Ge(o)Lo(cator): Geographic Information Extraction from Unstructured Text Data and Web Documents

Paolo Nesi, Gianni Pantaleo and Marco Tenti

Department of Information Engineering, DINFO

University of Florence,
Via S. Marta 3, 50139, Florence, Italy
Tel: +39-055-4796567, fax: +39-055-4796363

DISIT Lab
http://www.disit.org
paolo.nesi@unifi.it
gianni.pantaleo@unifi.it

9th Workshop on Semantic and Social Media Adaptation and Personalization
– SMAP 2014 –

1. Introduction – *Geographic Information Extraction: Application Areas*

- Automatic extraction and retrieval of geographic information from Web Domains and URLs is a field of large and increasing interest.

Application Areas:

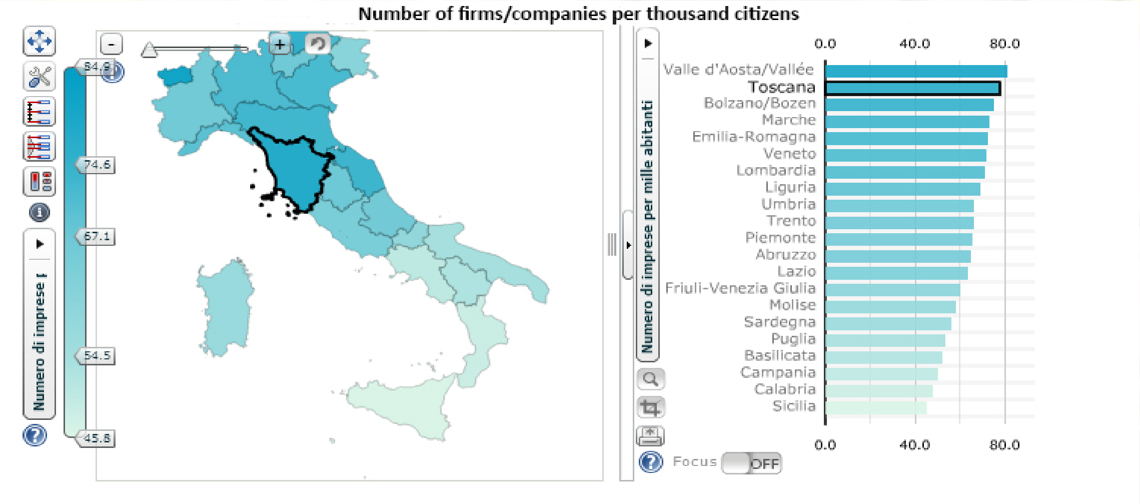
- Commercial and business interest: locating commercial services and operators, geo-marketing, geographic location of web sites visitors, fraud detection utilities, criminal investigation...

- Public Administration & Smart City services: locating PA services, spatial location of places of interest, tourism facilities, real-time traffic, mobility, parking information

1. Introduction – Main Issues and Problems

- Within typical Smart City services and applications, a major requirement is to geolocalize commercial operators and public utilities.
- Smart City services are often based upon unstructured Open Data representing services which are not always geolocalized

Only 30000 commercial operators appears to be active in the Tuscany region, according Public Administration Open Data and among GoodRelations Ontology users.



1. Introduction – *Main Issues and Problems*

- Additional information, with respect to PAs Open Data, is represented by web domains owned by commercial utilities and companies. The necessity arises to associate web domains to geographic information related to the owner/responsible of the web domain.

- Extract geographical information potentially related with physical/legal location of the web domain owner, handling ambiguities and minimizing noisy data.

