# Nelly Barret

Post-doctoral researcher at Politecnico di Milano, Italy

https://nelly-barret.github.io
https://orcid.org/0000-0002-3469-4149
https://www.linkedin.com/in/nelly-barret/

## 1 Curriculum Vitae

### 1.1 Personal information

- Nelly Barret
- Living Via Monfalcone, 42 - 20132 Milano (MI), Italy
- With French address being 31 rue colin, 69100 Villeurbanne, France
- Reachable on pro.nelly.barret@gmail.com

### 1.2 Education

| Date | Position | Laboratory | Funding |
|---|---|---|---|
| 04/2024 - 04/2025 | Post-doc | Politecnico di Milano, DEIB | European project |
| 01/2021 - 03/2024 | PhD | École Polytechnique & Inria Saclay, CEDAR team | Paris regional funds |
| 10/2020 - 12/2020 | CDD relais thèse | Inria Saclay, CEDAR team | Inria |
| 02/2020 - 07/2020 | M2 internship | LIRIS, BD team | LabEx IMU |
| 05/2018 - 07/2018 | L3 internship | LIRIS, BD team | LabEx IMU |

#### 1.2.1 Post-doc: distributed analyses of heterogeneous healthcare data

European project n°101136262 titled "BETTER: Better real-world health-data distributed analytics research platform" in the frame of the prgramme "Horizon Europe". Close collaborators:

- Pietro Pinoli, associate professor in DEIB at Politecnico di Milano (Milan, IT)
- Anna Bernasconi, associate professor in DEIB at Politecnico di Milano (Milan, IT)
- Boris Bikbov, medical doctor in DEIB and Politecnico di Milano (Milan, IT)

#### 1.2.2 PhD: user-oriented exploration of semi-structured datasets

PhD thesis funded by the Île-de-France region thanks to the DIM RFSI funds. My work has been awarded the accessit (2nd place) during the French conference BDA 2024. This awards one or two outstanding PhD theses in the data management community.

| | | |
|---|---|---|
| Ioana Manolescu | Research director<br>Inria Saclay and École Polytechnique (Palaiseau, FR) | PhD advisor |
| Karen Bastien | CEO of WeDoData | PhD co-advisor |
| Fatiha Saïs | Professor<br>Université Paris Saclay and LISN (Palaiseau, FR) | Présidente |
| Jean-Marc Petit | Professor<br>INSA Lyon and LIRIS (Villeurbanne, FR) | Rapporteur |
| Olivier Teste | Professor<br>Université Toulouse Jean Jaurès and IRIT (Toulouse, FR) | Rapporteur |
| Katja Hose | Professor<br>TU Wien (Vienne, AT) | Examinatrice |
| Stefano Ceri | Professor<br>Politecnico di Milano (Milan, IT) | Examinateur |
| Fatemeh Nargesian | Associate professor<br>Université de Rochester (New York, USA) | Examinatrice |

#### 1.2.3 Contrat pré-thèse: interpretation of complex objects in ConnectionLens graphs

Short term contract (3 months) to evaluate is the PhD fits the expectations of both the student and the advisor, while taking into account the COVID-19 (remote work mandatory).

#### 1.2.4 Master, AI track: predicting the environment of a neighbourhood

Master conducted between Sept. 2018 and July 2020, including 6 months of internship (Feb. to July 2020).

Internship funded by the LabEx IMU (Laboratoire d'Excellence – Intelligences des Mondes Urbains) and pursuing my Bachelor internship. Advisors:

- Fabien Duchateau, associate professor at Université Lyon 1 and LIRIS (Villeurbanne, FR)
- Franck Favetta, associate professor at Université de Lyon and LIRIS (Villeurbanne, FR)

During this multidisciplinary internship, I often collaborated with Nelly Duong, the CEO of Home in Love, a start-up helping salary mobility employees to find a new place to live based on their criteria and needs (neighbourhood, family, hobbies, ...), as well as with Behnaz Jullien (interne in psychology at Université Lyon 2), Wissame Laddada (post-doctoral researcher in computer science at Université Lyon 1), and Ludovic Moncla (associate professor in computer science at INSA Lyon).

#### 1.2.5 Bachelor: integrating geographic data for neighbourhood recommendation

Bachelor conducted between Sept. 2015 and July 2018, including 3 months of internship (May to July 2018).

