SIRIUS ANNUAL REPORT 2021

sfi Centre for Research-based Innovation
The Research Council of Norway

UiO : University of Oslo

Contents

Executive Summary

I wish to thank everyone that contributed to SIRIUS in 2021. Your continued support and valued contributions have led to measurable results in a challenging environment. For the second consecutive year our community of researchers and partners has worked through the COVID pandemic and demonstrated a high level of resilience. Once again thank you.

What is evident, is that our research programs and the interfaces between the research programs are strengthening. This is a result of participation in innovation projects, formal SIRIUS meetings between research programs and informal networking within our growing research community.

An example of this is the Ontology Engineering, Semantic Integration and Domain-Adapted Data Science research programs. During 2021 they have contributed to impactful innovation projects lead by our industrial partners. Of particular importance are innovation projects that support Asset Information Modelling, Strong AI through hybrid approaches and the OBDA subsurface pilot. These innovation projects draw on expertise and methods from two or more research programs. Further strengthening the knowledge sharing and interactions between research programs.

It is encouraging to see our researchers participate in these partner lead innovation projects. From these activities we see considerable scope to expand our future research programs. We have also identified some limitations in existing methodologies and technologies that are hindering the efficiency of digital solutions to solve industrial challenges. An example of this is the need to develop a graphical tool for building system orientated models. This will have a positive impact on the efficiency of asset information modelling.

Asset Information Modeling is still in its infancy, and we see a need for significant research in the digital aspect of engineering. The interest in the domain adapted data science research program is also growing, as users and investors start to better understand the importance of strong AI. We now need to materialize the research opportunity by re-engaging with our partners, not only to re-align on the industrial challenges but also evaluate and prioritize opportunities that can form the core of our future research.

SIRIUS has strong international cooperation across various academic, research and industrial sectors. Even with the challenges of COVID we have continued to maintain and

Arild Waaler

leverage our international network. This impactful collaboration has led to several research innovations and is one of the major strengths of SIRIUS. An example of this is the SIRIUS OBDA subsurface pilot. This was a joint research project utilizing the Volve dataset that Equinor generously made available for research and the Norwegian Petroleum Directorate fact pages. The project aims to reduce the time required for subsurface experts to access, integrate and clean data. The project was demonstrated at the SIRIUS General Assembly in November 2021.

I would also like to congratulate the 12 students that completed their master's degrees and the three PhD candidates that successfully defended their thesis with SIRIUS in 2021. Our research community are looking forward to welcoming new master's students and PhD candidates into SIRIUS during 2022. We have seen over the previous years that the educational aspect from these programs is not only significant in terms of the individual's growth, but also a considerable contribution to our research.

Hopefully, we are now through the worst of the COVID pandemic. We see many research opportunities and once again we can meet face to face. Together we can identify future challenges and define the future research required to address these challenges. I am looking forward to working with our community of researchers and partners in 2022 to deliver on our research plan and develop our future research program.

SIRIUS RESEARCH PROGRAMS

Analysis of Complex Systems: Formal methods for executable design

Now a days industry is experiencing more and more the need for digitization and digital transformation. Supporting technologies to assist industry into such process needs to deal with complex systems. Characteristics of such systems are that they are usually very difficult to understand and analyse due to many interdependent processes happening at the same time. In the oil & gas domain, we can observe such systems in emergent supporting technologies such e.g., digital twins, big networks of heterogeneous autonomous systems, parallel processing of big data, etc. **Formal methods are mathematical approaches to support the rigorous specification, design, and verification of the development of digital systems. They are very powerful to capture the behaviours and interactions of complex systems, helping with the understanding of what is currently happening and to further develop predictive and prescriptive analysis. Currently the use of Formal Methods is being more and more accepted in industry. This program use and research formal methods for system specifications and formal methods to describe, predict and prescribe the behaviours and interactions of system executions based on the analysis of models.**

- **Objective 1 Theoretical development:**

  Analysis of formal models can span from lightweight automated simulations to heavyweight complex non-automated verification techniques. We aim to explore the middle ground, which we call systematic model and multi-model exploration.

- **Objective 2 Tool development:**

  Tool development (both back- and front- end) for the developed methods. Knowledge extraction, visualization, and fill gaps in industries/market are important for tool development.

In the past year, our research program has continue maturing the tools and methods developed in the group and has started to explore other application domains. We now detail some of these activities.

ABS - Abstract Behavioural Specification Language

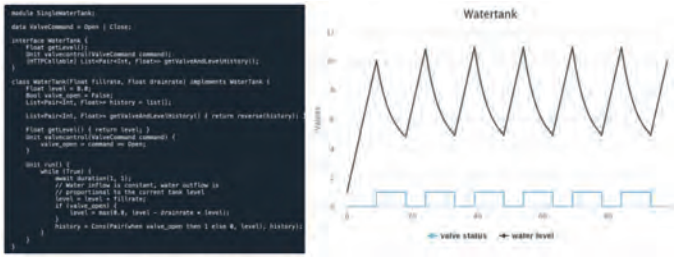


Figure 1: Specification and Simulation of a physical model in ABS.

ABS[1] (Abstract Behavioural Specification Language) is a language for executable design. ABS is a method and language for modelling, analysing and simulating distributed timed, resource-aware systems. In addition to supporting modelling of functional behaviour and distributed algorithms and systems, ABS supports the modelling of resource restrictions and resource management. It combines implementation-level specifications with verifiability, high-level design with executability, and formal semantics with practical usability. It is a concurrent, object-oriented, modelling language that features functional datatypes and supports model variability based on feature models and delta- oriented specifications. Deployment modelling can be based on high-level deployment models. The ABS system supports the modelling of resource-aware and resource-restricted systems and provides a range of techniques for model exploration and analysis, based on formal semantics.

ABS was designed to be easy to read and use by industrial programmers. It is an open-source[2] research project that is used in teaching and research, including industrial innovation research at Sirius.

For the end user, the main features and application areas of ABS are:

- Discrete-Event Simulation of timed, resource-aware systems
- Custom visualization of live simulation data

Silvia Lizeth Tapia Ta

Rudolf Schlatte

- Data export of simulation results into other tools
- Model exploration using formal analysis tools

Highlights of 2021

In 2021, the research group has developed a 90-minute ABS tutorial[3], presented at the DisCoTec 2021 conference. The tutorial focused on the ABS language, its approach to modelling, and on tool usage. The ABS toolchain and an accompanying paper is currently being reviewed have been accepted for publication in the Elsevier journal Science of Computer Programming.

ABS is being used in the integrated digital planning[4] and digital twins at SIRIUS, as shown in Figure 1, in collaboration with Equinor.

SAT/SMT - Satisfiability of formulas

Satisfiability Problem (SAT) is the problem of deciding the satisfiability[5] of a Boolean formula. A Boolean formula is satisfiable if there exists an interpretation that assigns truth values to the Boolean atoms such that the formula evaluates to true. For example, the formula

$$((a \vee \neg b) \wedge (\neg a \vee c))$$

is satisfiable, possible solutions are: {¬a,¬b,¬c}, and {a,¬b,c}; while the formula

$$((a \vee \neg b) \wedge (\neg a \vee b) \wedge (a \vee \neg b))$$

is unsatisfiable.

Satisfiability Modulo Theories (SMT) is the extension of SAT. It is the problem of deciding the satisfiability[5] of a quantifier -free first-order formula with respect to some decidable theories. For example, for the theory of linear arithmetic over the rational (LRA), a SMT(LRA) is the problem of checking the satisfiability of a formula φconsisting in atomic propositions A1, A2, A3, ... and linear-arithmetic constraints over rational variables like

$$(2.1x\_1 - 3.4x\_2 + 3.2x\_3 \leq 4.2)$$

combined by means of Boolean operators ¬,φφφφφAn LRA-interpretation μ is a function which assigns truth values to Boolean atoms and rational values to numerical variables. μ satisfies in the theory LRA if and only if μ makes the formula φevaluate to true. φis LRA-satisfiable if and only of it has at least one LRA-interpretation μ that satisfies it. The satisfiability of a SMT formula depends on the theory it based on. For example, the qualifier-free formula

$$((y \leq 4x) ((y > 0) (y \leq -4x+4))) ((x \geq 1) (y \leq 2)) ((x < 1) (y = (5x-4)))$$

is satisfiable over the theory of linear arithmetic over rational (possible solution: {x=0.87,y=0.35}); but is unsatisfiable over the theory of linear arithmetic over integers.

Many popular problems can be solved by SAT/SMT-based approaches, such as: travelling salesman problem, graph colouring problem, sudoku, knight tour, and dominating set. There are many highly efficient SAT/SMT that can solve many problem instances with hundreds of thousands of variables and millions of constraints. Some state-of-the-art solvers are Z3, Yices, MathSAT, MaxSAT.

SAT/SMT methods are being used in the integrated digital planning demonstration. The work mainly focuses on applying SAT/SMT-based automated reasoning techniques to model, verify, and optimise the vessel schedules and cargo transport to improve the workflow of the planners, in a use case provided by Equinor.

SMOL - Semantic Micro Object Language

SMOL is a framework that integrates knowledge bases with object-oriented programming languages which uses Semantic Web technologies. The integration of semantic technologies directly into a programming language is used to ensure the correct usage of semantic meaning in the program through type systems and other tools. Through the clean separation between data modelling and programming, SMOL integrates smoothly with industry-strength frameworks and builds on the expertise already present for existing semantic technologies, such as SPARQL, OWL or RDF.

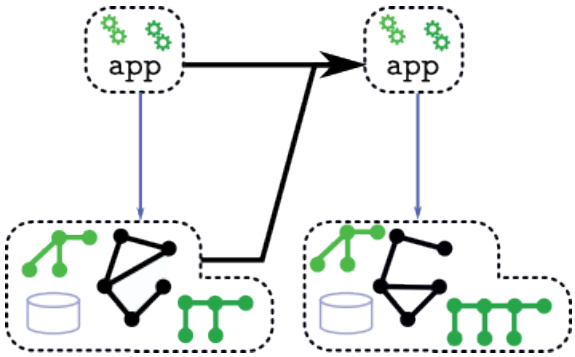
Recently, SMOL was extended by:

(1) a novel type system[6] that integrates query containment and description logic reasoning to ensure formal safety properties for SMOL programs,

(2) a new data loading mechanism using lazy data structures and future, akin to the mechanism found in the ABS modelling language,

(3) an integration of simulation units based on the industrial FMI standard which gives additional safety guarantees concerning correct connections between simulation units through both static type system and dynamic runtime monitoring using SHACL shapes.

SMOL is being developed in the context of the project PeTwin, as a co-simulation framework to orchestrate different Functional Mock-ups Units[7] (FMUs), developing methods towards predictive and prescriptive analysis for Digital Twins.



Combining Rewriting logic with Semantic Technology

Rewriting logic[8] (RL) is a logical framework in which other logics can be represented and a framework for transition systems, in which many different models of concurrency, distributed algorithms, programming languages, and different systems can be naturally represented, executed and analysed as rewrite theories[9], which include a set of rewriting rules expressing state transitions. Semantic Technologies covers different techniques to attach semantic meaning to data, as a formal and conceptual description of a relevant domain. The combination of formal methods and Semantic Technology is used in this project to capture the knowledge of the geology domain and to do multi-scenario reasoning on geological processes. More details about this combination can be found in the geological multi-scenario reasoning project at SIRIUS[10].

REMARO: Reliable Marine Robotics

REMARO
RELIABLE AI FOR MARINE ROBOTICS

REMARO[11] is a European Training Network (ETN) on Reliable artificial intelligence for marine robotics. It started in December 2020 and it is funded by the EU Horizon 2020. REMARO ETN is a consortium of experts in submarine AI, software reliability and marine safety certification (DNV). REMARO ETN is created to educate 15 PhD students, two of them at University of Oslo (UiO). REMARO will develop technology for AI methods with quantified reliability, correctness in specifications, models, tests, and analysis & verification for autonomous systems.

The two PhD topics at UiO will focus on safety in self-adaptive meta-controllers and on robustness of controllers with evolving knowledge-based systems. Technology developed in REMARO will be relevant for Digital Twins.

Smart Journey Mining (SJM)

SJM[12] is a research project on digitalization of services. It started in 2021 and it is funded by the The Research Council of Norway. SJM is a consortium of experts in service science, process mining and formal modelling and analysis. SJM is funding one PhD students at University of Oslo (UiO) and one postdoc at SINTEF. The PhD topic at UiO will focus on the development of predictive and prescriptive analysis to improve the user experience of services[13]. Technology developed in SJM can be relevant for the integrated digital planning and digitalization of information systems at SIRIUS.

Team Members

Chi Mai Nguyen	Gianluca Turin
Rudolf Schlatte	Chinmayi Baramashetru
Violet Pun	Paul Kobialka
Einar Johnsen	Juliane Päßler
Ingrid Yu	Tobias John
S. Lizeth Tapia Tarifa	Erik Voogd
Eduard Kamburjan	

Collaboration

Semantic Technology Digital Twins

Contact

S. Lizeth Tapia Tarifa, email: sltarifa@ifi.uio.no

[1] Tool documentation: https://abs-models.org/
[2] Tool: https://github.com/abstools/abstools
[3] Tutorial: https://link.springer.com/chapter/10.1007/978-3-030-78142-2_1
[4] Webpage: https://sirius-labs.no/integrated-digital-planning/
[5] Book: https://www.iospress.com/catalog/books/handbook-of-satisfiability-2
[6] Paper: http://ceur-ws.org/Vol-2954/paper-19.pdf
[7] Webpage: https://fmi-standard.org/
[8] Webpage: https://en.wikipedia.org/wiki/Rewriting
[9] Paper: https://www.sciencedirect.com/science/article/pii/S1567832612000707
[10] Webpage: https://sirius-labs.no/geological-assistant/
[11] Project: https://remaro.eu/
[12] Project: https://www.sintef.no/en/projects/2021/smart-journey-mining-towards-successful-digitalisation-of-services/
[13] Paper: https://ieeexplore.ieee.org/document/9592471

Ontology Engineering

The digital transformation of the industry depends on rich information models in order to support automation of specialized and knowledge intensive tasks. These models must be intelligible and usable by both computers and humans and should ideally represent the concepts and relationships in a manner to which domain experts are accustomed. This way users and systems may explore and extract implicit information from data through the help of automated reasoning without the need for understanding the technical details of how and where the data is stored.

Martin G. Skjæveland

However, the construction, maintenance, and use of such a model, called an ontology, are far from straight forward. Creating and maintaining a high-quality ontology requires close collaboration between domain experts, information modellers, and ontology experts to ensure that the model works as intended. Furthermore, an ontology quickly becomes a very complex artefact in order to express and make use of all the desired information objects. This makes maintaining the ontology a real issue.

The aim of the ontology engineering research program is to **develop tools and methods that improve the usability, efficiency and quality of ontology development, maintenance and use in the industry, by**

- **lowering the barrier for domain experts to understand, build, and use ontologies without the support of ontology experts.**
- **providing programmers and information modellers with powerful interfaces for interacting with and exploiting the knowledge captured in the ontology with existing software platforms.**
- **equipping ontology experts with powerful tools to oversee the development of the ontology**

Work in the research program is primarily performed in two projects: the *pattern-based ontology engineering project and the information modelling framework* project.

The pattern-based ontology engineering project has developed the **Reasonable Ontology Templates (OTTR)** framework. OTTR is a language and framework for representing and instantiating recurring patterns for engineering ontologies. This allows building and interfacing with the ontology at a higher level of abstraction than what is possible using the current standard ontology language OWL. This includes:

- Building ontologies and knowledge bases by instantiating templates;
- presenting, transferring and visualising the knowledge base as a set of template instances at different levels of abstraction; and
- securing and improving the quality and sustainability of the knowledge base via structural and semantic analysis of the templates used to construct the knowledge base.

Members of the project are Martin G. Skjæveland, Leif Harald Karlsen, Christian Kindermann, Oliver Stahl. The framework is available as open source specifications and applications which are in active use by several industrial partners in and outside of SIRIUS, including DNV GL, Aibel, Grundfos and CapGemini.

For more information about OTTR, including interactive examples, specifications and research papers, see its homepage http://ottr.xyz.

In addition to the OTTR framework, the project works on research topics for identifying and characterising patterns in ontologies. OWL ontologies are built and maintained on the basis of all sorts of methods and methodologies using a wide range of tools. As ontologies are primarily published as sets of axioms, their underlying design principles often

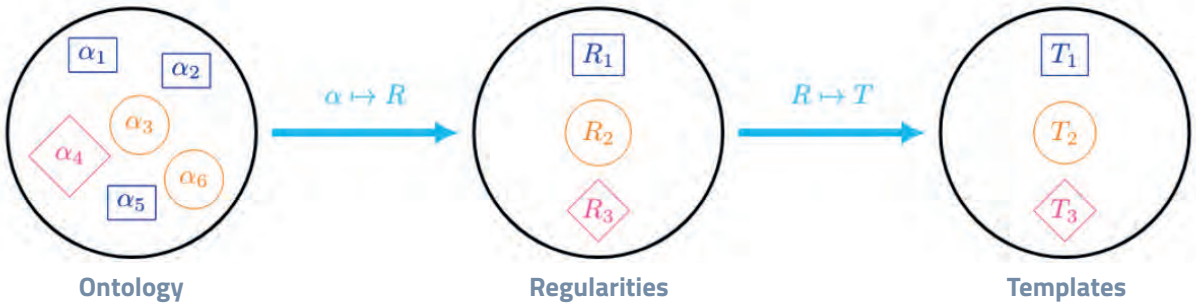


Figure 1: Reverse-engineering ontologies: Automated regularity extraction combined with automated template creation.

remain opaque. However, a principled and systematic ontology design is likely to be reflected in regularities for axioms. Identifying such regularities may help to recover and unveil conscious design choices and otherwise recurring modelling practices.

This reverse-engineering of ontologies provides a starting point for high-level language services, e.g., automated rewritings from OWL into OTTR for the purpose of redundancy removal, model verification, data validation, only to name a few.

