

User-oriented exploration of semi-structured datasets

Nelly Barret

3rd year PhD student

Supervised by Ioana Manolescu

Inria Saclay and Institut Polytechnique de Paris

October 9, 2023



Context: data is the new gold (1/3)



Context: data is the new gold (1/3)



Context: data is the new gold (2/3)

Our digital world comes:

- In **various contexts**: science, health, political life
- At **various scales**: home, city, country, world
- By **different actors**: scientists, businesses, policy makers
- With **different needs**, constraints, abilities

We are **overwhelmed** by (raw) data, we need:

- Data-driven applications
- Data journalism
- Knowledge graphs
- Artificial “intelligence”
- ...



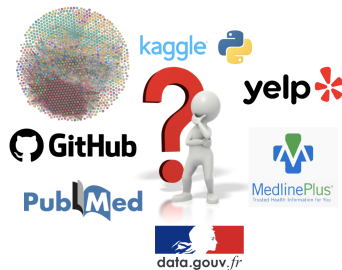
Context: data is the new gold (3/3)

Very **heterogeneous** data:

- Mainly RDF (1K datasets in the LODC)
- Also: XML, JSON, relational, Property Graph...

Detection of **entities** of interest:

- People, Place, email, ...



With **heterogeneous** data, users need:

- 1 A **uniform** integration, *view*
- 2 **Efficient** algorithms and applications
- 3 A global **understanding**, *description*
- 4 Interesting **entity connections**

Create a unique data graph

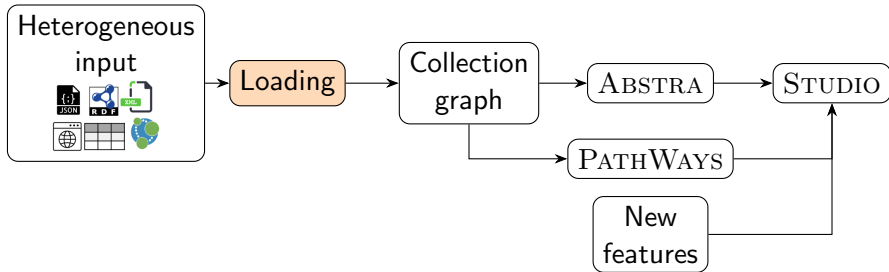
*"A **uniform** integration, view"*

Angelos Anadiotis
IPP, EPFL

Oana Balalau
Inria, IPP

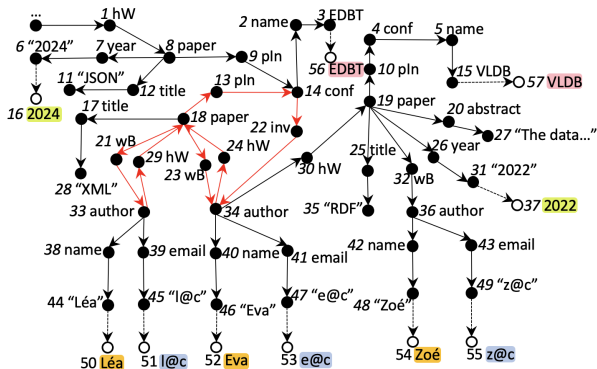
Ioana Manolescu
Inria, IPP

et al...
INSEC, ...



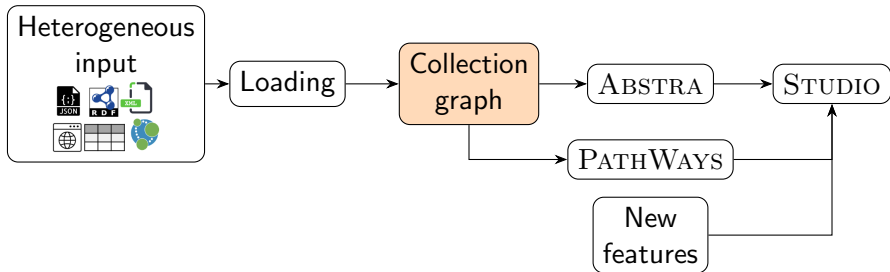
Graph construction

- Ingest any dataset into a **directed graph** (•, →)
- Extract **named entities**, NEs, from the graph values (○, -->):
 - *Temporal*: **date**, time reference
 - *Web*: URI, **email address**, hashtag, Twitter citation
 - *Complex entities*: **People**, Place, **Organization**



