Modelling Data with semantics - from Astronomy to Ethnography.

A semantic-based method for storing and sharing qualitative data in digital humanities

A. Tricoche, F. Weber, C.M. Zwölf

Research data: more issues than benefits?

Data

Issues

Methodology

Ethical

Technical

Research data: more issues than benefits?

Data

Issues

- Methodology
  - Unique definition of metadata and data meanings
  - Data production workflow documentation
  - Data citation
- Ethical
- Technical

Research data: more issues than benefits?

Data

Issues

- Methodology
  - Unique definition of metadata and data meanings
  - Data production workflow documentation
  - Data citation
- Ethical
  - Protection of confidential data
  - Reuse of data produced by third parties
  - Peer review/check & sharing
- Technical

Research data: more issues than benefits?

Data

Issues

- Methodology
  - Unique definition of metadata and data meanings
  - Data production workflow documentation
  - Data citation
- Ethical
  - Protection of confidential data
  - Reuse of data produced by third parties
  - Peer review/check & sharing
- Technical
  - Building software & hardware data infrastructure
  - Development and maintenance costs for software
  - Development and maintenance costs for the data center
  - Long term preservation of the scientific know how

Research data: more issues than benefits?

Fragmented solutions may not cope with the Data-deluge related issues

The proposed solution

- Directly deals with the three issues-categories
- Is generic and may be transposed to other disciplines

ArchEthno2020 is the operational prototype implementing the methods described hereafter.

The ArchEthno2017 experience

- Tool for archiving qualitative sociology data
  - Allowing comparison between different use-cases
  - Case study: how the family organisation is impacted by dependency of one member?
  - A two level study:
    - the case (family) for answering the scientific question, the context (medical institution) for evaluating the socio-economic environment and validation statistics on the cases.

- The tool is technically built over a relational database. The schema guarantees:
  - Space for highlighting observers reactions
  - To build the scientific use-case with the specific metadata
  - To separate the use-case from the context
  - To protect the knowledge processing workflow

Adapting the existing tool to new use-cases is difficult and time consuming

- Change the database schema
- Change the data-entry and the data-extraction interfaces

Toward a new approach

Is the context for Astronomy & Ethnography interaction

- Multidisciplinary Research University
- Federated actions for solving research-data linked issues
  - Data documenting & sharing are common issues.

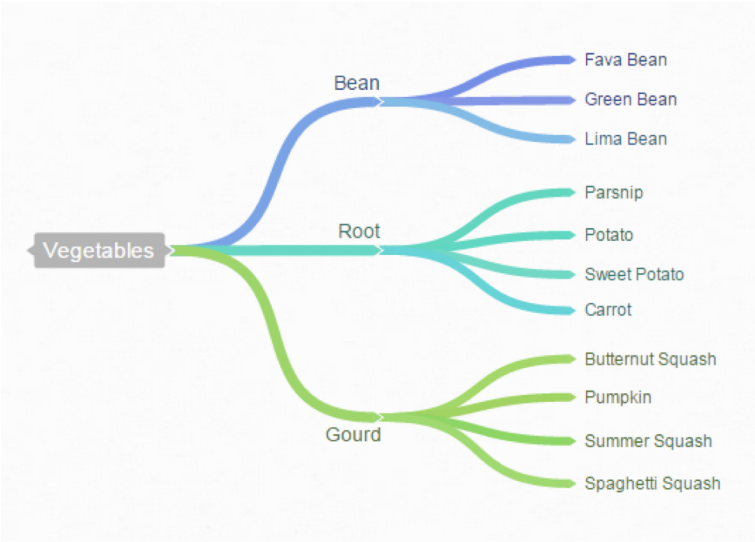
Toward a new approach

PSL UNIVERSITÉ PARIS

Is the context for Astronomy & Ethnography interaction

- Multidisciplinary Research University
- Federated actions for solving research-data linked issues
  - Data documenting & sharing are common issues.

