UNIVERSITÄT BERN

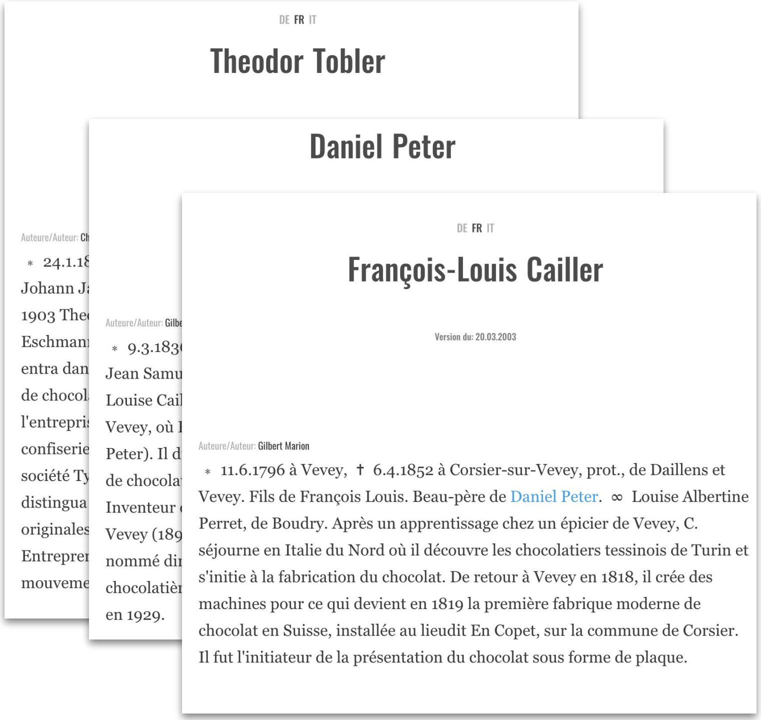
From Text to Graph

Automatic knowledge extraction and semantification of texts

60th CIDOC CRM SIG Meeting
03.04.2025

Stephen Hart, Unibe

Knowledge Extraction: The Goal of the KNEX tool



DE FR IT

Theodor Tobler

Auteure/Auteur: Ch

⁎ 24.1.18
Johann J
1903 Theo
Eschmann
entra dan
de chocol
l'entrepri
confiserie
société Ty
distingua
originales
Entrepren
mouveme

Daniel Peter

Auteure/Auteur: Gilbe

⁎ 9.3.183
Jean Samu
Louise Cail
Vevey, où l
Peter). Il d
de chocola
Inventeur
Vevey (189
nommé dir
chocolatièr
en 1929.

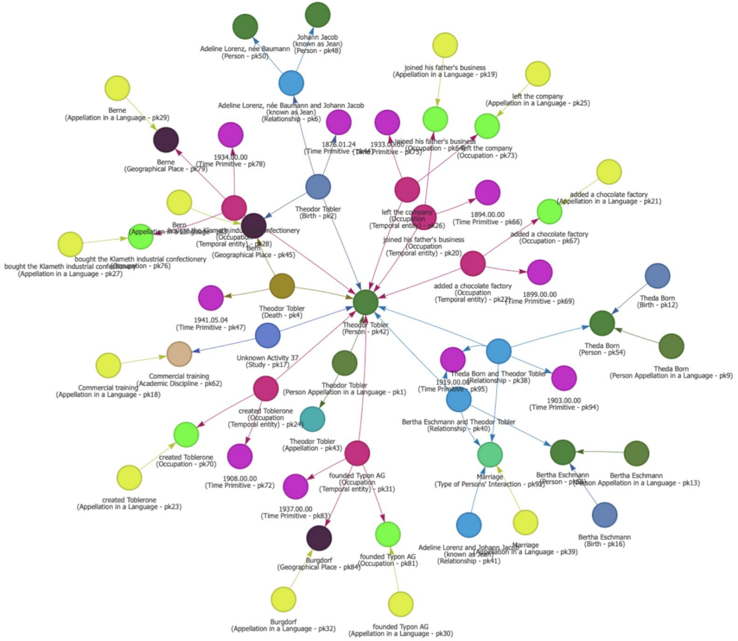
DE FR IT

François-Louis Cailler

Version du: 20.03.2003

Auteure/Auteur: Gilbert Marion

⁎ 11.6.1796 à Vevey, † 6.4.1852 à Corsier-sur-Vevey, prot., de Daillens et Vevey. Fils de François Louis. Beau-père de Daniel Peter. ∞ Louise Albertine Perret, de Boudry. Après un apprentissage chez un épicier de Vevey, C. séjourne en Italie du Nord où il découvre les chocolatiers tessinois de Turin et s'initie à la fabrication du chocolat. De retour à Vevey en 1818, il crée des machines pour ce qui devient en 1819 la première fabrique moderne de chocolat en Suisse, installée au lieudit En Copet, sur la commune de Corsier. Il fut l'initiateur de la présentation du chocolat sous forme de plaque.


Adeline Lorenz, née Baumann (Person - pk50)
Johann Jacob (known as Jean) (Person - pk48)
joined his father's business (Appellation in a Language - pk19)
left the company (Appellation in a Language - pk25)
Berne (Appellation in a Language - pk29)
Adeline Lorenz, née Baumann and Johann Jacob (known as Jean) (Relationship - pk6)
1876.01.24 (Time Primitive - pk44)
1933.00.00 (Time Primitive - pk75)
joined his father's business (Occupation - pk64)
left the company (Occupation - pk73)
Berne (Geographical Place - pk79)
1934.00.00 (Time Primitive - pk78)
Theodor Tobler (Birth - pk2)
left the company (Occupation (Temporal entity) - pk26)
1894.00.00 (Time Primitive - pk66)
added a chocolate factory (Appellation in a Language - pk21)
added a chocolate factory (Occupation - pk67)
Bern (Appellation in a Language - pk3)
bought the Klameth industrial confectionery (Occupation (Temporal entity) - pk28)
Bern (Geographical Place - pk45)
joined his father's business (Occupation (Temporal entity) - pk20)
bought the Klameth industrial confectionery (Occupation - pk76)
bought the Klameth industrial confectionery (Appellation in a Language - pk27)
added a chocolate factory (Occupation (Temporal entity) - pk22)
1899.00.00 (Time Primitive - pk69)
Theda Born (Birth - pk12)
1941.05.04 (Time Primitive - pk47)
Theodor Tobler (Death - pk4)
Theodor Tobler (Person - pk42)
Unknown Activity 37 (Study - pk17)
Theda Born (Person - pk54)
Theda Born (Person Appellation in a Language - pk9)
Commercial training (Appellation in a Language - pk18)
Commercial training (Academic Discipline - pk62)
Theodor Tobler (Person Appellation in a Language - pk1)
Theda Born and Theodor Tobler (Relationship - pk38)
1919.00.00 (Time Primitive - pk95)
1903.00.00 (Time Primitive - pk94)
created Toblerone (Occupation (Temporal entity) - pk24)
Bertha Eschmann and Theodor Tobler (Relationship - pk40)
created Toblerone (Occupation - pk70)
Theodor Tobler (Appellation - pk43)
1908.00.00 (Time Primitive - pk72)
founded Typon AG (Occupation (Temporal entity) - pk31)
Marriage (Type of Persons' Interaction - pk92)
Bertha Eschmann (Person - pk58)
Bertha Eschmann (Person Appellation in a Language - pk13)
created Toblerone (Appellation in a Language - pk23)
1937.00.00 (Time Primitive - pk83)
Burgdorf (Geographical Place - pk84)
founded Typon AG (Occupation - pk81)
Adeline Lorenz and Johann Jacob (known as Jean) (Relationship - pk41)
Marriage (Appellation in a Language - pk39)
Bertha Eschmann (Birth - pk16)
Burgdorf (Appellation in a Language - pk32)
founded Typon AG (Appellation in a Language - pk30)

