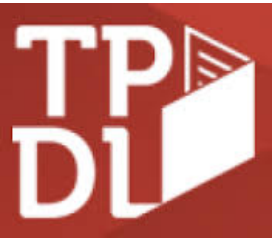

# Digital Geolinguistics: On the use of Linked Open Data for Data-Level Interoperability between Geolinguistic Resources

Giorgio Maria Di Nunzio

Dept. of Information Engineering, University of Padua

3rd InternationalWorkshop on Semantic Digital Archiving

September 26th, Valletta Malta



DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

# Outline

- Background
- Projects
- Linked Open Data Approach



# Background



# Linguistics

- Studies aspects of human language
  - Morphology
  - Syntax
  - Phonology
- Understand the fundamental principles that underlie
  - language differences
  - language innovation
  - language variation in time and space



# Dialect

- The variation of a language is called dialect
- The study of these variations has led to the constitution of two research fields:
  - Dialectology, concerned with grammatical, lexical and phonological features that correspond to regional areas
  - Dialectometry, concentrates on the regional distribution of dialect similarities
    - http://www.washingtonpost.com/wp-srv/special/national/us-language-map/

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

# Dialect Maps

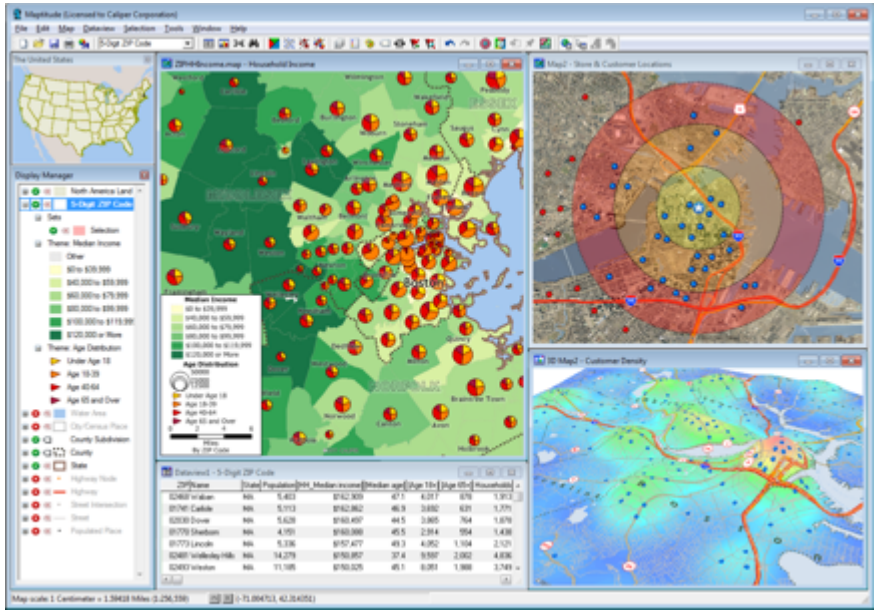
- Since the end of the XIX century, linguists have produced extensive cartographic work, most notably in the form of linguistic atlases

- Manuel Alvar López presented in 1981 an automated linguistic atlas which highlighted the advantages of a computerized versus manually drawn and reproduced atlas (!)

# Dialect Maps

# Geographic Information Systems

- In the last thirty years, modern Geographic Information System (GIS) have provided efficient analysis of spatial data in many fields.



DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

# Geolinguistics 1/2

- Geolinguistics is an interdisciplinary field which incorporates language maps
  - depicting spatial patterns of language location
  - or the results of processes that lead to language change

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

# Geolinguistics 2/2

- The synergy between geography and linguistics
  - Breton and Schiffman (1991, Geolinguistics: language dynamics and ethnolinguistic geography)

- They described the process through which a geographic thought becomes a tool for linguists:

  *In analyzing the distribution in space and in society of the facts of language, the linguist employs the methods of geography, cartography and the establishment of correlations and causalities between spatial phenomena.*

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

# Linguistic Atlas

- It has proved to be a vital tool and product of geolinguistics
- The World Atlas of Languages Structures (WALS)
  - first linguistic feature atlas on a world-wide scale.
  - 160 maps showing the geographical distribution of structural linguistic features
- http://wals.info/

DEPARTMENT OF
INFORMATION
ENGINEERING
UNIVERSITY OF PADOVA

# Projects



# European Projects 1

- CLARIN, Common Language Resources and Technology Infrastructure
- Research Infrastructure that was selected for the European Research Infrastructures Roadmap by ESFRS
  - create an infrastructure which makes language resources and technology available and readily usable to scholars of all disciplines.
- http://www.clarin.eu/vlo



# European Projects 2

- EdiSyn
  - Funded by the European Science Foundation
  - Establish a European network of researchers using similar standards with respect to methodology of data collection, data storage and annotation, data retrieval and cartography.