Internship also funded by the LabEx IMU, and in collaboration with Nelly Duong (CEO of Home in Love), Loïc Bonneval (associate professor in sociology at Université Lyon 2) and Aurélien Gentil (PhD student in sociology at Université Lyon 2). Advisors:

- Fabien Duchateau, associate professor at Université Lyon 1 and LIRIS (Villeurbanne, FR)
- Franck Favetta, associate professor at Université de Lyon and LIRIS (Villeurbanne, FR)

### 1.3 Awards

My PhD work has been awarded by the BDA conference via an accessit (2nd place) to the PhD award, awarding one or two PhD with significant contributions in the domain of data management.

# 2 Collective responsibilities

## 2.1 Member of conference committees

**1 organization committee:** CMLS 2025 (workshop – conference ER [A])

**4 program committee:** DBKDA 2025 [N/A], ADBIS 2025 [C], iSAILS 2025 (workshop – conference CAiSE [A]), demos BDA 2025 [national]

## 2.2 Review activity

I have been an external reviewer for the conferences and journals in *italic*. I have often made several reviews for the journal of a given year.

**6 journals:** GigaScience 2025 [Q1], JDIQ 2025 [Q2], CMC 2025 [Q3], PLOS ONE 2024 [Q1], BMC Medical informatics and Decision Making 2024 [Q1], Scientific Reports 2024 [Q1]

**4 conferences:** *ICDE 2025 [A*]*, *CIDR 2023 [A]*, *EDBT 2023 [A]*, *ICWE 2021 [B]*

## 2.3 Participation to working groups

**Diversity, Equity, Inclusion.** Since 2023, I co-lead the SCOUT action of the DEI working group (Diversity, Equity, Inclusion) with Madhulika Mohanty (young researcher at Inria Saclay) and Sujaya Maiyya (assistant professor at Université de Waterloo). This action aims at facilitating DEI efforts in the database community. My main proposal is a list allowing authors to check whether their submission is inclusive. We are now integrating it to EasyChair and CMT. We then extended this list to oral presentations (notably for conferences). All members of the DEI initiative are preparing an activity report, which should be submitted to the journal SIGMOD Record.

**FAIRification of genomic annotations.** Since end of 2024, I am a member of the working group "FAIRification of genomic annotations" being part of RDA (Research Data Alliance). This working group aims at defining novel methods to make genomic annotations and data more interoperable. I co-lead two actions: (*i*) define a strategy to publicly and permanently store harmonized genomic metadata with Svainung Gundersen (computer scientist at ELIXIR) and Evan Christensen (PhD in bio-informatics at Université d'Utah); and (*ii*) define new models for representing genomic annotations with Sveinung Gundersen and Adam Wright (researcher at Cancer institute of Ontario).

## 2.4 Dissemination and vulgarization

The slides are available on my web page: `https://nelly-barret.github.io/talks.html`

I presented my PhD and post-doctoral researches in many teams, including:

- "Heterogeneous datasets: a tale of integration and exploration" – BD team, LIRIS, Lyon – Jan. 2025
- "Integrating and exploring heterogeneous datasets" – DS team, DEIB, Milan – April 2024
- "Semi-structured data user exploration" – DS team, LIB, Dijon – Jan. 2024
- "Semi-structured data user exploration" – LAHDAK team, LISN, Paris Saclay – Oct. 2023

I also vulgarized my research with different audiences:

- "Real-world health-data distributed analytics across hospitals" – Girls@CSE[1] – 2024

[1]Event inviting female researchers in computer science to inspire young female to pursue a scientific career.