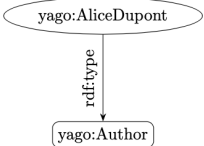
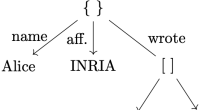
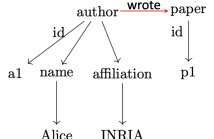
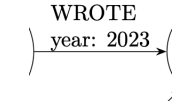
Create a compact representation of the data graph

“Efficient algorithms and applications”



A uniform view of data formats

Each data format has its own specificities:

RDF (URIs)	JSON (ε labels)	XML (ID-IDREF)	PG (edge attrs)
			

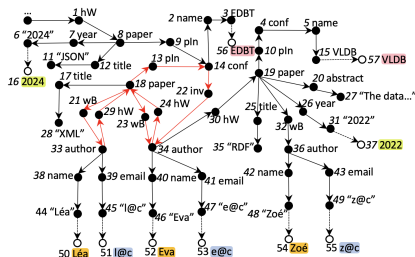
But, we **encode** the same logic:

- **Record:** piece of data, an object
- **Value:** record with no children
- **Same-kind records:** schema or *intuitive* order
- **Relationship:** how records relate

Compact representation (summarization)

Three **equivalence relations**:

- Per label for *XML*
- Per path for *JSON* and *relational data*
- Per type or edge neighbourhood for *RDF* and *PG* [GGM20]



$$\left\{ \begin{array}{l} EC_1 \\ EC_2 \\ EC_3 \\ EC_4 \\ EC_i \end{array} \right\} \left\{ \begin{array}{l} \{N_8, N_{18}, N_{19}\} \\ \{N_{33}, N_{34}, N_{36}\} \\ \{N_4, N_5\} \\ \{N_2, N_5, N_{38}, N_{40}, N_{42}\} \\ \dots \end{array} \right\}$$

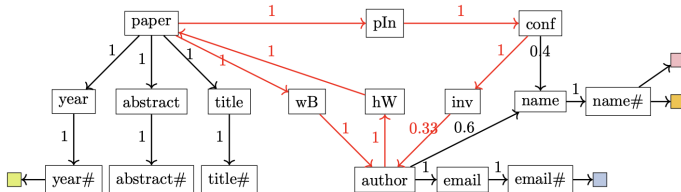
The collection graph

One **collection node** for each equivalence class

One **collection edge** $C_s \rightarrow C_t$:

- Between two collection nodes if a data edge exists
- *Edge transfer factor*: $\frac{|C_t \rightarrow C_s|}{|C_t|}$
- *At-most-one*: 1:1 cardinality

An **entity profile** for each **leaf collection node**: presence of entities



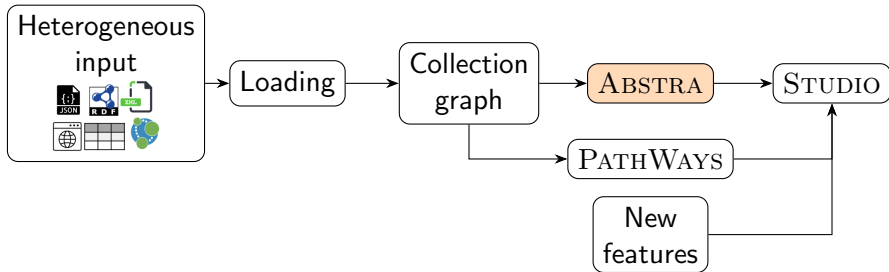
Build an Entity-Relationship model

"A global **understanding**, *description*"

Nelly Barret
Inria, IPP

Prajna Upadhyay
Inria

Ioana Manolescu
Inria, IPP



ABSTRA: get an overview of the data

Problem statement

How to produce a **compact** and **expressive** description out of **any** dataset?

- ① A **high-level, global description**, easy to grasp for **NTUs**
- ② Focus on the data **meaning** more than the **syntax**

⇒ Retrieve / build **the Entity-Relationship model** behind any dataset

ABSTRA: get an overview of the data

Problem statement

How to produce a **compact** and **expressive** description out of **any** dataset?

