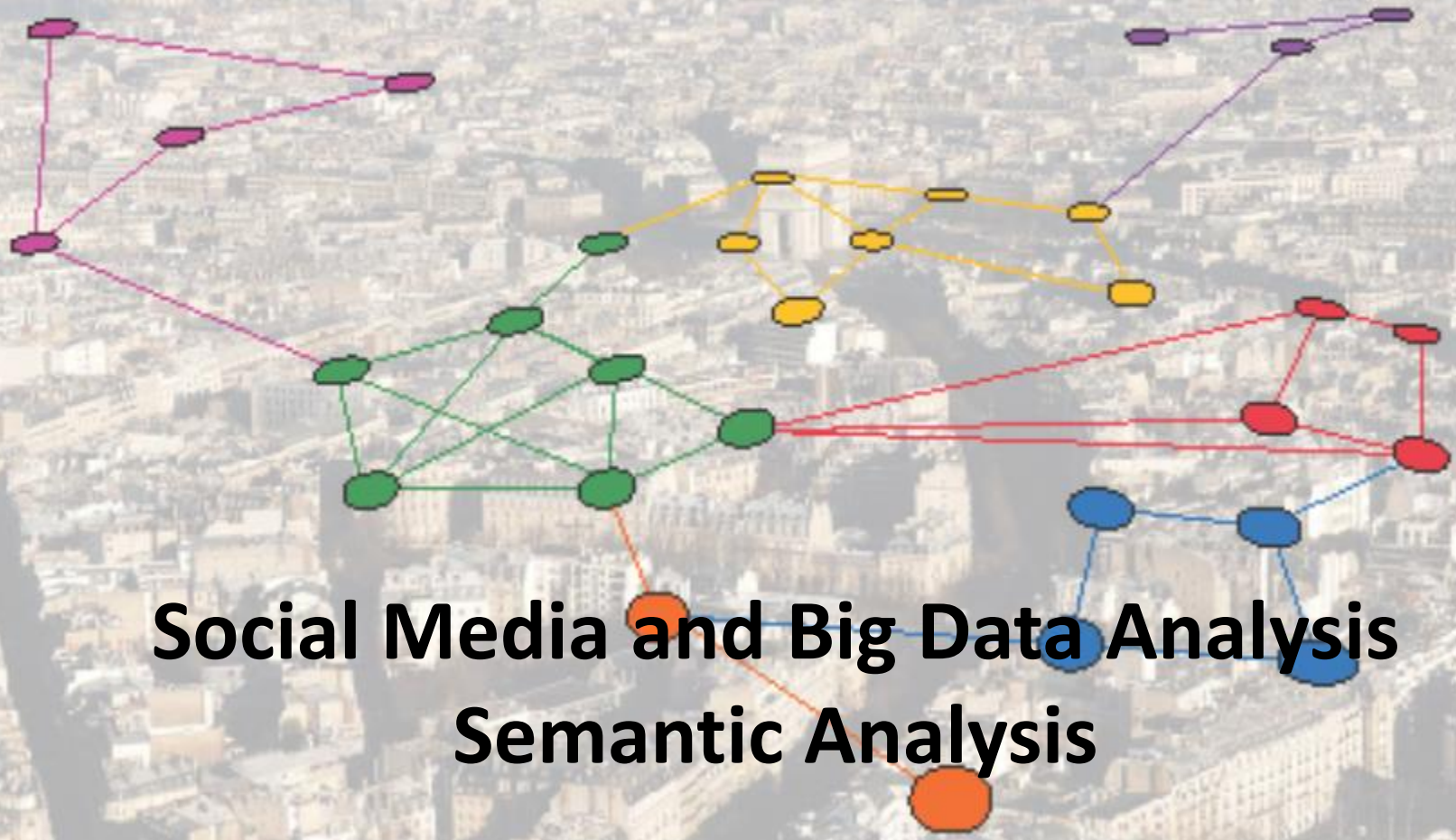
COST TU1305 London meeting 9.11.16



Social Media and Big Data Analysis
Semantic Analysis

Ainhoa Serna, Mondragon Unibertsitatea

Agenda

Data collection

- Data sources: from where? Is it Open? In what form?
- Specific Characteristics of the data (variables)
- How it is collected technically?
- Type of data: text/ geolocation/?
- Quality of data
- Social Media or Social Network?

Data sources

from where? Is it Open? In what form?

