Heterogeneous datasets
A tale of integration and exploration

Nelly Barret

Postdoctoral researcher
Data Science group
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano

January 24, 2025

Short bio

My CS background:

- Bachelor @ Univ. Lyon
- Master, AI track @ Univ. Lyon
- PhD @ Inria Saclay and Ecole Polytechnique
- Post-doc @ Politecnico di Milano (Italie)

My thesis was about **user-oriented exploration of semi-structured data**.
My post-doc is about **enabling federating analyses of health data**.

Université Claude Bernard Lyon 1

Inria

POLITECNICO MILANO 1863

Outline

1. Motivation: data integration and exploration problems
2. PhD: exploring unknown semi-structured datasets
3. Post-doc: healthcare analytics across hospitals
4. Systems developed
5. Conclusion

Outline

Different settings, different needs

Structured data models:

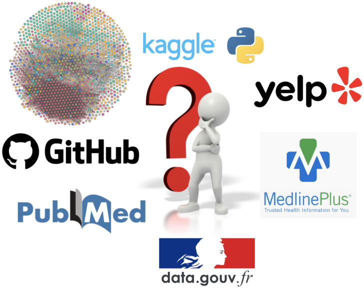
- **Tables**
- **Relational** databases

Semi-structured data models:

- **XML** documents
- **JSON** documents
- **RDF** graphs
- **Property** graphs

Unstructured data models:

- **Text**
- **Images**



Different settings, different needs

Various domains:

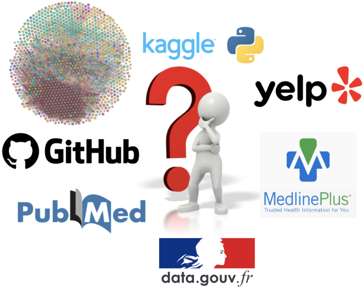
- Health
- Journalism
- Transports, ...

Sensitivity levels:

- Enforce privacy rules
- EU GDPR rules

Several actors/users:

- Different skills
- Time/money constraints



Different settings, different needs

Various domains:

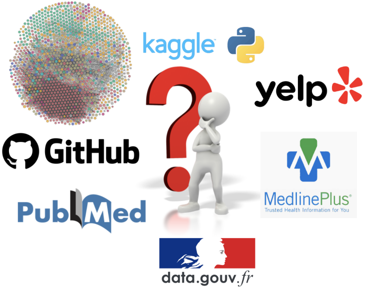
- Health
- Journalism
- Transports, ...

Sensitivity levels:

- Enforce privacy rules
- EU GDPR rules

Several actors/users:

- Different skills
- Time/money constraints



Dataset integration and exploration is hard: large, complex, irregular

Outline

What does the dataset describe?

- Real-world objects and relationships between them

What does the dataset describe?

- Real-world objects and relationships between them
- Traditional setting: Entity-Relationship models [RG03]

What does the dataset describe?

- Real-world objects and relationships between them
- Traditional setting: Entity-Relationship models [RG03]
- Need to compute them from the dataset!

What does the dataset describe?

```xml
<person id="person1">
  <name>Alice</name>
  <address>
    <street>2, Second Street</street>
    <province>Georgia</province>
    <country>USA</country>
  </address>
  <mailbox>
    <mail from="person1@test.fr" to="person2@test.fr">
      <parlist>
        <listitem><text>Task 1</text></listitem>
        <listitem>
          <parlist>
            <listitem><text>Sub task 1</text></listitem>
            <listitem><text>Sub task 2</text></listitem>
            <listitem><text>Sub task 3</text></listitem>
          </parlist>
        </listitem>
      </parlist>
    </mail>
```

- Real-world objects and relationships between them
- Traditional setting: Entity-Relationship models [RG03]
- Need to compute them from the dataset!
- What about semi-structured data models (nesting)?

What does the dataset describe?

```xml
<person id="person1">
  <name>Alice</name>
  <address>
    <street>2, Second Street</street>
    <province>Georgia</province>
    <country>USA</country>
  </address>
  <mailbox>
    <mail from="person1@test.fr" to="person2@test.fr">
      <parlist>
        <listitem><text>Task 1</text></listitem>
        <listitem>
          <parlist>
            <listitem><text>Sub task 1</text></listitem>
            <listitem><text>Sub task 2</text></listitem>
            <listitem><text>Sub task 3</text></listitem>
          </parlist>
        </listitem>
      </parlist>
    </mail>
```

- Real-world objects and relationships between them
- Traditional setting: Entity-Relationship models [RG03]
- Need to compute them from the dataset!
- What about semi-structured data models (nesting)?
- Keep it simple and of controllable size

Thesis problem and research contributions

Thesis problem statement

How to facilitate **user exploration** of **unknown heterogeneous semi-structured datasets**?

Thesis problem and research contributions

Thesis problem statement

How to facilitate **user exploration** of **unknown heterogeneous semi-structured datasets**?

Abstra: semi-structured data overviews [BMU22, BMU24] 

- Automatically compute lightweight Entity-Relationship diagrams
- Ideal for first-sight dataset discovery

Thesis problem and research contributions

Thesis problem statement

How to facilitate **user exploration** of **unknown heterogeneous semi-structured datasets**?

Abstra: semi-structured data overviews [BMU22, BMU24]

- Automatically compute lightweight Entity-Relationship diagrams
- Ideal for first-sight dataset discovery

PathWays: interesting Named Entity connections [BGLM23b, BGLM23a, BGLM25]

- Compute and rank entity paths in and across datasets
- Ideal for exploring connections within and across datasets

Related work

Data summarization

- Structural
  - Quotient [GGM20, KC10, MS99] (the one we adopt to build $\mathcal{G}$)
  - Non-quotient [GW97]
- Pattern mining [ZLVK16]
- Statistical [HS12]
- Hybrid [RGSB17]

Schema inference

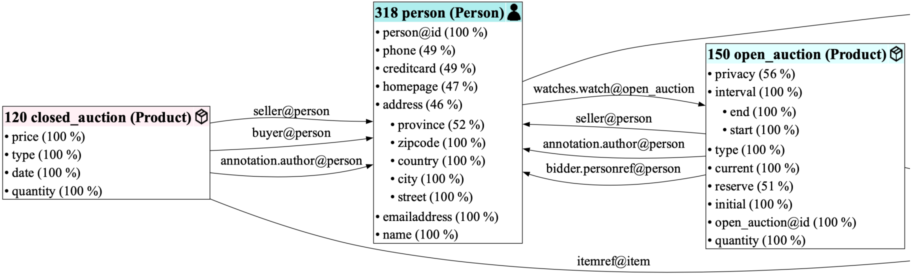
- XML [CGS11]
- JSON [BCGS19]
- RDF [GLSW22]
- PG [LBH21]

- Data summarization and schema inference are tied to one data model
- Schemas are often not suited to NTUs

A JSON schema from social network data using [BCGS19]

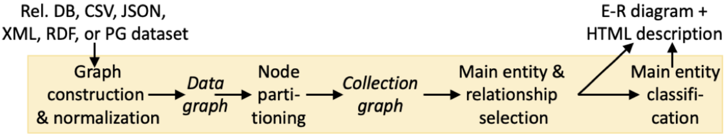

```
▾ __Content:
  ▾ _id:
    ▾ __Content:
      ▾ $oid:
          __Kind:                                   "StrType"
      __Kind:                                       "Record"
  ▾ code:
      __Kind:                                       "NumType"
  ▾ event:
    ▾ __Content:
      ▾ 0:
        ▾ __Content:
          ▾ action:
              __Kind:                               "StrType"
          ▾ attachments:
            ▾ __Content:
              ▾ __Content:
                ▾ 0:
                  ▾ __Content:
                    ▾ audio:
                      ▾ __Content:
                        ▾ 0:
                          ▾ __Content:
                            ▾ album_id:
                                __Kind:             "NumType"
                            ▾ artist:
                                __Kind:             "StrType"
                            ▾ content_restricted:
                                __Kind:             "NumType"
                            ▾ date:
                                __Kind:             "NumType"
                            ▾ duration:
                                __Kind:             "NumType"
                            ▾ genre_id:
                                __Kind:             "NumType"
                            ▾ id:
                                __Kind:             "NumType"
                            ▾ lyrics_id:
                                __Kind:             "NumType"
                            ▾ owner_id:
                                __Kind:             "NumType"
```

What does the dataset describe?



The Abstra approach

1. Integrate all data sources in a graph (ConnectionLens) [ABC+22]
2. Summarize the graph
3. Among summary nodes, identify entities and their attributes
4. In the summary, identify relationships between the entities
5. Propose a simple category to each entity (best-effort)



Background: from heterogeneous data to data graphs

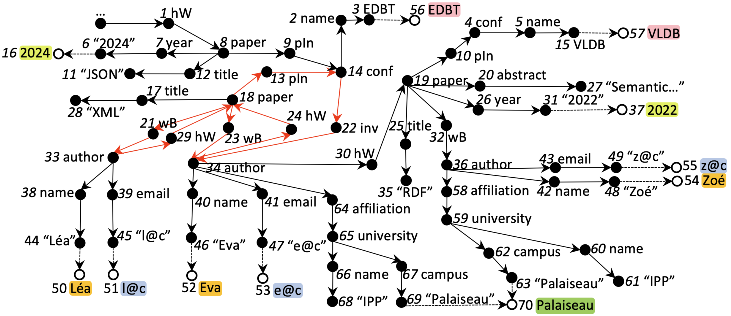
ConnectionLens [ABC+22]:

1. Ingests any dataset into a **directed graph**
    - Generic, flexible, fine granularity

Background: from heterogeneous data to data graphs

ConnectionLens [ABC+22]:

1. Ingests any dataset into a **directed graph**
   - Generic, flexible, fine granularity
2. Extracts **Named Entities** (NEs) from all text nodes
   - date, email address, People, Place, Organization, ...



Data graph summarization

We need a **compact representation of large data graphs**

Data graph summarization

We need a **compact representation of large data graphs**

Challenges:

- Heterogeneous graphs originate from different data models
- Node and/or edge labels may be empty

Data graph summarization

We need a **compact representation of large data graphs**

Challenges:

- Heterogeneous graphs originate from different data models
- Node and/or edge labels may be empty

We aim for a **quotient graph summary**:

- Based on **equivalence** between nodes of the original graph
- We prefer **small summaries** (number of nodes)

Quotient summarization across data models

Each data model has its own syntax:

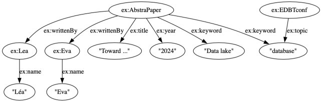
XML

```xml
<root>
    <paper id="p1" writtenBy="a1,a2">
        <title>Toward ...</title>
        <year>2024</year>
        <keyword>Data lake</keyword>
        <keyword>database</keyword>
    </paper>
    <author id="a1">
        <name>Léa</name>
        <affiliation>...</affiliation>
    </author>
    <author id="a2">
        <name>Eva</name>
    </author>
</root>
```

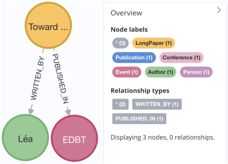
JSON

```json
{
    "paper": {
        "title": "Toward ...",
        "year": 2024,
        "keyword": ["Data lake", "database"],
        "writtenBy": {
            "author": {
                "name": "Léa",
                "affiliation": {}
            }
        }
    }
}
```

RDF



PG



Summarization based on same-kind nodes

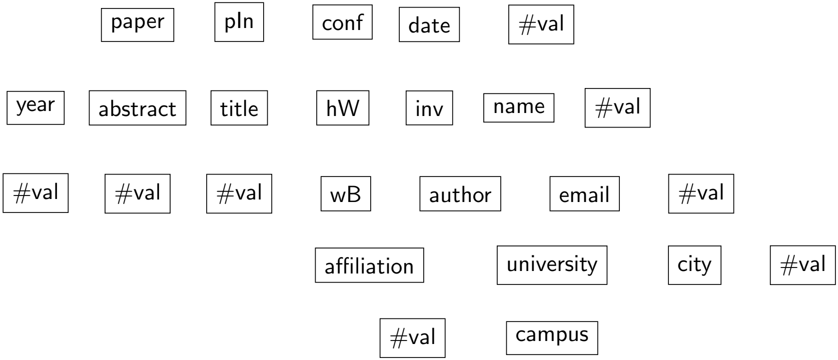
We identify **node kinds** in each model based on the respective best practices for data design:

- XML: elements with the same **label** (or type)
- JSON: nodes on the same **path from the root**
- RDF [GGM20]: depending on **node type(s)** or, if absent, **incoming and outgoing properties**
- PG: adaptation of the above [GGM20]

