ercim-news.ercim.eu

Number 133 April 2023

ERCIM NEWS

Special theme:

Data Infrastructures and Management

Also in this issue

Research and Innovation: Survival Skills in the Digital Era: Helping Teachers to Develop Digital Competences

Editorial Information

ERCIM News is the magazine of ERCIM. Published quarterly, it reports on joint actions of the ERCIM partners, and aims to reflect the contribution made by ERCIM to the European Community in Information Technology and Applied Mathematics. Through short articles and news items, it provides a forum for the exchange of information between the institutes and also with the wider scientific community. This issue has a circulation of about 2,000 printed copies and is also available online ,at https://ercim-news@ercim.eu.

ERCIM News is published by ERCIM EEIG
BP 93, F-06902 Sophia Antipolis Cedex, France
+33 4 9238 5010, contact@ercim.eu
Director: Philipp Hoschka, ISSN 0926-4981

Contributions
Contributions should be submitted to the local editor of your country

ERCIM News online edition: *https://ercim-news.ercim.eu/*

Next issue:
July 2023: Explainable AI

Subscription
Subscribe to ERCIM News by sending an email to en-subscriptions@ercim.eu

Cover photo by Mathew Schwartz on Unsplash

SPECIAL THEME

RESARCH AND INNOVATION

ANNOUNCEMENTS / IN BRIEF

Introduction to the Special Theme

Data Infrastructures and Management

by Manolis Terrovitis (ATHENA RC)
and Dominik Ślęzak (University of Warsaw)

Data infrastructures are critical for efficient handling of large-scale datasets in any domain, and they are required for industrial, commercial, scientific and policy-making purposes. Robust data infrastructures ensure data security and integrity, which is essential for compliance and maintaining the trust of users and stakeholders. They allow for scalability and flexibility, which is needed to handle the constant influx of new data and changing the data processing pipelines. A well-designed data infrastructure ensures that data is properly integrated, standardised, and secure, which is crucial for accurate and reliable analysis.

The last decade has witnessed an explosion in data infrastructures that are available from private parties. We can also see bold initiatives by the European Commission to offer data infrastructures for scientific purposes (European Open Science Cloud, OpenAIRE, European Data Infrastructure etc.) and regulatory approaches to create common data spaces in the EU (e.g., Common European Health Data Space). The development of data infrastructures poses multiple scientific, governance and social challenges that have attracted great interest from the scientific community but also from regulators. This special issue features contributions addressing the aforementioned challenges that can be grouped in the areas as follows.

Experiences from Existing Data Infrastructures

Candela et al. (p. 6) present virtual research environments (VREs) of the D4Science infrastructure. The VREs are web-based, collaborative, and open-science-enablers, allowing seamless access to datasets and services. Bardi et al. (p. 8) present the ARIADNE infrastructure, which provides archaeological researchers with a suite of tools and services to support different phases of the research life cycle. These include a knowledge base of 3.5 million archaeological resources, virtual research environments for data analysis, data sharing and collaboration, as well as tools for data visualisation. Bardi and Benassi (p. 9) describe how the IPERION HS (heritage science) research infrastructure makes its outputs open and accessible to the community. Its impact is monitored with the help of services of the OpenAIRE infrastructure, such as Zenodo, Connect and Monitor. They let IPERION HS store research resources in a FAIR (findable, accessible, interoperable, reusable) way, make research outputs discoverable, and track the success of its Open Science strategy. Avramo et al. (p. 11) present the data portal of the European Plate Observing System (EPOS), which integrates different data, metadata, software and services into one platform for solid-earth sciences in Europe. The data portal is built around the FAIR principles and is designed to ensure sustainable and universal use and reuse of multidisciplinary solid-earth science data.