1. Social Media
2. Official Statistics (INE - Statistics National Institute, Eustat - Basque Institute of Statistics,..)
3. ODE Open Data Euskadi (open data portal of the Basque Government)
4. Blogs, webs..

- All Official data is OPEN. Social Media data & blogs,.. through API (Application Program interface) or Scraper
- Heterogeneous Format: XML, KML, Excel, String, JSON, API-XML, RSS, ODS..

Data sources

Social Media:

1. Facebook
2. Twitter
3. Digital newspapers
4. TripAdvisor
5. minube

Data sources

ODE Open Data Euskadi

Tourism Dataset in the Basque Country

- Hotels
- Rural accommodation
- Beaches
- Restaurants, steakhouses and cider houses
- Accommodation
- Cultural heritage: religious buildings, castles and points of interest
- Travel agencies
- Health and wellness Tourism: thalasso, spas and resorts

Data sources

ODE Open Data Euskadi

Traffic Dataset

- Traffic events in real time in Basque Country
- History traffic incidents that occurred in the Basque Country since 2006

Transport Mode Dataset

ODE Open Data Euskadi

Recursos de transporte y movilidad de Euskadi

Acceder a los datos en formato: API RSS KML XLS (266 KB) XML (387.12 KB) JSON (306.55 KB)

Datos de Recursos Turísticos *28/09/2016*

Compañías de transporte de Euskadi

Acceder a los datos en formato: API RSS KML XLS (179.5 KB) XML (262.79 KB) JSON (208.09 KB)

Datos de Recursos Turísticos *23/09/2016*

Estaciones de transporte de Euskadi

Acceder a los datos en formato: API RSS KML XLS (32.5 KB) XML (40.55 KB) JSON (32.09 KB)

Datos de Recursos Turísticos *19/09/2016*

Base Topográfica Armonizada a escala 1:5.000 de Gobierno Vasco. BTA

Acceder a los datos en formato: SHP WMS PDF

Datos Geográficos *31/12/2015*

Base Topográfica Armonizada a escala 1:400.000 de Gobierno Vasco.

Acceder a los datos en formato: SHP WMS

Specific Characteristics of the data (variables)

- language, date, timestamp, transport mode, country, city, geoloc(lat-lon), transport condition, punctuality, price, route, road, service, polarity analysis (positivity-negativity-neutral), …etc.

How it is collected technically?

- Ad-hoc software through API's and Crawling+ scraping techniques.

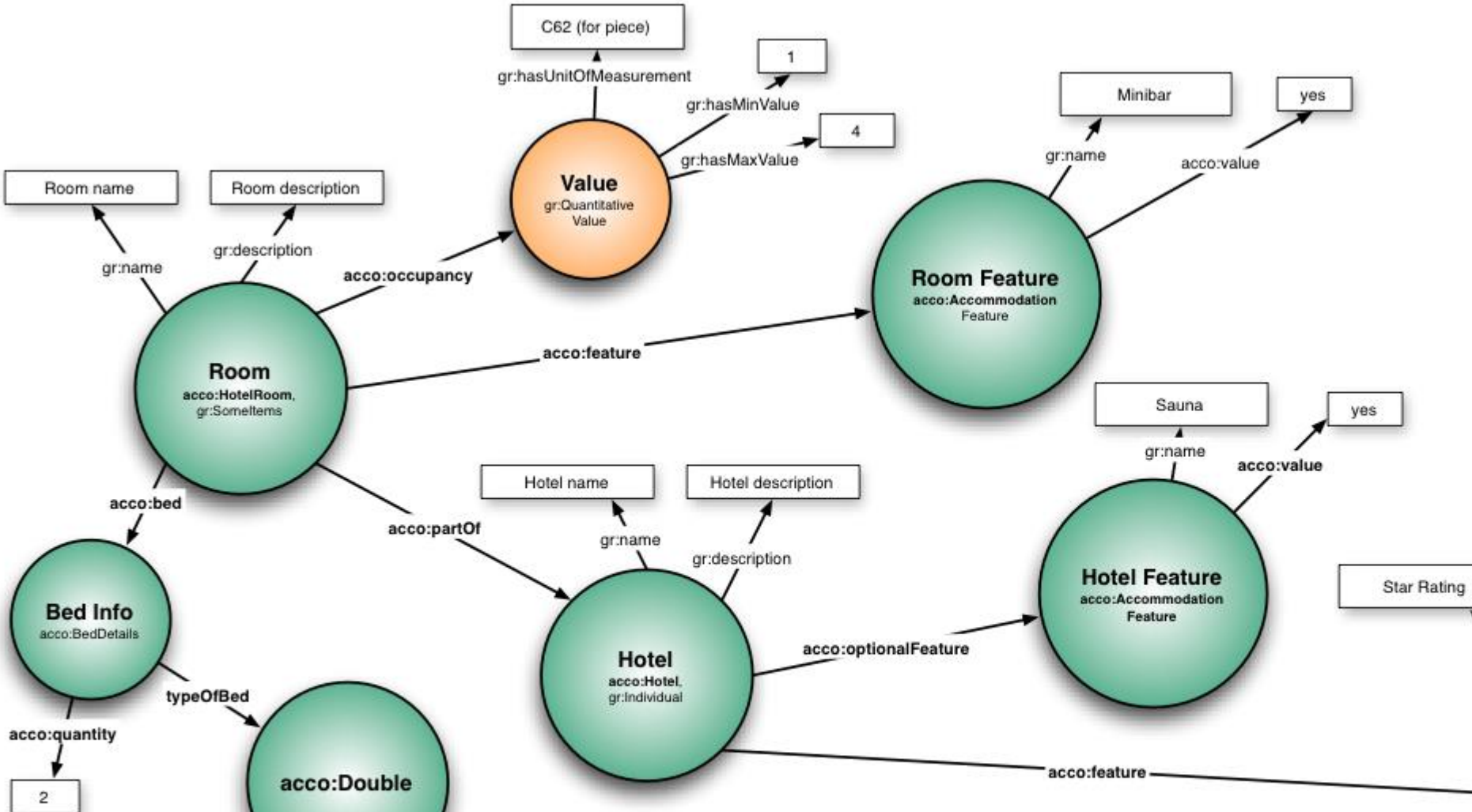
six main phases:

1. semantic discovery (source identification),
2. data acquisition,
3. data preparation for analysis,
4. data curation,
5. data storage
6. and data visualization

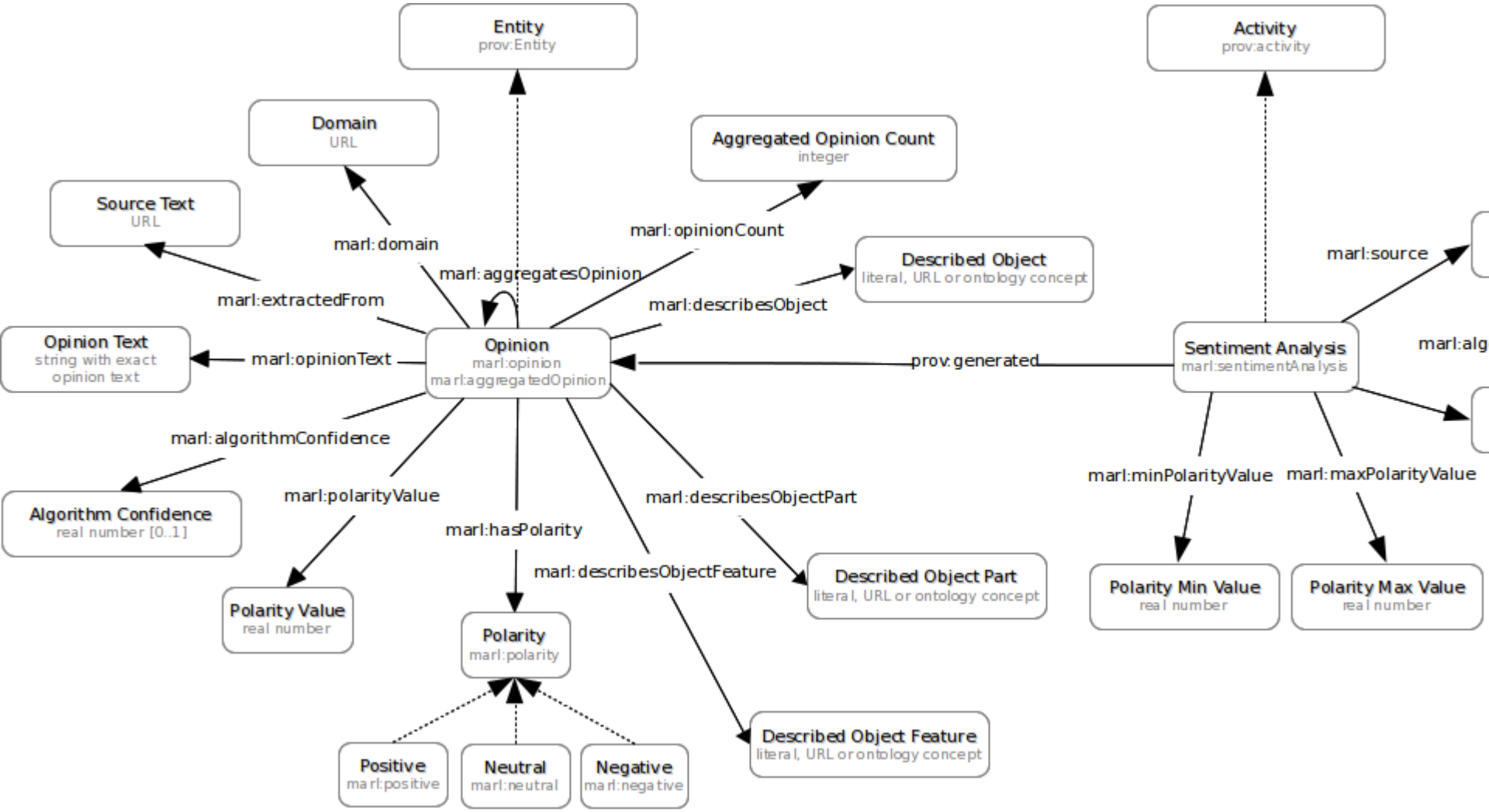
Data Curation

- The categorization is aided by a text-mining tool developed by Mondragon Unibertsitatea based on Freeling and the Wordnet lexical database.
  - two upper ontologies - WordNet and SUMO
  - two tourism ontologies: ACCO, QALL-ME
  - to add the feeling of the reviews, two ontologies were mapped: Marl (subjective opinions domain) and SIOC (Semantically-Interlinked Online Communities).