Inspired by the semantic web and based on SKOS
(W3C **S**tandard **S**imple **Knowledge** **O**rganisation **S**ystem, SKOS)



Toward a new approach

Is the context for Astronomy & Ethnography interaction

- Multidisciplinary Research University
- Federated actions for solving research-data linked issues
  - Data documenting & sharing are common issues.

Inspired by the semantic web and based on SKOS
(W3C **S**tandard **S**imple **Knowledge** **O**rganisation **S**ystem, SKOS)

One define the concepts linked with research data and metadata

These concepts are kept into a *thesaurus.*

The *thesaurus* is serialized in a **SKOS/RDF** (XML) file

Describing the approach

SKOS file

Sustainability: The scientific know-how is preserved.
Inequivocability: unambiguous definitions are given *una tantum*.

Flexibility : The conceptual model is imbedded into the SKOS model, not into a SQL schema

Describing the approach

SKOS file

Sustainability: The scientific know-how is preserved.
Inequivocability: unambiguous definitions are given *una tantum*.

Flexibility : The conceptual model is imbedded into the SKOS model, not into a SQL schema

With the SKOS model, a simple table may fit for all the data-types.

SKOS	Value
Concept « Name of the author»	Gustave
Concept «Surname of the author »	Flaubert

Evolutions in the conceptual model (e.g. new concepts) won't have consequence on the table structure and already existing content.

Describing the approach

SKOS file

Sustainability: The scientific know-how is preserved.
Inequivocability: unambiguous definitions are given *una tantum*.

Flexibility : The conceptual model is imbedded into the SKOS model, not into a SQL schema

With the SKOS model, a simple table may fit for all the data-types.

Persistent unique Identifier	SKOS	Value
7a8a34f4-a7aa-11e7-abc4	Concept « Name of the author»	Gustave
a975361a-a7aa-11e7-abc4	Concept «Surname of the author »	Flaubert

Evolutions in the conceptual model (e.g. new concepts) won't have consequence on the table structure and already existing content.

Describing the approach

SKOS file

Sustainability: The scientific know-how is preserved.
Inequivocability: unambiguous definitions are given *una tantum*.

Flexibility : The conceptual model is imbedded into the SKOS model, not into a SQL schema

With the SKOS model, a simple table may fit for all the data-types.

Persistent unique Identifier	SKOS	Value
7a8a34f4-a7aa-11e7-abc4	Concept « Name of the author»	Gustave
a975361a-a7aa-11e7-abc4	Concept «Surname of the author »	Flaubert

One may cite with a fine grained granularity data using PIDs.

Describing the approach

SKOS file

Sustainability: The scientific know-how is preserved.
Inequivocability: unambiguous definitions are given *una tantum*.

Flexibility : The conceptual model is imbedded into the SKOS model, not into a SQL schema

With the SKOS model, a simple table may fit for all the data-types.

Persistent unique Identifier	SKOS	Value	Confidential level
7a8a34f4-a7aa-11e7-abc4	Concept « Name of the author»	Gustave	Public
a975361a-a7aa-11e7-abc4	Concept «Surname of the author »	Flaubert	Private

One may cite with a fine grained granularity data using PIDs.

Describing the approach

SKOS file

Sustainability: The scientific know-how is preserved.
Inequivocability: unambiguous definitions are given *una tantum*.

Flexibility : The conceptual model is imbedded into the SKOS model, not into a SQL schema

With the SKOS model, a simple table may fit for all the data-types.

Persistent unique Identifier	SKOS	Value	Confidential level
7a8a34f4-a7aa-11e7-abc4	Concept « Name of the author»	Gustave	Public
a975361a-a7aa-11e7-abc4	Concept «Surname of the author »	Flaubert	Private

One may cite with a fine grained granularity data using PIDs.

Sensitive data are encrypted and may be displayed only by trusted users:

- ✓ Confidentiality
- ✓ Scientist may secure their own work
- ✓ Legal aspect are covered.

Describing the approach

SKOS file

Sustainability: The scientific know-how is preserved.
Inequivocability: unambiguous definitions are given *una tantum*.

