UMR IRISA

Activity Report 2024

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
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
Virginie Sans, Associate Professor, ISTIC Univ. Rennes - Rennes,
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 (from September 2024)
Matthieu Branthôme, Associate Professor, IUT Lannion, Univ. Rennes - Lannion (from September 2024)

LRU / ATER

Manel Chehibi, ATER ESIR (1/9/2024-31/08/2025)

International Associates

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

PhD students Arthur Hoarau, Departemental and regional funding (until August 2024)
Erwan Vincent, CIFRE KEOLIS
Aymen Bazouzi, CominLabs Clara project
Johan Le Boursicaud, scholarship and team's own ressources
Dorra Sassi, scholarship and regional funding
Mathis Ronzon, CIFRE ALTEN
Lea Driessens, scholarship
Victorien Desbois (PhD at ENS Rennes) CIFRE NewLogUp (from September 2024)

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 organise 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, costs 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.

On 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]. Other potential direction is to consider a vector-space embedding of entire relational databases [Gro20] that could open entire 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). Our in-database mechine 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).

3.1.1 Precedability Prediction Between Open Educational Resources

The abundance of Educational Resources (ERs) has allowed people to have access to a vast amount of knowledge. However, it can be difficult, for both educators and learners, to navigate through these resources. One way to facilitate navigation is to identify useful relations between these resources. This can improve the teaching and learning experiences by allowing the users to go from one resource to another based on the identified relations, such as precedence. In this work, we introduce the notion of precedability between educational resources; whether a resource A can precede another resource B. Then, we propose a two-step method to identify precedability relations between educational resources. Our method structures the educational resources in an enriched Knowledge Graph (KG). Then, it uses a Graph Neural Network (GNN) model to predict precedability relations. Our method performed better than multiple baselines on different benchmarks. [6]

3.1.2 ConstrucTED: Constructing Tailored Educational Datasets from Online Courses

Researchers are actively involved in developing various systems to support education, including recommender systems. However, to create and evaluate such systems, they require rich and versatile datasets about educational content. At times, the available

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

data proves insufficient, leading researchers to invest significant time in crafting personalized web scrapers for additional data retrieval. The generated datasets are often task-specific and may be time-consuming to adapt to future tasks. Additionally, researchers may encounter licensing issues when using courses from different providers. Furthermore, researchers prefer evaluating their methods through diverse tests, involving datasets with varying characteristics. However, this diversity is not commonly found in most available datasets, at least not explicitly so. To address these challenges, we introduce ConstrucTED, a tool built on top of Google APIs, enabling the efficient creation of custom educational datasets from YouTube pl aylists. This allows datasets to be tailored to specific characteristics such as a predetermined number of courses, coverage of specific topics, or courses from a particular university. ConstrucTED creates datasets from video course transcripts, providing a ready-to-use solution that significantly shortens the time required to create such datasets. The resulting datasets are versatile and suitable for tasks like classification and learning path creation.[7]

3.2 IA and transport public

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 to understand 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 methodology in the case of the city Rennes, France, and based on the data from the operator KEOLIS. (under review)

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

It is a challenging work 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 predicting 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)

3.3 Named entity discovery in a large noisy heterogeneous diachronic corpus

Participants: Oivier Ridoux, Clément Morand (LISN, Paris-Saclay).

Size, noise (eg. OCR errors), heterogeneity and diachronicity are well-known difficulties in the discovery of named entities in historical corpora. The corpus of the science popularization magazine LA NATURE gathers all these difficulties. However, this corpus is of great values for digital humanities studies that cover the 1870-1970 period, which for several authors is the period of the creation of the modern world, and of its development. As an example of named entities that can be found in scientific literature, we have studied the case of binominal named entities (aka, Latin names for living entities, or Linnean names, or taxons). This type of named entities suffers from all difficulties we have mentioned above: noise (ocerization of a badly scaned text), heterogenicity (natural science article are a minority in the corpus, in fact any matter is a minority), and diachronicity (linnean taxonomy changes everytime, but writing practices change a lot too). We have shown that a combination of a pattern based approach (ie. the authorized declensions of Latin names) and of a statistic based approach (ie. these names are altogether rare) computes efficiently good quality results (ie. high precision and recall, and small computation time), even when compared with modern IA-based approaches [5].

