Named Entities
Names, Dates, People, and Places

`<persName>Christian Steiner</persName>`

CLARIAH-AT
TEI Introductory School
10.-13. September 2019

Venue:

Centre for Information Modelling (ZIM)
Elisabethstraße 59/III, SR 81.31

What is a (named) entity?

- information object that can be denoted with a proper name

- material or immaterial, concrete or abstract

- names, dates, people, places

- a thing: a car, an account, a person, a condition, an event …

- contains semantics (Semantic Web)
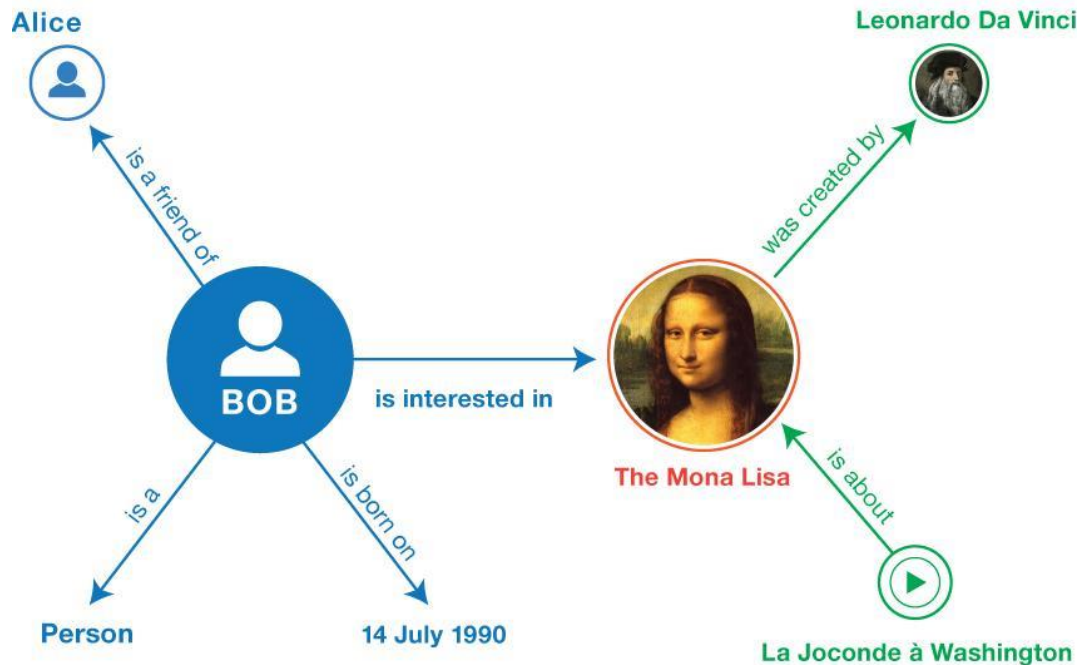

Named Entity Recognition

Why annotate named entities?

- enables complex search queries

- enables the ability to normalize the entities

- enables automatic generation of indices (register of persons, places …)

- enables linking to commentaries

- enables linking to authority files

- enables the usage of the Semantic Web
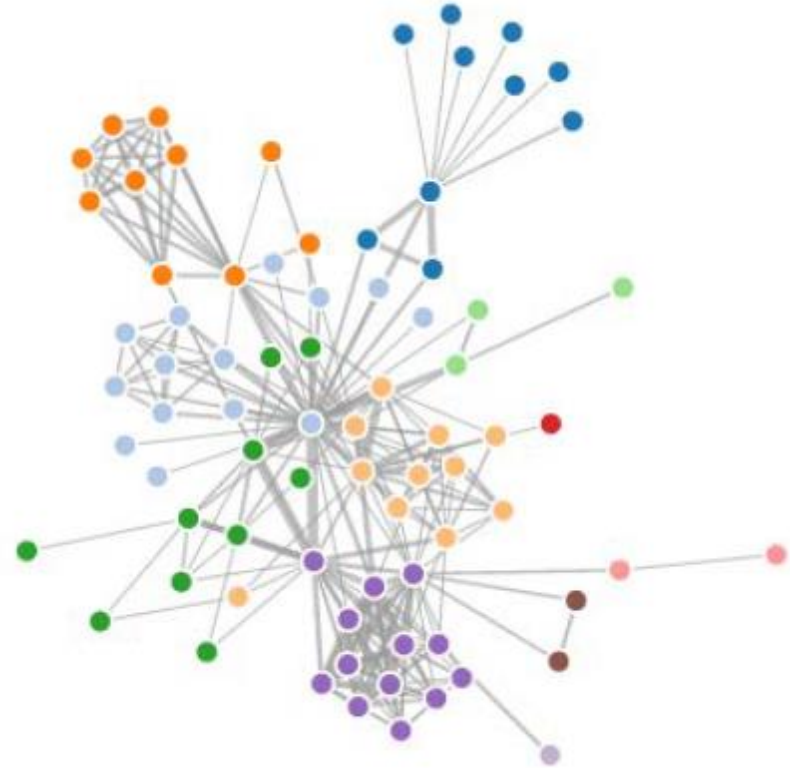
Semantic Web (RDF): Subject - Predicate - Object

