Part 2
State of the Art

1 Part 0 - Course Organization
2 Part 1 - The Reuse Problem
3 Part 2 - State of the Art
4 Part 3 - Knowledge Graphs
5 Part 4 - Entity Base
6 Part 5 - The iTelos Methodology
7 Part 6 - KG Evaluation and Exploitation

Part 2.1
Type of information

1 Types of source
2 Types of language
3 Types of data schema
4 Types of data values
5 What does data represent?

Existing knowledge representation

- The stratified view of the information is not a novelty.
- The need to represent valuable knowledge has led, in the past, to several studies on:
  - the data sources and their capacity to provide information;
  - the language used by the data;
  - the modelling of the information carried by the data (data schema, ontologies);
  - the shape of the data and its values (data types and formats).
- The data reuse problem, rooted into the representation of the information, can be addressed only by understanding what type of information we can reuse, nowadays.

Types of source

- There are many different types of sources providing information which can be turned into data to be re-used.
- However, different source provide different types of information or data.
  - Sometimes the data is expressed with a clear and formal language (data format):
    - **Structured data**: CSV, Excel, JSON, XML, databases
  - Sometimes the "language" used by the source is not easily processable, thus making it complicated to transform its information into data:
    - **Textual data**: PDF, TXT
  - Sometimes the source does not use a formal language, making it very difficult to transform its information into data:
    - **Media data**: images, videos, sounds

Types of source

- The key point is that,
- we can distinguish between **data source** and **information source**:
  - **Data source**: a source that provides processable data represented with a formal language (data format).
  - **Information source**: a source that provides data hard to be processed, not described with a formal language (media data).

Types of source

- In both the two cases, described above, the information provided by the source (Producer) needs to be transformed into data that fit the Consumer's purpose.
  - **Data source case**: more easy processable information, low information loss.
  - **Information source case**: hard-to-process information cause high information loss, due to the consumer perception and interpretation.

- **For this reason the problem of choosing the right source, and to deal with the information it is able to provide, is not so trivial!**

Part 2.2
Type of language

1 Types of source
2 Types of language
3 Types of data schema
4 Types of data values
5 What does data represent?

Language and Concepts (1)

- Before to see which are the current available language resources, we need to understand what is a language and how it is represented as a digital data.
- We can start to define a **language dataset as a set of concept representations**. Those concepts used to give a **meaning** to the data we need to (re)use.

Language and Concepts (2)

- For example:
  - To properly (re)use the dataset in the figure below, we need to understand **the meaning expressed by the words**: "department" and "type".
  - Meanings which are uniquely expressed by **concepts**.
    - "department": a university facility working on education, research and innovation activities.
    - "type": the type of the department specifies the work area, like computer science (INF) or biology (BIO).

Departments			
ID	**name**	**address**	**type**
1	DISI	via x 3	INF
2	Sociology	via y 2	SOC
3	economy	via z 6	ECO

Concepts Representation (1)

- A concept is represented by:
  - a **word**: a term or a composition of terms, which shape the concept (i.e., "Car", "Person", "Electric engine");
  - a **gloss**: a textual description (plus examples) of the concept that helps the reader to disambiguate its meaning.
- A concept can be linked (semantically related to) other concepts in two ways:
  - **Hyponymy**: is a relation to a more generic concept (i.e., "Woman" is a more general concept than "Daughter").
  - **Hypernym**: is a relation to a more specific concept (i.e., "Artificial lake" is a more specific concept than "Lake").

Concepts Representation (2)

- Based on the above, the following definitions have to be considered.
  - **Synonymy:** A word with the same (or nearly the same) meaning (i.e., *sense*) as another word. e.g., car, auto, automobile, etc.
  - **Polysemy:** The coexistence of many possible meanings for a word. e.g., 645 distinct meanings of the word *run*.
  - **Synset:** A synset is a set of synonyms that are, in principle, interchangeable for a particular sense of a word. e.g., {car, auto, automobile, motorcar}.
  - **Subsumption:** A classification of concepts from the general (i.e., *hypernym*) to the specific (i.e., *hyponym*) via `IS-A` relation. e.g., spoon `IS-A` cutlery.
  - **Lexical Gap:** The absence of a word in a particular language where it is present in another. e.g., *Malga* in Italian absent in English.

Language Representation

- A language data is, as a consequence, a set of concepts defined by their words, synonyms and glosses, linked together by semantic relationships.
- Here below some examples of language data sources.

Language data sources (1)

- We can distinguish between two types of language data sources:
  - online dictionaries: list of terms
  - Wordnet like resources: network of concepts.
  - The wordnet-like resources are more useful for data integration and data reuse purposes, due their structure connecting words, concepts, and meaning.

Example

- Global Wordnet Association
- WordNet
- Open Multilingual WordNet
- Datascientia/UKC (forthcoming)

Global WordNet Association

Figure: Global WordNet Association[1]

[1] http://globalwordnet.org/

WordNet Home

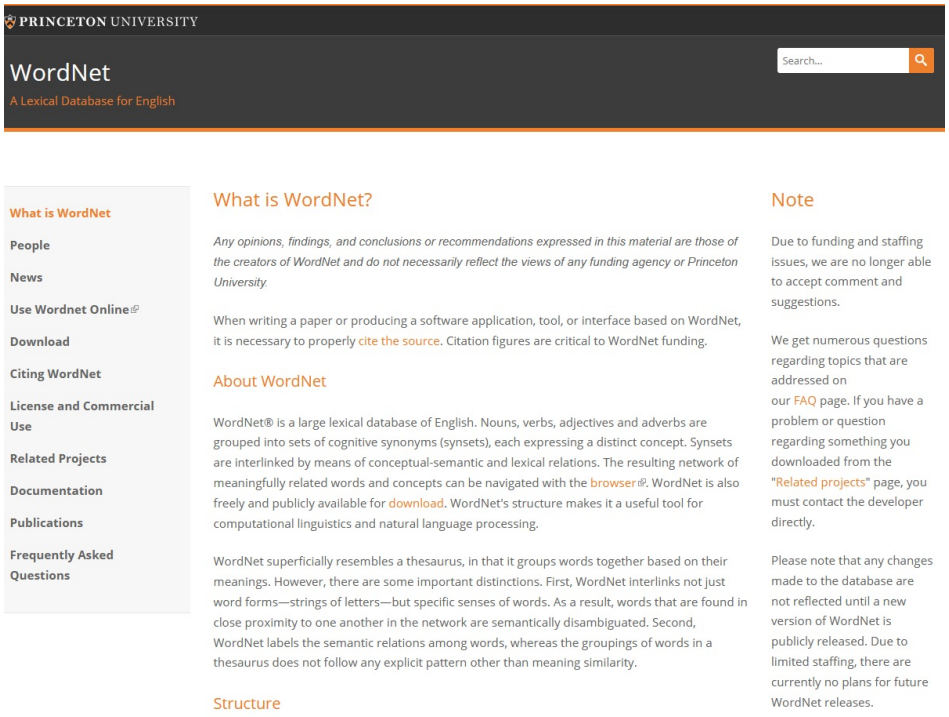
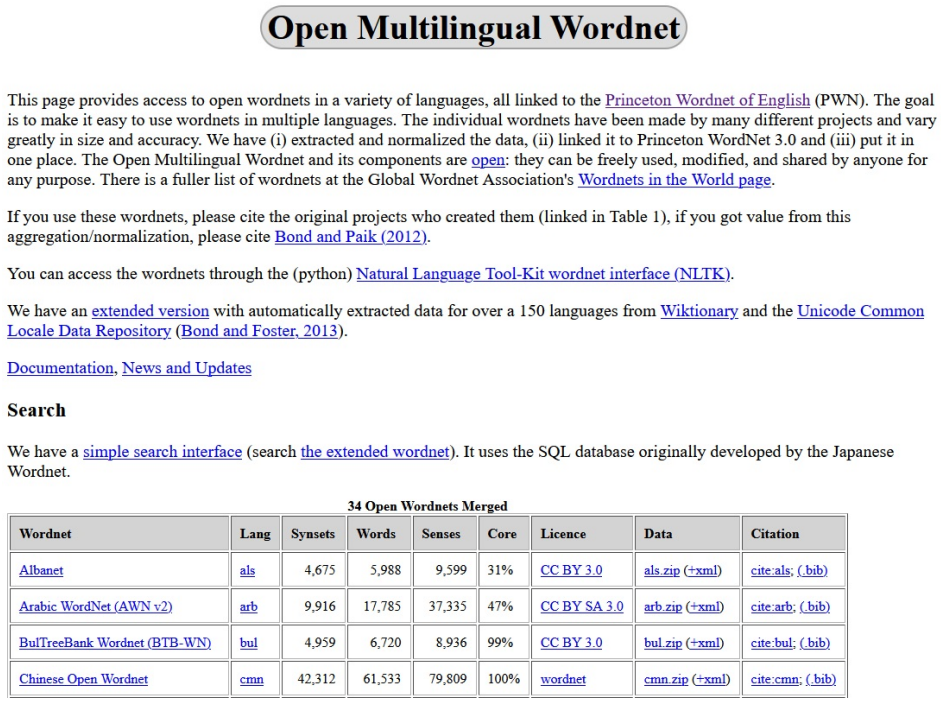


