UMR IRISA

Activity Report 2023

Team DRUID

Rennes - Lannion

D7 – Data and knowledge management

Université de Rennes | ENS 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
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
Arnaud Martin, Professor, IUT Lannion, Univ. Rennes - Lannion, passed away on August 21, 2023
Zoltan Miklos, Associate 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

LRU / ATER

Mathieu Chambe ISTIC, Univ Rennes, Rennes, (09/2022-08/2023)
Matthieu Branthome, LRU, ENSAT, Univ. Rennes - Lannion (from december 2023)

International Associates

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

PhD students Fancois Mentec, CIFRE ALTEN (defense 30/11/2023)
Arthur Hoarau, Departemental and regional funding
Erwan Vincent, CIFRE KEOLIS
Aymen Bazouzi, CominLabs Clara project
Johan Le Boursicaud, scholarship and team's own ressources
Dorra Sassi, scholarship and regional funding (official starting date delayed to early 2024 because of visa problems)

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 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 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.

[BN03]

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.

[BN03] M. Belkin, P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation", *Neural Comput. 15*, 6, June 2003, p. 1373–1396, https://doi.org/10.1162/089976603321780317.

[GR17] M. Grohe, M. Ritzert, "Learning first-order definable concepts over structures of small

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) 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 EMBEDD-ER : EMBEDDing Educational Resources Using Linked Open Data

Participants: Aymen Bazouzi, Zoltan Miklos, Mickaël Foursov. Joint work with Hoël Le Capitaine, LS2N, Nantes..

There are a lot of educational resources publicly available online. Recommender systems and information retrieval engines can help learners and educators navigate in these resources. However, the available educational resources differ in format, size, type, topics, etc. These differences complicate their use and manipulation which raised the need for having a common representation for educational resources and texts in general. Efforts have been made by the research community to create various techniques to homogeneously represent these resources. Although these representations have achieved incredible results in many tasks, they seem to be dependent on the writing style and not only on the content. Furthermore, they do not generate representations that reflect a semantic representation of the content. In this work, we present a new task-agnostic method (EMBEDD-ER) to generate representations for educational resources based on document annotation and Linked Open Data (LOD). It creates representations that are focused on the content, compact, and can be generalized to unseen resources without requiring extra training. The resulting representations encapsulate the information found in the resources and project similar resources closer to one another than to non-similar ones. Empirical tests have shown promising results both visually and in a subject classification task. [14]

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

3.2 HDR-LFNet: Inverse tone mapping using fusion network

Participants: Mathieu Chambe, Zoltan Miklos. Joint work with Ewa Kijak, Kadi Boutouche, Remi Cozot, Olivier Le Meur.

To capture the real-world luminance values, High Dynamic Range (HDR) image processing has been developed. HDR images have a richer content than the widely-used Standard Dynamic Range (SDR) images, and are used in a number of situations, e.g. in film industry. As HDR displays are more and more commercially available, we need to be able to process HDR images as well as SDR ones (for example, devising denoising algorithms, inpainting or anti-aliasing). The most powerful methods to process images are now deep neural networks. However, the training of such networks calls for a lot of images, and HDR images datasets are relatively small. One way to generate HDR images is inverse tone mapping operators (iTMOs). They are algorithms that expand the dynamic range of SDR images. In this paper, we propose HDR-LFNet, a novel iTMO, and its HDR training dataset. Our method merges several existing handcrafted iTMOs, combined in a supervised neural network to produce an HDR output. Our lightweight network requires less training images than state-of-the-art methods, and has faster training phase. Besides, the quality of the generated images is similar to the state-of-the-art. We present the architecture as well as the subjective and experimental evaluations of our method. [5]

3.3 Identification of the factors that impact the commercial speed of urban buses

Participants: Erwan Vincent, Zoltan Miklos, Simon Malinowski (Linkmedia team).

We work on identifying the factors that impact the commercial speed of urban buses, based on data analysis techniques. We have made important advances on this topics, and we are in the process of consolidating an article on this topic. Besides this work we seek to improve the commercial speed prediction methods. We have developed a novel method based on graph neural networks. We are finalising our experiments and we plan to publish a second article on these advances.

3.4 AI4HR Recruiter: a job recommender system for internal recruitment in consulting companies

Participants: Francois Mentec, Zoltan Miklos.

