Activity Report 2020

Team SemLIS

Semantics, Logics, Information Systems

for Data-User Interaction

D7 – Data and Knowledge Management

Abstract. The main objective of the SemLIS team is **to bring back to users the power on their data**. It aims at facilitating data-user interaction by making users more autonomous and agile, by providing flexibility and expressivity, and yet control and confidence in the information system. It should support users in the semantic representation of heterogeneous data, and in the collaborative acquisition of domain knowledge. Its scientific foundations are logics and formal languages for knowledge representation and reasoning, the Semantic Web, information systems, natural language processing, symbolic data mining, and user-data interaction. A key idea is to reconcile the power of formal languages and the usability of natural language and interaction. On the application side, the focus will be put on social sciences and on business intelligence.

Keywords: information systems, knowledge representation, logics, formal languages, natural language processing, data mining, user-data interaction, business intelligence, group decision and negotiation.

1 Team composition

Researchers and faculty

Peggy Cellier, Associate Professor (HDR), INSA Rennes
Mireille Ducassé, Professor, INSA Rennes
Sébastien Ferré, Professor, Univ. Rennes 1, *head of the team*
Annie Foret, Associate Professor (HDR), Univ. Rennes 1
Olivier Ridoux, Professor, Univ. Rennes 1

Associate members

Shridhar B. Dandin, Sarala Birla University, Ranchi, India
Archil Elizbarashvili, Ivane Javakhishvili Tbilisi State University, Georgia (since October)

PhD students

Hugo Ayats (since October)
Francesco Bariatti
Aurélien Lamercerie, co-supervised with team HYCOMES

Administrative assistant

Gaëlle Tworkowski

2 Overall objectives

2.1 Overview

In a context of ever-increasing volumes of data and knowledge, both in quantity and in diversity (Big Data), **the main objective of SemLIS is to bring back to users the power on their data**. By users we mean any individual or group who has a strong interest over some data, and the need to exploit them in order to derive new knowledge and to take decisions. That includes tasks such as search, authoring, data mining, and business intelligence. Those data can range from the personal data of an individual to the information systems of large companies, through project management inside a team. We take a subjective view on "Big Data" where the complexity does not lie in efficiently performing a given task on a large volume of data (e.g., query evaluation), but in enabling users to perform tasks that could not be anticipated (e.g, query formulation). In that subjective view, "Big" only means an amount of data that is too large or too complex for users to grasp and analyze by hand or by simple tools (e.g., spreadsheets).

Our objectives fit in the scope of axis 26 (human-machine collaboration) of challenge 7 (society of information and communication) of the national strategy for research. We particularly agree with the notion of man-machine collaboration, where the machine is not supposed, in our view, to *replace* humans by full automation, but rather to *support* them in information-intensive tasks. In this view, both the human and the machine should learn one from the other.

> *One will review the human-computer interaction in the light of natural human behavior and progress in the decisional and operational autonomy of machines. To develop a real collaboration between man and machine, research on self-learning process between man and machine must be amplified. The machine should adapt to unpredictable aspects of user behavior, and develop a greater wealth of interactions for "intelligent" automation.*

That main objective of **bringing back to users the power on their data** can be decomposed into five high-level objectives:

AUTO (O1): to make users **autonomous and agile** in the process of exploiting data and knowledge by avoiding intermediates (e.g., database administrators);

SEM (O2): to facilitate the **semantic** representation and alignment of heterogeneous and multi-source data;

FLEX (O3): to provide **flexibility** by enabling out-of-schema data acquisition, and continuous evolution of the data schema;

CON (O4): to provide **control and confidence** in the information system by promoting transparency and predictability of system actions;

COLL (O5): to support the **collaborative** acquisition and verification of data and knowledge.

Those objectives are the different facets of a unique approach that targets user guidance as a trade-off between full automation (aka. artificial intelligence) and no automation (aka. adhoc programming). We are conscious that this set of objectives is ambitious but we think we can address them because we do not target the hard problems of full automation, and because we now have an effective design pattern, ACN (Abstract Conceptual Navigation) [Fer14a], to encapsulate an expressive formal language into data-user interaction and natural language.

2.2 Scientific foundations

A distinctive aspect of our team is the application of formal methods coming from software engineering and theoretical computer science (formal languages and grammars, logics, type theory, declarative programming languages, theorem proving) to artificial intelligence tasks (knowledge representation and reasoning, data mining, user-data interaction). This is explained by the combination of a theoretical background shared by permanent members and a real interest for data and their users. Some members, Olivier Ridoux and Mireille Ducassé, have had a long research experience in software engineering in general, and in logic programming in particular. Annie Foret studies different variants of substructural logics for the analysis of natural languages. Peggy Cellier did her PhD thesis on the application of data mining to the localization of faults in programs [CDFR18]. Sébastien Ferré relies on formal languages to formalize user-data interaction models, and to prove usability properties such as the safeness and completeness of user guidance.

We briefly describe the scientific foundations of the team, organized by high-level research topics, along with references to a few former contributions in each topic.

2.2.1 Knowledge Representation and Querying

The team uses symbolic approaches, and in particular the Semantic Web technologies [AvH04,HKR09]. Indeed, those are an active research domain, and provide W3C standards for concepts introduced by widely recognized formalisms for knowledge representation: e.g., Datalog [CGT89], description logics [BCM+03], or conceptual graphs [CM08]. The Semantic Web defines languages for the representation of facts and rules (RDF, RDFS, OWL, SWRL), and for their querying (SPARQL). Moreover, the Semantic Web has an active community, both in academy and in industry. That research domain solicits competencies in formal languages (syntax and semantics), in logics, and in automated

[AvH04] G. Antoniou, F. van Harmelen, *A Semantic Web Primer*, MIT Press, 2004.

[HKR09] P. Hitzler, M. Krötzsch, S. Rudolph, *Foundations of Semantic Web Technologies*, Chapman & Hall/CRC, 2009.

[CGT89] S. Ceri, G. Gottlob, L. Tanca, "What you Always Wanted to Know About Datalog (And Never Dared to Ask)", *IEEE Trans. Knowl. Data Eng. 1*, 1, 1989, p. 146–166.

[BCM+03] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, P. F. Patel-Schneider (editors), *The Description Logic Handbook: Theory, Implementation, and Applications*, Cambridge University Press, 2003.

[CM08] M. Chein, M.-L. Mugnier, *Graph-based knowledge representation: computational foundations of conceptual graphs*, *Advanced Information and Knowledge Processing*, Springer, 2008.

reasoning.

2.2.2 Natural Language Processing

Here again, the team uses symbolic approaches. One task is to extract structured and semantic information from texts. The employed techniques are: a) categorial grammars [MR12] associating syntactic/semantic types to words, b) Montague grammars [DWP81] associating grammars, lambda calcul, and logic, and c) sequential patterns [AS95]. Those techniques can be used for syntactic/semantic analysis of sentences, for Information Extraction (IE), and for defining Controlled Natural Languages (CNL) [Kuh13]. In those topics, we have for instance contributed to the learnability of pregroup grammars [BFT07], and their extension with option and iteration [BDF12], to a CNL (SQUALL) for querying and updating RDF graphs [Fer14b], and to the discovery of linguistic patterns from texts [BCCC12].

2.2.3 Symbolic Data Mining

The team has competencies in the conception and application of symbolic data mining algorithms, in particular for sequential patterns, and their application to texts. It also has competencies in learning the grammar of natural languages from a structured corpus [BFT07, FB19]. Moreover, the LIS team was scientifically founded on Formal Concept Analysis (FCA) [GW99]. It produced FCA-based contributions for data mining [CFRD08] and machine learning [FR02], as well as for data exploration [FH12].

2.2.4 User-Data Interaction

Because of the importance that we give to user-data interaction, the team invested into techniques that enable to structure and reason on those interactions. We can refer, in particular, to faceted search [ST09] (often used in e-commerce platforms), On-Line Analytical Processing (OLAP, often used in business intelligence) [CCS93], and Geographical

[MR12] R. Moot, C. Retoré, *The Logic of Categorial Grammars: A Deductive Account of Natural Language Syntax and Semantics*, *FoLLI-LNCS*, Springer, 2012, `https://hal.archives-ouvertes.fr/hal-00829051`.