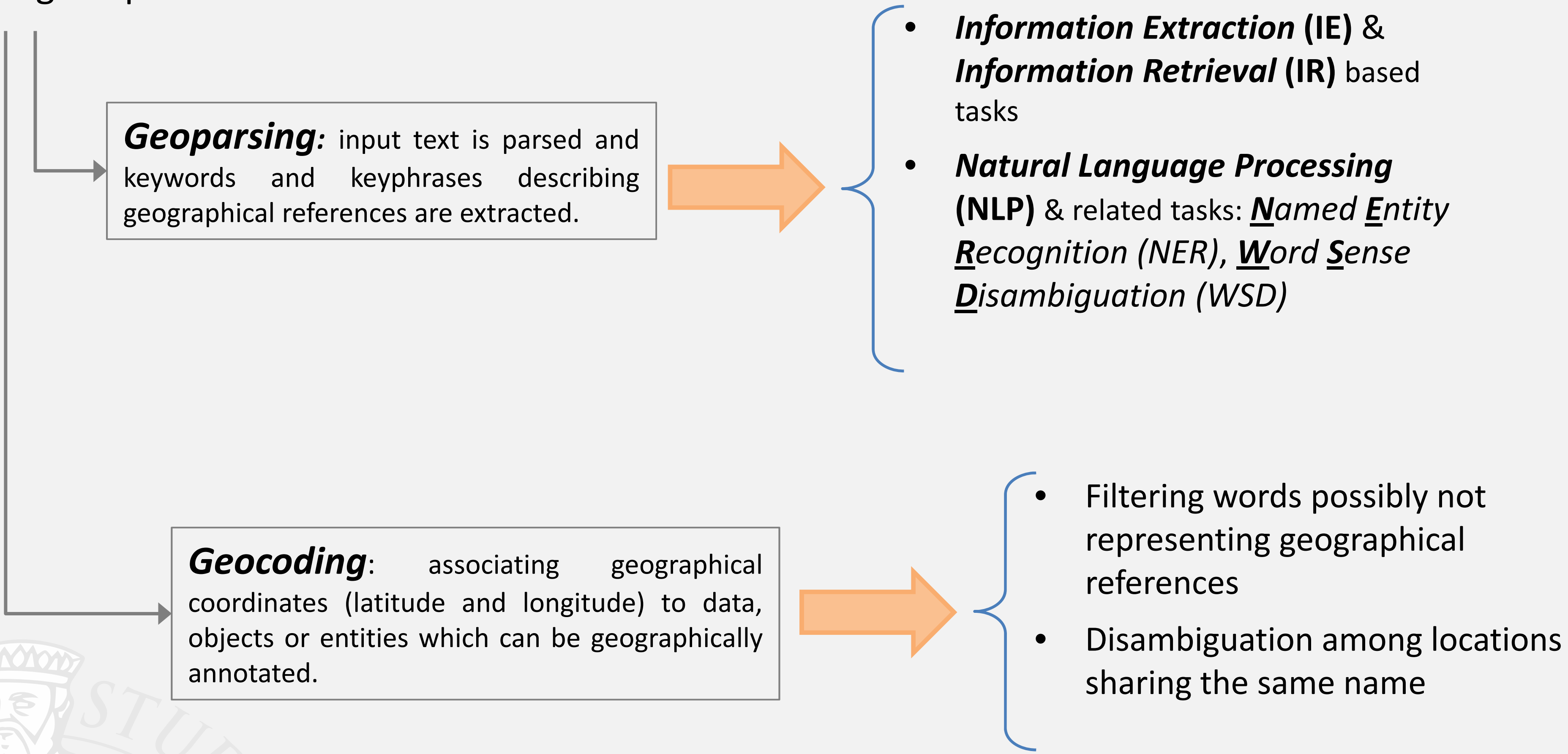
- Many Smart City services require a very high spatial resolution (at civic number level). For this purposes, Existing *IP-based / Whois* geolocation tools are not suitable (they retrieve a geographic information associated with the location of the Internet Service Provider, not the physical location of the organization / company / service.

- In addition, different web domains and their hosts are commonly assigned to a single, centralized Provider.

- Web pages and documents are usually composed by collections of unstructured, natural language text, giving raise to several types of linguistic ambiguities (e.g., 2,100 toponym labels exactly matching the name "*San Antonio*" in the database of the USA National Geospatial-Intelligence Agency).

1. Introduction – *Geographic Information Retrieval*

➢ ***Geographic Information Retrieval (GIR):*** Extraction of information involving some kind of geo-spatial reference**.**