- "From data to journalism" – Lycée international de Palaiseau[2], Palaiseau – 2024
- "Artificial Intelligence: a tool for journalism" – Forum CFI[3] and DataJournos[4], Paris – 2023
- "Research in data integration: the ConnectionLens case" – RJMI[5], online – 2022

### 2.5 Institutional responsibilities

| Date | Service |
|---|---|
| 2024-now | Manage the LinkedIn page of the "Data Science" group (DEIB) at Politecnico di Milano where I regularly post content (PhD defenses, interviews, exciting research results) |
| 2021-2024 | Manage the CEDAR team seminars at Inria Saclay |
| 2023-2024 | Represent Inria PhD students in Institut Polytechnique de Paris |
| 2021 | Co-develop and deploy the website for internships at École Polytechnique, in collaboration with the École IT team |

## 3 Teaching

### 3.1 Teaching assistant

Table 1 lists teaching duties I did. They have been done at École Polytechnique as a teaching assistant. My contribution to the internship website for École Polytechnique in fall 2021 has been converted to 24h of teaching duty for its beneficits.

| Name | Year | Public | Level | Volume | Nature | Language |
|---|---|---|---|---|---|---|
| CS project | Spring 2024 | Bachelor | L2 | 4h | Defenses | EN |
| Machine Learning | Spring 2023 | Bachelor | L2 | 20h | TP | EN |
| Programmation basics | Fall 2022 | Ingénieur | 2A | 40h | TP | FR |
| | | | | 4h | Tutorat | FR |
| Object-oriented programming | Spring 2022 | Ingénieur | 1A | 36h | TP | FR |

Table 1: List of teaching duties that I conducted at École Polytechnique as a teaching assistant. 1A and 2A refer to first and second year of the Ingénieur cursus, while L2 refers to the second year of the Bachelor cursus.

### 3.2 Intern supervision

| Name | Affiliation | Level | Year | Topic |
|---|---|---|---|---|
| 1. Shay Pripstein | École Polytechnique | Bachelor 3A | Winter 2023 | Target data acquisition in open repositories |
| 2. Tudor Enache | École Polytechnique | Bachelor 3A | Winter 2023 | From simple to property graphs |
| 3. Nikola Dobricic | École Polytechnique | Bachelor 2A | Summer 2023 | Graph queries for abstractions |
| 4. Jia-Jean Law | École Polytechnique | Ingénieur 3A | Winter 2022 | Optimization of entity-to-entity data paths |
| 5. Antoine Gauquier | IMT Nord-Europe | M1 | Summer 2022 | Path exploration of ConnectionLens graphs |

Table 2: List of interns I co-advised during my PhD.

During my PhD, I co-advised **5 students** with Ioana Manolescu (PhD advisor) and Madhulika Mohanty (young researcher at CEDAR) – see Table 2. They worked on ConnectionLens [ABC+22], Abstra [C1], Pathways [J1], et ConnectionStudio [C2] (see Section 4.2.2).

[2]The international highschool of Palaiseau opened a Cycle Pluridisciplinaire d'Enseignement Supérieur (CPES), being like a Bachelor, in the domains of Data for Science, Society and Health. https://www.lipsp.fr/le-post-bac-au-lipps/

[3]Canal France International is a French media development agency. They organised a forum for media, journalists and computer scientists. https://cfi.fr/fr

[4]A consortium of 40 French journalistes eager to include more data science in their works. https://datajournos.fr

[5]The Rencontres Jeunes Mathématiciennes et Informaticiennes are an event between young female whose aspirations are to pursue scientific careers and female researchers in computer science and mathematics. https://filles-et-maths.fr/rjmi/

1. We automatically matched named entities from ConnectionLens graphs with the corresponding Wikidata page. I participated to the weekly meetings of Shay, explained him how to conduct experiments, and helped him for his final report. Estimated total: **50%**.

2. We converted Abstra E-R diagrams to property graphs. I participated to Tudor's weekly meetings and read his report; I also had additional meetings with him to help him understand well the his internship topic. Estimated total: **50%**.

3. We allowed Abstra users to simply formulate queries using the E-R diagrams. I participated to some of the Nikola's weekly meetings, especially to integrate his work in ConnectionStudio and answer his questions on Abstra. Estimated total: **30%**.

4. We proposed a multi-query optimization algorithm for evaluating more efficiently the paths enumerated in Pathways. I helped Jia-Jean in the implementation of the algorithm and its integration in Pathways during individual meetings. Estimated total: **40%**.

5. This works corresponds to the premisses of Pathways where we enumerate the set of paths connecting named entities of interest. I regularly co-advised Antoine by answering his questions and help him with the implémentation. Estimated total: **30%**.