We obtain a **partition** over the graph: a set of equivalence classes

The summary (collection graph) $\mathcal{G}$

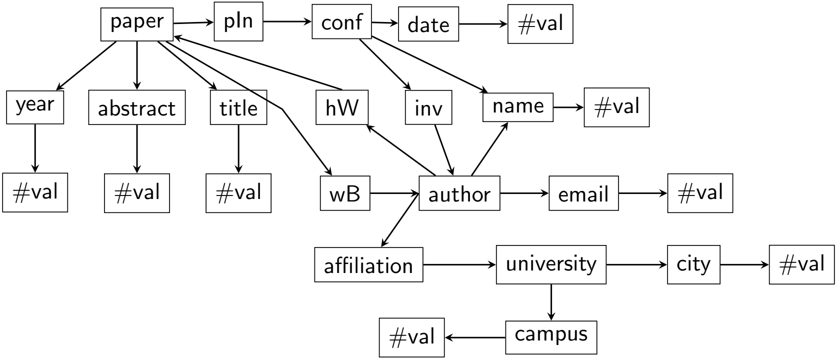
Collection node for each equivalence class



The summary (collection graph) $\mathcal{G}$

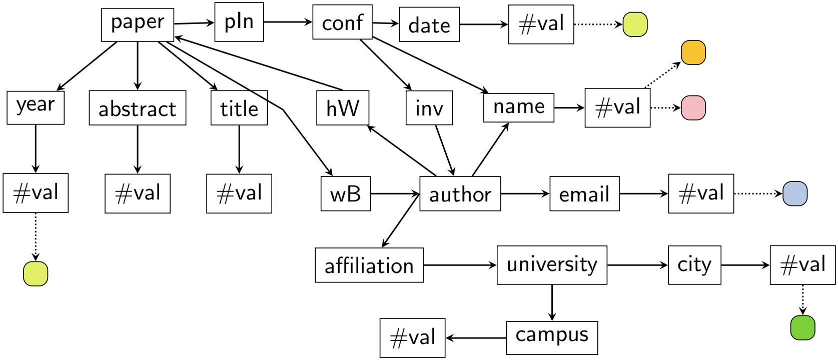
Collection node for each equivalence class

Collection edge $C_s \to C_t$ if a data edge exists

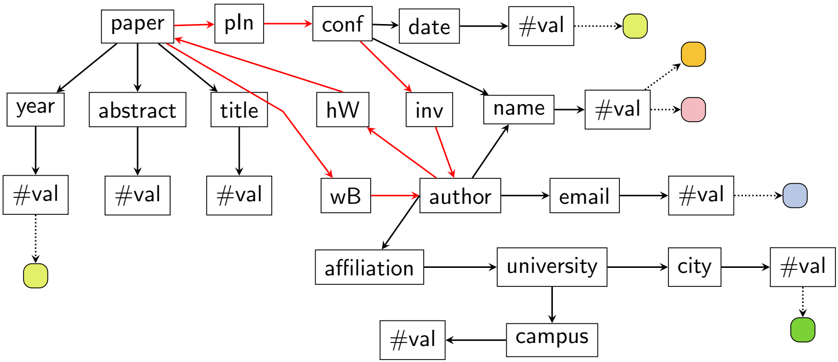


The summary (collection graph) $\mathcal{G}$

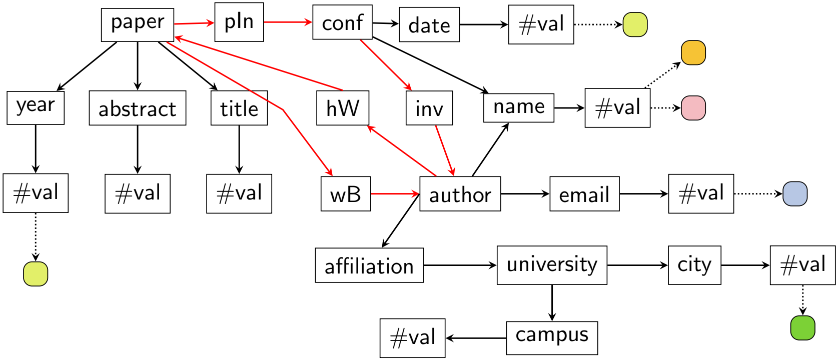
Collection node for each equivalence class
Collection edge $C_s \rightarrow C_t$ if a data edge exists
Entity profile for each **leaf collection node**: reflects NEs in the leaves



Identifying entities in the collection graph $\mathcal{G}$

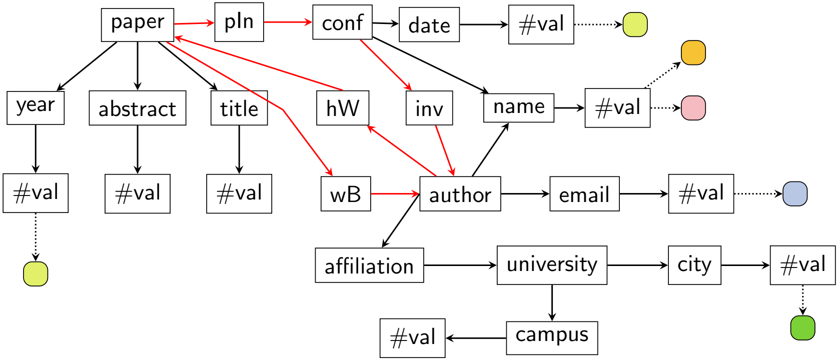


Identifying entities in the collection graph $\mathcal{G}$



Which collections represent entities in the E-R diagram?

Identifying entities in the collection graph $\mathcal{G}$



Which collections represent entities in the E-R diagram?

Which collections represent entity attributes?

Requirements and algorithm

We need an algorithm to identify entity roots and attributes for the E-R diagram

- For complex, potentially cyclic, collection graphs

Requirements and algorithm

We need an algorithm to identify entity roots and attributes for the E-R diagram

- For complex, potentially cyclic, collection graphs

Greedy selection of few entities in $\mathcal{G}$

1. Assign a score to each collection node
2. While less than E_{max} entity roots, or data coverage $< cov_{min}$
    1. Elect the next highest-scored eligible collection node as an entity root
    2. Compute its boundary (set of attributes)
    3. Update the collection graph to reflect the selection of an entity
    4. Recompute the scores

How to score a collection node?

Reflect the weight of this node and its structure in the dataset

1. w_{desc_k}, w_{leaf_k}: # descendants, leaf descendants, at depth k

How to score a collection node?

Reflect the weight of this node and its structure in the dataset

1. w_{desc_k}, w_{leaf_k}: # descendants, leaf descendants, at depth k
   - ⊗ Not clear how to pick k

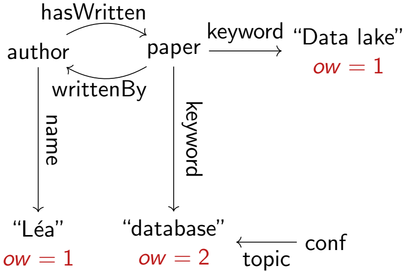
How to score a collection node?

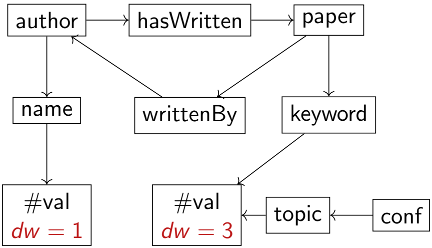
Reflect the weight of this node and its structure in the dataset

1. w_{desc_k}, w_{leaf_k}: # descendants, leaf descendants, at depth k
2. Directed Acyclic Graph (DAG) rooted in each node: w_{DAG}

Data weight

Own weight *ow* of a leaf node: its in-degree
Data weight *dw* of a leaf collection node: the sum of its nodes' *ow*

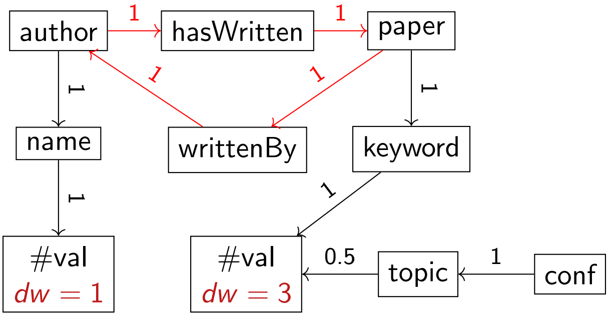




Data weight DAG propagation

Leaf collection *dw* is propagated back to all ancestors which are not in a cycle

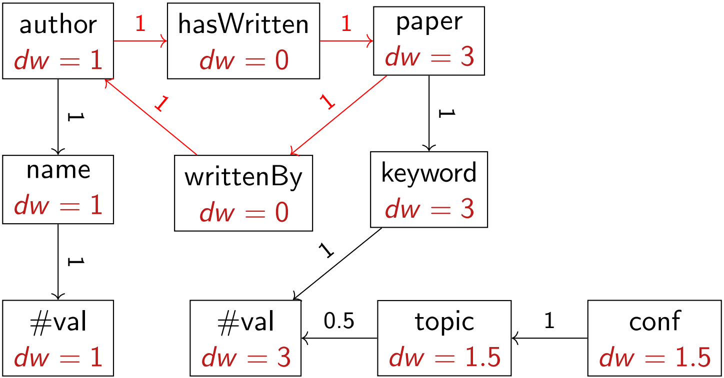
- **Edge transfer factor:** $\frac{|\text{nodes in } C_t \text{ having a parent in } C_s|}{|C_t|}$



Data weight DAG propagation

Leaf collection *dw* is propagated back to all ancestors which are not in a cycle

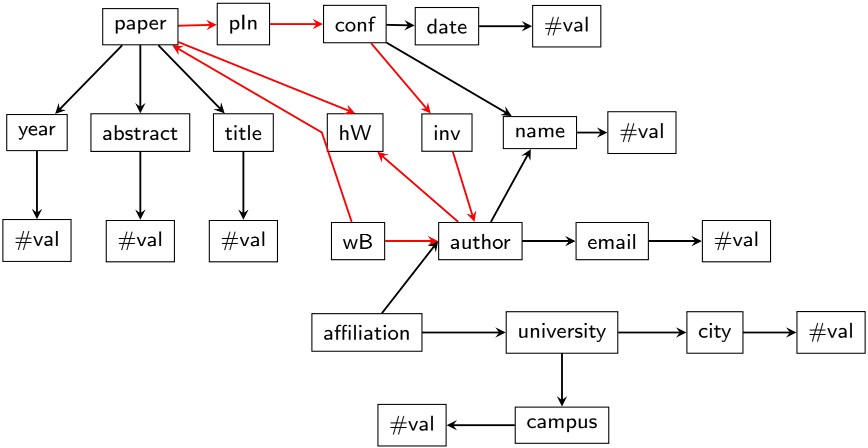
- **Edge transfer factor:** $\frac{|\text{nodes in } C_t \text{ having a parent in } C_s|}{|C_t|}$



How to score a collection node?

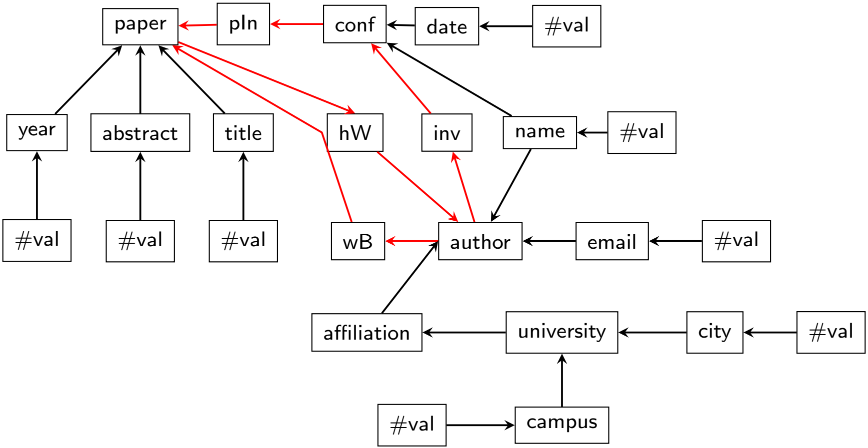
1. w_{desc_k}, w_{leaf_k}: # descendants, leaf descendants, at depth k
2. Directed Acyclic Graph (DAG) rooted in each node: w_{DAG}
3. $w_{PageRank}$: PageRank algorithm on $\mathcal{G}$

