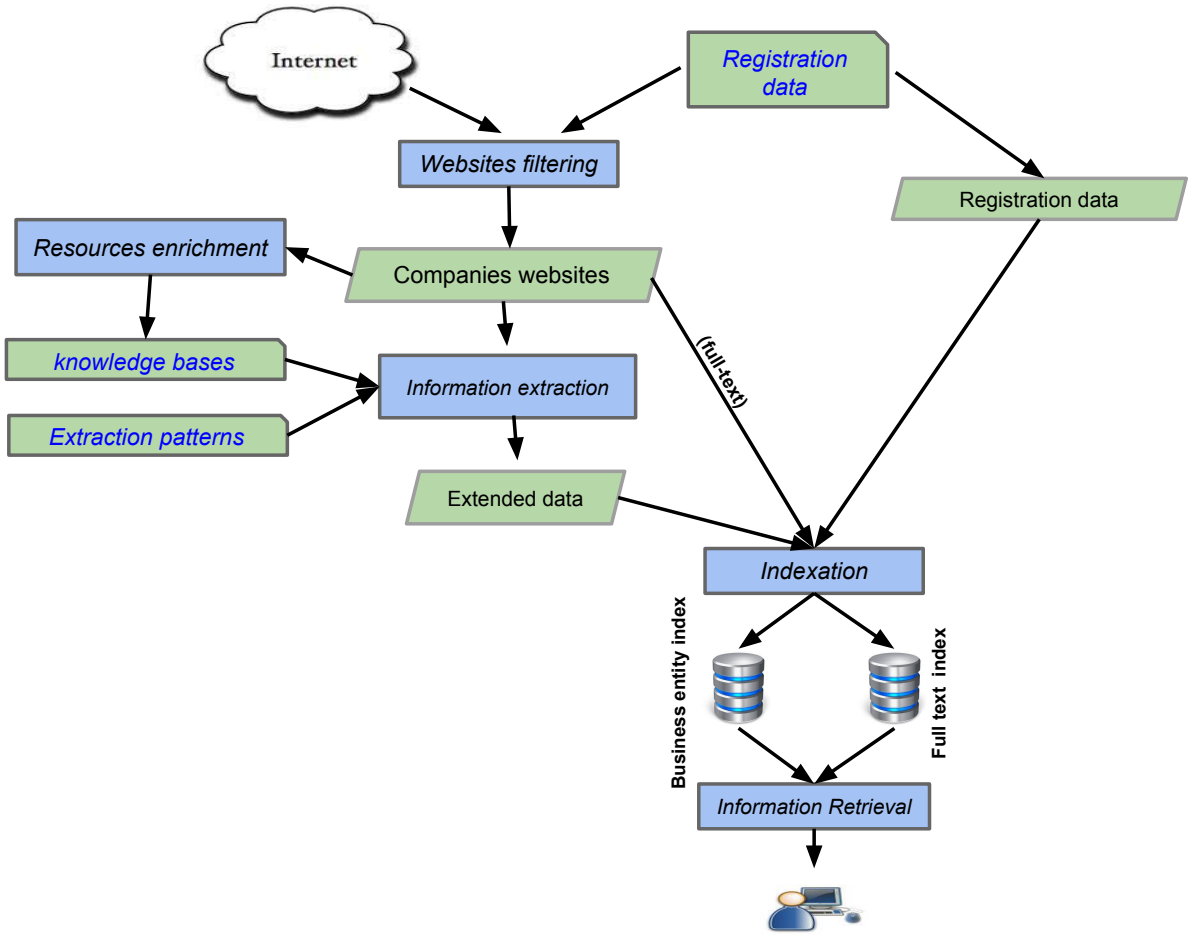
Latent Dirichlet Allocation (LDA) is used for clustering (Blei et al. - 2003)



Information Extraction



Indexation and RI



Different approaches

Three main approaches for Address Extraction:

- **Ontology-based**:
  Borges et al. (2007): Recognition, extraction and geocoding of Brazilian addresses in web pages.

- **Learning-based**:
  Loos et Biemann (2008): Use of CRF algorithm for extraction of addresses in web pages.

- **Pattern-based**:
  Ahlers and Boll (2008): Extraction and validation of German addresses in web pages.

Our Proposal

Major issue: Identification of the street name.