[DWP81] D. R. Dowty, R. E. Wall, S. Peters, *Introduction to Montague Semantics*, D. Reidel Publishing Company, 1981.

[AS95] R. Agrawal, R. Srikant, "Mining Sequential Patterns", *in : Proceedings of the Eleventh International Conference on Data Engineering, ICDE '95*, IEEE Computer Society, p. 3–14, 1995.

[Kuh13] T. Kuhn, "A Survey and Classification of Controlled Natural Languages", *Computational Linguistics*, 2013.

[GW99] B. Ganter, R. Wille, *Formal Concept Analysis — Mathematical Foundations*, Springer, 1999.

[ST09] G. M. Sacco, Y. Tzitzikas (editors), *Dynamic taxonomies and faceted search, The information retrieval series*, Springer, 2009.

[CCS93] E. Codd, S. Codd, C. Salley, *Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate*, Codd & Date, Inc, San Jose, 1993.

Information Systems (GIS) [LT92]. In those topics, we have for instance contributed to the exploration of geographical data [BFRQ08], to the discovery of functional dependencies and association rules with OLAP cubes [AFR10], and to the extension of faceted search to RDF graphs [FH12].

2.3 Application Domains

The application field of SemLIS is widely open as it covers the field of the Semantic Web. According to a study done in September 2011, the Semantic Web that is available as Linked Open Data (LOD) counts 30 billions triples covering many domains: e.g., life sciences, media, governmental organizations, publications, geography. In addition to those public data, we can count the numerous internal data of companies and other organizations, as well as personal data. Social networks and wikis are yet another source of semantic data: e.g., photo annotations, relationships between people, restaurant ratings.

The approach to applications of the team is to first design generic information systems, then to evaluate the generic design on different use cases or domains, and finally to specialize and adapt it to a particular application if need be. This follows software engineering of reusability and orthogonality.

Our past and current experiences and collaborations have led us to target in priority the large domains below. In particular, we target users in the middle of the spectrum going from pure IT people to the general public, i.e., individuals and groups who are experts in a domain that implies data and knowledge management. Our objective is to enable those users to perform tasks that normally require IT technical competencies.

Social Sciences. Here, users are often other researchers in domains that have been strongly impacted by the increasing availability of digital data: e.g., geography, linguistics, law, group decision and negotiation. Our objective is not to solve their own scientific problems, but to make those users more autonomous and more efficient in the management and exploration of their data, and to guide them in the knowledge extraction process.

Business Intelligence. Here, users are groups of various sizes (e.g., teams, committees, companies, organizations) collaborating around one or several projects (e.g., strategic orientation, recruitment process). Our priority will go to small- to medium-sized groups because our emphasis is on expressivity rather than scalability. The objective is to enable a group to capitalize facts and knowledge continuously, to analyze data for self-evaluation or diagnostic, and help in decision making. To be effective, those functions should be coupled with information systems and private social networks.

[LT92] R. Laurini, D. Thompson, *Fundamentals of Spatial Information Systems*, Elsevier, Academic Press Limited, 1992.

3 Scientific achievements

3.1 Data-driven Assessment of Structural Evolution of RDF Graphs

Participants: Carlos Bobed, Pierre Maillot, Peggy Cellier, Sébastien Ferré.

Since the birth of the Semantic Web, numerous knowledge bases have appeared. The applications that exploit them rely on the quality of their data through time. In this regard, one of the main dimensions of data quality is conformance to the expected usage of the vocabulary. However, the vocabulary usage (i.e., how classes and properties are actually populated) can vary from one base to another. Moreover, through time, such usage can evolve within a base and diverges from the previous practices. Methods have been proposed to follow the evolution of a knowledge base by the observation of the changes of their intentional schema (or ontology); however, they do not capture the evolution of their actual data, which can vary greatly in practice.

We have proposed a data-driven approach to assess the global evolution of vocabulary usage in large RDF graphs [3]. Our proposal relies on two structural measures defined at different granularities (dataset vs update), which are based on pattern mining techniques. We have performed a thorough experimentation which shows that our approach is scalable, and can capture structural evolution through time of both synthetic (LUBM) and real knowledge bases (different snapshots and updates of DBpedia).

3.2 Widening for MDL-based Retail Signature Discovery

Participants: Clément Gautrais[1], Peggy Cellier, Matthijs van Leeuwen[2], Alexandre Termier[3].

Signature patterns have been introduced to model repetitive behavior, e.g., of customers repeatedly buying the same set of products in consecutive time periods. A disadvantage of existing approaches to signature discovery, however, is that the required number of occurrences of a signature needs to be manually chosen.

To address this limitation, we formalize the problem of selecting the best signature using the minimum description length (MDL) principle [18]. To this end, we propose an encoding for signature models and for any data stream given such a signature model. As finding the MDL-optimal solution is unfeasible, we propose a novel algorithm that is an instance of widening, i.e., a diversified beam search that heuristically explores promising parts of the search space. Finally, we demonstrate the effectiveness of the problem formalization and the algorithm on a real-world retail dataset, and show that our approach yields relevant signatures.

[1] KU Leuven - Catholic University of Leuven - Katholieke Universiteit Leuven
[2] LIACS - Leiden Institute of Advanced Computer Science [Leiden]
[3] Équipe Lacodam de l'IRISA

3.3 GraphMDL: Graph Pattern Selection based on Minimum Description Length

Participants: Francesco Bariatti, Peggy Cellier, Sébastien Ferré.

Many graph pattern mining algorithms have been designed to identify recurring structures in graphs. The main drawback of these approaches is that they often extract too many patterns for human analysis. Recently, pattern mining methods using the Minimum Description Length (MDL) principle have been proposed to select a characteristic subset of patterns from transactional, sequential and relational data. We have proposed an MDL-based approach for selecting a characteristic subset of patterns on labeled graphs [12, 11]. A key notion in this work is the introduction of ports to encode connections between pattern occurrences without any loss of information. Experiments show that the number of patterns is drastically reduced. The selected patterns have complex shapes and are representative of the data.

3.4 Graph-FCA: An Extension of Formal Concept Analysis to Knowledge Graphs

Participants: Sébastien Ferré, Peggy Cellier.

Knowledge graphs offer a versatile knowledge representation, and have been studied under different forms, such as conceptual graphs or RDF graphs in the Semantic Web. A challenge is to discover conceptual structures in those graphs, in the same way as Formal Concept Analysis (FCA) discovers conceptual structures in tables. FCA has been successful for analysing, mining, learning, and exploring tabular data, and our aim is to help transpose those results to graph-based data. Previous several FCA approaches have already addressed relational data, hence graphs, but with various limits.

We propose Graph-FCA [5] as an extension of FCA where a dataset is a hypergraph instead of a binary table. We show that it can be formalized simply by replacing objects by tuples of objects. This leads to the notion of "n-ary concept", whose extent is an n-ary relation of objects, and whose intent is a "projected graph pattern". We formally reconstruct the fundamental results of FCA for knowledge graphs. We describe in detail the representation of hypergraphs, and the operations on them, as they are much more complex than the sets of attributes that they extend. We also propose an algorithm based on a notion of "pattern basis" to generate and display n-ary concepts in a more efficient and more compact way. We explore a few use cases, in order to study the feasibility and usefulness of Graph-FCA. We consider two use cases: workflow patterns in cooking recipes and linguistic structures from parse trees. In addition, we report on experiments about quantitative aspects of the approach.

3.5 Practical Comparison of FCA Extensions to Model Indeterminate Value of Ternary Data

Participants: Priscilla Keip, Sébastien Ferré, Alain Gutierrez, Marianne Huchard, Pierre Silvie, Pierre Martin.