PageRank score of a collection graph node



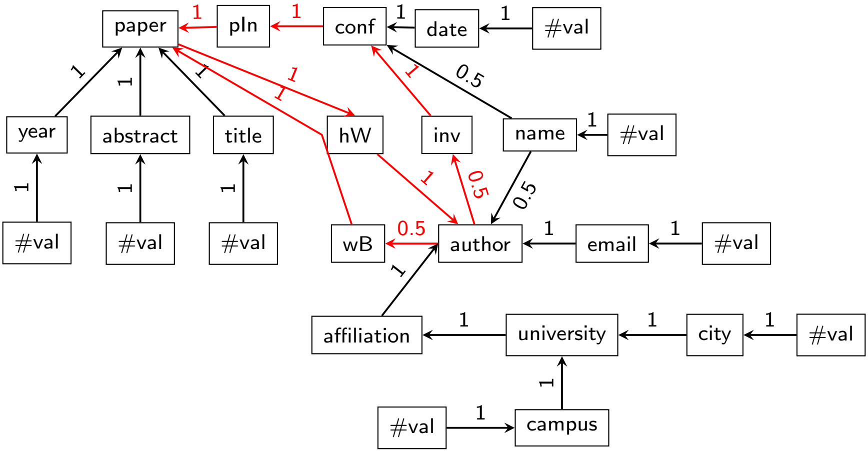
The collection graph $\mathcal{G}$

PageRank score of a collection graph node



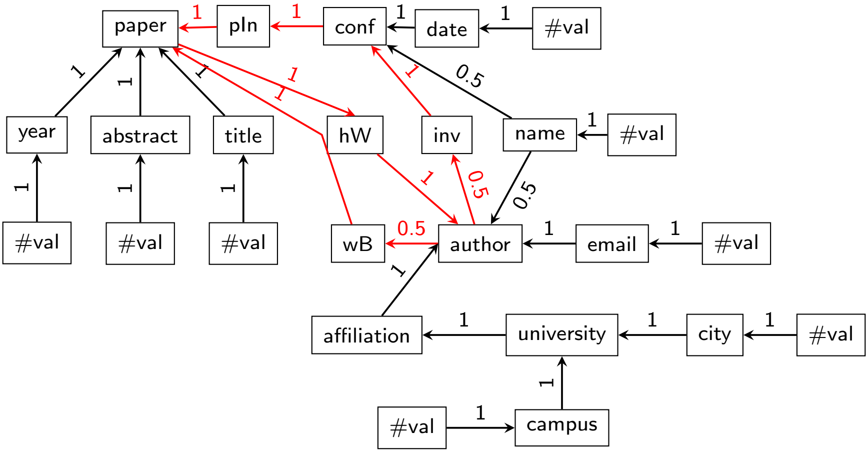
The reverse collection graph $\mathcal{G}_R$

PageRank score of a collection graph node



The reverse collection graph $\mathcal{G}_R$ with PR edge weights

PageRank score of a collection graph node



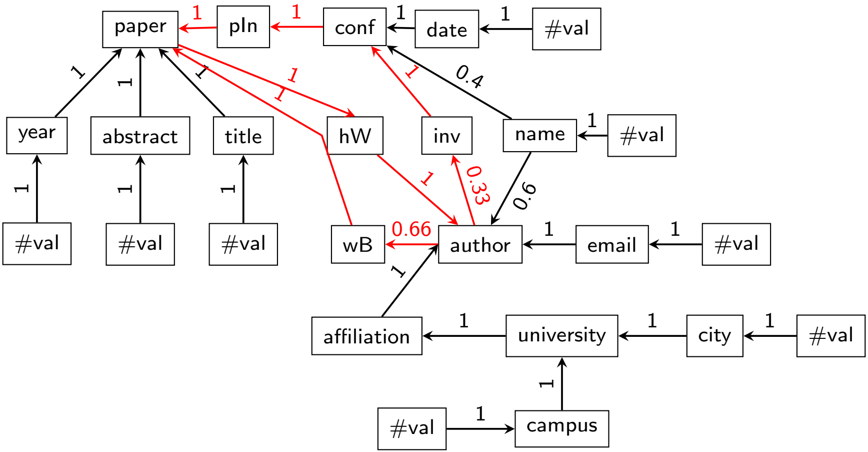
The reverse collection graph $\mathcal{G}_R$ with PR edge weights

Collections distribute their score based solely on their connectivity

How to score a collection node?

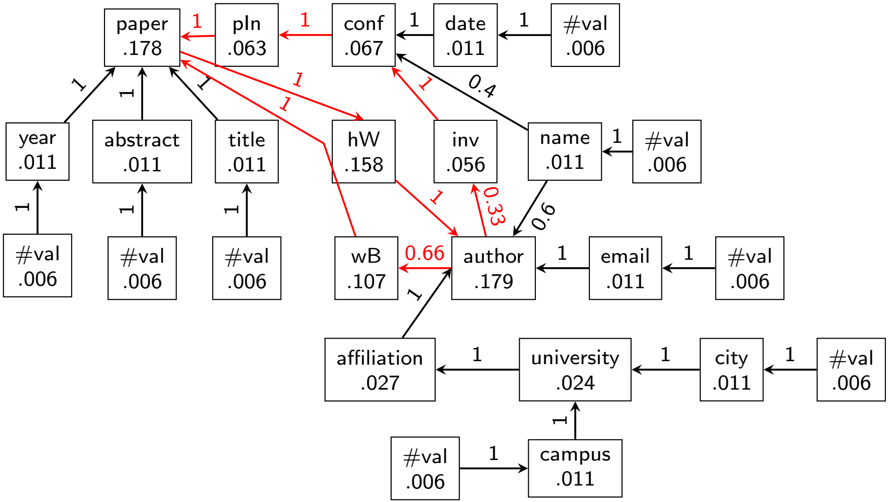
1. w_{desc_k}, w_{leaf_k}: # descendants, leaf descendants, at depth k
2. w_{DAG}: dw bottom-up propagation on $\mathcal{G}$ (outside cycles)
3. $w_{PageRank}$: PageRank algorithm on $\mathcal{G}$
4. $w_{dwPageRank}$: PageRank algorithm on $\mathcal{G}$ with dw-tuned PR edge weights
   - ✓ Reflects both the topology and where actual data is

The data-weighted PageRank score

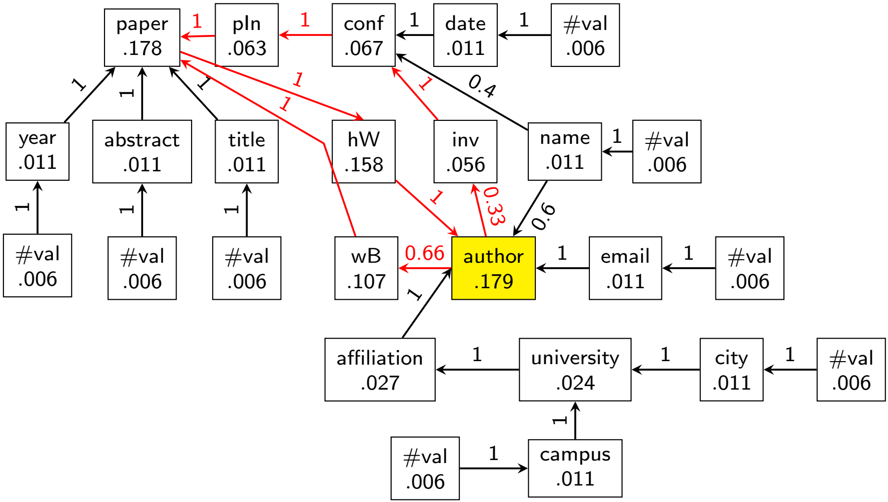


The reverse collection graph $\mathcal{G}_R$ with *dw*-tuned PR edge weights

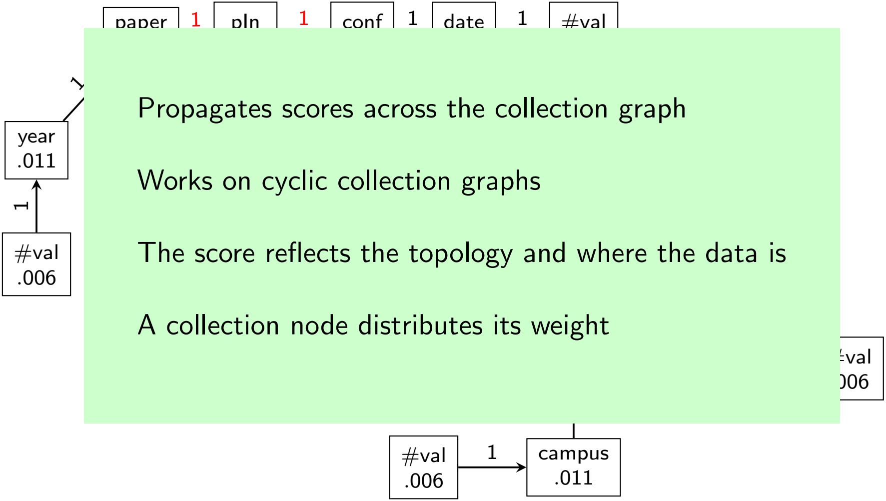
The data-weighted PageRank score



The data-weighted PageRank score



The data-weighted PageRank score



Propagates scores across the collection graph

Works on cyclic collection graphs

The score reflects the topology and where the data is

A collection node distributes its weight

How to compute an entity boundary?

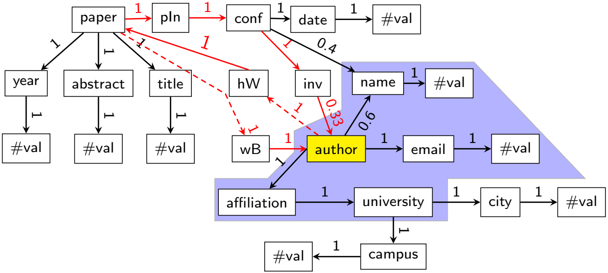
Collections in $\mathcal{G}$ representing attributes of this entity

How to compute an entity boundary?

Collections in $\mathcal{G}$ representing attributes of this entity

"Those that contribute to the entity's weight"

- The boundary may go far (for deep-structure entities)
- Easy to define for w_{desc_k}, w_{leaf_k}, w_{DAG}. Example for w_{desc_2}

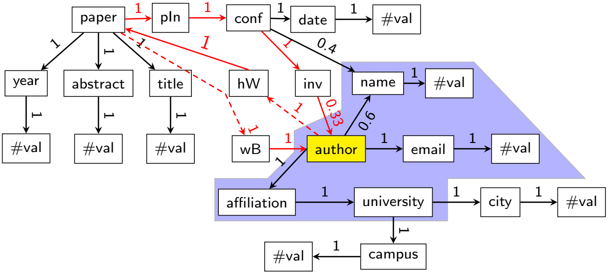


How to compute an entity boundary?

Collections in $\mathcal{G}$ representing attributes of this entity

"Those that contribute to the entity's weight"

- The boundary may go far (for deep-structure entities)
- Easy to define for w_{desc_k}, w_{leaf_k}, w_{DAG}. Example for w_{desc_2}



Does not apply for PageRank-based scores

Data-acyclic flooding boundary $bound_{dfl-ac}$

Idea: the collection nodes

- **Reachable** from the entity root

Data-acyclic flooding boundary $bound_{dfl-ac}$

Idea: the collection nodes

- **Reachable** from the entity root
- **Mainly** part of **this entity**
  - **Edge transfer factor** $\geq f_{min}$
  - **At-most-one:** each C_s node has at most one child in C_t

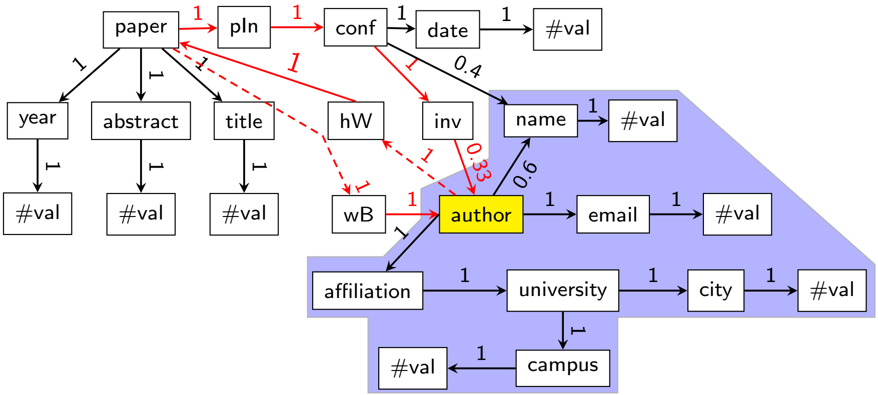
Data-acyclic flooding boundary $bound_{dfl-ac}$

Idea: the collection nodes

- **Reachable** from the entity root
- **Mainly** part of **this entity**
  - **Edge transfer factor** $\geq f_{min}$
  - **At-most-one:** each C_s node has at most one child in C_t
- The path between the entity root and this collection node is **not data cyclic**
  - If the path in $\mathcal{G}$ has no in-cycle edges
  - Or, the $\mathcal{G}$ path has in-cycle edges, but they are not in the data

Data-acyclic flooding boundary $bound_{dfl-ac}$

- **Reachable** from the entity root
- **Mainly** part of **this entity**
- The path is **not data cyclic**



How to update $\mathcal{G}$ after selecting an entity?

Reflect the allocation of data nodes and edges to one entity

How to update $\mathcal{G}$ after selecting an entity?

