UMR IRISA

Activity Report 2025

Team DRUID

Declarative & Reliable management of Uncertain user-generated Interlinked Data

D7 – Data and Knowledge Management

Université de Rennes | rennes | INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES | cnrs | Inria | Inserm | CentraleSupélec | IMT Atlantique Bretagne-Pays de la Loire École Mines-Télécom | Université Bretagne Sud ubs

1 Team composition

Researchers and faculty

Jean-Christophe Dubois, Associate Professor, IUT Lannion - Lannion
Mireille Ducassé, Professor Emeritus, INSA Univ. Rennes, Rennes
Annie Foret, Associate Professor, ISTIC Univ. Rennes - Rennes (Emeritus 2/2025)
Mickaël Foursov, Associate Professor, ISTIC Univ. Rennes - Rennes
David Gross-Amblard, Professor, ISTIC Univ. Rennes - Rennes
Yolande Le Gall, Associate Professor, IUT Lannion, Univ. Rennes - Lannion
Zoltan Miklos, Professor, ESIR Univ. Rennes - Rennes, **head of the team**
Olivier Ridoux, Professor, ISTIC Univ. Rennes - Rennes (retirement in 12/2025)
Virginie Sans, Associate Professor, ISTIC Univ. Rennes - Rennes (until 8/2025)
Constance Thierry, Associate Professor, IUT Lannion, Univ. Rennes - Lannion
Mohamed Ez-Zaouia, Associate Professor, IUT Lannion, Univ. Rennes - Lannion
Maroua Masmoudi Kotti, Associate Professor, ENSSAT, Univ. Rennes - Lannion
Matthieu Branthôme, Associate Professor, IUT Lannion, Univ. Rennes - Lannion

LRU / ATER

Manel Chehibi, ATER ESIR (1/9/2024-31/08/2025)
Claude Petit, LRU TIARe (1/9/2025-31/8/2028)

International Associates

Archil Elizbarashvili, Ivane Javakhishvili Tbilisi State University- TSU, Georgia

PhD students

Erwan Vincent, CIFRE KEOLIS (until February 2025)
Aymen Bazouzi, CominLabs Clara project (until February 2025)
Johan Le Boursicaud, scholarship and team's own ressources
Dorra Sassi, scholarship and regional funding
Mathis Ronzon, CIFRE ALTEN
Lea Driessens, scholarship
Sami Chtitah, scholarship
Victorien Desbois (PhD at ENS Rennes) CIFRE Numa (previous name: New-LogUp)

Administrative assistant

Gunther Tessier, Inria

2 Overall objectives

2.1 Overview

Recently, there is an increased interest in data management methods. Statistical machine learning techniques, empowered by the available pay-as-you-go distributed computing power, are able to extract useful information from certain data. The international press, being specialized or not, has echoed these remarkable results as a new spring for Artificial Intelligence in a broad sense. The data is sometimes even referred to as the "gold of the 21st century". In any area of business and science, one tries to construct huge datasets to be able to profit from the benefits of the Artificial intelligence revolution.

Our team works on questions of data management techniques to efficiently store, query and organize data. We also work on artificial intelligence techniques to extract knowledge, and to gain understanding from data, especially in the presence of uncertainties. Ideally this knowledge should be actionable to be able to provide services based on them (e.g. recommendations).

Unfortunately, data management and machine learning are often seen as different tasks. Machine learning primitives are not supported elegantly for now, in the data management dogma. For example, Machine Learning operators are seen for now as external procedures outside the query language, barely accounted for by the optimizer. Moreover, the knowledge extraction tasks are hard to design without understanding the available data, thus one should consider knowledge extraction as an interactive process, where users influence the process.

The above listed observations lead us to define the following goals for the DRUID team:

- Propose new query mechanisms, in particular for graph and network oriented data and to better integrate Machine Learning methods with the database logic and engines
- Propose interactive, human-in-the-loop data analysis and knowledge extraction methods even with uncertain data

2.2 Scientific foundations

Our team gathers specialists from data management, information extraction and belief functions, various bricks that contribute to our goal. As a common ground, for data management we will naturally elaborate on classical techniques: finite model theory, complexity theory, declarative or algebraic languages, execution plans, cost models, storage and indexing strategies. The theory of belief functions (also commonly referred to as Dempster-Shafer theory) allows to take simultaneously into account both uncertainty and imprecision on the data but also on the models. This theory is one of the most popular one among the quantitative approaches because it can be seen as a generalization of both classical probabilities and possibilities theories. Belief functions are especially developed for information fusion, pattern recognition and clustering.

Analytics in databases Making sense of large amounts of data and extracting useful information is a problem in various fields, in business context as well in various scientific domains. One needs to rely on a wide range of techniques (regression, clustering, embeddings, ...). A classical data analytics workflow is 1) to extract, to model imperfection and clean a data set, 2) to learn a model and to consider imperfection and 3) to make predictions. Such workflows are now very well handled in procedural languages such as Python or Scala, at various scales (*e.g.* Big Data in Spark).

While this approach works well, it does not make use the numerous achievements in the database field: when the data set is updated, the workflow has to be re-run (dynamicity problem), data are now much more evolved than classical numerical or categorical ones, such as graphs (data type problem), and machine learning operators are not first class citizen in database query languages (closure problem). Moreover, it is often impossible to formulate the "right" knowledge extraction or machine learning tasks, as it would require the knowledge of the large and heterogeneous datasets, *a priori*.

Our specific goal is to develop data management and -possibly interactive- data analysis methods for generic, uncertain and time-varying data (*e.g.* large evolving graphs). We will rely on graph signal processing [SNF+13], spectral graph theory[Chu97], graph neural networks [WPC+20], graph databases[RWE15], [BFVY18] and graph embedding techniques. We aim at modelling and querying graphs, with (temporal) integrity constraints, where graph analytics is first used to optimize data storage and evaluation. Machine learning techniques also allow to build realistic huge benchmark data sets, that do not exist for all domains.