This combined approach have since then been applied with success to other types of named entities, like chemical entities, or astronomical entities. Our current work is to design a digital humanities framework that would permit to navigate in a corpus via association of ideas (eg. "palmier" → "*Phoenix dactylifera*" → "dattier" → Tunisie → ...).

3.4 Studies in Computer and the Environment

3.4.1 Computing within Limits from out of the box

Participants: Olivier Ridoux.

Computing within Limits is a difficult concept for a domain whose concepts are often based on unlimited growth (eg. Moore-s law and integrated circuit development, but also much less trivially asymptotical complexity). We think that it could be helpful to step out the box, and examine very different domains that already have an history of coping with limits. In these other domains coping with limits has first required to define them, then to imagine novel ways to practice ones trade, then to observe the result, and repeat again. This research belongs as well to Computer Science, and to Social Science as Policy Making.

We have studied the case of the Common Fisheries Policy of the European Union, and tried to evaluate which elements of this policy could be relevant to the domain

of computing [12]. Three of these conclusions are that one must not be afraid to oversimplify the problem, one must not wait to have an excellent solution to apply it (*fail early ⇒ fix early*), and most importantly, one must install measument tools right from the beginning. A fourth conclusion comes from the observation that Common Fisheries Policy started more than 45 years ago, and that it already had several versions. The fourth conclusion is that all this takes time!

3.4.2 Popularization tools for Computer and the Environment

Participants: Olivier Ridoux, Hugo Depuydt (ENS Rennes), and various students.

Analysing the environmental impact of computing systems requires a knowledge of these systems that goes beyond the knowledge of the average citizen. However, laws are edicted (eg. WEEE at the UE level, or REEN and AGEC at the French level) whose rationals come from this deep understanding that is wanting to the average citizen. There are certainly many ways for filling this gap, but we believe that one such way is to build material demonstrators of the phenomenons the average citizen if often not aware of. Such demonstrators consist in small simplish models of otherwise vast and complex realities. They must be interactive in order to permit *what-if* experiments. And they must be very easy to observe. We have developped one such demonstrator, DemoJ (J like joule), to illustrate energy impacts of complex computing systems that in reality are planetwise. It consists in three boxes that correspond to the three tiers that are commonly used for analysing complex systems (namely, Terminal, Network, and Cloud). Each box displays its energy consumption and its temperature as a color gauge. Each box can run scenarios according to its role (eg. download a large/small file, or start a session with a chatbot). The gauges permit to see very concretly the effect of decisions taken at one layer on the other layers. This demonstrator has been presented at the Poool (La French Tech Rennes-Saint-Malo)[1], as well as during the 2024 CNRS Autumn School on ICT and the Environment[2] [16], and to the CNRS GDS ÉcoInfo working group[3].

We are currently developping another demonstrator for exposing the so-called static and dynamic energy consumptions of CPUs. While DemoJ is exposing the very large, this new demonstrator will expose the very small. Note that these demonstrators simplify the observation of complex phenomenons, but are themselves complex systems. In particular, their calibration to make phenomenons visible at a human scale requires a lot of attention.

3.5 Crowdsourcing

3.5.1 Automated Hierarchical Conflict Reduction for Environmental Crowdsourced Annotation Tasks using Belief Functions

Participants: Constance Thierry, David Gross-Amblard, Yolande Le Gall, and

[1] `https://lepoool.tech/collectif-numerique-responsable/`
[2] `https://eco-ict-2024.sciencesconf.org/`
[3] `https://ecoinfo.cnrs.fr`

Jean-Christophe Dubois.