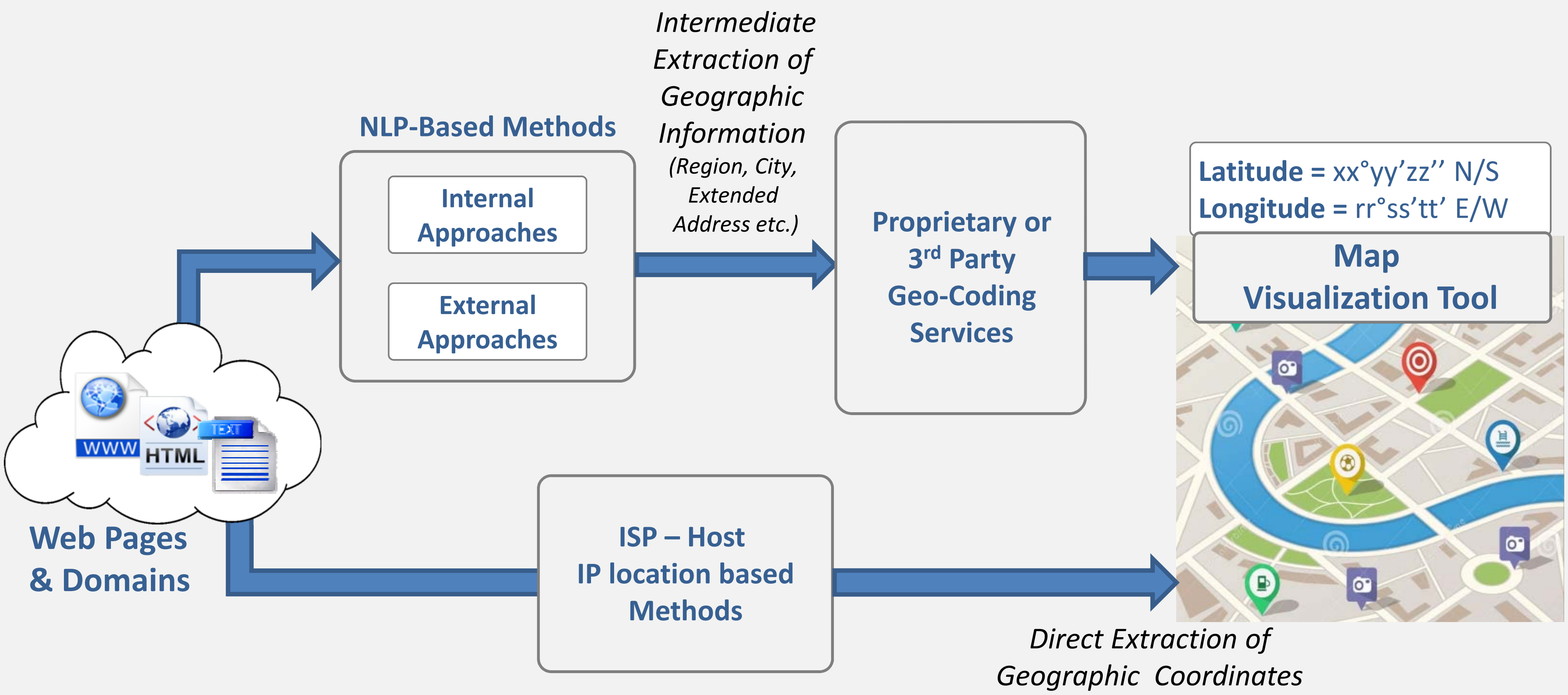
Geoparsing*:* input text is parsed and keywords and keyphrases describing geographical references are extracted.

- ***Information Extraction* (IE)** & ***Information Retrieval* (IR)** based tasks
- ***Natural Language Processing* (NLP)** & related tasks: ***N**amed **E**ntity **R**ecognition (NER), **W**ord **S**ense **D**isambiguation (WSD)*

Geocoding: associating geographical coordinates (latitude and longitude) to data, objects or entities which can be geographically annotated.

- Filtering words possibly not representing geographical references
- Disambiguation among locations sharing the same name

1. Introduction – *Geographic Information Retrieval*



NLP-Based Methods

Internal Approaches

External Approaches

Intermediate Extraction of Geographic Information
(Region, City, Extended Address etc.)