From a longer perspective, we would like to work on other aspects of database and machine learning integration. In particular, databases have efficient mechanisms for indexing and loading data to main memory and these could be better exploited to realize machine learning tasks. In some cases one could envisage that the machine learning tasks are realized inside the database systems and machine learning methods use database primitives[GR17]. Another potential direction is to consider a vector-space embedding of entire relational databases [Gro20] that could open entirely new ways to

[SNF+13] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains", *IEEE Signal Processing Magazine 30*, 3, May 2013, p. 83–98.

[Chu97] F. R. K. Chung, *Spectral Graph Theory*, American Mathematical Society, 1997.

[WPC+20] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P. S. Yu, "A Comprehensive Survey on Graph Neural Networks", *IEEE Transactions on Neural Networks and Learning Systems*, 2020, p. 1–21.

[RWE15] I. Robinson, J. Webber, E. Eifrem, *Graph Databases: New Opportunities for Connected Data*, edition 2nd, O'Reilly Media, Inc., 2015.

[BFVY18] A. Bonifati, G. H. L. Fletcher, H. Voigt, N. Yakovets, *Querying Graphs*, *Synthesis Lectures on Data Management*, Morgan & Claypool Publishers, 2018, `https://doi.org/10.2200/S00873ED1V01Y201808DTM051`.

[GR17] M. Grohe, M. Ritzert, "Learning first-order definable concepts over structures of small degree", *in : 32nd Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2017, Reykjavik, Iceland, June 20-23, 2017*, IEEE Computer Society, p. 1–12, 2017, `https://doi.org/10.1109/LICS.2017.8005080`.

[Gro20] M. Grohe, "word2vec, node2vec, graph2vec, X2vec: Towards a Theory of Vector Embed-

analyze data stored in such systems.

2.3 Application domains

Our natural applications are storing and querying large-scale semantic graphs for IOT (*e.g.* Maria Massri's thesis), Digital humanities (epistemology, understanding the evolution of ideas and scientific fields, (*e.g.* EPIQUE ANR project, completed), human resources management (*e.g.* François Mentec's thesis), and crowd management systems (HEADWORK ANR, completed) for "artificial artificial intelligence". Our work and results can be used to analyze the evolution of other types of networks (e.g. transportation network, e.g. Erwan Vincent and Gauthier Lyan's thesis) and in the areas of IA and education (Clara project and Dorra Sassi's ongoing PhD). Our in-database machine learning methods could have applications in various domains, including telecommunications or chemo-informatics.

3 Scientific achievements

3.1 IA and education

Participants: Aymen Bazouzi, Zoltan Miklos, Mickael Foursov, Hoel Le Capitaine (Nantes University), Matthieu Branthome, Mohamed Ez Zaouia, Constance Thierry, Dorra Sassi, David Gross-Amblard.

3.1.1 These Aymen Bazouzi

Open Educational Resources (OERs) are teaching, learning, and research materials destined for the public, allowing them to be freely used. They can be used by teachers to create new courses. Teachers can combine different OERs to achieve a specific learning objective. The CLARA project was launched to empower teachers by facilitating the creation of licensable educational resources based on existing ones. In this thesis, funded by the CLARA project, our goal is to enrich the CLARA educational corpus of OERs with useful relations between them, thus facilitating navigation for the teachers. To do so, several contributions were made in this thesis. First, the creation of a dataset construction tool that allows users to create their own tailored educational datasets from YouTube video transcripts. Second, the development of an OER-specific vector representation (embedding) that considers the specificities of OERs, which are content-centrality and the presence of semantic features. Third, the proposal of a querying method that retrieves OERs relevant to a list of keywords based on OER representations. Fourth, the design of a model that identifies possible precedence relations between pairs of OERs by using a Knowledge Graph (KG) and exploiting a Graph Neural Network (GNN). The contributions made work in harmony in order to enrich the CLARA corpus

dings of Structured Data", *CoRR abs/2003.12590*, 2020, `https://arxiv.org/abs/2003.12590`.

with educational resources, retrieve them, and identify possible relations between them, facilitating the navigation for the users. [1]

3.1.2 Codigen: Enhancing Personalized Programming Education with Visual Authoring and LLMs

Participants: Mohamed Ez zaouia.

Computer programming is an essential skill for the 21st century workforce. However, mastering programming is challenging, especially for beginners. Recently, digital tools have become popular in supporting programming education. However, creating engaging programming materials remains difficult due to the many factors that need consideration. We introduce CODIGEN, an interactive tool powered by large language models (LLMs) that integrates visual authoring and prompt engineering to support programming education with various learning strategies. A user study with 29 participants showed that CODIGEN can assist teachers in generating quality programming materials. Statistical analysis showed that CODIGEN supports learnersâ confidence and learning outcomes in programming. The perceived utility and engagement with the tool were also statistically significant. We discuss our findings to guide the design and use of LLM-powered authoring tools for educational purposes [6]

3.1.3 Adaptive Feedback for Programming Education

Participants: Matthieu Branthôme.

Learning to program is a challenging task that requires a high level of support from teachers. Adaptive feedback offers a way to support learners automatically and in a targeted manner, addressing their individual needs. The Pyrates learning platform was designed to introduce Python programming to secondary school students. Based on historical data of student-teacher interactions, we developed reinforcement learning models[7] and supervised classifiers [5] capable of selecting the most appropriate type of feedback in real time. The results [4] show that this approach significantly improves students' progression within the game and reduces teachers' workload, without, however, significantly impacting learning gains as measured by standardized tests. These findings highlight the potential of AI to support education at scale, while also identifying limitations that must be overcome to achieve finer-grained personalization of pedagogical feedback.

3.2 IA and public transportation

Participants: Erwan Vincent, Zoltan Miklos, Simon Malinowski (LINKMEDIA/MALT team).