Highlights of 2021

- A concept paper for the Information Modelling Framework was delivered by the READI Joint Industry Project
- IMF is implemented in the NOAKA field development project to develop, share and exchange overarching process and electricity supply models
- The IMF editor Mimir was released for use
- The 2nd OTTR user forum took place in June 2021 with more than 20 participants from a multitude of industrial companies and universities. Featuring talks from Veronika Heimsbakk (CapGemini), Basil Ell (SIRIUS) and Moritz Blum (Bielefeld University), and Johan Klüwer (DNV).
- The OTTR framework continues to receive attention from our industrial partners and is a central part of projects that apply semantic technologies.
- OTTR has been integrated with Semantic Mediawiki through a plugin developed by Basil Ell and colleagues/ students at Bielefeld University.

Several master students at UiO undertake a master thesis topic related to the OTTR project. The topics include:

- Developing a translation from OTTR templates to SPARQL queries to allow templates to be used as queries.
- Developing a functional term manipulation language that builts on and complements OTTR by supporting features such as concatenating strings to construct IRIs.
- Characterising and implementing OTTR constructs and the expansion mechanism with a relational database.
- Developing algorithms for synchronised updates and expansion of OTTR instance data.
- Developing a visual language for OTTR templates.

Information Modelling Framework

The Asset Information Modelling Framework (IMF) was initiated in the READI Joint Industry Project and has in 2021 been further developed, with participation from SIRIUS partners and researchers, in the NOAKA field development project. While the IMF project aims at innovation in the industry, SIRIUS researchers contribute with methods and theoretical foundation, advancing scalability of semantic technologies for complex industrial use cases.

IMF targets the need in the industry for developing more precise structuring of asset information. It aims to deliver methods and resources that enable the design of:

- Models of asset information so that questions that today can only be answered by discipline experts, can instead be answered by querying a computer,
- Industry commons resources, such as properties, types and classes, can be shared and used by all parties in the value chain,
- Machine to machine sharing through publishing so that pieces of asset information can be seamlessly reused in a «single source of truth» manner,
- Protocols for exchange of information and for dynamically allocating authoring rights in compliance with the needs of companies in the value chain.

IMF as fully implemented in a field development project will consist of several components, as illustrated in Figure 2. An IMF asset model digitally shared and exchanged in a cloud environment is linked to a reference data library, typically delivered by the Posc Caesar Association (PCA). The IMF model is created using software that is prototyped in the IMF project:

- The Mimir model builder is an editor that supports development of IMF models
- The Type editor supports development of common industry types for the objects in IMF models
- Mechanisms to link project specific data in engineering registers to the IMF model

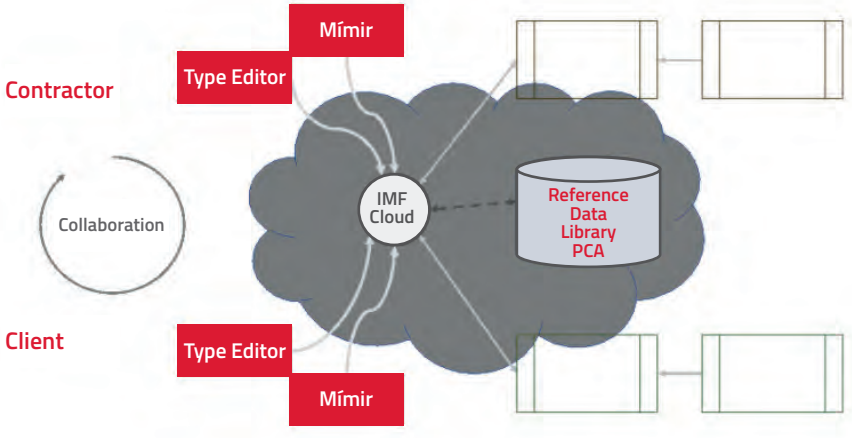


Figure 2: A vision of IMF implemented in a field development project.

IMF is based on best practice from industry and state of the art technology and research, advancing and sharpening principles from several standards and industry initiatives and integrating them in a coherent and scalable way. Key standards and technologies for IMF are summarized in Table 1.

Objective	Standards basis	Technologies
Describe the asset	ISO/IEC 81346	Mimir model builder
Exploit shared libraries	ISO 15926 / PCA RDL	Type editor
Represent models	OWL, SHACL	OTTR / Lutra
Publish models	RDF	Reference architecture
Exchange models	Industry 4.0	AAS adaptation

Table 1: Standards and technologies that form the basis for IMF.

Today, asset information is created using different methods, tools, and work processes, resulting in misalignment of information created in different contexts. IMF aims to overcome this misalignment by offering methods capturing the specific perspectives of any context in which information is created and relating information created in one context to information created in other contexts. Figure 3 illustrates an IMF model that evolves through a project lifecycle. The colors in the figure distinguish different aspects, a key concept that IMF borrows from the ISO/IEC 81346 standard. Primary aspects are requirements to function, requirements from location, and specification to product realization. The figure also shows two relations that are captured in IMF models:

- A hierarchical system of system decomposition which captures how the design space is repeatedly constrained by design decisions. The elements in the decomposition reflect different level of granularity in system descriptions.
- A relation between system objects of different aspects that captures transition from one stage in the project lifecycle to another. These different stages will typically introduce system objects at different levels of abstraction.

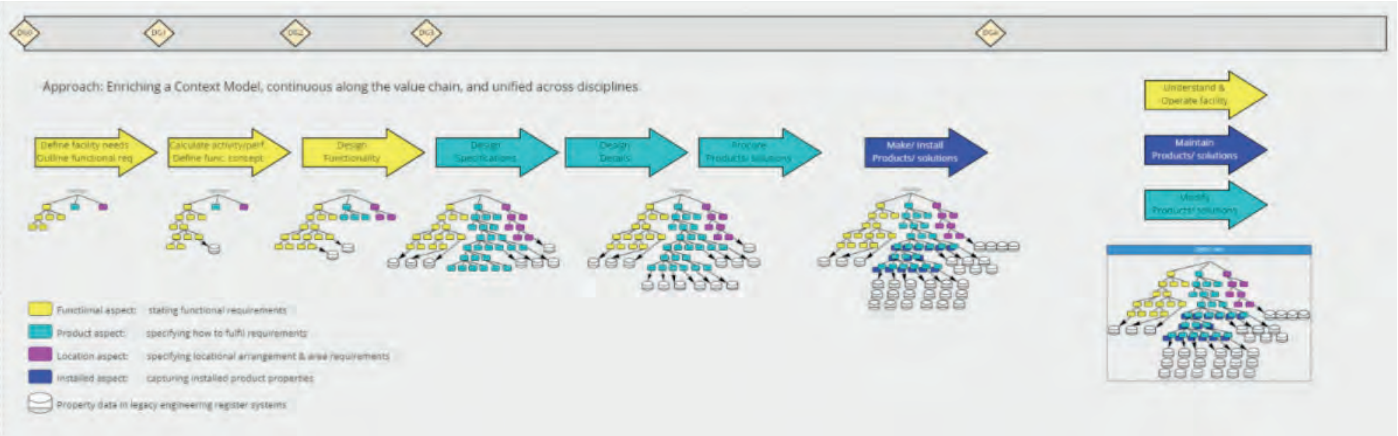


Figure 3: IMF models capture systems in context.

Adoption of the standards, listed as different layers in Table 1, presents several research challenges that SIRIUS researchers address.

First, most of these standards are not data standards. In order to serve the needs of data standards, the standards have to be significantly further detailed. This work must be based on solid and well understood principles that in many cases go beyond the principles on which the standards are based.

Second, the precision required for data standards makes it hard for domain expert to adopt them. It is important to develop languages, methods and tools that discipline experts in the industry can easily understand and use.

Third, the lack of inherent structure in RDF presents challenges for large RDF graphs. Further structure needs to be imposed. Our approach is based on OTTR templates.

Forth, the technology that IMF uses or develops is cutting edge and hence not mainstream. We develop a reference architecture that facilitates implementation of IMF. A key topic is management of identifiers.

Finally, we aim for a seamless integration of the technologies and standards in Figure 2 so that they form a layered group of languages. To achieve this, we want to show how one layer can be used to represent the layer above, while also admitting a higher degree of detail, in analogy to the way a technology stack operates.

A concept paper about IMF was produced in the READI Joint Industry project and can be downloaded from the web page of READI (https://readi-jip.org/wp-content/uploads/2021/03/Information-modelling-framework-V1.pdf). Expected result in 2022 is a comprehensive manual comprised of several parts:

- Modelling method and language
- Representation format
- Exchange protocols
- Reference architecture
- Management of change

Research publications and open-source software releases are also planned.

The SIRIUS contribution to the IMF project is a collaboration between SIRIUS researchers and SIRIUS partners Equinor, Aibel, Aker Solutions, Bouvet, Computas and Envester. The research work is organized under the Ontology Engineering program, while the industrial use case is organized in the Field development demonstration project.

Team members (from SIRIUS)

Arild Waaler

Erlend Fjøsna, Envester

Martin G. Skjæveland

Rustam Mehmandarov, Computas

> **IMF IS BASED ON** best practice from industry and state of the art technology and research, advancing and sharpening principles from several standards and industry initiatives and integrating them in a coherent and scalable way

Scalable Computing

The Scalable Computing (SC) research program is about making data access and processing faster to SIRIUS projects. This is achieved by building knowledge in High Performance Computing (HPC) and coupling this with scalable Cloud computing to support scalable big-data application processing. Specifically, we look at solutions for scalable and reconfigurable hardware, software design for parallel numerical simulations, and automatic cross-cloud application deployment and reconfiguration using hardware accelerators.

The target problems of HPC involve large scale computations that are beyond the capabilities of single laptop PCs or desktop computers. Moreover, close interactions often exist between the inherent components of these computations, thus the required hardware platforms for HPC are tightly coupled computer clusters, consisting of many powerful interconnected computers. The research topics of HPC encompass parallelization schemes, partitioning algorithms, communication overhead reduction strategies, software implementation and optimization techniques, use of heterogeneous clusters that consist of both conventional CPUs and cutting-edge hardware accelerators, in addition to adopting HPC for real-world applications.

The methodology on the HPC side is largely experimentation with different computing platforms built on technology from SIRIUS hardware partners and to evaluate the performance of these platforms for real applications of the SIRIUS application partners. The work will therefore involve experimental software design and hardware architectures for scalable computing ranging from accelerators to numerical methods. Stochastic combinatorial optimization is the methodology used for managing applications across different Cloud providers to allocate the application components where they give the best utility for the application owner, and to reconfigure in response to changing application execution context.

The Scalable Computing research have the following objectives:

Objective 1 Better and Flexible Execution Platforms:
The focus is on making advanced execution platforms available to all SIRIUS partners through open interfaces that can be used remotely allowing researchers without direct access to the computers to use and experiment with different hardware configurations for their applications.

Objective 2 Scalable Application Support:
On one side this will continue the support to open source for better numerical computations for reservoir simulations, and on the other side it will continue the development of Cross-Cloud and Multi-Cloud application management middleware.

Objective 3 SIRIUS Application Execution:
To evaluate the research delivered under the previous objectives, demanding real world applications from the SIRIUS partners may be tested using the software and the hardware available.

Activities and their contribution to the main objectives:

Objective 1:

- Work is ongoing to link UiO's Numascale computer with NREC to enable its use as Cloud HPC platform to demonstrate how HPC applications can benefit from using Cloud computing
- A new and flexible HPC cluster architecture based on PCIe has been established and is under testing.
- Selection has started for the hardware for a future Exascale experimental platform.

Objective 2:

- The research activities on numerical methods for reservoir simulations and the associated code optimizations have continued.
- The work on installing the Cloud management software Open Stack on the moved NUMAScale computer has started.

Håkon K. Stensland

Xing Cai

Geir Horn

- The MORPHEMIC project developing proactive and polymorphic Cross-Cloud application management has successfully passed its mid-term evaluation with a real-world application demonstration.

- Ongoing PhD project in utility optimising Cross-Cloud autonomic application management.

Objective 3:

- Interactions with Equinor on the reservoir simulations through two ongoing PhD projects.

- Identification of SIRIUS partner applications in need for scalability and the projects of Scalable Computing have been presented to the industrial partners on multiple occasions.

Cross-Cloud application management

The Cloud activities of Scalable Computing in SIRIUS was for the first three years focused on contribution to the MELODIC[1] Horizon 2020 project. MELODIC supports Cross-Cloud application management through the Application Programming Interface (API) offered by the various Cloud providers and it is thereby able to deploy Cloud computing instances of the application's components in the form of virtual machines (VMs), containers, and serverless functions. The result of the MELODIC project is a multi-Cloud application management platform. This is available as open source from the main European open-source community OW2[2], or as a supported and installed package on standard commercial terms. There are already several paying customers of MELODIC, and it currently has support for managing applications across all the big Cloud providers, Amazon Web Services (AWS[3]). Google Cloud[4], and the Azure Cloud[5] from Microsoft. Additionally, there is support for some smaller European Cloud providers, and the open-source Cloud infrastructure management platform OpenStack[6], which is used by most academic Cloud installations worldwide, among them the he Norwegian Research and Education Cloud (NREC[7]).

However, artificial intelligence (AI) applications may benefit from using specialized hardware when training the algorithms. These are accelerators like Graphical Processing Units (GPU), Field Programmable Gate Arrays (FPGA) or Tensor Processing Units (TPUs) tailored for TensorFlow[8] processing. Hence, an application component may come as a standard Central Processing Unit (CPU) artefact, which is currently being deployed by MELODIC, or as artefacts compiled for one or more of the hardware accelerators. Furthermore, the accelerated version will only be beneficial during the training of the AI components, and more costly to use in the Cloud than standard CPUs. It is therefore a need to switch between different component artefacts depending on the application's need, and the SIRIUS Scalable Computing team leads the effort in the Horizon 2020 project MORPHEMIC[9] to add this support.

The optimization of the application's Cloud deployment configuration in MELODIC is largely reactive as it is based on measured changes in the managed application's execution context. Acquiring Cloud resources may unfortunately take several minutes, and the execution context may therefore manage to change significantly before the reconfigured application is up and running on the new resources. MORPHEMIC aims at remedying this by *proactive adaptation* based on real time series prediction of the measurements of the running application's execution context and perform the optimization and reconfiguration early so that the Cloud resources will be available when they are needed by the application.

The target problems of HPC involve large scale computations that are beyond the capabilities of single laptop PCs or desktop computers

[1] https://h2020.melodic.cloud/ and https://melodic.cloud/
[2] https://gitlab.ow2.org/melodic
[3] https://aws.amazon.com/
[4] https://cloud.google.com/
[5] https://azure.microsoft.com/en-us/
[6] https://www.openstack.org/
[7] https://www.nrec.no/
[8] https://www.tensorflow.org/
[9] https://www.morphemic.cloud/
[10] https://opm-project.org
[11] https://en.wikipedia.org/wiki/NVM_Express

The MORPHEMIC project successfully passed the mid-term evaluation by European Commission in the autumn of 2021 demonstrating a real application running on the first release of the software platform. This is a major recognition given the Covid-19 restrictions in effect during the first half of the project. The SIRIUS researchers have successfully published four conference publications in 2021 related to MORPHEMIC, and work is ongoing on a further 6 publications expected for 2022. Work to integrate the HPC cluster from NumaScale into the NREC computing facilities allowing it to be used transparently by MELODIC and MORPHEMIC managed HPC applications started in 2020 will hopefully be completed early in 2022. The MORPHEMIC platform will be demonstrated at the SIRIUS General Assembly in the spring of 2022, and hopefully this will stimulate the industrial partners to test and evaluate advanced Cross-Cloud application management as it can provide significant performance improvements and cost savings for applications with variable resource need over longer execution times.

HPC support for reservoir simulation

This research topic aims to enable faster and more realistic reservoir simulations, which are the cornerstone in the workflow of oil reservoir management, thereby of vital importance to the entire oil and gas sector. The complexity and uncertainty of the subsurface geological properties require large-scale numerical computations that can only be treated with the technique of HPC. Besides simulating oil reservoirs, the same technique and software are also applicable to planning geological CO2 storage, which is becoming increasingly more important for reducing man-made climate changes. Specifically, the researchers in this topic investigate how to improve the various algorithmic and implemental aspects of such computations, so that reservoir simulations can more efficiently use modern computing hardware.

During 2021, SIRIUS researchers continued to collaborate actively with Equinor researchers in enhancing the performance and capabilities of the reservoir flow simulator inside the Open Porous Media[10] (OPM) framework. Several new numerical components were implemented and incorporated into the simulator, which has led to faster computing speed. Also, research was carried out to quantitively characterize the impact of these new components. With respect to the

application scenarios, the same flow simulator was also used to study geological CO2 storage. Moreover, a new research subject was started in 2021, with the overall aim of improving an entire simulation ensemble, instead of individual simulations. The rationale is that there exist similarities between the simulations inside an ensemble, so the numerical successes or failures from the completed simulations can help to improve the upcoming simulations. The current focus is on deriving a quantitative understanding of the level of accuracy in the simulated results. Two journal publications appeared in 2021, one on the numerical and implementational structure of OPM's flow simulator, the other on a quantitative analysis of the impact of various strategies related to the parallelization of the flow simulator.

Smart Scalable PCI Express

I/O resources like Non-Volatile Memory express[11] (NVMe), GPUs, FPGAs are today installed in many modern servers and computers. Today, these resources are only available to applications running on the same server without sub-optimal networks and software. The Smart Scalable PCI Express (PCIe) project's goal is to solve the network inefficiency and enable servers connected by PCIe Gen4 networks to access remote I/O resources and achieve the same performance as if the I/O devices were local.

In 2021, the project has used the eX3 infrastructure to help prototype an NVMe storage device driver using the Dolphin SmartIO API, which has been developed in conjunction with the Smart Scalable PCI Express project. Although the Device Lending component of SmartIO makes it possible to use existing device drivers, most device drivers are written in a way that assumes exclusive control over the device. Using Device Lending alone, a device may only be used by a single host at the time. To demonstrate software-enabled dis aggregation, we are implementing a distributed NVMe driver. As proof of concept, we have shown that a single NVMe device can be shared and operated by 30 cluster nodes simultaneously, without requiring single root input/output virtualization (SR-IOV) . This driver also demonstrates how multiple sharing aspects of our system may be combined by disaggregating (remote) GPU memory and enabling memory access optimizations.