Knex Context: The Geovistory Environment

LARHRA
UMR 5190

b
UNIVERSITÄT
BERN

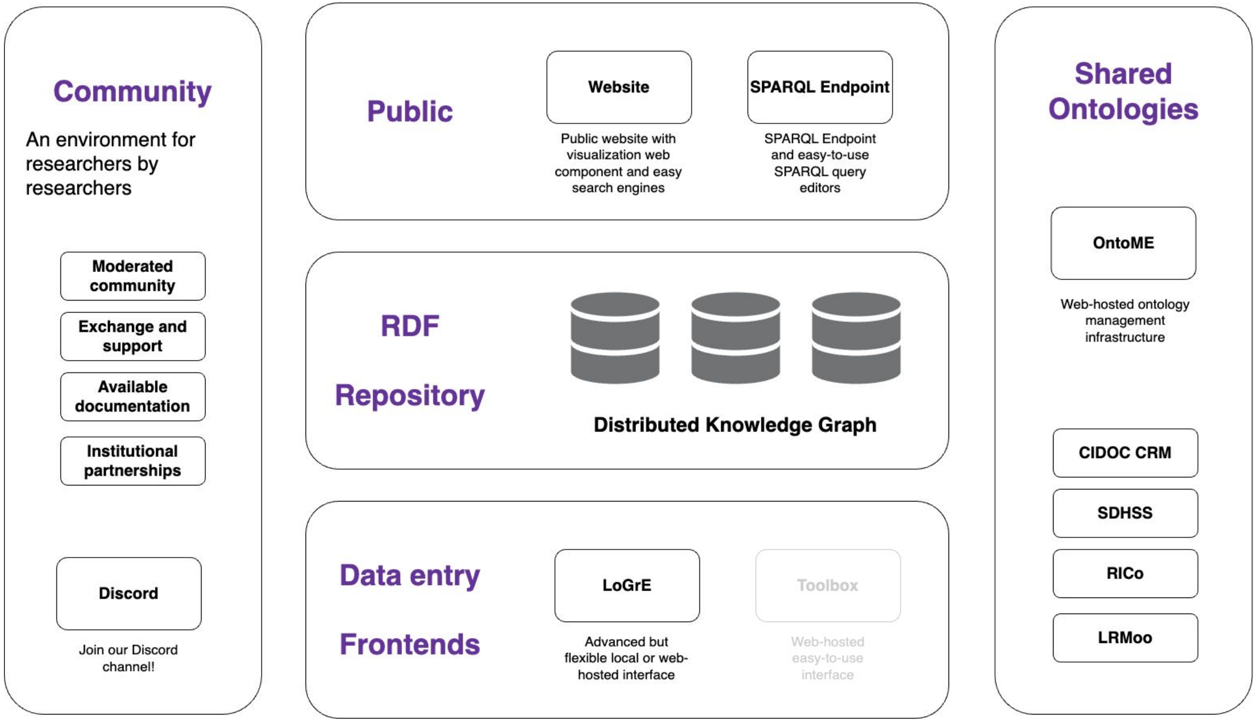
KleioLab

- L'ÉCOLE DES HAUTES ÉTUDES EN SCIENCES SOCIALES
- UNIVERSITÉ LYON III JEAN MOULIN
- ÉCOLE FRANÇAISE DE ROME *Histoire, Archéologie, Sciences sociales*
- UNI BASEL
- huygens instituut
- unine UNIVERSITÉ DE NEUCHÂTEL
- UNIVERSITÉ PARIS 1 PANTHÉON SORBONNE
- UNIVERSITÉ TOULOUSE Jean Jaurès

- unine UNIVERSITÉ DE NEUCHÂTEL
- Universität Zürich UZH
- DaSCH
- hep/ haute école pédagogique vaud
- UNI BASEL
- UNI FR
- UNIVERSITÉ DE FRIBOURG UNIVERSITÄT FREIBURG
- Freie Universität Berlin
- USI
- UNIVERSITÄT GREIFSWALD Wissen lockt. Seit 1456
- UNIVERSITÄT TRIER
- abes
- huygens instituut
- swissuniversities

https://www.geovistory.org

The Geovistory Environment



Community

An environment for researchers by researchers

Moderated community

Exchange and support

Available documentation

Institutional partnerships

Discord

Join our Discord channel!

Public

Website

Public website with visualization web component and easy search engines

SPARQL Endpoint

SPARQL Endpoint and easy-to-use SPARQL query editors

RDF Repository

Distributed Knowledge Graph

Data entry Frontends

LoGrE

Advanced but flexible local or web-hosted interface

Toolbox

Web-hosted easy-to-use interface

Shared Ontologies

OntoME

Web-hosted ontology management infrastructure

CIDOC CRM

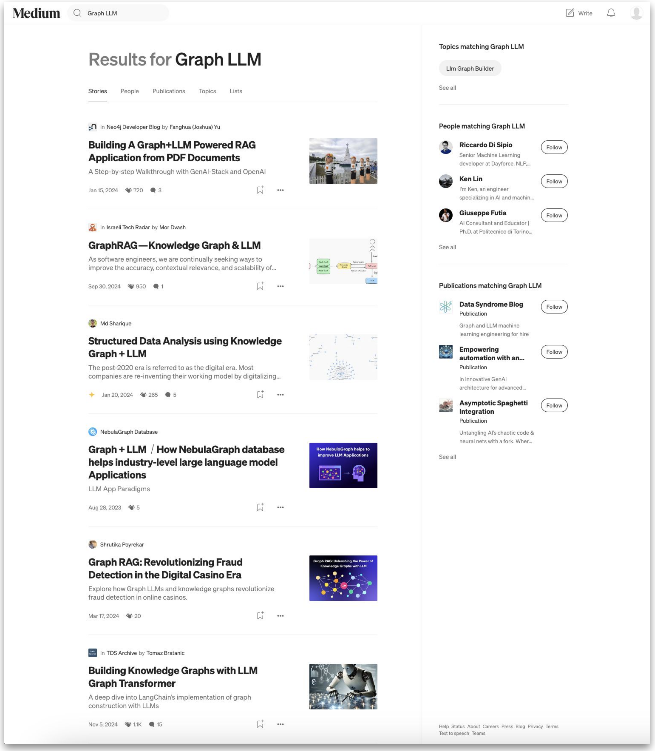
SDHSS

RICo

LRMoo

Knowledge Extraction: The Goal

- New LLM technologies has transformed the way of interacting with texts and generating new knowledge
  - Many new attempts at generating Knowledge Graphs from texts
- For the moment little attempt in the Humanities (but it's starting)
- Solutions existing lack Semantic Engineering, and creates the ontology from the text
  - **Our goal is to map the extracted data to CIDOC CRM and the SDHSS extension**



Results for **Graph LLM**

In Neo4j Developer Blog by Fanghua (Joshua) Yu

Building A Graph+LLM Powered RAG Application from PDF Documents

A Step-by-step Walkthrough with GenAI-Stack and OpenAI

Jan 15, 2024 · 720 · 3

In Israeli Tech Radar by Mor Dvash

GraphRAG—Knowledge Graph & LLM

As software engineers, we are continually seeking ways to improve the accuracy, contextual relevance, and scalability of...

Sep 30, 2024 · 950 · 1

Md Sharique

Structured Data Analysis using Knowledge Graph + LLM

The post-2020 era is referred to as the digital era. Most companies are re-inventing their working model by digitalizing...