- 1 A **high-level, global description**, easy to grasp for **NTUs**
- 2 Focus on the data **meaning** more than the **syntax**

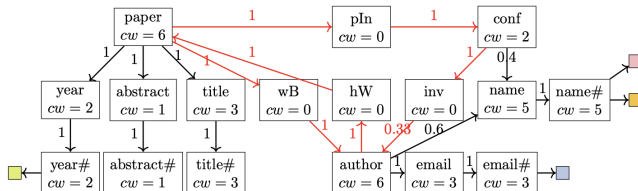
⇒ Retrieve / build **the Entity-Relationship model** behind any dataset

	Data Summarization	Schema inference	Abstra
Several data formats	×	~	✓
Content and structure	~	~	✓
No syntactic detail	✓	×	✓
First-sight discovery	~	×	✓

Main collections selection

Election of *few* main collections, representing mostly the dataset

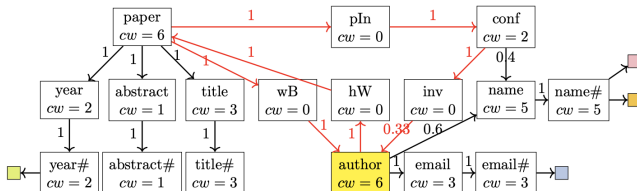
- ① Assign a **weight** to each collection
- ② While less than E_{max} main collections or data coverage $< cov_{min}$
 - ① Pick C^* , the next heaviest collection
 - ② Compute the **boundary** of C^*
 - ③ **Update** the collection graph to reflect the selection of C^*
 - ④ Recompute the weights
- ③ Compute relationships that are connecting the main collections



Collections weights, boundaries and graph updates

Collection weight

- W_{desc_k}
- W_{leaf_k}
- W_{DAG}
- $W_{PageRank}$
- $W_{dwPageRank}$



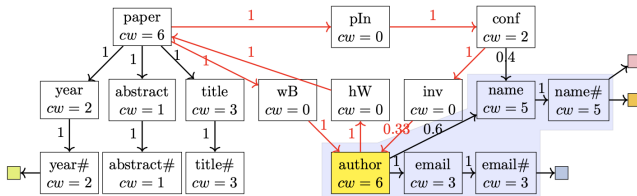
Collections weights, boundaries and graph updates

Collection weight

- w_{desc_k}
- w_{leaf_k}
- w_{DAG}
- $w_{PageRank}$
- $w_{dwPageRank}$

Boundary

- $bound_{desc}$
- $bound_{leaf}$
- $bound_{DAG}$
- $bound_{flood}$
- $bound_{acyclic-flood}$



Collections weights, boundaries and graph updates

Collection weight

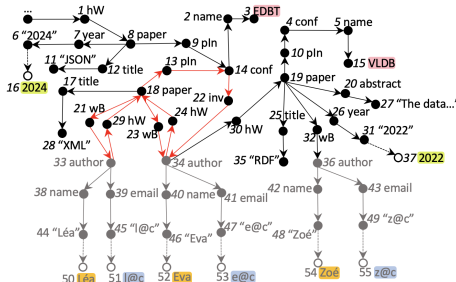
- W_{desc_k}
- W_{leaf_k}
- W_{DAG}
- $W_{PageRank}$
- $W_{dwPageRank}$

Boundary

- $bound_{desc}$
- $bound_{leaf}$
- $bound_{DAG}$
- $bound_{flood}$
- $bound_{acyclic-flood}$

Graph update

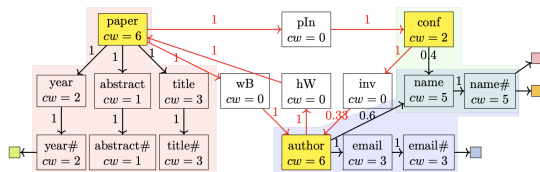
- $update_{boolean}$
- $update_{exact}$



Find relationships between main collections

Possible relationships

The **set of relationships** connecting a pair of collections is the set of their paths.



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

The final output in ABSTRA

<https://team.inria.fr/cedar/projects/abstra/>

ABSTRA description

file:///Users/helly/Desktop/CL-working-dir/tmp/description_xmark1_PR_SFLOOD_GRAPH_ENABLE_SCORE.html

Abstra Home About Help

Here's what your dataset xmark1 contains!