Advanced HW accelerators for HPC

Compute Express Link (CXL) is one of the proposed next-generation high-speed interconnect buses designed to connect CPUs, memory, and accelerators in servers, data-centres, and computers. The CXL standard is also backed by all the large hardware vendors such as Intel, AMD, ARM, Nvidia, etc.

This project has started investigating the potential of the CXL bus, both as a next-generation replacement for PCI Express and the possibility of using it as a next-generation chip for NUMAScale to connect accelerators, memory, and CPUs in an HPC environment. Activities in this project in 2021 have been preliminary discussions about the possibility of the CXL standard for multi-host communication or vendor-specific extensions for adding the functionality in future versions. This work will continue in 2022.

Team Members

Geir Horn

Xing Cai

Tor Skeie

Andreas Thune

Håkon Kvale Stensland

Marta Różańska

Thomas Hansen

Atle Vesterkjær

Einar Rustad

Hugo Kohmann

Domain-Adapted Data Science

Vision: We develop hybrid approaches that exploit both knowledge in data and knowledge in ontologies

Within the artificial intelligence (AI) research community and beyond that community there is interest in developing *strong AI*, which means intelligent machines that are indistinguishable from the human mind or that go beyond human-level intelligence (superintelligence). However, despite the impressive progress in the field over the last decades, we still do not know how to achieve strong AI.

Basil Ell

As an example of human capability, imagine a child who has seen the usual animals that live on Norwegian farms, such as sheep and horses, but has never seen a giraffe, neither in person nor in pictures. If the child has sufficient language skills, then you can tell them, before going to a zoo, that a giraffe is an animal that looks like a horse but with a very long neck. Then, at the zoo, most likely the child will correctly identify a giraffe as a giraffe with ease.

This example shows that humans can combine what they have learned from experience (in our example, what they have seen before) with declarative statements (the description of the similarities and differences between giraffes and horses). Machines are not yet good at that.

In the AI community, there has been a long-standing frontier, labeled as the discussion about «symbolic vs. non-symbolic AI». In symbolic AI, information is structured in ontologies and deductions are made via reasoning. In sub-symbolic AI, information is obtained from data, and deductions are made via machine learning (ML). One hypothesis is that machines will become better at learning if we can combine these two types of information in the learning process. That is, combine information in terms of declarative statements and ontologies with information encoded in data made available through statistics and machine learning.

The focus of the SIRIUS domain adapted data science program is exactly to develop approaches that combine the use of structured knowledge with learning from data in the machine learning process. On one hand, this means that we try to bridge the traditional divide between symbolic and sub-symbolic learning, developing what we refer to as hybrid approaches. On the other hand, as it turns out, hybrid approaches yield improved machine learning results, and especially on «not so big-data». Combined with the fact that symbolically represented knowledge can often be very small and concise, this is a powerful tool that makes machine learning available on datasets that are otherwise too small, or otherwise unfit, for classical machine learning tasks. The overarching goal is a general methodology for how data science tasks can be enhanced through the combined use of symbolic and sub-symbolic knowledge.

Another intriguing feature of hybrid approaches is that the presence of symbolic knowledge in the machine learning process may lead to more explainable predictions. Within our research program, we also develop novel hybrid approaches that identify and exploit these capabilities.

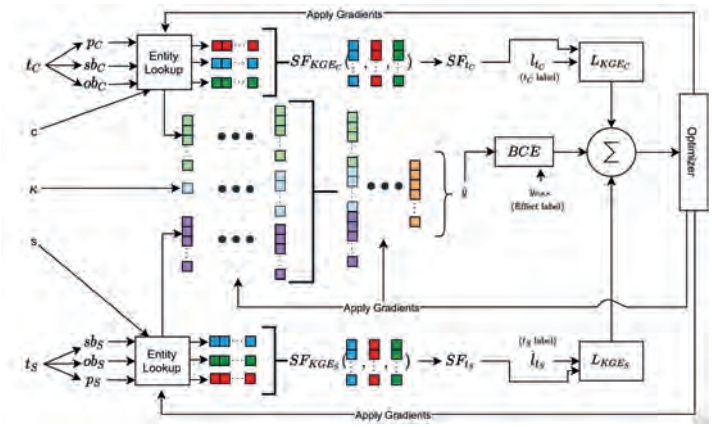
One term that we use to refer to a particular class of hybrid approaches is domain-adapted approaches. Very often in machine learning and data science tasks, data in the form of textual documents, images, or tables is processed, where making use of domain knowledge, for example in the form of an ontology, can improve the results.

In the context of our project *domain adapted data science pipeline*, we develop a catalog of situations related to domain adaptation and describe how these situations are related to each other. For example, consider the situation that an organization performs machine learning given a tabular dataset, is interested in improving the performance of the approach, has selected the most appropriate approach, and has tried out the usual performance improvement strategies such as hyperparameter optimization, but has so far not made use of domain knowledge such as taxonomies or ontologies. (Often it is a case that "a little semantics goes a long way", which means that the ontology needs not to be very extensive. Instead, a couple of statements can already make a significant difference.) One way to go forward can be to check whether openly available domain knowledge

exists (e. g., in the Wikidata knowledge graph) that can be used to contextualize the data so that the performance of the ML approach could be improved. Thus, we arrive at research questions such as: Given tabular data, how to find openly available domain knowledge in the Web of Data that could contextualize given tabular data? How to align tabular data to entities and properties in an RDF dataset (i. e., the domain knowledge represented in RDF format, which is the common format in the Semantic Web)? How to find out which parts of the external knowledge help most, so that when this information is improved it will have a significant impact on the performance? Which parts of the data have a negative impact on the performance and should thus be removed? How exactly can external knowledge be incorporated into an ML approach such as in a preprocessing step to improve the quality of existing training data? How can domain knowledge help in post-processing the output of the ML approach? How can the solution space exploration of the ML approach be guided by domain knowledge? How can the search space be pruned by making use of domain knowledge?

The situations and research questions that we collect can be ordered in two groups: those research questions that primarily address domain adaptation, such as, how to make use of (medical) taxonomies while training a (disease) classification model? Secondary research questions do not directly address domain adaptation but enable domain adaptation, e. g., embedding of knowledge graphs into vector spaces so that these can be processed by classical ML approaches, or they tackle the improvement of domain knowledge so that the performance of a domain adapted approach can be improved further, e. g., by anomaly detection in knowledge graphs and automatic knowledge graph completion.

The task of identifying and describing situations is both a top-down approach (guided by combinatorial exploration and brainstorming, zooming out and in) and a bottom-up approach (inspired by existing research). Not only does it allow us to structure existing approaches according to situations,



but it also lets us identify research questions that have so far not been addressed sufficiently. Our vision is that we can develop a methodology for domain adaptation, where an individual or organization browses the graph of situations to learn about how to realize domain adaptation or at least finds pointers to relevant sources such as publications.

Knowing what research could be done, which is an outcome of our activities, needs to be complemented by what is relevant for our SIRIUS partners, so that we can focus on those tasks or challenges especially relevant to our partners for mutual benefit. Therefore, we plan a stronger involvement of our partners in 2022 to prioritize our activities and to create a roadmap for research.

Beyond domain adaptation or improving domain knowledge that can then be used for domain adaptation, we actively work on a couple of other topics, within a group that also includes non-SIRIUS researchers at the UiO Department of Informatics (i. e., Anne-Marie George, Thomas Kleine Büning, and Meirav Segal). For example, we investigate in which way domain knowledge can assist reinforcement learning tasks. UiO researchers in the NRC project «Safe and Beneficial Artificial Intelligence» investigate active learning problems involving human and societal preferences such as learning preferences from interactions, collaborating effectively with humans, and making repeated decisions that are fair in the long term. Here, background knowledge, e. g., behavioral conventions, structure of preferences, features, and their relations of states of the world, might be able to improve the learning. Reversely, elicitation schemes might be able to gather and structure knowledge.

Furthermore, we carry out research related to explainable AI (XAI). The current approaches and technologies in XAI mostly focus on shedding light on the behavior of black-box machine learning models (like deep neural networks) by explaining their decisions to the users. However, the least work has been done towards employing the information provided by the explanations for enhancing the models concerning accuracy, fairness, and robustness in a systematic way. In SIRIUS, we have studied this research area and devised explanation-based frameworks for investigating the accuracy and robustness of black-box ML classification models [1].

Figure 1. Image taken from [2]. Fine tuning model architecture where each component is shown with inputs and outputs. tc and ts are knowledge graph triples relating to chemicals and species each with a score SF and loss l. c, s, and are the prediction input variables while ŷ is the predicted toxicity.

Team members

Basil Ell (program leader)	Ingrid Chieh Yu
Daniel Bakkelund	Martin Giese
Egor V. Kostylev	Jiaoyan Chen
Erik Bryhn Myklebust	Ole Magnus Holter
Ernesto Jimenez-Ruiz	Peyman Rasouli
Evgeny Kharlamov	Roxana Pop
Gong Cheng	Summaya Mumtaz

Within our research program, we have developed hybrid approaches, gained evidence for the benefits of hybrid approaches, and work towards developing novel hybrid approaches:

- Erik Bryhn Myklebust, Ernesto Jimenez-Ruiz, Jiaoyan Chen and colleagues have shown in the context of ecotoxicological effect prediction that the accuracy of predictions can be improved when domain knowledge is incorporated into the prediction model – see Figure 1 [2].

- Ole Magnus Holter and Basil Ell develop approaches that make use of domain knowledge in the context of semantic parsing of textual requirements. Their goal is to formally represent (parts of) the meaning of textual requirements, so that the meaning of requirements becomes more accessible to machines and the management of requirements can be improved [3].

- Egor V. Kostylev and colleagues study theoretical and practical connections between graph neural networks (GNNs), a modern structure-aware machine learning architecture, and classic logic-based knowledge representation formalisms. In particular, they designed a family of monotonic GNNs that allow for an efficient translation to Datalog logic-based language, and developed an efficient INDIGO system for knowledge graph completion [4, 5].

- In the context of a task relevant for the oil and gas industry, namely reservoir analogue identification, Summaya Mumtaz and Martin Giese have shown that a similarity measure based on the combination of domain knowledge (in the form of a taxonomy) with classical frequency-based features leads to significantly better results [6]. The disputation of her PhD thesis took place in November 2021.

- In the context of classification, based on a use case that is relevant to the oil and gas industry, namely that of excess inventory reduction, Daniel Bakkelund has developed theory and methodology for improved classification of interchangeable equipment, by integrating equipment structure awareness into classical methods for unsupervised machine learning [7]. Daniel will submit his PhD thesis in 2022.

- Jiaoyan Chen, Ernesto Jimenez-Ruiz, Ole Magnus Holter, Ian Horrocks and colleagues have developed an ontology embedding framework named OWL2Vec* that can embed symbolic knowledge in an OWL ontology into a vector space, so that the information can be consumed by machine learning algorithms. OWL2Vec* can be directly applied to ontology completion tasks such as subsumption prediction as well as to help address machine learning challenges, such as sample shortage, by injecting symbolic knowledge [8, 9].

- Actionable recourse (AR) techniques are a popular class of post-hoc interpretability approaches that help the users of ML models to obtain their desired decision from a machine learning model. Given an individual's preferences, an AR recommends feasible changes to their corresponding input that lead to the desired outcome by the model. To generate realistic ARs, it is important to capture and exploit the domain's information and the preferences of the users in the explanation process. Peyman Rasouli and Ingrid Chieh Yu are working on a model-agnostic framework that combines user/domain-level knowledge with model/data-level information to create plausible ARs that can guide individuals to obtain their desired decision from any ML classification and regression model in a simple and efficient manner.

- Current explainable artificial intelligence (XAI) techniques only rely on the observational data to analyze and explain the behavior of machine learning models. To increase the comprehensibility and faithfulness of explanations of ML models, hence, it is essential to exploit domain knowledge that bridges between the models and human concepts. Peyman Rasouli and Ingrid Chieh Yu aim to integrate domain knowledge (in the form of knowledge graphs and taxonomies) with structured/tabular data to provide more feasible, comprehensible, and faithful explanations.

- Gong Cheng, Evgeny Kharlamov investigated keyword-based exploration of knowledge graphs [10,11] and proposed a novel method to generate smart snippets or summaries of large-scale knowledge graphs. Baifan Zhou, Evgeny Kharlamov and colleagues from SIRIUS showed how to facilitate development

of ML models using semantic technologies [12]. Then, they investigated several practical aspects of knowledge graph management in connection to analytics and machine learning motivated by applications from Industry 4.0 [13,14]. That is, they showed how to scale usability of ML analytics and reshape industrial knowledge graphs. Moreover, Baifan and Evgeny consolidated a number of research directions into an advanced SIndAIS4 project (https://sirius-labs.no/sindais4-scaling-industrial-ai-with-semantics/) of SIRIUS that aims at Scaling Industrial AI with Semantics in four directions: human, data, methods, and applications. Within this project and together with Ahmet Soylu they selected several Bosch-funded interns – students of Ahmet – thus strengthening the Bosch-SIRIUS collaboration and disseminating it in two large Norwegian universities: NTNU and OsloMet.

- Basil Ell develops approaches to align symbolic data (i. e., ontologies) with sub-symbolic data (e. g., texts or tables). The alignment enables labeled training data to be generated via distant supervision for approaches such as information extraction (IE) for ontology population or natural language generation (NLG). Having symbolic and sub-symbolic data aligned means obtaining hybrid data that can be processed by hybrid approaches. **He received a best paper award at LDK 2021 – 3rd Conference on Language, Data and Knowledge**, for his work on mining association rules that help to bridge between text and data [15] – see Figure 2. Furthermore, he develops statistical approaches that are applied to symbolic data (KGs) for the purposes of identifying regularities and anomalies, for the prediction of missing facts, for the evaluation of the structural plausibility of facts, for bridging between structured and unstructured data (as in IE, question answering, NLG), and the structural classification of regions within graphs (which is similar to sequence labeling, but on graphs).

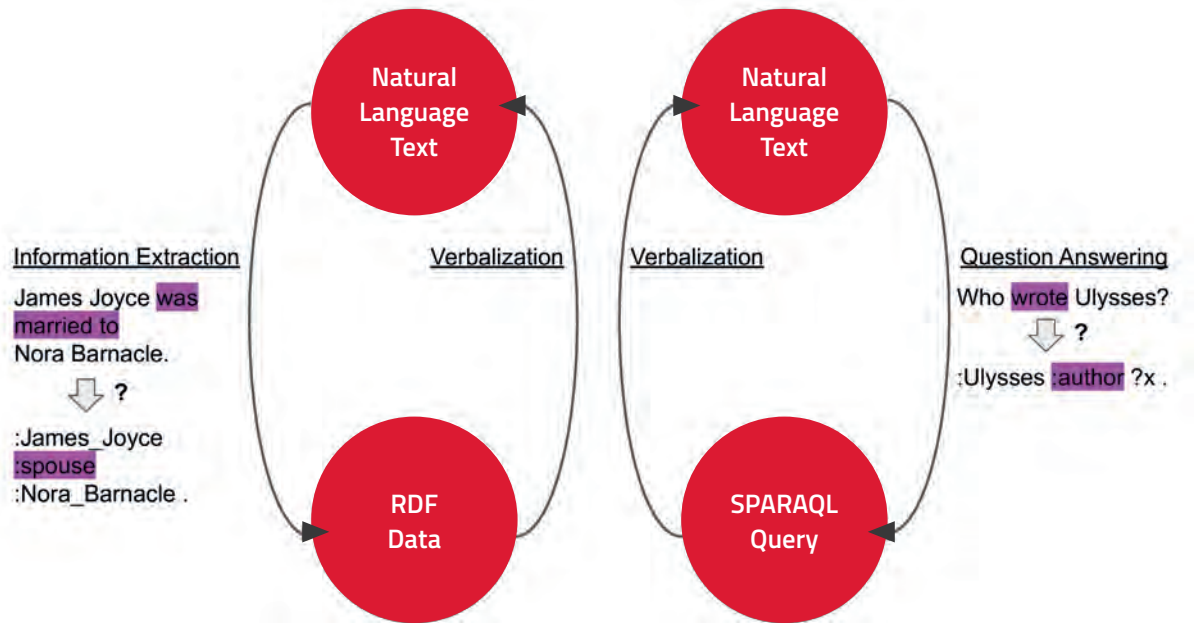


Figure 2. Image from [15]. There is a non-trivial lexical gap between expressions in natural language and terms in a knowledge graph, that needs to be bridged for a couple of tasks such as Information Extraction, Question Answering, and Verbalization (of RDF data or SPARQL queries).