Jan 20, 2024 · 265 · 5

NebulaGraph Database

Graph + LLM / How NebulaGraph database helps industry-level large language model Applications

LLM App Paradigms

Aug 28, 2023 · 5

Shrutika Poyrekar

Graph RAG: Revolutionizing Fraud Detection in the Digital Casino Era

Explore how Graph LLMs and knowledge graphs revolutionize fraud detection in online casinos.

Mar 17, 2024 · 20

In TDS Archive by Tomaz Bratanic

Building Knowledge Graphs with LLM Graph Transformer

A deep dive into LangChain's implementation of graph construction with LLMs

Nov 5, 2024 · 1.1K · 15

Topics matching Graph LLM

Llm Graph Builder

People matching Graph LLM

Riccardo Di Sipio
Senior Machine Learning developer at Dayforce. NLP,...

Ken Lin
I'm Ken, an engineer specializing in AI and machin...

Giuseppe Futia
AI Consultant and Educator | Ph.D. at Politecnico di Torino...

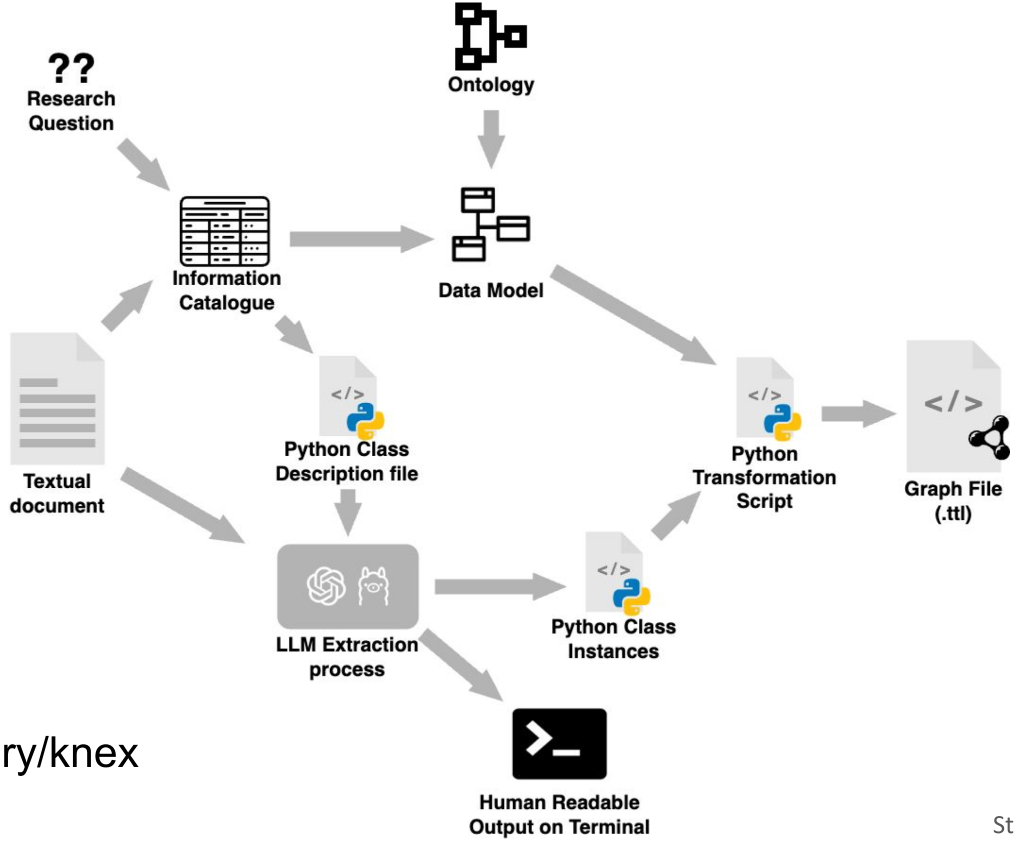
Publications matching Graph LLM

Data Syndrome Blog
Publication
Graph and LLM machine learning engineering for hire

Empowering automation with an...
Publication
In innovative GenAI architecture for advanced...

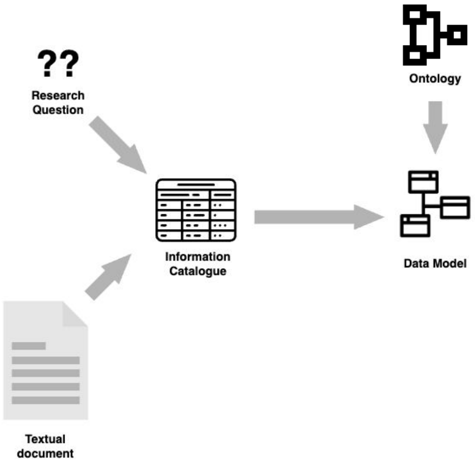
Asymptotic Spaghetti Integration
Publication
Untangling AI's chaotic code & neural nets with a fork. Wher...

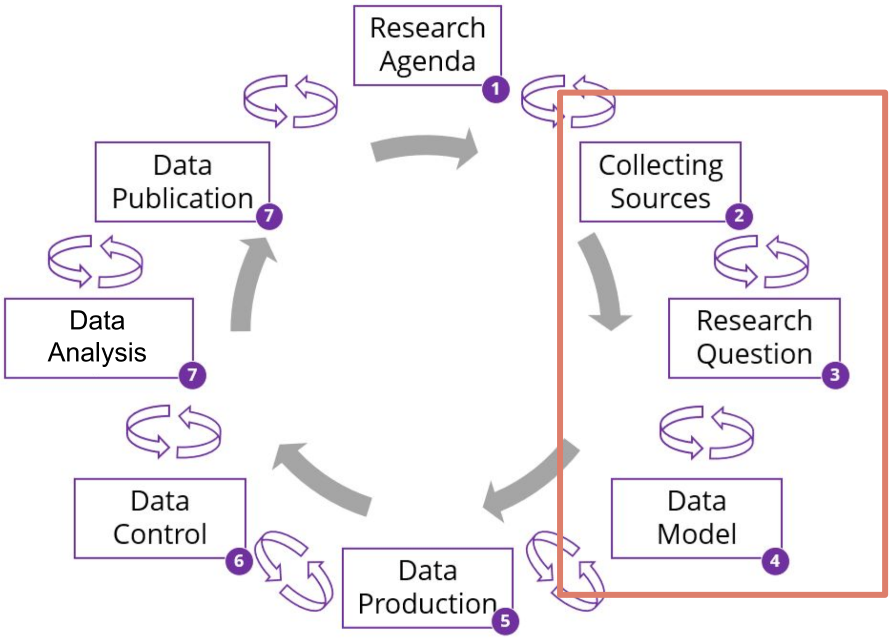
Knex: The Pipeline



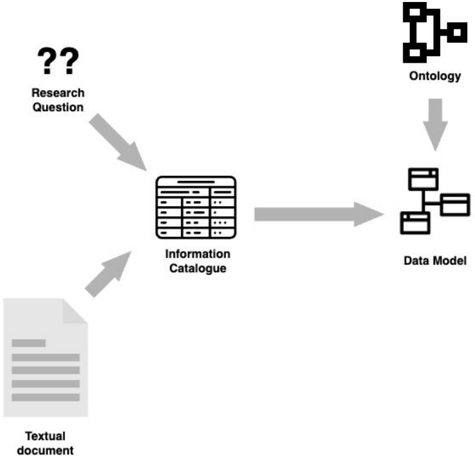
https://github.com/geovistory/knex

Developing the Data Model: Content Analysis





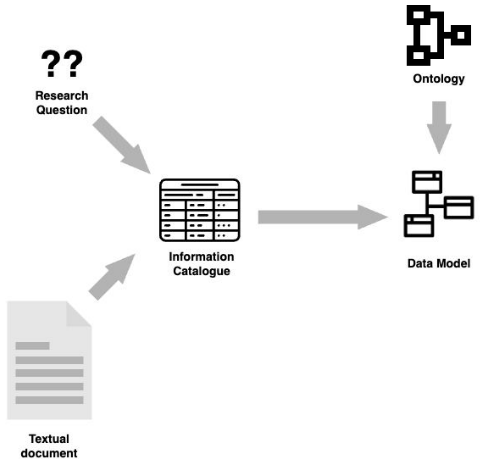
Developing the Data Model: Content Analysis