Description

Entities:

A collection of 59486 bidder having the following properties: ⚙️

- data (100%)
- increase (100%)
- time (100%)

A collection of 25500 person having the following properties: ⚙️

- name (100%)
- emailaddress (100%)
- phone (50%)
- homepage (50%)
- creditcard (50%)
- profile (50%) ⚙️
 - interest (294%)
 - education (51%)
 - business (100%)
 - profile@income (100%)
 - gender (50%)
 - age (50%)
- personid (100%)
- address (50%) ⚙️
 - street (100%)
 - city (100%)
 - country (100%)
 - province (49%)
 - zipcode (100%)

A collection of 21750 item having the following properties: ⚙️

- incategory (978%)
- payment (100%)
- itemid (100%)
- shipping (100%)
- quantity (100%)
- description (60%) ⚙️

Entity/Relationship schema

Abstraction of file:///data/datasets/abstraction_data/xmark1 ont (1392794 normalized nodes, 136 collections, 5 main collections, data coverage is 76%) with parameters PROP_PB, ROUND_SFLOOD, UPDATE_EXACT, ENABLE_SCORE

```
graph LR
    person((person (25500))) -- seller --> open_auction[open_auction (12000)]
    person -- buyer --> open_auction
    person -- annotation.author --> open_auction
    open_auction -- winner --> closed_auction[closed_auction (9750)]
    open_auction -- winner --> person
    open_auction -- winner --> item((item (21750)))
    open_auction -- annotation.author --> closed_auction
    closed_auction -- itemref --> item
```

Entities and their properties:

- person (25500)**: name, emailaddress, phone, homepage, creditcard, profile, personid, address
- open_auction (12000)**: quantity, initial, reserve, current, type, interval, open_auctionid, privacy
- closed_auction (9750)**: price, date, quantity, type
- bidder (bidder) (59486)**: data, increase, time
- item (21750)**: incategory, payment, itemid, shipping, quantity, description, mailbox, item@skatand, name, location

Enumerate entity paths

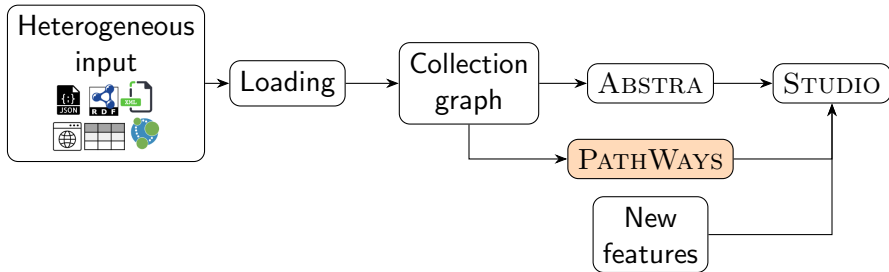
“Interesting **entity connections**”

Nelly Barret
Inria, IPP

Antoine Gauquier
IMT

Jia Jean Law
IPP

Ioana Manolescu
Inria, IPP



PATHWAYS: find interesting connections in the data

Problem statement

How to **interactively** explore **entity connections** in **heterogeneous datasets**?

- 1 No query writing, nor prior knowledge
- 2 A **tabular, high-level output**, easy to grasp for **NTUs**
- 3 Do it **efficiently** even if the data graph is large

⇒ Connect **named entities** (People, Places, ...) **in and across** datasets.

PATHWAYS: find interesting connections in the data

Problem statement

How to **interactively** explore **entity connections** in **heterogeneous datasets**?

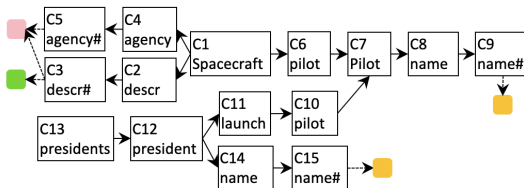
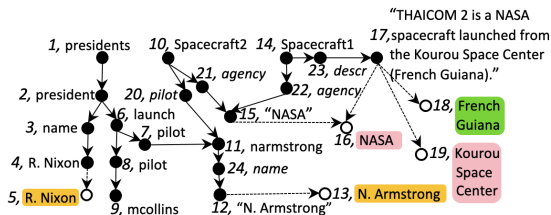
- 1 No query writing, nor prior knowledge
- 2 A **tabular, high-level output**, easy to grasp for **NTUs**
- 3 Do it **efficiently** even if the data graph is large

⇒ Connect **named entities** (People, Places, ...) **in and across** datasets.