Figure: WordNet Home[2]

[2] https://wordnet.princeton.edu/

Multilingual WordNet



Open Multilingual Wordnet

This page provides access to open wordnets in a variety of languages, all linked to the Princeton Wordnet of English (PWN). The goal is to make it easy to use wordnets in multiple languages. The individual wordnets have been made by many different projects and vary greatly in size and accuracy. We have (i) extracted and normalized the data, (ii) linked it to Princeton WordNet 3.0 and (iii) put it in one place. The Open Multilingual Wordnet and its components are open: they can be freely used, modified, and shared by anyone for any purpose. There is a fuller list of wordnets at the Global Wordnet Association's Wordnets in the World page.

If you use these wordnets, please cite the original projects who created them (linked in Table 1), if you got value from this aggregation/normalization, please cite Bond and Paik (2012).

You can access the wordnets through the (python) Natural Language Tool-Kit wordnet interface (NLTK).

We have an extended version with automatically extracted data for over a 150 languages from Wiktionary and the Unicode Common Locale Data Repository (Bond and Foster, 2013).

Documentation, News and Updates

Search

We have a simple search interface (search the extended wordnet). It uses the SQL database originally developed by the Japanese Wordnet.

34 Open Wordnets Merged

Wordnet	Lang	Synsets	Words	Senses	Core	Licence	Data	Citation
Albanet	als	4,675	5,988	9,599	31%	CC BY 3.0	als.zip (+xml)	cite:als; (.bib)
Arabic WordNet (AWN v2)	arb	9,916	17,785	37,335	47%	CC BY SA 3.0	arb.zip (+xml)	cite:arb; (.bib)
BulTreeBank Wordnet (BTB-WN)	bul	4,959	6,720	8,936	99%	CC BY 3.0	bul.zip (+xml)	cite:bul; (.bib)
Chinese Open Wordnet	cmn	42,312	61,533	79,809	100%	wordnet	cmn.zip (+xml)	cite:cmn; (.bib)

Figure: Open Multilingual WordNet Home[3]

[3] http://compling.hss.ntu.edu.sg/omw/

Universal Knowledge Core (UKC)

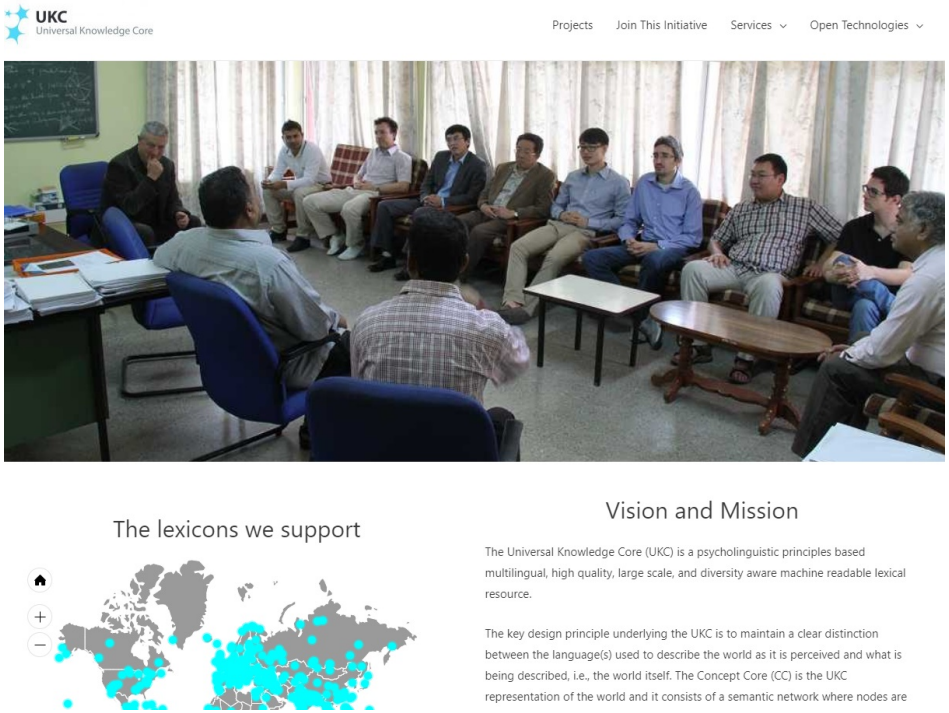


Figure: UKC Home[4]

[4] http://ukc.disi.unitn.it

Part 2.3
Type of data schema

1 Types of source
2 Types of language
3 Types of data schema
4 Types of data values
5 What does data represent?

Data schema and Ontology

- A data schema (or dataset schema) is the **model of the information** carried by a data (or a dataset). It can be extracted by a data (dataset) or defined separately in order to represent the relative (existing or future) data (ontology).

- Nowadays, many different models are available.

  - **Single Etypes**: representing a single entity types (ETypes) with its properties
    - "City" := <"name", "region", "province", "population", "area">

  - **Dataset schema**: representing the entity types (ETypes) and the relative attributes, considered for a specific dataset.

  - **Application specific ontologies**: ontological models indicating how to represent the data for specific application.

  - **Domain specific ontologies**: ontological models indicating how to represent the data for specific domain of interest.

  - **Top-level ontologies**: general ontological model used for large scale domain of interest.

Ontology repositories

- The different types of data schema and ontologies, led to a variety of ontology and schema providers, for different domain and applications.
- In the following slides, there are some examples of high quality ontology repositories.

Linked Open Vocabulary (LOV)