A typical crowdsourcing task is concept labeling, where participants annotate e.g. images using a list of predefined concepts. Recent popular campaigns for environmental bird monitoring even use hierarchies of concepts (taxonomies of species) to obtain the most precise labeling of bird images. But in most applications, volunteer opinions are isolated from each other, and decision is taken upon majority voting. In this work [14] we propose a new iterative labeling process where participants express their opinions together, on ascending levels of the taxonomy. Level changes are performed to minimize opinion conflict, according to the belief function theory. This complex task is orchestrated by a finite-state automaton driven by conflict measures. This model will soon be integrated into the Headwork platform to be tested during real crowdsourcing campaigns and to evaluate its performance. We particularly plan to conduct a study to determine the most relevant conflict calculation in this context. Our future work also focuses on adapting the level of taxonomy to be adopted according to the profile and expertise of the contributor: more specific for experts and more general for non-experts. [14]

3.5.2 Practical aspects of data-oriented crowdsourcing platforms for complex tasks and participants

Participants: David Gross-Amblard, Constance Thierry.

We pursue the study of the practical aspects of crowdsourcing platforms, through our proof-of-concept Headwork. Besides classical crowdsourcing features, Headwork eases the development of crowdsourcing campaigns through a full relational abstraction of relevant concepts (participants, skills, tasks, current answers, decision procedures, GUI, etc.). It allows in particular the orchestration of complex dynamic tasks using so-called tuple artifacts (i.e. finite-state automata which transition guards and actions are SQL-defined, on an evolving database). Our demo has been presented in selective venues [9], and is the recipient of the best demo at the national conference BDA 2023.

3.6 Linguistic data and NLP formalisms

3.6.1 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).

This contribution [15] to NLP4CALL 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, to test the formal concept analysis approach on such data.

3.6.2 Categorial Dependency Grammars extended with barriers

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

We have presented a new result [8] on CDG, a Categorial Grammar Formalism. 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)[4], while preserving advantages of the original CDG, in terms of expressivity, parsing and efficiency.

3.6.3 Incremental learning of iterated dependencies

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

In this journal paper [3] we study some learnability problems in the family of Categorial Dependency Grammars (CDG), a class of categorial grammars defining dependency structures. CDG is a formal system, where types are attached to words, combining the classical categorial grammars' elimination rules with valency pairing rules defining non-projective (discontinuous) dependencies; very importantly, the elimination rules are naturally extended to the so called "iterated dependencies" expressed by a specific type constructor and related elimination rules. This paper first reviews key points on negative results: even the rigid (one type per word) CDG cannot be learned neither from function/argument structures, nor even from dependency structures themselves. Such negative results prove the impossibility to define a learning algorithm for these grammar classes. Nevertheless, we show that the CDG satisfying reasonable and linguistically valid conditions on the iterated dependencies are incrementally learnable in the limit from dependency structures. We provide algorithms and also discuss these aspects for recent variants of the formalism that allow the inference of CDG from linguistic treebanks.

3.7 Dealing with imperfect data in Machine Learning

In response to the societal challenge of digital sobriety, our aim is to work with a reduced number of labelled observations, thereby also reducing labelling costs. Our work has enabled us to study the impact of adding information during the labelling phase on

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

classification performance. The studies and experiments carried out on the processing of such enriched labels have shown the relevance of this approach.

3.7.1 Mean distance for processing rich labels

Participants: Arthur Hoarau, Constance Thierry, Yolande Le Gall, and Jean-Christophe Dubois.

The prevalence of imperfections in data, characterized by uncertainty and imprecision, prompts the need for effective modeling techniques. In this work [11], we handle the problem of calculating the mean distance between elements of the same class, especially when class membership is uncertain and imprecise. Leveraging belief functions and a notion of similarity between elements, we propose a solution and validate its efficacy through experimental evaluations. Two proofs and experiments have been conducted to theoretically support these properties. It has been observed during our experiments that this solution can be useful under moderate noise, or imprecision, but it could be more appropriate to use unsupervised methods, such as K-means, when noise is significant. This approach solves many problems, and we have used it in particular in active learning.

3.7.2 Active learning from imperfect data

Participants: Arthur Hoarau, Yolande Le Gall, Jean-Christophe Dubois and Zoltan Miklos.