# 4 Research

## 4.1 Synthetic description of my research topics

My research works lie in the **integration, management and exploration of large, complex, heterogeneous data, often using AI tools**. The projects that I worked on have always been part of multidisciplinary contexts: urban planning, journalism, and health. Here, I briefly describe the main research topics that I enjoy working on and their main challenges.

**Integrating heterogeneous and complex data.** There exist mainly three approaches for data integration. The **mediator** [Wie92] converts data sources towards a conceptual model (e.g., the conceptual model of an app like Google Maps features entities such as *GeographicEntity*, *PointOfInteret*, *Trip*, etc.) using a set of rules. It brings a virtual integration to guarantee data freshness but requires to define many rules as well as pay the cost of conversion at each access/query. The **warehouse** [DM88] operates a physical integration of the data towards a conceptual model, which allows for very efficient access to the data. However, it does not fit dynamic data (periodic build). Finally, the **data lake** is introduced as a compromise between mediators and warehouses. A general definition is "flexible and scalable system which stores and manages raw, heterogeneous data coming in their original model and provides maintenance, querying and on-the-fly analyses with the help of rich metadata" [Hai+23]. A data lake is "schema-less" in the sense that residing data do not need to conform to a pre-defined schema (on the contrary to mediators and warehouses). Nevertheless, they are recent, require deep improvements in terms of research and development, notably on the representation, semantic and querying. These tasks are more complex because of the lack of structure and the few (or no) interoperability; thus, proposed algorithms and techniques have to be general, leveraging few or no hypotheses on the form and the quality of the data. Moreover, those approaches start to include AI tools but are not yet ready for a decentralized context.

**Interoperability between data and actors.** The lack of interoperability is becoming a major issue because data are produced independently by different actors and are shared without being homogenized, this limiting cooperation between actors, systems, and data. Fortunately, it is possible to increase the homogenization and interoperability of data. When domain experts do not have much IT skills, it is often common that they make data more interoperable manually, e.g., by modifying input data to homogenize names/values. However, this approach is only possible when data is small and are not dynamic. This is why automating (completely or partially) interoperability is crucial. Several approaches exist and choosing one mainly depends on the initial data quality and the wished automation level. Existing techniques include converting data to a unique data model [Put+23; Ong+17] (often use-case tailored) and the matching of concepts present in the data to those

described in dedicated vocabularies[6] or ontologies[7]. While mainly being focused on output data, the evaluation of the level and quality of interoperability should be tracked along the process [CMP24], notably to detect/fix errors in interoperabilization. This effort of improving interoperability takes place in the FAIR principles [Wil+16] (*Findable, Accessible, Interoperable, Reusable*), a set of guidelines for creating interoperable systems, based on standards, reusing concepts already defined in specialized vocabularies, etc. Interoperability is a crucial aspect for federated learning architectures and AI techniques, without which algorithms and data cannot cooperate.

**Data visualization, querying and analyses.** Finally, the two axes presented above would be as impactful without tools (interfaces) dedicated to users, whether they are domain experts or newbies. Those interfaces may take varied forms (Web application, algorithms, etc.) and be specific to an application domain or not. They include means to visualize and/or query and/or analyze data, but are often not extensible and made by and for researchers.

## 4.2 Description of 4 main works

Table 3 describes the whole set of my research contributions. In this section, I will detail 4 of them (Better, Abstra, Pathways and Predihood).

| Education | Project name | Architecture | AI | Publications | # cite |
|---|---|---|---|---|---|
| Post-doc | Better | Warehouse network | Learn. & Pred. | *[Z1; Z2] (under review)* | - |
| PhD | Abstra | Graph data lake | LLM | [M1; D2; C4; C1; W1] | 12 |
| | Pathways | Graph data lake | LLM | [M1; J1; C3; D1] | 4 |
| | ConnectionStudio | Graph data lake | - | [C2; N1; W2] | - |
| Master 2 | Predihood | Warehouse | Prediction | [J2; J3; C5] | - |
| Master 1 | GeoAlign | Mediator | - | [D3] | 3 |
| Licence | VizLiris | Warehouse | Pred. & Recomm. | [N2] | - |

Table 3: List of conducted projects in research, with the architecture type and the type of used AI tool. Abbreviations: learn=learning, pred=prediction, recomm=recommandation. # cite is the total number of citations.