Theodor Tobler

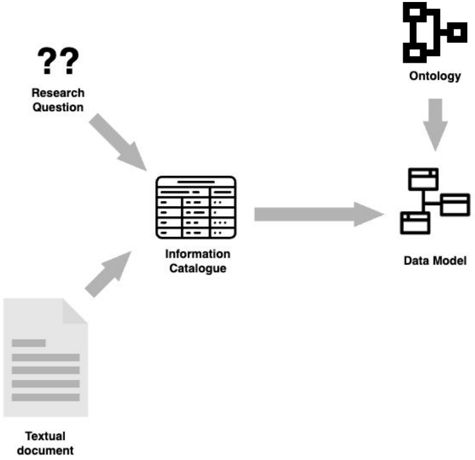
Theodor Tobler was born on 24.1.1876 in Bern and died on 4.5.1941 in Bern, prot., from Lutzenberg. Son of Johann Jacob (known as Jean), confectioner, and Adeline Lorenz, née Baumann. Married in 1903 Theda Born, daughter of Emil, architect (divorced in 1919), and married in 1919 Bertha Eschmann, daughter of Heinrich. Commercial training in Geneva and Venice. Tobler joined his father's business in 1894, adding a chocolate factory in 1899. In 1908, he created Toblerone. After leaving the company (1933), which was in need of restructuring, he bought the Klameth industrial confectionery in Bern in 1934. In 1937, he founded Typon AG in Burgdorf, which produced films for the graphics industry, and made a name for itself with new products and original advertising strategies. Member of the Bernese Masonic lodge A l'Espérance (from 1902). An entrepreneur and advocate of social reform, Tobler was also active in the pacifist and pan-European movements.

Developing the Data Model: Information Catalogue



Entity	Attributes
Person	Name has birth place -> Place Birth Date has father -> Person has mother -> Person studied -> Study has occupation -> Occupation
Study	has study place -> Place Study Date Study Domain
Occupation	has occupation place -> Place Occupation Date Related Group
Group	Group Name has founder -> Person
Place	Place Name

Developing the Data Model: Ontology

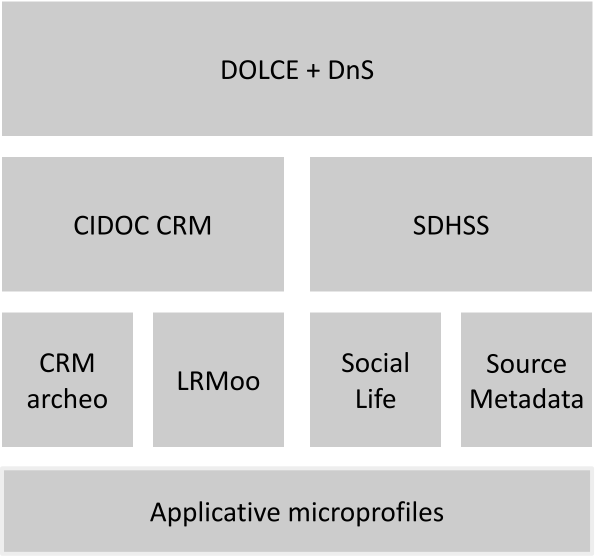


Foundational ontologies and modelling methodologies

Generic, domain related core ontology

Demain related extensions

Modular data models


DOLCE + DnS

CIDOC CRM

SDHSS

CRM archeo

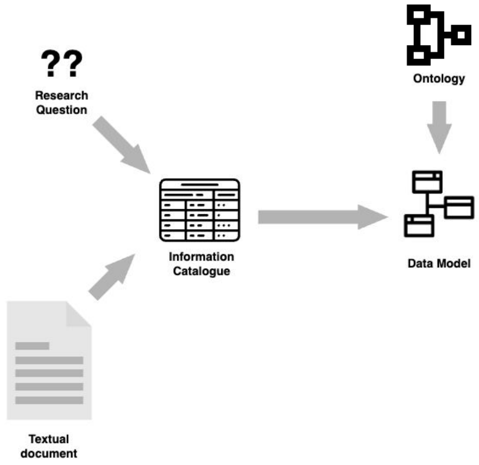
LRMoo

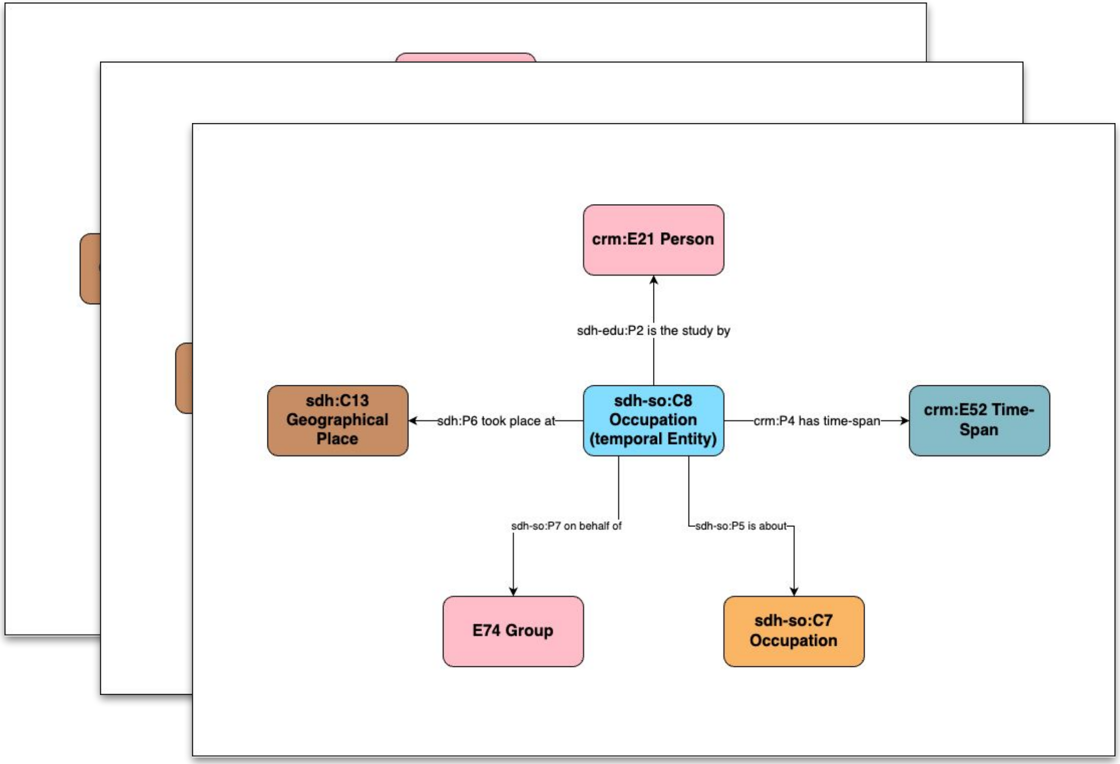
Social Life

Source Metadata

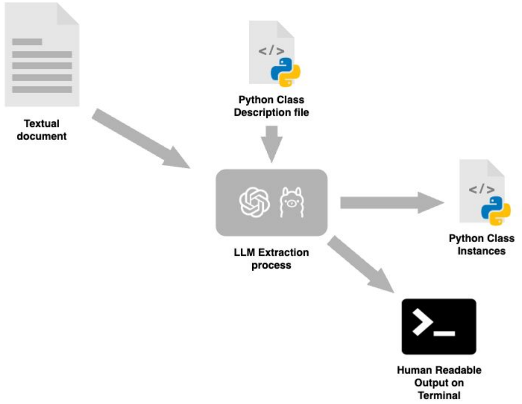
Applicative microprofiles

Developing the Data Model: Data Model





Knowledge Extraction Process


Textual document
Python Class Description file
LLM Extraction process
Python Class Instances
Human Readable Output on Terminal