The Knomana knowledge base brings together knowledge from the scientific literature on the use of plants with pesticidal or antibiotic effects on animals, plants, and human beings to propose protection solutions using local plants. In this literature, the elements of the 3-tuple (protected organism, protecting plant, pest) are named using the binomial nomenclature consisting of the genus name followed by the species name. In some instances, authors use the abbreviation "sp." in the singular or "spp." in the plural, as species name, to indicate the indeterminate status of the species for a guaranteed genus. To suggest protection solutions, the indeterminacy of the species has to be hypothesized based on assigning the sp./spp. to the other species in the same genus and conversely. In this work [19], we discuss the classification of ternary data containing some indeterminate values generated by three extensions of Formal Concept Analysis (FCA): Triadic FCA, Relational Concept Analysis (RCA), and Graph-FCA [5].

3.6 Application of Concepts of Neighbours to Knowledge Graph Completion

Participants: Sébastien Ferré.

The open nature of Knowledge Graphs (KG) often implies that they are incomplete. Knowledge graph completion (aka. link prediction) consists in inferring new relationships between the entities of a KG based on existing relationships. Most existing approaches rely on the learning of latent feature vectors for the encoding of entities and relations. In general however, latent features cannot be easily interpreted. Rule-based approaches offer interpretability but a distinct ruleset must be learned for each relation. In both latent- and rule-based approaches, the training phase has to be run again when the KG is updated. We have proposed a new approach [7] that does not need a training phase, and that can provide interpretable explanations for each inference. It relies on the computation of Concepts of Nearest Neighbours (C-NN) to identify clusters of similar entities based on common graph patterns. Different rules are then derived from those graph patterns, and combined to predict new relationships. We evaluate our approach on standard benchmarks for link prediction, where it gets competitive performance compared to existing approaches.

3.7 Concepts of Neighbours in RDF Graphs: a Jena Extension, and a Graphical User Interface

Participants: Nicolas Fouqué, Sébastien Ferré, Peggy Cellier.

Concepts of neighbors define a symbolic similarity between the entities of a knowledge graph. Starting with an entity, each concept of neighbors is a cluster of neighbor entities that share a common graph pattern centered on the entity. In this work [17], we recall the definitions of concepts of neighbors, and we present a Jena library extension whose API enables to compute concepts of neighbors for an RDF(S) Jena model. We also present a graphical user interface that enables a user to perform those computations in a simple and interactive way.

3.8 Conceptual Navigation in Large Knowledge Graphs

Participants: Sébastien Ferré.

A growing part of Big Data is made of knowledge graphs. Major knowledge graphs such as Wikidata, DBpedia or the Google Knowledge Graph count millions of entities and billions of semantic links. A major challenge is to enable their exploration and querying by end-users. The SPARQL query language is powerful but provides no support for exploration by end-users. Question answering is user-friendly but is limited in expressivity and reliability. Navigation in concept lattices supports exploration but is limited in expressivity and scalability.

In this work [8], we introduce a new exploration and querying paradigm, Abstract Conceptual Navigation (ACN), that merges querying and navigation in order to reconcile expressivity, usability, and scalability. ACN is founded on Formal Concept Analysis (FCA) by defining the navigation space as a concept lattice. We then instantiate the ACN paradigm to knowledge graphs (Graph-ACN) by relying on Graph-FCA, an extension of FCA to knowledge graphs. We continue by detailing how Graph-ACN can be efficiently implemented on top of SPARQL endpoints, and how its expressivity can be increased in a modular way. Finally, we present a concrete implementation available online, Sparklis, and a few application cases on large knowledge graphs.

3.9 How to interact with medical terminologies? Formative usability evaluations comparing three approaches for supporting the use of MedDRA by pharmacovigilance specialists

Participants: Romaric Marcilly, Laura Douze, Sébastien Ferré, Bissan Audeh, Carlos Bobed, Agnès Lillo-Le-Louët, Jean-Baptiste Lamy, Cédric Bousquet.

Background: Medical terminologies are commonly used in medicine. For instance, to answer a pharmacovigilance question, pharmacovigilance specialists (PVS) search in a pharmacovigilance database for reports in relation to a given drug. To do that, they first need to identify all MedDRA terms that might have been used to codify an adverse reaction in the database, but terms may be numerous and difficult to select as they may belong to different parts of the hierarchy. In previous studies, three tools have been developed to help PVS identify and group all relevant MedDRA terms using three different approaches: forms, structured query-builder, and icons. Yet, a poor usability of the tools may increase PVS' workload and reduce their performance. This study [10] aims to evaluate, compare and improve the three tools during two rounds of formative usability evaluation.

Methods: First, a cognitive walkthrough was performed. Based on the design recommendations obtained from this evaluation, designers made modifications to their tools to improve usability. Once this re-engineering phase completed, six PVS took part in a usability test: difficulties, errors and verbalizations during their interaction with the three tools were collected. Their satisfaction was measured through the System Usability Scale. The design recommendations issued from the tests were used to adapt the tools.

Results: All tools had usability problems related to the lack of guidance in the graphical user interface (e.g., unintuitive labels). In two tools, the use of the SNOMED-CT to find MedDRA terms hampered their use because French PVS were not used to it. For the most obvious and common terms, the icons-based interface would appear to be more useful. For the less frequently used MedDRA terms or those distributed in different parts of the hierarchy, the structured query-builder would be preferable thanks to its great power and flexibility. The form-based tool seems to be a compromise.

Conclusion: These evaluations made it possible to identify the strengths of each tool but also their weaknesses to address them before further evaluation. Next step is to assess the acceptability of tools and the expressiveness of their results to help identify and group MedDRA terms.

3.10 Guided Construction of Analytical Queries on RDF Graphs

Participants: Sébastien Ferré.

As more and more data are available as RDF graphs, the availability of tools for analytical queries beyond semantic search becomes a key issue of the Semantic Web. Previous work require the modelling of data cubes on top of RDF graphs. We propose an approach [15] that directly answers analytical queries on unmodified RDF graphs by exploiting the computation features of SPARQL 1.1 (aggregations, expressions). We rely on the NAF design pattern [Fer16a] to design a query builder user interface that is user-friendly by completely hiding SPARQL behind a verbalization in natural language; and responsive by giving intermediate results and suggestions at each step. Our evaluations show that our approach covers a large range of use cases, and scales well on large datasets.

3.11 A Proposal for Nested Results in SPARQL

Participants: Sébastien Ferré.

Tables are a common form of query results, notably in SPARQL. However, due to the flat structure of tables, all structure from the RDF graph is lost, and this can lead to duplicates in the table contents, and difficulties to interpret the results. We propose an extension of SPARQL 1.1 aggregations to get nested results, i.e. tables where cells may contain embedded tables instead of RDF terms, and so recursively [16].

3.12 Visualization of Databases

Participants: Shridhar B. Dandin, Mireille Ducassé.

Interpreting data with many attributes is a difficult issue. A simple 2D display, projecting two attributes onto two dimensions, is relatively easy to interpret but provides limited help to see multidimensional correlations. We propose a tool, ComVisMD, which displays, from a dataset, five dimensions in compact 2D maps. A map contains cells; each one represents an object from the dataset. In addition to the usual horizontal

and vertical projections and the use of colors, we offer holes and shapes. In order to compact the display, we partition objects according to two dimensions, grouping values of each dimension into up to seven categories. This year's work focused on two case studies covering two different domains, a cricket player dataset and a heart disease dataset. The cricket dataset has 15 attributes and 2170 objects. We showed how, using ComVisMD, correlations between variables can be found in an intuitive way. The heart disease dataset has 14 attributes and 297 objects. Blokh and Stambler, in the June 2015 issue of "Aging and Disease," state that individual attributes show little correlation with heart disease. Yet in combination the correlation improves dramatically. We showed how ComVisMD helps visualize those multidimensional correlations between four attributes and heart disease diagnosis. This activity, done in collaboration with an international partner, has led to a publication of a chapter in a book [4].

3.13 Categorial Grammars and NLP

Participants: Annie Foret, Aurélien Lamercerie.