Flexibility : The conceptual model is imbedded into the SKOS model, not into a SQL schema

- The XML file contain all the scientific know-how.
- This file is used for building dynamically the data-entry and data extraction interfaces, from generic piece of software

Ad-hoc

Generic software layer

Ad hoc GUIs

Describing the approach

SKOS file

Sustainability: The scientific know-how is preserved.
Inequivocability: unambiguous definitions are given *una tantum*.

Flexibility : The conceptual model is imbedded into the SKOS model, not into a SQL schema

- The XML file contain all the scientific know-how.
- This file is used for building dynamically the data-entry and data extraction interfaces, from generic piece of software
  - A unique software library may meet many communities needs
    - Software development process is shortened
    - Technical know-how shared trans-disciplinary
    - Technical architecture derived directly by the software solution.

Describing the approach

SKOS file

Sustainability: The scientific know-how is preserved.
Inequivocability: unambiguous definitions are given *una tantum*.

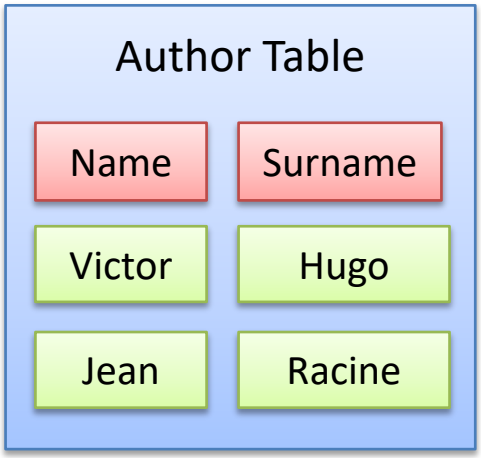
Flexibility : The conceptual model is imbedded into the SKOS model, not into a SQL schema

Inspired by the **Paramater Description Language** approach and the **PDL** software framework.

- The XML file contain all the scientific know-how.
- This file is used for building dynamically the data-entry and data extraction interfaces, from generic piece of software

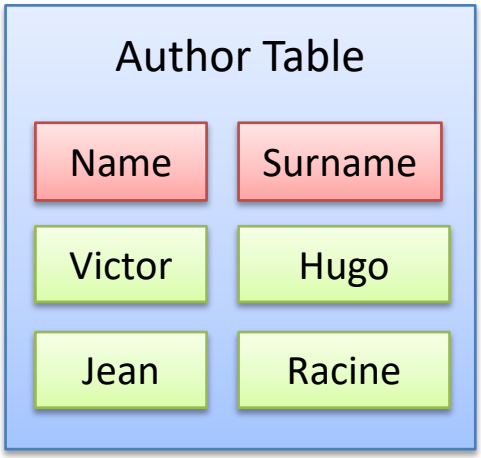
- A unique software library may meet many communities needs
  - Software development process is shortened
  - Technical know-how shared trans-disciplinary
  - Technical architecture derived directly by the software solution.

From relational DB to concepts – A working example



Let us consider a simple SQL table for storing information about authors.

From relational DB to concepts – A working example



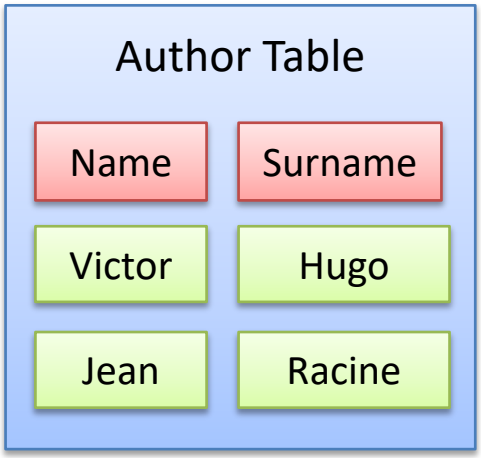
We would like to add further information:

- The birth date
- The death date (for death authors)
- If alive authors, the date of the last publication.