In 2021, we collaborated with Bosch Center for Artificial Intelligence, DNV, IBM Research, Samsung Research UK, TechnipFMC, The Alan Turing Institute, University of Lisbon, University of Malaga, and University of Oxford and we organized a couple of events:

> SemTab challenge: https://www.cs.ox.ac.uk/isg/challenges/sem-tab/
> Ontology Matching workshop: http://om2021.ontologymatching.org/
> OAEI evaluation campaign: http://oaei.ontologymatching.org/2021/
> NeSy workshop: https://sites.google.com/view/nesy20/

Highlights of 2021

- We organized 2 workshops (Ontology Matching workshop, NeSy workshop)
- We organised 2 challenges (OAEI and SemTab)
- We published more than 20 papers
- We won a best paper award at the LDK 2021 conference
- Three non-SIRIUS researchers joined our program
- Summaya Mumtaz completed her PhD
- We are organizing the Journal of Web Semantics Special Issue on «Automating Knowledge Graph Construction»
- Ernesto Jimenez-Ruiz gave a keynote at the Description Logics workshop and is co-organizing the Interest Group on Knowledge Graphs at The Alan Turing Institute

[1] Peyman Rasouli and Ingrid Chieh Yu: «Explainable Debugger for Black-box Machine Learning Models.» 2021 International Joint Conference on Neural Networks (IJCNN), IEEE, pp. 1-10, 2021. [https://ieeexplore.ieee.org/document/9533944]

[2] Erik B. Myklebust, Ernesto Jiménez-Ruiz, Jiaoyan Chen, Raoul Wolf, Knut Erik Tollefsen: «Prediction of Adverse Biological Effects of Chemicals Using Knowledge Graph Embeddings.» Semantic Web journal (2021). [http://www.semantic-web-journal.net/system/files/swj2804.pdf]

[3] Magnus Holter and Basil Ell: «Towards Scope Detection in Textual Requirements.» 3rd Conference on Language, Data and Knowledge (LDK 2021). [https://drops.dagstuhl.de/opus/volltexte/2021/14567/pdf/OASIcs-LDK-2021-31.pdf]

[4] David Jaime Tena Cucala, Bernardo Cuenca Grau, Egor V. Kostylev, Boris Motik: «Explainable GNN-Based Models over Knowledge Graphs.» International Conference on Learning Representations (ICLR 2022). [https://openreview.net/pdf?id=CrCvGNHAIrz]

[5] Shuwen Liu, Bernardo Cuenca Grau, Ian Horrocks, Egor V. Kostylev: «INDIGO: GNN-based inductive knowledge graph completion using pair-wise encoding.» The 34th Annual Conference on Advances in Neural Information Processing (NeurIPS 2021). [https://proceedings.neurips.cc/paper/2021/file/0fd600c953cde8121262e322ef09f70e-Paper.pdf]

[6] Summaya Mumtaz and Martin Giese: «Frequency-Based vs. Knowledge-Based Similarity Measures for Categorical Data.» AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering (1). 2020. [http://ceur-ws.org/Vol-2600/paper16.pdf]

[7] Daniel Bakkelund: «Order preserving hierarchical agglomerative clustering.» Machine Learning (2021). [https://link.springer.com/content/pdf/10.1007/s10994-021-06125-0.pdf]

[8] Jiaoyan Chen, Pan Hu, Ernesto Jiménez-Ruiz, Ole Magnus Holter, Denvar Antonyrajah, Ian Horrocks: «OWL2Vec*: embedding of OWL ontologies.» Machine Learning 110.7 (2001): 1813-1845. [https://link.springer.com/content/pdf/10.1007/s10994-021-05997-6.pdf]

[9] Jiaoyan Chen, Ernesto Jiménez-Ruiz, Ian Horrocks, Denvar Antonyrajah, Ali Hadian, Jaehun Lee: «Augmenting Ontology Alignment by Semantic Embedding and Distant Supervision.» Extended Semantic Web Conference (2021): 392-408. [https://openaccess.city.ac.uk/id/eprint/25810/1/]

[10] Yuxuan Shi, Gong Cheng, Trung-Kien Tran, Evgeny Kharlamov, Yulin Shen: «Efficient Computation of Semantically Cohesive Subgraphs for Keyword-Based Knowledge Graph Exploration.» WWW 2021: 1410-1421. [https://dl.acm.org/doi/pdf/10.1145/3442381.3449900]

[11] Yuxuan Shi, Gong Cheng, Trung-Kien Tran, Jie Tang, Evgeny Kharlamov: «Keyword-Based Knowledge Graph Exploration Based on Quadratic Group Steiner Trees.» IJCAI 2021: 1555-1562. [https://www.ijcai.org/proceedings/2021/0215.pdf]

[12] Baifan Zhou, Yulia Svetashova, Andre Gusmao, Ahmet Soylu, Gong Cheng, Ralf Mikut, Arild Waaler, Evgeny Kharlamov: «SemML: Facilitating development of ML models for condition monitoring with semantics.» Journal of Web Semantics 71: 100664 (2021). [https://reader.elsevier.com/reader/sd/pii/S1570826821000391]

[13] Baifan Zhou, Dongzhuoran Zhou, Jieying Chen, Yulia Svetashova, Gong Cheng, Evgeny Kharlamov: «Scaling Usability of ML Analytics with Knowledge Graphs: Exemplified with A Bosch Welding Case.» IJCKG 2021: 54-63. [https://dl.acm.org/doi/pdf/10.1145/3502223.3502230]

[14] Dongzhuoran Zhou, Baifan Zhou, Jieying Chen, Gong Cheng, Egor V. Kostylev, Evgeny Kharlamov: «Towards Ontology Reshaping for KG Generation with User-in-the-Loop: Applied to Bosch Welding.» IJCKG 2021: 145-150. [https://dl.acm.org/doi/pdf/10.1145/3502223.3502243]

[15] Basil Ell, Mohammad Fazleh Elahi, and Philipp Cimiano: «Bridging the Gap Between Ontology and Lexicon via Class-Specific Association Rules Mined from a Loosely-Parallel Text-Data Corpus.» 3rd Conference on Language, Data and Knowledge (LDK 2021). [https://drops.dagstuhl.de/opus/volltexte/2021/14569/pdf/OASIcs-LDK-2021-33.pdf]

[16] Xiaxia Wang, Gong Cheng, Tengteng Lin, Jing Xu, Jeff Z. Pan, Evgeny Kharlamov, Yuzhong Qu: «PCSG: Pattern-Coverage Snippet Generation for RDF Datasets.» ISWC 2021: 3-2.

[17] Roberto Avogadro, Marco Cremaschi, Ernesto Jiménez-Ruiz, Anisa Rula: «A Framework for Quality Assessment of Semantic Annotations of Tabular Data.» International Semantic Web Conference (2021): 528-545. [https://openaccess.city.ac.uk/id/eprint/26426/1/]

[18] Yuan He, Jiaoyan Chen, Denvar Antonyrajah, Ian Horrocks: «BERTMap: A BERT-based Ontology Alignment System.» AAAI 2022. [https://arxiv.org/abs/2112.02682]

[19] Jiaoyan Chen, Yuan He, Ernesto Jimenez-Ruiz, Hang Dong, Ian Horrocks: «Contextual Semantic Embeddings for Ontology Subsumption Prediction.» Submitted to KR 2022. [https://arxiv.org/abs/2202.09791]

[20] Peyman Rasouli and Ingrid Chieh Yu: «Analyzing and Improving the Robustness of Tabular Classiers using Counterfactual Explanations.» 20th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1286-1293, 2021. [https://ieeexplore.ieee.org/document/9679972]

Semantic Integration

Vision: The Semantic Integration research program will continue improving the software systems

Introduction

The Semantic Integration research program designs and develops scalable infrastructure that supports semantic integration using large ontologies (with many thousands of classes) and massive data sets (many billions of tuples) into Knowledge Graphs. It will demonstrate the efficacy of these tools through deployment in the demonstration projects. Specifically, we work with ontology reasoners capable of supporting the development of large-scale ontologies and semantic data stores which answer realistic ontology-based queries over massive data sets.

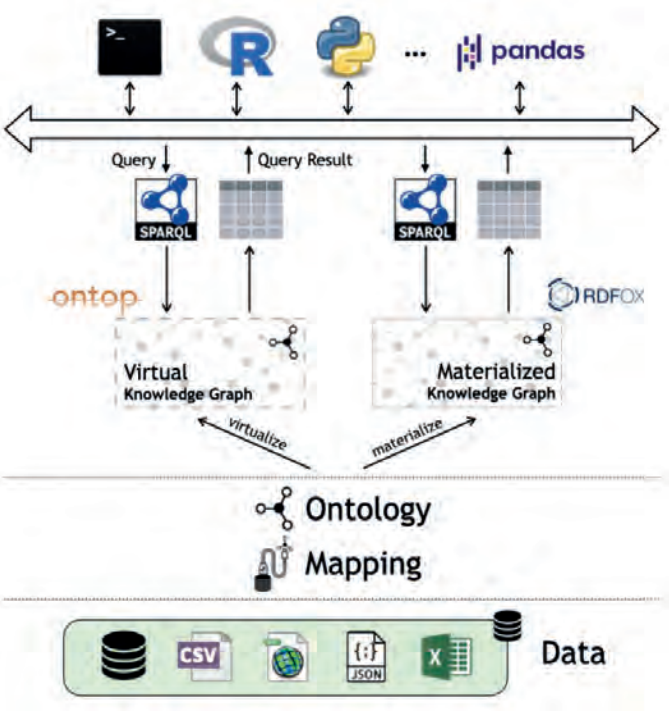


Figure 1 The Framework of Semantic Integration (OBDA)

Figure 1 shows the conceptual framework of Semantic Integration, also known as ontology-based data access (OBDA) [1][2]. At the bottom of this figure, in this project, we are working on integrating different kinds of data sources, which are typically legacy systems and might come in different forms, such as relational databases (DBs), or as files in various formats (such as CSV, XML, JSON, or proprietary formats). The objective is to semantically integrate these data sources into a Knowledge Graph consisting of a set of data assertions that use the vocabulary of classes and properties provided in the ontology. The data assertions in the KG are often obtained by mapping the data stored in various data sources to the terms of the ontology vocabulary. Intuitively, a mapping can be thought of as a collection of queries that are used to construct RDF triples using the classes and properties of the ontology by retrieving the necessary data from the sources.

Semantic Integration can be realized in two flavors:

- **Virtual Knowledge Graphs (VKGs).** In the virtual approach, the triples are not materialized in a separate triple store, but their presence in the KG is only virtual. Systems operating on VKGs are able to retrieve the data directly from the data sources only when it is required for a particular user query. In fact, query processing is delegated to the data sources. This is achieved by unfolding the mappings, thus translating user queries into queries over the data sources, whilst taking into account also the ontology background knowledge through a so-called query rewriting step. The advantage of VKGs is that information is always fresh and up-to-date with the data sources. For example, Ontop is a state-of-the-art Virtual Knowledge Graph system. Ontop implements the VKG technology, thus lowering the cost of typical data integration projects. In this way, companies and organizations can readily exploit the value of their data assets and make such data available for Business Intelligence and applications based on Machine Learning.

- **Materialized Knowledge Graphs (MKGs).** Despite the advantages of the virtual approach, it is sometimes convenient to actually materialize the triples. In such a case, we talk about Materialized Knowledge Graphs (MKGs). The main advantage of MKGs over VKGs is that usually a more predictable performance in query answering can be achieved, especially in those situations where

mappings are very complex and thus the unfolding of the virtual approach would give rise to complex queries over the data sources. This comes at the cost of maintaining a potentially very large MKG. For example, RDFox is a powerful system for Materialized Knowledge Graph. RDFox is a high-performance in-memory knowledge graph and semantic reasoner, optimised for speed and efficiency. Designed from the ground up with reasoning in mind, it outperforms other graph databases while also providing benefits and insights that cannot be achieved by alternatives.

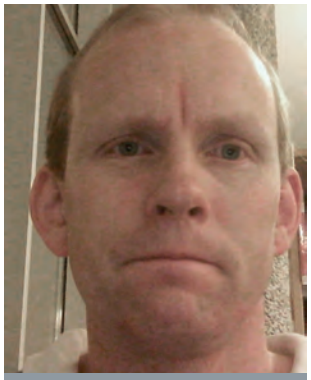
The top part of Figure 1 are various kinds of possible user interfaces for accessing the information in the data sources through the VKG. Indeed, Semantic Integration, through an explicit and non-ambiguous representation of the semantics of the data as KGs, promotes:

- interoperability among different scientific data platforms;
- knowledge discovery and data mining practices by exposing a conceptually sound view over a multiplicity of distinct and possibly non-interoperable data sources;
- the enrichment of the information originally present in the data sources, through the application of reasoning techniques that combine domain knowledge and data assertions.

Highlights of 2021

In 2021, the Semantic Integration research program has made significant improvements in the Knowledge Graph reasoning systems Ontop and RDFox in terms of new features and optimisation, and has introduced the Ontopic Studio, a novel designer for creating Virtual Knowledge Graph systems. We have continued our work on reasoning techniques in query answering, KG construction and curation. We have built a SIRIUS OBDA subsurface pilot and have successfully deployed in a few novel use cases, e.g., South Tyrol OpenDataHub, Dow Jones, and Festo. Finally, we have disseminated our work at major events such as KGC 2021 and SWAT4HCLS 2022.

Guohui Xiao



Dag Hovland

Team

Semantic Integration research program collaborates with the Oxford University and two SMEs Oxford Semantic Technologies and Ontopic. At **Oxford University**, the Knowledge Representation and Reasoning Group conducts research in knowledge representation formalisms, ontology languages, the design, implementation and optimisation of reasoning systems, and applications in areas such as e-Science and the Semantic Web. **Oxford Semantic Technologies (OST)** is an early-stage spin-out from the University of Oxford backed by leading investors including Samsung Ventures and Oxford Science Enterprises. OST enables leading organisations to turn their data into knowledge through our graph database and reasoning engine, RDFox. **ONTOPIC s.r.l.** is a young innovative SME, which was founded in April 2019, as the first spin-off of the Free University of Bozen-Bolzano. Ontopic provides consultancy services and develops software solutions for data and information integration. The key expertise of Ontopic is techniques and technologies for data access, management, and integration based on Virtual Knowledge Graphs (VKGs).

Team members (from SIRIUS)

Guohui Xiao

Dag Hovland

Stefano Germano

Software Tools

In 2021, the group has made significant progress in improving the engines Ontop and RDFox, and developed a new system Ontopic Studio for designing VKGs.

Ontop (https://ontop-vkg.org/) is the state-of-the-art open-source VKG engine. The Ontop project is hosted by the Free University of Bolzano, and is also commercially supported by the company Ontopic, which became SIRIUS partner in 2020. SIRIUS is continuing to contribute to the development of this tool. In 2021, there were two major releases of Ontop (v4.1 and v4.2) with an extensive list of new features and optimization techniques (https://ontop-vkg.org/guide/releases.html). Notably, Ontop v4.1 added supports for GeoSPARQL and Dremio, features a new query logger and the specification of custom WebAPIs, and embeds new optimizations, memory consumption decreased. Ontop v4.2 introduces support for Apache Spark, Ontop views, TBox information, time functions, robust data type extraction and, as usual, novel optimizations and many bug fixes. Ontop has been widely used in industrial and academic projects. According to SourceForge, it has been downloaded for a total of 62K times since 2015 and 16K times in 2021.

Recently, the Community Leader badge (for open-source projects that have reached the milestone of 50,000 total downloads) was awarded to Ontop by SourceForge.

RDFox (https://www.oxfordsemantic.tech/product) is a high-performance in-memory knowledge graph and semantic reasoner, optimised for speed and efficiency, initially developed by University of Oxford, and currently by Oxford Semantic Technologies. In 2021, along with performance enhancements, the latest version of RDFox brings a variety of new capabilities and updates. RDFox now supports SHACL, property paths and we include a release dedicated to the new Apple M1 ARM chip. These developments mean RDFox now has complete support for SPARQL 1.1. Moreover, the console has new slick features and the support for OWL and Solr has been updated for increased usability.

Ontopic Studio (https://ontopic.ai/en/ontopic-studio/) is Ontopic's flagship product developed starting from 2019 and first released in 2021. It allows users to easily create Virtual Knowledge Graphs from a multitude of sources, such as data lakes, databases, Excel files, and to quickly adapt to changes in the data landscape. It features a carefully

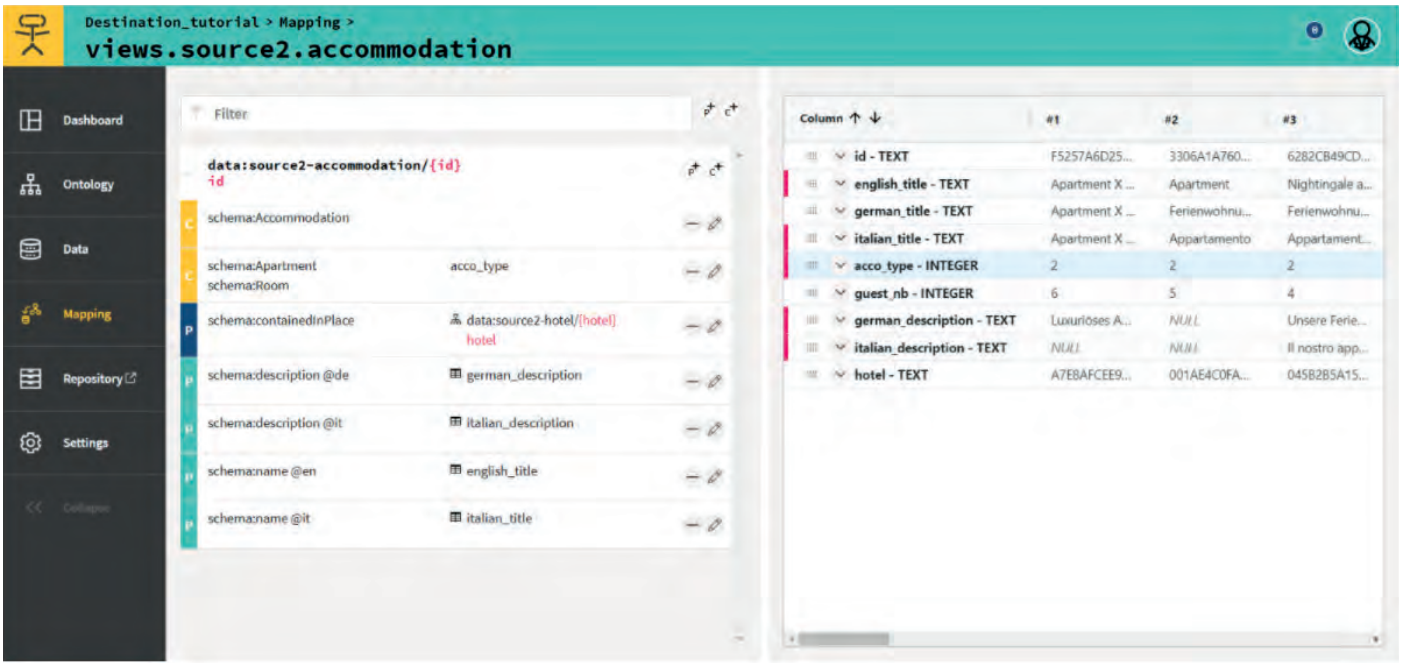


Figure 2. Mapping Editing in Ontopic Studio

designed user interface, grown out of a long experience in data integration projects is completely dedicated to the technology of Virtual Knowledge Graphs. Figure 2 is a screenshot of Ontopic Studio for editing Mapping. There are many innovative new functionalities that support the data architect for the whole lifecycle of the Knowledge Graph. Ontopic Studio internally relies on Ontop, the most advanced Open-Source VKG engine.