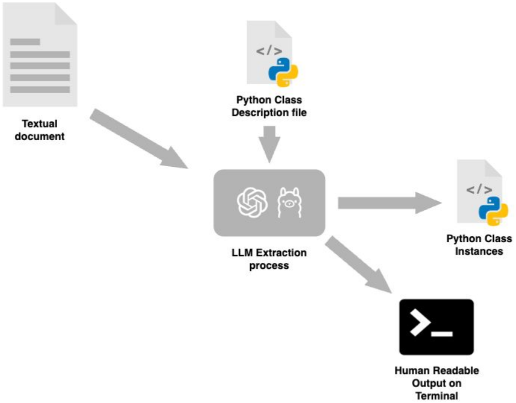
LLMs struggle to understand the event-centric model of CIDOC CRM

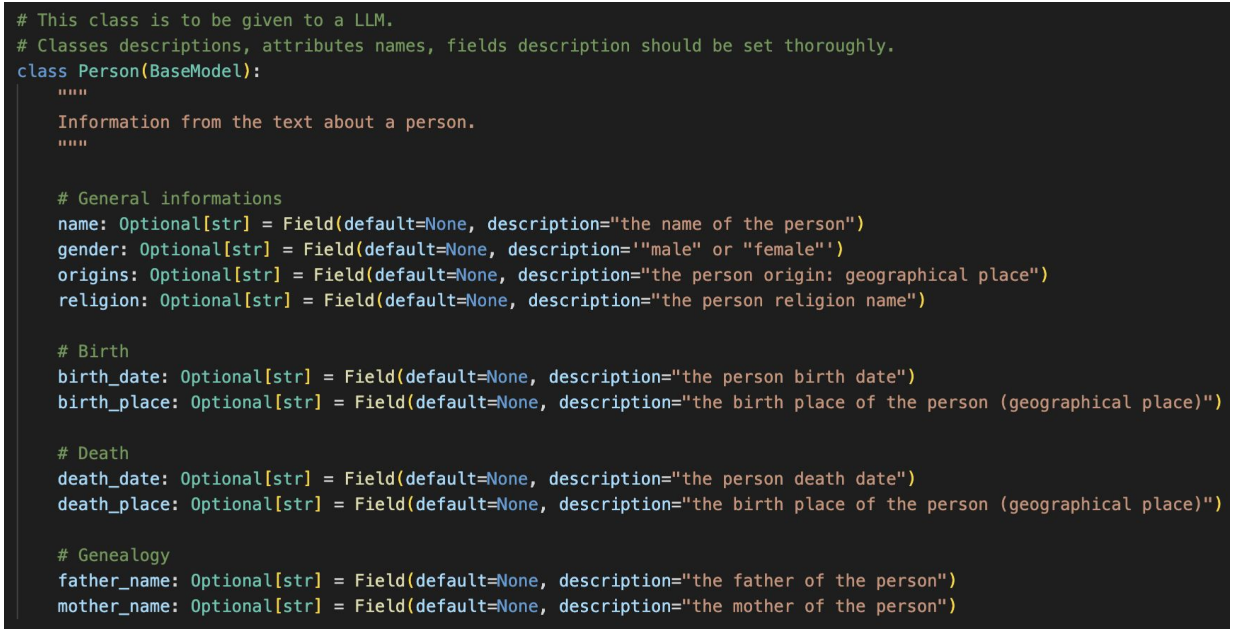
The solution adopted is to rely on **shortcuts**, that would then be transformed into the extended CIDOC CRM paths later on in the pipeline.

LLM steps:

- Identify each entity in the text
- For each entity, parse the text for extracting the shortcut assertions
- Store the assertions as Python class instances

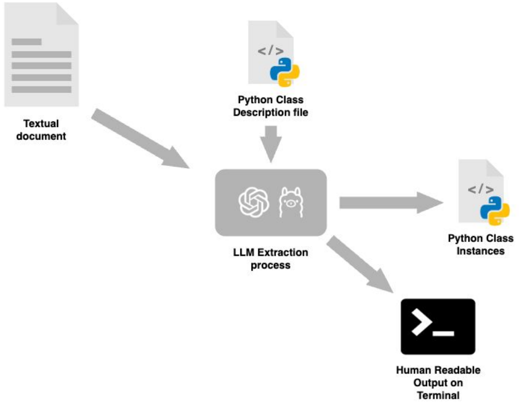
Knowledge Extraction Process: Class Description