Reflect the allocation of data nodes and edges to one entity

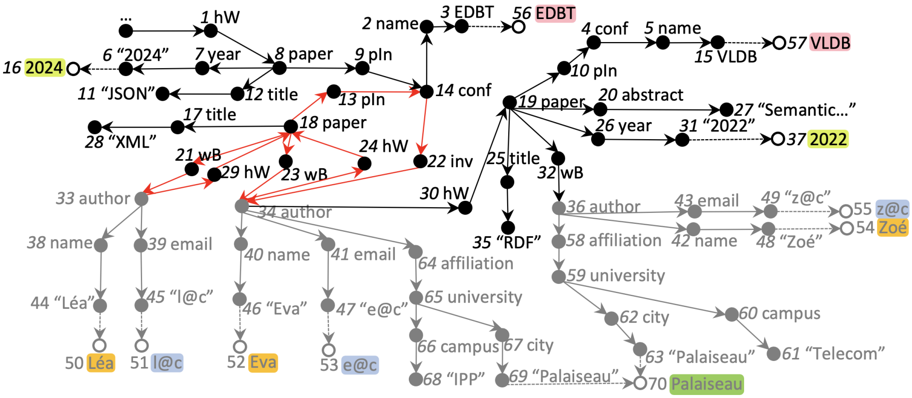
1. $update_{boolean}$
   - Collection nodes and edges in the boundary of the entity
     - Very efficient
     - Sufficient for w_{desc_k}, w_{leaf_k}, w_{DAG}

How to update $\mathcal{G}$ after selecting an entity?

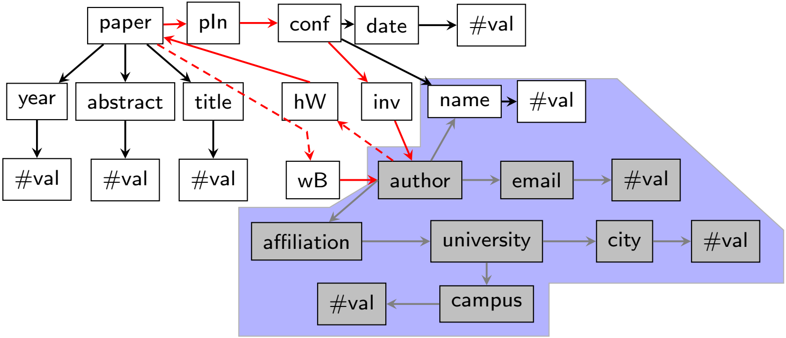
Reflect the allocation of data nodes and edges to one entity

1. $update_{boolean}$
   - Collection nodes and edges in the boundary of the entity
     - Very efficient
     - Sufficient for w_{desc_k}, w_{leaf_k}, w_{DAG}
2. $update_{exact}$
   - Graph nodes and edges
     - Much more costly
     - Required for $w_{PageRank}$, $w_{dwPageRank}$

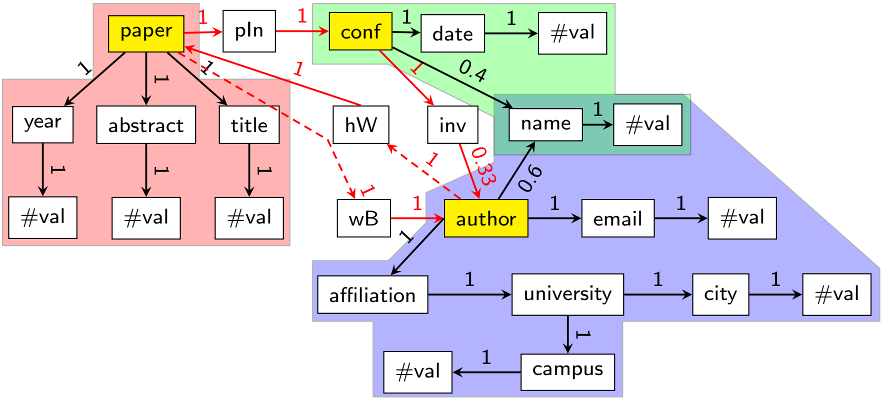
Exact graph update



Exact graph update

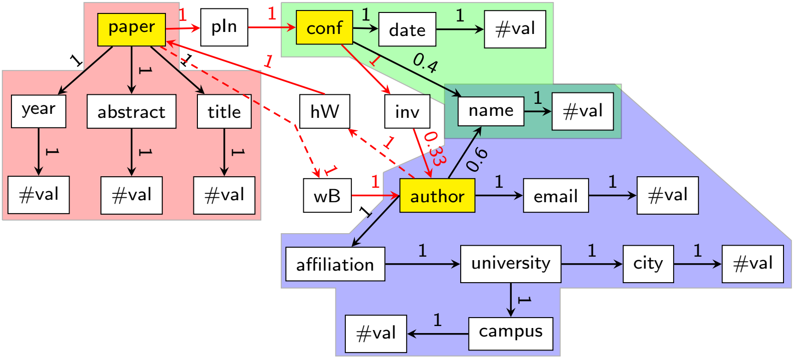


Selected entities and their boundaries



Finding relationships between entities

Relationship: a path from an entity to another



- paper → wB → author
- paper → pln → conf
- author → hW → paper
- conf → inv → author

Entity classification

Assign a semantic category to each entity

Input: an entity E, categories $\mathcal{K}$, semantic properties $\mathcal{P}$

- $\mathcal{K}$: Person, ScientificPaper, Event, Website, Mountain, ...
- $\mathcal{P}$: {`label:"address", domain:[Pers., Org.], range:[Place]`}, ...

Output: a category for E

Entity classification

Assign a semantic category to each entity

Input: an entity E, categories $\mathcal{K}$, semantic properties $\mathcal{P}$

- $\mathcal{K}$: Person, ScientificPaper, Event, Website, Mountain, ...
- $\mathcal{P}$: `{label:"address", domain:[Pers., Org.], range:[Place]}`, ...

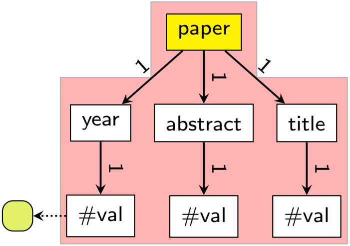
Output: a category for E

Algorithm:

- Compare:
  - The common name of all nodes in the entity root (if it exists) with $k \in \mathcal{K}$ (*conf, paper, author*)
  - Its attribute names with $p \in \mathcal{P}$ (*affiliation, email, ...*)
  - Its entity profiles with $p.range \in \mathcal{P}$ (■, ■, ■, ...)
- Each good match votes for one or few categories

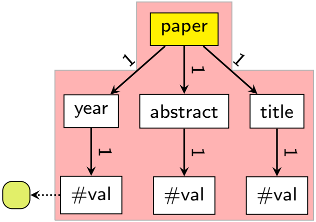
Entity classification

Name	Similar to	Votes for
paper	ResearchPublication (0.85) News (0.63)	ResearchPublication News



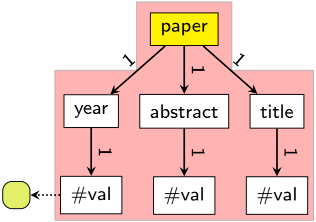
Entity classification

Attribute	Similar to	Votes for
abstract	abstract (1.0) summary (0.92) preface (0.47)	ResearchPublication Book
title	title (1.0) honorific title (0.87)	ResearchPublication Movie Person
year	year publication (0.85 +)	Event Book ResearchPublication, ...



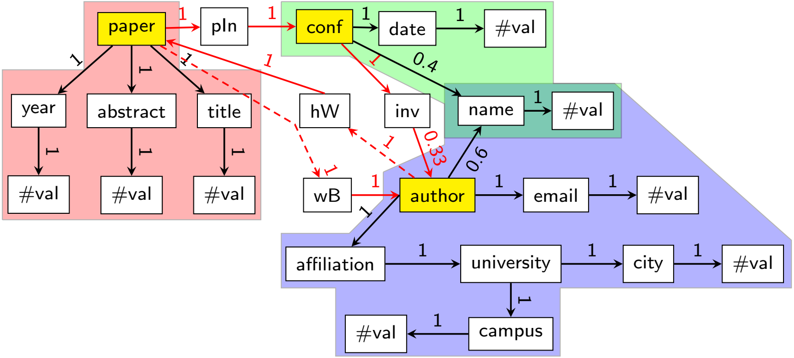
Entity classification

Attribute	Similar to	Votes for
abstract	abstract (1.0) summary (0.92) preface (0.47)	ResearchPublication Book
title	title (1.0) honorific title (0.87)	ResearchPublication Movie Person
year	year publication (0.85 +)	Event Book ResearchPublication, ...

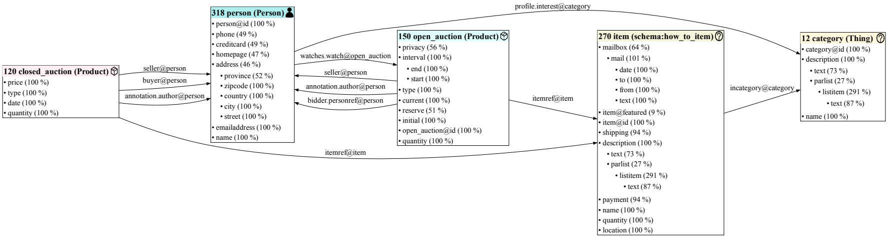


Entity classification

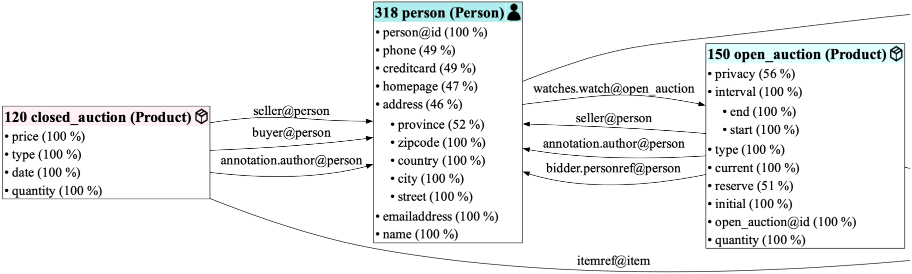
- paper nodes classified as `ResearchPublication`
- author nodes classified as `Researcher`
- conference nodes classified as `Event`



Abstra output: a lightweight Entity-Relationship diagram



Abstra output: a lightweight Entity-Relationship diagram



Experimental evaluation

On main **semi-structured** data models: 8 JSON, 7 RDF, 5 XML, 3 PG

- 10 synthetic, 13 real-world
- 5M to 14M nodes
- Collection graphs:
  - 26 to 4.8K collections
  - 14/23 have cycles

Experimental evaluation

On main **semi-structured** data models: 8 JSON, 7 RDF, 5 XML, 3 PG

- 10 synthetic, 13 real-world
- 5M to 14M nodes
- Collection graphs:
  - 26 to 4.8K collections
  - 14/23 have cycles

Graphs stored in PostgreSQL, algorithms in SQL and Java

Experimental evaluation

On main **semi-structured** data models: 8 JSON, 7 RDF, 5 XML, 3 PG

- 10 synthetic, 13 real-world
- 5M to 14M nodes
- Collection graphs:
  - 26 to 4.8K collections
  - 14/23 have cycles

Graphs stored in PostgreSQL, algorithms in SQL and Java

We evaluate:

1. Entity selection quality
2. Scalability

Entity selection quality with $(w_{dwPageRank}, bound_{fl-ac})$

Dataset name	$\|C\|$	$\|\mathcal{ME}\|$	$\|\mathcal{MR}\|$	cov	$\mathcal{ME}$	d_{max}	$\|\mathcal{ME}_i\|$
Mondial ⟳	168	5	8	0.85	City	3	3,152
					Province	3	1,455
					Country	4	231
					Organization	4	168
					River	4	135
PubMed	26	1	0	1.0	PubMedArticle	5	957
XMark1 ⟳	136	5	10	0.91	Person	4	25,500
					Item	7	21,750
					Open_Auction	8	12,000
					Closed_Auction	8	9,750
					Category	2	1,000
XMark4 ⟳	136	5	10	0.90	Person	4	102,000
					Item	7	87,000
					Open_Auction	8	48,000
					Closed_Auction	8	39,000
					Category	2	4,000
Wikimedia	59	2	0	1.0	Page	4	54,750
					Namespace	3	32

Entity selection quality with ($w_{dwPageRank}$, $bound_{fl-ac}$)

Dataset name	$\|C\|$	$\|\mathcal{ME}\|$	$\|\mathcal{MR}\|$	cov	$\mathcal{ME}$	d_{max}	$\|\mathcal{ME}_i\|$
Mondial ↻	168	5	8	0.85	City	3	3,152
					Province	3	1,455
					Country	4	231
					Organization	4	168
					River	4	135
PubMed	26	1	0	1.0	PubMedArticle	5	957
XMark1 ↻	136	5	10	0.91	Person	4	25,500
					Item	7	21,750
					Open_Auction	8	12,000
					Closed_Auction	8	9,750
					Category	2	1,000
XMark4 ↻	136	5	10	0.90	Person	4	102,000
					Item	7	87,000
					Open_Auction	8	48,000
					Closed_Auction	8	39,000
					Category	2	4,000
Wikimedia	59	2	0	1.0	Page	4	54,750
					Namespace	3	32

Entity selection quality with ($w_{dwPageRank}$, $bound_{fl-ac}$)