Proprietary or 3rd Party Geo-Coding Services

Latitude = xx°yy'zz'' N/S
Longitude = rr°ss'tt' E/W

Map Visualization Tool

Web Pages & Domains

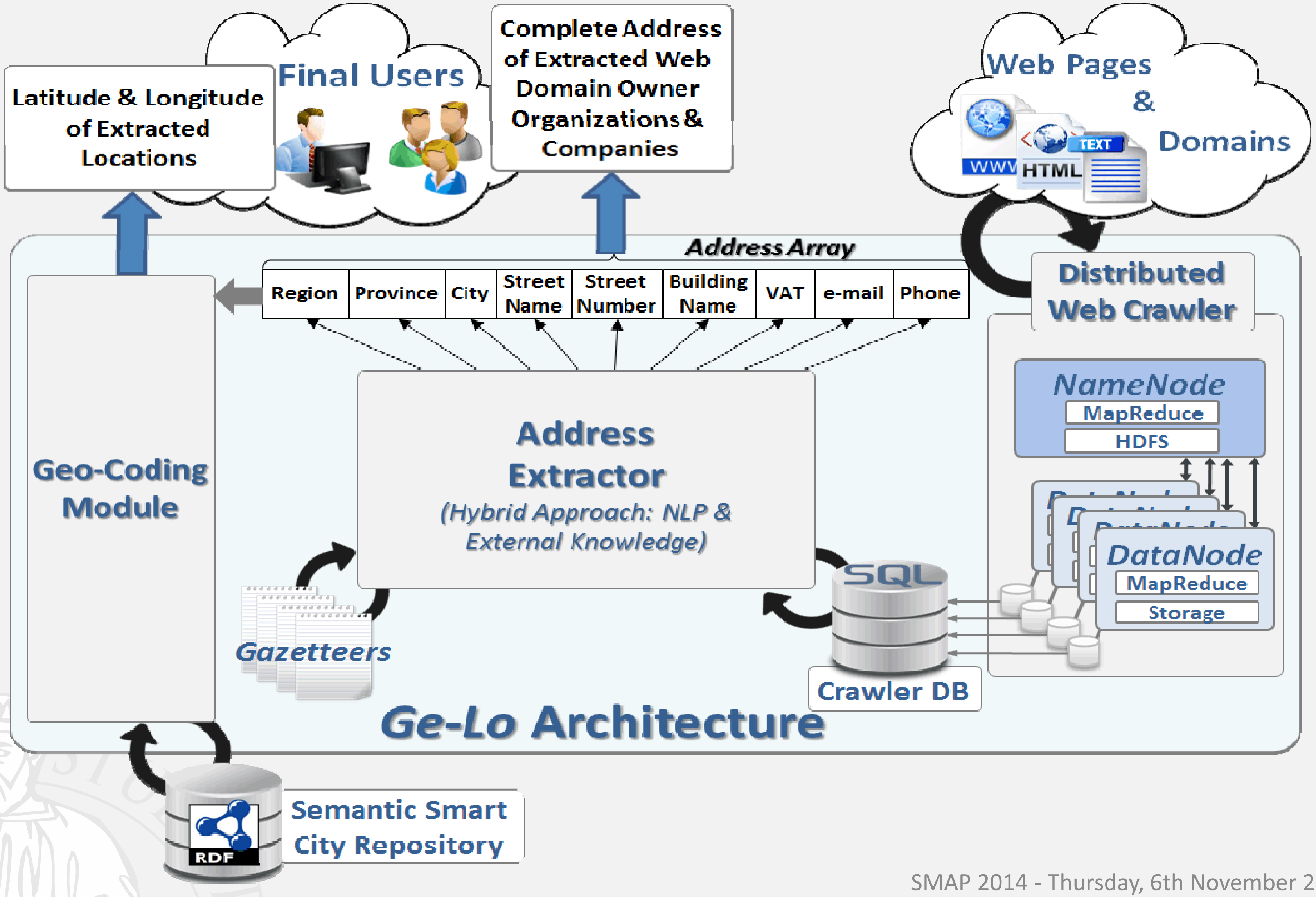
ISP – Host IP location based Methods

Direct Extraction of Geographic Coordinates

2. Ge(o)Lo(cator) System Description

- The ***GeLo*** framework has been designed and developed to achieve the following goals:
  - Mining, retrieving and geolocalizing web-domains associated to companies and research institutes located in the Tuscany region (Italy).
  - The extraction of geographical information is based on a ***hybrid approach***:
    - ***internal*** : *linguistic parsing, Part-of-Speech (POS) tagging, pattern based annotations.*
    - ***external***: *use of external annotated gazetteers containing names of provinces and cities in the Tuscany region, names of local commercial companies, research institutions, public administrations (PAs) and other types of organizations.*
  - 2 Different operating modes: ***Language-Dependen***t (Italian & English supported at the moment) and ***Language-Independent***.
  - Use of a ***semantic Smart City repository*** (RDF datastore), implemented at DISIT Lab, for the geocoding task.

2. Ge(o)Lo(cator) System Description – Architecture (1 of 5)

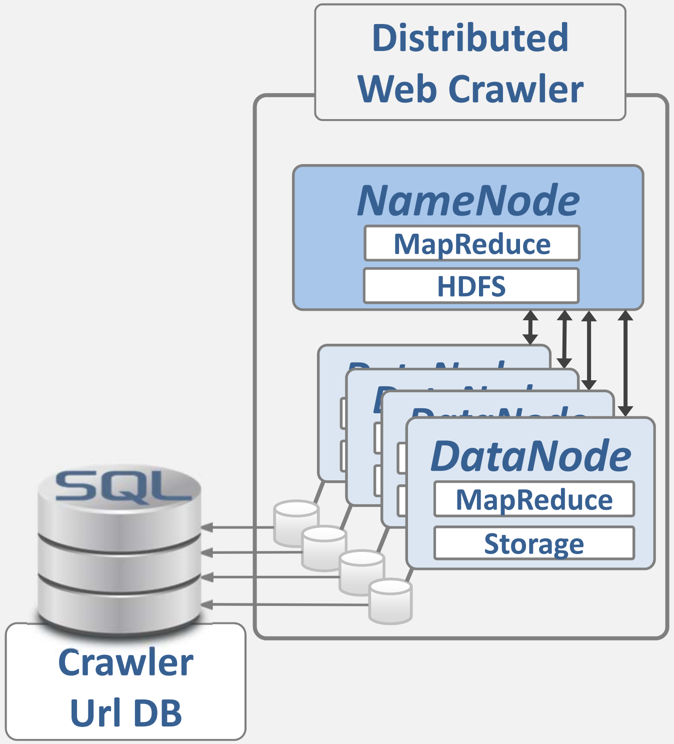


2. Ge(o)Lo(cator) System Description – *Architecture (2 of 5)*

Distributed Web Crawler

- Based on the open source ***Apache Nutch*** crawling tool.
- Distribuite and parallel crawling process based on the open source ***Apache Hadoop*** distributed filesystem (HDFS)