Recruiting job candidates is a crucial activity for consulting companies to fulfill their consulting activities for customers or internal projects. Due to the complexity of the recruitment process, companies adopt standardized resume formats to be able to identify or compare candidates. However, this process requires significant involvement from human resource specialists and project leaders. In this paper, we present AI4HR Recruiter, a recommender system designed to improve the internal recruitment process of consultants. The system uses artificial intelligence-based techniques, specifically pre-trained

sentence transformers. It aims at supporting recruiters in identifying and selecting candidates, rather than automating the process. We measure improvement in efficiency in terms of gains in recruiters time induced by the recommendation explanation presence. We also measure the impact on user's satisfaction of two different types of algorithms to generate recommendations. To do so, we have been conducting a two fold study in a real-life setting for six months, where on the one hand, telemetry data was gathered while recommendation process user interaction was being performed and on the other hand, surveys about user's satisfaction were conducted afterward. Our results evaluation demonstrates the effectiveness of the system in assisting recruiters in candidates selection process. [19]

3.5 Temporal graph databases

3.5.1 RTGEN++: A Relative Temporal Graph GENerator

Participants: Maria Massri, Zoltan Miklos.

Graph management systems have become popular for storing and querying graph-oriented data, and they are often evaluated with benchmarks based on large-scale graphs. However, obtaining such graphs is difficult due to their limited public availability. Several graph generators have been developed to address this challenge, producing synthetic graphs with characteristics similar to real-world graphs, such as degree distribution, community structure, and diameter. Generating synthetic graphs with constantly changing topology has received less attention despite its importance in developing useful benchmarks for temporal graph systems. In this paper, we present RTGEN++, a temporal graph generator that supports two evolution models. The first model generates temporal graphs by controlling the evolution of their degree distributions, using optimal transport methods to minimize the transformation effort. We also extend our method in order to consider the community structure of the generated graphs. The second model allows one to control the number of added and removed graph entities, thus enabling the modeling of the evolution of real-world graphs in many use cases. Our generator also includes a decorator that adds types and time-varying attributes to nodes and edges, enhancing the generated graphs and aligning with data platforms that use the property graph model. We validate our approach with experiments that demonstrate the reliability of the generated graphs in approximating ground-truth parameters. [9]

3.5.2 TLDKS

Participants: Maria Massri, Zoltan Miklos.

Graphs are a ubiquitous data model for capturing entities and their relationships. Since most graphs that model real-world networks evolve over time, efficiently managing temporal graphs is an important problem from both a theoretical and practical perspective. Querying the history of temporal graphs can lead to new applications such as object tracking, anomaly detection, and predicting future behavior. However, existing commercial graph databases lack native temporal support, hindering their usefulness in

these use cases. This paper introduces Clock-G, a temporal graph management system designed to handle the history temporal graphs. What differentiates Clock-G from other temporal graph management systems is its comprehensive approach, covering query language, query processing, and physical storage. We define T-Cypher, a temporal extension of Cypher query language, enabling user-friendly and concise querying of the graph's history. Additionally, we propose a query processor that utilizes temporal statistics collected from underlying temporal graphs to offer a good evaluation plan for T-Cypher queries. We also propose a novel storage technique that balances space usage and query evaluation time. [10]

3.6 Reasoning over time into models with DataTime

Participants: David Gross-Amblard, Gauthier Lyan..

Models at runtime have been initially investigated for adaptive systems. Models are used as a reflective layer of the current state of the system to support the implementation of a feedback loop. More recently, models at runtime have also been identified as key for supporting the development of full-fledged digital twins. However, this use of models at runtime raises new challenges, such as the ability to seamlessly interact with the past, present, and future states of the system. In this paper, we propose a framework called DATATIME to implement models at runtime which capture the state of the system according to the dimensions of both time and space, here modeled as a directed graph where both nodes and edges bear local states (i.e., values of properties of interest). DATATIME offers a unifying interface to query the past, present, and future (predicted) states of the system. This unifying interface provides (i) an optimized structure of the time series that capture the past states of the system, possibly evolving over time, (ii) the ability to get the last available value provided by the systemâs sensors, and (iii) a continuous micro-learning over graph edges of a predictive model to make it possible to query future states, either locally or more globally, thanks to a composition law. The framework has been developed and evaluated in the context of the Intelligent Public Transportation Systems of the city of Rennes (France). This experimentation has demonstrated how DATATIME can be used for managing data from the past, the present, and the future and facilitate the development of digital twins.[8]

3.7 Evolving response modelling in a crowdsourcing context

Participants: Arnaud Martin, Jean-Christophe Dubois, Yolande Le Gall, Constance Thierry.