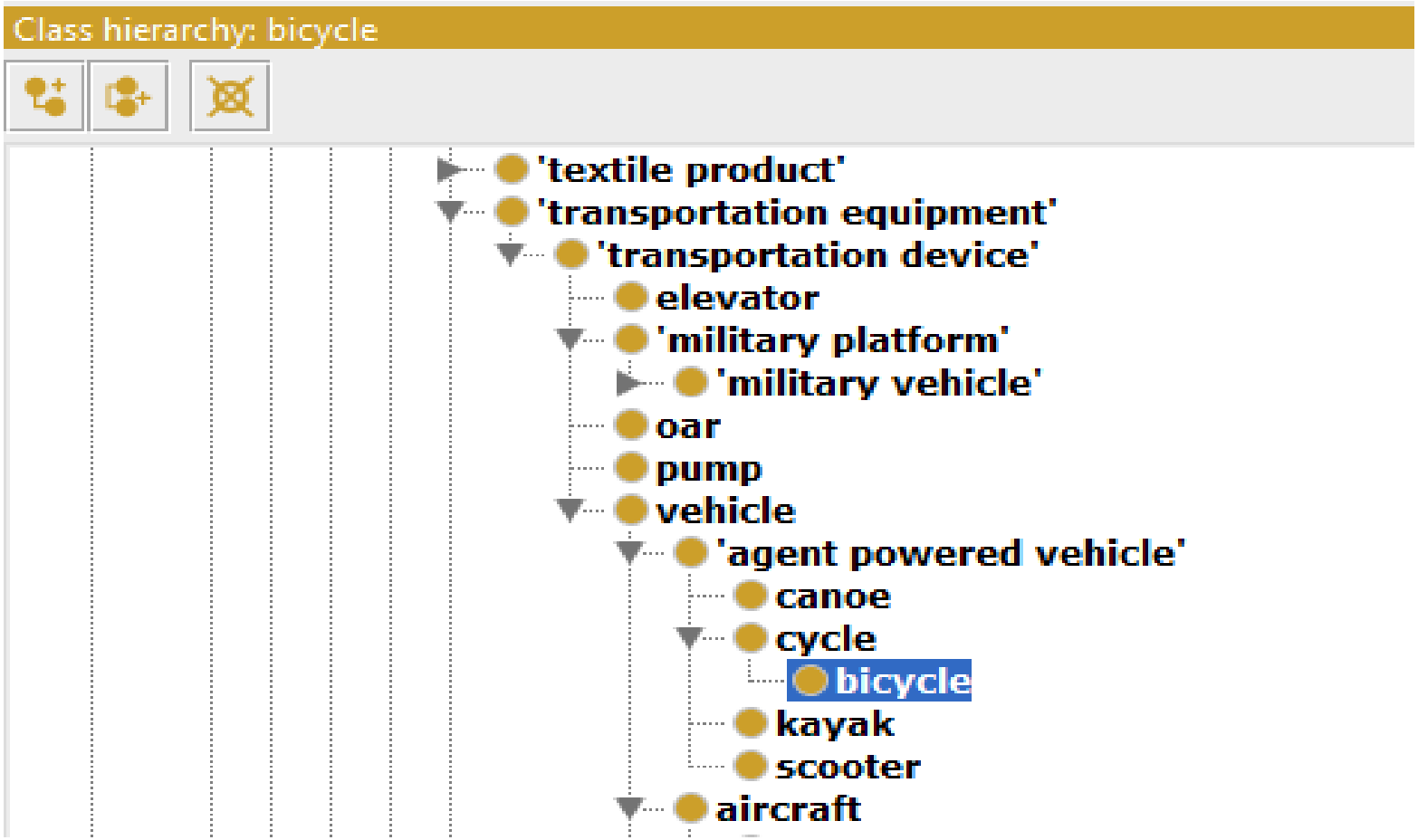
ACCO ontology



MARL ontology



SUMO ontology

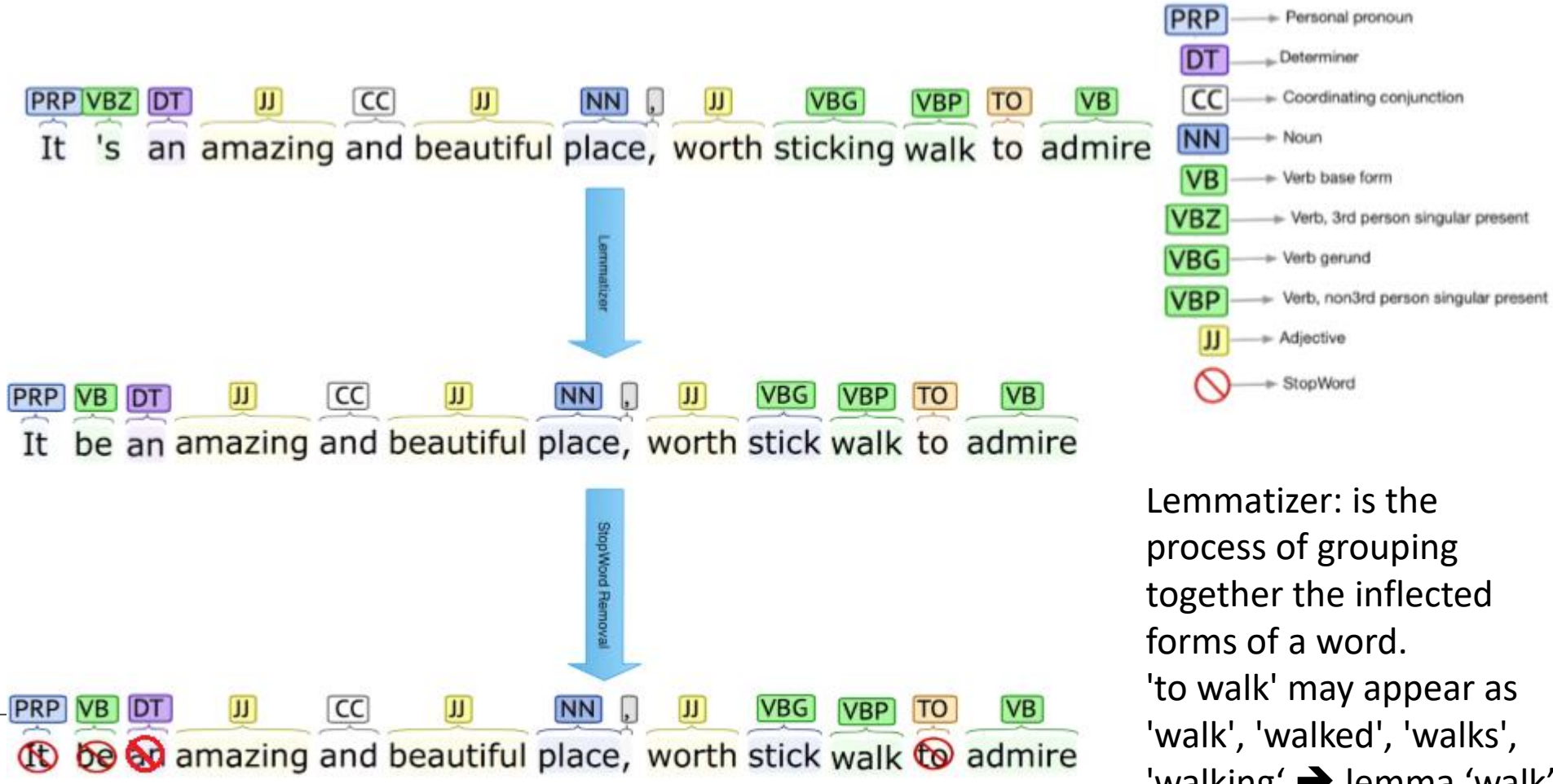