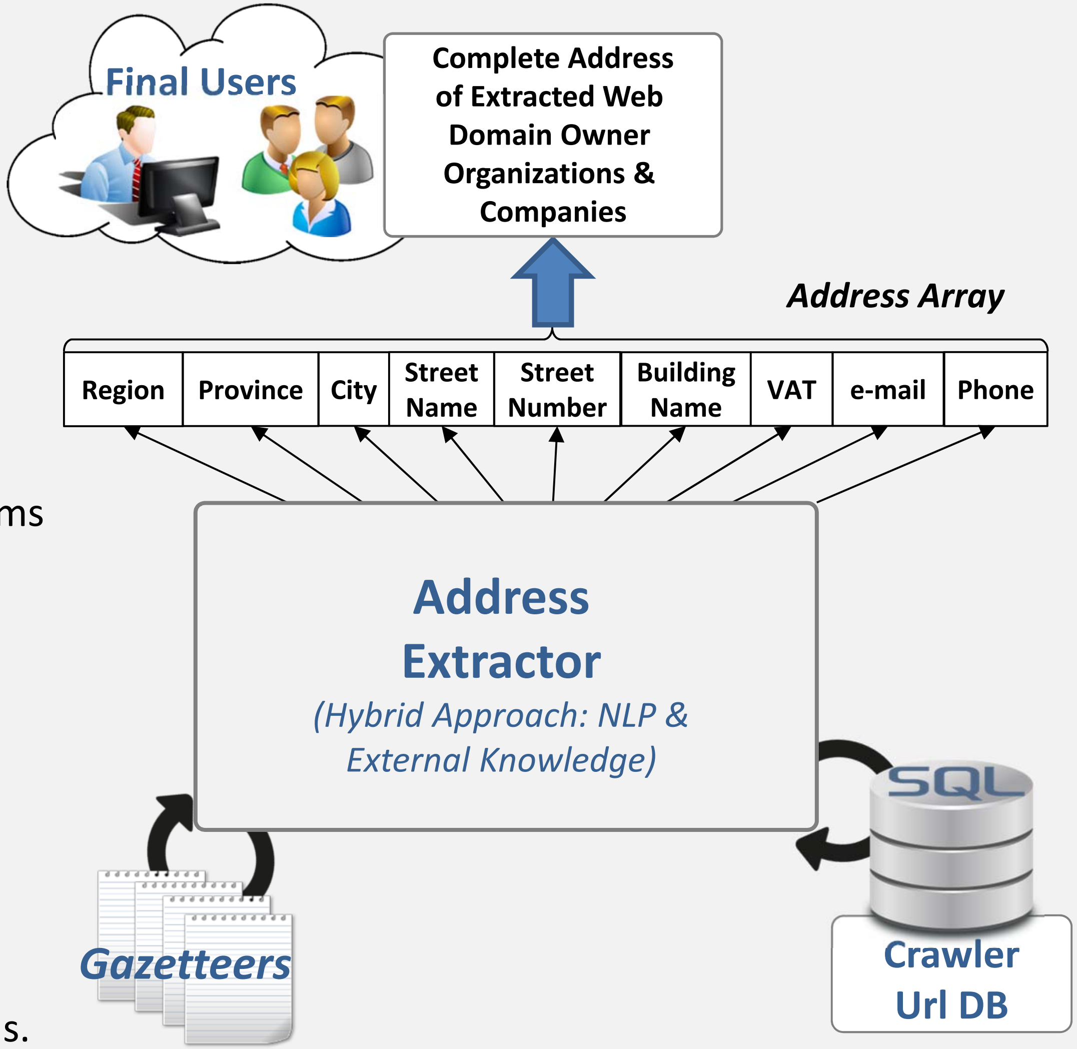
  MapReduce programming paradigm
- High scalability: HDFS and Data on a commodity cluster.
- More than 6 million URLs fetched (belonging to Universities, research institutes and companies in the area of Italian Tuscan Region).