Dataset name	$\lvert C\rvert$	$\lvert \mathcal{ME}\rvert$	$\lvert \mathcal{MR}\rvert$	cov	$\mathcal{ME}$	d_{max}	$\lvert \mathcal{ME}_i\rvert$
Mondial ⥁	168	5	8	0.85	City	3	3,152
					Province	3	1,455
					Country	4	231
					Organization	4	168
					River	4	135
PubMed	26	1	0	1.0	PubMedArticle	5	957
XMark1 ⥁	136	5	10	0.91	Person	4	25,500
					Item	7	21,750
					Open_Auction	8	12,000
					Closed_Auction	8	9,750
					Category	2	1,000
XMark4 ⥁	136	5	10	0.90	Person	4	102,000
					Item	7	87,000
					Open_Auction	8	48,000
					Closed_Auction	8	39,000
					Category	2	4,000
Wikimedia	59	2	0	1.0	Page	4	54,750
					Namespace	3	32

Entity selection quality with $(w_{dwPageRank}, bound_{fl-ac})$

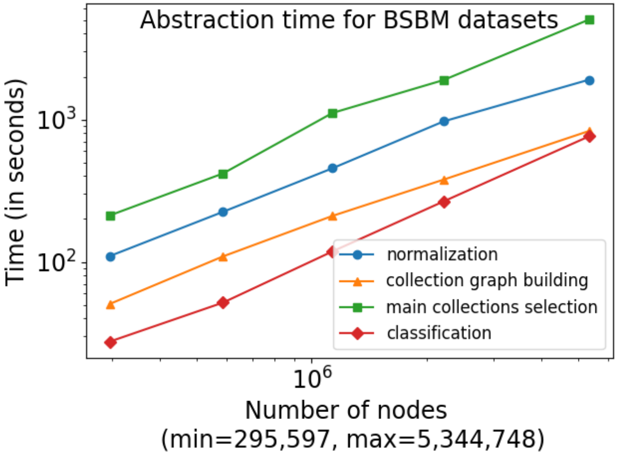
Dataset name	$\|C\|$	$\|\mathcal{ME}\|$	$\|\mathcal{MR}\|$	cov	$\mathcal{ME}$	d_{max}	$\|\mathcal{ME}_i\|$
Mondial ↻	168	5	8	0.85	City	3	3,152
					Province	3	1,455
					Country	4	231
					Organization	4	168
					River	4	135
PubMed	26	1	0	1.0	PubMedArticle	5	957
XMark1 ↻	136	5	10	0.91	Person	4	25,500
					Item	7	21,750
					Open_Auction	8	12,000
					Closed_Auction	8	9,750
					Category	2	1,000
XMark4 ↻	136	5	10	0.90	Person	4	102,000
					Item	7	87,000
					Open_Auction	8	48,000
					Closed_Auction	8	39,000
					Category	2	4,000
Wikimedia	59	2	0	1.0	Page	4	54,750
					Namespace	3	32

Entity selection quality with ($w_{dwPageRank}$, $bound_{fl-ac}$)

Dataset name	$\|C\|$	$\|\mathcal{ME}\|$	$\|\mathcal{MR}\|$	cov	$\mathcal{ME}$	d_{max}	$\|\mathcal{ME}_i\|$
Mondial ⟳	168	5	8	0.85	City	3	3,152
					Province	3	1,455
					Country	4	231
					Organization	4	168
					River	4	135
PubMed	26	1	0	1.0	PubMedArticle	5	957
XMark1 ⟳	136	5	10	0.91	Person	4	25,500
					Item	7	21,750
					Open_Auction	8	12,000
					Closed_Auction	8	9,750
					Category	2	1,000
XMark4 ⟳	136	5	10	0.90	Person	4	102,000
					Item	7	87,000
					Open_Auction	8	48,000
					Closed_Auction	8	39,000
					Category	2	4,000
Wikimedia	59	2	0	1.0	Page	4	54,750
					Namespace	3	32

Abstra selects frequent, coherent and semantically central entities

Experimental evaluation: scalability



Our abstraction method scales up linearly in the data size

Extending abstractions to Property Graphs

Property Graphs (PGs) are graphs whose nodes and edges may carry named attributes

- Model under standardization [ABD+23, ABD+21]
- Numerous industrial PG databases (Neo4J, Oracle)
- Widely used (the Offshore leaks database)

For interoperability, we **derive a PG schema from any (semi)structured dataset** following PG-Schema [ABD+23]

Deriving a PG schema from an abstraction

We need to accommodate to nested attributes:

- FLAT: wrap the nested attribute in a JSON object
- CUT: unfold each nested attribute in a PG node

1. For each Abstra entity *E*:
   1. Create a PG node type for *E*
   2. For each attribute *a*:
      1. If *a* is not nested: add *a* to the PG node
      2. If *a* is nested and nesting is FLAT: wrap *a*
      3. If *a* is nested and nesting is CUT: unfold *a*
2. For each Abstra relationship *R*:
   1. Create a PG edge type with corresponding PG nodes
3. If all *G* nodes and edges are in E-R: PG graph type is STRICT, else LOOSE

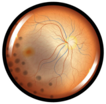
Extending abstractions to Property Graphs

```
CREATE GRAPH TYPE myGraphType STRICT {
  (paperType: Paper {
      title string,
      OPTIONAL year integer,
      OPTIONAL abstract string, ...
  })
  (authorType: Author {name string, email string, ...}),
  (confType: Conference {name string, year integer, ...}),

  (:authorType)-[edgeAuthorPaper: HasWritten]->(:paperType),
  (:paperType)-[edgePaperAuthor: WrittenBy]->(:authorType),
  (:paperType)-[edgePaperConf: PublishedIn]->(:confType),
}
```

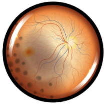
Outline

What does multi-source healthcare data has to reveal?



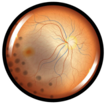
- Very low cooperation/normalization between medical centers
- Few patient data for rare diseases

What does multi-source healthcare data has to reveal?



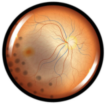
- Very low cooperation/normalization between medical centers
- Few patient data for rare diseases
- Traditional setting: warehouses [DM88]

What does multi-source healthcare data has to reveal?



- Very low cooperation/normalization between medical centers
- Few patient data for rare diseases
- Traditional setting: warehouses [DM88]
- Need to provide decentralized and federated analyses!

What does multi-source healthcare data has to reveal?



- Very low cooperation/normalization between medical centers
- Few patient data for rare diseases
- Traditional setting: warehouses [DM88]
- Need to provide decentralized and federated analyses!
- Leverage experts' knowledge + make it as automatic as possible

Post-doc problem and research contributions

Problem statement

How to **enable federated analyses of healthcare data** across institutions and national borders?

Post-doc problem and research contributions

Problem statement

How to **enable federated analyses of healthcare data** across institutions and national borders?

2 conceptual models for healthcare (meta)data [BBBP25] 

- A metadata model to collect expert's knowledge on their data
- An extensible and general data model to represent healthcare data

Post-doc problem and research contributions

Problem statement

How to **enable federated analyses of healthcare data** across institutions and national borders?

2 conceptual models for healthcare (meta)data [BBBP25]

- A metadata model to collect expert's knowledge on their data
- An extensible and general data model to represent healthcare data

An ETL to build interoperable healthcare databases [BBBP25]

- Produces an interoperable warehouse at each medical center
- Assesses interoperability along the ETL pipeline

Post-doc problem and research contributions

Problem statement

How to **enable federated analyses of healthcare data** across institutions and national borders?

2 conceptual models for healthcare (meta)data [BBBP25]

- A metadata model to collect expert's knowledge on their data
- An extensible and general data model to represent healthcare data

An ETL to build interoperable healthcare databases [BBBP25]

- Produces an interoperable warehouse at each medical center
- Assesses interoperability along the ETL pipeline

Also: a catalogue (browse and query) + a federated AI analysis platform

Related work

Data plateforms:

- EHDEN [PdGdK+23] for tabular data
- Also: OHDSI [HDS+15], UMG-MeDIC [PSS+23], etc

Conceptual models:

- OMOP [SRR+10] for observational data, also FHIR [fhi]

ETL pipelines:

- D-ETL [OKK+17], also EHDEN's ETL, OHDSI's ETL

Related work

Data plateforms:

- EHDEN [PdGdK+23] for tabular data
- Also: OHDSI [HDS+15], UMG-MeDIC [PSS+23], etc

Conceptual models:

- OMOP [SRR+10] for observational data, also FHIR [fhi]

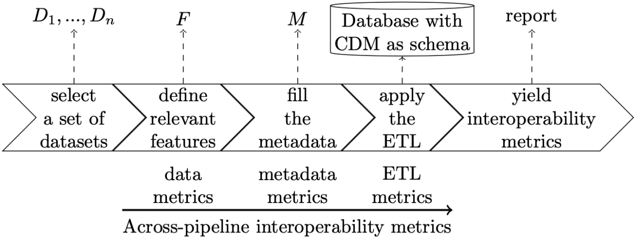
ETL pipelines:

- D-ETL [OKK+17], also EHDEN's ETL, OHDSI's ETL

- Data platforms are often tied to a single data type (tables, etc.)
- Conceptual models often design one kind of data (observational, etc.)
- ETLs provide limited interoperability and require time from experts

The I-ETL approach

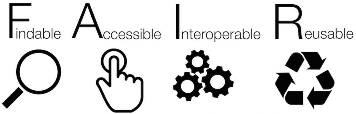
1. Analyze datasets and extract their metadata
2. Create an interoperable database in each medical center
3. Assess interoperability along the pipeline
4. Allow federated analyses of data across centers



Interoperability as in FAIR principles

FAIR principles are guidelines for good data management [WDA+16]:

- (F) Findable: search for (indexed) resources based on identifiers
- (A) Accessible: access data with standard protocols, even after data dies
- (I) Interoperable: integrate and refer to datasets following FAIR principles
- (R) Reusable: reuse datasets in other settings using provenance, etc.



From datasets analysis to metadata

Metadata: each dataset can be described by a set of features

- Name, definition, type, unit, values, ...
- Specified by medical experts

	A	B	C	D	E	F	G	H	I	J	K	L
1	**line**	**SampleBarcode**	**2MBC**	**ADO**	**Ala**	**DateOfBirth**	**Weight**	**ENFeed**	**Etnicity**	**Gest.**	**Sex**	**id**
2	0	20LD042587		0.374	463.711	2021-12-19	3370	0	Caucasian	39	M	4.2165648176126E+018
3	1	20LD050321		0.659	372.249	2021-12-12	4050	1	Caucasian	40	F	-1.25680706520154E+18
4	2	20LD810743		0.208	815.699	2021-12-20	2425	0	Italy	36	M	-7.58913922052569E+18
5	3	20LD811192		0.32	328.315	2021-12-15	3430	0	Caucasian	39	M	6.73808832627831E+18
6	4	20LD811194	0.089	0.295	504.553	2021-12-14	3170	0	Italy	41	M	7.1694429792529E+018
7	5	20LD811195		0.427	379.091	2021-12-14	2810	0	Caucasian	40	M	3.3926325222509E+018
8	6	20LD811196		0.242	378	2021-12-15	3510	1	Asian	40	M	-5.74936744641016E+18
9	7	20LD811197		0.293	274.644	2021-12-15	4200	1	Morocco	39	M	3.81562481054663E+17
10	8	20LD811198		0.463	265.08	2021-12-02	3000	0	India	39	M	-5.97288917349607E+18
11	9	20LD811199	0.165	0.607	297.13	2021-12-15	2500	1	Caucasian	38	M	-6.68549945408787E+18

Tabular data of clinical measurements and phenotypic information

From datasets analysis to metadata

Metadata: each dataset can be described by a set of features

- Name, definition, type, unit, values, ...
- Specified by medical experts

	A	B	C	D	E
1	**name**	**description**	**vartype**	**dimension**	**Accepted values**
2	SampleBarcode	Sample ID	str		
3	2MBC	2-methylbutyrylcarnitine	float	microM	
4	ADO	adenosine	float	micromol/L	
5	Ala	alanine	float	micromol/L	
6	DateOfBirth	Date of birth	datetime64		
7	Weight	Baby weight	int	g	
8	ENFeed	Enteral feeding	category		0, NA, 1
9	Etnicity	Geographical origin	category		
10	Gest.	Gestational age	int	settimane e giorni	
11	Sex	Sex	category		M, F
12	id	Patient ID	int		

The metadata obtained from the tabular data

From datasets analysis to metadata

Metadata: each dataset can be described by a set of features

- Name, definition, type, unit, values, ...
- Specified by medical experts

	A	B	C	D	E
1	**name**	**description**	**vartype**	**dimension**	**Accepted values**
2	SampleBarcode	Sample ID	str		
3	2MBC	2-methylbutyrylcarnitine	float	microM	
4	ADO	adenosine	float	micromol/L	
5	Ala	alanine	float	micromol/L	
6	DateOfBirth	Date of birth	datetime64		
7	Weight	Baby weight	int	g	
8	ENFeed	Enteral feeding	category		0, NA, 1
9	Etnicity	Geographical origin	category		
10	Gest.	Gestational age	int	settimane e giorni	
11	Sex	Sex	category		M, F
12	id	Patient ID	int		

The metadata obtained from the tabular data

- What if "ethnicity" is referred to as "race" in another dataset?
- What if datasets refer to "Homme" / "Femme" vs. "Male" / "Female"?