Crowdsourcing is a widespread method for outsourcing tasks to a crowd of contributors. These simple tasks are often formulated by multiple choice questionaries (MCQs) to which the contributor has to give a precise answer. We hypothesize that offering the contributor to fill in an imperfect answer and to evolve it is more profitable than a single precise answer. This year, we had proposed a model for the evolutionary answers of contributors to MCQs in crowdsourcing platforms. In order to realize this modeling we use the theory of belief functions. The model and experiments conducted on real

data from crowdsourcing campaigns are presented at the KES conference [22]. Our experiments show that modelling evolutionary responses using consonant mass functions improves the quality of the results obtained when aggregating responses compared with majority voting.

3.8 Developments around Logical Information Systems

Participants: Mireille Ducassé, Olivier Ridoux, with Peggy Cellier, Sébastien Ferré from IRISA team LACODAM, and Eric Wong, University of Texas at Dallas.

Formal Concept Analysis aims at finding relations among entities based on Lattice Theory. It suggests to extract sets of attributes as descriptors of entities and let Lattice Theory do the rest. We have shown in the past how to extend this basic principle to logical descriptors and a lattice based on logical deduction. This leads to Logical Concept Analysis (LCA). And we have shown how to base full-grown information systems on LCA. This leads to Logical Information Systems [**?**], known as LIS. Then we have studied various ways of implementing LIS, e.g., as file systems or as web services, extending LIS, e.g., to graphs, and applying LIS to various domains, e.g., Geographic Information Systems, Software Engineering, and Digital Humanities. Concerning Digital Humanities, we have recently shown how to apply a web-based LIS to the archives of the popularisation magazine La Nature (1873-1960, 150 volumes, 80 000 pages), and in particular, how to treat references to biological entities in the magazine [11, 21]. Concerning Software Engineering, we have recently published a book chapter on the localization of software faults [4].

3.8.1 Extracting ODRL Digital Right Representations from License Texts using AMR

Participants: Malo Revel, Aurélien Lamercerie, Annie Foret and Zoltan Miklos.

Keywords: Content Extraction, Automated Semantic Analysis, Semantic Graph, AMR (Abstract Meaning Representation), License Rights, ODRL (Open Digital Rights Language).

Licenses of digital resources describe rights and duties for users. If the licenses are expressed in natural language, as it is frequently the case, it is hard to reason and verify the license compatibility to specific uses. We propose an automatic end-to-end workflow for extracting Open Digital Rights Language (ODRL) representations from textual license documents. This process uses AMR semantic representations as an intermediate; it adapts a tool that performs a semantic transduction analysis, using formal rules. This work focuses on deontic modalities expressing the permissions and obligations of the user. We provide a proof of concept and discuss experiments.

This work was undertaken as part of the CLARA project, and is also discussed in Malo Revel Master2 internship. It was presented at ASAIL' 2023 (Automated Semantic Analysis of Information in Legal Text)[20].

3.8.2 Categorial Grammars for NLP

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

We have obtained and submitted a new result on 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)[1], while preserving advantages of the original CDG, in terms of expressivity, parsing and efficiency.

3.9 Belief functions

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

3.9.1 On computing evidential centroid through conjunctive combination: an impossibility theorem

The theory of belief functions (TBF) is now a widespread framework to deal and reason with uncertain and imprecise information, in particular to solve information fusion and clustering problems. Combination functions (rules) and distances are essential tools common to both the clustering and information fusion problems in the context of TBF, which have generated considerable literature. Distances and combination between evidence corpus of TBF are indeed often used within various clustering and classification algorithms, however their interplay and connections have seldom been investigated, which is the topic of this paper. More precisely, we focus on the problem of aggregating evidence corpus to obtain a representative one, and we show through an impossibility theorem that in this case, there is a fundamental contradiction between the use of conjunctive combination rules on the one hand, and the use of distances on the other hand. Rather than adding new methodologies, such results are instrumental in guiding the user among the many methodologies that already exist. To illustrate the interest of our results, we discuss different cases where they are at play. Impact Statement-Within the theory of belief functions, both distances and conjunctive combination rules can be used to achieve very similar purposes: evaluating the conflict between sources, performing supervised or unsupervised learning in presence of evidential information, or more simply obtaining a synthetic representation of multiple items of information. However, the results obtained by both approaches may show some inconsistency between them. This paper provides some insight as to why this may happen, showing that