2. Ge(o)Lo(cator) System Description – *Architecture (3 of 5)*

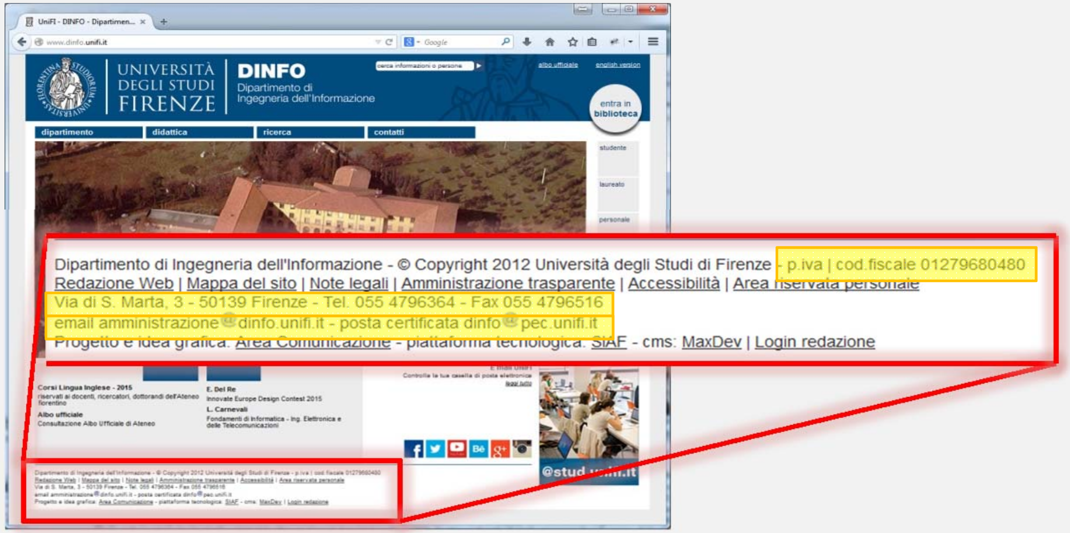
Address Extractor (1)

- Hybrid approach:
  - NLP-based, linguistic rules, POS-tagging (based on the Open Source **GATE** Framework)
  - Use of external gazetteers providing names of Italian cities, regions, companies, abbreviations for address items and identifiers.
- Address Extraction Algorithm steps:
  - Search for geographic information in HTML specific tags (footer, header).
  - If no useful information is retrieved, the system continues searching in the rest of the page (unstructured text) by means of specifically designed patterns.



2. Ge(o)Lo(cator) System Description – *Architecture (4 of 5)*

Address Extractor (2)



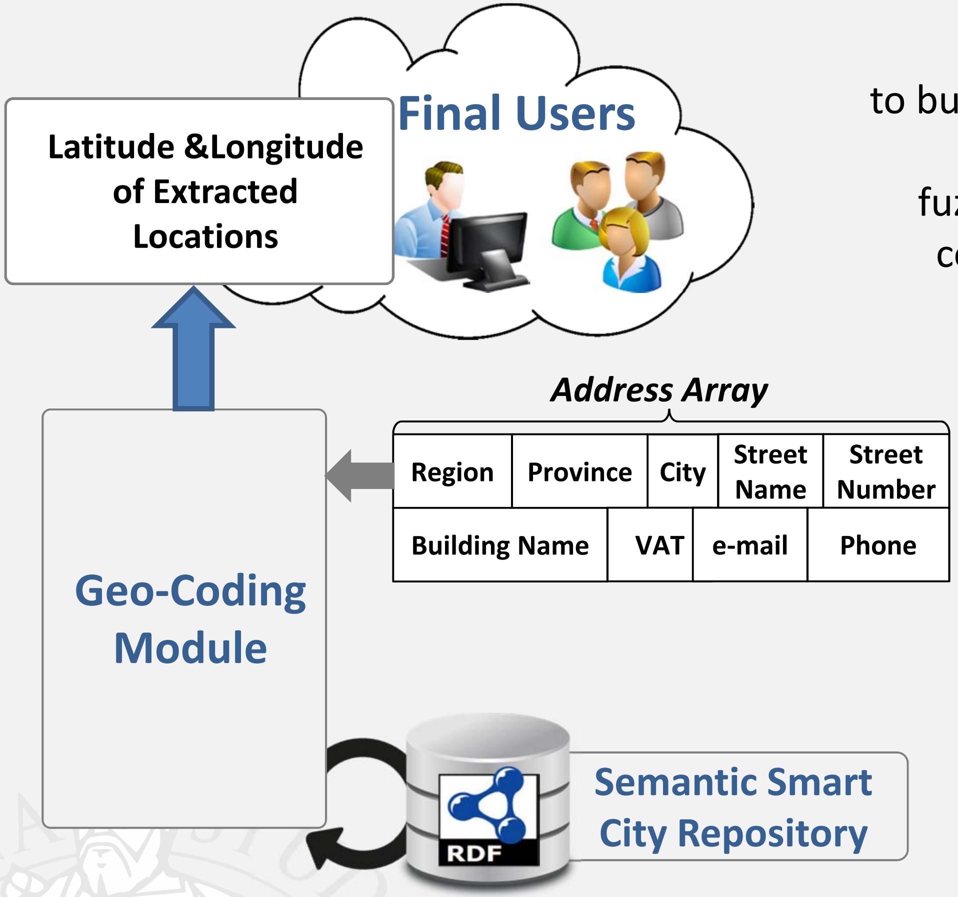
Dipartimento di Ingegneria dell'Informazione - © Copyright 2012 Università degli Studi di Firenze - p.iva | cod.fiscale 01279680480
Redazione Web | Mappa del sito | Note legali | Amministrazione trasparente | Accessibilità | Area riservata personale
Via di S. Marta, 3 - 50139 Firenze - Tel. 055 4796364 - Fax 055 4796516
email amministrazione@dinfo.unifi.it - posta certificata dinfo@pec.unifi.it
Progetto e idea grafica: Area Comunicazione - piattaforma tecnologica: SIAF - cms: MaxDev | Login redazione