- http://www.meertens.knaw.nl/edisyn/searchengine/

- The Edisyn search engine (make different dialectal databases comparable)
  - "in practice has proven to be unfeaseable"



# Linguistic Projects 1

- The Open Language Archives Community (OLAC)
- worldwide network dedicated to
  - collecting information on language resources (field notes, grammars, audio/video recording, descriptive papers)
  - developing standard protocols for interoperability
- http://www.language-archives.org/

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

# Linguistic Projects 2

- General Ontology for Linguistic Description (GOLD)
- First ontology to be designed specifically for linguistic description on the Semantic Web.
- It proposes a solution to the lack of interoperability between linguistic projects and projects designed specifically for NLP applications.
- http://linguistics-ontology.org/

# Geolinguistic Projects 1

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

- The ALD: Linguistic Atlas of Dolomitic Ladinian and Neighbouring Dialects

- http://ald2.sbg.ac.at/a/index.php/en/the-project/

# Geolinguistic Projects 2

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

- VIVaio Acustico delle Lingue e Dialetti d'Italia (VIVALDI)

- http://www2.hu-berlin.de/vivaldi/index.php?id=0004&lang=de



# Geolinguistic Projects 3

- ALAVAL: The Atlas Linguistique Audiovisuel du Francoprovençal Valaisan