A part of our approach is to consider several classes of categorial grammars and discuss their learnability. We consider learning as a symbolic issue in an unsupervised setting, from raw or from structured data, for some variants of Lambek grammars and of categorial dependency grammars. In that perspective, we discuss for these frameworks different type constructors and structures, some limitations (negative results) but also some algorithms (positive results) under some hypothesis. On the experimental side, we also consider the Logical Information Systems approach, that allows for navigation, querying, updating, and analysis of heterogeneous data collections where data are given (logical) descriptors. Categorial grammars can be seen as a particular case of Logical Information System.

This general approach had been discussed by A. Foret at the 2018 LACompling conference (invited talk), and the post-conference paper [FB19]. The CDG (categorial dependency grammar) case is revisited and presented in details in our forthcoming paper in the Journal of Machine Learning [2].

This is also under experiment on recent linguistic data in the universal dependency format.

The approach has also been studied for the construction of formal representations of natural language texts. The mapping from a natural language to a logical representation is realized with a grammatical formalism, linking the syntactic analysis of the text to a semantic representation.

3.14 Meaning Representation and Semantic Transduction

Participants: Aurélien Lamercerie, Annie Foret.

Semantic representations provide an interesting intermediary between the natural expression of statement and their processing by automatic methods. They allow to formally capture the meaning of texts, making them more accessible for automatic

processing. We propose to exploit this kind of representations for documents analysis. Specifically, we provide a methodology to link natural language statements to formal models that can be exploited in a given context.Thus, the use of semantic structures allows the construction of pivot representations. From these representations, we define an analysis process named semantic transduction. The central idea is reflected by a series of transformations on the interpretation of a semantic graph, whose execution is guided using transduction patterns. This technique, which applies to any structure that can be reduced to a labeled graph, thus opening the way for the composition of simple, readable and adaptable processes for interpreting language statements natural.

In particular, we use Abstract Meaning Representation (AMR), supplemented by a transformation process to achieve the expected formal definitions [22]. We target the behavioral aspect of the specifications for cyber-physical systems, i.e. any type of system in which software components interact closely with a physical environment. In this way, the challenge would be to provide assistance to the designer. So, we could simulate and verify, by automatic or assisted methods, "systems" specifications expressed in natural language. We have proposed a new construction to meet this need, namely Deterministic Propositional Acceptance Automata [20], a formalism with good properties and adapted to integrate into a complete processing chain starting from statement in natural language.

4 Software development

4.1 Software development

4.1.1 Sparklis

Participants: Sébastien Ferré, Pierre Maillot.

Sparklis [Fer17] is a Web user interface that works on top of SPARQL endpoints, i.e. semantic data repositories. It is not tied to a particular endpoint, and works with any endpoint provided that it grants public access. The principle of Sparklis is to let users see and explore data and build expressive queries in natural language at the same time. A SPARQL query is built at the same time but it is only visible at the bottom of the page, for curious expert users. Users don't need to know the data schema, and discover it on the fly. They don't need to write anything, apart from filter values (e.g., matching keywords), which ensures that none of lexical, syntactic, and schema errors are introduced. Sparklis covers a large fragment of SPARQL: graph patterns, optional, union, negation, ordering, aggregation, main filters (string matching, inequalities and intervals, language or datatype). By default, Sparklis connects to DBpedia, a semantic version of the Wikipedia encyclopedia, and several other datasets are available: e.g., Mondial (geographical data), Bretagne tourism (touristic information in Brittany), Wikidata, Nobel prizes.

In 2020, the UI of Sparklis has again been further improved for the purpose of the PEGASE project, following a user study made by project partners who are human factor specialists (Laura Douze and Romaric Marcilly from CIC-IT / Evalab Lille).

4.1.2 Graph-FCA: computation and graphical display of concepts

Participants: Sébastien Ferré, Peggy Cellier.

Graph-FCA is a command-line tool for the computation and graphical display of Graph-FCA concepts from knowledge graphs, also known as multi-relational data. Graph-FCA [5] is an extension of Formal Concept Analysis (FCA) [GW99] to knowledge graphs where objects are nodes, and attributes are the labels of hyperedges. The intension of a Graph-FCA concept can be seen as a conjunctive query, combining a graph pattern and a projection tuple. The extension of a Graph-FCA concept equals the set of answers of the intension, and the intension is the most specific query for those answers.

The repository of the tool contains the source code, executable programs, a user manual, and examples of inputs and outputs. The inputs are textual files whose syntax is inspired by λProlog [BBR99] and RDF/Turtle, and the outputs are both textual and graphical (DOT and SVG files). Graph-FCA has been applied to genealogical data, descriptions of cooking recipes, environmental and health data, and linguistic data (parse trees). As an alternative to downloading the tool, a web service is available on A||GO.

4.1.3 SQUALL: a Semantic Query and Update High-Level Language

Participants: Sébastien Ferré.

SQUALL (Semantic Query and Update High-Level Language) is a controlled natural language (CNL) for querying and updating RDF graphs [Fer14b]. The main advantage of CNLs is to reconcile the high-level and natural syntax of natural languages, and the precision and lack of ambiguity of formal languages. SQUALL has a strong adequacy with RDF, and covers all constructs of SPARQL, and most constructs of SPARQL 1.1. Its syntax completely abstracts from low-level notions such as bindings and relational algebra. It features disjunction, negation, quantifiers, built-in predicates, aggregations with grouping, and n-ary relations through reification.

SQUALL is available as a Web application at `http://servolis.irisa.fr/squall/` under two forms: one that translates SQUALL sentences to SPARQL, and another one that directly return query answers from a SPARQL endpoint.

4.1.4 PEW: Possible World Explorer

Participants: Sébastien Ferré, Sebastian Rudolph.

The Possible World Explorer (PEW) [Fer16b] targets ontology designers, and aims to help them correct and complete their ontologies. It reuses the query-based faceted search principles of Sewelis for exploring the "possible worlds" (i.e., models) of an OWL

[GW99] B. Ganter, R. Wille, *Formal Concept Analysis — Mathematical Foundations*, Springer, 1999.

[BBR99] C. Belleannee, P. Brisset, O. Ridoux, "A Pragmatic Reconstruction of λProlog", *The Journal of Logic Programming 41*, 1999, p. 67–102.

ontology. Users are guided in the incremental construction of class expressions, such that only satisfiable classes are reachable. All classes made of qualified existential restrictions, nominals, intersections, unions, and atomic negations are reachable.

PEW not only supports the exploration of an ontology's possible worlds, but also supports its completion by the addition of axioms. When a class is found satisfiable, and this contradicts domain knowledge (e.g., a man that is not a person), the undesirable possible worlds can be excluded ("pew pew!") by asserting an axiom saying that this class is unsatisfiable (e.g., every man is a person). This could be made a game, where the player would strive to exclude as many undesirable worlds as possible. The benefits are to complete the ontology with more knowledge, and therefore to improve its deduction power.

In addition to completing existing ontologies, PEW also allows the edition of ontologies *de novo*. It allows for the extension of the signature by creating new classes, properties, and individuals; and it covers a wide range of OWL axioms.

4.1.5 Scikit-mine

Participants: Rémi Adon[4], Peggy Cellier, Alexandre Termier[5].

The tool is a library developed in Python in the context of the Inria ADT project Standim/Scikit-mine. The goal of Scikit-mine is to provide and widespread the recent and promising MDL-based pattern mining approaches. A challenge is also to be compatible with the very well-known Scikit-learn[6] in which pattern mining methods are missing. The library is available online[7].

4.1.6 TermLis

Participants: Annie Foret.

TermLis (2015-) is a collection of Logical information contexts for terminological resources (possibly with workflows) as an application of the Logical Information System approach to this field. The current version is to be used with Camelis.

4.1.7 Kartu-Verbs

Participants: Mireille Ducassé, Archil Elizbarashvili.