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

the two approaches are definitely at odds, and that using distances is, for instance, incompatible with some fundamental notions of the theory of belief functions, such as the least commitment principle. We illustrate the importance of the studied differences on problems such as k-centroid clustering, and discuss the importance of interpretations in such problems, which is rarely done in the literature. [12]

3.9.2 BSC: Belief Shift Clustering

t is still a challenging problem to characterize uncertainty and imprecision between specific (singleton) clusters with arbitrary shapes and sizes. In order to solve such a problem, we propose a belief shift clustering (BSC) method for dealing with object data. The BSC method is considered as the evidential version of mean shift or mode seeking under the theory of belief functions. First, a new notion, called belief shift, is provided to preliminarily assign each query object as the noise, precise, or imprecise one. Second, a new evidential clustering rule is designed to partial credal redistribution for each imprecise object. To avoid the âuniform effectâ and useless calculations, a specific dynamic framework with simulated cluster centers is established to reassign each imprecise object to a singleton cluster or related meta-cluster. Once an object is assigned to a metacluster, this object may be in the overlapping or intermediate areas of different singleton clusters. Consequently, the BSC can reasonably characterize the uncertainty and imprecision between singleton clusters. The effectiveness has been verified on several artificial, natural, and image segmentation/classification datasets by comparison with other related methods. [13]

3.9.3 Evidential random forest

In machine learning, some models can make uncertain and imprecise predictions, they are called evidential models. These models may also be able to handle imperfect labeling and take into account labels that are richer than the commonly used hard labels, containing uncertainty and imprecision. This paper proposes an Evidential Decision Tree, and an Evidential Random Forest. These two models use a distance and a degree of inclusion to allow the model to group observations whose response elements are included in each other into a single node. Experimental results showed better performance for the presented methods compared to other evidential models and to recent Cautious Random Forests when the data is noisy. The models also offer a better robustness to the overfitting effect when using datasets that are effectively uncertainly and imprecisely labeled by the contributors. The proposed models are also able to predict rich labels, an information that can be used in other approaches, such as active learning. [7]

3.9.4 Datasets with Rich Labels for Machine Learning

Most datasets used for classification use hard labels. In this paper, five new datasets labeled with uncertainty and imprecision by crowdsourcing contributors are presented. Richer labels are modeled with the theory of belief functions, which generalizes several reasoning frameworks with uncertainty, such as possibilities or probabilities. These datasets can be used with classical models using hard labels but also with probabilis-

tic, fuzzy or even evidential models. Several concrete application cases are presented, for which these new datasets provide a useful knowledge representation of the user's uncertainty.[18]

4 Software development

4.1 HEADWORK platform

Participants: David Gross-Amblard.

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 CDG-UD software

Participants: Annie Foret, Denis Bechet (Univ. Nantes), Valérie Bellynck (Univ. Grenoble).

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

This software, in the domain of computational linguistics, aims at providing tools for handling universal dependencies (UD) with Categorial Dependency Grammars (CDG). On the linguistic data side, Universal dependencies has become a common annotation framework for linguistic data, providing over 200 treebanks in over 100 languages. On the grammatical framework side, categorial grammars have good formal properties (as to the syntax-semantics interface and their grammatical inference potential), this project chose the Categorial Dependency Grammar (CDG) framework, that relies on a dependency format (parsing a sentence in CDG generates a dependency tree or a dependency a graph).