New Reasoning Techniques

CQ answering [5][6][7]. We continued our work on optimising CQ answering over unrestricted OWL 2 ontologies. This effort resulted in the development of two tools:

- RSAComb, an optimised implementation of a combined approach technique for conjunctive query answering over RSA ontologies.
- The next iteration in the development of PAGOdA, now integrating RSAComb to improve the computation of query bounds.

We extended our research on ontology approximations; we proposed a novel technique to compute an RSA restriction of an unrestricted ontology that maintains completeness w.r.t. CQ answering. The most recent evaluation of the system provides a fair and extensive analysis of the tools' performance and capabilities.

KG construction, curation, etc. [8][9][10][11]. We mainly worked on ontology alignment by combining modern deep learning techniques such as pre-trained language models and distant supervision with traditional ontology alignment systems such as LogMap. Specifically, we developed LogMap-ML which is an extension for LogMap with higher results especially w.r.t. Recall. It uses LogMap anchor mappings filtered by disjointness-based rules to train a mapping prediction model which uses two classes' embeddings or their paths' embeddings as the input. We also developed a new system named BERTMap which totally gives up the lexical mapping part of LogMap but uses the pre-trained language model BERT to fully utilize the textual information. BERTMap can achieve state-of-the-art performance on the tasks of the OAEI LargeBio track.

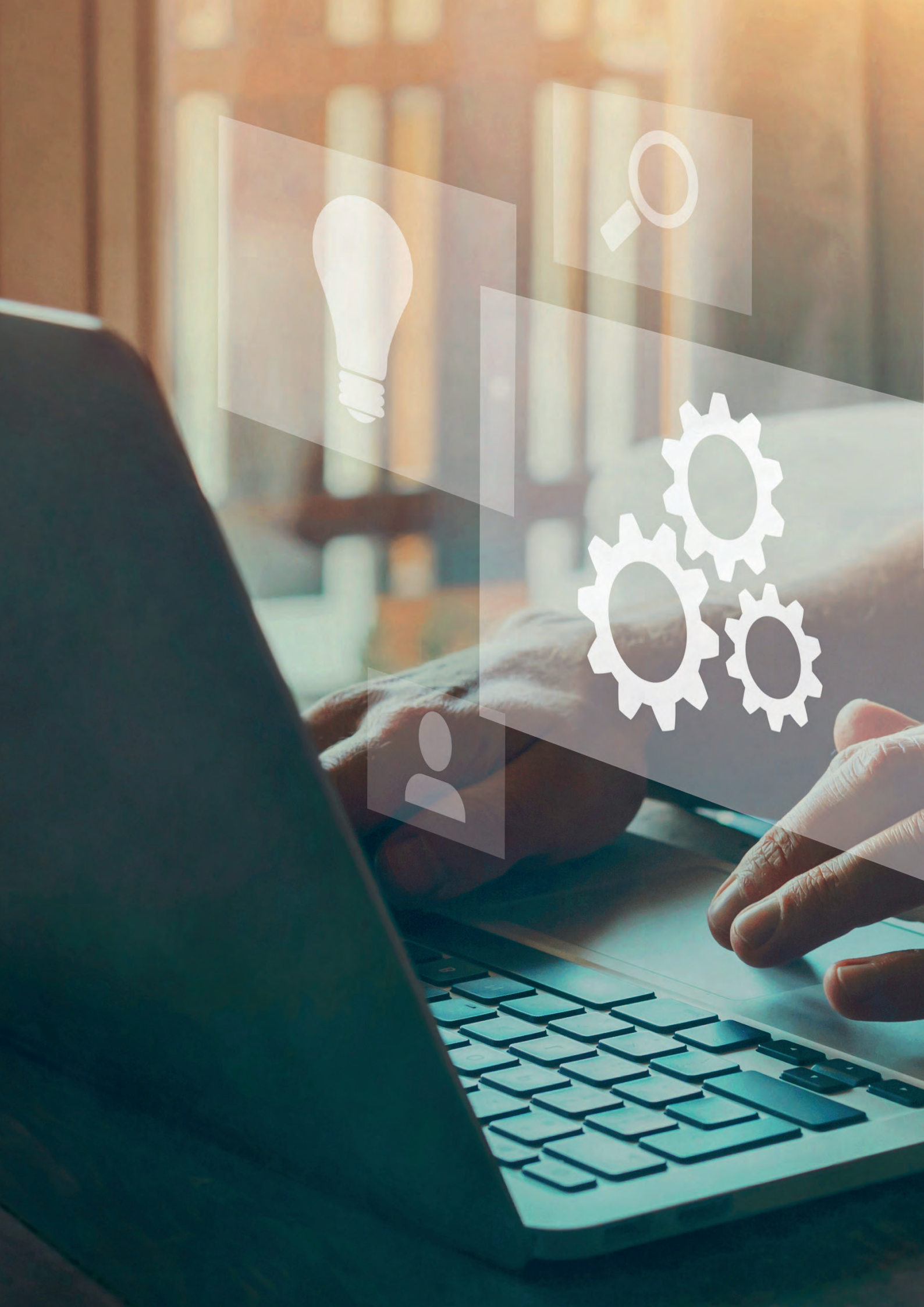
We also continued the works of knowledge graph correction and OWL ontology embeddings. Our previous Web conference paper on erroneous fact correction is extended with the case of correcting mapping assertions, and the extension is accepted by Semantic Web Journal. Our Word2Vec-based ontology embedding method OWL2Vec* is finally published in Machine Learning Journal.

Use Cases

Our approach has been successfully deployed in a few novel use cases. We have developed the SIRIUS OBDA subsurface pilot, and applied it to the South Tyrolean OpenDataHub, LinkedGeoData, Dow Jones, and Festo.

The SIRIUS OBDA subsurface pilot project is addressing these shortcomings and aim to significantly broaden the applicability of the approach for use in subsurface projects. In the past, the Optique project was focused on developing OBDA on a relational database at Equinor, which is no longer an active database. Further, that database contains proprietary data, and access to that is now restricted. This resulted in a significant setback for researchers to continue further extermination on extending OBDA capabilities. In 2021 we established a large in-house relational database from the publicly available G&G datasets (mainly by processing the Volve dataset https://data.equinor.com & NPD FactPages https://factpages.npd.no). This database is now being utilized in various other internal and external research projects where G&G data is being used for experimentation, e.g., DigiWell at USN is also using this database for research purposes. We have adapted mapping and Ontology from Optique, and carried out Integration & testing. SIRUS OBDA Subsurface V1.0 was demonstrated at the SIRIUS General Assembly in November 2021.

South Tyrol OpenDataHub Knowledge Graph (https://sparql.opendatahub.bz.it/) is a joint project between NOI Techpark and Ontopic for publishing South Tyrolean tourism data as a Knowledge Graph. LinkedGeoData (http://linkedgeodata.org/) is an effort to add a spatial dimension to the Semantic Web. LinkedGeoData uses the information collected by the OpenStreetMap project and makes it available as an RDF knowledge base according to the Linked Data principles [3]. The Festo case study (https://uploads-ssl.webflow.com/5ed7f18d11a068aa460ce2e9/5f5252796dd12f613510c1eb_Festo%20Case%20Study.pdf) shows how Festo was able to completely transform the related internal data processes to reduce the time to provide satisfactory specifications from hours to seconds using RDFox. OST's partner Derivo integrated RDFox in Festo's Semantic Platform. Finally, for Dow Jones, RDFox enhances the scope and capability of various products, from The Wall Street Journal to competitor listings in the S&P 500 index (https://www.oxfordsemantic.tech/blog/dow-jones-enhances-product-line-with-semantic-innovation).



Tutorials

3 May 2021 «Integrating Data through Virtual Knowledge Graphs with Ontop». Diego Calvanese, Benjamin Cogrel, Guohui Xiao. Knowledge Graph Conference 2021, Virtual.
10 Jan, 2022 «FHIR RDF Data Transformation and Validation Framework and Clinical Knowledge Graphs: Towards Explainable AI in Healthcare». Harold Solbrig, Guohui Xiao, Eric Prud'Hommeaux. Half day tutorial at 13th International SWAT4HCLS Conference (Semantic Web Applications and Tools for Healthcare and Life Sciences). Leiden, The Netherlands.

Vision

In 2022, the Semantic Integration research program will continue improving the software systems (Ontop, RDFox, Ontopic Studio) involved in the project, develop novel reasoning techniques, and deploy our approach in more significant use cases.

[1] *Diego Calvanese, Linfang Ding, Alessandro Mosca, and Guohui Xiao.*
Realizing ontology-based reusable interfaces for data access via virtual knowledge graphs.
In Proceedings of the 14th Biannual Conference of the Italian SIGCHI Chapter, CHItaly '21, Bozen-Bolzano, Italy, and online (www), July 11-13, 2021, pages 35:1–35:5. ACM, 2021.

[2] *Diego Calvanese, Davide Lanti, Tarcisio Mendes De Farias, Alessandro Mosca, and Guohui Xiao.*
Accessing scientific data through knowledge graphs with Ontop.
Patterns, 2(10):100346, 2021.

[3] *Linfang Ding, Guohui Xiao, Albulen Pano, Claus Stadler, and Diego Calvanese.*
Towards the next generation of the LinkedGeoData project using virtual knowledge graphs. Journal of Web Semantics, 2021.

[4] *Guohui Xiao and Julien Corman.*
Ontology-mediated SPARQL query answering over knowledge graphs.
J. of Big Data Research, 23, 2021.

[5] *Federico Igne, Stefano Germano, and Ian Horrocks.*
Computing CQ Lower-Bounds over OWL 2 Through Approximation to RSA.
20th International Semantic Web Conference - Research Track, ISWC 2021, Virtual Event, October 24-28, 2021.

[6] *Federico Igne, Stefano Germano, and Ian Horrocks.*
Computing CQ Lower-Bounds over OWL 2 Through Approximation to RSA - Extended Abstract.
20th International Semantic Web Conference - Posters, Demos and Industry Tracks, ISWC 2021, Virtual Event, October 24-28, 2021.

[7] *Federico Igne, Stefano Germano, and Ian Horrocks.*
RSAComb: Combined Approach for CQ Answering in RSA.
34th International Workshop on Description Logics, DL 2021, Bratislava, Slovakia, September 19-22, 2021.

[8] *Yuan He, Jiaoyan Chen, Denvar Antonyrajah and Ian Horrocks.*
BERTMap: A BERT-based Ontology Alignment System.
AAAI 2022.

[9] *Jiaoyan Chen, Ernesto Jimenez-Ruiz, Ian Horrocks, Denvar Antonyrajah, Ali Hadian, Jaehun Lee.*
Augmenting Ontology Alignment by Semantic Embedding and Distant Supervision.
ESWC 2021.

[10] *Jiaoyan Chen, Ernesto Jiménez-Ruiz, Ian Horrocks, Xi Chen, Erik Bryhn Myklebust.*
An assertion and alignment correction framework for large scale knowledge bases.
Semantic Web Journal. https://content.iospress.com/articles/semantic-web/sw210448

[11] *Jiaoyan Chen, Pan Hu, Ernesto Jimenez-Ruiz, Ole Magnus Holter, Denvar Antonyrajah, Ian Horrocks.*
Owl2vec: Embedding of owl ontologies.*
Machine Learning 110.7 (2021): 1813-1845.

Industrial Digital Transformation

Our vision is to produce knowledge in close collaboration with industrial stakeholders

Thomas Østerlie

Increased availability and use of digital data hold the potential to transform how organizations work and collaborate as well as transforming their products and services. Yet, digital transformation is not easy. Both experience and research show that digital transformation frequently fall short of meeting expectations.

There are many reasons for this, but an often-overlooked aspect is the need to cultivate the organizational preconditions necessary for realizing digital data and technologies' transformative potential. What these preconditions entail in practice varies greatly, and remains object of much scientific scrutiny across disciplines concerned with current shifts towards data-centric and -driven forms of work and organizing.

Our vision is to produce knowledge in close collaboration with industrial stakeholders, that

(1) inform collaborating companies in planning and organizing for digital transformation, and

(2) advance scientific knowledge.

We pursue this vision through active engagement with digital transformation initiatives, where we apply our methods and theoretical frameworks to address important and difficult challenges and issues. Our focus is the relationship between social and technical factors during development, implementation, and use of digital data and technologies.

Our scientific objective is to develop empirically grounded insights and theory on digital transformation in general, with particular emphasis on transitions towards data-centric and -driven forms of work and organizing. We pursue this across multiple levels of scale from micro-level studies of data-centric work practices, via adoption and implementation of tools and data at the company level, to large-scale technological change at the industry level.

We have pursued our vision and scientific objective through a series of studies and projects throughout Sirius' lifecycle. During 2021, we pursued our research through two ongoing projects. These projects demonstrate the breadth of our research, from innovation to activities oriented towards basic research.

Digitalization of LCI exchange

This project addresses the problem complex associated with digitalization of LCI exchange. It is conducted in close collaboration with key industry stakeholders through our participation in the NORSOK Z-TI expert group.

Background. Structured and machine-readable life-cycle information (LCI) is a pre-requisite for the transition to data-centric and -driven approaches to design, construction, operations, and maintenance of oil and gas installations. Yet, much of it is today stored and exchanged between companies' internal legacy systems in form of unstructured data such as digital documents and images. What exists of structure information exchange is between proprietary systems on formats governed by single providers rather than as standards at the industry level.

While there is shared understanding of the need for structured and machine-readable LCI standards throughout the oil and gas industry, there is little to no coherence across companies to jointly solve central challenges associated with this. The result is that digitalization of LCI exchange is driven through disparate activities within individual enterprises, through larger capital expense projects or across various joint industry initiatives.

Results. This joint work has resulted in **two innovations** so far: (1) a novel standardization strategy and (2) a mechanism for industry-wide coordination of standardization. These two innovations are part of the NORSOK national strategy for digitalization of LCI exchange, to be approved during spring 2022.

Future work will focus on evaluating and improving upon the innovations, as well as further development and detailing of key principles underpinning the national digitalization strategy. There is also ongoing work to introduce

these innovations as basis for international standardization initiatives.

Relationship with other activities. Our work in NORSOK Z-TI builds and extends upon engagement in the READI JIP, predominantly through Mina Hagshenas PhD project. We also have an ongoing PhD project on digital tool support for early-stage design of subsurface facilities. This PhD is funded through the BRU21 program.

Further reading

Haghshenas, M. and T. Østerlie (2020). «Coordinating innovation in digital infrastructure: The case of transforming offshore project delivery», in Agrifoglio, R., R. Lamboglia, D. Mancini, and F. Ricciardi (Eds.), *Digital Business Transformation: Organizing, Managing, and Controlling in the Information Age*, Springer Verlag.

Haghshenas, M. and T. Østerlie (2020). «Navigating towards a digital ecosystem: The case study of offshore infrastructure industry», in Proceedings of the 11th Scandinavian Conference on Information Systems (SCIS2020)

Haghshenas, M. and T. Østerlie (2019). «Digital infrastructure innovation vs. digital innovation in infrastructure: Digital transformation in the offshore construction industry», in *Proceedings of the 6th edition of The Innovation in Information Infrastructure (III) workshop*.

Geological data preparation

This project is a collaboration with the Sirius research programs on semantic integration and digital geoscience, with Equinor as industrial partner.

Background. Analytics applied on vast quantities of digital data holds the potential for faster and better decisions. However, this is of limited use ifthe time it takes to prepare the data exceeds the decision window. This is the situation facing oil and gas companies in transitioning from traditional subsurface evaluation techniques towards data-driven decision-making. While existing data preparation techniques work perfectly with traditional (small data) subsurface evaluation, they do not scale to big data settings.

Results. This project is a continuation of our sustained engagement with transitions towards data-centric approaches in oil and gas exploration. This is a topic we have been engaged with since the beginning of Sirius; initially through the Geological Assistant demonstration that we initiated together with Equinor and Schlumberger, and later through sustained collaboration with Equinor's data management community. Our output from this has been predominantly theoretical results, but has formed and continues to form the basis for innovations developed by the other research programs.

Future work. We will contribute to evaluating and improving the prototypes to be delivered by the other research programs as part of this study.

Further reading

Parmiggiani, E., T. Østerlie, and P.G. Almklov (2022). «In the backrooms of data science», *Journal of the Association for Information Systems*, 23(1), pp.139-164.

Monteiro, E., T. Østerlie, E. Parmiggiani, and M. Mikalsen (2018). «Quantifying quality: Towards a post-humanist perspective on sensemaking», in Schultze, U., M. Aanestad, M. Mähring, and K. Riemer (Eds.), *Living with monsters? Social implications of algorithmic phenomena, hybrid agency, and the performativity of technolgy*, Springer Verlag.

Østerlie, T., Parmiggiani, E., and Monteiro, E. (2017). «Information infrastructure in the face of irreducible uncertainty», in *Proceedings of the 5th edition of Innovation in Information Infrastructure (III) workshop*.

Team members

Thomas Østerlie

Eric Monteiro

Elena Parmiggiani

Mina Haghshenas

Analysis of Digital Twins

A digital twin is typically a system which collects data about a physical asset such as a plant or a reservoir.

The digital twin is a vision for a technology, originally conceived for NASA's space program, enabling industry to significantly improve the life-cycle management of physical assets. A digital twin is typically a system which collects data about a physical asset (such as a plant or a reservoir), continuously revises this data set through, e.g., updates reflecting changes to the asset's structure and sensor data reflecting the physical asset's state and uses this data to monitor and make predictions about the physical asset. The digital twin can be thought of as a three-layered structure: the data sources, an information layer, and an insight layer. Industrial focus is today mainly on collecting data into shared, and increasingly structured, data sets which we think of as the information layer, and on providing dashboard-like insights into the system.

This project will focus on analysis support for digital twins, by building or combining tools which can leverage the information layer into insights. The purpose of these tools can be to reproduce and explain past events, to explore alternatives for decision making, to prepare for incidents or to optimize production. A central goal for this project is to combine semantics, behavioral and conceptual modelling techniques, and analysis methods in the context of digital twins.

Methodological background for the work is an integration of ontology-based conceptual modelling techniques, formal methods, and data-driven techniques for system analysis.