- http://www2.unine.ch/dialectologie/page-8174.html



# Geolinguistic Pojects 4

- SoundComparisons

- http://www.soundcomparisons.com/



# Geolinguistic Pojects 5

- LL-MAP: Language and Location - Map Accessibility Project

- http://www.llmap.org/



- Atlas of everyday German language

- http://www.atlas-alltagssprache.de/

DEPARTMENT OF
INFORMATION
ENGINEERING
UNIVERSITY OF PADOVA

# Linked Open Data Approach

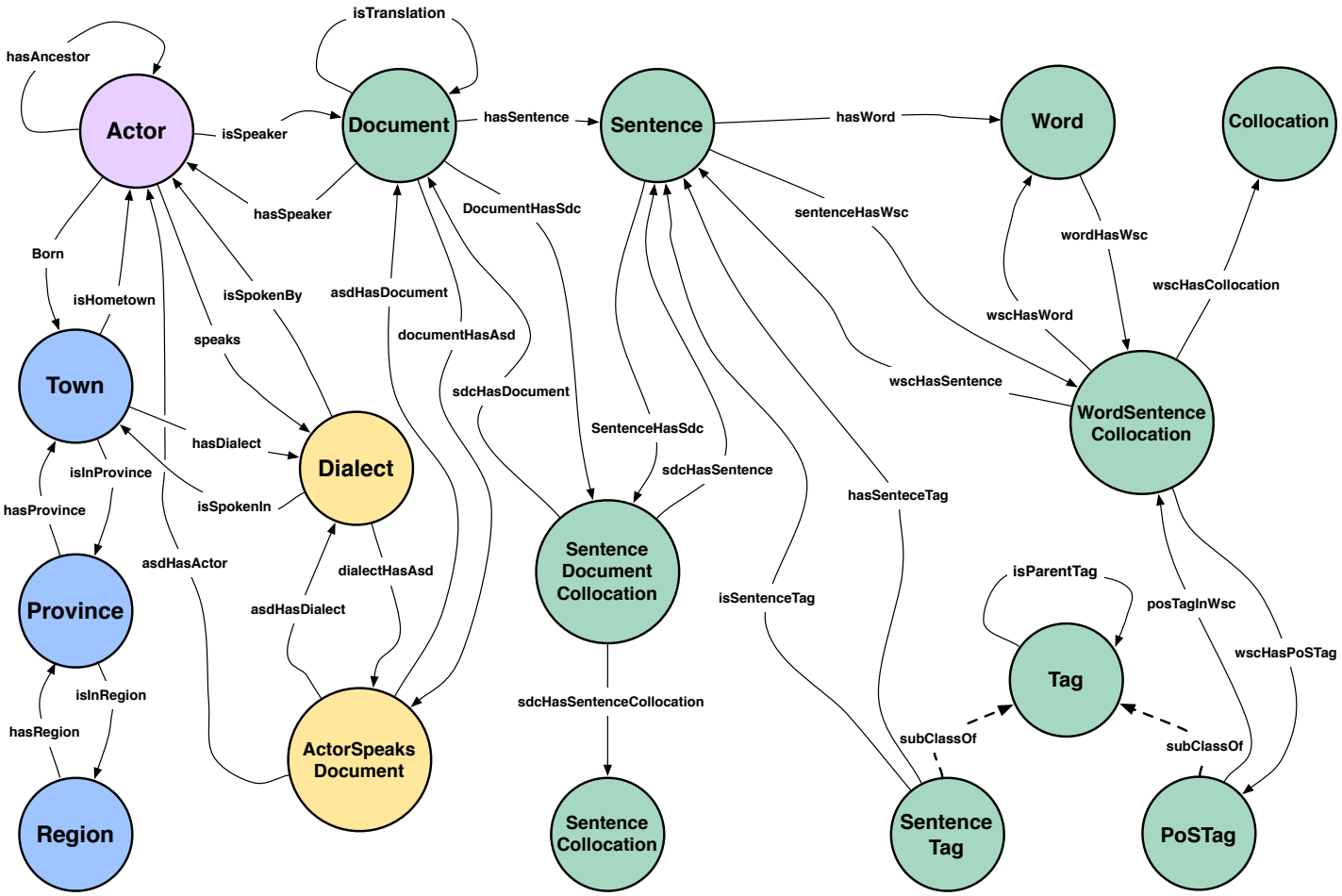


# Ontology for Geolinguistic Resources

- The common ground defined by current European geolinguistic projects allows us to infer the fundamental classes and properties necessary to define an ontology for modeling and representing geolinguistic resources.

- Three major areas
  - Geographic
  - Derivation
  - Taggin

# Ontology for Geolinguistic Resources





# Ontology for Geolinguistic Resources

- http://ims.dei.unipd.it/websites/ASIt/RDF/asit-schema.rdf

- This ontology is the starting point for modeling and describing geolinguistic resources because:
  - it provides general-purpose concepts and relationships;
  - it is extendable by adding more fine-grained classes;
  - it permits an easy mapping from existing linguisticprojects and publicly available databases.



# Browsing RDF

- http://svrims2.dei.unipd.it:8080/asit-enterprise/do/page/Region/Lombardia

- RDF search with LODLive
- http://svrims2.dei.unipd.it:8080/asit-enterprise/do/rdfSearch