The software is available on gitlab.inria.fr[3]. It consists in a module (CDG-pattern) that contains patterns related to CDG enabling corpus exploration with grew, a Graph Rewriting software for NLP (available at http://parse.grew.fr/); a second module (CDG-transform) contains a grs (graph rewriting systems) for transforming a corpus, adding CDG constructs to it. Examples are also provided. This work was partly described at CLTW-LREC 2022[4][17]. It has been applied to Breton data and could help in further studies within the SmartFCA ANR Project (WP4).

4.3 Kartu-Verbs : A Linked-data Knowledge Base of Inflected Georgian Verbs

Participants: Mireille Ducassé, Archil Elizbarashvili (TSU).

The Georgian language has a complex verbal system, both agglutinative and inflectional, with many irregularities. Inflected forms of a given verb can differ greatly from one another and it is still a controversial issue to determine which lemmas should represent a verb in dictionaries. Verb tables help people to track lemmas starting from inflected forms but these tables are tedious and error-prone to browse. We propose Kartu-Verbs, a Semantic Web base of inflected Georgian verb forms. For a given verb, all its inflected forms are present. Knowledge can easily be traversed in all directions: from Georgian to French and English; from an inflected form to a masdar (a verbal noun, the form that comes closest to an infinitive), and conversely from a masdar to any inflected form; from component(s) to forms and from a form to its components. Users can easily retrieve the lemmas that are relevant to access their preferred dictionaries. Kartu-Verbs can be seen as a front-end to any Georgian dictionary, thus bypassing the lemmatization issues [15]. KartuVerbs contains more than 5 million inflected forms related to more than 16 000 verbs; there are more than 80 million links in the base. Response times are acceptable when running on a private machine, thus validating the feasibility of the linked-data approach. There is still a need to validate, correct and expand data. Considering the mass of data, this requires tools. We are working on a process to transform textual structured knowledge into semantic linked data, applied to Georgian verbal knowledge. The process successively applies improvement tools. A specific one, using decision tree technique, complement occasional missing values. The results of the prediction are not 100% satisfactory. We are setting up a page of DRUID crowd-sourcing platform, HEADWORK (see 4.1), in order to get user-suggested data

[3]https://gitlab.inria.fr/foret/cdg-ud
[4]http://lrec-conf.org/proceedings/lrec2022/workshops/CLTW4/index.html

and expertise from professionals of the domain. The scripts produced so far are freely available[5]. They can be adapted to other applications to help transform data produced for given objectives into other data suited for different objectives [16].

4.4 Pinfig: A Visual Authoring Tool for Learning, Writing, Collaboration, and Personal Information Management

Participants: Mohamed Ez-zaouia.

Pinfig (https://pinfig.com/) is a collaboration visual authoring tool for interactive and multimodal content. It leverage a canvas+palettes interface, visualization, and a set of content authoring plugins. It can support different contexts of use, such as note-taking, content creation and presentation, and personal information and knowledge management (PIM, PKM).

Pinfig aims to enable users to author documents, view, filter, analyze, and make sense of a large corpus of documents, as well as create visual representations of content. Another goal of Pinfig is to integrate natural language processing and generative artificial intelligence tools to assist users in specific tasks, such as learning, writing, sensemaking, and collaboration.

5 Contracts and collaborations

5.1 National Initiatives

5.1.1 Project Clara

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

- Project type: Cominlabs
- Dates: 2021.12-2024.12
- 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] https://github.com/aelizbarashvili/KartuVerbs

5.1.2 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.

5.2 International collaborations

To build the Kartu-Verbs platform (see Section 4.3), Mireille Ducassé works in close collaboration with the Computer Science department of Ivane Javakhishvili Tbilisi State University - TSU, Georgia

5.3 Bilateral industry grants

5.3.1 Cifre ALTEN (until July 2023)

We collaborate with the company ALTEN, who finance in the form of a CIFRE contract the PhD thesis of Francois Mentec. Our collaboration focuses on the use of artificial intelligence techniques for recruitment and human resources tasks in general. In our collaborations we try to propose new ways how the artificial intelligence methods can support the work of recruiters. We do not intend to replace human recruiters or automatically affect consultants to project, rather to support the work of RH agents.

This collaboration was prolonged, as the advancement of thesis of Francois Mentec is slower as expected, however the contract will terminate in July 2023. However, we are discussing the preparation of a second CIFRE contract, on a different application domain. Annie Foret and Olivier Ridoux (ex-Semlis team) were involved in the discussions and they will be potentially co-supervisors of this thesis.