3.2.1 Commercial speed impact factors identification for a public urban transport network

Public transport operators of large cities seek continually to improve the quality of their services. Specifically, they are interested in understanding the different factors that can impact the commercial speed of the public bus services. In this work, we propose a methodology that can identify such factors from collected service operational data using machine learning and data analysis techniques. We demonstrate our methods in the case of the city Rennes, France, and based on the data from the operator KEOLIS. [12], [15].

3.2.2 Graph Neural Network strategy for predicting the commercial speed of urban buses

It is a challenging task to accurately predict traffic conditions. However, this specific task is a necessary prerequisite for work related to urban traffic planning. For local specialists in the field, this makes it a very intriguing tool. In this study, we suggest a new Graph Neural Network (GNN) architecture that is able to predict bus commercial speeds. The data come from a large database that we previously created using substantial data collected on the STAR transit network in the Rennes metropolitan area, operated by the Keolis Rennes Company. To verify how effective our data and GNN model are for this purpose, we train a selection of models using these data, including multiple GNN versions, a more "classical" regression model , and a baseline established through expert work. Additionally, the complex temporal and spatial structure of large public transport networks may be captured by the model architecture we have suggested. This prediction task can be used to predict unknown routes (project evaluation) or current routes (network upgrade and/or supervision). (under review)

3.2.3 PhD thesis of Erwan Vincent: Machine learning for analyzing and predicting the quality of service of urban bus transportation

Improving the efficiency of public transport is a crucial issue for the organizations responsible for its operation. Today, with the advent of big data and new machine learning methods, it has become much easier to administer transport services and understand and predict performance criteria such as commercial speed. Predicting such criteria could improve the networkâs fluidity, performance, and the satisfaction of users and operators. However, current methods are fallible and sometimes fail to capture the complexity of interactions between the various factors influencing network performance. This highlights the need for operators to identify more advanced and efficient approaches. The main challenge lies in identifying these factors. Furthermore, the spatio-temporal relationships between these factors and commercial speed complicate estimation. Using Graph Neural Networks (GNNs) therefore appears to be a potential solution. Nevertheless, their applications in this task remain little explored in the literature. This thesis proposes a GNN-based approach to predicting bus commercial speed, which is one of the main performance indicators of public transport networks, using real-world datasets from a bus transport network. Several experiments demonstrate that this approach

outperforms traditional methods in terms of accuracy and robustness. The results obtained pave the way for more efficient and explanatory management of complex urban bus transport. [2]

3.3 IA and health

Participants: Maroua Masmoudi Kotti.

3.3.1 Machine Learning-Driven Annotation for Healthcare Management

In healthcare management, the manual assessment of documents to verify compliance with legal regulations is labor-intensive, error-prone, and time-consuming. These documents often consist of multiple paragraphs, and the accurate annotation of these paragraphs is essential for ensuring alignment with specific legal requirements and facilitating efficient regulatory oversight. To address these challenges, this research proposes a novel approach for automating the annotation of paragraphs within care support documents using Natural Language Processing (NLP) models and a knowledge graph (KG) to compute semantic similarity between legal articles and document paragraphs, enabling efficient annotation for compliance. Experiments on real-world healthcare documents show the method significantly improves annotation accuracy, reduces compliance verification time, and enhances similarity confidence when integrating a KG. These results highlight the potential of combining NLP and KGs to automate compliance assessment in healthcare [8].

3.4 IA and software testing

Participants: Matthis Ronzon, Annie Foret, Zoltan Miklos, Olivier Ridoux.

Functional software testing is essential to ensure that software meets user expectations. Our ambition is to enable business experts who have extensive domain knowledge but limited software engineering skills, to realize the functional software tests, through formulating test case descriptions in natural language. To meet this challenge, we propose a framework called GUIDE (Guided User-driven Interactive Description Enhancement), which leverages language models to improve functional software test descriptions written in natural language. Our framework implements an intermediate step based on a structured language (Gherkin1) that is a language widely used for software tests. We translate test descriptions written by the business expert to this language using language models. We automatically evaluate the quality of the description based on the generated Gherkin. When this quality is insufficient, GUIDE initiates an interactive and personalized assistance process, delivering targeted advice to help business experts enhance and improve their test case descriptions. We evaluated our approach through a case study based on test cases for a human resources management related software, written in French. We recorded a 26number of descriptions required per test objective to reach the desired quality level thanks to the advice generated. [9]

3.5 IA and graph databases

Participants: Johan Le Boursicaud, David Gross-Amblard, Zoltan Miklos, Calude Petit.

We work on the integration of ML techniques into graph databases. In this context, we study machine learning algorithms on heterogeneous graphs. We have developed a metapath selection technique for heterogeneous graph neural networks. Our method MSHGNN uses a two-stage automatic learning framework for node classification. We have a publication under review and two other publications in preparation.

3.6 Studies in Computer and the Environment

3.6.1 Computing within Limits from out of the box

Participants: Olivier Ridoux.

Planetary boundaries The obligations of moderation Global warming is no longer just a hypothetical threat; it is here. Its consequences are real. For example, it is estimated that around 60,000 deaths will be attributable to heat waves in Europe in 2025, and that extreme weather events will cost more than 40 billion euros in the same year. Faced with this growing threat, some are calling for restraint. By its very nature, this term evokes virtue, but as virtue can easily become hypocrisy, we must be wary. What follows shows that it is possible to quantify the impact of human activities on the planet, to quantify the maximum activity that the planet can sustain, and even to quantify the differences in impact between regions. We can deduce target activity values from this. For example, the Paris Agreement popularized such a value for the impact of global warming: carbon neutrality after 2050 (i.e., an average reduction of more than 8% per year) in order to limit global warming to less than 2 ° C. Such target values can be deduced for many other impacts. As a result, sustainability takes on a new dimension. It is no longer a question of being sustainable to a degree that each individual considers satisfactory, but rather of being sustainable to a degree that complies with these limit values. It is no longer a matter of individual virtue, but rather a massively collective, globally collective issue. An examination of these limits reveals a very large inequality between regions of the world: a contemporary inequality, but also an inequality in historical trajectories. Taking these inequalities into account is a constant stumbling block at major global meetings on the subject (e.g., the COPs). In this regard, Western countries (or countries in the global North) cultivate a view of other countries that does not help to build truly effective agreements. For example, we now know that the Paris Agreement is not delivering on its promises. It is therefore important to also hear other voices from the global South. [14]

3.7 Crowdsourcing

3.7.1 Optimizing Action Significance in Citizen Science

Participants: David Gross-Amblard, Constance Thierry, Sami Chtitah, Noe Vincent, Simon Chollet (ECOBIO).