- ➢ HTML tags are often used to contain and display administrative and physical information of the company/organization owner / responsible / representative or domain (*footer* and *header*).

- ➢ For complete address extraction, the following Address Patterns have been defined and searched:
  - *General (high level detail) address pattern*:
    `[REGION] + [PROVINCE] + [POSTAL_CODE] + [CITY].`
  - *Specific (low level detail) address pattern*:
    `[STREET_IDENTIFIER] + [STREET_NAME] + [STREET_NUMBER].`
  - Pattern for retrieving address attributes in the form of internal block (e.g.: “Scala A, Interno 4”):
    `[INNER_BLOCK_ID1] + [ID1_VALUE] + [INNER_BLOCK_ID2] + [ID2_VALUE].`
  - *Pattern for extracting geographical coordinates* (if tpresent):
    `[LATITUDE_IDENTIFIER] + [LATITUDE_VALUE] + [LONGITUDE_IDENTIFIER] + [LONGITUDE_VALUE].`

2. Ge(o)Lo(cator) System Description – *Architecture (5 of 5)*

Geocoding Module


Final Users
Latitude &Longitude of Extracted Locations
Address Array
Geo-Coding Module
RDF
Semantic Smart City Repository

Region	Province	City	Street Name	Street Number
Building Name	VAT	e-mail	Phone	

- Sparsely populated ***Address Array*** fields are used to build an expanded query which is sent to the Smart City semantic repository SPARQL endpoint (provided with fuzzy search capabilities), in order to extract geographic coordinates of retrieved locations, where it is possible.

- ***Ge-Lo*** system can be easily adapted for working with any other geocoding repository or services, outside the Tuscany Region domain (3rd parties *OpenStreetMap*, *Google Maps*, *Google Geocoding* APIs…).

- Collecting also nongeographic, high level features like labels, (e.g. the building name), allows the system to potentially retrieve correct coordinates even for domains with missing or incomplete address or other fields.

3. Validation

- Dataset & Ground Truth: subset of 100000 web URLs from the Crawler Database. Manual evaluation upon a Ground Truth generated by manual annotation of geographic information.

- Evaluation of ***Ge-Lo*** system geocoding performances through standard IR parameters: *Precision*, *Recall* and *F-Measure*:

$$Precision = \frac{TP}{TP+FP}; \qquad Recall = \frac{TP}{TP+FN}; \qquad F-Measure = 2\frac{Precision \cdot Recall}{Precision+Recall}.$$

True Positive (TP): geographical information provided by ***Ge-Lo*** matches the place/location which is perceived by the human evaluators.

False Positive (FP): geographical information provided by ***Ge-Lo*** does not match the place or location which is per-ceived by the human evaluators.

False Negative (FN): ***Ge-Lo*** is not able to extract any geographical informa-tion, while human evalua-tors actually does.

True Negative (TN): neither ***Ge-Lo*** nor the human evaluators are able to find geographical information for a web domain.

- Two Steps Validation:
  - A. *Evaluation of Complete Address Array Extraction*
  - B. *Evaluation of Geographic Coordinates Extraction*