	Keyword search	Graph query	Reachability query	PATHWAYS
No query writing	✓	✗	✗	✓
Tabular output	~	~	✗	✓
Efficient	✗	✓	✓	✓

Scenario and terminology

- A **data (entity) path** is a path in the *data graph*
- A **collection (entity) path** is a path in the *collection graph*
- The evaluation of a collection path leads to a set of data paths



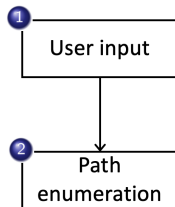
Collection (entity) path enumeration

1

User input

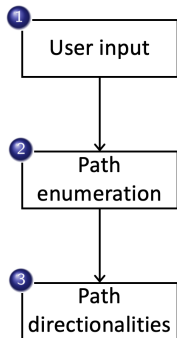
- (τ_1, τ_2) ; max path length; non-specific connections

Collection (entity) path enumeration



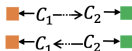
- (τ_1, τ_2) ; max path length; non-specific connections
- **Enumerate** all collection paths using the user input
- **Regardless of edge direction**

Collection (entity) path enumeration

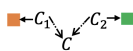


- (τ_1, τ_2) ; max path length; non-specific connections
- **Enumerate** all collection paths using the user input
- **Regardless of edge direction**

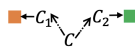
unidirectional



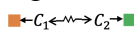
shared-sink



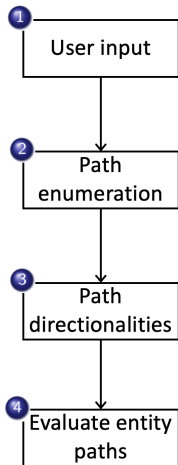
shared-root



general



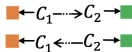
Collection (entity) path enumeration



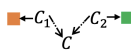
- (τ_1, τ_2) ; max path length; non-specific connections

- **Enumerate** all collection paths using the user input
- **Regardless of edge direction**

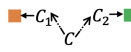
unidirectional



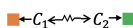
shared-sink



shared-root



general

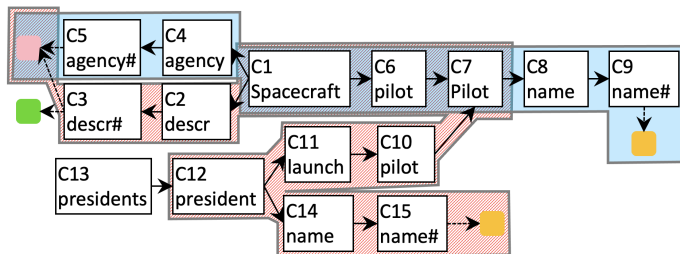


- **Evaluate** selected collection paths into data paths

Optimized data paths computation (1/2)

Assumption: enumerated collection paths (largely) overlap

- There exist common sub-paths between collection paths
- Common sub-paths should be evaluated only once as *views*
 - Saves computation time
- Collection paths are rewritten using *views*
 - Reduces the number of joins



Optimized data paths computation (2/2)

Greedily select the *most profitable* views to materialize

Input: collection paths \mathcal{P} , candidate views \mathcal{V}

Output: a set of views, a set of rewritings

① While there are some $v \in \mathcal{V}$:

- ① For each pair (p, v) , compute $ben(p, v) \leftarrow costEval(p) - costEval(p, v)$
- ② Store v_{max} , the view maximizing $ben(v) \leftarrow \sum ben(p, v) - costMat(v)$
- ③ For each path p , rewrite it, if possible, using v_{max}

■ \leftarrow agency# \leftarrow agency \leftarrow **Spacecraft** \rightarrow **pilot** \rightarrow **Pilot** \rightarrow name \rightarrow name# \rightarrow ■

■ \leftarrow agency# \leftarrow agency \leftarrow **v** \rightarrow name \rightarrow name# \rightarrow ■

```
SELECT le.label, C5.label, C4.label, v.C1label, v.C6label, v.C7label, C8.label, C9.label, re.label
FROM nEntities le, nodes C5, edges C4, view v, edges C8, nodes C9, nEntities re
WHERE le.leafId=C5.id and C4.t=C5.id and C4.s=v.C1id and C8.s=v.C7id and C8.t=C9.id and re.leafId=C9.id
and le.type = ■ and re.type = ■;
```