Citizen Science, where volunteers are invited to collect information according to a scientific protocol, has seen many successes on online platforms such as Zooniverse or EyeWire. In this work, we propose to go further by asking citizens to undertake actions that have repercussions on the physical world. An action is considered effective if its impact on a real-world measurement is statistically significant. But since Citizen Science is mainly voluntary, it is necessary to take into account the effort expended for a given action. In this work, we model the problem of statistically significant citizen action and propose the algorithm that maximizes both the statistical significance of actions and the willingness to participate. Our experiments on synthetic datasets show that our algorithm offers better performances than natural baselines.

3.7.2 Sami Chtitah's Phd: Declarative and participative approaches for citizen-oriented digital twins

Participants: David Gross-Amblard, Constance Thierry, Sami Chtitah, Benoit Combemale (Diverse Team).

Digital twins offer numerous promises. By integrating data from a variety of sources (sensors, vehicles, etc.), they can be used to model complex, hybrid, and geo-referenced systems such as a factory in operation, a growing city, a river flow network, etc. Thus, digital twins are part of the family of decision-making tools, offering in particular the ability to simulate scenarios (what-if analysis). Apart from these technical aspects, a faithful digital twin is also a tool for communication and shared understanding of a territory. However, due to their youth, digital twins suffer from several limitations. On the one hand, their development still relies on low-level programming methods, where every detail must be explicitly specified. This approach makes development costly and difficult to verify and maintain. On the other hand, existing digital twins rely primarily on data from hardware devices, typically sensors. In doing so, they neglect an important aspect: the human populations living in contact with the simulated object (a city, a territory, etc.).

In this thesis, we aim to address these two limitations using techniques developed in the data management community, based on two questions.

1. Are the declarative approaches used in data management (SQL, GQL languages, etc.) and their formalization (relational algebra, Datalog, regular path queries, etc.) capable of capturing the essential elements of modeling and executing a digital twin? In other words, is it possible to fully specify a digital twin using rules rather than procedural programs? If successful, this approach would provide an elegant bridge between the design of a digital twin and the management of the data that is supposed to feed its operation.

2. More recently, participatory data management systems (crowd management systems) have been proposed to place human participants at the heart of data production. It is therefore legitimate to ask whether such approaches also make it possible to enrich the dataset used by a digital twin through human contributions (surveys), ideally triggered as the system evolves (permanent dynamic surveys). Taking human participants into account in this way would make it possible to go beyond passive sensors.

3.8 Linguistic data and NLP formalisms

3.8.1 Enhancing a multi-faceted Breton verb-centered resource to help a language learner

Participants: Annie Foret, Pêr Morvan (former student) and Erwan Hupel (univ. Rennes 2).

We recently contributed to NLP4CALL, our article builds on two recent resources for Breton, a verb-centered database and a set of sentences in the universal dependencies (UD) format. Our focus is on Breton, an endangered language in the Celtic family. We provide an analysis of the resource on verbs and show how it can be connected and transformed to a multi-faceted system intended to help a learner in a flexible way. We discuss several scenarios.

The work development and outputs are available at `https://gitlab.inria.fr/foret/termlis/-/tree/main/Breton/Verbs`. Progress are also discussed within the SmartFCA project (01/2025), to test the formal concept analysis approach on such data.

3.8.2 Categorial Dependency Grammars extended with barriers

Participants: Annie Foret and Denis Béchet (univ. Nantes).

We have presented at MCLP 2025 (International Conference on Mathematical and Computational Linguistics for Proofs, part of the EuroProofNet Symposium) a new result [3] on CDG, a Categorial Grammar Formalism, in the continuity of NLPCL 2024. We consider the family of Categorial Dependency Grammars (CDG), as computational grammars for language processing. CDG are a class of categorial grammars defining dependency structures. They can be viewed as a formal system, where types are attached to words, combining the classical categorial grammars' elimination rules with valency pairing rules that are able to define non-projective (discontinuous) dependencies.
Whereas the problem of closure under iteration is open for the original version of CDG, we define "CDG extended with barriers", an extended version of the original CDG, that solves this formal issue. We provide a rule system and we show that the extended version defines an Abstract Family of Languages (AFL)[1], while preserving advantages of the original CDG, in terms of expressivity, parsing and efficiency.

[1] closure under certain operations, see https://encyclopediaofmath.org/wiki/AFL_operations

3.9 Uncertainty management

Participants: David Gross-Amblard, Constance Thierry, Dorra Sassi.

3.9.1 Credal Knowledge Tracing for Imprecise and Uncertain MCQ

Online quizzes constitute a classical building block of digital learning environments. Their answers can be aggregated using Bayesian Knowledge Tracing (BKT), a widely used model for student knowledge evaluation. But traditional quiz systems force the user to provide precise answers, limiting the expression of her doubt, uncertainty or imprecision. To address these limitations, we introduce a novel approach called Credal Knowledge Tracing (CrKT)[10], based on the Dempster-Shafer Theory. In this setting, a doubting learner is not constrained to choose randomly an answer but can choose different ones simultaneously, and can weight each of them with a level of certainty. In this work, we provide a definition of CrKT and compare it with alternative approaches. Using a multi- attempt simulation framework, our experiments demonstrate that CrKT converges more rapidly to knowledge mastery than BKT, highlighting CrKT's potential to enhance intelligent tutoring systems.

3.10 Vehicle routing problems

Participants: Victorien Desbois, Zoltan Miklos, Francois Schwarzentruber (ENS Lyon), Ocan Sankue (DEVINE team).