Example

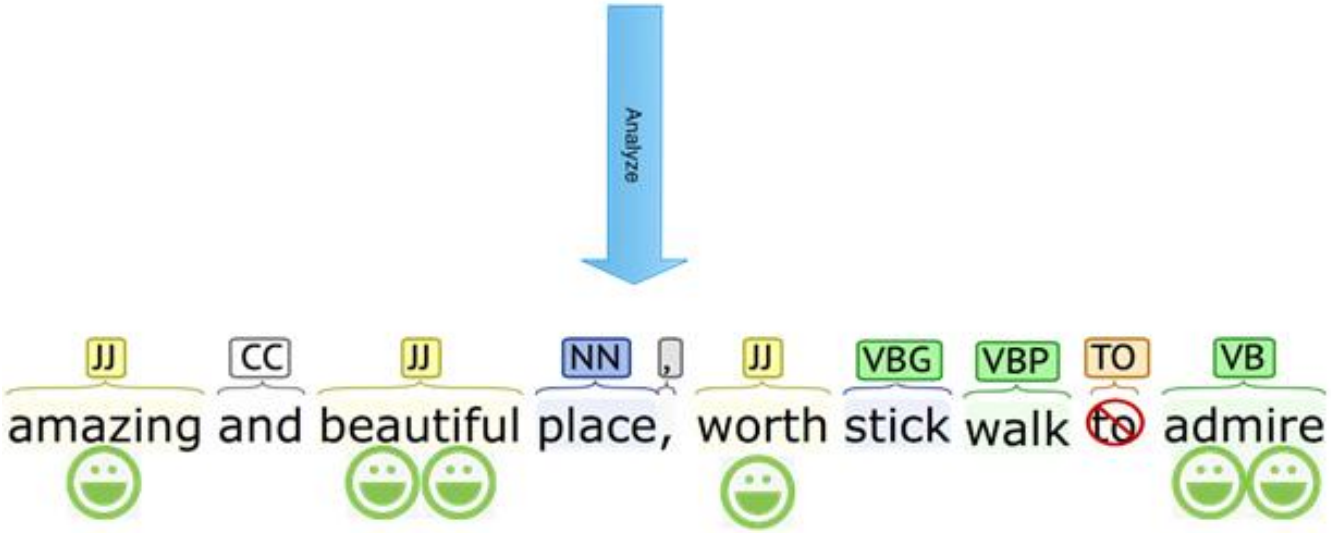
Opinion #13: 2013-12-02 21:50:24.0

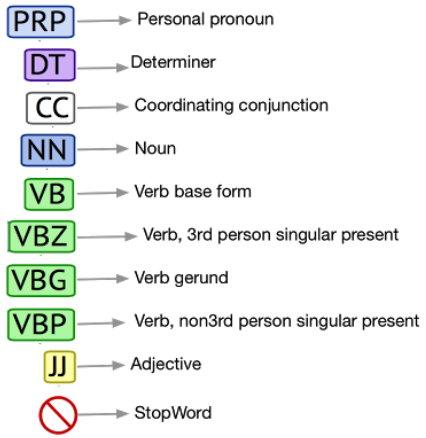
An amazing place

It's an amazing and beautiful place, worth sticking walk to admire !!