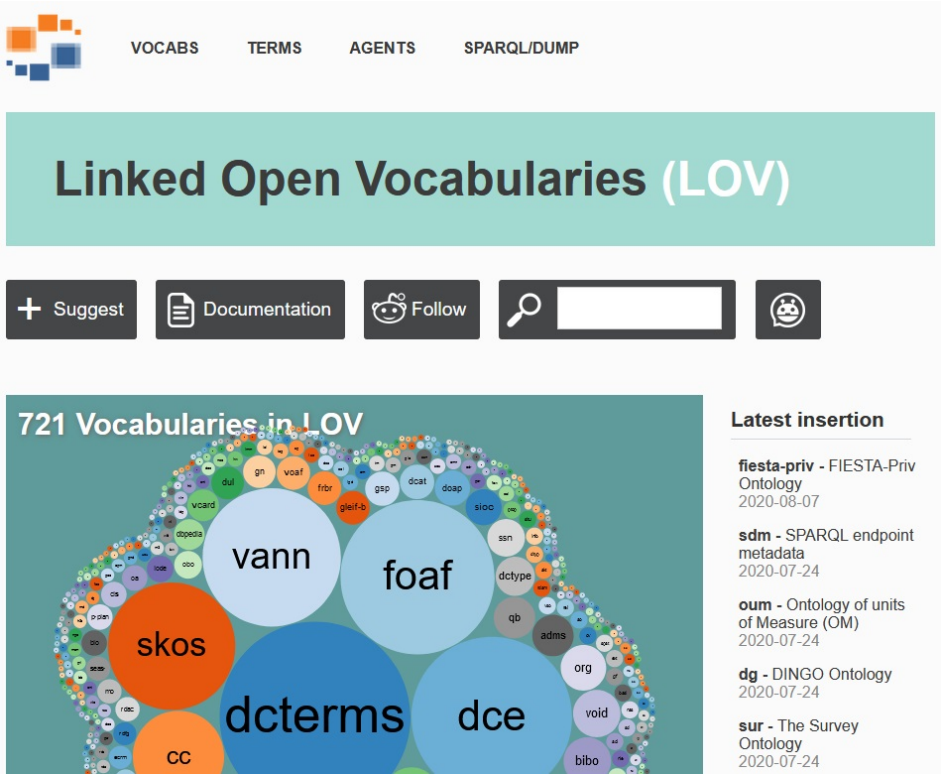


Figure: Linked Open Vocabulary[5]

[5] https://lov.linkeddata.es/dataset/lov/

Schema.org

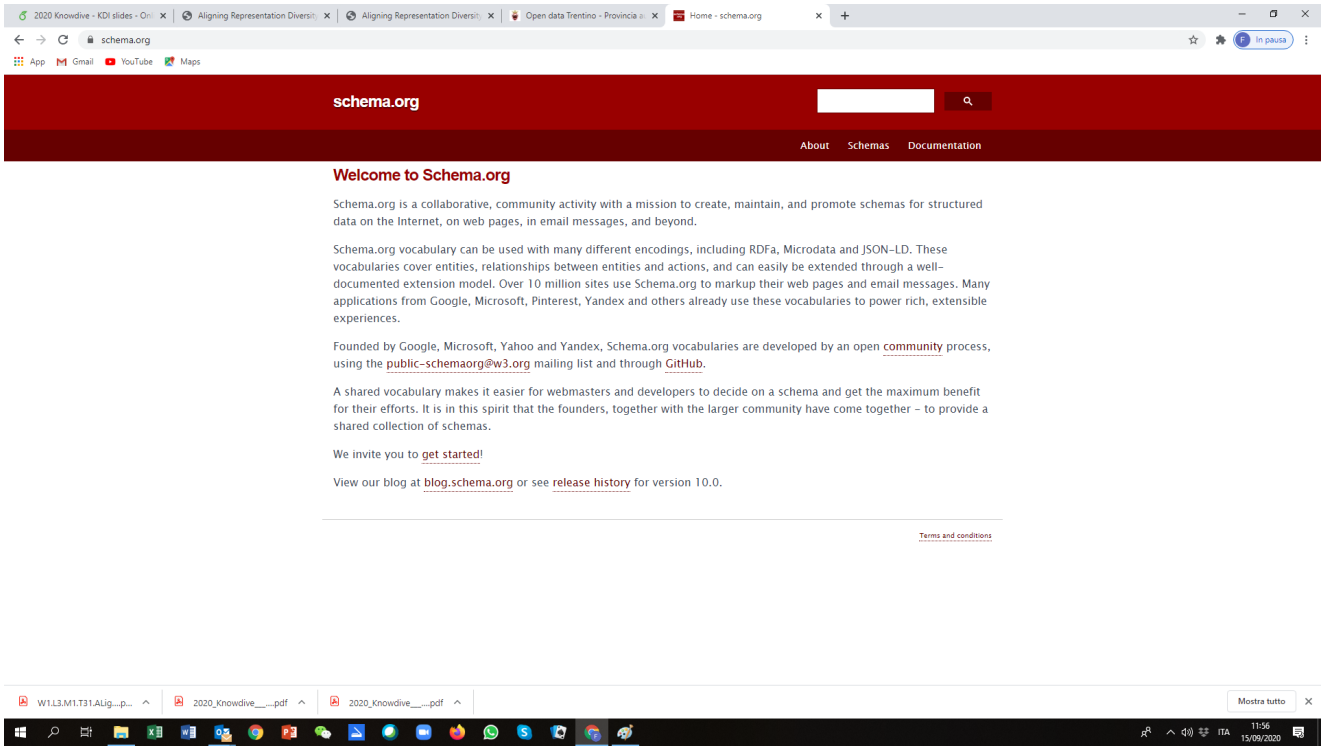


Figure: Schema.org[6]

[6] http://www.schema.org/

DBpedia

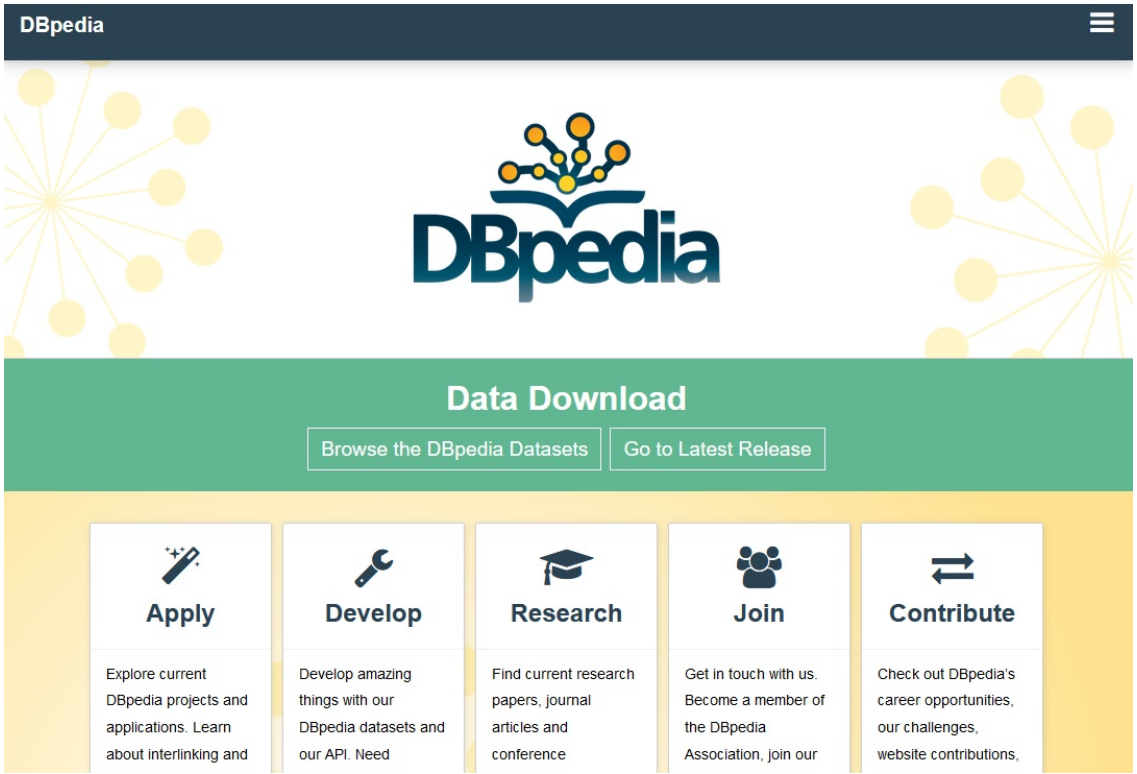


Figure: DBpedia Home[7]

[7] https://wiki.dbpedia.org/

DataScientia - LiveKnowledge

Figure: LiveKnowledge catalog[8]

[8] https://datascientiafoundation.github.io/LiveKnowledge/

Part 2.4
Type of data values

1 Types of source
2 Types of language
3 Types of data schema
4 Types of data values
5 What does data represent?

Data values and Dataset types

- The last layer considered by the stratification of the information, is the one that represents **the value of a data**.
- The resources defined by this layer are the **more common type of dataset**, those that are most used to carry information.
  - tabular datasets, JSON datasets, XML datasets, etc ...
- It is easy to understand how many types of such datasets exists nowadays!