The Georgian language has a complex verbal system, both agglutinative and inflectional, with many exceptions. It is still a controversial issue to determine which lemmas should represent a verb in dictionaries. Verb tables help neophytes to track lemmas starting from inflected forms but if in paper documents they are tedious and error-prone to browse. We propose Kartu-Verbs, a Semantic Web base of inflected Georgian

[4]Engineer, Lacodam team
[5]Lacodam team
[6]`https://scikit-learn.org/stable/`
[7]`https://github.com/scikit-mine/scikit-mine`

verb forms. For a given verb, all inflected forms are present. Knowledge can easily be traversed in all directions: from Georgian to French and English; from an inflected form to a verbal noun that represent a verb ("masdar"), 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 dictionary. Kartu-Verbs can be seen as a front-end to any Georgian dictionary, thus bypassing the lemmatization issues. An article illustrates in detail how to use Kartu-Verbs [14]. Our base, in its current state, is already a successful proof of concept. It has proven helpful to learn about Georgian verbs. It can be accessed at https://www-semlis.irisa.fr/software/georgian-verb-inflected-forms-base/. Collaboration with a researcher from Ivane Javarishvili Tbilisi State University has started last Autumn to enlarge the scope of the tool.

4.1.8 Ares: Abstract Requirement Extraction for Systems

Participants: Aurélien Lamercerie.

Using formal methods to assist system design relies on expected behaviors modeling. The construction of these representations requires extracting behavior rules, called requirements, generally defined in a specification document. ARES (Abstract Requirement Extraction for Systems) meets this need starting from a statement of requirements in natural language. This tool operates an intermediate semantic representation (AMR), and converts it into exploitable formal requirements to model the behaviors of systems.

4.1.9 GraphMDL Visualizer: Interactive Visualization of Graph Patterns

Participants: Francesco Bariatti, Peggy Cellier, Sébastien Ferré.

Pattern mining algorithms allow to extract structures from data to highlight interesting and useful knowledge. However, those approaches can only be truly helpful if the users can actually understand their outputs. Thus, visualization techniques play a great role in pattern mining, bridging the gap between the algorithms and the users. We have developed GraphMDL Visualizer [13], a tool for the interactive visualization of the graph patterns extracted with GraphMDL, a graph mining approach based on the MDL principle (see 3.3).

GraphMDL Visualizer is structured according to the behavior and needs of users when they analyze GraphMDL results. The tool has different views, ranging from more general (distribution of pattern characteristics), to more specific (visualization of specific patterns). It is also highly interactive, allowing the users to customize the different views, and navigate between them, through simple mouse clicks. GraphMDL Visualizer is freely available online.

5 Contracts and collaborations

5.1 National Initiatives

5.1.1 PEGASE: Improved Pharmacovigilance and Signal Detection with Groupings

Participants: Sébastien Ferré, Annie Foret, Peggy Cellier.

- Project type: ANR
- Dates: 2016–2021
- PI institution: Univ. Rennes 1
- Other partners: LIMICS (INSERM U1142), Regional Centers for Pharmacovigilance in 4 University Hospitals (Besançon, Lille, Paris HEGP, Toulouse), CIC-IT Evalab

The SemLIS team was invited to join the PEGASE project for its Sparklis software, as a way to reconcile the formal aspect of Semantic Web languages, and the need for usability for the end-users, here pharmacovigilance experts.

The mission of those experts is to collect, annotate, store, analyze, and prevent the undesirable effects of drugs. They rely on the MedDRA terminology (Medical Dictionary for Regulatory Activities) to annotate new cases, and to retrieve former cases. An important issue is the large size of MedDRA (about 20,000 terms), and the fact that several terms must generally be used to retrieve all relevant cases from the base. A Semantic Web version of that terminology, the OntoADR ontology, already exists. It allows the precise querying of MedDRA with formal languages like SPARQL. The objective of the project is to develop and compare several user interfaces enabling pharmacovigilance experts to navigate and query the terminology in order to identify the relevant terms.

The leader of the project is Cédric Bousquet from SSPIM ("Service de santé publique et de l'information médicale") and CHU St Etienne. The project gathers computer scientists from LIMICS (INSERM U1142) and IRISA, pharmacovigilance experts from 4 regional centers (Besançon, Lille, Paris HEGP, Toulouse), and ergonomists in the medical domain from CIC-IT Evalab.

In 2020, he UI of Sparklis was further improved based on the first user study performed in 2019. A new version of the PEGASE knowledge base was developed to account for an evolution of the OntoADR ontology. This base was also made lighter to focus on the information needs of pharmacovigilance experts.

5.1.2 LangNum-br-fr: a DGLF-LF "Langue et numérique" Project

Participants: Annie Foret (coordinator), Karen Kechis (2018-2019), Pierre Morvan (2019-2020), Pierre Martinet (2021).

- Project type: Ministère de la culture, DGLF

- Dates: 2018, 2020
- PI institution: Univ. Rennes 1
- Other partners: Univ. Rennes 2, LIG (Grenoble)

This project (led by Annie Foret) is funded by the "Delegation générale à la langue française et aux langues de france" (DGLF-LF, French culture minister) in the theme "languages and digital" and concerns the French-Breton language pair. The general approach of the scientific project is multidisciplinary, involving computer scientists specialized in natural language processing [Partner A: IRISA and Rennes 1 University, Partner B: LIG Grenoble, Partner C: IT Laboratory in Tours), linguists specialized in Celtic languages [Partner D: CRBC and Rennes2] and specialists in ICT usage [Partner E: Loustic Laboratory]. This work includes technical design work (partners A, B, C in TAL), linguistic work (CRBC) and work on usages (Loustic).

The current challenge is to improve and develop resources and tools for Breton, in coordination between different disciplines, and with a pedagogical concern. A state of the art on tools and resources, and new proposals can be found in our previous contributions. Before defining a software development (a processing chain), an analysis of usages and needs is undertaken with support from a specific Loustic project involving one month engineer.

5.1.3 GTnum : Artificial Intelligence and Education

The proposal described in `https://chaireunescorel.ls2n.fr/2020/10/23/gt-numerique-cest-parti/` for a thematic group “L’impact de l’Intelligence Artificielle à travers l’Éducation Ouverte”, has been accepted by the French "Ministère de l’Éducation Nationale". The project is co-animated by Fahima Djelil (Brest) and Colin de la Higuera.

The SemLIS Team is a participant of this working group.

5.2 Collaborations

- Since the end of 2019, Peggy Cellier is involved in the ADT project SKM in collaboration with Alexandre Termier, Laurent Guillo and Rémi Adon (Engineer on the project since December 2019) about the development of a library of pattern mining tools compatible with the Scikit-learn python library.
- Mireille Ducassé collaborates with Ivane Javakhishvili Tbilisi State University, in Georgia (Caucasus). An informal collaboration is starting with Tina Margalitadze from the Lexicography team in relation with the Kartu-verbs project (see 4.1.7). She also collaborates with Sarala Birla University, India, on visualization of databases (see 3.12).
- Annie Foret collaborates with LS2N (research lab. Nantes), TALN team (Natural Language Processing), she is a member of “Agence Universitaire de la Francophonie” (AUF), LTT network on “Lexicologie, terminologie et traduction”. Annie Foret is member of ATALA (Association pour le Traitement automatique des Langues), and of SIF (Société Informatique de France).

6 Dissemination

6.1 Promoting scientific activities

6.1.1 Scientific Events Organisation

General Chair, Scientific Chair

- Sébastien Ferré acted as President of the Program Committee of the French-speaking days of Knowledge Engineering (IC) 2020 [1], which was part of PFIA 2020. Due to COVID-19, the conference was held virtually.