3.10.1 Dynamic vehicle routing problems and resource-constrained shortest paths

We develop new algorithms for dynamic vehicle routing problems, and we analyse special variants of the vehicle routing problem, with the help of ressource-constrained shortes path framework. The work, in the context of the PhD thesis, involves the analysis of theoretical questions and also experimental research. We have several papers in preparation.

4 Software development

4.1 HEADWORK platform

Participants: David Gross-Amblard, Constance Thierry.

Crowdsourcing relies on potentially huge numbers of on-line participants to resolve data acquisition or analysis tasks. It is an exploding area that impacts various domains, ranging from scientific knowledge enrichment to market analysis support. But currently, existing crowd platforms rely mostly on low level programming paradigms, rigid data models and poor participant profiles, which yields severe limitations. The

low- level nature of existing solutions prevents the design of complex data acquisition workflows, that could be executed, composed, searched and even be proposed by participants them- selves. Taking into account the quality, uncertainty, inconsistency and representativeness of participant contributions is still an open problem. Methods for assigning a task to the correct participant according to his trust, motivation and expertise, automatically improving crowd execution time, computing optimal participant rewards, are missing. Similarly, usual crowd campaigns produce isolated and rigid data sets: A flexible and common data model for the produced knowledge about data and participants could allow participative knowledge acquisition. To overcome these challenges, Headwork[2] defines:

- Rich workflow, participant, data and knowledge models to capture various kind of crowd applications with complex data acquisition tasks and human specificities.
- Methods for deploying, verifying, optimizing, but also monitoring and adapting crowd- based workflow executions at run time.

To reach its goals, Headwork relied on two experts of large participative knowledge acquisition platforms: Cesco (Museum National d'Histoire Naturelle), Wirk (Foule-Factory), Valda (INRIA Paris), Druid (Rennes 1), Links (Inria-Lille), Sumo (Inria-Bretagne), Spirals (Inria-Lille).

Over the period of this report the platform is live in Beta version, open source (AGPL), and holding several experimental crowd campaigns. The overall project on GitLab has now 850 commits, 17 members, 7000 PHP lines.

4.2 Visual authoring tools

Participants: Mohamed Ez-zauoia.

The team is developing and maintaining Pinfig (https://pinfig.com) a visuospatial tool powered AI that incorporates end-user authoring, visualization, and prompt engineering techniques within a single user interface. Pinfig enables users to create, personalize, and organize multimodal content for insight, sensemaking, and knowledge work.

In addition, the team contributed to Mixap (https://mixap-lium.univ-lemans.fr/, `https://forge.apps.education.fr/lium/mixap`), a visual authoring tool for augmented reality. Mixap enables users to author multimodal content and activities for augmented reality using direct and visual interactions, similar to the concept of PowerPoint.

4.3 Termlis on linguistic data

Participants: Annie Foret.

[2]https://headwork.irisa.fr

The `https://gitlab.inria.fr/foret/termlis` site is an extension of the initial TermLis software at http://www.irisa.fr/LIS/softwares/TermLis/ as an application of Logical Information Systems to terminological resources.

Work under development includes :
- a section on Breton data, including data presented in [**?**] to help Breton learners ;
- data preparation for the `https://www.smartfca.org/` project.

5 Contracts and collaborations

5.1 International Initiatives

5.1.1 Multi-Dimensional Exploration of Media Collection Metadata

Participants: Omar Shahbaz Khan, Aaron Duane, Hariz Hasnan, Noe Le Blavec, Pierre Ouvrard, Johan Verdon, Laurent d'Orazio, Constance Thierry, and Jónsson Björn Þór.

This work is part of a PHC Jules Vernes project between France and Iceland. The aim of the project was to offer a multidimensional exploration of metadata from multimedia collections [11]. Two user interfaces, a virtual reality explorer and a web explorer, were designed to explore three different collections of images from diaries, general video clips, and music using various metadata attributes. Exploration is currently done through strict element searches; our future work will focus on implementing fuz

5.2 National Initiatives

5.2.1 Project Clara

Participants: Zoltan Miklos [Contact point], Annie Foret, Mickaël Foursov, David Gross-Amblard.

- Project type: Cominlabs
- Dates: 2021.12-2025.12 (The project duration was prolonged: The original date of the end of the project was 2024.12.)
- Coordinator: Patricia Serano (Nantes Université)
- Funding: 113 000 euros (IRISA)
- PI institution: CominLabs
- Other partners: LS2N (Nantes), IRISA (Rennes)

CLARA project aims to empower teachers to facilitate the creation of licensable educational resources based on existing ones. Our approach will suggest a relevant set of educational resources such that these are coherent with a course sketch and have compatible licenses. The main challenges we will face are how to enrich a network of educational resources using AI algorithms, and how to guarantee a minimal set of license-compatible educational resources relevant to a given course goal with query

relaxation techniques. We will exploit educational resources provided by the French Ministry of Education and the X5-GON project.

5.2.2 Project SmartFCA

Participants: Mireille Ducassé, Annie Foret.

- Project type: ANR
- Dates: 2021.12-2025.12 (48 month)
- Coordinator: Marianne Huchard, LIRMM
- Other partners: CIRAD/AîDA Montpellier, France (resp. Pierre Martin), INFOLOGIC R&D Lyon, France (resp. Mehdi Kaytoue), ICube Strasbourg, France (resp. Florence Le Ber), IRISA Rennes, France (resp. Sébastien Ferré), L3i La Rochelle, France (resp. Karell Bertet), LORIA Nancy, France (resp. Amedeo Napoli)

SmartFCA project relies on Formal Concept Analysis (FCA), a mathematical framework based on lattice theory and aimed at data analysis and classification. FCA, which is closely related to pattern mining in knowledge discovery (KD), can be used for data mining purposes in many application domains, e.g. life sciences and linked data. Moreover, FCA is human-centered and provides means for visualization and interaction with data and patterns.