Example of complex address

Z.I. du Phare - Mérignac
3, Impasse Rudolf Diesel, Bât A - BP 50227, 4ème Etg
F-33708, Mérignac Cedex, France

Process

Observation of a sample of 160 websites and pattern identification

Example of rule:

$$\begin{aligned}\textbf{Adresse} \rightarrow\ & CA?\ \ ((BP\ \ CS) \mid (CS\ \ BP) \mid BP \mid CS)?\\ & NV?\ \ NVo\ \ CA?\ \ ((BP\ \ CS) \mid (CS\ \ BP)\\ & \mid BP \mid CS)?\ \ ((CP\ C) \mid (C\ CP))\ \ NC?\ \ D?\ \ P?\end{aligned}$$

Illustration

Open Inventor is a high-performance 3D software development toolkit (SDK) for professional applications in Medical, CAD & Engineering, Oil & Gas and Mining. It is a subsidiary of FEI. The European headquarters are located in Mérignac:
Z.I. du Phare - Mérignac
3, Impasse Rudolf Diesel, Bât A - BP 50227, 4ème Etg
F-33708, Mérignac Cedex
France

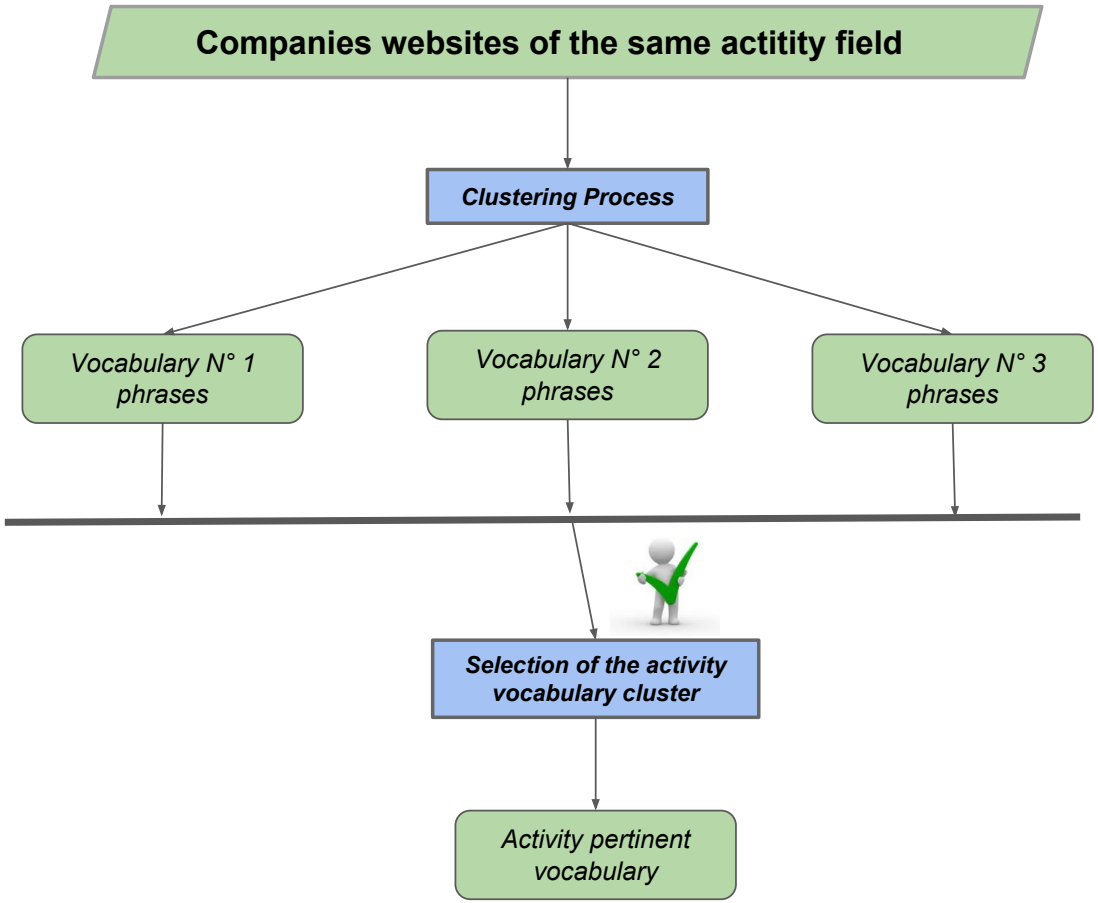
Illustration

Open Inventor is a high-performance 3D software development toolkit (SDK) for professional applications in Medical, CAD & Engineering, Oil & Gas and Mining. It is a subsidiary of FEI. The European headquarters are located in Mérignac:
Z.I. du Phare - Mérignac
3, Impasse Rudolf Diesel, Bât A - BP 50227, 4ème Etg
F-33708, Mérignac Cedex
France