The work has synergies with, and feeds technology to the PeTWIN and other Digital Twins projects.

Objectives

- Understand the design space for coupling behavioral and conceptual models.
- Develop methods that combine structured information with behavioral analyses.
- Understand how conceptual modelling can be used to integrate analyses results.
- Develop experience with semantics foundations for co-simulation.
- Develop methods for decision making with digital twins.

Activities

- Develop a formal theory of coupled behavioral and conceptual models.
- Develop prototype tool for programming with semantics.

- Develop methods for coupling simulators by means of semantics and constraint representations.
- Develop methods for exploring semantics to express «what-if» scenarios in multi-model simulation and analysis.
- Collaborate with and disseminate results through the PeTWIN demonstration cases.

Highlights of 2021

A prototype domain specific language SMOL has been developed, which supports high-level ways to program the orchestration of simulation units which implement the FMI standard. SMOL seamlessly integrates semantic technology with programming constructs for interacting with simulators. SMOL has a formally defined semantics and a prototype runtime implemented in Kotlin, a dialect of Java. SMOL addresses the composition problem for simulators with different domain models through lifting semantic technologies to enable correct configuration and orchestration of connections inside the digital twin. By embracing semantic technologies, formal methods can contribute to the development of provably correct digital twins.

Einar Broch Johnsen

Eduard Kamburjan

Team members

Einar Broch Johnsen	Vidar Norstein Klungre
David Cameron	Rudolf Schlatte
Martin Giese	Silvia Lizeth Tapia Tarifa
Eduard Kamburjan	

Selected Publications for Further Reading

- Eduard Kamburjan, Vidar Norstein Klungre, Rudolf Schlatte, Einar Broch Johnsen, Martin Giese: *Programming and Debugging with Semantically Lifted States.* In: ESWC 2021. Springer 2021.
- Eduard Kamburjan, Egor V. Kostylev: *Type Checking Semantically Lifted Programs via Query Containment under Entailment Regimes.* Description Logics 2021. CEUR Workshop Proceedings 2021.

Selected SIRIUS Programs and Projects

PeTwin

SIRIUS DEMONSTRATION PROJECTS

Cross-Domain Application

The main objective of cross-domains is to demonstrate how methods and tools created by SIRIUS' researchers obtained from working in the oil&gas domain transfer to other areas in industry, biomedicine, and environmental applications. In 2021, there were three main areas of focus for cross-domain: Healthcare, Earth Science, and Biodiversity.

- Transfer to Healthcare

We have previously focused on semantic integration and semantic mapping tasks (in the earlier BIGMED project) as well as formal methods for process planning (health organization logistics). In 2021, we conducted scoping work and consortium building activities for data-driven decision making in healthcare policymaking. This work resulted in cross-domain applications from the Ontology Engineering and Data Science research programs. Scoping activities in ontology-based data integration, included exploration of healthcare-related OBDI mappings and requirements for an AI Knowledge Base, and Explainable AI. A processing pipeline for unstructured data was also mapped out as a part of these activities.

- The Earth Science project covers research by the Execution Modelling and Analysis group on cloud computing.

This project is concerned with the development of distributed deep learning techniques for the analysis of satellite images, focussing on the detection of sea ice. The project is run by the CIRFA SFI in Tromsø, with Einar Broch Johnsen as part of the supervision team for a PhD student. The PhD project is now in the completion phase.

- The Norwegian Biodiversity Information Centre (Artsdatabanken) is a project related to ontology engineering research (OTTR and modeling) and ODBI.

In 2021, SIRIUS delivered a report to the Norwegian Biodiversity Information Centre covering design, development, reuse, and implementation of ontologies in environmental science as well as ontology-based data integration (ODBI) for a new service, TraitBank. The Norwegian TraitBank is Norwegian Biodiversity Information Centre's solution for describing, connecting, and displaying data on species and nature type traits. This growing resource containing Norwegian species' trait data will be used for knowledge-based

Laura Slaughter

conservation and research. The work completed by SIRIUS included piloting implementation of semantic technology for the Norwegian Biodiversity Information Centre's data on species and an initial ontology constructed for the TraitBank. The result of this work was a step towards streamlining ontology reuse/construction and implementation of OTTR templates. In addition, we focused on integrating TraitBank's data with other NBIC resources, specifically data for red-listed and alien species. Our vision is to demonstrate the tools built by the ontology engineering group that apply to environmental science data.

For further information about this work, contact Laura Slaughter or Leif Harald Karlsen.

Highlights of 2021

- Biodiversity applications and further work on ontologies.
- Submission of proposal with Norwegian Cancer Registry. Scoping activities in biomedical applications

Data preparation is key to data-driven decision making. Oil and gas companies are transitioning towards more data-driven decision-within the subsurface domain. By visualizing large volumes of complex data through dashboards and other forms of business analytics techniques, decision-makers are to make decisions faster and with greater confidence.

GeoDataPrep

However, such data-driven decision-making is moot if the time spent preparing subsurface data for analysis and visualization far exceeds the time saved by decision-makers.

GeoDataPrep project targets data preparation workflows necessary for dashboarding and business analytics in the subsurface domain.

The point of departure for the GeoDataPrep project was the observation that data scientists in Equinor spend inordinately much of their time and energy preparing subsurface data for Big Data analytics. This observation supplements Sirius' ongoing engagement with OSDU as a repository for storing and retrieving large data sets for Big Data analytics by targeting the use side of collecting and making large subsurface datasets available for analytics.

In this project, we are mainly targeting the issue of naming variability in geology and well log metadata.

Naming variability is something geoscientists are well acquainted with and have developed robust and reliable methods to handle. However, these methods and practices do not scale to the high volumes of data used for data analytics. The initial study of this research project showed that dashboarding projects could spend several man-years just harmonizing existing naming variability within large sets of data to the extent that the algorithms used can run and give useful results.

Some key observations from this study

- The data preparation is a large part of the work for making dashboards, often not easily visible to other people in the organization. The time needed is often hard to estimate ahead of time.
- Data preparation includes harmonizing the data and inferring lacking data to an extent that the algorithms can run and give useful results. Several man-years in preparation for a single dashboard is not unusual.
- Variation in names must be aligned, a problem that is easy for a domain specialist when the scale is small enough, for example, all wellbore logs from a handful of wells, but unmanageable when the scale is large enough, for instance the Norwegian continental shelf.

The specific use cases in this project

- Metadata associated with the Well log data. This data is produced during and after drilling and is frequently used to study the petrophysical properties of the subsurface for further hydrocarbon exploration and production.
- The data in the geological knowledgebases; a key information in subsurface studies but is attributed to frequent and complex naming variability. One of the main reasons for this variability is the fact that a single object in the subsurface can be defined, described and data stored by different names representing particular aspects of the geological domain.

Highlights of 2021

In 2021, Researchers in SIRIUS worked on

- A detailed study on understanding the problem. SIRIUS researchers conducted a series of interviews at Equinor to understand better issues surrounding data preparation.
- Creating a demonstrator for naming variably in the geological knowledgebases. We used semantic reasoners to infer and enrich information for a particular object from different G&G aspects.
- Transformation of the publicly available G&G dataset (Vovle Field) to a relational database to be used as the main source of data for this project.
- Using Machine learning techniques to predict the missing/uncertain part in the curve information block header.

Subsurface Data access and analytics

Demonstration: Subsurface data access & analytics

SIRIUS is building on the Ontology based data access technology developed in the EU project Optique to demonstrate how G&G data sources like enterprise databases, application databases , NPD factPages, DISKOS and OSDU can be integrated and developed into digital platforms for exploration, research and innovation. Once this data is opened up, it needs to be analysed. For this reason, we are also working with image analysis, data science and natural language applications in sub-surface data management.

Subsurface Data Access

SIRIUS is working on a vision of providing a platform for innovation in the sub-surface. A recent book by Andrew McAfee and Erik Brynjolfsson, Machine Platform Crowd traced the role of platforms as enablers for innovation by crowds of workers. We believe that there is a need to open up subsurface data to researchers and innovators to try out their ideas on real data. We also believe that national data repositories, like DISKOS, have the potential to provide such a platform. However, for this to be done, we need to improve access to the data and allow it to be linked with data in other databases. We also need to improve access to unstructured text information in these databases. SIRIUS has several active projects to address the subsurface data access challenges.

Project 1: OBDA Subsurface Pilot

Exploration digital transformation is about overcoming the bottleneck of data access and increasing the quality of interpretations by means of the better use of data. The data access bottleneck is substantial as up to 70% of exploration experts' time is spent finding, accessing, integrating, and cleaning data before analysis can even start. (Putting the FOCUS on Data, W3C Workshop on Semantic Web in Oil & Gas Industry, Jim Crompton)

One possible approach to address this challenge is to extend the OBDA (Ontology-based data access) theory and tools to support the data access challenges for the sub-surface data. OBDA was extended in the Optique project to meet the needs of the oil & gas industry, but the solution has failed to be adopted due to its technological limitations. For example, a significant obstacle is that the current OBDA can only provide access to data stored in databases. This scenario suits the particular use-case used at that time (Slegge database) at Equinor (Ontology-Based Data Access to Slegge, ISWC 2017). Further, this database is no longer active, and access to that is now restricted for further experimentation.

Adnan Latif

In 2021, we developed the V1.0 of the OBDA subsurface pilot. This work includes setting up a large relational database from the publicly available G&G datasets, mappings, ontology, and integration. This Pilot was demonstrated at the SIRIUS GA in fall 2021. In 2022, we aim to work with SIRIUS partners to evaluate the developed Pilot, both for usefulness and usability and work further on extending the OBDA capabilities based on the feedback.

See more details on this project at page 47.

Project 2: Geo Data Prep

Oil and gas companies are transitioning towards more data-driven decision-within the subsurface domain. By visualizing large volumes of complex data through dashboards and other forms of business analytics techniques, decision-makers are to make decisions faster and with greater confidence. However, such data-driven decision-making is moot if the time spent preparing subsurface data for analysis and visualization far exceeds the time saved by decision-makers.

This project targets data preparation workflows necessary for dashboarding and business analytics in the subsurface domain. In particular, we target the issue of naming variability in well logs. Naming variability is something geoscientists

are well acquainted with and have developed robust and reliable methods to handle. However, these methods and practices do not scale to the high volumes of data used for data analytics. The initial stages of this research project showed that dashboarding projects could spend several man-years just harmonizing existing naming variability within large sets of well logs to the extent that the algorithms used can run and give useful results.

See more details on this project at page 43.

Project 3: SIRIUS Subsurface Lab

Equinor has released a big dataset from the Volve field; this dataset consists of a variety of structured and unstructured subsurface data. This can be used to establish a subsurface laboratory that can be further used to prototyping various projects and run experiments requiring Multiphysics data.

In 2021, we cleaned this dataset and transformed it into a relational database. This database is now being used in several SIRIUS and DigiWell projects (as a G&G database). In 2022, we aim to deploy an API to access this database and create a sandbox environment where researchers can connect their prototypes directly to this database and run experiments. Even after the SIRIUS life, this work will produce a long-term asset for data science, computer science, and geoscience researchers.

Subsurface Data Analytics

Faster access to relevant data is of interest only if the data can be used to create insight and drive decisions. The Exploration scoping workshop identified several challenges related to information extraction and structured and unstructured data usage. Some of them are given below.

Unstructured Data- Documents Domain experts spend a massive amount of time annotating corpora to train supervised statistical learning models for unstructured data.

Unstructured Data – Images Finding a geological image based on its technical content from a large image database is difficult. Geoscientists use the keyword search on the textual content of source documents to find relevant Images.

Structured Data Geoscientists use Reservoir Analogues to estimate the missing or uncertain reservoir parameters. Finding, selecting appropriate analogues, and extracting inferences depend on the team´s experience and limited human capacity.

To address problem number 1, a PhD project was recently completed on domain adopted knowledge extraction from Oil and Gas documents. The methodology developed in this project supports a significant reduction in the time and effort required in creating training sets using domain adaptation techniques.

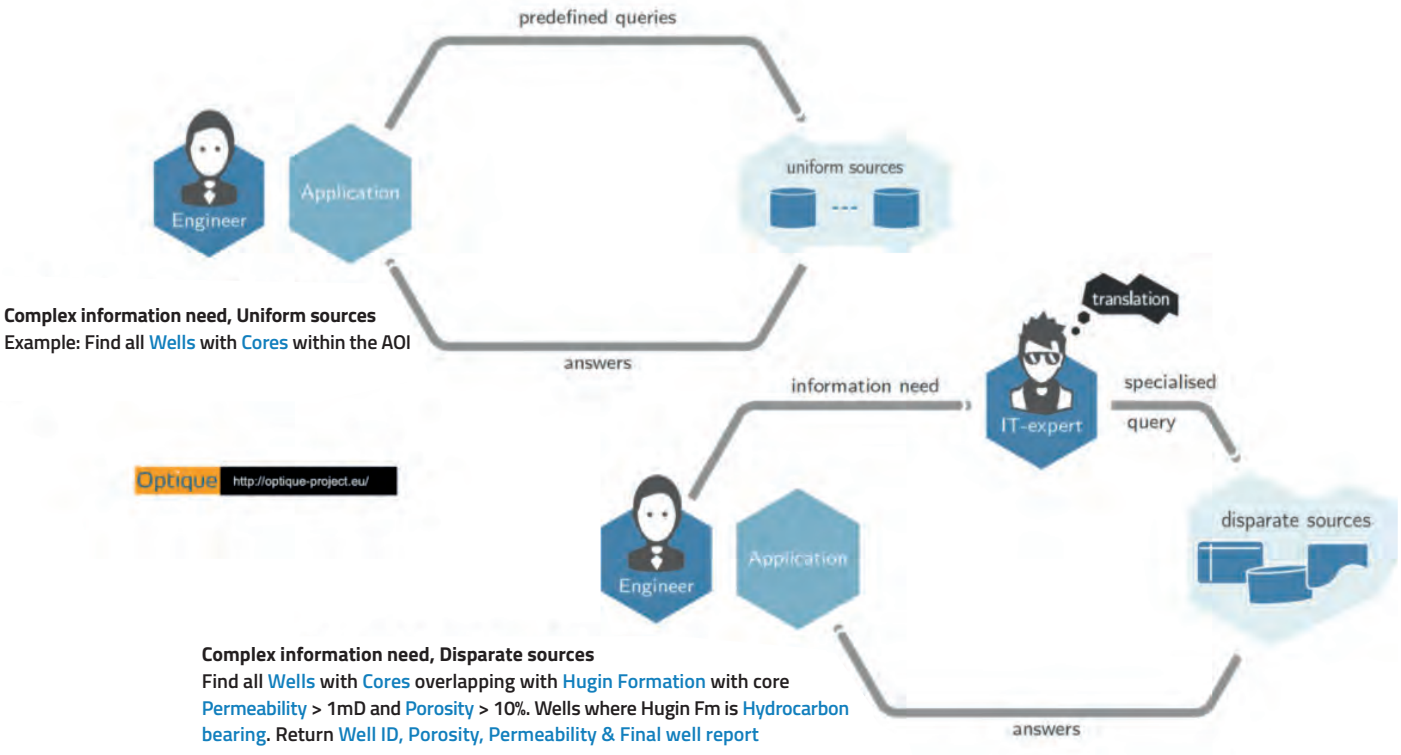
For problem number 2, an innovation project was initiated in 2019, and a prototype is developed. This tool supports executing complex queries to find geological images based on the geological content embedded in the images and significantly reduces the time and effort required to find the most relevant images and corresponding documents.

For problem number 3, a PhD project was started in 2017 and completed in 2021. In this project, techniques are developed to identify and quantify formal domain knowledge, thus predicting more accurate parameters for exploration modelling. The main objective is to extend a Machine learning model that can incorporate Oil & Gas domain information and recommend analogues to a reliable extent.

SIRIUS OBDA Subsurface Pilot

Subsurface digital transformation is about overcoming the bottleneck of data access and increasing the quality of interpretations by means of the better use of data. The data access bottleneck is substantial as up to 70% of subsurface experts' time is spent finding, accessing, integrating, and cleaning data before analysis can even start. (Putting the FOCUS on Data, W3C Workshop on Semantic Web in Oil & Gas Industry, Jim Crompton)

Viewed from the Geoscientist, *it is hard to get an overview of all available data related to an area of interest*, as this data is spread over different applications and many internal and external data sources. No unified view is, as a rule, available up front, though Project Data Managers (PDMs) assist. *It is difficult to extract data from databases;* should complex queries have to be written, Central Data Managers (CDMs) typically assist. *It is challenging to extract data and information based on geological and petrophysical attributes (see the user scenario example below)* as it is not possible to execute these types of queries simultaneously on multiple data sources. *It is challenging to integrate datasets before analysis can start:* this is often tedious manual work that the geo-scientists must do themselves. *It is incredibly difficult to extract data and knowledge from the text documents* as there are very few tools that can deal with the contents of unstructured documents and reports. Geoscientists are well aware of the limitations of the workflow. As a result, valuable analyses on data are too often not performed, and possibilities in data are too often not detected.