Actually it is now possible to deal with complex data such as intervals, sequences, trajectories, trees, and graphs. Research in FCA is dynamic, but there is still room for extensions of the original formalism. Many theoretical and practical challenges remain. Actually there does not exist any consensual platform offering the necessary components for analyzing real-life data. This is precisely the objective of the SmartFCA project to develop the theory and practice of FCA and its extensions, to make the related components inter-operable, and to implement a usable and consensual platform offering the necessary services and workflows for KD. In particular, for satisfying in the best way the needs of experts in many application domains, SmartFCA will offer a " Knowledge as a Service " (KaaS) component for making domain knowledge operable and reusable on demand.

Linguistic Data are part of WP4 (resp. M. Kaytoue, P. Martin). The variety of partner use cases will serve as the experimental ground for designing the steps and the procedures supporting a common methodology. The specific use cases will be studied and a user manual will be prepared.

5.3 Bilateral industry grants

5.3.1 Cifre KEOLIS (February 2022 - January 2025)

We collaborate with the company Orange, which finances in the form of a CIFRE contract for the Ph.D. thesis of Erwan Vincent. KEOLIS is the company that operates the public transport service of the city of Rennes. The company would like to understand

the factors that influences the commercial speed of the bus services, based on collected data. In particular, they would like to understand these questions, to prepare and to provision the Trambus service that they should operate from 2030. We collaborate with the company, to develop suitable prediction and simulation methods. This work is a collaboration with Simon Malinowski (who changes his team from LINKMEDIA to MALT in 2024).

5.3.2 Cifre ALTEN (June 2024 - June 2027)

We collaborate with the company ALTEN, who finance in the form of a CIFRE contract the PhD thesis of Mathis Ronzon. Thesis co-directors: Zoltan Miklos, Annie Foret and Olivier Ridoux. Our collaboration focuses on the use of artificial intelligence techniques for enabling domain experts to be involved in programming tasks without the specific competences in coding. Our initial focus in this project is software testing.

6 Dissemination

6.1 Promoting scientific activities

6.1.1 Scientific Events Organisation

Member of the Organizing Committees

- M. Branthôme was workshop co-leader for the EIAH2025 conference in Lille (France)
- M. Ez-zaouia, workshop: *Réalité augmentée au service de la pédagogie*. Journées des libertés numériques (JDLN), BU. Université de Nantes, 2024, France.
- M. Ez-zaouia, workshop: *what authoring tools to democratize augmented reality?*. IHM, 2023, France. `https://mixap-lium.univ-lemans.fr/atelier-ihm-2023/`

6.1.2 Scientific Events Selection

Member of Conference Program Committees

- A. Foret: DCAI'25 (PC member of Special Session on Computational Linguistics, Information, Reasoning, and AI (CLIRAI))
- D. Gross-Amblard: CIKM'2024, CIKM'2025, WWW'2026
- Z. Miklos: WWW'2026, EGC'2026, SOCIT'2025, WSDM'2026, SIMbig'2025, CIKM2025, ECAI'2025, ITSC'2025, ECML PKDD'2025, IV'2026
- M. Branthôme : AIED'2025
- M. Masmoudi Kotti : KES'2025, OCTA'2025, MEDES'2025 (PC member of Workshop on AI for Semantic-Based Knowledge Ecosystems)

- M. Ez-zaouia: CHI LBW'2024
- C. Thierry: IACD-EGC'2026

Reviewer

- M. Masmoudi Kotti : OCTA'2025
- M. Ez-zaouia: IUI'2024, CHI'2024
- A. Foret: book CLIRAI 2024, volume of "Studies in Computational Intelligence" (Computational Linguistics, Information, Reasoning, and AI 2024)

6.1.3 Journal

Reviewer - Reviewing Activities

- D. Gross-Amblard: VLDB Journal,
- Z. Miklos: VLDB Journal, Neural Processing Letters, Data Mining Knowledge Discovery, Journal of King Saud University Computer and Information Sciences

6.1.4 Leadership within the Scientific Community

- David Gross-Amblard:
  - member of BDA board[3]
- David Gross-Amblard, Zoltan Miklos
  - In charge of the website of the French research in database community (`https://bdav.irisa.fr/`)
- Olivier Ridoux
  - member of CNRS GDS ÉcoInfo[4] and of "Collectif Numérique responsable" of the Poool (La French Tech Rennes-Saint-Malo)[5]

6.1.5 Scientific Expertise

- Zoltan Miklos: ANRT, ECOS NORD
- Constance Thierry : ANR ICCER

[3] `http://bdav.org`
[4] `https://ecoinfo.cnrs.fr`
[5] `https://lepoool.tech/collectif-numerique-responsable/`

6.2 Teaching, supervision

6.2.1 Teaching

- Our team is in charge of most of the database-oriented courses at University of Rennes (ISTIC department and ESIR Graduate Engineering school), with courses ranging from classical databases to business intelligence, database theory, MapReduce paradigm, or database security and privacy. We also teach several modules related to machine learning or artificial intelligence.
- IUT Lannion (Bachelor)
  - Constance Thierry, "AI tools"
  - Yolande Le Gall, "Database Conception" and "Web Technologies"
  - Jean-Christophe Dubois, "Web Technologies" and "Programming"
  - Matthieu, "Data Mining", "Object-Oriented Programming", "Web Development", "Advanced Web Programming"
- ISTIC
  - Bachelor
    - Annie Foret, "Outils formels pour l'informatique" (OFI)
    - Olivier Ridoux, "Principles of Computer Systems", "Computer and the Environment", "Language Theory", and "Web Technologies"
    - Mickaël Foursov, "Programming" and "IHM programming"
    - David Gross-Amblard, "Advanced Databases"
  - Master
    - Zoltan Miklos, "Artificial Intelligence",
    - David Gross-Amblard, "OLAP and NoSQL databases", "Database security"
    - Mickaël Foursov and David Gross-Amblard, "Information Intelligence",
    - Olivier Ridoux, "Logic Programming" and "Project engineering",
    - Mickaël Foursov, "Android Programming",
    - Maroua Masmoudi kotti, "Analysis and data mining of massive data", ENSSAT,
    - Claude Petit, "Operational Research".
- ENSSAT (Master level)
  - Maroua Masmoudi kotti, "Analysis and data mining of massive data", ENSSAT.
- ESIR (Bachelor level)
  - Zoltan Miklos "Databases"
  - "Introduction to artificial intelligence"
- ESIR (Master level)
  - Zoltan Miklos, "Data management for BigData"