### 4.2.1 Better: AI analyses of heterogeneous healthcare data

Better is a European project which will lead to: (*i*) a network of interoperable warehouses in partner hospitals; (*ii*) a decentralized and secured platform for data exploration, querying, and to create learning algorithms for the federated analyses of the warehouses in the network. The 3 main contributions of this project are:

**Two general conceptual models for data and metadata.** We have first proposed a general conceptual model to represent varied healthcare data. It build upon the notions of *feature* (observed variable) and *record* (value obtained for a given feature), which allows to represent all types of data brought by hospitals in a similar manner. Moreover, each feature is associated to a unique identifier (described in specialized vocabularies, e.g., SNOMED-CT for clinical features, ou OrphaNet for rare diseases). The main advantage of this is to group different versions of a single concept, e.g., "sexe" in French and "sex" in English are two linguistic versions of the same concept. Next, we have proposed a conceptual model for the metadata, which facilitates the gathering of knowledge experts have on the features (description, unit, type, etc.).

**An algorithm to convert existing data to our model.** We have next developed an ETL algorithm (Extract-Transform-Load) to automatically create a (more) interoperable database from the input datasets. It build on our two general conceptual models and is, *a fortiori*, general.

**A catalogue for exploration.** Our third contribution is a catalogue which: (*i*) lists ingested datasets and resulting databases; (*ii*) allows to easily formulate queries to inspect data more precisely; (*iii*) facilitates the creation of federated AI algorithms. This catalogue is the entry point for medical experts who want to better understand

[6]Set of concepts/values, each of them being properly defined and associated to a unique identifier.
[7]Model which, for an application domain, defines concepts, their properties and relationships.

rare diseases of interest. Of course, the query engine and the federated algorithmes are executed in a secured environment, being of type PHT [Bey+20] (Personal Health Train).

#### 4.2.2 Abstra and Pathways: user-oriented exploration of graphs built from complex and heterogeneous data

ConnectionLens [ABC+22] is a data lake which ingests structured (tables), semi-structured (JSON, XML, RDF, etc), and unstructured (text) data. Internally, it uses a graph to represent the different data source. However, the obtained graph is often large and complex, thus its exploration is tedious for users, even more id they are not familiar with the data or IT. To help them in their exploration task, my PhD work proposes:

**Automatique structured summary (Abstra).** We build structured summaries looking like Entity-Relationship diagrams starting from the graphs produced by ConnectionLens. For this, we first adapted existing structural summarization techniques [GW97; Baa+19; GGM20] to obtain collections (groups) of equivalent nodes. Next, we have proposed a greedy algorithm which: (*i*) assigns a score to each collection; (*ii*) elects the collection with the highest score as an entity; (*iii*) determines collections representing the attributes of the elected entity; (*iv*) start again without elected collections. Relationships between entities are next computed by enumerating their connections in the graph. We defined several scoring functions, the most advanced one combining the graph topology (using PageRank [Pag+99]]) and the data distribution in the graph. Likewise, we defined several methods to determine the attributes of selected entities. To finish, we also developed a classification module which assigns a semantic category to each entity; this is particularly useful when the entities are unlabelled. The result is shown as en E-T diagram, showing entities, their attributes (possibly nested) and their relationships.

**Enumeration and ranking of named entity paths (Pathways).** We have developed a module based on the LLM ChatGPT to extract named entities (people names, places, companies, etc.) in the graph value nodes. Next, we have proposed a method which enumerates all paths connecting two types of NEs, e.g., all paths connecting people and companies. This method builds on: (*i*) an algorithm which efficiently enumerates all paths in the Abstra summary; (*ii*) rank the summary paths according to their reliability ("are the found NEs correct?") and their force ("is the information diluted over the path?"); (*iii*) evaluates the paths that are enough reliable and strong on the underlying data graph. The result presents the "best" paths of the summary and their associated data paths evaluated on the real graph.