Illustration

Open Inventor is a high-performance 3D software development toolkit (SDK) for professional applications in Medical, CAD & Engineering, Oil & Gas and Mining. It is a subsidiary of FEI. The European headquarters are located in Mérignac:
Z.I. du Phare - Mérignac
3, Impasse Rudolf Diesel, Bât A - BP 50227, 4ème Etg
F-33708, Mérignac Cedex
France

Process Flow



Activity field: 4391A - Carpentry Work

- 148 websites for 770 web pages
- 906,000 2-grams with 48 000 distinct ones
- 3 topics.

Vocabulary 1

Bargassat Simon
Dominique Pascal
Sinom Couverture
Mr Marc
Laborde JP
Dombon Willy
Charpente Charpent
Maison bois
Gouttière habillage
Corrihons Pierre

Vocabulary 2

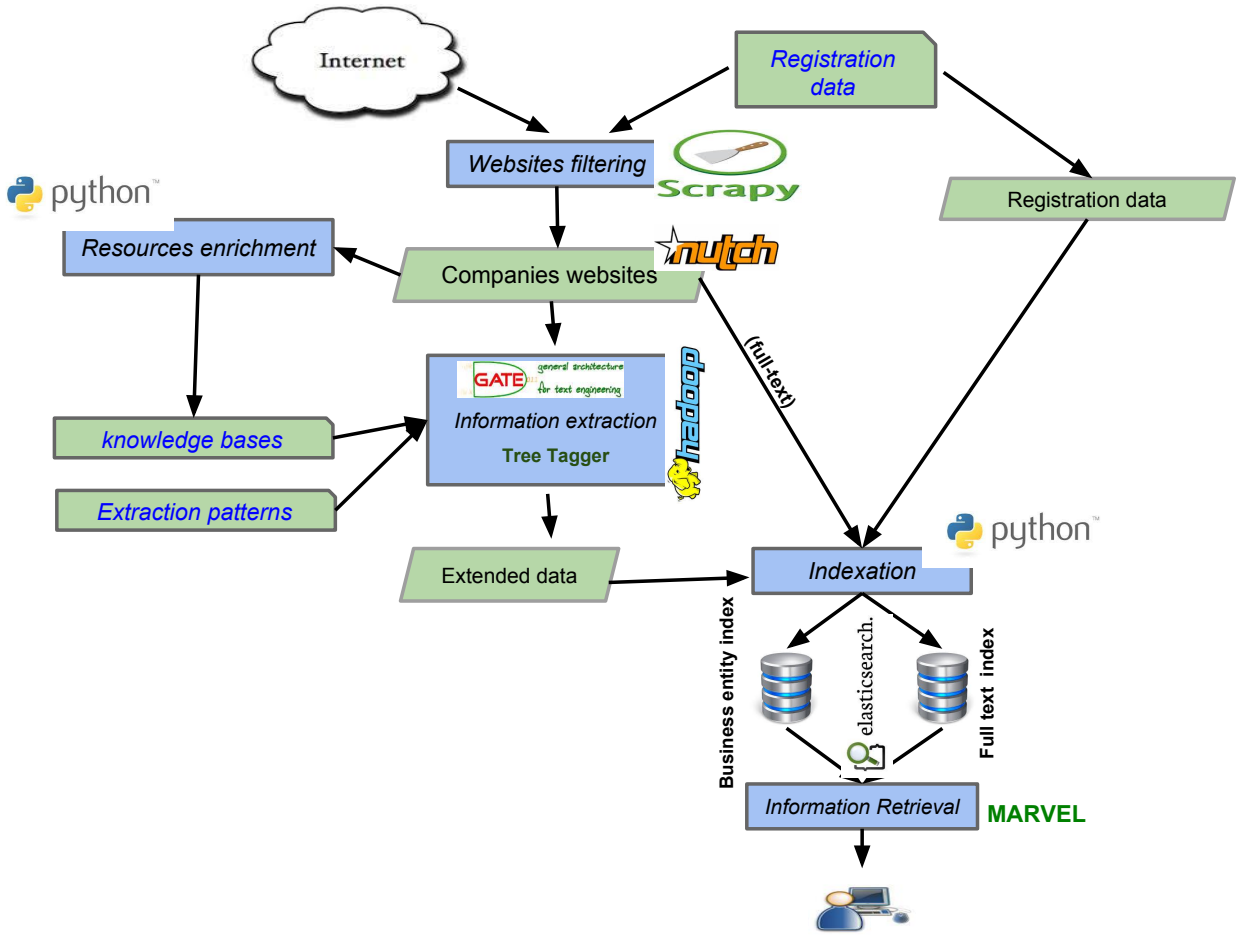
Charpente chêne
Chantier Saint
ossature bois
Entreprise charpente
Toiture ardoise
Charpente Couverture
Charpente traditionnelle
Zingage toiture
Charpente industrielle
Rénovation bois