6.1.2 Scientific Events Selection

Member of Conference Program Committees

- Sébastien Ferré and Peggy Cellier are members of the Editorial Board of the International Conference on Formal Concept Analysis (ICFCA).
- Peggy Cellier was a member of the program committee of several conferences:
  - ICCS (Concept Lattices and Applications),
  - SDM (SIAM International Conference on Data Mining)

  She also served as a "Senior PC" for EGC.
- Sébastien Ferré was a member of the program committee of several conferences and workshops:
  - WWW (The Web Conference),
  - IJCAI (Int. Joint Conf. Artificial Intelligence),
  - ESWC (Semantic Web),
- Annie Foret was a member of the following program committees :
  - FG 2020 [postponed] (Formal Grammar International Conference). "FG provides a forum for the presentation of new and original research on formal grammar, mathematical linguistics and the application of formal and mathematical methods to the study of natural language."
  - ICGI 2020 [postponed, member of ICGI 2021 PC] The International Conference on Grammatical Inference (ICGI) is "the major forum for presentation and discussion of original research papers on all aspects of grammar learning."
  - NLA 2020 Special Session on Natural Language and Argumentation 2020 (NLA'20) at DCAI 2020: International Conference on Distributed Computing and Artificial Intelligence.
  - Annie Foret made a review for ECAI 2020 (European Conference on Artificial Intelligence)

6.1.3 Journal

Reviewer - Reviewing Activities

- Sébastien Ferré made reviews for the following journals:
  - JODS (Journal of Data Semantics)
  - TOIS (Transactions on Information Systems)
  - SOCO (Soft Computing)
  - ASE (Automated Software Engineering)
- In 2020-2021, Annie Foret is a reviewer for the Journal of Logic, Language and Information (https://www.springer.com/journal/10849).

6.1.4 Invited Talks

- In December 2020, Annie Foret gave an *invited talk* at "Édition 2020 des journées scientifiques du GdR LIFT" , where LIFT stands for « Linguistique Informatique, Formelle et de Terrain ». Her talk was On Categorial Grammatical Inference and Logical Information Systems. The summary is as follows. We consider several classes of categorial grammars and discuss their learnability. We consider learning as a symbolic issue in an unsupervised setting, from raw or from structured data and treebanks for some variants of Lambek grammars and of categorial dependency grammars. In that perspective, we discuss for these frameworks different type constructors and structures, some limitations (negative results) but also some algorithms (positive results) under some hypothesis. On the experimental side, we also consider the Logical Information Systems approach, that allows for navigation, querying, updating, and analysis of heterogeneous data collections where data are given (logical) descriptors. Categorial grammars can be seen as a particular case of Logical Information System.

6.1.5 Research Administration

- Olivier Ridoux is head of the AI transversal axis of IRISA.
- Sébastien Ferré is a member of the committee of the DKM scientific department (Data and Knowledge Management) at IRISA.
- Since September 2018, Peggy Cellier is in charge of the Irisa Ph.D. students at IRISA, i.e. she is involved in the "commission du personnel" and organizes the selection of Ph.D. students for ministerial grants (contrats doctoraux). She is also an elected member of the “Conseil de Composante IRISA/INSA” at INSA and an elected member of the "Conseil de laboratoire" at IRISA.

  She served as an external member of the selection committee for an associate professor position at INSA Lyon.

- Since the end of 2018, Sébastien Ferré is a member of the scientific committee of ABES, the Agency of Libraries in Higher Education, as an expert in Semantic Web technologies. Only one meeting took place (virtually) in 2020 because of COVID-19, about "collection and exposition of the metadata of scientific publications" (November 13th).
- Hugo Ayats and Francesco Bariatti take part in the organisation of monthly scientific seminars for the DKM department at IRISA and the organisation of the yearly "DKM day".
- Annie Foret is the team correspondant for the (new) GDR TAL.

6.1.6 Other services

- Mireille Ducassé takes part in the Mentoring program of IRISA as a mentor of a younger colleague.
- Olivier Ridoux is a member of the EcoInfo CNRS service group (GDS) on sustainable development and information technology (aka Green IT).
- Peggy Cellier is "secrétaire" of "Revue de Traitement automatique des langues"[8] since 2019.

6.2 Teaching, supervision

6.2.1 Administration

- Peggy Cellier organized the bibliographic and internship defense for the Research Master in Computer Science (SIF).

  She has also been involved in the IDPE (Ingénieur diplômé par l'état) diploma.

  She also helped the three persons in charge of each year at Computer Science department at INSA (3INFO, 4INFO and 5INFO) in the process of student selection for the options through the use of two tools (Wallet, Whishlis).

- Mireille Ducassé is the dean of international affairs of INSA Rennes since December 2010. As such, she is a member of the direction of INSA Rennes. She is an active member of the international relations committee of Groupe INSA. She is tightly involved in the working committee regarding international affairs for the constitution of UniR, the forthcoming University of Rennes.

  She is, in particular, responsible for exchange programs involving around 400 student mobilities and 30 staff mobilities per year. She set up a number of dual degrees programs over the past years. She supervises an Erasmus+ consortium for Groupe INSA and International credit mobility programs with *Tbilisi State University*, *Akaki Tsereteli State University* of Kutasi and *Georgian technical University* of Tbilisi in Georgia ; *Université Euro-Méditerranéenne de Fès* and *Institut National des Postes et Télécommunications* in Morrocco ; *Institut de Technologie*

[8] `https://www.atala.org/revuetal`

du Cambodge in Cambodgia, : *Université Cheikh Anta Diop* of Dakar and Université Gaston Berger de Saint Louis in Senegal ; as well as *Université Libanaise* in Lebanon. She is directly in charge of the management of the projects with Georgia.

- Sébastien Ferré is vice-director of the MIAGE at ISTIC.

  Along with Simon Malinowski, he created a new track on Data Science in the EIT Digital Master School at Univ. Rennes 1. The first master year opened this year in September, and it shares teaching resources with Master Miage.

  As of November, he is responsible of the first year of Master Miage, which includes four tracks: classic, alternance, EIT Data Science, and EIT Financial Technologies. They have 73 students in total this year.

- Annie Foret is an elected member of the scientic committee of ISTIC/Rennes 1. She is a member of the IRISA local committee on sustainable development. She was responsible of the internships of computer science students (Master 1 IL and SSR) until september 2018. In 2018-2021 she is responsible with Olivier Ridoux of the second year computer science studies at Rennes 1 university (the group has nearly 200 students).

  She participated in the recruitment committees of external candidates to the L2info level.

- Olivier Ridoux is an elected member of the administration board of ISTIC (CS and Electronic engineering departement of University of Rennes 1). He is co-head, with Annie Foret, for the second year CS studies (bachelor).

6.2.2 Teaching

- At INSA, Peggy Cellier is responsible of four courses: *Databases and web development* (Licence 3 INFO), *Databases* (Licence 3 Math) *Data Mining* (Licence 3) and *Advanced Database and Semantic Web* (Master 2). She also teaches some other courses: *Database* (Licence 2), *Use and functionalities of an operating system* (Licence 3).

  At master 2 SIF, she teaches in English 4 hours in the data mining course (DMV). In addition she gives a lecture of 2 hours also in master 2 SIF about "Qu'est-ce qu'une thèse, un doctorat, un·e doctorant·e ?".

- Mireille Ducassé, at INSA Rennes, is responsible of two courses, taught in English if international students are present: *Constraint Programming* at Master 1 level, as well as *User-centered Design* at Master 2 level. The latter course, entirely given online, has been open to a group of 30 Indian students from SBU for **virtual exchange**, a première at INSA Rennes.

- Sébastien Ferré teaches symbolic data mining, Semantic Web, and compiler techniques at the master level. He also teaches functional programming at license level. He also teaches Semantic Web at master 2 level at ENSAI Rennes (15h, 22 students).

This year, he created with Simon Malinowski two new teaching units for the new EIT Data Science track: basics of data analysis with Python, and technological watch.

- Annie Foret teaches university courses including formal logic and formal methods for computer scientists, XML technology and related notions and databases at ISTIC and ESIR, Rennes.
- Aurélien Lamercerie teaches compiler techniques at the master level. He also teaches scientific programming and principles of information systems at license level.
- Francesco Bariatti this year taught taught Database (Licence 2) and Data mining (Licence 3) at INSA Rennes.
- Hugo Ayats taught Database (License 2) at INSA Rennes.
- Olivier Ridoux teaches formal language theory and compiler design at ESIR and ISTIC, and logic and constraint programming, operating system, and epistemology at ISTIC. He participated in the ISTIC program for high-school teachers on Turing machines and on Green IT.