- Bob is a Person.
- Bob is a friend of Alice.
- Bob is born on the 4th of July 1990.
- Bob is interested in the Mona Lisa.
- The Mona Lisa was created by Leonardo da Vinci.
- The video 'La Joconde à Washington' is about the Mona Lisa.

http://example.org/bob

https://www.wikidata.org/wiki/
Special:EntityData/Q12418

Alice
http://example.org/alice#me

Leonardo Da Vinci
http://dbpedia.org/resource/
Leonardo_da_Vinci

foaf:knows

dcterms:creator

BOB

foaf:topic_interest

http://example.org/bob#me

dcterms:title

The Mona Lisa
http://www.wikidata.org/entity/Q12418

"Mona Lisa"

dcterms:subject

rdf:type

schema:birthDate

Person
foaf:Person

"1990-07-04"^^xsd:date

La Joconde à Washington
http://data.europeana.eu/item/04802/243FA
8618938F4117025F17A8B813C5F9AA4D619
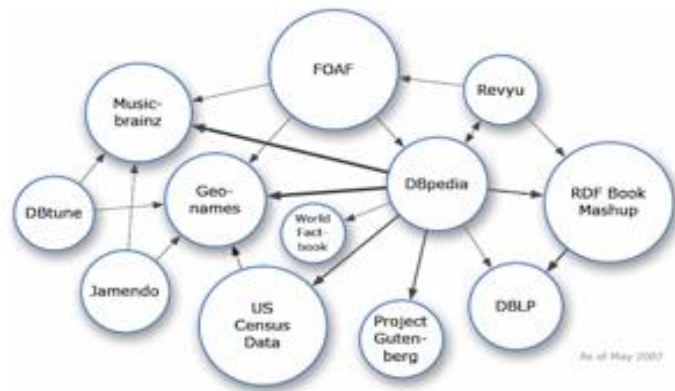
May 2007

Exercise

Try to annotate all entities in your exercise document

Use `<rs>` and `<name>` with the attribute `type`

How could you then get a quick overview of all entities?

Annotation of entities with TEI (simple)

referring strings with `@type`:

`<rs>` (referencing string) contains a general purpose name or referring string.

`<name>` (name, proper noun) contains a proper noun or noun phrase.

```
<name type="person"></name>

<name type="organization"></name>

<name type="place"></name>

<name type="event"></name>
```

Annotation of entities with TEI (detailed)

`<persName>` (personal name) contains a proper noun or proper-noun phrase referring to a person, possibly including one or more of the person's forenames, surnames, honorifics, added names, etc.

`<orgName>` (organization name) contains an organizational name.

`<placeName>` (place name) contains an absolute or relative place name.

`<geogName>` (geographical name) identifies a name associated with some geographical feature such as Windrush Valley or Mount Sinai.

`<objectName>` (name of an object) contains a proper noun or noun phrase used to refer to an object.

Annotation of personal names

`<surname>`

`<forename>`

`<addName>` (additional name) nickname or alias

`<roleName>` official title or rank/position of a person

`<genName>` contains a name component used to distinguish
otherwise similar names on the basis of the relative ages or
generations of the persons named:

`<nameLink>` connecting phrase or link:

```
<persName>
 <forename>Louis</forename>
 <genName>XIV</genName>
 <nameLink>of France</nameLink>
</persName>
```

Annotation of places

`<district>` contains the name of any kind of subdivision of a settlement, e.g. University district

`<settlement>` contains the name of a settlement such as a city or village, e.g. Graz

`<region>` contains the name of an administrative unit such as a state or province, e.g. Styria

`<country>` contains the name of a geo-political unit, e.g. Austria

`<bloc>` contains the name of a geo-political unit with 2 or more countries, e.g. the European Union

Annotation of dates

`<date>` contains a date in any format.

`<time>` contains a phrase defining a time of day in any format.

`<offset>` marks that part of a relative temporal or spatial expression which indicates the direction of the offset between the two place names, dates, or times involved in the expression.

```
<date when="--12-02">
 <date>A week</date>
 <offset>before</offset>
 <date when="--12-09">
  <date type="occasion">my birthday</date> on <date>9th December</date>
 </date>
</date>
```

Annotation of dates and times

@when

@from @to

@notBefore @notAfter

@period

@dur

Exercise

This time concentrate on the proper nouns and dates

Annotate in detail the persons and places and dates in the letter

Use `<persName>` and `<placeName>` and `<date>` and their child elements

References (Links)

- instead of annotating within the text itself it makes more sense to use **references**

- these attributes help us do so:

  Att.global.linking: `@corresp, @sameAs, @prev, @next,` etc.