Lemmatizer: is the process of grouping together the inflected forms of a word.
'to walk' may appear as 'walk', 'walked', 'walks', 'walking' ➔ lemma 'walk'





TTW: Trip conditions: Time and Weather

F&B: Food and Beverage

SOC: Social Determinants

SFT: Safety

REC: Natural and cultural resources

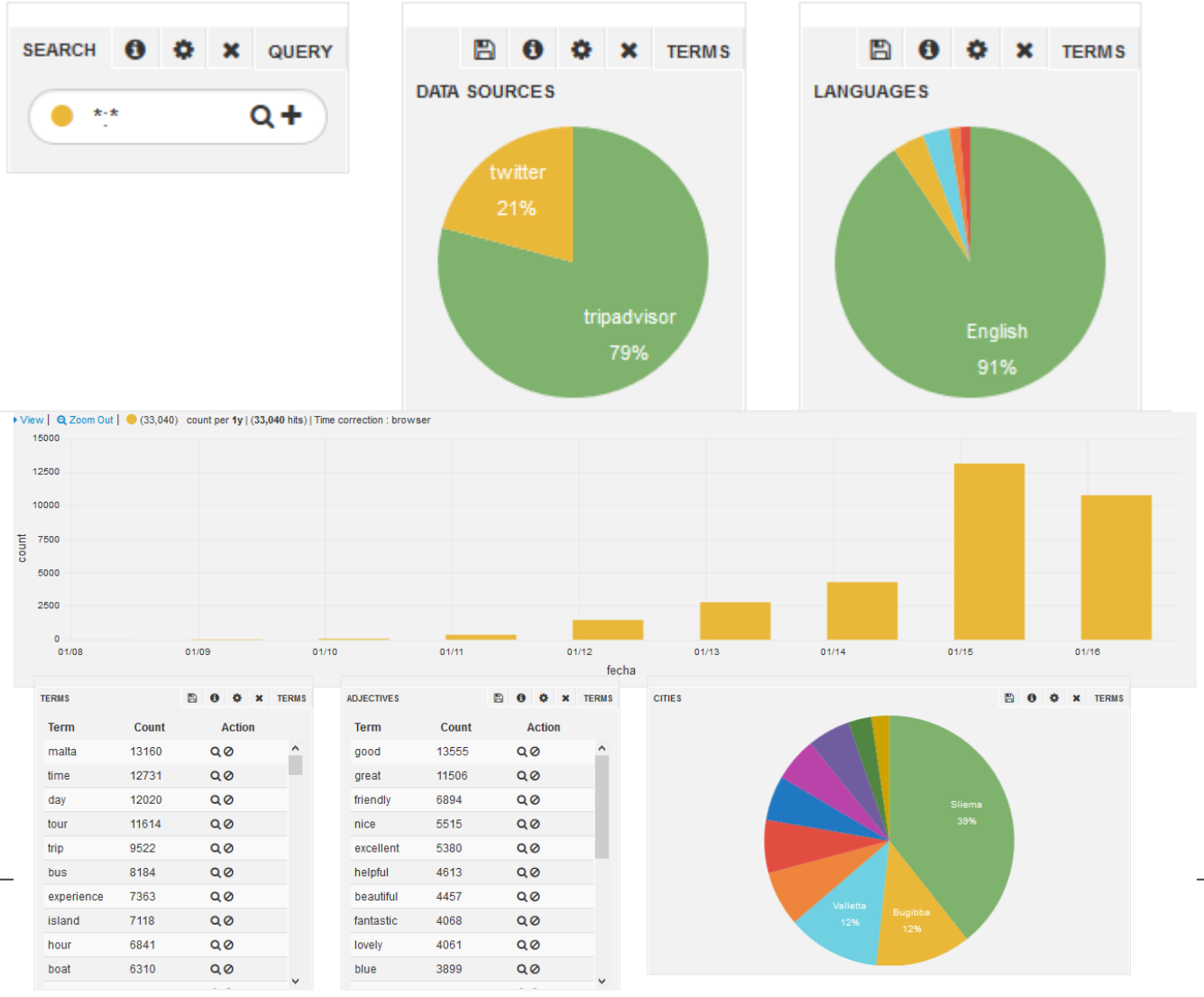
STO: Suggestion to other

PFM: Performance

INF: Infrastructure and socio-economic environment

ATM: Atmosphere

Visualization



Type of data: text/ geolocation/?

- Text,
- Geolocation

Example:

- *Content: Very nice place, comfortable, cozy room, clean and .....*
- Geolocation: *35.921249,14.489501*

Quality of data

- Twitter: poor information, difficult to find something of value. Identify Twitter channels whose theme is the interest for analysis. For instance: @Maltese_Traffic; https://twitter.com/Maltese_Traffic
- TripAdvisor: Excellent (rich data and a large sample, many languages)
- Minube: Excellent (rich data but only spanish language)
- Official statistics: one year later and small sample

Social Media or Social Network?