Privacy and Security for Research Infrastructures

Salant (p.16) describes the development of a privacy-aware framework for fine-grained data access that enables secure, policy-driven data exchange with sophisticated access control. Krenn at al. (p. 17) discuss the benefits of joint computations on data from various sources. The authors developed verifiably privacy-preserving protocols based on multi-party computation, which allows parties to jointly evaluate an arbitrary function without revealing anything about the input data. Spanakis et al. (p. 19) describe the secure and trustworthy platform for cross-discipline federation of data using self-sovereign technologies and homomorphic encryption. The article presents the platform's architecture and highlights its key layers. Albanese et al. (p. 21) present the E-CORRIDOR framework for multi-modal transportation systems. It provides secure services to passengers and transport operators by implementing collaborative privacy-aware edge-enabled information sharing, analysis and protection as a service. The framework is based on the concept of data sharing agreement, which is a digital contract that defines a set of data sharing constraints. Anciaux and Bouganim (p. 23) propose a three-layer logical architecture for extensive and secure Personal Data Management Systems. It allows individuals to manage and control their personal data while allowing third-party applications to access it. The paper also discusses challenges related to handling large volumes of personal data, protecting data of a community of users and retrieving third-party data. Carreras et al. (p. 24) present the components of a trusted ecosystem for sharing medical data in a secure and privacy-preserving manner. In particular, the data anonymisation and functional encryption modules are discussed.

Artificial Intelligence and Data Infrastructures

Hemker et al. (p. 26) propose a modularised approach for reducing complexity in the AI life cycle and describing its data-related components. The platform follows the principles of the Unix philosophy, solving problems with small and effective tools while retaining full control over all parts of the process, treating every data as files, and using Data Version Control to handle their execution in the correct order and to avoid redundant calculations on the unchanged stages. Hoseini and Quix (p. 27) describe SEDAR, a semantic data lake for the integration of heterogeneous datasets and machine learning (ML). SEDAR's key element is semantic metadata management. The system's generic ingestion interface can deal with any external data source and incorporates data capture with data versioning and automatic metadata extraction, while ML-related artefacts are embedded into the lake to allow for the coherent development of data preparation and ML pipelines. Ballhausen et al. (p. 33) present a new approach to managing high-quality digital cultural content throughout its life cycle, called curatorial companionship. The approach combines domain knowledge from diverse fields to refine and select digital cultural artefacts, and adopts a structured, iterative process for the creation of art by generative AI. Hummel et al. (p. 29) present a research agenda on network intelligence to support new digitised applications while also being sustainable. The paper discusses the novel networked applications, network intelligence methods, and major challenges in the field. Trasarti et al. (p. 30) present a research infrastructure that provides data and facilities to researchers, and services to firms and public administrations to develop tools based on ethics and fairness principles. It encourages interdisciplinary studies and promotes European principles in social analysis, offering an innovative and free platform that combines AI and social issues to perform large-scale social mining experiments within a legal and ethical framework of responsible data science. Renault and Hitzelberger (p. 32) describe the establishment of a high-performance data analytics and AI testing facility in Luxembourg with the aim of supporting digitisation and Industry 4.0 projects. The facility, which is based on research and technology infrastructure, provides a test-before-invest approach that allows companies to determine the worthiness of the technology for their specific business purposes, tailor their AI and data analytics projects, and make informed investment decisions.

Data Management and Governance

Massa et al. (p. 35) propose a data-aware and declarative solution for determining service-based application placements over the Cloud-IoT continuum while meeting functional and non-functional application requirements. The solution considers the characteristics of the data processed by the application, such as security needs, volume, velocity, transmission rate, and sources and targets, and uses a continuous reasoning approach to reduce the size of the placement problem instances at runtime. Vinju (p. 14) introduces the Rascal metaprogramming language as a solution to the need for up-to-date, easy-to-use and easy-to-combine instruments for collecting data about software and the software development processes. Maillot et al. (p. 36) present the IndeGx framework that is designed to create an index of knowledge graphs in the form of linked open datasets and provide descriptions of them for humans and machines to understand their content, quality and compliance with standards. Those descriptions are generated by extraction from a SPARQL endpoint and represented in RDF, providing a transparent, declarative, collaborative and extensible framework to be used in various use cases. Stefanidis et al. (p. 38) present the federated data sharing/trading and monetisation platform for secure, trusted and controlled exchange and usage of proprietary data assets and data-driven intelligence. The platform employs federated data discovery, distributed ledger technologies, data non-fungible tokens and AI-driven data quality assessment to build trust among data providers, data owners and data consumers. Marazakis and Louloudakis (p. 13) highlight the contribution of the RISER project, which aims to develop the first all-European RISC-V cloud server infrastructure, enhancing Europe's strategic autonomy in open-source technologies. RISER will leverage and validate open hardware high-speed interfaces and a fully featured operating system environment, enabling the integration of low-power components, including RISC-V processor chips, in an energy-efficient cloud architecture.