#### 4.2.3 Predihood: supervision prediction of the environment of neighbourhoods

The multidisciplinary project Home in Love aims at facilitating the real estate research during professional mobility, e.g., transfer or work-study program. The two main research questions of this project are: (*i*) how to simply qualify the environment of a neighbourhood?; and (*ii*) how to predict the neighbourhood environment using supervised learning? For this, we have at hand around 600 indicators for each of the 50k French neighbourhoods provided by the INSEE[8]. However, so many indicators do not allow us to simply describe a neighbourhood, nor to make good and explainable predictions. We first proposed six environment variables (EV), each with a limited number of values, e.g., the EV "landscape" takes values in "urban", "green areas", "forest" ou "countryside". Next, we proceeded as follows:

**Indicator selection for AI prediction.** We first proposed an algorithm which, for each EV, computes lists of indicators being the most useful to its prediction. This algorithm removes unnecessary indicators (descriptive or too specific). Next, it generates 7 lists containing from 10 to 100 indicators: this allows to only use the most important indicators for each EV and improves explainability. The selection of indicators builds on the combination of 3 dedicated techniques: a correlation matrix to detect fully correlated indicators, as well as the algorithms RandomForest and Extra Tree Classifier to identify indicators that are ranked as highly important in the prediction by both algorithms. Those techniques also allow to quantify the importance of an indicator in the prediction on top of improving explainability).

[8]Institut National de la Statistique et des Études Économiques

**Supervised prediction of EV.** After manually annotating 300 neighbourhoods (over 50k), our approach can predict the 6 EV (building type, usage, landscape, social class, morphologic position, geographic position) for any neighbourhood in France. After having chosen a neighbourhood, users select a prediction algorithm which is used to predict the value of the 6 EV according to the 7 indicator lists. Next, the most frequent predicted value for each EV is assigned as the final value.

**Generic interface for comparing AI algorithms.** In order to facilitate the tuning and comparison of AI algorithms (mostly for prediction or recommendation), we also designed a generic interface allowing to use, test, and compare the Scikit-learn algorithms as well as home-made ones. The genericity allows to easily add hand-made algorithms and to use personal datasets, which highly facilitates the testing phase of such algorithms.

## 4.3 Scientific publications and softwares

My publications sum to **18** and are all available onmy ORCID record[9]. For each publication, the first author is underlined and top-tier conference in my research domain are **in bold**. Titles are links to the online articles.

### 4.3.1 Peer-reviewed international journals

[J1] Nelly Barret, Antoine Gauquier, Jia-Jean Law, and Ioana Manolescu. "Finding meaningful paths in heterogeneous graphs with PathWays". In: *Information Systems (**IS 2024**) [Q1]* (2025).

[J2] Nelly Barret, Fabien Duchateau, and Franck Favetta. "Predihood: an open-source tool for predicting neighbourhoods' information". In: *Journal of Open Source Software (JOSS 2021) [score d'impact: 3]* (2021).

[J3] Nelly Barret, Fabien Duchateau, Franck Favetta, Aurélien Gentil, and Loïc Bonneval. "An Environmental Study of French Neighbourhoods". In: *Data Management Technologies and Applications (CCIS 2021) [Q4]*. 2021.

### 4.3.2 Peer-reviewed international conferences

[C1] Nelly Barret, Ioana Manolescu, and Prajna Upadhyay. "Computing Generic Abstractions from Application Datasets". In: *International Conference on Extending Database Technology (**EDBT 2024**) [A]*. 2024.

[C2] Nelly Barret, Simon Ebel, Théo Galizzi, Ioana Manolescu, and Madhulika Mohanty. "User-friendly exploration of highly heterogeneous data lakes". In: *International Conference on Cooperative Information Systems (CoopIS 2023) [B]*. 2023.

[C3] Nelly Barret, Antoine Gauquier, Jia Jean Law, and Ioana Manolescu. "Exploring Heterogeneous Data Graphs Through Their Entity Paths". In: *European Conference on Advances in Databases and Information Systems (ADBIS 2023) [C]*. 2023.

[C4] Nelly Barret, Ioana Manolescu, and Prajna Upadhyay. "Abstra: toward generic abstractions for data of any model". In: *ACM International Conference on Information & Knowledge Management (**CIKM 2022**) [A]*. 2022.

[C5] Nelly Barret, Fabien Duchateau, Franck Favetta, and Loïc Bonneval. "Predicting the Environment of a Neighborhood: A Use Case for France". In: *International Conference on Data Science, Technology and Applications (DATA 2020) [C]*. 2020.

[9] Exception the two ongoing submissions, which are only accessible on my web page.