6.2.3 Supervision

- PhD in progress: **Hugo Ayats**, From prediction to automation with an explainable and user-centric AI, application to the construction of knowledge graphs from texts, started October 2020, supervised by Sébastien Ferré (50%) and Peggy Cellier (50%)
- PhD in progress: **Francesco Bariatti**, Semantic Lifting of Complex Data by Hypergraph Compression, started October 2018, supervised by Sébastien Ferré (50%) and Peggy Cellier (50%)
- PhD in progress: **Aurélien Lamercerie**, From texts carrying deontic modalities to their formal representations, started November 2017, supervised by Annie Foret and Benoît Caillaud[9]
- PhD in progress: **Josie Signe**, Animal welfare : characterizing the diversity between and within livestock farming situations with data mining methods used on information from dairy herd sensors, started September 2020, supervised by Yannick Lecozler (25%), Alexandre Termier (25%), Peggy Cellier (25%) and Véronique Masson (25%)
- research internship (M2): **Hugo Ayats**, on "Relation Extraction with Concepts of Neighbours", 5 months, supervised by Peggy Cellier and Sébastien Ferré
- research internship (M2): **Josie Signe**, on "Subgroup discovery for time series in precision agriculture", 5 months, supervised by Peggy Cellier, Christine Largouet, Véronique Masson and Alexandre Termier

[9] Team Hycomes - IRISA

- research internship (M2): **Thibaud Balem**, on "Customization and adaptation of mobile gaming experience", 5 months, supervised by Tassadit Bouadi, Peggy Cellier, Matthieu Marionneau and Alexandre Termier
- research internship: **Théo Losekoot** and **Matthieu Gillet** from ENS Rennes are on a research internship with SemLIS on the semantic elevation of the archive of the magazine LA NATURE: about 100 years of a weekly publication from about 1870 to 1970. The aim is to render the result as RDF data on the SparkLIS platform. Possible applications are researchs in history and epistemology of the period, and a first step toward a virtual museum of science.
- research internship: **Samuel Bouaziz** from ENS Rennes is on a research internship with SemLIS on the reversible computing. The aim is to animate reversible computations. Possible applications are research and teaching on reversible computing.
- internship (M1): **Jérémy Angora**, on "Development of a Web application based on Sparklis", 2 months, supervised by Sébastien Ferré.
- internship (LP USETIC, Rennes 2): **Pierre Morvan**, on "languages and digital" projects to help learners, 12 weeks in 2019-2020, supervised by Annie Foret.
- internship: **Tamar Sharabidze and Beka Chachua** from Tbilisi State University, as well as **Tamari Kldiashvili** from Akaki Tsereteli University in Kutaisi, Georgia, on the Kartu-Verbs project (see 4.1.7), 5 months, supervised by Mireille Ducassé.

6.2.4 Juries

- Sébastien Ferré served as president in the PhD committee of Lolita Lecompte on "Structural variant genotyping with long read data", supervised by Dominique Lavenier and Claire Lecompte, at University Rennes 1, on 04/12/2020.

  He also served in the “CSID” of Camille Guerry (Univ. Rennes 1) and Priscilla Keip (IMT Mines-Alès).
- Peggy Cellier served as examiner in the PhD committee of Clément Dalloux on "Text mining and information extraction in clinical data", Vincent Claveau supervisor, at University Rennes 1, on 07-12-2020.

  She also served in the “CSID” of Cheikh Brahim (Univ. Rennes 1), Cyrielle Mallart (Univ. Rennes 1), Erwan Bourrand (Univ. Rennes 1), Grégory Martin (Univ. Rennes 1) and Priscilla Keip (IMT Mines-Alès).
- Annie Foret served as jury member for the *PHD* of Luyen Ngoc Lê, 15-09-2020, on "French Language DRS Parsing", supervised by Yannis HARALAMBOUS and Philippe LENCA, IMT Atlantique Brest.
- Annie Foret is in the pre-PhD committee (CSI) of Hugo Talibart, in bioinformatics, on ”Learning grammars with long-distance correlations on proteins”, supervised by J. Nicolas and F. Costes at IRISA-Rennes.

- Annie Foret is in the pre-PhD committee (CSI) of Mathilde Régnault, on "Annotation et analyse de corpus hétérogènes", in Paris, within the PROFITEROLE ANR project (PRocessing Old French Instrumented TExts for the Representation Of Language Evolution), supervised by Sophie Prévost at Lattice – ENS.

6.3 Popularization

- In March 2020, within the framework of "A la découverte de la recherche", supervised par Rennes Rectorate, Mireille Ducassé was planned to give two presentations in high schools about "Logical Information Systems: Artificial Intelligence to Leverage Natural Intelligence". They have been postponed due to Covid19.

7 Bibliography

P. Allard, S. Ferré, O. Ridoux, "Discovering Functional Dependencies and Association Rules by Navigating in a Lattice of OLAP Views", *in : Concept Lattices and Their Applications*, M. Kryszkiewicz, S. Obiedkov (editors), CEUR-WS, p. 199–210, 2010.

N. Béchet, P. Cellier, T. Charnois, B. Crémilleux, "Discovering Linguistic Patterns Using Sequence Mining", *in : Int. Conf. on Computational Linguistics and Intelligent Text Processing (CICLing)*, A. F. Gelbukh (editor), *LNCS*, *7181*, Springer, p. 154–165, 2012.

D. Béchet, A. Dikovsky, A. Foret, "Categorial grammars with iterated types form a strict hierarchy of k-valued languages", *Theor. Comput. Sci. 450*, 2012, p. 22–30.

O. Bedel, S. Ferré, O. Ridoux, E. Quesseveur, "GEOLIS: A Logical Information System for Geographical Data", *Revue Internationale de Géomatique 17*, 3-4, 2008, p. 371–390.

D. Béchet, A. Foret, I. Tellier, "Learnability of Pregroup Grammars", *Studia Logica 87*, 2-3, 2007.

P. Cellier, M. Ducassé, S. Ferré, O. Ridoux, "Data Mining for Fault Localization: towards a Global Debugging Process", *Research report*, INSA RENNES ; Univ Rennes, CNRS, IRISA, France, 2018, `https://hal.archives-ouvertes.fr/hal-02003069`.

P. Cellier, S. Ferré, O. Ridoux, M. Ducassé, "A Parameterized Algorithm to Explore Formal Contexts with a Taxonomy", *Int. J. Foundations of Computer Science (IJFCS) 19*, 2, 2008, p. 319–343.

M. Ducassé, P. Cellier, "Using Bids, Arguments and Preferences in Sensitive Multi-unit Assignments: A p-Equitable Process and a Course Allocation Case Study", *Journal of Group Decision and Negotiation 25*, 6, 2016, p. 1211–1235.

A. Foret, D. Béchet, "On Categorial Grammatical Inference and Logical Information Systems", *in : Logic and Algorithms in Computational Linguistics 2018, Series:*

Advances in Intelligent Systems and Computing, Studies in Computational Intelligence, 2019, https://hal.inria.fr/hal-02462675v1/bibtex.

S. Ferré, P. Cellier, "Graph-FCA in Practice", *in : Int. Conf. Conceptual Structures (ICCS) - Graph-Based Representation and Reasoning*, O. Haemmerlé, G. Stapleton, C. Faron-Zucker (editors), *LNCS 9717*, Springer, p. 107–121, 2016, https://hal.inria.fr/hal-01405491.

S. Ferré, *Reconciling Expressivity and Usability in Information Access - From Filesystems to the Semantic Web*, Habilitation thesis, Matisse, Univ. Rennes 1, 2014, Habilitation à Diriger des Recherches (HDR), defended on November 6th.

S. Ferré, "SQUALL: The expressiveness of SPARQL 1.1 made available as a controlled natural language", *Data & Knowledge Engineering 94*, 2014, p. 163–188.

S. Ferré, "Bridging the Gap Between Formal Languages and Natural Languages with Zippers", *in : The Semantic Web (ESWC). Latest Advances and New Domains*, H. Sack, E. Blomqvist, M. d'Aquin, C. Ghidini, S. P. Ponzetto, C. Lange (editors), *LNCS 9678*, Springer, p. 269–284, 2016, https://hal.inria.fr/hal-01405488.