  - Zoltan Miklos, "Artificial Intelligence"
- ENSAI (National School of Statistics):
  - David Gross-Amblard, "NoSQL databases", "Database security",
  - Claude Petit, "Mathematical Statistics", "Machine Learning".
- ENS Rennes:
  - David Gross-Amblard, "Préparation à l'agrégation d'Informatique" (databases)
  - Olivier Ridoux, "Epistemology and Philosophy of Sciences"
  - Olivier Ridoux, CNRS summer school, on GreenIT: Teaching Computers and the Environment [13]
- International
  - Mireille Ducassé taught a 3 weeks Logic Programming course at the Computer Science department of Ivane Javakhishvili Tbilisi StateUniversity - TSU, Georgia, with the support of Erasmus Progam 2020-1-FR01-KA107-079346.
  - Mickaël Foursov taught a course on Multimedia Mobile at the Félix Houphouët-Boigny University in Ivory Coast (2nd year Miage program).
  - Constance Thierry taught a course on Belief functions theory at the British University in Egypt.

6.2.2 Administration

- Mickaël Foursov is director of special program in Business Informatics (MIAGE)
- Mickaël Foursov is in charge of the double diploma (Bachelor's et Master's in MIAGE) between University of Rennes and the Félix Houphouët-Boigny University in Ivory Coast
- Mickaël Foursov is in charge of International Relations at ISTIC
- Yolande La Gall is member of the Institute Council at IUT Lannion
- Yolande La Gall is manager of work-study program support (until september 2024)
- Yolande Le Gall is in charge of company contacts for the second and third work-study years of the BUT in computer science (until september 2024)
- Jean-Christophe Dubois is manager of the second work-study years of BUT in computer science (until september 2024)
- Jean-Christophe Dubois is in charge of the timetable for the first year of the computer science BUT at IUT Lannion
- Zoltan Miklos is the responsable of the option Information Systems, ESIR
- Constance Thierry is Member of the Board of Directors of Société Informatique de France
- Constance Thierry is responsable of the third year students of the computer science BUT at IUT Lannion
- Constance Thierry is a member of the lannion iut research committee
- Mattheu Branthôme is in charge of the timetable for the second and third year of the computer science BUT at IUT Lannion
- David Gross-Amblard is responsible of M1 Miage, and also in charge of M1 Miage and M2 Miage Intership (classic track), Correspondant for students with disabilities, and member of the ISTIC council.

6.2.3 Supervision

- (thesis completed: Feb 2025) Aymen Bazouzi, Recommendations of educational resources, through graph representation learning. Supervised by Zoltan Miklos (director), Mickaël Foursov and Hoel Le Capitaine (Nantes University, laboratoire LS2N), project Clara (Cominlabs)
- (thesis completed: Dec 2025) Erwan Vincent, Simulation and prediction of public transport services, especially, high-quality bus services, from February 2022, Zoltan Miklos (director) and Simon Malinowski (Linkmedia team). (Cifre KEOLIS Rennes),
- Johan Le Boursicaud: Machine learning in graph databases, David Gross-Amblard (co-director/supervisor), Zoltan Miklos, (co-director/supervisor).
- Dorra Sassi: Modeling Human Learning through Imperfect Responses and Elicitation, (director) David Gross-Amblard, (director, co-supervisor) Constance Thierry (co-supervisor)
- Mathis Ronzon: Towards a no-code approach to test-case generation, (directors) Zoltan Miklos and Thierry Roger (ALTEN), (co-directors) Annie Foret and Olivier Ridoux. (CIFRE, ALTEN).
- Lea Driessens: AI-Powered Visual Authoring Tools for Multimodal and Interactive Agent-Based Activities, (director) Zoltan Miklos
- Mohammed Oussama Seddini: Studying and Designing Visual Authoring Tools Powered by Generative AI to Support Teachers and Learners in Creating Augmented Reality Activities (director) Iza Marfisi, (co-supervisor) Mohamed Ezzaouia. These a l'Univ Le Mans
- Victorien Desbois, Vehicle routing problems, Zoltan Miklos (co-director, from September 2024), François Schwarzentrouber (ENS Lyon, co-director), thesis at ENS Rennes cifre NUMA (ex. NewLogUp)
- Moncef Naji, An LLM-based approach for medical e-consent, Hajer Zghal (director), Maroua Masmoudi Kotti (co-supervisor), thesis at Cergy University (Phd defense in February 2026)
- Sami Chtitah, Declarative and participative approach for citizen-oriented digital twins, David Gross-Amblard (co-director and advisor with Benoit Combemale, Diverse team), Constance Thierry (advisor), from Sep. 2025.

6.2.4 Juries

- Z. Miklos: PhD examinateur: Ginwa Fakih, Nantes University
- Z. Miklos: PhD examinateur/rapporteur: Eric Behar, Institut Polytechnique Telecom Paris Sud
- Z. Miklos: PhD examinateur/rapporteur: Alexander Selzer, Technische Universitaet Wien (TU Wien), Vienna, Austria
- Z. Miklos: Manoe Kieffer rapporteur (2025) / examinateur (2026): Nantes University
- Z. Miklos: Moncef Naji rapporteur (2025) / examinateur (2026): Cergy University

- Y. Le Gall: PhD defense, (examinatrice), September 2024, Institut Montpellierain Alexander Grothendieck (IMAG), Montpellier

- Z. Miklos: CSID Simon Lévêque, Univ Rennes, ENSSAT Lannion (FOTON)

- Z. Miklos: CSID Clement Bardet, Univ Rennes

- Z. Miklos: CSID Erik Yeghan, University of Strasbourg

- Z. Miklos: CSID Tooryanand Seetohul, Univ Rennes

- Z. Miklos: Selection committee (COS), Chair Prof Junior (CPJ), co-president, Univ Claude Bernard, Lyon

- A. Foret: PhD examinateur: Maxime Guillaume, November 2025, Nancy University.