Vocabulary 3

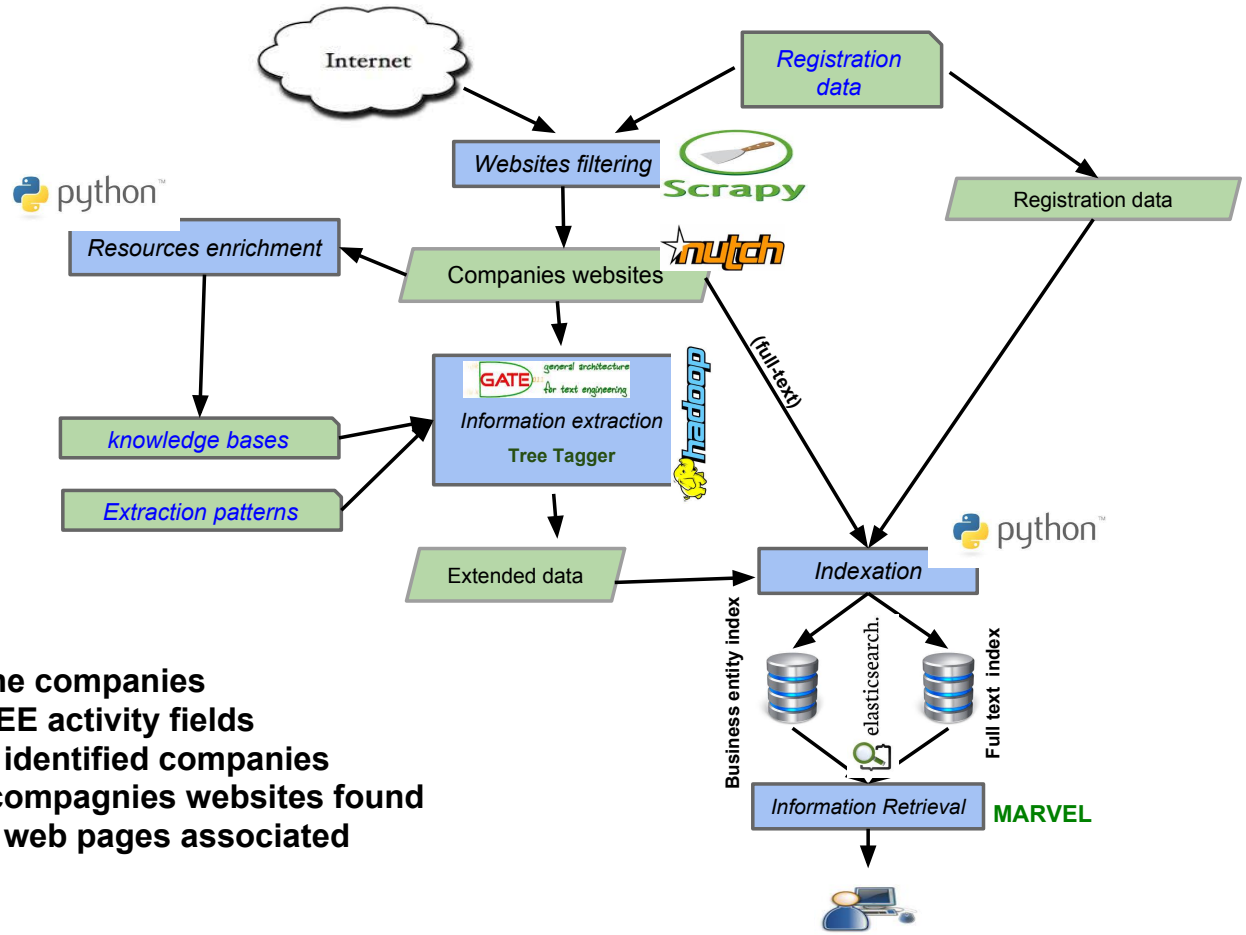
Construction Landreau
Maison industrielle
Charpente Bordes
Décoration construction
Aménagements extérieurs
Orx charpente
Reportage art
Ossature maison
Bois surélévations
Sarl Labouyrie