Data catalogs

- Being **the reuse of data** the main problem addressed by this course,
- a key role, regarding the data value layer resources, is played by the **data catalogs**.
- A data catalog is a web portal (or web-app) designed and implemented for the **distribution of data**.
  - A web-base access point **for different data repositories**;
  - providing smart data **search** and user friendly **navigation**.
- **IMPORTANT**: a data catalog manage only the **metadata**, used to describe the datasets that are, instead maintained in a **separate** repository.
  - In other words a catalog, describes and provides the access to datasets that are administrated separately.

Open Data Hub

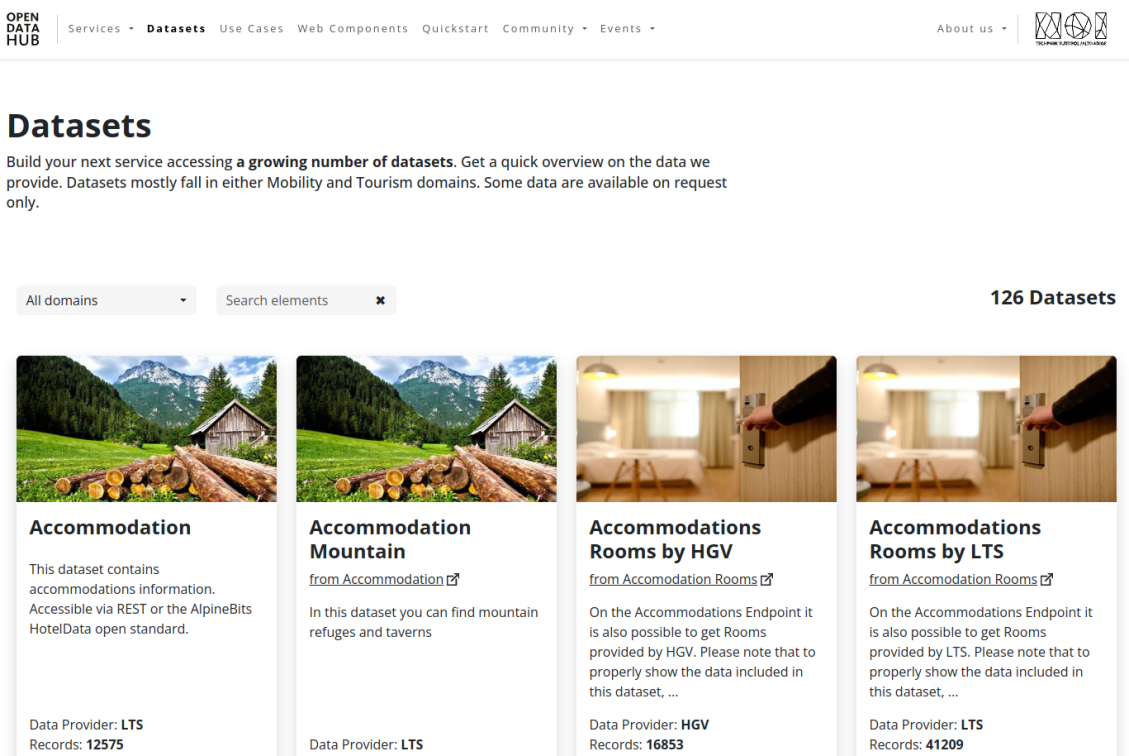


Figure: Open Data Hub[9]

[9] https://opendatahub.com/datasets/

Open Data Trentino

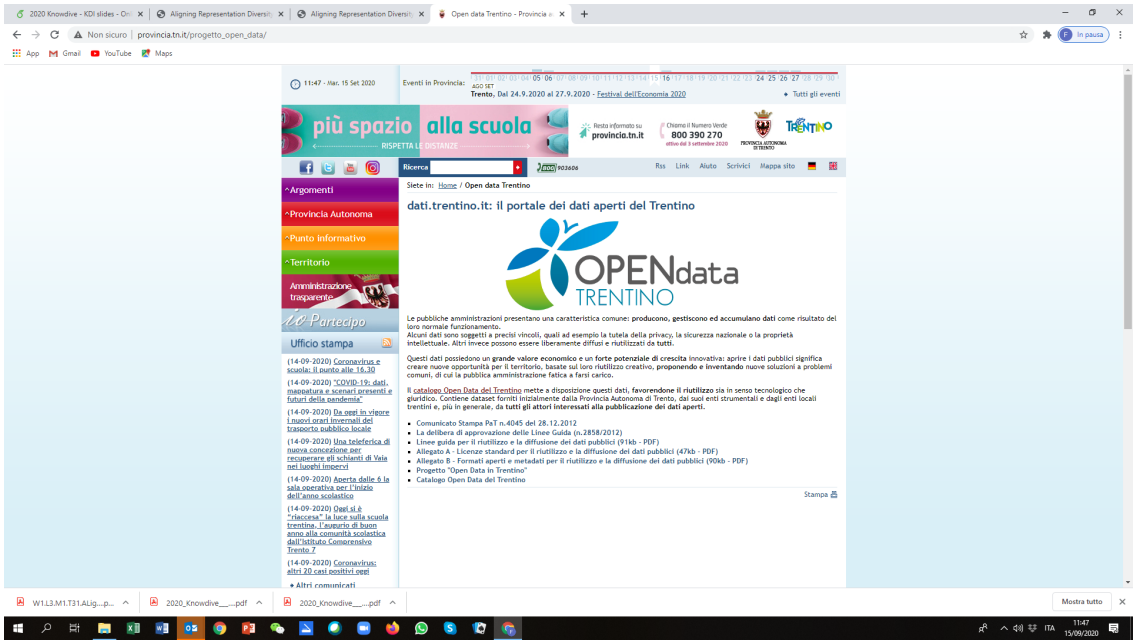


Figure: Open Data Trentino[10]

[10] http://dati.trentino.it/

OpenStreetMap

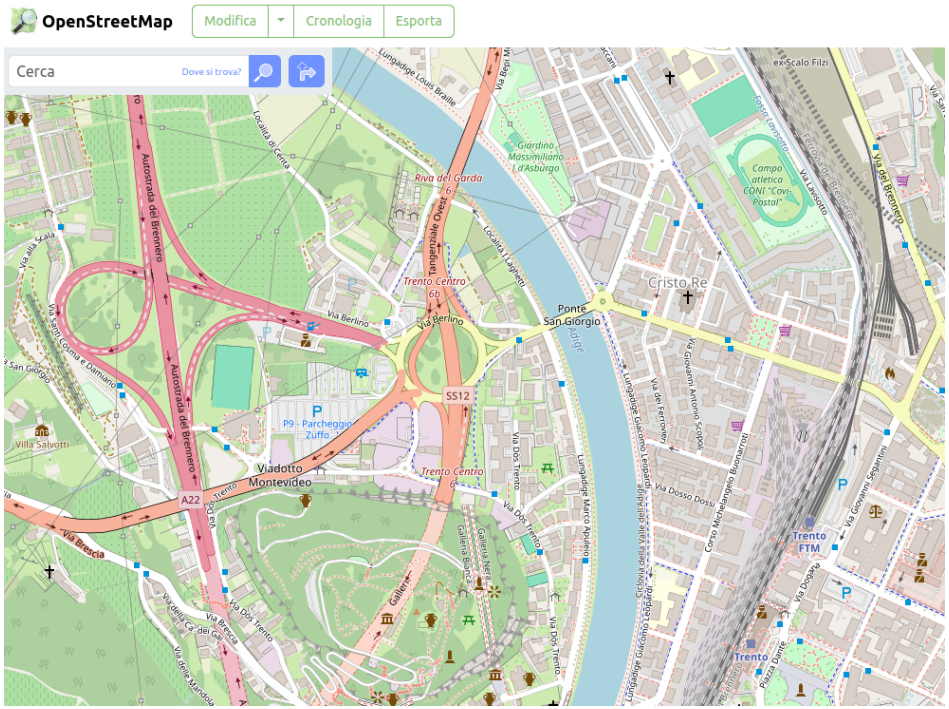


Figure: OpenStreetMap[11]