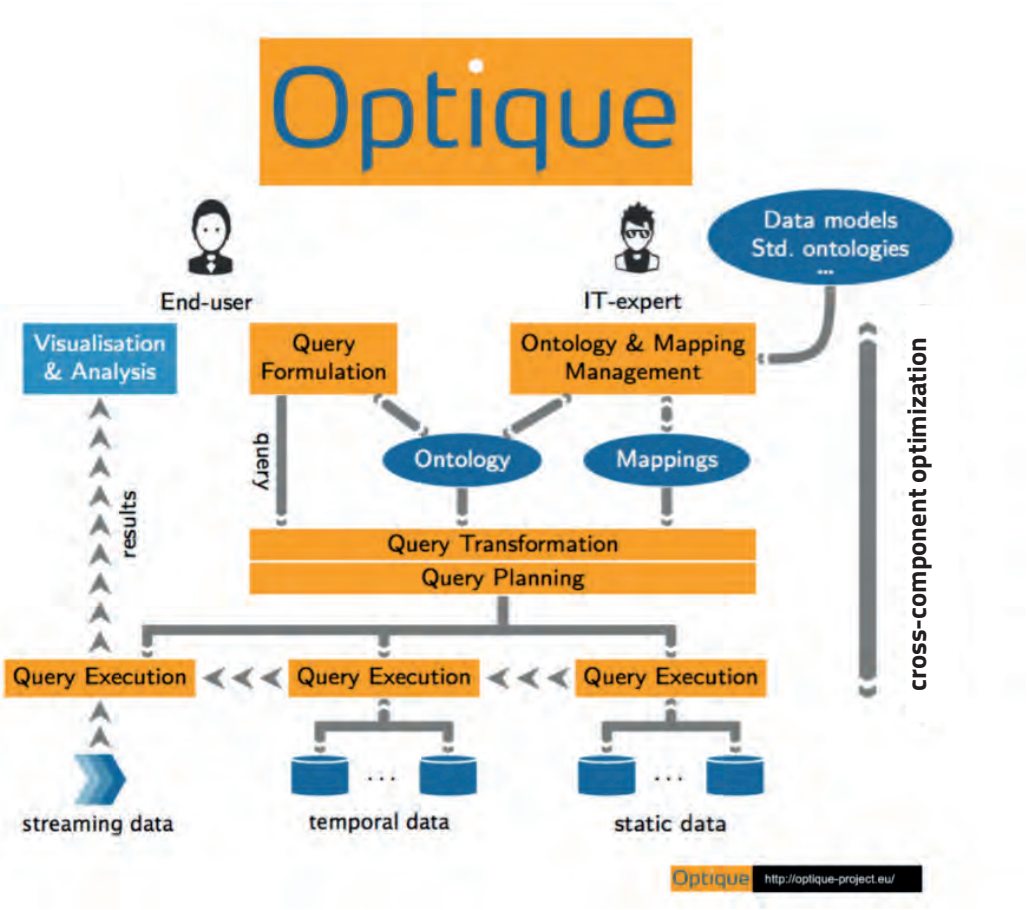


Complex information need, Uniform sources
Example: Find all Wells with Cores within the AOI

Complex information need, Disparate sources
Find all Wells with Cores overlapping with Hugin Formation with core Permeability > 1mD and Porosity > 10%. Wells where Hugin Fm is Hydrocarbon bearing. Return Well ID, Porosity, Permeability & Final well report

It is urgently needed to build competence and tools for the *exploration data wrangling*. An exploration data wrangler has competence in both geoscience and digital technologies. This competence is crucial to integrate the workflows of geoscientists, PDMs and CDMs and plays an important role in enabling digital transformation in exploration work practices. Along with supporting exploration teams with the routine data access tasks, the data wranglers will efficiently exploit opportunities brought by new IT technologies. This includes efficient handling of critical tasks such as identifying relevant data sources, developing complex ad-hoc queries over federated databases, and retrieving information from reports stored as text documents. In these ways, the data wrangler can bring data much closer to the project teams and give geoscientists a radically better possibility of extracting data and information with the exact specification (in terms of complex geological and petrophysical attributes) they need for their subsurface evaluation.

For the data wrangler to be less dependent on the CDMs and PDMs than the geoscientists are today we need to capture the special knowledge of the CDMs and PDMs and buildt his into data wrangling tools. Asuccessful attempt in this direction was Optique, a 14M Euro EU project that finished in 2016. Optique showed that geoscience knowledge could be reliably captured in a knowledge graph (or an ontology) and reusable mappings from CDMs could efficiently connect this knowledge graph to data in data-bases. Optique then demonstrated that complex queries over several federated data sources (including EPDS, NPD FactPages, Open Works installations, GeoChemDB, CoreDB and DDR) could be easily written and efficiently executed. Since the process was fully automated, tasks that normally would take several days could, with the Optique platform, be performed in minutes.



Optique showed the potential to transform the way data is gathered and analyzed by streamlining the workflow and making it more user-friendly. However, Optique has also revealed shortcomings that impede the realization of its full potential: (i) limitation to relational databases, (ii) lack of built-in support for quantitative analytics, (iii) lack of access to unstructured data, and (iv) limited tool support for constructing and maintaining the necessary ontology and mappings.

The SIRIUS OBDA subsurface pilot project is addressing these shortcomings and aim to significantly broaden the applicability of the approach for use in subsurface projects.

Highlights of 2021

- [subsurface part of] the Optique was focused on developing OBDA on a relational database at Equinor (Ontology-Based Data Access to Slegge, ISWC 2017), which is no more an active database. Further, that database contains proprietary data, and access to that is now restricted. This resulted in a significant setback for researchers to continue further extermination on extending OBDA capabilities. In 2021 we established a large in-house relational database from the publicly available G&G datasets (mainly by processing the Volve dataset https://data.equinor.com & NPD FactPages https://factpages.npd.no). This database is now being utilized in various other internal and external research projects where G&G data is being used for experimentation, e.g., DigiWell at USN is also using this database for research purposes.
- Adapt mapping and Ontology from Optique
- Integration & testing
- SIRUS OBDA Subsurface V1.0 was demonstrated at the SIRIUS General Assembly in November 2021

Plans for 2022

- Evaluation of the pilot with Geoscientists, PDMs & CDMs at partner companies
- Experimentation on using OTTR (https://ottr.xyz) for building and maintaining Ontology and mappings.
- Experimentation on extending the data sources to NoSQL data
- Integration with OSDU & potentially with DISKOS
- Design, Development, Deployment and Evaluation – OBDA Subsurface Pilot V2.0

Digital Design Basis

Prototyping a Shared Data Model for Early-phase Field Development | **The Digital Design Basis project concluded this year and provided valuable experience for further work in the READI and IMF projects. The common data model, written using OWL and RDF was used to demonstrate how design basis information from an operator could be made available to engineering applications used by different suppliers. The project brought together Lundin Energy, AkerBP, Equinor, Aker Solutions, TechnipFMC and Aize. The work was organized as a SIRIUS Innovation project. The project results are freely available at the SIRIUS web site. Results were presented at the Advances in Process Digitalization conference and will be published in *Digital Chemical Engineering*.**

David Cameron

The Digital Design Project started in 2019 and concluded in mid 2021. Since the end of the project, we have worked on preparing the results for publication and building on its results in the READI and IMF projects. We have developed and demonstrated a common digital model representation of the information in early-phase design bases for oil & gas field developments. The scope of the project was to develop a proof of concept for a Digital Design Basis that supports data-centric rather than document-based engineering.

The project established a standards-based data model that holds data about both the design basis and functional requirements decided by an operator. This model that can be implemented in any relevant software tools in a concept study, to ensure that information shared between operators and EPC vendors, with their different software tools, have the same meaning and understanding. The model is based on a common digitalized language for communication along the field development supply chain.

Semantic modelling made this representation possible and allowed data to be entered in a structured way and be consumed by engineering applications. We have validated the basic approach, which builds on reusing existing semantic models where possible. We have also demonstrated the feasibility of mixing the modelling approaches defined by ISO15926 and ISO/IEC81346.

We believe the industry needs to have more projects like this, where consortia along the supply chain work with academia and software vendors to agree on interoperability standards by working on real, non-trivial problems. Fortunately, it appears that the European Union, World Economic Forum, and International Organization of Oil & Gas Producers agree with this goal.

Our approach here is not restricted to the oil & gas industry. The system breakdown and modelling of fluid properties can be extended straightforwardly to chemical, fine chemicals, and energy applications. A good first step would be an extension of the RDS for Oil & Gas to ensure that it covers the unit operations in these other domains. We are working further with the READI partners to do this.

This work was experimental, where we were seeking to prove that recent advances in system modelling and ontologies could be applied to a real design basis problem. This meant the integrations with tools tended to be pragmatic rather than user-friendly and scalable. Further work is needed to provide the tools that are necessary to integrate models like this into engineering workflows.

Semantic technology tools are too low-level to be used by practising engineers. OTTR templates have addressed some of these usability challenges, but further work must focus on developing a set of tools to simplify configuration of the model and access to data.

A graphical tool is needed for building system-oriented models by selecting nodes and connecting these nodes with topological and semantic relationships. This tool should also allow the configuration of design basis data in a guided, but flexible sequence. This interface can exploit the semantic

content of the model to provide flow and check consistency.

This interactive tool must be supplemented with tools that allow data to be entered into the DDB in bulk, using tabular data. OTTR provides some of this functionality, but this needs to be lifted up into interactive tools. Mappings need to be developed towards common process simulators and engineering design databases. We need to both read design variables from the DDB and write calculated results back to the DDB. Here we need to work together with the industry so that our models and vendor's models converge over time to an actual or de-facto standard. We are cautiously optimistic about the possibilities of this being successful. We see that many influential vendors are interested in exposing their data using semantic schemas and open formats.

We should aim for work practices where a DDB harvests data from engineering tools without intervention from the engineer. The proof-of-concept has also helped us to develop a more systematic approach to defining the digital design basis. Time and organizational constraints meant that this first effort was more inductive than deductive. The modelling was driven by the data we had to represent, and we then drew systematic conclusions from the solutions developed.

The lessons from this project have been taken up in further, ambitious initiatives by each of the partners. We are contributing to the revisions of READI IMF and RDS for Oil & Gas. This work aims to address the tooling challenges above in the context of several on-going field development projects. It is also developing a formal systematization of modelling and use of data in engineering projects. The results of this work have also been taken into the development of the forthcoming Part 14 of the ISO15926 standard. We hope that these initiatives together will provide elements for establishing a practical, scalable framework for sharing information in the process engineering sector.

PeTWIN

Collaboration with Brazil around Digital Twins in Oil & Gas | The PeTWIN project was finally able to kick off at the end of 2021. Work has been made difficult by COVID-19, as we have not been able to collaborate in person with our partners in USA and Brazil. However, we have made good initial progress, with three post-doctoral researchers and a PhD fellow in place. We held three workshops and a course, held by Shell, on agile development. The work in Oslo is looking at the combination of semantics with formal simulation of cyber-physical systems, linkages to knowledge representation in the READI and IMF projects, and semantic modelling of subsurface in production.

PeTWIN is a Petromaks Brazil project, financed by the Research Council of Norway. It is part of a program where the Research Council collaborates with FINEP, the Brazilian national innovation funding organization, to finance projects with Brazilian and Norwegian partners. PeTWIN's Brazilian Partners are the Federal University of Rio Grande do Sul (UGRGS) and the Libra Field Development Consortium, hosted by Petrobras. The Norwegian Partners are the University of Oslo, Shell and Equinor. The project started in late 2020 and will run until 2024. The project team in Oslo consists of David Cameron, project manager, Martin Giese, scientific leader, three post-doctoral researchers (Eduard Kamburjan, Christian Kindermann and Vidar Klungre) and a PhD fellow (Irina Pene). The project team is supplemented by an internally financed PhD fellow (Yuanwei Qu) and a post-doctoral researcher financed by the Digiwell project (Baifan Zhou). Digiwell is a Petromaks project, led by the University of South Eastern Norway, with University of Oslo, Equinor, Kongsberg Digital, SINTEF, MIT and Imperial College as partners. Coordination between PeTWIN and Digiwell is beneficial to both projects, as they have overlapping interests.

The aim of PeTWIN is literally to write the book about digital twins in the oil and gas industry. We plan to publish a text-book on this topic in 2024. We have prepared the table of contents and are negotiating with publishers. We are doing this by working on the fundamentals of digital twins in engagement with applications in Petrobras / Libra, Shell and Equinor. The use cases cover the field management chain, from reservoir monitoring, through production optimization to facility operations and maintenance. Our researchers bring complementary skills to specific aspects of the digital twin puzzle. For example, Eduard Kamburjan is researching how we can link semantic data access and integration with dynamic simulation of cyber-physical systems. He has worked this year with a Modelica library for facility simulation that was provided by Equinor. This work promises new ways to link our semantic technologies to complex digital twin models of facilities in a way that allows us to ensure the quality of the models and their results. Eduard is working also with Vidar Klungre, who provides the semantic technologies skills needed to realize the work.

Irina Pene and Yuanwei Qu are working on semantic modelling at the other end of the production chain, namely the sub-surface. Here we have a linkage to the Exploration demonstrations, in that the modelling work done by them is coordinated with the semantic models used in that work package. Our hope is that we can demonstrate the use of these models in modelling and using the knowledge from monitoring of fields in production.

Christian Kindermann joined PeTwin late in 2021. He will be looking at the application of semantic domain modeling and reasoning to digital twins, drawing on the results of the READI and IMF projects. We have seen that it would be profitable to use the results or the READI and IMF work as a framework for structuring and delivering digital twins. Our vision is still to provide common tools and standard knowledge models that can benefit Libra, Shell and Equinor.

SIRIUS Partners

SIRIUS draws together a consortium of leading industrial organisations across the oil & gas value chain, including operators (Equinor), service companies (Schlumberger, Aibel, Aker Solutions, Robert Bosch GmbH, DNV and TechnipFMC) and IT companies (Computas, Dolphin Interconnect Solutions, TietoEvry, IBM, Kadme, Numascale, OSISoft, SAP, Billington Process Technology, ONTOPIC, Oxford Semantic Technologies, Prediktor, Envester As and MetaPhacts). In addition, the mentioned companies collaborate with researchers from the University of Oslo, NTNU, the University of Oxford, Simula Research Laboratories and SINTEF.

Envester As and **MetaPhacts** joined SIRIUS as the new partners from June 2021.

Envester

Envester AS is a newly established company with the purpose of advancing the innovation and digitalisation of the energy business and related technology ecosystem. They are heavily involved in work related to semantic asset modelling and the application of the READI results in the oil and gas industry. In SIRIUS, Envester will contribute a substantial amount of in-kind in the Information modelling framework related projects and background technology to the laboratory.

MetaPhacts

MetaPhacts is a Germany-based company delivering metaphactory – a platform that empowers customers to accelerate their knowledge graph journey and drive knowledge democratization, improve data literacy and reach smarter business decisions with data. The metaphacts team offers unmatched experience and know-how around enterprise knowledge graphs for our clients in areas such as pharma and life sciences, engineering and manufacturing, energy, finance, business, and cultural heritage.

Defended PhDs

Summaya Mumtaz

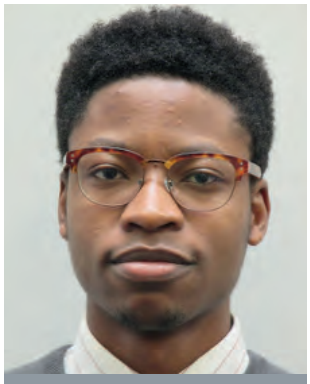
Summaya Mumtaz

Main research findings: The real-world application of Artificial Intelligence/machine learning techniques is challenging. Most of the standard machine learning approaches depend heavily on large amounts of historical data. However, in real-world complex use cases, the data vary across several dimensions which makes it challenging to find a sufficient amount of quality data. Particularly, in low-resource domains, not enough training data is available, which affects the machine learning model's performance. In many disciplines a significant amount of prior knowledge about the domain is available, often in the form of a taxonomy or a hierarchy. For instance, a disease hierarchy in the medical domain that classifies diseases into different groups based on similar symptoms. We have experimented in three domains by adding domain knowledge: recommending hydrocarbon reserves in the oil and gas industry, grouping similar words in natural language, and patient mortality prediction in health care domain. Our research has shown that addition of domain knowledge (taxonomy) in the given scenarios where little training data is available, can improve the performance of the prediction task.

Shiji Bijo

Shiji Bijo

Main research findings: Multicore architectures aim to improve execution speed of software through parallel computations. Large-scale multicore systems have massively parallel hardware architectures. The development of parallel programs which can exploit the multicore architecture is a non-trivial task for the software industry. Analysing the effect of different architectures during software development helps to uncover cases that may affect the performance. This thesis studies at the theoretical level how the underlying architecture and data movements between cores and memory systems may influence the program performance. The main contribution of this thesis is a detailed formal model of multicore architectures along with an associated proof-of-concept analysis tool. One of the main challenges in characterising the core-memory communication patterns in multicore systems is the consistency of shared data in different memory levels. The formalisation is used to guarantee consistency of shared data for all architectures that can be expressed in our formal model. The formal model captures interactions between cores and memory and the tool provides a model-based simulation environment to examine the effect of different multicore architectures on performance during parallel executions.



Temitope Ajileye

Temitope Ajileye

Abstract: Many RDF systems support reasoning with Datalog rules via materialisation, where all conclusions of RDF data and the rules are precomputed and explicitly stored in a preprocessing step. As the amount of RDF data used in applications keeps increasing, processing large datasets often requires distributing the data in a cluster of shared-nothing servers. While numerous distributed query answering techniques are known, distributed materialisation is less well understood. In this paper, we present several techniques that facilitate scalable materialisation in distributed RDF systems. First, we present a new distributed materialisation algorithm that aims to minimise communication and synchronisation in the cluster. Second, we present two new algorithms for partitioning RDF data, both of which aim to produce tightly connected partitions, but without loading complete datasets into memory. We evaluate our materialisation algorithm against two state-of-the-art distributed Datalog systems and show that our technique offers competitive performance, particularly when the rules are complex. Moreover, we analyse in depth the effects of data partitioning on reasoning performance and show that our techniques offer performance comparable or superior to the state of the art min-cut partitioning, but computing the partitions requires considerably less time and memory.

International Activity and Dissemination

International cooperation:

SIRIUS has strong international cooperation across various academic, research and industrial sectors. Some of the highlights on SIRIUS International collaboration is as follows:

- Partnerships at the Center of the University of Oxford and the City University London. Focus areas have for this collaboration is around semantic integration, databases and otology engineering.
- Collaboration with researchers at Birkbeck College, University of London. This collaboration is around the SIRIUS OBDA subsurface pilot project (Ontology-based data access for the subsurface data).
- Collaboration with UFRGS in Porto Alegre, Brazil. This collaboration is mainly focused on students mobility, partly funded by DIKU and DIKU / CAPES:
- Membership and participation in the EU's Big Data Value Association PPP and A.SPIRE PPP as part of Horizon 2020.
- Continuation for PETROMAKS / FINEP project PeTWIN.
- Participation in SINOS 'collaboration with Brazil.
- A new international partner joined SIRIUS in 2021: metaPhacts GmbH.