Prototype

Implementation technologies



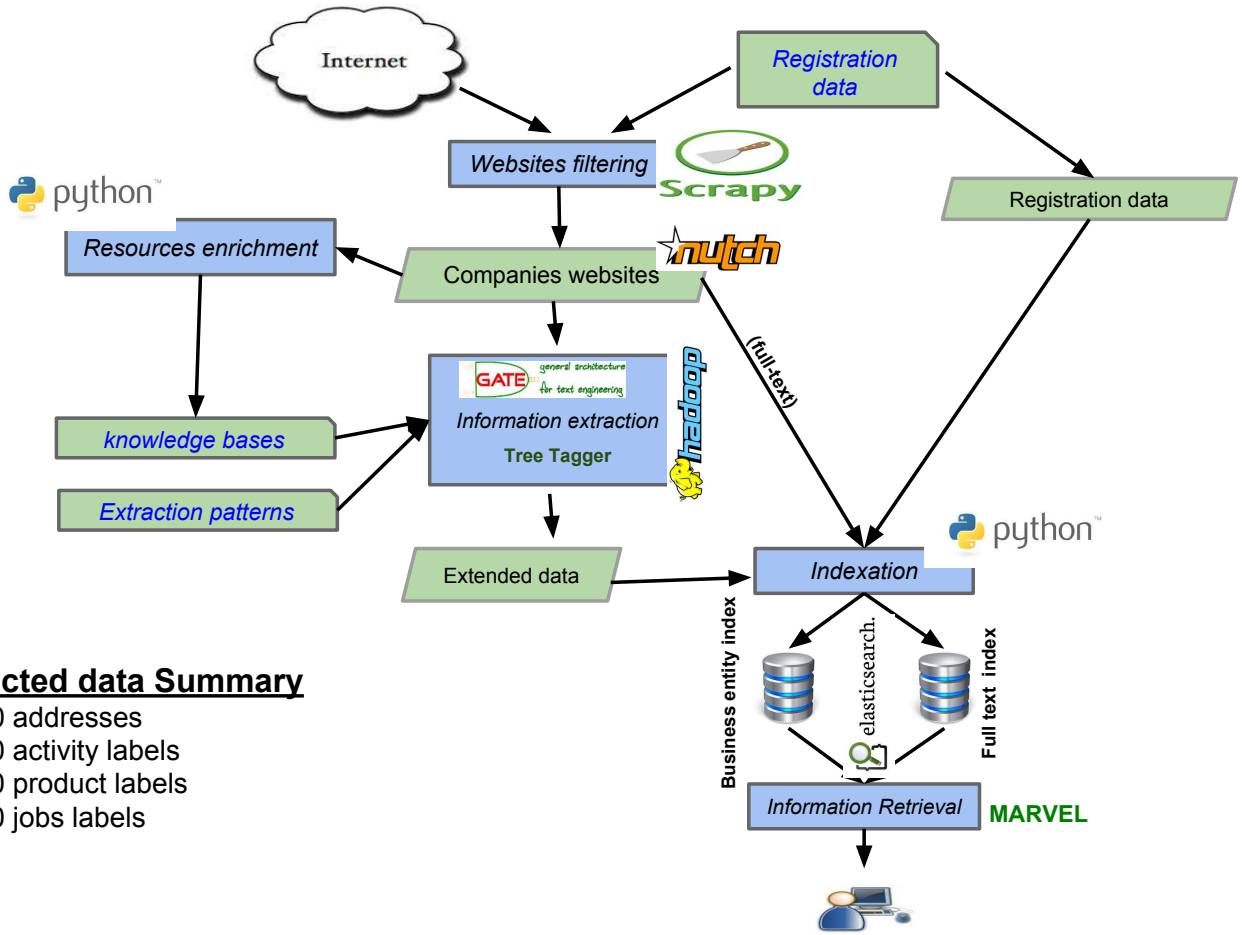
Process Flow Input



+ **Aquitaine companies**
+ **212 INSEE activity fields**
+ **115,000 identified companies**
+ **22,000 compagnies websites found**
+ **550,000 web pages associated**