The SQL approach is not flexible and data modelling is not optimal

- One has to define a “death date” column also for alive authors – nonsense.
- The schema has to change and the already stored data may be corrupted meanwhile.
- This situation is even worst if multiple tables & foreign keys are involved.

From relational DB to concepts – A working example



We would like to add further information:

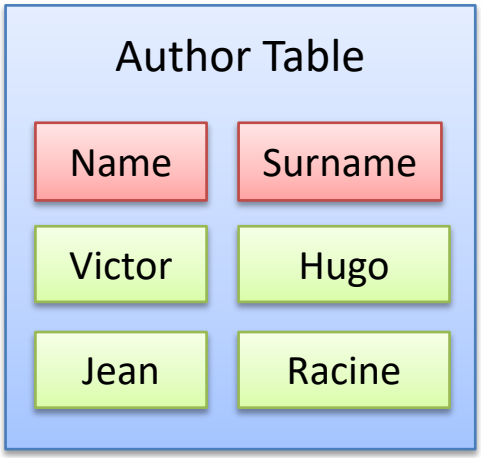
- The birth date
- The death date (for death authors)
- If alive authors, the date of the last publication.

The SQL approach is not flexible and data modelling is not optimal

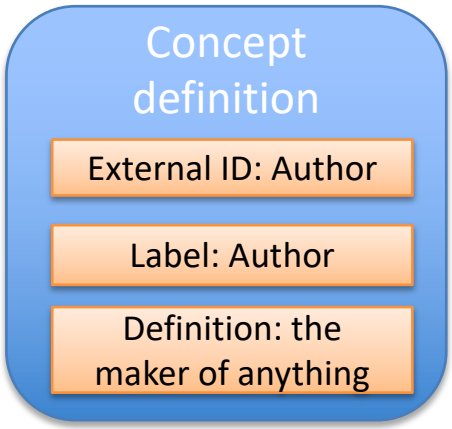
- One has to define a “death date” column also for alive authors – nonsense.
- The schema has to change and the already stored data may be corrupted meanwhile.
- This situation is even worst if multiple tables & foreign keys are involved.

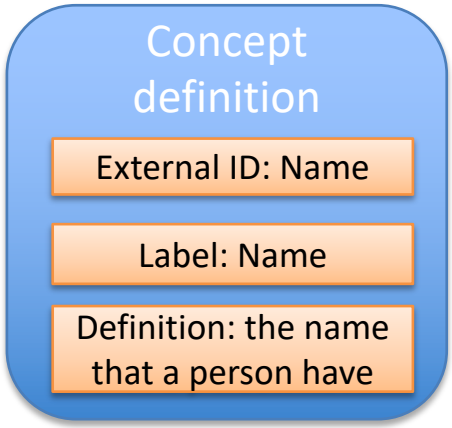
Let us see how SKOS concepts may help us in this particular case…

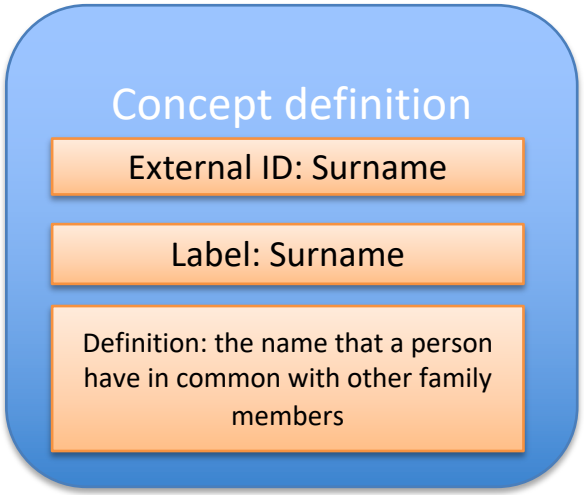
From relational DB to concepts – A working example



We define three concepts:







From relational DB to concepts – A working example

Author Table

Name	Surname
Victor	Hugo
Jean	Racine

Author Table (Skos based version)

Entry_ID	ID Concept	Value	Parent entry ID
1	Author		
2	Name	Victor	1
3	Surname	Hugo	1
4	Author		
5	Name	Jean	4
6	Surname	Racine	4

We define three concepts:

Concept definition

- External ID: Author
- Label: Author
- Definition: the maker of anything

Concept definition

- External ID: Name
- Label: Name
- Definition: the name that a person have

Concept definition

- External ID: Surname
- Label: Surname
- Definition: the name that a person have in common with other family members

From relational DB to concepts – A working example

Author Table

Name	Surname
Victor	Hugo
Jean	Racine

Author Table (Skos based version)

Entry_ID	ID Concept	Value	Parent entry ID
1	Author		
2	Name	Victor	1
3	Surname	Hugo	1
4	Author		
5	Name	Jean	4
6	Surname	Racine	4

For adding the additional required information (dates of birth, death, last activity) we define three additional concepts:

Concept definition

External ID: BirthDate

Label: BirthDate

Definition: the date of birth in format jj/mm/aaaa (Gregorian Calendar)

Concept definition

External ID: DeathDate

Label: DeathDate

Definition: the date of death in format jj/mm/aaaa (Gregorian Calendar)

Concept definition

External ID: LastActivityDate

Label: LastActivityDate

Definition: the date of last activity in format jj/mm/aaaa (Gregorian Calendar)

From relational DB to concepts – A working example

Concept definition

External ID: DeathDate

Label: DeathDate

Definition: the date of death in format jj/mm/aaaa (Gregorian Calendar)

Concept definition

External ID: BirthDate

Label: BirthDate

Definition: the date of birth in format jj/mm/aaaa (Gregorian Calendar)

Concept definition

External ID: LastActivityDate

Label: LastActivityDate

Definition: the date of last activity in format jj/mm/aaaa (Gregorian Calendar)

Author Table (Skos based version)

Entry_ID	ID Concept	Value	Parent entry ID
1	Author		
2	Name	Victor	1
3	Surname	Hugo	1
4	Author		
5	Name	Jean	4
6	Surname	Racine	4
7	BirthDate	26/02/1802	1
8	DeathDate	22/05/1885	1
9	BirthDate	22/12/1639	4
10	DeathDate	21/04/1699	4

From relational DB to concepts – A working example

- The proposed example may be easily generalized to more complex configurations.
- We adopted this method for migrating all the data from the ArchEthno2017 SQL structure to the actual ArchEthno2020 SKOS version
- A quick demonstration…

Author Table (Skos based version)

Entry_ID	ID Concept	Value	Parent entry ID
1	Author		
2	Name	Victor	1
3	Surname	Hugo	1
4	Author		
5	Name	Jean	4
6	Surname	Racine	4
7	BirthDate	26/02/1802	1
8	DeathDate	22/05/1885	1
9	BirthDate	22/12/1639	4
10	DeathDate	21/04/1699	4

A word about confidentiality

- From the technical point of view confidentiality level is an Integer N.
  - Any user whose accreditation level is greater than N may display the information
  - If accreditation level is smaller than N, data are not conveyed to user.

- Who may decide (and how)
  - the confidentiality level for each specific datum?
  - The access write for each user?

- We have started thinking about an ethical committee for taking these decisions
  - Are other data-practitioners having similar issues?

N.B.

- Data protection does not imply to remove confidential data
- Data discovery and data-processing tools related to confidential data should not be accessible for users with no sufficient permission.

Concluding remarks

The described approach meets the main

- methodological
- ethical
- technical

Issues experienced by teams working with research data.

The proposed method is based on semantic web

- And may be adopted by other communities
- Reducing development/maintenance costs
- Compliant with the RDA registry type (CF. Registry Data Type WG and Data Fabric IG outputs)

Thanks to José Sastre, to Jean-Robert Dantou, to Maxime Tissier who contributed to this work.