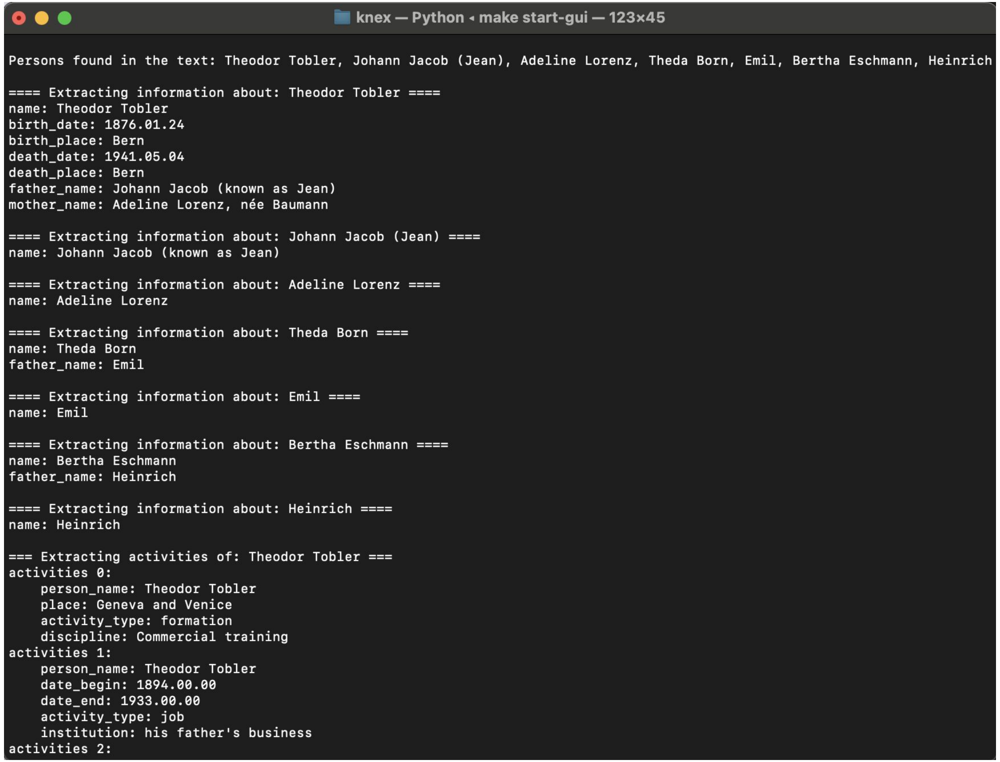

Textual document
Python Class Description file
LLM Extraction process
Python Class Instances
Human Readable Output on Terminal



```
# This class is to be given to a LLM.
# Classes descriptions, attributes names, fields description should be set thoroughly.
class Person(BaseModel):
    """
    Information from the text about a person.
    """

    # General informations
    name: Optional[str] = Field(default=None, description="the name of the person")
    gender: Optional[str] = Field(default=None, description='"male" or "female"')
    origins: Optional[str] = Field(default=None, description="the person origin: geographical place")
    religion: Optional[str] = Field(default=None, description="the person religion name")

    # Birth
    birth_date: Optional[str] = Field(default=None, description="the person birth date")
    birth_place: Optional[str] = Field(default=None, description="the birth place of the person (geographical place)")

    # Death
    death_date: Optional[str] = Field(default=None, description="the person death date")
    death_place: Optional[str] = Field(default=None, description="the birth place of the person (geographical place)")

    # Genealogy
    father_name: Optional[str] = Field(default=None, description="the father of the person")
    mother_name: Optional[str] = Field(default=None, description="the mother of the person")
```

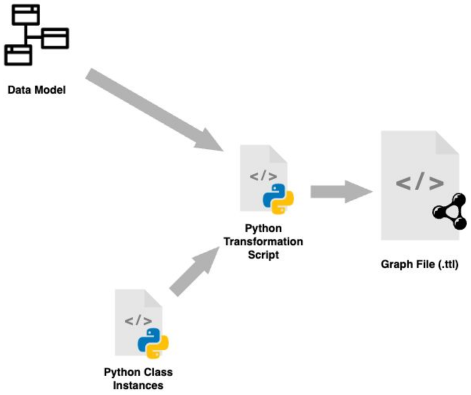
Knowledge Extraction Process: Output

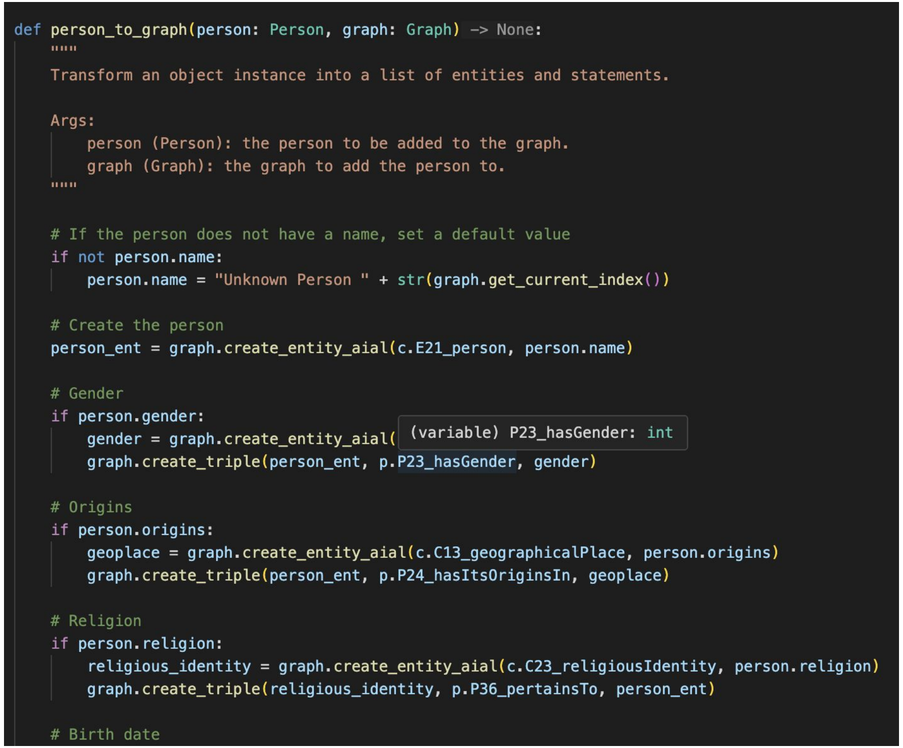

Textual document
Python Class Description file
LLM Extraction process
Python Class Instances
Human Readable Output on Terminal



```
Persons found in the text: Theodor Tobler, Johann Jacob (Jean), Adeline Lorenz, Theda Born, Emil, Bertha Eschmann, Heinrich

==== Extracting information about: Theodor Tobler ====
name: Theodor Tobler
birth_date: 1876.01.24
birth_place: Bern
death_date: 1941.05.04
death_place: Bern
father_name: Johann Jacob (known as Jean)
mother_name: Adeline Lorenz, née Baumann

==== Extracting information about: Johann Jacob (Jean) ====
name: Johann Jacob (known as Jean)

==== Extracting information about: Adeline Lorenz ====
name: Adeline Lorenz

==== Extracting information about: Theda Born ====
name: Theda Born
father_name: Emil

==== Extracting information about: Emil ====
name: Emil

==== Extracting information about: Bertha Eschmann ====
name: Bertha Eschmann
father_name: Heinrich

==== Extracting information about: Heinrich ====
name: Heinrich

=== Extracting activities of: Theodor Tobler ===
activities 0:
    person_name: Theodor Tobler
    place: Geneva and Venice
    activity_type: formation
    discipline: Commercial training
activities 1:
    person_name: Theodor Tobler
    date_begin: 1894.00.00
    date_end: 1933.00.00
    activity_type: job
    institution: his father's business
activities 2:
```

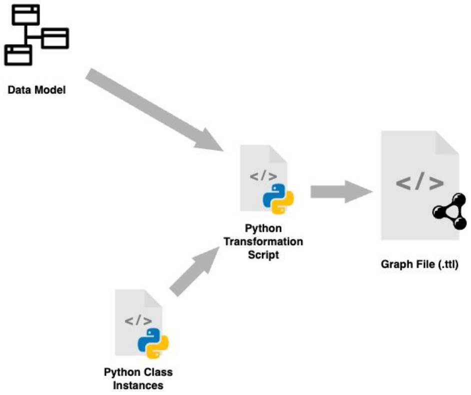
Data Transformation Process: Conversion Script