5.3.2 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 the LINKMEDIA team.

6 Dissemination

6.1 Promoting scientific activities

6.1.1 Scientific Events Organisation

Member of the Organizing Committees

- 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

- M. Ez-zaouia: PC member, CHI LBW, ACM Conference on Human Factors in Computing Systems (CHI), 2023
- A. Foret: PC member: ICGI'2023, LLMaNLPinAI at ICAART'2024

- D. Gross-Amblard: PC member: BDA'2023 (Demo)
- Z. Miklos: PC member: CIKM'2023, WSDM'2023, DSAA'20223, EGC'2023, ECMLPKDD'2023

Reviewer

6.1.3 Book

Member of the Editorial Boards

- A. Martin: Advances in Knowledge Discovery and Management vol. 10 [1]

6.1.4 Journal

Reviewer - Reviewing Activities

- D. Gross-Amblard: Neural Processing Letters
- Z. Miklos: VLDB Journal, Engineering Applications of Artificial Intelligence

Reviewer

- C. Thierrry : Expert Systems With Applications

6.1.5 Leadership within the Scientific Community

- Arnaud Martin:
  - president of EGC society[6]
- David Gross-Amblard:
  - member of BDA board[7]
- 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[8]
- Olivier Ridoux

[6] `http://www.egc.asso.fr`
[7] `http://bdav.org`
[8] `https://gdr-tal.ls2n.fr/`

  - member of CNRS GDS ÉcoInfo[9] and of "Collectif Numérique responsable" of the Poool (La French Tech Rennes-Saint-Malo)[10]

6.1.6 Scientific Expertise

- Zoltan Miklos : ANR, ANRT, European Commission

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, "Principes des systèmes informatiques", "Numérique éco-responsable", "Théorie des langages", and "Technologies Web"
    * Mickaë Foursov, "Programming" and "IHM programming"
  - Master
    * Zoltan Miklos, "Artificial Intelligence",
    * David Gross-Amblard, "OLAP and NoSQL databases",
    * Mickaël Foursov, David Gross-Amblard and Olivier Ridoux, "Information Intelligence",
    * Olivier Ridoux, "Logic Programming"
    * Mickaël Foursov, "Android Programming"
- Master: Arnaud Martin, "Data mining and data fusion", M2 research, ENSSAT.
- ESIR (Master level)
  - Zoltan Miklos, "Data management for BigData" and "Artificial Intelligence"
- ENSAI (National School of Statistics) David Gross-Amblard, NoSQL databases
- ENS Rennes:
  - David Gross-Amblard, "Préparation à l'agrégation d'Informatique" (databases)
  - Olivier Ridoux, "Epistémologie et philosophie des sciences" and "Sobriété numérique"
- International

[9] `https://ecoinfo.cnrs.fr`)
[10] `https://lepoool.tech/collectif-numerique-responsable/`

- 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).

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 manager of work-study program support
- Yolande Le Gall is in charge of company contacts for the second and third work-study years of the BUT in computer science.
- Jean-Christophe Dubois is manager of the second work-study years of BUT in computer science.
- Arnaud Martin was the director at IUT Lannion (September 2021-2023)
- Zoltan Miklos is the responsable of the option Information Systems, ESIR
- Constance Thierry Member of the Board of Directors of Société Informatique de France
- Annie Foret and Olivier Ridoux are co-heads of the second year of the computer science degree at ISTIC, Univ. Rennes.

6.2.3 Supervision

- PhD completed (juin 2023): Mathieu Chambe: Zoltan Miklos (co-director), Kadi Bouatouche (co-director emeritius), Remi Cozot (Univ Litoral) Improving Image Quality using High Dynamic Range and Aesthetics Assessment, (PhD stipend, ATER) [2]
- PhD completed: Francois Mentec, Explainable recommender system for recruiters and consulting companies, Zoltan Miklos. He defended on the 30 November 2023 [3]
- 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: (co-director) Zoltan Miklos, (co-director) David Gross-Amblard, Machine learning in graph databases

- Dorra Sassi: Modeling Human Learning through Imperfect Responses and Elicitation, (director) David Gross-Amblard, (co-supervisor) Constance Thierry
- Master Research Internship completed 2023 : Malo Revel, (co-supervisor) Annie Foret, (co-supervisor) Zoltan Miklos, on "Contributions à l'analyse automatisé des licences. Cas du français et des ressources pédagogiques".