[11] https://www.openstreetmap.org/#map=16/46.0751/11.1204

DataScientia - LiveData

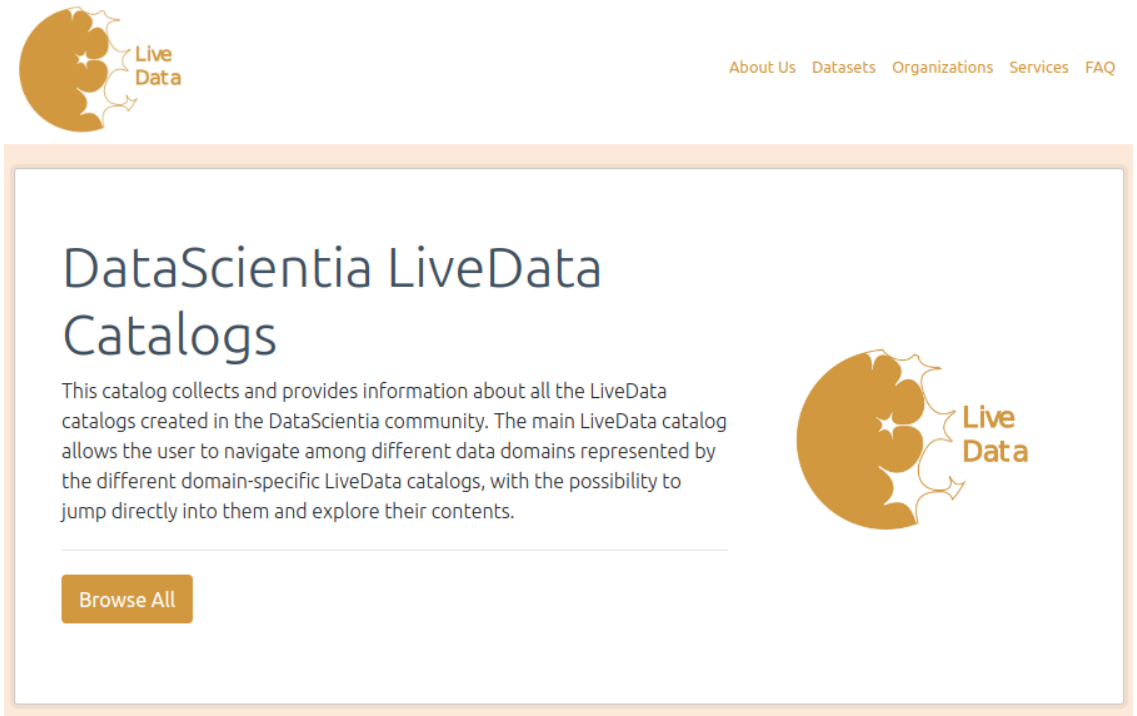


Figure: LiveData catalog[12]

[12] https://datascientiafoundation.github.io/LiveData/

Data catalogs

- The examples above are only some of the several data catalogs available online.
- A lot of catalogs exist, for different domains of interest, topics, purposes.
- There are **high quality and low quality** data catalogs, depending by:
  - the **user experience** provided;
  - the **quality of the data** distributed;
  - the **quality of the metadata** used to describe the data.
- The choice of the data catalogs to collect reusable resources, is crucial to mitigate the cost of the data reuse.

Part 2.5
What does data represent?

1 Types of source
2 Types of language
3 Types of data schema
4 Types of data values
5 What does data represent?

What does data represent?

- The following slides provide an overview about the general content of the data.
- We can summarize such content into the below dimensions:
  - Space and Time data
  - Space related to time
  - Time related to space
  - Reference context data
  - Personal context data
  - Stream data

Space and Time

- The world can be viewed as continuously evolving three-dimensional spatial regions, namely space. Within each spatial region, there is a set of events that are continuously evolving over time.
- We show how data the represents **Space** and **Time** dimensions in the world.

Space (1)

- Spatial data represents a set of places, each can be viewed as a **location** and also an **entity**. For example, BUC can be a place as a library or a building entity.
- P**art-of relations exsit among places**. For example, BUC is part of Trento, Trento is part of Trentino, Trentino is part of Italy, etc.

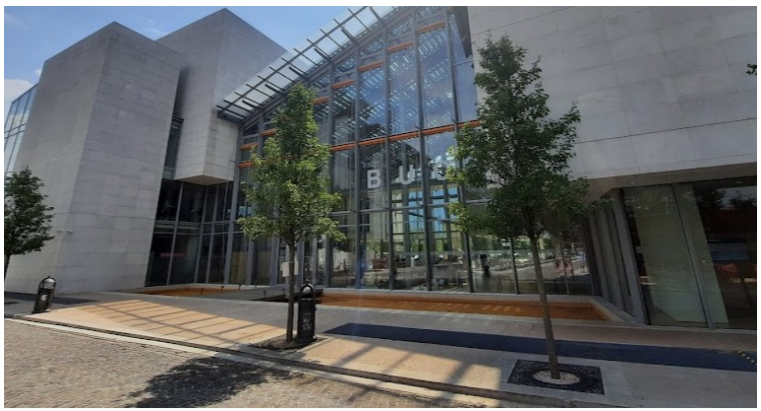


Figure: BUC (Picture is from Google Map)

Space (2)

- Places can be organized by a **hierarchy** based on IS-A/PART-OF property.

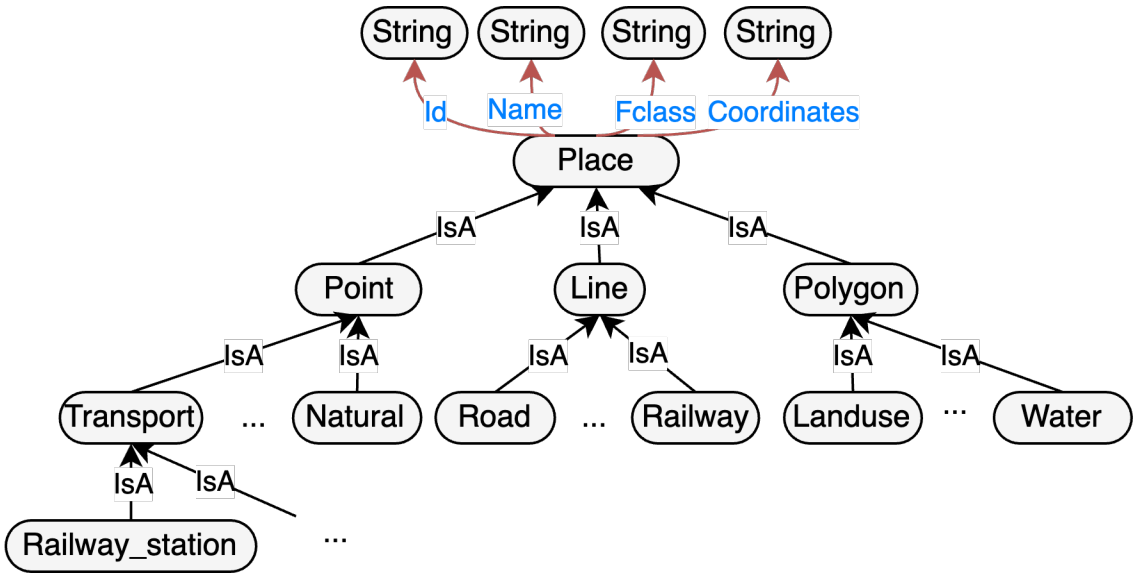


Figure: The OpenStreetMap (OSM) Hierarchy, classified by place categories[13].