Data path results in PATHWAYS

<https://team.inria.fr/cedar/projects/pathways/>

PathWAYS

← → ↺ 📄 localhost:8080/gul/pathways 100% ☆ ⬇ 📁 >> ⋮

PathWAYS Home About Help

Load a PathWAYS result from database

pathways_pubmedcoi

- (PERSON, ORGANIZATION), max 100 paths of max size 20 Load
- (PERSON, LOCATION), max 100 paths of max size 20 Load

Run PathWAYS on a dataset

Enter a database name:

Left entity type:

Right entity type:

Run Pathways

Result

Sort queries by length Sort queries by number of associated data paths Hide/show queries without associated data paths

▶ Name#val — Name — Author — AuthorList — PubmedArticle — CoiStatement — CoiStatement#val (860 data paths)

ID	Name#val	Name	Author	AuthorList	PubmedArticle	CoiStatement	CoiStatement#val
2901	Giampiero Mazzaglia	Name	Author	AuthorList	PubmedArticle	CoiStatement	... Bayer ...
2931	Giampiero Mazzaglia	Name	Author	AuthorList	PubmedArticle	CoiStatement	... Pfizer ...
5531	Paolo Angelo Cortesi	Name	Author	AuthorList	PubmedArticle	CoiStatement	... Bayer ...
5561	Paolo Angelo Cortesi	Name	Author	AuthorList	PubmedArticle	CoiStatement	... Pfizer ...

▶ CoiStatement#val — CoiStatement — PubmedArticle — AuthorList — Author — Affiliation — Affiliation#val (480 data paths)

▶ Name#val — Name — Author — AuthorList — PubmedArticle — ArticleTitle — ArticleTitle#val (8 data paths)

▶ Name#val — Name — Author — Affiliation — Affiliation#val (71 data paths)

▶ CoiStatement#val — CoiStatement — PubmedArticle — ArticleTitle — ArticleTitle#val (12 data paths)

Authors: Nelly Barret @ Inria, Antoine Gauquier @ IMT Nord Europe, Jean Law @ Ecole Polytechnique, Ioana Manolescu @ Inria

Main contact: nelly.barret@inria.fr

Nelly Barret (Inria)

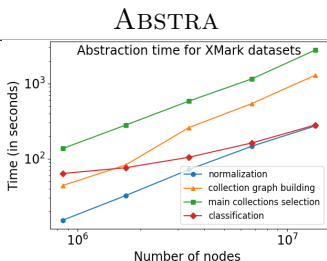
Semi-structured Data Exploration

October 9, 2023

29 / 38

Quick overview of experiments

On widely-used **open data formats**: JSON, RDF, XML and PG.



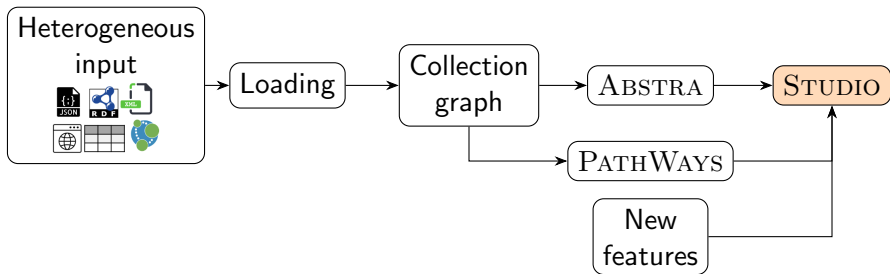
- User study
- Comparison to schemas

PATHWAYS

(τ_1, τ_2)	T_0	$T = T_R + T_{Q_{NV}}$	$s = T_0/T$
(τ_P, τ_O)	250.36	4.10	61×
(τ_P, τ_L)	37.29	19.38	2×
(τ_L, τ_O)	151.29	20.47	7×
(τ_P, τ_P)	152.59	44.27	3×
(τ_L, τ_L)	169.64	71.63	2×
(τ_O, τ_O)	317.92	23.24	13×

- # paths: 0 to very high
- 4 path “shapes”

Future work, takeaways and open questions



Future work: STUDIO

STUDIO: a data lake for ingesting, querying, cleaning and understanding heterogeneous data

- French media are interested (DataJournos, CFI)

Connection Studio Projects



Sort by

Project name

END SESSION

CREATE A PROJECT

Project Cac

1 files

Created on: 2023-07-13 11:32:13

Latest file addition: 2023-07-13 11:32:13

MANAGE

Project Cac40

1 files

Created on: 2023-07-05 16:12:38

Latest file addition: 2023-07-05 16:12:38

MANAGE

Project Hatvp Cac

2 files

Created on: 2023-07-11 16:03:48

Latest file addition: 2023-07-11 16:39:39

MANAGE

Project Hatvp Cac40

2 files

Created on: 2023-07-05 15:46:07

Latest file addition: 2023-07-05 16:25:52

MANAGE

Project Hatvpssmall

No files uploaded yet, add one!