In conclusion, this special theme highlights the advancements made in the field of data management through exploration of novel research and technologies. The projects showcased in this issue contribute to the development of effective data management environments and techniques that are necessary for addressing the increasing complexity of data generated in today's world. The papers provide valuable insights and promote further research on data management and infrastructure carried out in Europe.

Please contact:
Manolis Terrovitis
ATHENA RC, Greece
mter@athenarc.gr

Dominik Ślęzak
University of Warsaw, Poland
slezak@mimuw.edu.pl

D4SCIENCE: A Unique Infrastructure Delivering Virtual Research Environments as a Service

by Leonardo Candela (CNR-ISTI), Donatella Castelli (CNR-ISTI) and Pasquale Pagano (CNR-ISTI)

Nowadays, research challenges – often based on the collaborative analysis of a large amount of data – require suitable infrastructures and user-facing solutions promoting multidisciplinary collaboration and appropriate communication and sharing of data, processes, and outcomes. The D4Science infrastructure and its virtual research environments proved to be a viable and effective solution for many communities of practice and use cases.

Several initiatives are ongoing to develop infrastructures for supporting designated communities by providing them with data and computing capacity. D4Science [L1] is an infrastructure specifically conceived to deliver virtual research environments (VREs) via the as-a-Service paradigm [1,2]. Its development started 18 years ago as a testbed, and over time it progressed towards an operational infrastructure through the support of a series of EU Commission-funded projects.

D4Science-based VREs are web-based, community-oriented, collaborative, user-friendly, open-science-enabler working environments for scientists and practitioners willing to work together to perform a specific (research) task.

Each VRE manifests in a unifying web application hosted in a web gateway (and a set of application programming interfaces (APIs)). The application comprises several components made available by portlets organised in custom pages and menu items running in a plain web browser. Every component aims to provide VRE users with facilities implemented by relying on one or more services, possibly provisioned by diverse providers. Every VRE offers seamless access to the datasets and services of interest for the designated community while hiding the diversities originating from various resource providers.

Among the facilities each VRE offers, some basic ones allow VRE users to perform their tasks collaboratively, namely: (a) a workspace component to organise and share any digital artefact of interest, (b) a social networking component to communicate with co-workers by posts and replies, (c) a data analytics platform to share and execute analytics methods by relying on a distributed and scalable computing infrastructure, and (d) a catalogue component to document and publish any digital artefact worth sharing.

Along its lifetime, the D4Science infrastructure has supported the delivery of VREs for very diverse communities and usage scenarios. The creation of these VREs has been a continuous process. While some VREs were dismissed upon the completion of the activity that had motivated their deployment, others have been maintained to support continuously ongoing activities.