#### 4.3.3 Peer-reviewed international workshops

[W1] 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 in International Conference on Data Engineering (**ICDE 2024**) [A*]*. 2024.

[W2] Oana Balalau, Nelly Barret, Simon Ebel, Théo Galizzi, Ioana Manolescu, and Madhulika Mohanty. "Graph lenses over any data: the ConnectionLens experience". In: *SEAGRAPH workshop in International Conference on Data Engineering (**ICDE 2024**) [A*]*. 2024.

#### 4.3.4 Peer-reviewed national conferences

[N1] Nelly Barret, Simon Ebel, Théo Galizzi, Ioana Manolescu, and Madhulika Mohanty. "Exploration utilisateur de lacs de données très hétérogènes". In: *Conférence Francophone sur l'Extraction et la Gestion des Connaissances (EGC 2024) [national]*. 2024.

[N2] Nelly Barret, Fabien Duchateau, Franck Favetta, Maryvonne Miquel, Aurélien Gentil, and Loïc Bonneval. "À la recherche du quartier idéal". In: *Conférence Francophone sur l'Extraction et la Gestion des Connaissances (EGC 2019) [national]*. 2019.

#### 4.3.5 Demonstrations

[D1] Nelly Barret, Antoine Gauquier, Jia Jean Law, and Ioana Manolescu. "PathWays: entity-focused exploration of heterogeneous data graphs". In: *The Semantic Web: ESWC Satellite Events (ESWC 2023) [B]*. 2023.

[D2] Nelly Barret. "Facilitating Heterogeneous Dataset Understanding". In: *Conférence sur la Gestion de Données – Principes, Technologies et Applications (BDA 2020) [national]*. 2021.

[D3] Nelly Barret, Fabien Duchateau, Franck Favetta, and Ludovic Moncla. "Spatial entity matching with GeoAlign". In: *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL 2019) [A]*. 2019.

#### 4.3.6 Under review

[Z1] Nelly Barret, Anna Bernasconi, Boris Bikbov, and Pietro Pinoli. "I-ETL: an interoperability-aware health (meta)data pipeline to enable federated analyses". In: *Under minor revision at BMC Medical Informatics and Decision Making (**BMC 2025**) [Q1]*. 2025.

[Z2] Nelly Barret, Anna Bernasconi, Cinzia Cappiello, Giacomo Palu, and Pietro Pinoli. "Leveraging profiling to bridge healthcare silos for federated analyses". In: *Submitted to International Conference on Advanced Information Systems Engineering (Caise 2025 – Forum track) [A]*. 2025.

#### 4.3.7 Manuscripts

[M1] Nelly Barret. "User-oriented exploration of semi-structured datasets". PhD thesis. Institut Polytechnique de Paris, 2024.

| System name | Access | Audience | Evolution | Duration | Role | Language | LOC |
|---|---|---|---|---|---|---|---|
| 1. Better | at the project end | experts | long term | 1y | leader | Python | 6k |
| 2. Abstra | Git Inria | yeam | basic | 3y | leader | Java | 10k |
| 3. PathWays | Git Inria | team | basic | 2y | leader | Java | 4k |
| 4. ConnectionStudio | Git Inria | universe | long term | 1y | contributor | Java, JS | 25k |
| 5. Predihood | GitLab | universe | basic | 6m | leader | Python, JS | 3k |
| 6. GeoAlign | on demand | universe | basic | 6m | leader | PHP, JS | 4k |
| 7. VizLIRIS | Forge Lyon 1 | universe | basic | 3m | leader | Python, JS | 1k |

Table 4: List of the softwares that I developed. The link associated to each system corresponds to its web page; the access links are for the code (open-source). The categories are from Inria [Can+21].

#### 4.3.8 Softwares and their impact

1. For the **Better** project, I developed the script which automatically creates interoperable databases. It is now under deployment in the 7 partner hospitals. In parallel, the catalogue is developed by Noosware, an IT company. I am now working on the query engine for the catalogue.

2. **Abstra** implements or adapts existing structural summary techniques, including [GW97; GGM20; Bon+22]. I fully developed Abstra's code (algorithms and interactive E-R diagrams). Since then, Abstra has been integrated into ConnectionStudio.