The metadata model

We aim for a conceptual model for **expressive and interoperable metadata**

- **Name**: the feature name
- **Vocabulary**: a vocabulary name
- **Code**: the code of the term in the selected vocabulary
- **Kind**: phenotypic, clinical, genomic, ...
- **DataType**: *string*, *integer*, *numeric*, *boolean*, *category*, ...
- **Unit**: to interpret values when the data type is numeric;
- **Categories**: list of discrete values for categorical features
- **Visibility**: *public*, *anonymized*, *private*

Associate metadata to vocabularies

Vocabularies are dictionaries of concepts/values uniquely identified

- SNOMED_CT [SPSW01], LOINC [HRM+98], OMIM [HSA+05], ...

We associate each feature and categorical value to an existing vocabulary code → more interoperability

	A	B	C	D	E	F	G
1	**ontology**	**ontology_code**	**name**	**description**	**vartype**	**dimension**	**Accepted values**
2	loinc	57723-9	SampleBarcode	Sample ID	str		
3	loinc	30531-8	2MBC	2-methylbutyrylcarnitine	float	microM	
4	CLIR	M-004315	ADO	adenosine	float	micromol/L	
5	loinc	53150-9	Ala	alanine	float	micromol/L	
6	snomed ct	184099003	DateOfBirth	Date of birth	datetime64		
7	loinc	56056-5	Weight	Baby weight	int	g	
8	snomed ct	1217195001	ENFeed	Enteral feeding	category		0: (snomed_ct, 373067005) NA: (snomed_ct, 276727009) 1: (snomed_ct, 373066001)
9	loinc	46463-6	Etnicity	Geographical origin	category		
10	loinc	49051-6	Gest.	Gestational age	int	settimane e giorni	
11	snomed ct	734000001	Sex	Sex	category		M: (snomed_ct, 248153007) F: (snomed_ct, 248152002)
12			id	Patient ID	int		

The healthcare data model

We need a **general, extensible healthcare data model**

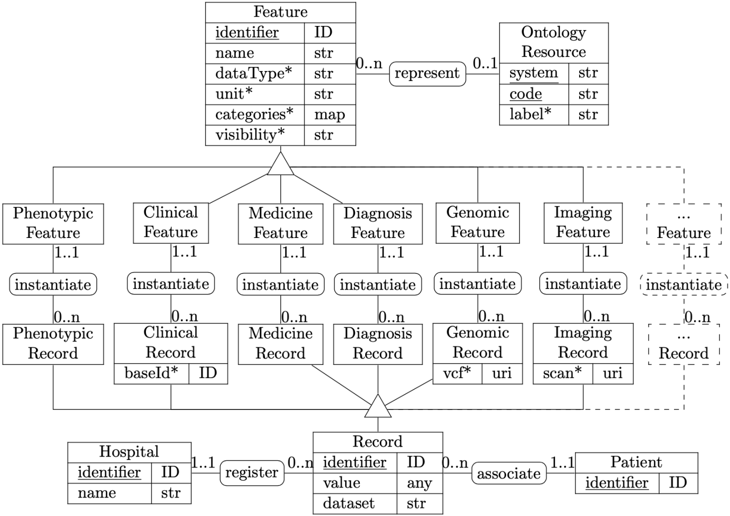
Challenges:

- Use-cases bring very different kinds of data
- Experts' metadata needs to be represented

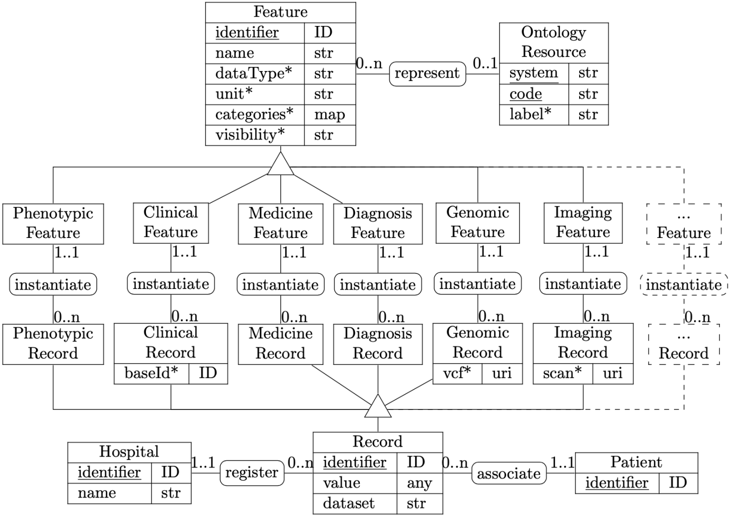
We aim for a **conceptual model**:

- Based on the notions of **features** and **records**
- Will be populated automatically by an **ETL**

General, extensible healthcare conceptual data model



General, extensible healthcare conceptual data model



How to automatically populate this data model with hospitals data?

Populate the data model within each hospital

The ETL algorithm: harmonizing data towards a target data model

1. **Extract**
   - Read the metadata
   - Read the datasets
2. **Transform**
   - Each patient is **anonymized** → a Patient instance
   - Each **metadata** variable → a Feature instance
   - Each **data** value is made **interoperable** + **anonymized** → a Record instance
     - **Interoperability**: cast based on type, identify NA, etc.
     - **Anonymization**: remove day from dates, etc.
3. **Load**
   - Load Patient, Feature, Record instances
   - Index instances

Example: the “Sex” Feature


```
{
  _id: ObjectId('678b8d94ad9674aed4cf2ea6'),
  name: 'sex',
  categories: [
    {
      system: 'http://snomed.info/sct',
      code: '248153007',
      label: 'Male'
    },
    {
      system: 'http://snomed.info/sct',
      code: '248152002',
      label: 'Female'
    }
  ],
  data_type: 'category',
  dataset_gid: 'http://better-health-project.eu/datasets/94ca1863-3815-46d7-bf28-0beeb3526073',
  description: 'Sex',
  domain: { accepted_values: [ 'm', 'f' ] },
  entity_type: 'phenotypicFeature',
  identifier: 63183,
  ontology_resource: {
    system: 'http://snomed.info/sct',
    code: '734000001',
    label: 'Biological sex'
  },
  visibility: 'PUBLIC'
}
```

Example: a “Sex” Record


```
{
  _id: ObjectId('678b8d96ad9674aed4cf3018'),
  registered_by: 1,
  entity_type: 'phenotypicRecord',
  instantiates: 63183,
  has_subject: 19,
  dataset: 'http://better-health-project.eu/datasets/94ca1863-3815-46d7-bf28-0beeb3526073',
  identifier: 63554,
  value: {
    system: 'http://snomed.info/sct',
    code: '248152002',
    label: 'Female'
  }
},
```

Interoperability assessment

We need to **assess intra- and extra-interoperability** (subset of FAIR)

Challenges:

- Interoperability should be assessed from the start [CMP24]
- Both metadata and data need to be assessed

We aim for a **set of interoperability metrics**:

- Evaluate input data and metadata interoperability
- Assign a score to each metric, rather than a single score [CMP24]

Interoperability assessment

Step	Metric
Data	(A1) Ratio of selected features (A2) Ratio of datasets that do not require dedicated extraction
Metadata	(M1) Features with both non-empty ontology *name* and *code* (M2) Features with non-empty *dataType* (M3) Features with non-empty *visibility* (M4) Categorical features with non-empty set of *categories* (M5) Numerical features with non-empty *unit*
ETL	(E1) Presence of non-empty *label* in Ontology Resource (E2) Values for which interoperability implementation has succeeded (E3) Correspondence of numerical values *unit* and Feature *unit* (E4) Presence of categorical value in the Feature *categories* (E5) Records with known Hospital references (E6) Records with known Patient references (E7) Records with known Feature references

The higher the metric is, the better

Anchor our metrics in FAIR principles

1. (Meta)data use a [...] broadly applicable language for **knowledge representation**
   - Our data model can be implemented within any type of database
   - Metadata can be easily specified using a tabular file
2. (Meta)data use **FAIR-first vocabularies**
   - Associate metadata variables and categories to vocabulary resources
   - Use of widely used vocabularies in healthcare domain
3. (Meta)data include **qualified references** to other data and metadata
   - To db instances: references to a patient, a hospital, and a Feature
   - To the data: from which dataset the value comes

I-ETL at work in the Better project

7 clinical centers across Europe

I-ETL is **under deployment** at each center → 7 interoperable databases

I-ETL at work in the Better project

7 clinical centers across Europe

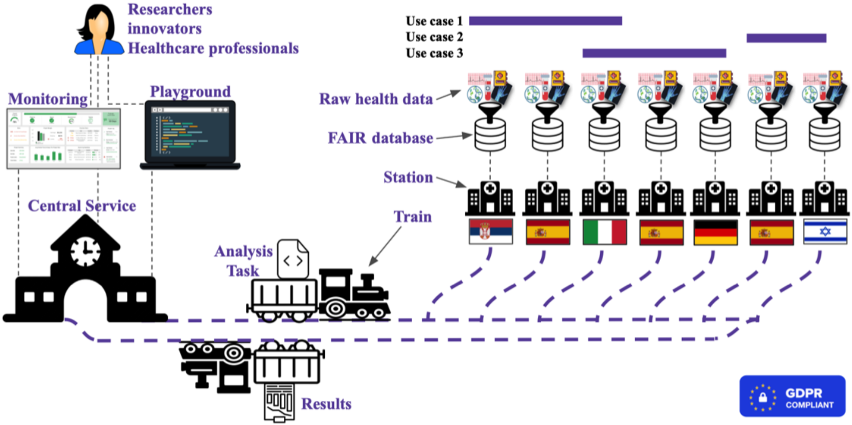
I-ETL is **under deployment** at each center → 7 interoperable databases

Working on:

- Designing a **catalogue** to:
  - List available datasets and their associated metadata
  - Explore datasets and their aggregated data
    - With visualizations and queries
- Designing a **decentralized federated learning plateform**
  - To run federated AI algorithms
  - Secured because no data leaves centers, only aggregates

Better plateform: decentralized federated learning

Based on the **Personal Health Train**: stations (centers), trains (queries), central station (results aggregation) → no data leaves centers = privacy preservation



Outline

Systems developed (1/2)

Abstra for data abstraction:
65 Java core classes, 10K LOC
Published in EDBT 2024 [BMU24]
Demonstrated at CIKM and BDA 2022 [BMU22]

PathWays for NE-to-NE paths:
18 Java core classes, 4K LOC
Published in ADBIS 2023 [BGLM23a], Info. Sys [BGLM25]
Demonstrated at ESWC and BDA 2023 [BGLM23b]

Systems developed (1/2)

Abstra for data abstraction:
65 Java core classes, 10K LOC
Published in EDBT 2024 [BMU24]
Demonstrated at CIKM and BDA 2022 [BMU22]

PathWays for NE-to-NE paths:
18 Java core classes, 4K LOC
Published in ADBIS 2023 [BGLM23a], Info. Sys [BGLM25]
Demonstrated at ESWC and BDA 2023 [BGLM23b]

ConnectionStudio for NTU data exploration:
Web interface by CEDAR engineers
Published in CoopIS 2023 [BEG+23]
Demonstrated at BDA and SEAGRAPH 2024 [BEMM24, BBE+24]
Also to journalists at **DataJournos** (40) and **CFI** (60)

Systems developed (2/2)

I-ETL for interoperable healthcare databases:
31 Python core classes, 8K LOC (*restricted access*)
Reviewed at BMC Med. Info. & Decision Making [BBBP25]
Under deployment in the 7 medical centers of the project

Data catalogue and decentralized platform:

- *Under development by an IT company*
- With collaboration of Better technical partners

Outline

Takeaways and next steps (1/2)

In my PhD, we introduced:

1. A unified view over heterogeneous semi-structured data models
2. Abstra: a dataset abstraction system for semi-structured data
3. PathWays: an entity-focused exploration system
4. ConnectionStudio: a comprehensive data lake exploration tool

Takeaways and next steps (1/2)

In my PhD, we introduced:

1. A unified view over heterogeneous semi-structured data models
2. Abstra: a dataset abstraction system for semi-structured data
3. PathWays: an entity-focused exploration system
4. ConnectionStudio: a comprehensive data lake exploration tool

Next steps:

- Migrate data graphs into PG graphs reusing [BEMM24]
- Enrich extracted NEs with RDF knowledge bases
- Propose an end-to-end data processing/exploration pipeline

Takeaways and next steps (2/2)

In my post-doc, we introduced:

1. Two general, extensible conceptual models for healthcare
2. I-ETL: an algorithm to build interoperable databases
3. The platform: a catalogue and federated learning algorithms

Takeaways and next steps (2/2)

In my post-doc, we introduced:

1. Two general, extensible conceptual models for healthcare
2. I-ETL: an algorithm to build interoperable databases
3. The platform: a catalogue and federated learning algorithms

Next steps:

- Automatically find vocabulary resources for the metadata
- Design a query engine for underlying databases
- Propose general visualizations and interactions

References I

Angelos Anadiotis, Oana Balalau, Catarina Conceicao, et al.
Graph integration of structured, semistructured and unstructured data for data journalism.
Inf. Systems, 104, 2022.