- **Social media** are computer-mediated technologies that allow individuals, companies, NGOs, governments, and other organizations to view, create and share information, ideas, career interests, and other forms of expression via virtual communities and networks.
- social media facilitate the development of online social networks by connecting a user's profile with those of other individuals and/or groups.

- A **social network** is a social structure made up of a set of social actors (such as individuals or organizations), sets of dyadic ties, and other social interactions between actors.

Social Media or Social Network?

- Social media is a technology that allows the creation of relationships among individual by means of social networks

For its target audience and theme

- Horizontal social networks are those aimed at all user **without a defined topic**: **Facebook**, **Twitter**, Pinterest,…
- Vertical social networks: They are designed on the basis of a **defined topic**. Its aim is to gather around a defined subject to a specific group.
  - Professional Vertical social networks: They are aimed at generating professional relationships among users. The most representative examples such as Linked In.
  - Leisure Vertical social networks: the goal is to bring together groups that develop leisure, sport, videogame players, fans, etc. The most representative examples are **TripAdvisor**, Wipley, **Minube**, Dogster, Last.FM and Moterus.

Cross questions (40 min)

Data analysis (5min each*5=25 min)

- Tools used for analysis
- Analysis potential: advantages of the technique
- Analysis limits & boundaries: Problems
- Compare to traditional survey?
- Relevance for urban / transportation analysis and models
- Description ability?
- Ability to forecast?
- Ability to use in models?

Tools used for analysis

–Open Source Analyzer

–Open Source language detector

– A spell checker (the texts are corrected using a spell corrector)

–Upper level (SUMO,..) and domain level ontologies (QALLME,..)and MARL sentiment analysis ontology and Lexical resource Database WordNet and SentiWordNet,…etc.

–Ad-hoc software for Natural Language Processing adapted to the research domain

–An enterprise search engine

- A Visual Analytics tool

Analysis potential : advantages of the technique

- Automatic Semantic classification
- Semantic Discovery, to allow the discover of new Data Sources
- Speed, big data processing in a short time, which is crucial for certain social networks.
- Automatic Visualizations, it facilitates the understanding of the results
- Customizable dashboards
- Knowledge inference capabilities
- Automatic Sentiment analysis for many languages
- Usability and dashboard visualization

Analysis limits & boundaries: Problems

- API's limitations:
  - for instance 2500 tweets maximum
  - API's update or deprecated for instance Facebook August 7 changed its API and FQL requests no longer supported. API dependent code needs to be fixed.

- Limitations about some Languages for the sentiment analysis:
  - Language identification
  - Language Analyzer available

- Accuracy of the automatic analysis (NLP hot topic in research area, not 100% accuracy)
- Free text, more difficult to categorize the data (freedom text, bad syntax, using abbreviations, emoticons,…)

Compare to traditional survey?

- Accuracy of the automatic analysis
- Larger sample data
- Richer data
- Speed
- Automatic sentiment analysis (what, how,…positivity, negativity,..
- Automatic reports
- Automatic summaries
- Automatic statistics

- Social media it is no representative of the whole population nowadays
- Lack of sociodemographic data in many cases

Relevance for urban / transportation analysis and models

- **The number of topics related to mobility is relevant on Traveller social networks** (TripAdvisor,..) since social media is a great tool for communication and meeting point. The analysis of that type of information can be very important for travel behaviour analysis.

- Information and Communication Technologies (ICT) **offer the opportunity to improve traditional survey methods to collect travel behaviour data**, decreasing bias in the data, reducing respondent burden, and increasing data quality.

- this approach enriches the data of the traditional surveys, extends traditional analysis with Big-Data methods, using Semantic with data mining algorithms and Natural Language Processing techniques to extract urban mobility information from Social Media data

Relevance for urban / transportation analysis and models

Predictions based on machine learning techniques are possible because there are historical and behavioral patterns.

Explanatory models of mobility in general seek to find factors that influence, and they can provide insights to implement transport policies.

What are the gaps? Which complementary data you need and/or can have from regular survey?

- Mainly demographics data: gender, age, region/province where live,
- Personal information about:
  - level of education, occupation,
  - Living arrangements: on your own, with children, with parents,...
  - newspapers regularly read, magazines,..
  - the las two holidays outside their country
  - if access websites/apps/social media on a regular basis
  - the main motivations for visiting
  - about the trip booking
  - when they decide to make the trip
  - what is the first airport of departure
  - aeroline used,..