3. Validation – Results

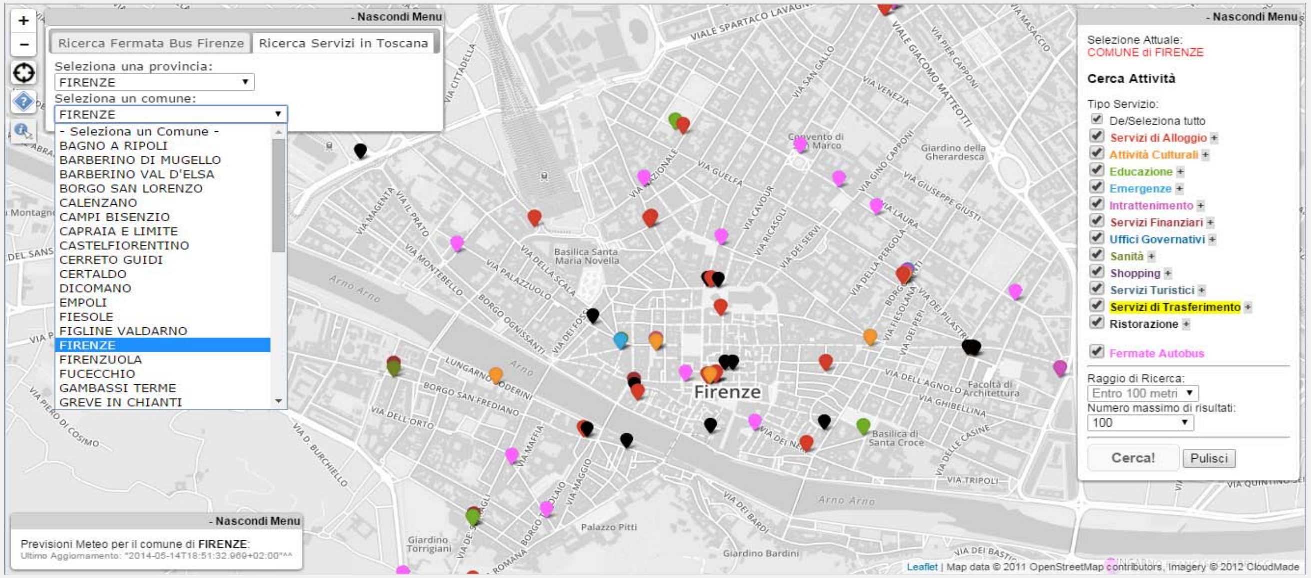
Evaluation Tasks	TP	FP	FN	TN	Precision	Recall	F-Measure
(1) Address Array Extraction	74.5%	7.8%	5.9%	11.8%	90.5%	92.7%	91.6%
(2a) Geographic Coordinates Extraction (Smart City Semantic Repository)	57.8%	4.7%	29.5%	8.0%	92.5%	66.2%	77.1%
(2b) Geographic Coordinates Extraction (Google Geocoding)	48.9%	31.1%	11.1%	8.9%	61.1%	81.5%	69.8%

- *Precision* rate for geographic coordinates extraction (employing the Smart City Semantic Repository) has increased, with respect to the value obtained in the evaluation of address array extraction.
- Slightly decreasing *TN* rate for Test (2a) with respect to Test (1): exploiting the extraction of high level features (such as building names) allows the system to obtain correct coordinates even for domains with incomplete Address Array.
- *Recall* rate for Test (2a) significantly decrease with respect to Test (1). This is due mainly to the noise generated by the supplementary logic and the extended semantic queries required to obtain the geographical coordinates.
- Higher *Recall* rate achieved when using the Google Geocoding APIs (2b): Google Repository is by far larger than DISIT Smart City RDF datastore, so that it is able to index a huge amount of resources, even if this can affect the precision rate.

4. Conclusions & Future Work

- Web domains related to commercial utilities which have been geolocalized by the ***Ge-Lo*** system will be integrated in the Smart City ontology realized by the DISIT Lab for the Sii-Mobility Project: a platform for management and monitoring of services related to the Florence Municipality transport systems.

- Geographical Information extracted with ***Ge-Lo*** will contribute to enrich and reconcile semantic repositories for existing services monitoring commercial activities in the Tuscany Region, such as the DISIT Service Map.



Ricerca Fermata Bus Firenze | Ricerca Servizi in Toscana

Seleziona una provincia: FIRENZE

Seleziona un comune: FIRENZE

- Seleziona un Comune -
BAGNO A RIPOLI
BARBERINO DI MUGELLO
BARBERINO VAL D'ELSA
BORGO SAN LORENZO
CALENZANO
CAMPI BISENZIO
CAPRAIA E LIMITE
CASTELFIORENTINO
CERRETO GUIDI
CERTALDO
DICOMANO
EMPOLI
FIESOLE
FIGLINE VALDARNO
FIRENZE
FIRENZUOLA
FUCECCHIO
GAMBASSI TERME
GREVE IN CHIANTI

Previsioni Meteo per il comune di **FIRENZE**:
Ultimo Aggiornamento: "2014-05-14T18:51:32.969+02:00"^^

Selezione Attuale:
COMUNE di FIRENZE

Cerca Attività

Tipo Servizio:
☑ De/Seleziona tutto
☑ Servizi di Alloggio +
☑ Attività Culturali +
☑ Educazione +
☑ Emergenze +
☑ Intrattenimento +
☑ Servizi Finanziari +
☑ Uffici Governativi +
☑ Sanità +
☑ Shopping +
☑ Servizi Turistici +
☑ Servizi di Trasferimento +
☑ Ristorazione +

☑ Fermate Autobus

Raggio di Ricerca: Entro 100 metri

Numero massimo di risultati: 100

Thanks for Your Attention !