Renzo Angles, Angela Bonifati, Stefania Dumbrava, George Fletcher, Keith W Hare, Jan Hidders, Victor E Lee, Bei Li, Leonid Libkin, Wim Martens, et al.
Pg-keys: Keys for property graphs.
In *Proceedings of the 2021 International Conference on Management of Data*, pages 2423–2436, 2021.

Renzo Angles, Angela Bonifati, Stefania Dumbrava, George Fletcher, Alastair Green, Jan Hidders, Bei Li, Leonid Libkin, Victor Marsault, Wim Martens, et al.
PG-Schema: schemas for property graphs.
Proceedings of the ACM on Management of Data, 1(2):1–25, 2023.

Nelly Barret, Boris Bikbov, Anna Bernasconi, and Pietro Pinoli.
I-etl: an interoperability-aware health (meta)data pipeline to enable federated analyses.
Under review in BMC Medical Informatics and Decision Making, 2025.

Oana Balalau, Nelly Barret, Simon Ebel, Théo Galizzi, and Madhulika Manolescu, Ioana an Mohanty.
Graph lenses over any data: the ConnectionLens experience.
In *SEAGraph workshop*, 2024.

Mohamed Amine Baazizi, Dario Colazzo, Giorgio Ghelli, and Carlo Sartiani.
Parametric schema inference for massive JSON datasets.
VLDB J., 28(4), 2019.

References II

Nelly Barret, Simon Ebel, Théo Galizzi, Ioana Manolescu, and Madhulika Mohanty.
User-friendly exploration of highly heterogeneous data lakes.
In Mohamed Sellami, Maria-Esther Vidal, Boudewijn F. van Dongen, Walid Gaaloul, and Hervé Panetto, editors, *Cooperative Information Systems - 29th International Conference, CoopIS 2023, Groningen, The Netherlands, October 30 - November 3, 2023, Proceedings*, volume 14353 of *Lecture Notes in Computer Science*, pages 488–496. Springer, 2023.

Nelly Barret, Tudor Enache, Ioana Manolescu, and Madhulika Mohanty.
Finding the PG schema of any (semi)structured dataset: a tale of graphs and abstraction.
In *SEAGraph workshop*, 2024.

Nelly Barret, Antoine Gauquier, Jia Jean Law, and Ioana Manolescu.
Exploring heterogeneous data graphs through their entity paths.
In *Advances in Databases and Information Systems*, volume 13985 of *Lecture Notes in Computer Science*, pages 163–179. Springer, 2023.

Nelly Barret, Antoine Gauquier, Jia Jean Law, and Ioana Manolescu.
PATHWAYS: entity-focused exploration of heterogeneous data graphs (demonstration).
In *ESWC*, 2023.

Nelly Barret, Antoine Gauquier, Jia Jean Law, and Ioana Manolescu.
Exploring heterogeneous data graphs through their entity paths.
Information Systems, 2025.

Nelly Barret, Ioana Manolescu, and Prajna Upadhyay.
ABSTRA: toward generic abstractions for data of any model (demonstration).
In *CIKM*, 2022.

References III

Nelly Barret, Ioana Manolescu, and Prajna Upadhyay.
Computing generic abstractions from application datasets.
In *EDBT*, 2024.

Clair Blacketer, Erica A Voss, Frank DeFalco, Nigel Hughes, Martijn J Schuemie, Maxim Moinat, and Peter R Rijnbeek.
Using the data quality dashboard to improve the EHDEN network.
Applied Sciences, 11(24):11920, 2021.

Dario Colazzo, Giorgio Ghelli, and Carlo Sartiani.
Schemas for safe and efficient XML processing.
In *ICDE*. IEEE Computer Society, 2011.

Leonardo Candela, Dario Mangione, and Gina Pavone.
The FAIR assessment conundrum: Reflections on tools and metrics.
Data Science Journal, 23(1), 2024.

Barry A. Devlin and Paul T. Murphy.
An architecture for a business and information system.
IBM systems Journal, 27(1):60–80, 1988.

The FHIR framework.
`https://hl7.org/fhir/summary.html`. Accessed 21 November 2024.

François Goasdoué, Pawel Guzewicz, and Ioana Manolescu.
RDF graph summarization for first-sight structure discovery.
The VLDB Journal, 29(5), April 2020.

References IV

Benoît Groz, Aurélien Lemay, Slawek Staworko, and Piotr Wieczorek.
Inference of shape graphs for graph databases.
In *ICDT*, volume 220, 2022.

Roy Goldman and Jennifer Widom.
DataGuides: enabling query formulation and optimization in semistructured databases.
In *VLDB*, 1997.

George Hripcsak, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R Rijnbeek, et al.
Observational health data sciences and informatics (OHDSI): opportunities for observational researchers.
In *MEDINFO 2015: eHealth-enabled Health*, pages 574–578. IOS Press, 2015.

Stanley M Huff, Roberto A Rocha, Clement J McDonald, Georges JE De Moor, Tom Fiers, W Dean Bidgood Jr, Arden W Forrey, William G Francis, Wayne R Tracy, Dennis Leavelle, et al.
Development of the logical observation identifier names and codes (LOINC) vocabulary.
Journal of the American Medical Informatics Association, 5(3):276–292, 1998.

Katja Hose and Ralf Schenkel.
Towards benefit-based RDF source selection for SPARQL queries.
In *Proceedings of the 4th International Workshop on Semantic Web Information Management*, pages 1–8, 2012.

Ada Hamosh, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick.
Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders.
Nucleic acids research, 33(suppl_1):D514–D517, 2005.

References V

Shahan Khatchadourian and Mariano P Consens.
ExpLOD: summary-based exploration of interlinking and RDF usage in the Linked Open Data Cloud.
In *Extended semantic web conference*, pages 272–287. Springer, 2010.

Hanâ Lbath, Angela Bonifati, and Russ Harmer.
Schema inference for property graphs.
In *EDBT*, 2021.

Tova Milo and Dan Suciu.
Index structures for path expressions.
In *International Conference on Database Theory*, pages 277–295. Springer, 1999.

Toan C Ong, Michael G Kahn, Bethany M Kwan, Traci Yamashita, Elias Brandt, Patrick Hosokawa, Chris Uhrich, and Lisa M Schilling.
Dynamic-ETL: a hybrid approach for health data extraction, transformation and loading.
BMC medical informatics and decision making, 17:1–12, 2017.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd.
The PageRank citation ranking: Bringing order to the web.
Technical report, Stanford InfoLab, 1999.

Daniel Puttmann, Rowdy de Groot, Nicolette de Keizer, Ronald Cornet, et al.
Assessing the FAIRness of databases on the EHDEN portal: A case study on two Dutch ICU databases.
International Journal of Medical Informatics, 176:105104, 2023.

Felipe Pezoa, Juan L Reutter, Fernando Suarez, Martín Ugarte, and Domagoj Vrgoč.
Foundations of JSON schema.
In *Proceedings of the 25th international conference on World Wide Web*, pages 263–273, 2016.

References VI

Marcel Parciak, Markus Suhr, Christian Schmidt, Caroline Bönisch, Benjamin Löhnhardt, Dorothea Kesztyüs, and Tibor Kesztyüs.
Fairness through automation: development of an automated medical data integration infrastructure for fair health data in a maximum care university hospital.
BMC Medical Informatics and Decision Making, 23(1):94, 2023.

Raghu Ramakhrishnan and Johannes Gehrke.
Database Management Systems (3rd edition).
McGraw-Hill, 2003.

Matteo Riondato, David García-Soriano, and Francesco Bonchi.
Graph summarization with quality guarantees.
Data mining and knowledge discovery, 31:314–349, 2017.

Michael Q Stearns, Colin Price, Kent A Spackman, and Amy Y Wang.
SNOMED clinical terms: overview of the development process and project status.
In *Proceedings of the AMIA Symposium*, page 662. American Medical Informatics Association, 2001.

Paul E Stang, Patrick B Ryan, Judith A Racoosin, J Marc Overhage, Abraham G Hartzema, Christian Reich, Emily Welebob, Thomas Scarnecchia, and Janet Woodcock.
Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership.
Annals of internal medicine, 153(9):600–606, 2010.

Resource Description Framework (RDF).
https://www.w3.org/RDF/.

References VII

The XML data model.
https://www.w3.org/XML/Datamodel.html.

W3C XML Document Type Specification.
https://www.w3.org/TR/REC-xml/#dt-doctype, 2008.

W3C XML Schema Definition Language (XSD).
https://www.w3.org/TR/xmlschema11-1/, 2012.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al.
The FAIR guiding principles for scientific data management and stewardship.
Scientific data, 3(1):1–9, 2016.

Mussab Zneika, Claudio Lucchese, Dan Vodislav, and Dimitris Kotzinos.
Summarizing linked data RDF graphs using approximate graph pattern mining.
In *19th International Conference on Extending Database Technology*, 2016.

The relational data model

According to [RG03]:

- A **relational schema** is a set of relations
- Each **relation** has a name and set of named attributes with their domain
- A **primary key** is a subset of attributes to uniquely identify a tuple
- A **foreign key** is a reference to a primary key

paper		
id	title	abstract
P1	RDF	W3C...
P2	XML	Data...
P3	JSON	Nodes...

wrote		
authorId	paperId	year
A1	P1	2023
A2	P1	2003
A3	P2	2013

author		
id	name	affiliation
A1	Alice	INRIA
A2	Bob	IPP
A3	Carl	IPP

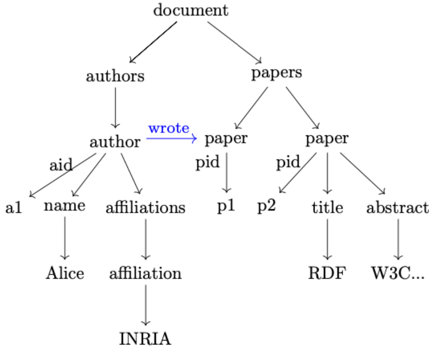
The XML data model

According to the W3C [W3Cb], a **tree** of:

- A (single) **document node**
- **Element nodes** with non-ϵ labels, possibly with named attributes
- **Text nodes**, carrying values, are children of element nodes

Possibility do define a DTD [W3C08] or an XSD [W3C12]

```xml
<document>
  <authors>
    <author aid="a1" wrote="p1">
      <name>Alice</name>
      <affiliations>
        <affiliation>INRIA</affiliation>
      </affiliations
    </author>
  </authors>
  <papers>
    <paper pid="p1"/>
    <paper pid="p2">
      <title>RDF</title>
      <abstract>W3C...</abstract>
    </paper>
  </papers>
</document>
```

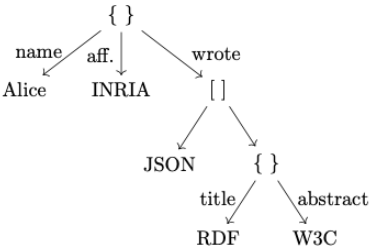


The JSON data model

According to [PRS+16], a **tree** where a node can be:

- A **map** (ϵ label, one or more a key-value elements)
- An **array** (ϵ label, zero or more child nodes)
- A **value** (a string)

```
{
  "name": "Alice",
  "aff.": "INRIA",
  "wrote": [
    "JSON",
    {
      "title": "RDF",
      "abstract": "W3C..."
    }
  ]
}
```



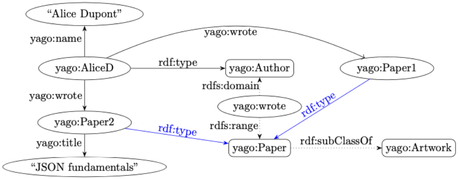
The RDF data model

According to the W3C [W3Ca], an RDG graph contains triples <s,p,o> where:

- *s* and *p* are **resource identifiers** (URIs)
- *o* can be a resource identifier or a literal (string)
- Also: blank nodes for anonymous resources (internal ID)

Add **semantic information** with ontologies (incl. RDFS, OWL)

```
<yago:AliceD> <yago:name> "Alice Dupont" .
<yago:AliceD> <rdf:type> <yago:Author> .
<yago:AliceD> <yago:wrote> <yago:Paper1> .
<yago:AliceD> <yago:wrote> <yago:Paper2> .
<yago:Paper1> <rdf:type> <yago:Paper> .
<yago:Paper2> <yago:title> "JSON fundamentals" .
<yago:wrote> <rdfs:domain> <yago:Author> .
<yago:wrote> <rdfs:range> <yago:Paper> .
<yago:Paper> <rdf:subClassOf> <yago:Artwork> .
```