MANAGE

Project Pubmed

1 files

Created on: 2023-07-05 09:46:07

Latest file addition: 2023-07-05 09:46:07

MANAGE

Project Recac40

1 files

Created on: 2023-07-12 23:24:56

Latest file addition: 2023-07-12 23:24:56

MANAGE


Future work: STUDIO

STUDIO: a data lake for ingesting, querying, cleaning and understanding heterogeneous data


Explore


Connection Studio


Uploaded files





 Project: Hatvp Cac

Uploaded files

 ADD



 DISPLAY ADVANCED OPTIONS

ID	File	Path	Type	Creation date	
1	hatvp-cleaned.xml	file:/Users/helly/Documents/boulot/theseNelly/connection-lens/.connection-studio/demo-CFI	XML	2023-07-11 16:03:48+02	 
2	Cac40.csv	file:/Users/helly/Documents/boulot/theseNelly/connection-lens/.connection-studio/demo-CFI	CSV	2023-07-11 16:39:39+02	 

Future work: STUDIO

STUDIO: a data lake for ingesting, querying, cleaning and understanding heterogeneous data

Explore

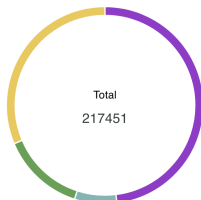
Connection Studio Statistics



Project: Hatvp Cac

Entities distribution by type

< Identified entities >



Total
217451

● Number of dates ● Number of Persons ● Number of Places
● Number of Organizations ● Number of hashtags

Entity cloud




Future work: STUDIO

STUDIO: a data lake for ingesting, querying, cleaning and understanding heterogeneous data

Explore

Connection Studio

Data view 

Project: Hatvp Cac

Select a file

hatvp-cleaned.xml

?

Select a path

declarations.declaration.general.dateDebutMandat#val.extract:d

Show the query

EVALUATE THE QUERY

SAVE CHANGES

Path 1

declaration.general.declarant.nom#val

Starting variable

decla

Ending variable

name

Path 2

declaration.general.mandat.label#val

Starting variable

decla

Ending variable

mandateType

Join

☒ Required ☐ Optional

Path 3

declaration.general.dateDebutMandat#val.extract:d

Starting variable





decla

Ending variable

mandateStart

Join

☒ Required ☐ Optional

 COLUMNS  FILTERS  DENSITY  EXPORT

decla	name	mandatetype	mandatestart
237676	abbassia hakem	elu local ou membre d'un établissement public de coopération intercommunale	03/07/2020
1836220	abdlatif ammar	elu local ou membre d'un établissement public de coopération intercommunale	10/07/2020

Future work: STUDIO

STUDIO: a data lake for ingesting, querying, cleaning and understanding heterogeneous data

Explore

Connection Studio Search

Project: Recac40

Airbus Engie



DISPLAY ADVANCED OPTIONS

Always display edge labels

RESULTS

NEIGHBORS

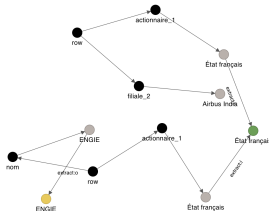
N°2

13 Nodes found
From 1 source
Score: 0.25

N°3

7 Nodes found
From 1 source
Score: 0.25

N°4



● Data node
● Location
● Organization


Takeaways and open questions

- ABSTRA: a dataset abstraction system for heterogeneous data
- PATHWAYS: an entity-focused exploration system
- STUDIO: a user-oriented data lake for data exploration


ABSTRA	PATHWAYS	STUDIO
EDBT 2024	ADBIS 2023	CoopIS 2023
		

Further opportunities

Nelly BARRET

 nelly.barret@inria.fr

 <https://pages.saclay.inria.fr/nelly.barret/>

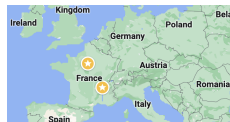
 Inria Saclay & Institut Polytechnique de Paris
Palaiseau



LYON



PALaiseau



?