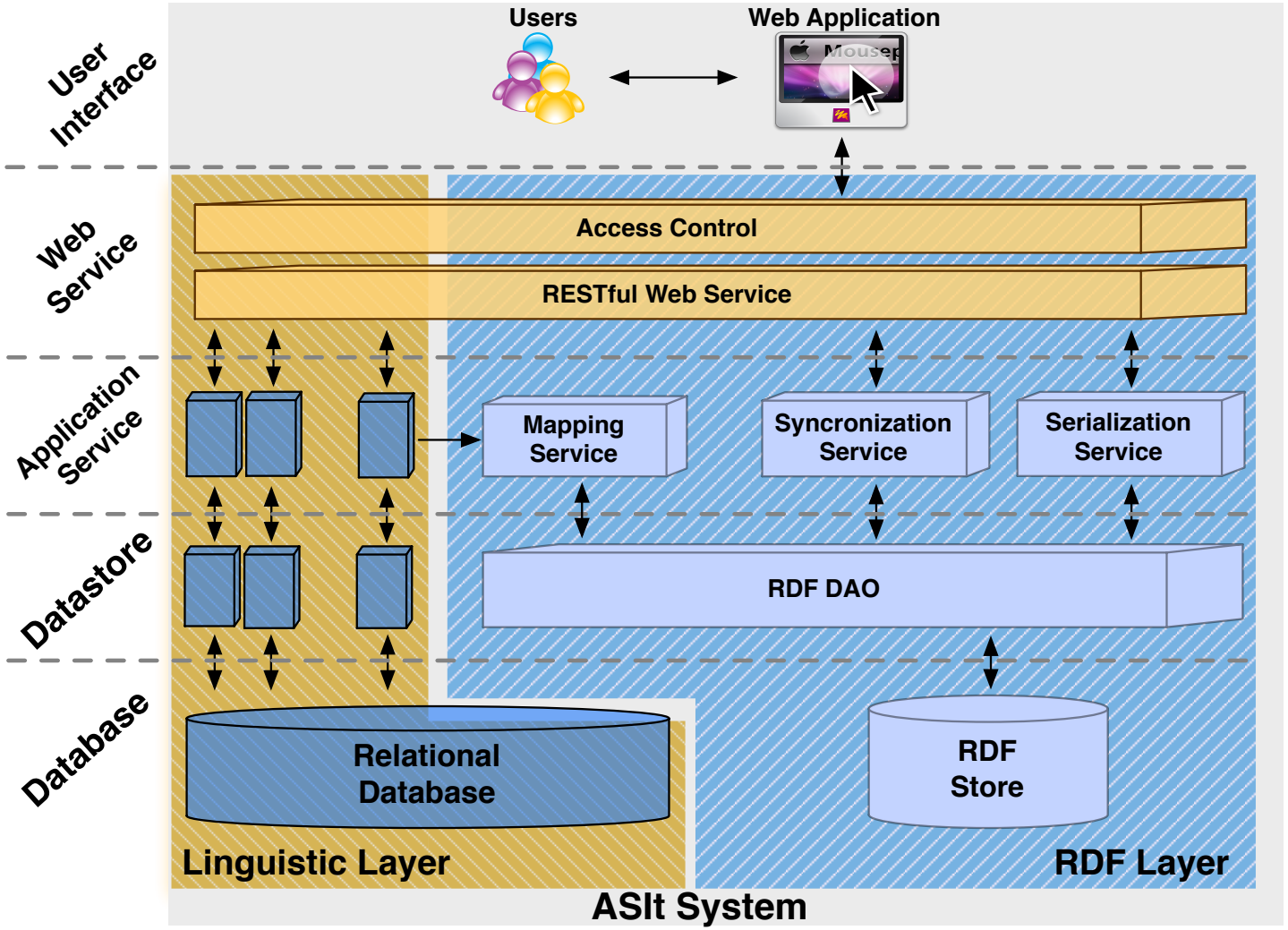
# SPARQL Querying

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

http://svrims2.dei.unipd.it:8080/asit-enterprise/do/sparqlGui

```
PREFIX asit: <http://purl.org/asit/terms/>
SELECT DISTINCT ?q ?s ?sp ?t
WHERE {
?s asit:hasSentenceTag
<http://purl.org/asit/resource/SentenceTag/QP_aggiunto> .
?s asit:hasSentenceTag
<http://purl.org/asit/resource/SentenceTag/clit_climb> .
?q asit:hasSentence ?s .
?s asit:sentence ?t .
?qt asit:isTranslation ?q .
?sdc asit:sdcHasSentence ?s .
?sdc asit:sdcHasDocument ?q .
?sdc asit:sdcHasSentenceCollocation ?sc .
?sc asit:sentencePosition ?sp
} ORDER BY ?q ?sp
```