6.2.4 Juries

- Z. Miklos: Davide Mario Longo, PhD defense, (rapporteur), July 2023, TU Wien, Vienna, Austria
- Z. Miklos: CSID Duc Thinh Ngo, IMT Atlantique Nantes
- A. Foret: selection committee for the recruitment of a teacher-researcher (COS), April-May 2023, ENSSAT, Lannion
- O. Ridoux: selection committee for the recruitment of a teacher-researcher (COS), April-May 2023, IUT Lannion

6.2.5 Awards

- The HEADWORK platform received the best demo award at BDA'2023.

6.3 Popularization

- David Gross-Amblard: "Tools for scientific research", at Lycée Descartes, Rennes (Terminal students in geopolitics, April 12, 2023)
- Z. Miklos: "What is artificial intelligence?", ISFEC Bretagne, Rennes, Continuing education for primary and secondary school teachers, co-animation of a 2-days course (8-9 March 2023)
- Z. Miklos: "What is artificial intelligence?", collège Charles Le Goffic, Lannion, 22 March 2023, presentation in two classes (6ieme)
- Z. Miklos: "What is artificial intelligence?", 7 July 2023, lecture at the pedagogical retreat of IFSEC Bretagne, for the instructions of teacher training programs
- Z. Miklos: "What is artificial intelligence?", 8 December 2023, presentation for the teachers (300+ teachers of primary and secondary schools)
- C. Thierry: An interactive unplugged computer workshop at the Festival of Science in Lannion
- C. Thierry: Participation in the organisation of the ADA LOVELACE challenge and member of the jury. This challenge is aimed at secondary school girls in Brittany to raise their awareness of IT.

- O. Ridoux: Communications on Green IT issues to Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE), Journées informatique de Institut national de physique nucléaire et de physique des particules (CNRS, IN2P3/IRFU), Journée annuelle du Réseau Aramis.

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] M. Chambe, *Improving Image Quality using High Dynamic Range and Aesthetics Assessment*, Theses, Univ Rennes, June 2023, https://hal.science/tel-04187749.

[3] F. Mentec, *Explainable Job Recommender Systems for Recruiters and Consulting Companies*, Theses, Université de Rennes, November 2023, https://hal.science/tel-04393964.

Articles in referred journals and book chapters

[4] P. Cellier, M. Ducassé, S. Ferré, O. Ridoux, W. E. Wong, "Data Mining-Based Techniques for Software Fault Localization", *in : Handbook of Software Fault Localization*, 1, Wiley, April 2023, p. Chapitre 7, https://hal.science/hal-04403506.

[5] M. Chambe, E. Kijak, Z. Miklos, O. Le Meur, R. Cozot, K. Bouatouch, "HDR-LFNet: Inverse tone mapping using fusion network", *Computers and Graphics 114*, August 2023, p. 1–12, https://hal.science/hal-04150900.

[6] M. Ez-Zaouia, R. Carrillo, "The Group Folding Effect: The Role of Collaborative Process Structuring and Social Interaction in Group Work", *ACM Trans. Comput.-Hum. Interact. 31*, 2, jan 2024, https://hal.science/hal-04194789.

[7] A. Hoarau, A. Martin, J.-C. Dubois, Y. Le Gall, "Evidential Random Forests", *Expert Systems with Applications*, June 2023, p. 120652, https://hal.science/hal-04121835.

[8] G. Lyan, J.-M. Jézéquel, D. Gross-Amblard, R. Lefeuvre, B. Combemale, "Reasoning over Time into Models with DataTime", *Software and Systems Modeling*, December 2022, p. 1–25, https://inria.hal.science/hal-03921928.

[9] M. Massri, Z. Miklos, P. Raipin, P. Meye, A. Bouchra Pilet, T. Hassan, "RT-GEN++: A Relative Temporal Graph GENerator", *Future Generation Computer Systems 146*, September 2023, p. 139–155, https://hal.science/hal-04439108.