- D.Gross-Amblard: CSID Hritika Kathuria, Ecole Polytechnique.

6.3 Popularization

- Z. Miklos "Appréhender et enseigner les enjeux de l'intelligence artifcielle en education", Public: teachers of primary and secondary schools (private), IFSEC Bretagne, 5/03/2025.

- Z. Miklos Congrès Départemental de FCPE (Fédération des Conseils de Parents d'Elèves des Écoles Publiques), Les enjeux de l'intelligence artificielle (IA): quelle vision pour la scolarité et l'avenir de nos enfants? Rennes, Lycée Brequigny, 21/06/2025

- C. Thierry and A. Foret animated several sessions on IT topics and jobs at Lycée de Guingamp in March 2025 dedicated to secondary school girls

- C. Thierry participation in the organisation of the "Journée Fille Math et informatique"

- C. Thierry participating in the organisation of the ada lovelace challenge to raise awareness of IT among secondary school girls

- C. Thierry took part in the panel discussion "AI and Environment" during the Research and Innovation event organized by the Côtes-d'Armor Departmental Council. Nov 2025.

- M. Branthôme took part in the panel discussion "AI and Education" during the Research and Innovation event organized by the Côtes-d'Armor Departmental Council. Nov 2025.

7 Bibliography

Doctoral dissertations and "Habilitation" theses

[1] A. Bazouzi, *Combining educational ressources using graph representation learning*, Theses, Université de Rennes, February 2025, https://theses.hal.science/tel-05242520.

[2] E. Vincent, *Apprentissage automatique pour l'analyse et la prédiction de la qualité de service des transports de bus urbains*, Theses, Université de Rennes, December 2025, https://theses.hal.science/tel-05456485.

Publications in Conferences and Workshops

[3] D. Béchet, A. Foret, "Categorial Dependency Grammars Extended with Typed Barriers", *in : International Conference on Mathematical and Computational Linguistics for Proofs (MCLP 2025)*, Roussanka Loukanova and Axel Ljungström, Orsay, Saclay, France, September 2025, https://hal.science/hal-05467165.

[4] M. Branthôme, S. Lallé, "Impact of Adaptive Feedback on Learning Programming with a Serious Game in High Schools' Classes", *in : 33rd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '25)*, ACM (editor), New York, NY, United States, June 2025, https://hal.science/hal-05024217.

[5] M. Branthôme, S. Lallé, "Learning from Teachers: AI-Driven Feedback for a High School Python Serious Game", *in : The 12th 2025 ACM Learning @ Scale Conference*, Palermo, Italy, July 2025, https://hal.science/hal-05049369.

[6] M. Ez-Zaouia, Y. Boumarafi, "Codigen: Enhancing Personalized Programming Education with Visual Authoring and LLMs", *in : Proceedings of the 2025 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, Raleigh, United States, 2025, https://hal.science/hal-05185391.

[7] B. Kirouchenassamy, A. Yessad, S. Jolivet, M. Branthôme, S. Lallé, V. Luengo, "Learning Feedback Policy from Historical Data: An Offline Approach within Pyrates", *in : Lecture Notes in Computer Science*, *Lecture Notes in Computer Science*, *15877*, Springer Nature Switzerland, p. 165–178, Palerme, Italy, July 2025, https://hal.science/hal-05158735.

[8] M. Naji, M. Masmoudi, H. Baraka, H. B. Zghal, "Machine Learning-Driven Annotation for Healthcare Management", *in : Advanced Information Networking and Applications*, *Lecture Notes on Data Engineering and Communications Technologies*, *247*, Springer Nature Switzerland, p. 354–364, Barcelona, Spain, April 2025, https://hal.science/hal-05290036.

[9] M. Ronzon, T. Roger, Z. Miklos, A. Foret, "GUIDE: A Framework for Improving Functional Software Test Descriptions with Language Models", *in : Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, Mourad Abbas and Tariq Yousef and Lukas Galke, Association for Computational Linguistics, p. 229–241, Odense, Denmark, August 2025, https://hal.science/hal-05462671.

[10] D. Sassi, D. Gross-Amblard, C. Thierry, "Credal Knowledge Tracing for Imprecise and Uncertain MCQ", *in : Symposium on Intelligent Data Analysis (IDA 2025)*, Konstanz (Allemagne), Germany, May 2025, https://hal.science/hal-05106052.

[11] O. Shahbaz Khan, A. Duane, H. Hasnan, N. Le Blavec, P. Ouvrard, J. Verdon, L. d'Orazio, C. Thierry, B. Þ. Jónsson, "Multi-Dimensional Exploration of Media Collection Metadata", *in : MMM 2025 - International Conference on Multimedia Modeling*, Nara, Japan, January 2025, https://hal.science/hal-05136933.

[12] E. Vincent, Z. Miklos, S. Malinowski, "Commercial speed impact factors identification for a public urban bus transport network", *in : IEEE International Conference on Intelligent Transportation Systems (ITSC) 2025*, Gold Coast, Australia, September 2025, https://hal.science/hal-05461919.

Miscellaneous

[13] O. Ridoux, "Teaching Computers and the Environment", Lecture, October 2024, https://hal.science/hal-04738732.

[14] O. Ridoux, "Les limites planétaires", Lecture, October 2025, https://hal.science/hal-05339290.

[15] E. Vincent, Z. Miklos, S. Malinowski, "Commercial speed impact factors identification for a public urban bus transport network", working paper or preprint, May 2025, https://hal.science/hal-05461649.