Obtained results

- **Enrichment of activity resource classes related to "roof" (covering, zinc-work, capentry)**
- **26 labels added to the resource**



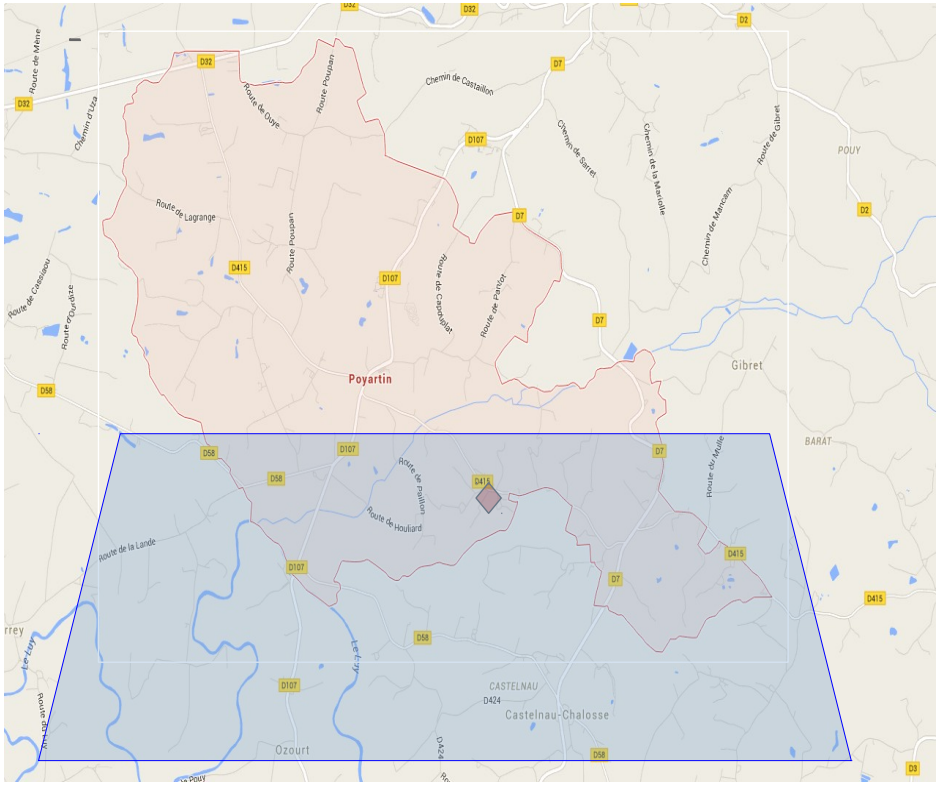
Extracted data Summary

- 30,000 addresses
- 44,000 activity labels
- 12,000 product labels
- 28,000 jobs labels

"oak beams south of Poyartin"

"What ?" "oak beams" → Carpentry work

"Where ?" "south of Poyartin"



Legend

South of Poyartin

Poyartin

S.E.E Laborde Jean Pierre

Address: 1323 Route Abbadie, 40380 Poyartin.

Activities: Covering
Carpentry
Zinc-work

Products: roofs

Jobs : -

Website: http://www.see-laborde.fr/

Phone Number: 05.58.98.69.19

Email: contact@see-laborde.fr

Conclusion

Conclusion

- Research areas explored : Learning, Information Extraction (pattern-based and knowledge-based approaches), etc.
- Proposition of an address extraction process
- Development of a prototype which illustrates the feasibility of the proposed approach

Future works will focus on :

- Evaluation of the website filtering and information extraction processes
- Extension of the enrichment to all the activity ontology
- Evaluation of the service with a set of representative queries

Thank you for your attention !

Similar services

Limits of these services

- They are supplied with:
  - Manually recorded data mostly
  - Partner companies data or bought one.
  - Open data
- They do not take into consideration topological relations in the spatial interpretation of the information needs.