Recent studies in active learning, and more precisely in uncertainty sampling, have focused on the decomposition of model uncertainty into reducible and irreducible uncertainties. The aim of our work [4][10] is to simplify the computational phase while removing the dependence on observations. Crucially, the inherent uncertainty in the labels is considered, i.e. the uncertainty of the oracles. Two strategies are proposed, sampling by Klir uncertainty, which tackles the exploration – exploitation dilemma, and sampling by evidential epistemic uncertainty, which extends the concept of reducible uncertainty within the evidential framework, both using the theory of belief functions. Experimental results in active learning demonstrate that our method can outperform uncertainty sampling.

3.7.3 Quantifying and Leveraging Uncertain and Imprecise Answers in Multiple Choice Questionnaires for Crowdsourcing

Participants: Constance Thierry, Gery Casiez, Jean-Christophe Dubois, Yolande Le Gall, Sylvain Malacria, Thomas Pietrzak, Arnaud Martin, Pierrick Uro.

Questionnaires are efficient for collecting numerous user feedback. However, the reliability of results is a major issue, even with honest participants. Indeed, they face situations of doubt, and usually do not have the option to express their hesitations. We describe in our work [13] a user study in which we provide participants with the

possibility to give 1) a certitude rate 2) imprecise answers, 3) both a certitude rate and imprecise answers. Firstly, we observe that contributors express their hesitations consistently: there is a correlation between the task difficulty on the one hand, and the uncertainty and imprecision of the answer, on the other hand. Secondly, our results demonstrate the effectiveness of the decision-making process by using this additional information with the belief functions theory. Indeed, this process helps to reduce the error rate and fewer participants are required to reach a satisfactory correct answers rate.

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[5] 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.

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

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.

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 [15] to help Breton learners ;
- data preparation for the `https://www.smartfca.org/` project.

5 Contracts and collaborations

5.1 International Initiatives

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) -signature pending

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

- Z. Miklos: The 35th Australasian Database Conference, ADC'2024, Proceedings co-chair `https://adc-conference.github.io/2024/organisation`
- M. Branthôme and M. Ez-zaouia, workshop : *Practical statistics for Technology Enanced Learning*. RJC-EIAH, 2024, 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

- D. Gross-Amblard: CIKM'2024

- Z. Miklos: ECAI'2024, CIKM'2024, WSDM'2024, ECMLPKDD'2024 , EGC'2024, SOCIT'2024, ASONAM'2024
- M. Branthôme : AIED'2024
- M. Masmoudi Kotti : KES'2024
- M. Ez-zaouia: CHI LBW'2024
- C. Thierry: DECOR'2024 , IACD-EGC'2025

Reviewer

- M. Ez-zaouia: IUI'2024, CHI'2024

6.1.3 Journal

Reviewer - Reviewing Activities

- D. Gross-Amblard: Neural Processing Letters, IEEE Transactions on Knowledge and Data Engineering (TKDE)
- Z. Miklos: Transactions on Large-scale Knowledge-based Systems, Data and Knowledge Engineering, Engineering Applications of Artificial Intelligence, Ain Shams Engineering Journal
- M. Branthôme : STICEF (Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation)

6.1.4 Leadership within the Scientific Community

- David Gross-Amblard:
  - member of BDA board[6]
- David Gross-Amblard, Zoltan Miklos
  - In charge of the website of the French research in database community (`https://bdav.irisa.fr/`)
- Annie Foret
  - local correspondant of GDR TAL[7]
- Olivier Ridoux
  - member of CNRS GDS ÉcoInfo[8] and of "Collectif Numérique responsable" of the Poool (La French Tech Rennes-Saint-Malo)[9]

[6] `http://bdav.org`
[7] `https://gdr-tal.ls2n.fr/`
[8] `https://ecoinfo.cnrs.fr`
[9] `https://lepoool.tech/collectif-numerique-responsable/`

6.1.5 Scientific Expertise

- Zoltan Miklos: ANRT

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.
- ISTIC
  - Bachelor
    - Annie Foret, "Outils formels pour l'informatique" and "Bases de données"
    - 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"
    - Constance Thierry, "AI tools"
    - Yolande Le Gall, "Database Conception" and "Web Technologies"
    - Jean-Christophe Dubois, "Web Technologies" and "Programming"
  - 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",
    - Annie Foret, projects "Information Systems for Educational Resources".
    - Maroua Masmoudi kotti, "Analysis and data mining of massive data", ENSSAT.
- 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
- 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 [16]
- 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 in charge of the timetable for the third year of the computer science BUT at IUT Lannion
- Constance Thierry is a member of the lannion iut research committee
- Annie Foret and Olivier Ridoux are co-heads (until september 2024) of the second year of the computer science degree at ISTIC, Univ. Rennes.
- Mattheu Branthôme is in charge of the timetable for the second year of the computer science BUT at IUT Lannion
- David Gross-Amblard is responsible of M2 Miage Interships (classic track), Correspondant for students with disabilities, and member of the ISTIC council.