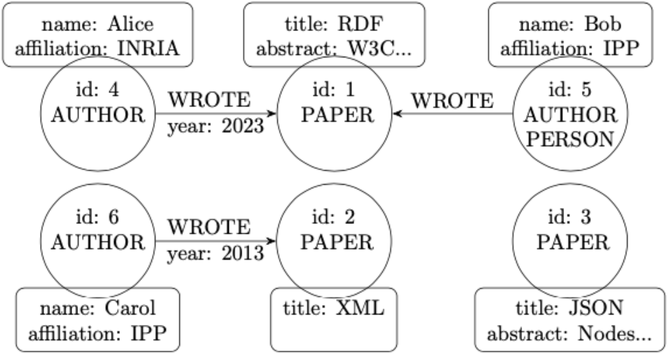


The property graph data model

A **node** is a structured record with:

- 0..n labels (types)
- 0..n properties (key-values)
- Records with the same type set may have different properties

A **relationship** is a directed labeled edge; possibly have attributes



Data summarization techniques

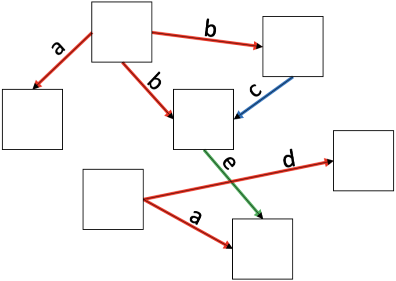
Build structured and concise summaries out of datasets; many approaches for semi-structured

- Structural approaches: groups of equivalent nodes; different notions of node similarity
  - Quotient summaries: groups based on an equivalence relation
  - Non-quotient summaries: other means (dataguides, etc)
- Pattern mining approaches: discovery of patterns
- Statistical approaches: counts over data (classes, properties, value types, etc)
- Hybrid approaches: combine above methods

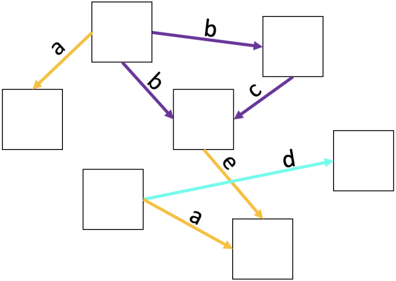
Schema inference techniques: build a schema s.t. the data conforms to it

RDF quotient graph summarization [GGM20]

- Source clique: set of outgoing properties co-occuring together on at least one node
- Target clique: set of incoming properties co-occuring together on at least one node



Properties “a”, “b”, “d” are in the same source clique



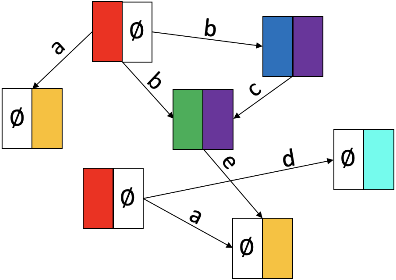
Properties “a” and “e” are in the same target clique

(c) Pawel Guzewic

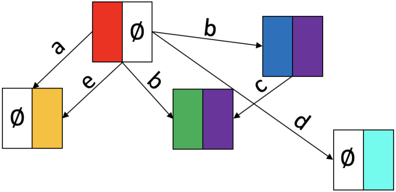
Strong summary [GGM20]

Strong S summary:

- Two nodes are S equivalent iff they have both the same source and target cliques



Source and target cliques for each node



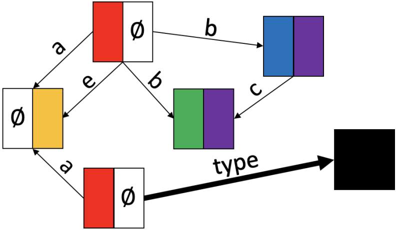
Strong summary

(c) Pawel Guzewic

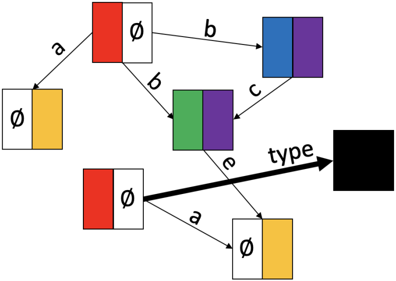
Typed-strong summary [GGM20]

Typed-strong TS summary:

- Two typed nodes are TS equivalent iff they have the same type set
- Two untyped nodes are TS equivalent iff they have both the same source and target cliques



Source and target cliques for each node + an RDF type



Typed-strong summary

(c) Pawel Guzewic

The PageRank algorithm [PBMW99]

Well-known algorithm to compute scores in a graph:

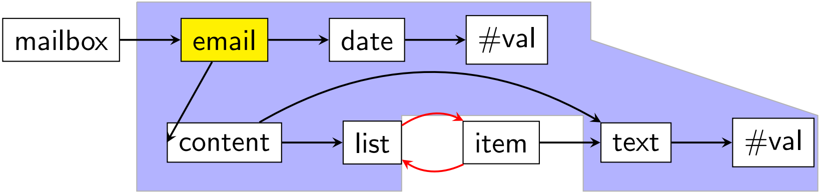
- Based on propagation along edges
- Regardless of the graph structure (complexity, cycles)
- A node is important if it is pointed by many important nodes

Key steps:

1. Each node has an initial score of $\frac{1}{|G|}$
2. Each edge has a weight of $\frac{1}{\text{source node out-degree}}$
3. Propagate scores until convergence

Note: PageRank scores reflect the graph topology only (independently from node weights)

Data-acyclic flooding boundary



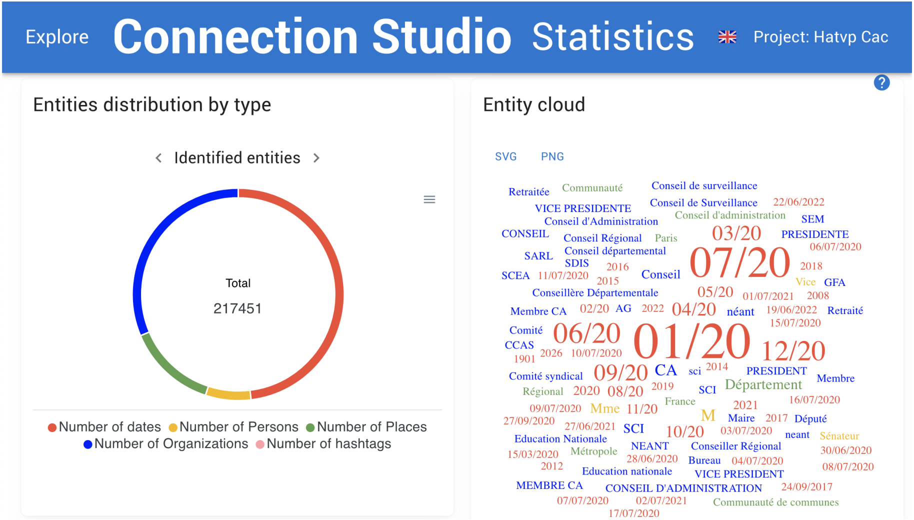
The boundary is truncated due to cyclic collection edges

Entity classification time

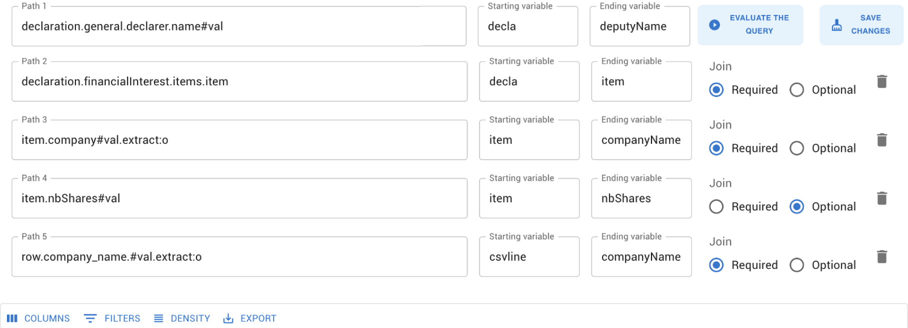
The classification time is composed of:

- Loading the Word2Vec semantic model
    - Constant, 4-8 seconds
- Comparing entity attributes with semantic properties
    - Varies with the number of entities and their number of attributes
    - May vary in a generated dataset of different sizes (different entity roots)
- Computing entity profiles
    - Linear in the input size

A comprehensive data exploration tool for NTUs



A comprehensive data exploration tool for NTUs



decla	deputyname	item	companyname	nbshares	csvline
2660	alain pierre marie rousset	2743	sanofi	1200	352
1470	edouard courtial	1511	lvmh	29013	248
1470	edouard courtial	1543	michelin	162179	261

The EHDEN platform [PdGdK+23, BVD+21]

"European Health Data and Evidence Network"

Consortium of 15 partners across 10 countries

- Mapped their data to the OMOP data model
  - Semi-automatic mapping
  - The system proposes mappings
  - Experts have to select/correct them
- Produced 98 databases
- Asses quality through a DQ (Data Quality Dashboard)

The OHDSI platform [HDS+15]

"Observational Health Data Sciences and Informatics" (said Odyssey, child from OMOP)

International collaboration for open-source data analytics on healthcare networks

Build tools for data exploration and evidence generation

- Achilles: interactive reports and statistics
- Hermes: vocabulary browsing and related searches
- Heracles: build cohorts to assess clinical features on poopulations
- Homer: risk identification by exploring many clinical dimensions

UMG-MeDIC [PSS+23]

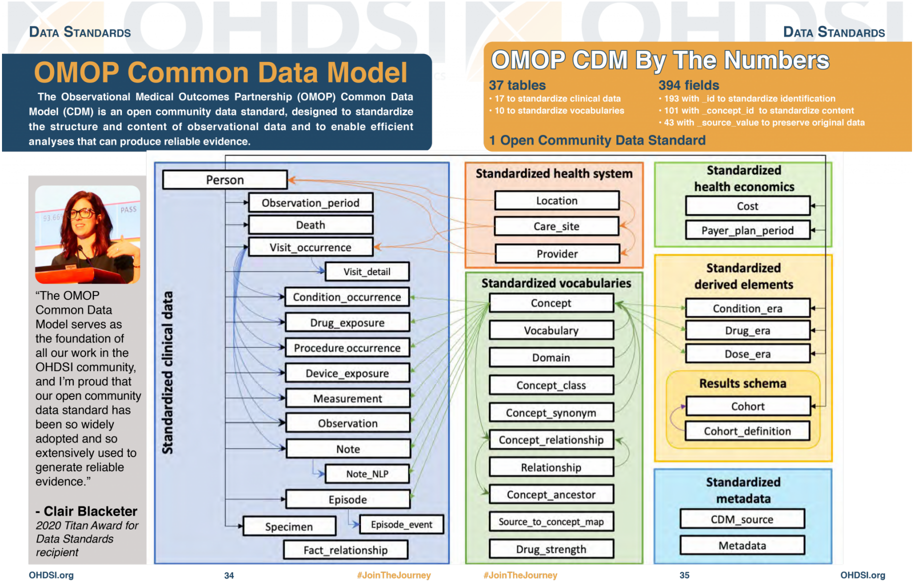
Medical data integration center; relies on Medical Informatics Initiative (MI-I) funds and HiGHmed consortium

Create a technical and legal framework for cross-site secondary use of routine healthcare data

Aim for high compliance with FAIR Principles but data integration workflows are complex and inefficient when done manually

- Operates on a continuous flow of data ($\neq$ individual datasets)
- Periodic integration of new data
- A central relational database with anonymized data
- Combine individual pre-processing tasks into workflows
- Require that each task is documented with “meta-data”

The OMOP data model [SRR+10]



D-ETL [OKK+17]

"Dynamic-ETL": semi-automatic ETL to map source and target data models

- Creation of an ETL specification document (vocabularies, data schema, definitions, conventions)
- Data extraction from initial sources and validation
- D-ETL rules writing ($T_1 \bowtie T_2 \text{ on } T_1.a = T_2.b$)
- Conversion of rules to SQL statements
- Testing rules on data; iterate if not satisfying

Creating new codes with post-coordination

Some healthcare concepts do not have a specific code
SNOMED-CT introduces post-coordination as a compositional grammar

A post-coordinated code = a sequence of existing codes with operators


```
{
  _id: ObjectId('6790ad18d74f95d3f0d0fc82'),
  name: 'tyroid_mother',
  data_type: 'bool',
  dataset_gid: 'http://better-health-project.eu/datasets/a161a1ba-bc51-4cd2-81f4-f8f1b2fed9ec',
  description: 'Mother thyroid disease',
  entity_type: 'phenotypicFeature',
  identifier: 63188,
  ontology_resource: {
    system: 'http://snomed.info/sct',
    code: '14304000:116154003=72705000',
    label: 'Disorder of thyroid gland:Patient=Mother'
  },
  timestamp: ISODate('2025-01-22T08:32:24.000Z'),
  visibility: 'PUBLIC'
}
```