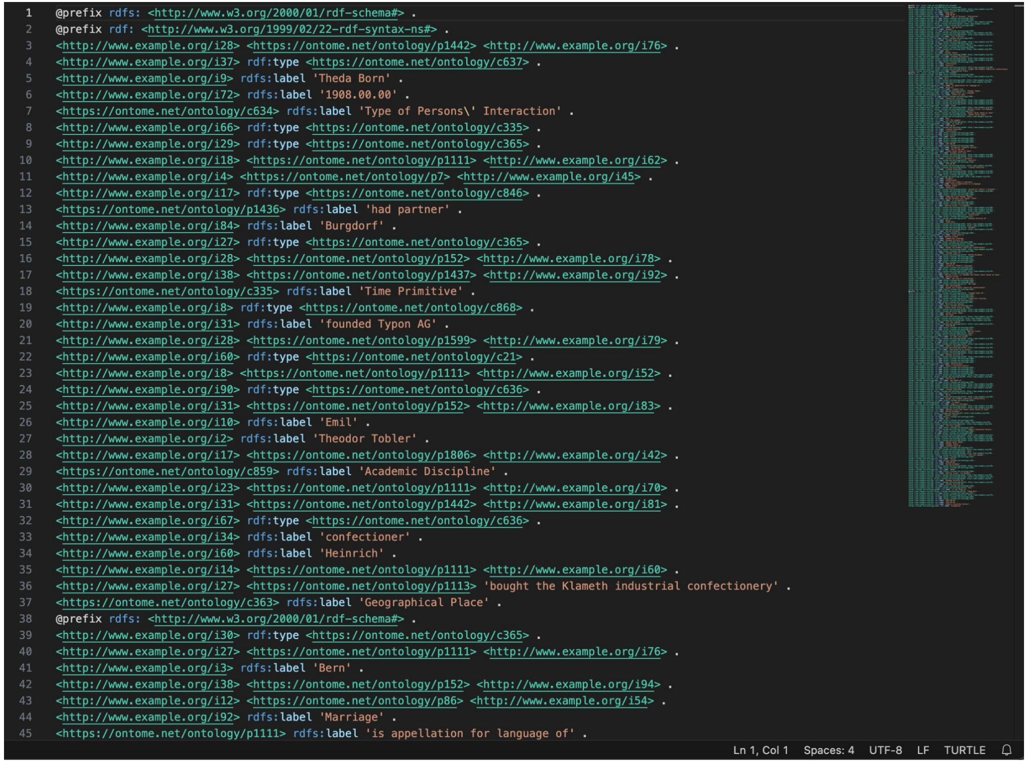




```
def person_to_graph(person: Person, graph: Graph) -> None:
    """
    Transform an object instance into a list of entities and statements.

    Args:
        person (Person): the person to be added to the graph.
        graph (Graph): the graph to add the person to.
    """

    # If the person does not have a name, set a default value
    if not person.name:
        person.name = "Unknown Person " + str(graph.get_current_index())

    # Create the person
    person_ent = graph.create_entity_aial(c.E21_person, person.name)

    # Gender
    if person.gender:
        gender = graph.create_entity_aial(
        graph.create_triple(person_ent, p.P23_hasGender, gender)

    # Origins
    if person.origins:
        geoplace = graph.create_entity_aial(c.C13_geographicalPlace, person.origins)
        graph.create_triple(person_ent, p.P24_hasItsOriginsIn, geoplace)

    # Religion
    if person.religion:
        religious_identity = graph.create_entity_aial(c.C23_religiousIdentity, person.religion)
        graph.create_triple(religious_identity, p.P36_pertainsTo, person_ent)

    # Birth date
```

Data Transformation Process: Generated Graph





```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
<http://www.example.org/i28> <https://ontome.net/ontology/p1442> <http://www.example.org/i76> .
<http://www.example.org/i37> rdf:type <https://ontome.net/ontology/c637> .
<http://www.example.org/i9> rdfs:label 'Theda Born' .
<http://www.example.org/i72> rdfs:label '1908.00.00' .
<https://ontome.net/ontology/c634> rdfs:label 'Type of Persons\' Interaction' .
<http://www.example.org/i66> rdf:type <https://ontome.net/ontology/c335> .
<http://www.example.org/i29> rdf:type <https://ontome.net/ontology/c365> .
<http://www.example.org/i18> <https://ontome.net/ontology/p1111> <http://www.example.org/i62> .
<http://www.example.org/i4> <https://ontome.net/ontology/p7> <http://www.example.org/i45> .
<http://www.example.org/i17> rdf:type <https://ontome.net/ontology/c846> .
<https://ontome.net/ontology/p1436> rdfs:label 'had partner' .
<http://www.example.org/i84> rdfs:label 'Burgdorf' .
<http://www.example.org/i27> rdf:type <https://ontome.net/ontology/c365> .
<http://www.example.org/i28> <https://ontome.net/ontology/p152> <http://www.example.org/i78> .
<http://www.example.org/i38> <https://ontome.net/ontology/p1437> <http://www.example.org/i92> .
<https://ontome.net/ontology/c335> rdfs:label 'Time Primitive' .
<http://www.example.org/i8> rdf:type <https://ontome.net/ontology/c868> .
<http://www.example.org/i31> rdfs:label 'founded Typon AG' .
<http://www.example.org/i28> <https://ontome.net/ontology/p1599> <http://www.example.org/i79> .
<http://www.example.org/i60> rdf:type <https://ontome.net/ontology/c21> .
<http://www.example.org/i8> <https://ontome.net/ontology/p1111> <http://www.example.org/i52> .
<http://www.example.org/i90> rdf:type <https://ontome.net/ontology/c636> .
<http://www.example.org/i31> <https://ontome.net/ontology/p152> <http://www.example.org/i83> .
<http://www.example.org/i10> rdfs:label 'Emil' .
<http://www.example.org/i2> rdfs:label 'Theodor Tobler' .
<http://www.example.org/i17> <https://ontome.net/ontology/p1806> <http://www.example.org/i42> .
<https://ontome.net/ontology/c859> rdfs:label 'Academic Discipline' .
<http://www.example.org/i23> <https://ontome.net/ontology/p1111> <http://www.example.org/i70> .
<http://www.example.org/i31> <https://ontome.net/ontology/p1442> <http://www.example.org/i81> .
<http://www.example.org/i67> rdf:type <https://ontome.net/ontology/c636> .
<http://www.example.org/i34> rdfs:label 'confectioner' .
<http://www.example.org/i60> rdfs:label 'Heinrich' .
<http://www.example.org/i14> <https://ontome.net/ontology/p1111> <http://www.example.org/i60> .
<http://www.example.org/i27> <https://ontome.net/ontology/p1113> 'bought the Klameth industrial confectionery' .
<https://ontome.net/ontology/c363> rdfs:label 'Geographical Place' .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
<http://www.example.org/i30> rdf:type <https://ontome.net/ontology/c365> .
<http://www.example.org/i27> <https://ontome.net/ontology/p1111> <http://www.example.org/i76> .
<http://www.example.org/i3> rdfs:label 'Bern' .
<http://www.example.org/i38> <https://ontome.net/ontology/p152> <http://www.example.org/i94> .
<http://www.example.org/i12> <https://ontome.net/ontology/p86> <http://www.example.org/i54> .
<http://www.example.org/i92> rdfs:label 'Marriage' .
<https://ontome.net/ontology/p1111> rdfs:label 'is appellation for language of' .
```

The Use Case and Demo

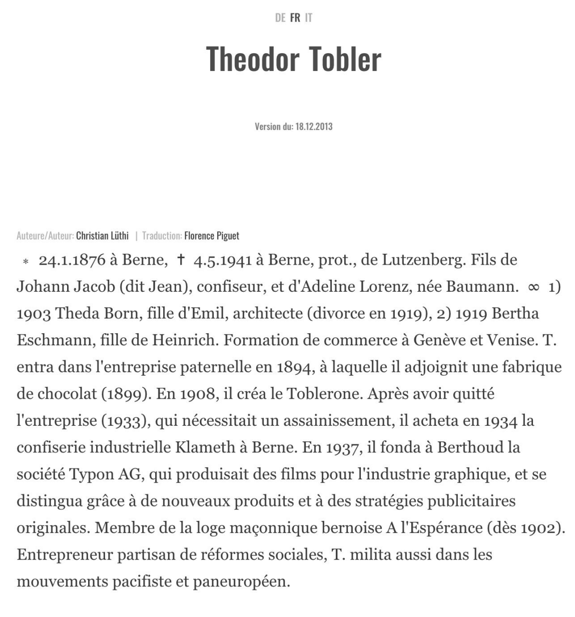
The sources

Historical Dictionnary of Switzerland:

https://hls-dhs-dss.ch/

Documenting 25394 people and 2559 families in the form of textual descriptive biographies.

The case of Theodor Tobler, inventor of the *toblerone*.