6.2.3 Supervision

- (thesis completed: June 2024) Arthur Hoarau, Active learning of imprecise and uncertain data, Jean-Christophe Dubois, Yolande Le Gall, Arnaud Martin, Zoltan Miklos (from September 2023)
- 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)
- Erwan Vincent, Simulation and prediction of public transport services, especially, high-quality bus services, from February 2022, Zoltan Miklos (co-director), Guillaume Gravier (co-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, (co-supervisor) Mohamed Ezzaouia.
- 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

6.2.4 Juries

- Z. Miklos: PhD defense, (rapporteur), November 2024, Sorbonne University (LIP6), Paris
- Y. Le Gall: PhD defense, (examinatrice), September 2024, Institut Montpellierain Alexander Grothendieck (IMAG), Montpellier
- Z. Miklos: CSID Duc Thinh Ngo, IMT Atlantique Nantes
- Z. Miklos: CSID Simon Lévêque, ENSSAT Lannion (FOTON)
- A. Foret: selection committee for the recruitment of a teacher-researcher (COS), April-May 2024, ENSSAT, Lannion
- A. Foret: selection committee for the recruitment of a teacher-researcher (COS), April-May 2024, IUT, Lannion

- D. Gross-Amblard: Phd Defense (reviewer), Tiange ZHU, Cnam Paris (October 22, 2024)
- D. Gross-Amblard: Phd Defense (reviewer), Julien Mirval, UVSQ Paris-Saclay (March 11, 2024)

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 Bretage, 13-14/03/2024.
- Z. Miklos Formation "Découvrir l'intelligence artificielle", Maison pour la science (MPLS) Bretagne, Public: teachers of primary and secondary schools (public), 15/04/2024
- Z. Miklos Journée régionale des colleges: "la Main à la Pâte", Maison pour la science (MPLS), 24/01/2024, Présentation de vulagrisation sur l'IA.
- A. Foret co-animated a session on "Comment diversifier les activités sur les bases de données ?" within the "journées du MENJ, programme national de formation, les rendez-vous de l'informatique" (resp. inspecteur général Jean-Marie Chesneaux), Rennes, april 2024.
- C. Thierry participation in the Fête de la Science 2024
- C. Thierry participation in the organisation of the SIF mediation school
- C. Thierry participating in the organisation of the ada lovelace challenge to raise awareness of IT among secondary school girls

7 Bibliography

Books and Monographs

[1] R. Jaziri, A. Martin, A. Cornuéjols, E. Cuvelier, F. Guillet, *Advances in Knowledge Discovery and Management*, *Studies in Computational Intelligence*, *1110*, Springer Nature Switzerland, 2024, https://inria.hal.science/hal-04455676.

Doctoral dissertations and "Habilitation" theses

[2] A. Hoarau, *Apprentissage Actif de Données Incertaines et Imprécises*, Theses, Université de Rennes, June 2024, https://hal.science/tel-04723819.

Articles in referred journals and book chapters

[3] D. Béchet, A. Foret, "Incremental learning of iterated dependencies", *Mach. Learn. 113*, 5, 2024, p. 2841–2875, https://doi.org/10.1007/s10994-021-05947-2.

[4] A. Hoarau, V. Lemaire, Y. Le Gall, J.-C. Dubois, A. Martin, "Evidential uncertainty sampling for active learning", *Machine Learning*, 2024, https://hal.science/hal-04623676.