[10] M. Massri, Z. Miklos, P. Raipin, P. Meye, "Clock-G: Temporal Graph Management System", *in : Lecture Notes in Computer Science*, *Transactions on Large-Scale Data- and Knowledge-Centered Systems*, *14160*, LIV, Springer, September 2023, p. 1–40, https://hal.science/hal-04439031.

[11] C. Morand, O. Ridoux, "Reconnaître les noms binomiaux des espèces biologiques dans un corpus ancien", *Interstices*, September 2023, https://inria.hal.science/hal-04231214.

[12] Y. Zhang, S. Destercke, Z. Zhang, T. Bouadi, A. Martin, "On computing evidential centroid through conjunctive combination: an impossibility theorem", *IEEE Transactions on Artificial Intelligence*, June 2022, p. 1–10, https://hal.science/hal-03698839.

[13] Z.-W. Zhang, Z.-G. Liu, A. Martin, K. Zhou, "BSC: Belief Shift Clustering", *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2022, p. 1–13, https://hal.science/hal-03816204.

Publications in Conferences and Workshops

[14] A. A. Bazouzi, M. Foursov, H. Le Capitaine, Z. Miklos, "EMBEDD-ER : EMBEDDing Educational Resources Using Linked Open Data", *in : 15th International Conference on Computer Supported Education (CSEDU), In Proceedings of the 15th International Conference on Computer Supported Education - Volume 1, ISBN 978-989-758-641-5, ISSN 2184-5026, Volume 1, ISBN 978-989-758-641-5, ISSN 2184-5026*, p. 439–446, Prague, Czech Republic, April 2023, https://hal.science/hal-04037990.

[15] M. Ducassé, A. Elizbarashvili, "Finding Lemmas in Agglutinative and Inflectional Language Dictionaries with Logical Information Systems: The Case of Georgian verbs", *in : Proceedings of XX EURALEX International Congress*, A. Klosa-Kückelhaus, S. Engelberg, C. Möhrs, P. Storjohann (editors), Mannheim: IDS-Verlag, 2022. Demonstration.

[16] A. Elizbarashvili, M. DucassÃ©, M. Khachidze, M. Tsintsadze, "From Structured Textual Data to Semantic Linked-data for Georgian Verbal Knowledge", *in : Proceedings of the eLex 2023 conference*, M. Medved, M. Mechura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubicek (editors), Lexical Computing CZ s.r.o. Brno, Czech Republic, 2023, https://elex.link/elex2023/.

[17] A. Foret, D. Béchet, V. Bellynck, "Iterated Dependencies in a Breton treebank and implications for a Categorial Dependency Grammar", *in : Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, Marseille, France, June 2022, https://hal.science/hal-03903545.

[18] A. Hoarau, C. Thierry, A. Martin, J.-C. Dubois, Y. Le Gall, "Datasets with Rich Labels for Machine Learning", *in : 2023 IEEE International Conference on Fuzzy Systems (FUZZ)*, IEEE, p. 1–6, Incheon, France, August 2023, https://hal.science/hal-04453391.

[19] F. Mentec, Z. Miklos, S. Hervieu, T. Roger, "AI4HR Recruiter: a job recommender system for internal recruitment in consulting companies", *in : 27th International Conference on Knowledge Based and Intelligent Information and Engineering Sytems (KES 2023), 225*, p. 1231–1240, Athens, Greece, September 2023, https://hal.science/hal-04438253.

[20] M. Revel, A. Lamercerie, A. Foret, Z. Miklos, "Extracting ODRL Digital Right Representations from License Texts using AMR", *in : ASAIL 2023 - Automated Semantic Analysis of Information in Legal Text*, p. 1–7, Braga, Portugal, September 2023, https://inria.hal.science/hal-04449668.

[21] O. Ridoux, C. Morand, "Extraction dans des textes anciens d'entités nommées de type binômes de la classification linnéenne du vivant : une étude de cas", *in : Extraction*

et Gestion des Connaissances (EGC) 2023, *RNTI-E-39*, Lyon, France, 2023, https://hal.science/hal-04447919.

[22] C. Thierry, A. Martin, Y. Le Gall, J.-C. Dubois, "Modeling evolutionary responses in crowdsourcing MCQ using belief function theory", *in : 27th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2023)*, Athens, Greece, September 2023, https://inria.hal.science/hal-04214522.