[13] https://download.geofabrik.de/osm-data-in-gis-formats-free.pdf

Time

- Temporal data represents a **sequence of events** evolving over time. For example, personal Google Calendars, proceedings of conferences, time diaries from people, and sensor data from phones.

userid	questiontimestamp2	what	where	withwhom	mood
0	05-11 17:02:19 UTC	Studying	Other Library	Friend(s)	4
0	05-11 18:02:19 UTC	Studying	Other Library	Friend(s)	5
0	05-11 19:02:19 UTC	Shopping	Shop, supermarket, etc	Relative(s)	5
0	05-11 20:02:19 UTC	Eating	Relatives Home	Relative(s)	4
0	05-11 21:02:19 UTC	Social life	Relatives Home	Relative(s)	3
0	05-11 22:02:19 UTC	Social life	Relatives Home	Relative(s)	4
0	05-11 23:02:19 UTC	Selfcare	Relatives Home	Alone	4
0	05-12 00:02:19 UTC	Sleeping	Relatives Home	Alone	4

timestamp	latitude	longitude	altitude	provider	userid
2024/9/2 7:50	45.8825	11.032	-1	gps	0
2024/6/2 7:51	45.8825	11.032	-1	gps	0
2024/6/2 7:52	45.8825	11.032	-1	gps	0
2024/6/2 7:53	45.8825	11.032	-1	gps	0
2024/6/2 7:54	45.8826	11.0322	-1	gps	0
2024/6/2 7:55	45.8826	11.0322	-1	gps	0
2024/6/2 7:56	45.8826	11.0322	-1	gps	0
2024/6/2 7:57	45.8827	11.0312	-1	gps	0

Figure: Time diaries (left) and GPS data (right).

(Data values are artificial because of privacy and security concerns)

Temporal Knowledge Graph Completion (TKGC)[14]

- KGs are structured data that typically contain a set of facts, these facts are annotated with temporal labels. Each KG can contain spatial information, evolving over time. TKGC is a way to organize data considering space and time dimensions.

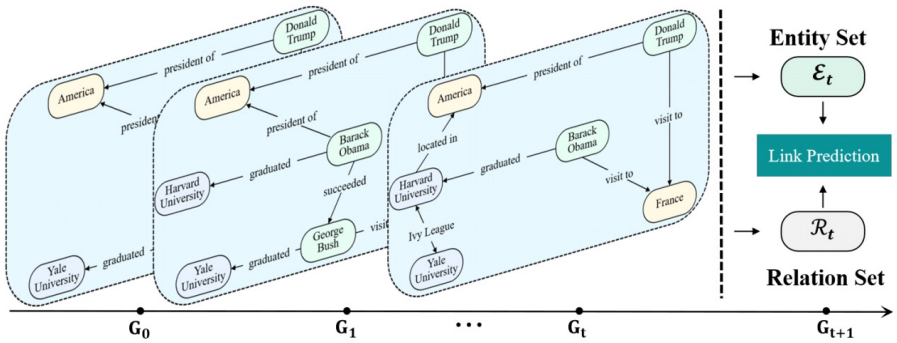


Figure: A type of TKGC example, where Gt denotes the static Knowledge Graph at time t.

[14] Wang, Jiapu, et al. A Survey on Temporal Knowledge Graph Completion: Taxonomy, Progress, and Prospects.arXiv e-prints, 2023.

Space-Time

- Enquire space information over time.

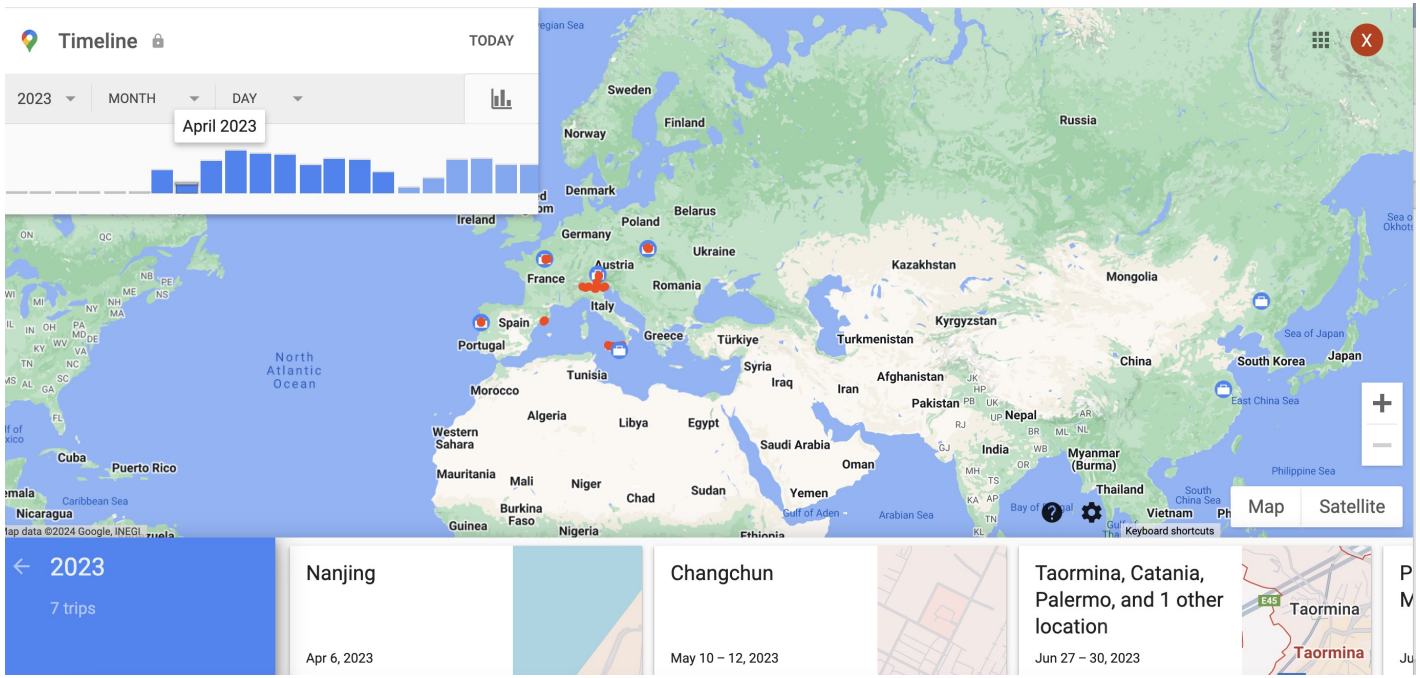


Figure: Google Maps Timeline[15] Example.

[15] https://timeline.google.com/maps/timeline

Time-Space

- Enquire time information over spaces.

userid	questiontimestamp2	what	where	withwhom	mood
0	05-11 17:02:19 UTC	Studying	Other Library	Friend(s)	4
0	05-11 18:02:19 UTC	Studying	Other Library	Friend(s)	5
0	05-11 19:02:19 UTC	Shopping	Shop, supermarket, etc	Relative(s)	5
0	05-11 20:02:19 UTC	Eating	Relatives Home	Relative(s)	4
0	05-11 21:02:19 UTC	Social life	Relatives Home	Relative(s)	3
0	05-11 22:02:19 UTC	Social life	Relatives Home	Relative(s)	4
0	05-11 23:02:19 UTC	Selfcare	Relatives Home	Alone	4
0	05-12 00:02:19 UTC	Sleeping	Relatives Home	Alone	4

Figure: Time Diaries Example.

- In *Other Library*, where user *0* was *Studying*.
- In *Shop, supermarket, etc*, user *0* was Shopping.
- In *Relatives Home*, user *0* was having: *Eating*, *Social life*, *Selfcare* and *Sleeping*.

Context

- Data have the ability to represent ***Context***[16]. Context is defined as **any information that can be used to characterise the situation of an entity**[17].

- In this course, we recognized and classified 2 types of context:
  - Reference Context
  - Personal Context

[16] Giunchiglia, Fausto, et al. A context model for personal data streams. APWeb-WAIM, 2022.