3. For **Pathways**, I contributed to the development of the multi-query optimization algorithm for evaluating paths, developed the new ChatGPT-based NE extractor, and implemented the path interestingness metrics (reliability and force). Pathways has also been integrated into ConnectionStudio.

4. **ConnectionStudio** is a software suite developed by the CEDAR team and integrates other softwares developed by myself and the team: ConnectionLens [ABC+22] for the graph integration, Abstra [C1] for the E-R summaries and Pathways [J1] for the entity paths. It also proposed new dedicated features for the exploration (statistics, queries, etc). I mainly participated to the management of the underlying systems (ConnectionLens, Abstra, Pathways) and to the test of the interface developed by the team engineers. While still being in development, ConnectionStudio has already been showcased to many journalists, who have shown a clear interest.

5. I developed the 2 interfaces of **Predihood**: the one for predicting EV and the one to test AI algorithms. I also took care of the INSEE indicators harvesting as well as their manual pre-processing. The real estate experts from HomeInLove internally use our prediction interface for select few neighbourhoods of interest (those matching the client wishes).

6. **GeoAlign**[10] is a prototype for automatic alignment of geographic points of interest (sightseeing, restaurants, universities, etc.) thanks to a personalizable formula and an estimation of the matching quality. I fully implemented algorithms and the Web interface, following the MVC design pattern (Model-View-Controler).

7. **VizLIRIS**[11] is an interface to help with: (*i*) the recommendation of neighbourhoods similar to an input neighbourhood; and (*ii*) the clustering of similar neighbourhoods. I implemented both interfaces (recommendation and clustering). VizLIRIS is now used internally by HomeInLove, like Predihood.


[ABC+22] Angelos Anadiotis, Oana Balalau, Catarina Conceicao, et al. "Graph integration of structured, semistructured and unstructured data for data journalism". In: *Information Systems* 104 (2022).

[Baa+19] Mohamed Amine Baazizi, Dario Colazzo, Giorgio Ghelli, and Carlo Sartiani. "Parametric schema inference for massive JSON datasets". In: *The VLDB Journal* 28.4 (2019).

[Bon+22] Angela Bonifati, Stefania-Gabriela Dumbrava, Emile Martinez, Fatemeh Ghasemi, Malo Jaffré, Pacome Luton, and Thomas Pickles. "DiscoPG: Property Graph Schema Discovery and Exploration". In: *PVLDB* 15.12 (2022).


[10] Work not presented here for the sake of brevity.
[11] Work not presented here for the sake of brevity.

[CMP24] Leonardo Candela, Dario Mangione, and Gina Pavone. "The FAIR Assessment Conundrum: Reflections on Tools and Metrics". In: *Data Science Journal* 23.1 (2024).

[Can+21] Anne Canteaut, Miguel Angel Fernández, Luc Maranget, Sophie Perin, Mario Ricchiuto, Manuel Serrano, and Emmanuel Thomé. "Software Evaluation". PhD thesis. Inria, 2021.

[DM88] Barry A. Devlin and Paul T. Murphy. "An architecture for a business and information system". In: *IBM systems Journal* 27.1 (1988), pp. 60–80.

[GGM20] François Goasdoué, Pawel Guzewicz, and Ioana Manolescu. "RDF graph summarization for first-sight structure discovery". In: *The VLDB Journal* 29.5 (2020).

[GW97] Roy Goldman and Jennifer Widom. "DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases". In: *VLDB*. 1997.

[Hai+23] Rihan Hai, Christos Koutras, Christoph Quix, and Matthias Jarke. "Data Lakes: A Survey of Functions and Systems". In: *TKDE* (2023).

[Ong+17] Toan C Ong et al. "Dynamic-ETL: a hybrid approach for health data extraction, transformation and loading". In: *BMC medical informatics and decision making* 17 (2017), pp. 1–12.

[Pag+99] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. *The PageRank citation ranking: Bringing order to the web.* Tech. rep. Stanford InfoLab, 1999.

[Put+23] 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". In: *International Journal of Medical Informatics* 176 (2023), p. 105104.

[Wie92] Gio Wiederhold. "Mediators in the architecture of future information systems". In: *Computer* 25.3 (1992), pp. 38–49.

[Wil+16] Mark D Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific data* 3.1 (2016), pp. 1–9.