DE FR IT

Theodor Tobler

Version du: 18.12.2013

Auteure/Auteur: Christian Lüthi | Traduction: Florence Piguet

* 24.1.1876 à Berne, † 4.5.1941 à Berne, prot., de Lutzenberg. Fils de Johann Jacob (dit Jean), confiseur, et d'Adeline Lorenz, née Baumann. ∞ 1) 1903 Theda Born, fille d'Emil, architecte (divorce en 1919), 2) 1919 Bertha Eschmann, fille de Heinrich. Formation de commerce à Genève et Venise. T. entra dans l'entreprise paternelle en 1894, à laquelle il adjoignit une fabrique de chocolat (1899). En 1908, il créa le Toblerone. Après avoir quitté l'entreprise (1933), qui nécessitait un assainissement, il acheta en 1934 la confiserie industrielle Klameth à Berne. En 1937, il fonda à Berthoud la société Typon AG, qui produisait des films pour l'industrie graphique, et se distingua grâce à de nouveaux produits et à des stratégies publicitaires originales. Membre de la loge maçonnique bernoise A l'Espérance (dès 1902). Entrepreneur partisan de réformes sociales, T. milita aussi dans les mouvements pacifiste et paneuropéen.

The case of Theodor Tobler

Theodor Tobler was born on 24.1.1876 in Bern and died on 4.5.1941 in Bern, prot. of Lutzenberg. Son of Johann Jacob (known as Jean), confectioner, and Adeline Lorenz, née Baumann. 1) 1903 Theda Born, daughter of Emil, architect (divorced in 1919), 2) 1919 Bertha Eschmann, daughter of Heinrich. Commercial training in Geneva and Venice. T. joined his father's business in 1894, adding a chocolate factory in 1899. In 1908, he created Toblerone. After leaving the company (1933), which was in need of restructuring, he bought the Klameth industrial confectionery in Bern in 1934. In 1937, he founded Typon AG in Burgdorf, which produced films for the graphics industry, and made a name for itself with new products and original advertising strategies. Member of the Bernese Masonic lodge A l'Espérance (from 1902). An entrepreneur and advocate of social reform, T. was also active in the pacifist and pan-European movements.

The Geovistory Environment

https://www.geovistory.org

The Geovistory Team

LARHRA
UMR 5190

ub
b UNIVERSITÄT BERN

KleioLab

L'ÉCOLE DES HAUTES ÉTUDES EN SCIENCES SOCIALES

UNIVERSITÉ LYON III JEAN MOULIN

ÉCOLE FRANÇAISE DE ROME
Histoire, Archéologie, Sciences sociales

UNI BASEL

huygens instituut

unine
UNIVERSITÉ DE NEUCHÂTEL

UNIVERSITÉ PARIS 1 PANTHÉON SORBONNE

UNIVERSITÉ TOULOUSE Jean Jaurès

unine
UNIVERSITÉ DE NEUCHÂTEL

Universität Zürich UZH

DaSCH

hep/ haute école pédagogique vaud

UNI BASEL

UNI FR

UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Freie Universität Berlin

usi

UNIVERSITÄT GREIFSWALD
Wissen lockt. Seit 1456

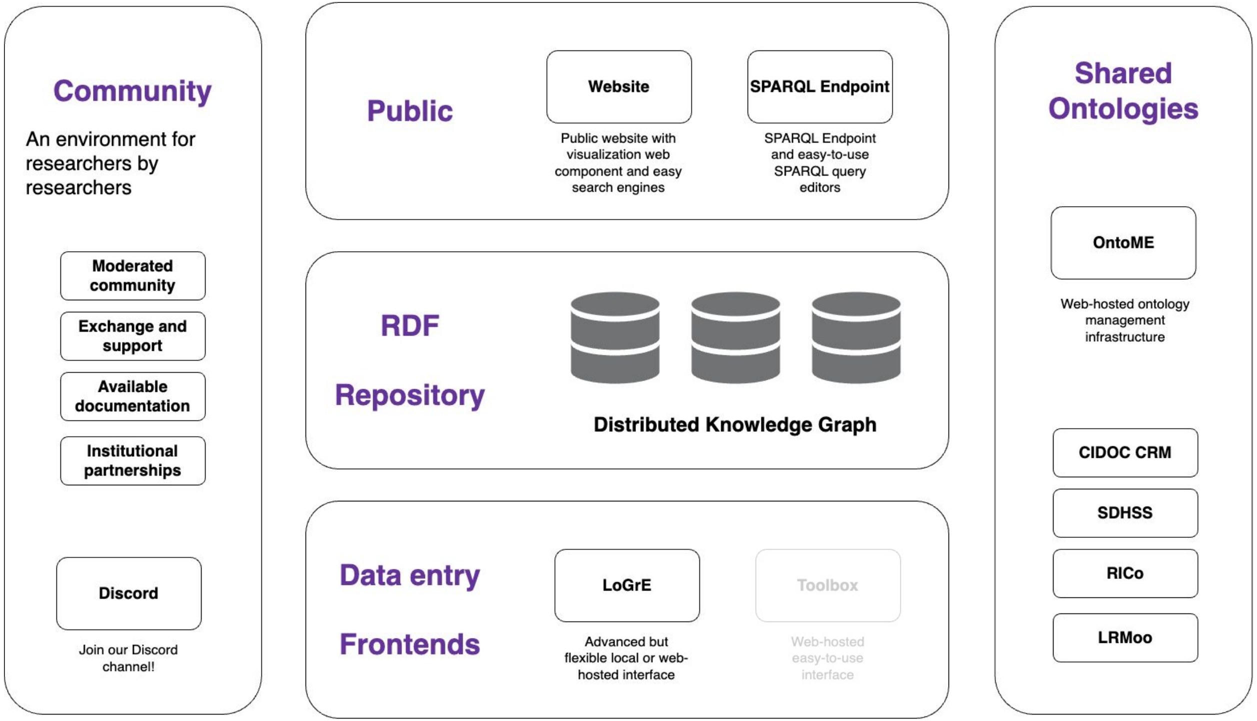
UNIVERSITÄT TRIER

abes

huygens instituut

swissuniversities

The Geovistory Environment



Community

An environment for researchers by researchers

Moderated community

Exchange and support

Available documentation

Institutional partnerships

Discord

Join our Discord channel!

Public

Website

Public website with visualization web component and easy search engines

SPARQL Endpoint

SPARQL Endpoint and easy-to-use SPARQL query editors

RDF

Repository

Distributed Knowledge Graph

Data entry

Frontends

LoGrE

Advanced but flexible local or web-hosted interface

Toolbox

Web-hosted easy-to-use interface

Shared Ontologies

OntoME

Web-hosted ontology management infrastructure

CIDOC CRM

SDHSS

RICo

LRMoo

Conclusions

Challenges

- The risk of hallucination is rather low
  - It's mitigated with the class description before transforming the assertions into triples
  - Reduced by the parsing of the text for each entity identified (trying to create all the assertions at the same time created a lot of errors), but it makes the process resource intensive
- The creation of the Python Class Description file is done manually and adapted for each data transformation process:
  - This allows the creation of a more precise Knowledge Graph, especially in line with the research questions, but requires a lot of resources on our side.
- Close monitoring on the evolution of the technologies is needed as they evolve very quickly

Next Steps

1. Make Knex more generic and reusable:
   - In order to make Knex more reusable beyond our team, we need develop process to generate Class Description files, such as with the use of SHACL ontological profiles (managed in the OntoME platform)
2. Find more efficient LLM processes:
   - Most of the LLM resources go in the multiple parsings of the text for generating the class instances. Finding ways to parse the text only once would significantly reduce the costs.
3. Integrate Knex in the graph management tool Logre:
   - Develop a more flexible and reusable GUI