Horizon 2020 and European Cooperation

- The H2020 project Melodic (731664) was expanded with another project, Morphemic (871643). This project contributes infrastructure for and content to the SIRIUS HPC research program.
- SIRIUS, through Einar Broch Johnsen is, a partner in REMARO (a Marie Curie project that looks at using formal methods to ensure that autonomous robotics is safe and reliable).
- SIRIUS submitted ten project applications for H2020 this year:
- SIRIUS is a partner and WP leader in ONTOCOMMONS, an H2020 CSA for the use of semantics in materials technology.
- SIRIUS is a partner in an H2020-SC1-PHE-CORONA-VIRUS-2020-2-NMBP project Eur3ka, which will use ICT to build flexible responses to pandemics in the industry.
- Positioning work in an EU project has been done through membership in BDVA (Big Data Value Association) and A.SPIRE PPP for the process industries.
- SIRIUS is participating in a Nordic Interoperability Cooperation. We collaborate with universities and companies in Sweden (Luleå) and Finland (Tampere) to establish EU projects in this important subject area.

Priority Partners: Brazil and the United States

SIRIUS is leading a new INTPART project, DSYNE, collaborating with the University of Southeast Norway, the Federal University of Rio Grande do Sul, the Federal University of Espirito Santo, the University of Houston and the Stevens Institute of Technology.

SIRIUS is the coordinator, together with UFRGS, for the PeTWIN project, a collaboration project within digital twins for field management. Industrial partners are Equinor, Shell and Petrobras.

SIRIUS has used SIU UTFORSK and UTFORSK / CAPES funds to support the collaboration with the research group of Prof. Mara Abel at UFRGS. This group has complementary expertise in applying semantic technologies in oil and gas.

List of Staff

Name	Main research area	Institution/Funding
Key Researchers		
Arild Waaler	Knowledge Representation	University of Oslo
Martin Giese	Knowledge Representation	University of Oslo
Einar Broch Johansen	Execution Modelling & Analysis	University of Oslo
Ingrid Chieh Yu	Execution Modelling & Analysis	University of Oslo
Ian Horrocks	Knowledge Representation	University of Oxford
Boris Motik	Databases	University of Oxford
Eric Monteiro	Working Practices	NTNU
Jan Tore Lønning	Natural Language Processing	University of Oslo
Geir Horn	Scalable Computing	University of Oslo
Laura Slaughter	University of Oslo	University of Oslo
Evgeny Kharlamov	Knowledge Representation	University of Oslo/Bosch
Basil Ell	Natural Language Processing	University of Oslo
Xing Cai	Scalable Computing	Simula Research
Dirk Hesse	Domain-Adapted Data Science	University of Oslo/Equinor
Egor Kostylev	Domain-Adapted Data Science	University of Oslo
Soylu Ahmet	Knowledge Representation	University of Oslo/NTNU
Martin Skjæveland	Knowledge Representation	University of Oslo
David Cameron	Knowledge Representation	University of Oslo
Adnan Latif	Knowledge Representation	University of Oslo
Visiting Researchers		
Fabricio Henrique Rodrigues	Knowledge Representation	UFRGS
Baifan Zhou	Knowledge Representation	Bosch
Vinicius Graciolli	Knowledge Representation	UFRGS
Postdoctoral researchers with financial support from the Centre budget		
Thomas Østerlie	Industrial Digital Transformation	NTNU
Dag Hovland	Knowledge Representation	University of Oslo
Laura Slaughter	Knowledge Representation	University of Oslo
Ernesto Jiménez-Ruiz	Knowledge Representation	University of Oslo
Rudi Schlatte	SIRIUS Laboratory	University of Oslo
Violet Pun	Analysis of Complex Systems	University of Oslo
Jiaoyan Chen	Knowledge Representation	University of Oslo
Medha Atre	Knowledge Representation	University of Oslo
Jieying Chen	Knowledge Representation	University of Oslo
Chi Mai Nguyen	Analysis of Complex Systems	University of Oslo
Daniel Lupp	Knowledge Representation	University of Oslo

Name	Main research area	Institution/Funding
Stefano Germano	Knowledge Representation	University of Oxford
Jens Otten	Knowledge Representation	University of Oslo
Postdoctoral researchers working on projects in the centre with financial support from other sources		
Lizeth Tapia	Analysis of Complex Systems	RCN FRINATEK
Elena Parmiggiani	Working practices	Digital Oil Programme
Crystal Chang Din	Analysis of Complex Systems	RCN FRINATEK
Vidar Klungre	Knowledge Representation	RCN PETROMAKS
Eduard Kamburjan	Analysis of Complex Systems	RCN PETROMAKS
Christian Kindemann	Knowledge Representation	RCN PETROMAKS
PhD students with financial support from the Centre budget		
Vidar Klungre	Knowledge Representation	
Alessandro Ronca	Databases	University of Oxford
Lars Tveito	Analysis of Complex Systems	University of Oslo
Mina Haghshenas	Work Practices	NTNU
Temitope Ajileye	Knowledge Representation	University of Oxford
Sigurd Kittilsen	Analysis of Complex Systems	University of Oslo
Andreas Thune	Scalable Computing	Simula
Frederico Igne	Knowledge Representation	University of Oxford
Marta Rozanska	Scalable Computing	University of Oslo
Ratan Bahadur Thapa	Domain-Adapted Data Science	University of Oslo
Ole Magnus Holter	Natural Langauge Processing	University of Oslo
Peyman Rasouli	Analysis of Complex Systems	University of Oslo
Yuanwei Qu	Digital Geosciences	University of Oslo
Erik Hide Sæternes	Scalable Computing	Simula
PhD students working on projects in the centre with financial support from other sources		
Shiji Bijo	Scalable Computing	UpScale EU Project
Johanna Beate Stumpf	Analysis of Complex Systems	RCN FRINATEK
Daniel Lupp	Knowledge Representation	University of Oslo
Leif Harald Karlsen	Knowledge Representation	University of Oslo
Anatasia Gkolfi	Analysis of Complex Systems	RCN FRINATEK
Daniel Bakkelund	Domain-Adapted Data Science	University of Oslo
Gianluca Turin	Analysis of Complex Systems	RCN FRINATEK
Summaya Mumtaz	Domain-Adapted Data Science	University of Oslo
Chinmayi Prabhu Baramashetru	Analysis of Complex Systems	University of Oslo
Irina Pene	Digital Geosciences	RCN PETROMAKS

Name	Main research area	Institution/Funding
Master degree students		
Per Kulseth Dahl	Geological Assistant	University of Oslo
Shuvo Mahmuda	Probabilistic Soft Logic (PSL) Farmework University of Oslo in Ontology Alignment Tasks	University of Oslo
Nils Petter Opsahl Skrindebakke	Explaining Machine Learning Predictions University of Oslo Using Semantics	University of Oslo
Victoria Varzhel	Mapping Health Registry Variables to Biomedical Ontologies	University of Oslo
Sondre Skaflem Lunde	Analysis of Complex Systems	University of Oslo
Gaute Berge	Analysis of Complex Systems	University of Oslo
Vegar Skaret	Geological Assistant	University of Oslo
Benjamin Edward Oliver	Ontology Engineering	University of Oslo
Frank Hestvik	Ontology Engineering	University of Oslo
Fredrik Rømming	Ontology Engineering	University of Oslo
Torgeir Lebesbye	Analysis of Complex Systems	University of Oslo
Ida Sandberg Motzfeldt	Analysis of Complex Systems	University of Oslo
Eirik Halvard Sæther	Analysis of Complex Systems	University of Oslo
Nina Gjersøyen Løkamoen	Knowledge Representation	University of Oslo
Hafsa Bajwa	Knowledge Representation	University of Oslo
Ahmed Abdulrahman Hussein Abbas	Knowledge Representation	University of Oslo
Marianne Andresen	Ontology Engineering	University of Oslo
Birgitte Løwe Johnsen	Ontology Engineering	University of Oslo
Marlen Jarholt	Ontology Engineering	University of Oslo
Ingeborg Storesund Nes	Ontology Engineering	University of Oslo
Preben Zahl	Ontology Engineering	University of Oslo
Magnus Wiik Eckhoff	Ontology Engineering	University of Oslo
Justyna Ozog	Domain-Adaped Data Science	University of Oslo
Eivind Grønlie Guren	Domain-Adaped Data Science	University of Oslo
Henrik Syversen Johansen	Domain-Adaped Data Science	University of Oslo

Publications in 2021

Avogadro, Roberto; Cremaschi, Marco; Jimenez-Ruiz, Ernesto; Rula, Anisa. A Framework for Quality Assessment of Semantic Annotations of Tabular Data. Lecture Notes in Computer Science (LNCS) 2021 ;Volum 12922 UiO

Blomqvist, Eva; Hahmann, Torsten; Hammar, Karl; Hitzler, Pascal; Hoekstra, Rinke; Mutharaju, Raghava; Poveda-Villalon, Maria;Shimizu, Cogan; Skjæveland, Martin G; Solanki, Monika; Svátek, Vojtch; Zhou, Lu. Advances in Pattern-Based Ontology Engineering. IOS Press 2021 (ISBN 978-1-64368-174-0) 395 s UiO

Cameron, David B. Digital Twins for Science. The Science of Digital Twins. dScience Seminar; 2021-11-25 UiO

Cameron, David B.; Falk, Kristin; Kokkula, Satyanarayana (Satya). Towards Digital Requirements for Transformation in the Natural Resources Industries White Paper from the DSYNE Network Workshop, 9th-10thFebruary 2021. Oslo: SIRIUS Centre for Research-Based Innovation 2021 15 s UiO USN

Chen, Jiaoyan; Hu, Pan; Jimenez-Ruiz, Ernesto; Holter, Ole Magnus; Antonyrajah, Denvar; Horrocks, Ian. OWL2Vec*: embedding of OWL ontologies. Machine Learning 2021 ;Volum 110.(7) s.1813-1845 UiO

Chen, Jiaoyan; Jimenez-Ruiz, Ernesto; Horrocks, Ian; Antonyrajah, Denvar; Hadian, Ali; Lee, Jaehun. Augmenting Ontology Alignment by Semantic Embedding and Distant Supervision. Lecture Notes in Computer Science (LNCS) 2021 ;Volum12731 UiO

Chen, Jiaoyan; Jimenez-Ruiz, Ernesto; Horrocks, Ian; Chen, Xi; Myklebust, Erik Bryhn. An Assertion and Alignment Correction Framework for Large Scale Knowledge Bases. Semantic Web Journal 2021NIVA UiO

Chen, Jieying; Ma, Yue; Peñaloza, Rafael; Hui, Yang. Union and Intersection of all Justifications (Extended Abstract). the 34th International Workshop on Description Logics (DL 2021); 2021-09-19 -2021-09-22 UiO

Damiani, Ferruccio; Hähnle, Reiner; Kamburjan, Eduard; Lienhardt, Michael; Paolini, Luca. Variability Modules for Java-like Languages. I: SPLC '21: Proceedings of the 25th ACM International Systems and Software Product LineConference - Volume A. Association for Computing Machinery (ACM) 2021 ISBN 978-1-4503-8469-8 UiO

Elahi, Mohammad Fazleh; Ell, Basil; Grimm, Frank; Cimiano, Philipp. Question Answering on RDF Data based on Grammars Automatically Generated from Lemon Models. Technical University of Aachen 2021;Volum 2941.0 s. CEUR Workshop Proceedings(1)UiO

Ell, Basil; Elahi, Mohammad Fazleh; Cimiano, Philipp. Bridging the Gap Between Ontology and Lexicon via Class-Specific Association Rules Mined from a Loosely-Parallel Text-Data Corpus. I: 3rdConference on Language, Data and Knowledge (LDK 2021). Zaragoza, Spain: Schloss Dagstuhl-Leibniz-Zentrum für Informatik. 2021 ISBN 978-3-95977-199-3. s.33:1-33:21 UiO

Halvorsrud, Ragnhild; Mannhardt, Felix; Johnsen, Einar Broch; Tapia Tarifa, Silvia Lizeth. Smart Journey Mining for Improved Service Quality. I: IEEE International Conference on Services Computing, SCC 2021. IEEE 2021 ISBN 978-1-6654-1683-2. s.367-369 UiO SINTEF

Holter, Ole Magnus; Ell, Basil. Towards Scope Detection in Textual Requirements. I: 3rd Conference on Language, Data and Knowledge (LDK 2021). Zaragoza, Spain:Schloss Dagstuhl-Leibniz-Zentrum für Informatik. 2021 ISBN 978-3-95977-199-3 UiO

Jimenez-Ruiz, Ernesto. Ontology Alignment and the two DLs (Keynote). 34th International Workshop on Description Logics; 2021-09-19 - 2021-09-22 UiO

Kamburjan, Eduard; Grätz, Lukas. Increasing Engagement with Interactive Visualization: Formal Methods as Serious Games. Lecture Notes in Computer Science (LNCS) 2021s.43-59 UiO

Kamburjan, Eduard; Klungre, Vidar; Schlatte, Rudolf; Johnsen, Einar Broch; Giese, Martin. Programming and Debugging with Semantically Lifted States (Full Paper). Oslo: Universitetet i Oslo. Institutt for informatikk 2021 (ISBN 978-82-7368-464-6) 21 s. Conference proceedings (Universitetet i Oslo. Institutt for informatikk)(499) UiO

Kamburjan, Eduard; Kostylev, Egor. Type Checking Semantically Lifted Programs via Query Containment under

Entailment Regimes. CEUR Workshop Proceedings 2021 ;Volum2954 UiO

Kamburjan, Eduard; Schlatte, Rudolf; Johnsen, Einar Broch; Tapia Tarifa, Silvia Lizeth. Designing Distributed Control with Hybrid Active Objects. I: Leveraging Applications of Formal Methods, Verification and Validation: Tools and-Trends - 9th International Symposium on Leveraging Applications of Formal Methods, ISoLA 2020, Rhodes, Greece, October 20-30, 2020,Proceedings, Part IV. Springer 2021 ISBN 978-3-030-83722-8. s.88-108 UiO

Kindermann, Christian; Lupp, Daniel P.; Skjæveland, Martin G; Karlsen, Leif Harald. Formal Relations over Ontology Patterns in Templating Frameworks. I: Advances in Pattern-Based Ontology Engineering. IOS Press 2021 ISBN978-1-64368-174-0. s.120-133 UiO

Koopmann, Patrick; Chen, Jieying. Deductive Module Extraction for Expressive Description Logics. In Proceedings of the 29th International Joint Conference on Artificial Intelligence(IJCAI 2020); 2021-01-01 - 2021-01-30 UiO

Mikalsen, Marius; Monteiro, Eric. Acting with inherently uncertain data: practices of data-centric knowing. Journal of the AIS 2021 ;Volum 22.(6) s.1-21 SINTEF NTNU

Nolano, Gennaro; Elahi, Mohammad Fazleh; De Buono, Maria Pia; Ell, Basil; Cimiano, Philipp. An Italian Question Answering System based on grammars automatically generated from ontology lexica. Aachen, Germany: Technical Universityof Aachen 2021 0 s. CEUR Workshop Proceedings(0) UiO

Otten, Jens. The nanoCoP 2.0 Connection Provers for Classical, Intuitionistic and Modal Logics. Lecture Notes in Computer Science (LNCS) 2021 ;Volum12842. s.236-249 UiO

Schlatte, Rudolf; Johnsen, Einar Broch; Kamburjan, Eduard; Tapia Tarifa, Silvia Lizeth. Modeling and Analyzing Resource-Sensitive Actors: A Tutorial Introduction. Lecture Notes in Computer Science (LNCS) 2021 UiO

Skjæveland, Martin G. The Core OTTR Template Library. I: Advances in Pattern-Based Ontology Engineering. IOS Press 2021 ISBN 978-1-64368-174-0. s.378-393 UiO

Tena Cucala, David; Cuenca grau, Bernardo; Horrocks, Ian. Pay-as-you-go consequence-based reasoning for the description logic SROIQ. Artificial Intelligence 2021 ;Volum 298

Thapa, Ratan Bahadur; Giese, Martin. A Source-to-Target Constraint Rewriting for Direct Mapping. I: The Semantic Web – ISWC 2021. 20th International Semantic Web Conference,ISWC 2021, Virtual Event, October 24–28, 2021, Proceedings. Springer Nature 2021 ISBN 978-3-030-88361-4. s.21-38 UiO

Zhou, Baifan; Zhou, Dongzhuoran; Chen, Jieying; Svetashova, Yulia; Cheng, Gong; Kharlamov, Evgeny. Scaling Usability of ML Analytics with Knowledge Graphs: Exemplified with A BoschWelding Case. The 10th International Joint Conference onKnowledge Graphs (IJCKG'21); 2021-12-06 - 2021-12-08 UiO

Zhou, Dongzhuoran; Zhou, Baifan; Chen, Jieying; Cheng, Gong; Kostylev, Egor; Kharlamov, Evgeny. Towards Ontology Reshaping for KG Generation with User-in-the-Loop: Applied to Bosch Welding. The 10th International Joint Conference onKnowledge Graphs (IJCKG'21); 2021-12-06 - 2021-12-08 UiO

Annual Accounts

Costs

All figures in 1000 NOK	2015	2016	2017	2018	2019	2020	2021
Personnel and indirect costs	539	5188	10087	12378	20922	21783	22101
Purchase of research services	-	600	2113	3415	6166	6356	5796
Equipment	-	31	122	-	-	-	-
Other operational costs	62	9505	10424	14131	17202	17124	11189
Total Sum	**601**	**15324**	**22746**	**29924**	**44290**	**45263**	**39087**

Funding

All figures in 1000 NOK	2015	2016	2017	2018	2019	2020	2021
Research Council	-	4168	8180	12430	16961	17805	14853
University of Oslo	601	2141	5593	6387	7721	11758	8048
Public partners	-	144	813	107	1137	1350	-
Private partners	-	8510	7910	11000	17700	14350	15247
International partners	-	361	250	-	771	-	-
Total Sum	**601**	**15324**	**22746**	**29924**	**44290**	**45263**	**39087**

Photo credits
Shutterstock
Unsplash
The Noun Project

Editor
Maunya Doroudi Moghadam

Layout
FÆRD AS

SIRIUS

ANNUAL REPORT 2021

Sirius-labs.no