S. Ferré, "Semantic Authoring of Ontologies by Exploration and Elimination of Possible Worlds", *in : Int. Conf. Knowledge Engineering and Knowledge Management*, *LNAI 10024*, Springer, 2016, https://hal.inria.fr/hal-01405502.

S. Ferré, "Sparklis: An Expressive Query Builder for SPARQL Endpoints with Guidance in Natural Language", *Semantic Web: Interoperability, Usability, Applicability 8*, 3, 2017, p. 405–418.

S. Ferré, A. Hermann, "Reconciling faceted search and query languages for the Semantic Web", *Int. J. Metadata, Semantics and Ontologies 7*, 1, 2012, p. 37–54.

S. Ferré, O. Ridoux, "The Use of Associative Concepts in the Incremental Building of a Logical Context", *in : Int. Conf. Conceptual Structures*, G. A. U. Priss, D. Corbett (editor), *LNCS 2393*, Springer, p. 299–313, 2002.

Books and Monographs

[1] S. Ferré (editor), *IC 2020 : 31es Journées francophones d'Ingénierie des Connaissances, Angers, France, June 29 - July 3, 2020*, 2020.

Articles in referred journals and book chapters

[2] D. Béchet, A. Foret, "Incremental learning of iterated dependencies", *Journal of Machine Learning*, to appear.

[3] C. Bobed, P. Maillot, P. Cellier, S. Ferré, "Data-driven Assessment of Structural Evolution of RDF Graphs", *Semantic Web: Interoperability, Usability, Applicability 11*, 2020, p. 831–853, http://www.semantic-web-journal.net/content/data-driven-assessment-structural-evolution-rdf-graphs-0.

[4] S. B. Dandin, M. Ducassé, "ComVisMD-Compact 2D Visualization of Multidimensional Data: Experimenting with Two Different Datasets", *in : Intelligent Learning for Computer Vision*, H. S. et al. (editor), *Lecture Notes on Data Engineering and Communications Technologies*, *61*, Springer Nature Singapore Pte Ltd, 2021, https://hal.archives-ouvertes.fr/hal-03131685.

[5] S. Ferré, P. Cellier, "Graph-FCA: An extension of formal concept analysis to knowledge graphs", *Discrete Applied Mathematics 273*, 2020, p. 81–102, http://www.sciencedirect.com/science/article/pii/S0166218X19301532.

[6] S. Ferré, M. Kaytoue, M. Huchard, S. O. Kuznetsov, A. Napoli, *A guided tour of artificial intelligence research*, *II*, Springer, 2020, ch. Formal Concept Analysis: from knowledge discovery to knowledge processing (Chapter 13), p. 411–445, https://www.springer.com/gp/book/9783030061661.

[7] S. Ferré, "Application of Concepts of Neighbours to Knowledge Graph Completion", *Data Science: Methods, Infrastructure, and Applications*, 2020, To appear, https://datasciencehub.net/paper/application-concepts-neighbours-knowledge-graph-completion-0.

[8] S. Ferré, "Conceptual Navigation in Large Knowledge Graphs", *in : Complex Data Analysis with Formal Concept Analysis*, R. Missaoui, L. Kwuida, and T. Abdessalem (editors), Springer, 2021, To appear.

[9] F. Lécuyer, V. Gouranton, A. Lamercerie, A. Reuzeau, B. Arnaldi, B. Caillaud, "Unveiling the implicit knowledge, one scenario at a time", *Visual Computer*, 2020, p. 1–12, https://hal.inria.fr/hal-02879083.

[10] R. Marcilly, L. Douze, S. Ferré, B. Audeh, C. Bobed, A. Lillo-Le-Louët, J.-B. Lamy, C. Bousquet, "How to interact with medical terminologies? Formative usability evaluations comparing three approaches for supporting the use of MedDRA by pharmacovigilance specialists", *BMC Medical Informatics and Decision Making 20*, 261, 2020, https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-01280-1.

Publications in Conferences and Workshops

[11] F. Bariatti, P. Cellier, S. Ferré, "GraphMDL : sélection de motifs de graphes avec le principe MDL", *in : Extraction et Gestion des Connaissances (EGC)*, Bruxelles, Belgium, 2020, https://hal.inria.fr/hal-02511412.

[12] F. Bariatti, P. Cellier, S. Ferré, "GraphMDL: Graph Pattern Selection based on Minimum Description Length", *in : Symposium on Intelligent Data Analysis (IDA)*, 2020, https://hal.inria.fr/hal-02510517.

[13] F. Bariatti, P. Cellier, S. Ferré, "GraphMDL Visualizer: Interactive Visualization of Graph Patterns", *in : Graph Embedding and Mining (GEM), an ECML-PKDD workshop*, 2020, https://gem-ecmlpkdd.github.io/papers/GEM2020_paper_7.pdf.

[14] M. Ducassé, "Kartu-Verbs: A Semantic Web Base of Inflected Georgian Verb Forms to Bypass Georgian Verb Lemmatization Issues", *in : First Proceedings of XIX EURALEX Conference*, Z. Gavriilidou (editor), Euralex association, November 2020, https://euralex.org/publications/.

[15] S. Ferré, "Construction guidée de requêtes analytiques sur des graphes RDF", *in : Atelier Web des Données*, Bruxelles, Belgium, 2020, https://hal.inria.fr/hal-02452395.

[16] S. Ferré, "A Proposal for Nested Results in SPARQL", *in : ISWC 2020 Posters, Demos, and Industry Tracks*, K. Taylor, R. Gonçalves, F. Lecue, J. Yan (editors), *CEUR Workshop Proceedings*, *2721*, p. 114–119, 2020, http://ceur-ws.org/Vol-2721/paper527.pdf.

[17] N. Fouqué, S. Ferré, P. Cellier, "Concepts de voisins dans les graphes RDF : Une extension Jena et une interface graphique", *in : Extraction et Gestion des Connaissances (EGC)*, A. Cornuéjols, E. Cuvelier (editors), *RNTI*, *E-36*, Éditions RNTI, p. 483–490, 2020, http://editions-rnti.fr/?inprocid=1002617.

[18] C. Gautrais, P. Cellier, M. van Leeuwen, A. Termier, "Widening for MDL-Based Retail Signature Discovery", *in : Advances in Intelligent Data Analysis XVIII - 18th International Symposium on Intelligent Data Analysis, IDA 2020, Konstanz, Germany, April 27-29, 2020, Proceedings*, M. R. Berthold, A. Feelders, G. Krempl (editors), *Lecture Notes in Computer Science*, *12080*, Springer, p. 197–209, 2020.

[19] P. Keip, S. Ferré, A. Gutierrez, M. Huchard, P. Silvie, P. Martin, "Practical Comparison of FCA Extensions to Model Indeterminate Value of Ternary Data", *in : Int. Conf. Concept Lattices and Their Applications*, F. J. Valverde-Albacete, M. Trnecka (editors), *CEUR Workshop Proceedings*, *2668*, CEUR-WS.org, p. 197–208, 2020, http://ceur-ws.org/Vol-2668/paper15.pdf.

[20] A. Lamercerie, B. Caillaud, "An Algebra of Deterministic Propositional Acceptance Automata (DPAA)", *in : FDL 2020 - Forum on specification & Design Languages*, p. 1–8, Kiel, Germany, September 2020, https://hal.archives-ouvertes.fr/hal-02971772.

[21] A. Lamercerie, "ARES : un extracteur d'exigences pour la modélisation de systèmes", *in : EGC 2020 - Extraction et Gestion des Connaissances (Atelier - Fouille de Textes - Text Mine)*, p. 1–4, Bruxelles, Belgium, January 2020, https://hal.archives-ouvertes.fr/hal-02971727.

[22] A. Lamercerie, "Transduction sémantique pour la modélisation de système", *in : PFIA 2020 - Plate-Forme de l'Intelligence Artificielle (PFIA), rencontres RJCIA*, p. 1–6, Angers, France, June 2020, https://hal.archives-ouvertes.fr/hal-02971742.