# Application





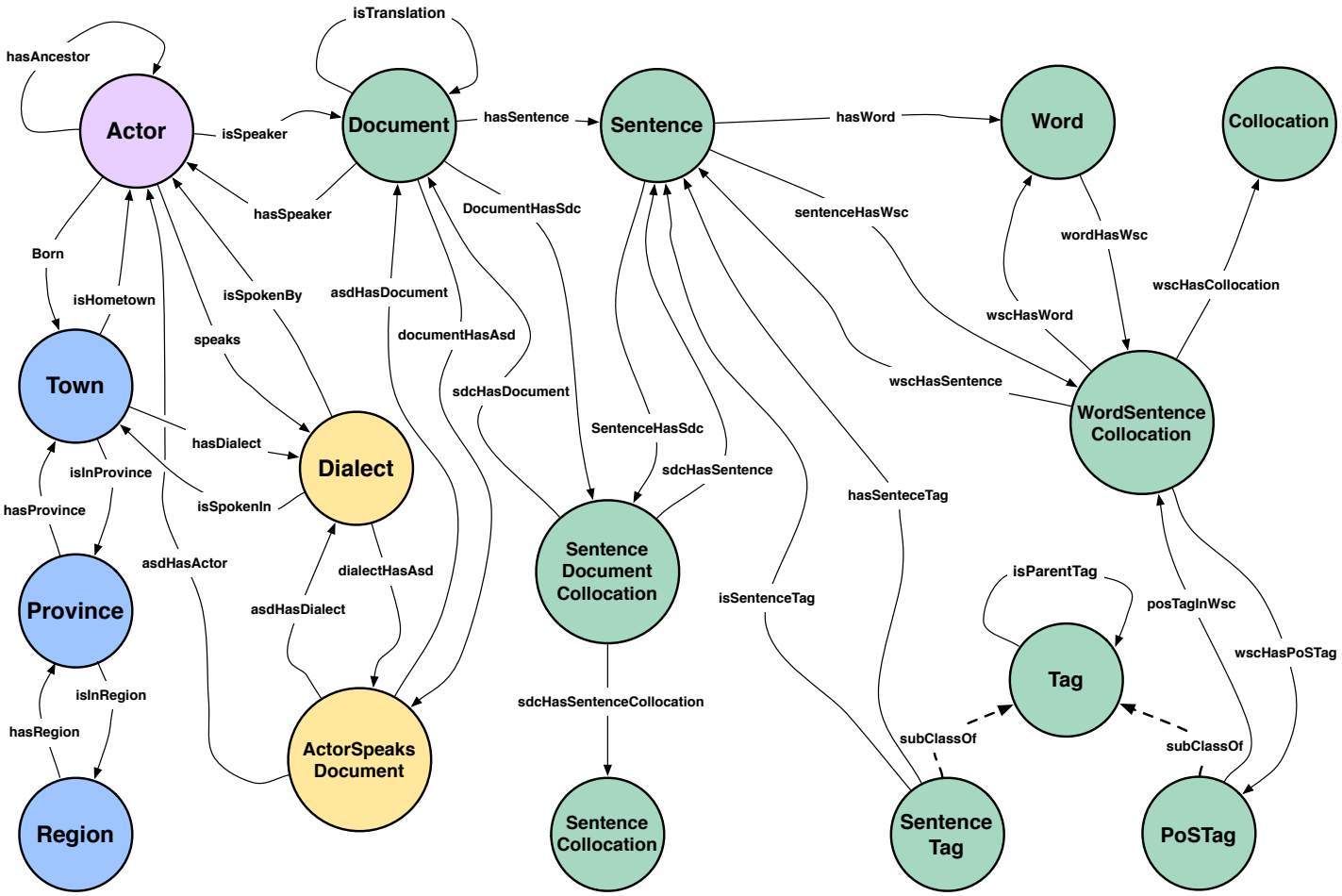
# Application

- http://svrims2.dei.unipd.it:8080/asit-enterprise/

DEPARTMENT OF
INFORMATION
ENGINEERING
UNIVERSITY OF PADOVA

# Discussion

# Discussion


isTranslation
hasAncestor
Actor
isSpeaker
Document
hasSentence
Sentence
hasWord
Word
Collocation
hasSpeaker
DocumentHasSdc
sentenceHasWsc
wordHasWsc
wscHasCollocation
Born
isHometown
isSpokenBy
asdHasDocument
wscHasWord
speaks
documentHasAsd
wscHasSentence
Town
sdcHasDocument
WordSentence Collocation
hasDialect
SentenceHasSdc
Dialect
isInProvince
sdcHasSentence
hasSenteceTag
isSpokenIn
hasProvince
Sentence Document Collocation
asdHasActor
dialectHasAsd
isParentTag
isSentenceTag
posTagInWsc
Province
asdHasDialect
Tag
wscHasPoSTag
isInRegion
sdcHasSentenceCollocation
hasRegion
subClassOf
subClassOf
ActorSpeaks Document
Region
Sentence Collocation
Sentence Tag
PoSTag

DEPARTMENT OF
INFORMATION
ENGINEERING
UNIVERSITY OF PADOVA

# Thank you



# References


- Di Buccio, E., Di Nunzio, G.M., Silvello, G.: A system for exposing linguistic linked open data. In Zaphiris, P., Buchanan, G., Rasmussen, E., Loizides, F., eds.: TPDL. Volume 7489 of Lecture Notes in Computer Science., Springer (2012) 173-178
- Di Buccio, E., Di Nunzio, G.M., Silvello, G.: An open source system architecture for digital geolinguistic linked open data. In Aalberg, T., Papatheodorou, C., Dobreva, M., Tsakonas, G., Farrugia, C.J., eds.: TPDL. Volume 8092 of Lecture Notes in Computer Science., Springer (2013) 438-441
- Di Buccio, E., Di Nunzio, G.M., Silvello, G.: A curated and evolving linguistic linked dataset. Semantic Web 4(3) (2013) 265-270
- Di Buccio, E., Di Nunzio, G.M., Silvello, G.: A linked open data approach for geolinguistics applications. International Journal of Metadata, Semantics and Ontologies , Special Issue on Metadata for e-Science and e-Research. 2013. In press.