  Att.canonical: `@ref, @key`

  Att.pointing: `@target`

  Att.naming: `@nymRef`

  Att.referencing: `@cRef`

Datatype `teidata.pointer` (=URI) - except `@key`

References to internal lists

Normalization (e.g. orthography)

Additional information

External references (e.g. to authority files)

```xml
<persName ref="#p.1">Desmond Fitzgerald</persName>
```

```xml
<listPerson>
   <person xml:id="p.1">
      <persName><forename>Desmond</forename> <surname>Fitzgerald</surname></persName>
      <idno type="URI" subtype="Wikidata">https://www.wikidata.org/entity/Q1200811</idno>
      <idno type="URI" subtype="Viaf">https://viaf.org/viaf/2430144/</idno>
   </person>
</listPerson>
```

References to external resources

```
<persName ref="https://www.wikidata.org/entity/Q1200811">Desmond
Fitzgerald</persName>


<persName ref="wd:Q1200811">Desmond Fitzgerald</persName>


<placeName ref="https://www.geonames.org/2964574">Dublin</placeName>


<placeName ref="geo:2964574">Dublin</placeName>
```

Authority control / Knowledge base

In library science, **authority control** is a process that organizes bibliographic information, for example in library catalogs by using a single, distinct spelling of a name (heading) or a numeric identifier for each topic. The word authority in authority control derives from the idea that the names of people, places, things, and concepts are authorized, i.e., they are established in one particular form (Wikipedia)

A **knowledge base** (KB) is a technology used to store complex structured and unstructured information used by a computer system (Wikipedia)

Most important authority files / KB

- [Wikidata](#)
- [DBpedia](#)
- [Geonames](#)
- [VIAF (Virtual International Authority File)](#)
- [GND (Gemeinsame Normdatei)](#)
- [LCCN (Library of Congress Control Number)](#)

Every domain has its own authority files!

How to find authority control or KB IDs?

1) Wikipedia

| Desmond FitzGerald navigational boxes | [show] |

Authority control 🖉 BNF: cb167162890 ⟑ (data) ⟑ · ISNI: 0000 0000 5440 8803 ⟑ · LCCN: nr2006013899 ⟑ · LNB: 000136382 ⟑ · NLA-person: 874409 ⟑ · SNAC: w6672p4x ⟑ · VIAF: 2430144 ⟑ · WorldCat Identities (via VIAF): 2430144 ⟑

2) Search the authority files or knowledge bases, e.g. http://viaf.org/, geonames.org, wikidata.org, https://portal.dnb.de etc.

3) Use special tools for so called *reconciliation* (matching): OpenRefine

Why use external resources?

Structured Data

Unstructured Data

Knowledge Graph
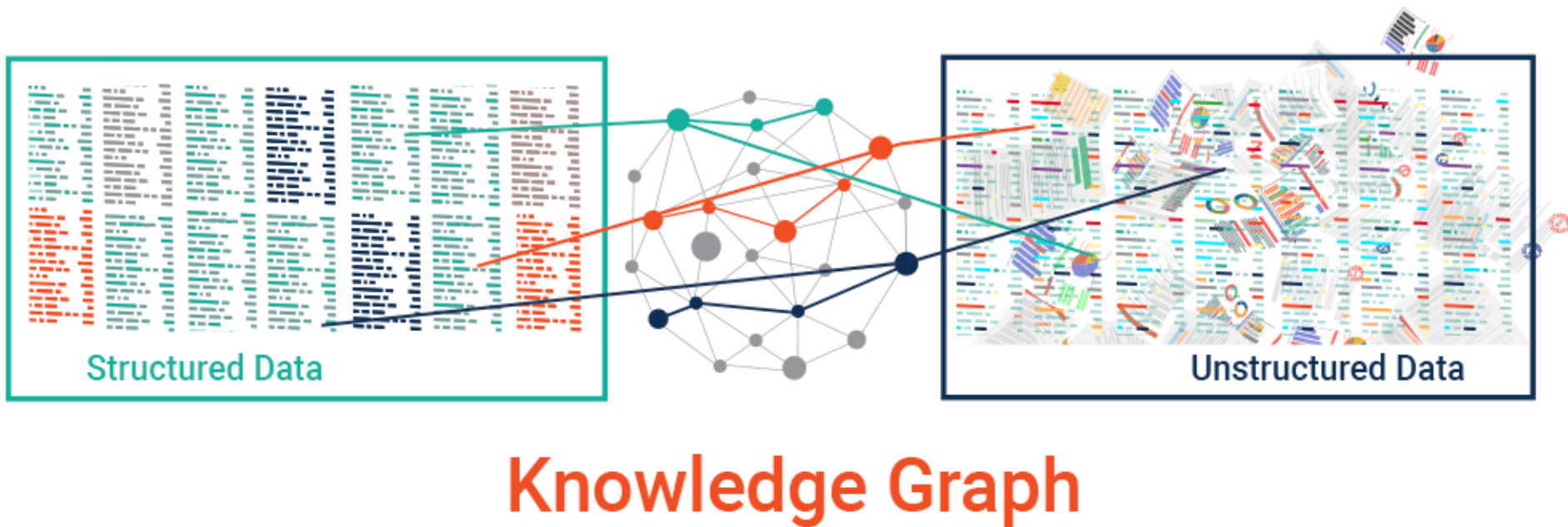
Exercise

Find authority controlled or Knowledge Base resources for the named entities you annotated

Choose a referencing system (internal list / direct external references) of your choice and use it for these resources

What advantages/disadvantage does each approach have?

How could you now find all the references in your letter?