[17] A. K. Dey. Understanding and using context. Personal and ubiquitous computing, 2001.

Reference Context (1)

- A reference context provides **a person-independent objective** all-encompassing view of a third-party observer.
- A reference context keeps track of the **environment** within which people are operating, defined in terms of a reference observation period and a reference location.

Reference Context (2)

- Examples of reference observation periods: Italy public holiday[18], Coordinated Universal Time (UTC).
- Examples of reference locations: OpenStreetMap(OSM)[19], Point of Interest in Trentino[20].

[18] https://www.wearedevelopers.com/magazine/italy-public-holidays
[19] https://www.openstreetmap.org
[20] https://dati.trentino.it/dataset/punti-di-interesse-del-trentino

Personal Context (1)

- A personal context encodes the **person's subjective view of the world**.
  - One personal context is for one person in a time period, e.g., from 9 to 10 am, where she is, what she is doing, why she does it, who she is with, her mood, etc.
  - Different users can perceive different personal contexts, even if they are involved in the same situation, e.g., in a lecture, the presentation activity for a teacher, or the studying activity for a student.

Personal Context (2)

- Personal contexts are built from the user-provided data for describing the current situations. For example:
  - human answers to machine questions: user labels[21], user time diaries[22].
  - Sensor data, e.g., GPS, Accelerometer data from phones[21,23].
  - Human self-reports[24].

[21] http://extrasensory.ucsd.edu/
[22] https://datascientiafoundation.github.io/LivePeople/datasets/2018-SU2-Trento-Time%20Diaries/
[23] https://datascientiafoundation.github.io/LivePeople/datasets/
[24] https://drive.google.com/file/d/1yY8RNaWO_eh4-UnXHkL2jpZld2739K3K/view

Context Unification [25] (1)

Definition

$$C = C_R \uplus \{C_P\} \quad (1)$$

where C is observation context; C_R is a reference context; $\uplus$ is the context unification operator; $\{C_P\}$ is the set, one or more, of personal contexts under consideration.

- Context unification operates in two macro steps:
  - Unification between C_R and $\{C_P\}$, one C_P at the time.
  - Pairwise unification between any two C_P's after they are unified with C_R.

[25] Fausto Giunchiglia, Xiaoyue Li. Big-Thick Data generation via reference and personal context unification, arXiv:2409.05883

Context Unification (2)

- The unification of two personal contexts always with respect to the reference context. Context unification exploits three specific types of unification:
  - **Etype and Property Unification (EPU)**. This is typical problem of ontology/ schema alignment;
  - **Spatio-Temporal Unification (STU)**. The spatio-temporal coordinates of entities are exploited. We have two types of results. The first is the recognition of two entities as being the same entity. The second is the spatial relation holding at a certain time between entities, e.g., a building near a bus stop.
  - **Entity Unification (EU)**. This is done using specific entity properties, different from spatio-temporal properties, mainly name and identifier.

Context Observation Queries (1)

- Context unification can be seen as defining a general purpose meta-process for the desired purpose. Hence, we classify the possible observation purposes into four main groups, as a function of what one is interested in observing.
  - **Reference (R) enquiries**
  - **Personal (P) enquiries**
  - **Personal-Reference (PR) enquiries**
  - **Reference-Personal (RP) enquiries**

Context Observation Queries (2)

- **Reference (R) enquiries.** Its goal is to know the details of the reference context. R-enquiries are posed only to know the details of the reference context, independently of the dynamics which may occur inside it.
  - 'Where is the bus stop near the bar named Bar Sport?'
  - 'What are the supermarkets near BiBa's Restaurant?'

Context Observation Queries (3)

- **Personal (P) enquiries.** Its goal is to know about what people have done in a certain period of time, including also their subjective view of what happened. P-enquiries are only to one or a sequence of $C_P s$, with no possibility of reference to entities of C_R which are not part of C_P.
  - 'Which places did you go in a certain period?'
  - 'What were you doing there?'
  - 'What was your mood?'

  The answers are associated with a person's subjective locations (e.g., shopping place), events (e.g., shopping) and moods (e.g., happy).

Context Observation Queries (4)

- **Personal-Reference (PR) enquiries.** Its goal is to explore how the inside of the reference context evolves as a function of the entities which populate it within a certain period of time. PR enquiries are posed to the observation context focusing on C_R and are about its state as a function of the activities of one or more C_P's.
  - 'How many people are in the Biba's restaurant during weekends?'
  - 'How many attractions in Trento have involved you during the last week?'

- **Reference-Personal (RP) enquiries.** Its goal is to explore the environment around a person and its impact on her. These are enquiries about one or more persons posed to the observation context C.
  - 'What was your mood when you were in the Coop supermarket?'
  - 'Which friends of you were in the Biba's restaurant during the dinner, yesterday?'

What is a Stream

2 Part 2 - State Of the Art
- Part 2.1 - Type of source
- Part 2.2 - Type of language
- Part 2.3 - Type of data schema
- Part 2.4 - Type of data values
- Part 2.5 - What does data represent?
- Part 2.6 - What is a Stream?

What is a Stream?

Definition

Streams are the continuous surge of events that are happening in time and space.

Streams are complex, continuous objects. To handle them in a digital world those objects undergo a series of processes.

- **sampling**: we need to make sample of the events;
- **approximation**: we need to measure the event stream evolution through sensors, introducing errors:
  - *discretization* of values (e.g. due to the precision of the sensor, the resolution of the DAC, etc.)
  - *semantic approximation*: contexts themselves are multi-dimensional streams. If context is not fully aligned with the data there might be a misinterpretation.
- **windowing**: the stream is *unbounded*, but our memory is finite. We need to consider a *finite part* of the stream at each moment.

Stream datasets

We have an approximation of the real-world stream.

- Context does change / there are multiple contexts.
- The dataset is continuously updated.
- The purpose might change.

While very general, is **not possible** to create a lossless representation in knowledge graph.

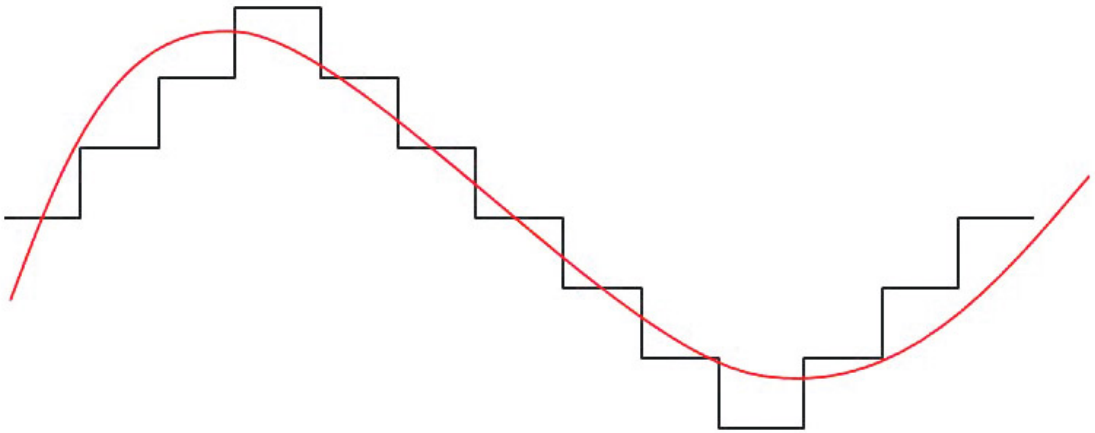


Figure: Real-word stream vs Stream dataset

Classic datasets

What we have in classic datasets is a "snapshot" in time and space of a stream.

- Context doesn't change.
- The dataset doesn't change.
- The purpose is fixed.

In this scenario, is **possible** create an exhaustive knowledge graph for that specific purpose.

Streams: Timeseries vs Graph Streaming

Timeseries:

- just a set of measurement;
- sampling is generally periodic;
- lack of explicit context.
- Ex. Stocks, sensor, etc...