[5] C. Morand, O. Ridoux, "CRI : A Competent Reader Imitator for detecting binomial names in an historical corpus", *Lingvisticae investigationes : International Journal of Linguistics and Language 47*, 1, October 2024, p. 30–67, https://hal.science/hal-04764787.

Publications in Conferences and Workshops

[6] A. Bazouzi, H. Le Capitaine, Z. Miklos, M. Foursov, "Precedability Prediction Between Open Educational Resources", *in : International Conference on Information Technology for Social Good (GoodIT '24)*, Bremen, Germany, September 2024, https://hal.science/hal-04654407.

[7] A. Bazouzi, Z. Miklos, M. Foursov, H. Le Capitaine, "ConstrucTED : Constructing Tailored Educational Datasets From Online Courses", *in : 16th International Conference on Computer Supported Education (CSEDU)*, SCITEPRESS - Science and Technology Publications, p. 645–652, Angers, France, May 2024, https://hal.science/hal-04510615.

[8] D. Béchet, A. Foret, "Categorial Dependency Grammars extended with barriers (CDGb) yield an Abstract Family of Languages (AFL)", *in : Computer Science & Information Technology (CS & IT)*, D. C. Wyld, D. Nagamalai (editors), *14*, 17, David C. Wyld and Dhinaharan Nagamalai, p. 53–66, Copenhagen, Denmark, September 2024, https://hal.science/hal-04570932.

[9] D. Gross-Amblard, M. Tommasi, I. Rakotoniaina, C. Thierry, R. Singh, L. Jacoboni, "HEADWORK: a Data-centric Crowdsourcing Platform for Complex Tasks and Participants", *in : Proceedings 27th International Conference on Extending Database Technology, EDBT 2024, Paestum, Italy, March 25 - March 28*, L. Tanca, Q. Luo, G. Polese, L. Caruccio, X. Oriol, D. Firmani (editors), OpenProceedings.org, p. 778–781, 2024, https://doi.org/10.48786/edbt.2024.68.

[10] A. Hoarau, V. Lemaire, A. Martin, J.-C. Dubois, Y. Le Gall, "Méthode crédibiliste pour l'extraction d'incertitudes sans dépendance aux observations", *in : EGC 2024 - 24ème conférence francophone sur l'Extraction et la Gestion des Connaissances*, Dijon, France, January 2024, https://hal.science/hal-04623877.

[11] A. Hoarau, C. Thierry, J.-C. Dubois, Y. Le Gall, "A mean distance between elements of same class for rich labels", *in : Belief Functions : Theory and Applications, 2024*, Belfast, United Kingdom, September 2024, https://hal.science/hal-04623863.

[12] O. Ridoux, "ICT within limits compared with fisheries within limits", *in : Undone Computer Science*, Université de Nantes, Inria, and LORIA Université de Lorraine, Nantes (France), France, February 2024, https://hal.science/hal-04867654.

[13] C. Thierry, G. Casiez, J.-C. Dubois, Y. Le-Gall, S. Malacria, T. Pietrzak, A. Martin, P. Uro, "Quantifying and Leveraging Uncertain and Imprecise Answers in Multiple Choice Questionnaires for Crowdsourcing: Quantifier et exploiter des réponses incertaines et imprécises dans des questionnaires à choix multiples pour le crowdsourcing", *in : Proceedings of the 35th Conference on l'Interaction Humain-Machine*, p. 1–10, 2024.

[14] C. Thierry, D. Gross-Amblard, Y. Le Gall, J.-C. Dubois, "Automated Hierarchical Conflict Reduction for Crowdsourced Annotation Tasks using Belief Functions", *in : 8th International Conference on Belief Functions (BELIEF)*, Belfast, United Kingdom, September 2024, https://hal.science/hal-04746967.

Miscellaneous

[15] A. Foret, E. Hupel, P. Morvan, "Enhancing a multi-faceted Breton verb-centered resource to help a language learner", NLP4CALL2024 : Natural Language Processing for Computer-assisted Language Learning, October 2024, Poster, https://hal.science/hal-04869508.

[16] O. Ridoux, "Teaching Computers and the Environment", Lecture to École d'automne du CNRS : Eco-ICT 2024, France, October 2024, https://hal.science/hal-04738732.