Graph Streams:

- a state of a set of entities;
- generally event-based;
- context is defined.
- Ex. DBPedia

Note!

Streaming entities is likely built upon timeseries. It offers an higher abstraction and solves representation issues (i.e. Shannon's theorem).

Common issues

1. **Entity Resolution**: recognize entities and relationships (harder between datasets).
2. **Datatype alignment**: give homogeneous format and measure unit (e.g. date conversion).
3. **Conceptualization**: evaluate information in context, extracting knowledge (e.g. WSD to extract concept from text).

Stream issues

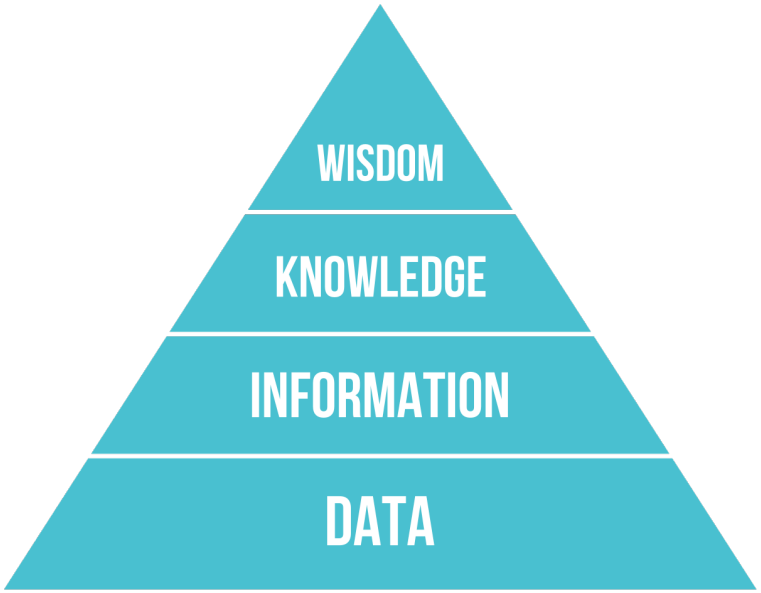
- Single Context, Single Actor
  1. **Sampling rate alignment**Tu et al. 2020
- Single Context, Multiple Actors
  1. **Sampling rate alignment**
  2. **Entity Resolution** (new value or new actor)
- Multiple Context, Single Actors
  1. **Sampling rate alignment**
  2. **Context Drift**Cobb and Van Looveren 2022Bontempelli et al. 2022
- Multiple Context, Multiple Actors
  1. **Sampling rate alignment**
  2. **Entity Resolution** (new value or new actor)
  3. **Context Drift**

Note!

Since context and data are changing, also techniques need to adapt!

Other stream issues

- **Velocity**: generally the requirement is real-time (*online*) processing.
- **Volume**: generally is not possible to store all the values of the stream.
- Dealing more with low level **data** rather than **information**.



Project environment

- Your project lives in a controlled environment.
- Some of the mentioned problem are already solved for you.
- Data is not changing, but we will **simulate** it.
- Biggest problem will be **updates** and (maybe) **alignment**.

Sampling rate alignment

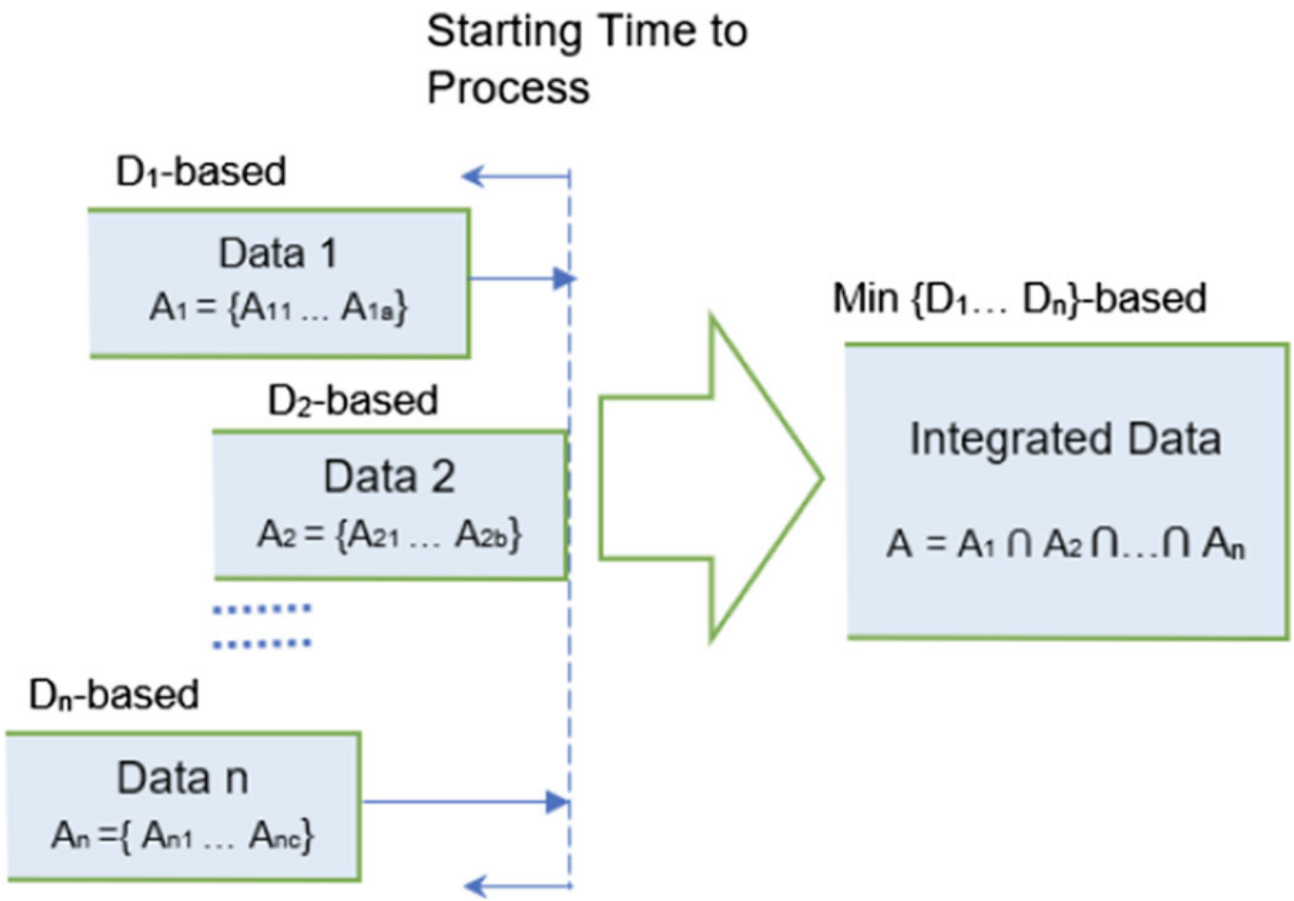


Fig. 1 A scenario of the IoT streaming data integration from multiple sources

References

- Bontempelli, Andrea et al. (July 2022). "Human-in-the-loop handling of knowledge drift". In: *Data Mining and Knowledge Discovery* 36.5, pp. 1865–1884. ISSN: 1573-756X. DOI: 10.1007/s10618-022-00845-0. URL: http://dx.doi.org/10.1007/s10618-022-00845-0.
- Cobb, Oliver and Arnaud Van Looveren (2022). "Context-Aware Drift Detection". In: DOI: 10.48550/ARXIV.2203.08644. URL: https://arxiv.org/abs/2203.08644.
- Tu, Doan Quang et al. (July 2020). "IoT streaming data integration from multiple sources". In: *Computing* 102.10, pp. 2299–2329. ISSN: 1436-5057. DOI: 10.1007/s00607-020-00830-9. URL: http://dx.doi.org/10.1007/s00607-020-00830-9.