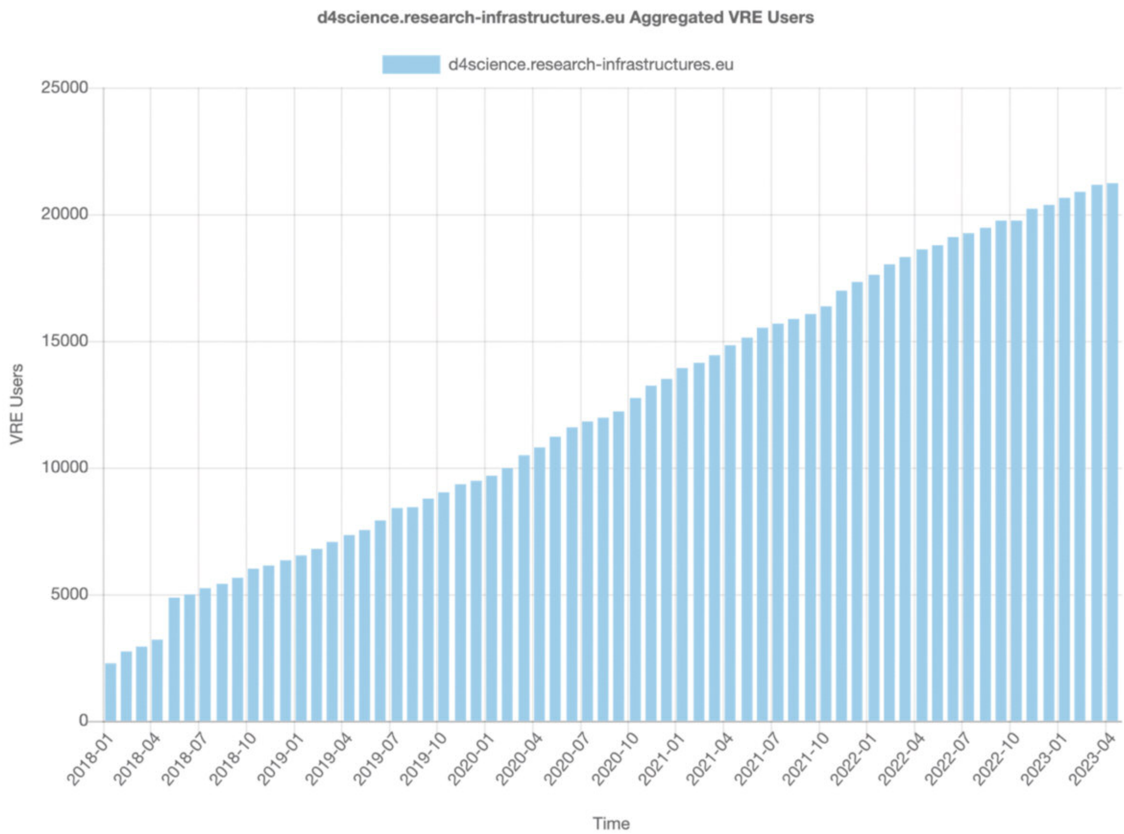

Figure 1: D4Science Users.

Currently D4Science operates 187 VREs (with others to come) made available by 20 thematic gateways [L2]. These environments support the activities and tasks of diverse communities of practice centred around organisations like FAO, ESFRI Research Infrastructures, EU and national projects, and operating in different thematic areas from marine science to social science, humanities, agri-food, health, and geothermal science. The D4Science user base counts over 20,000 users from almost all over the world. Figure 1 displays the constantly growing trend of the user base in the past five years.

From the analysis of the requests and feedback received from this large user base it emerges that the following key features make the D4Science infrastructure a unique and effective solution for many use cases.

The VREs as-a-Service delivery approach is, to a large extent, the most relevant feature of the D4Science solution for most user communities. Most do not have the necessary resources and personnel to deploy, host and maintain such environments. Often they are also looking for solutions to help them to minimise the "time-to-market", i.e. the time in which they can start using the VRE to support their specific activities. D4Science implements a VRE distribution model in which it hosts the whole application and provides it to users over the Internet as-a-Service [1]. The advantage of this design choice is that the actual management of the IT solution is in the hands of expert operators who manage it by providing reliable services, leveraging economies of scale, and using elastic approaches to scale. A new VRE is created by using a wizard to select the VRE's functional constituents among those available. The software components' deployment and configuration implementing the selected functionalities are completely automatic. It leads to a new and ready-to-use VRE made available through one of the gateways operated by D4Science.

The system of systems approach is paramount to promote the establishment of synergies with several service providers and to enlarge the capacity and service offering exploitable when creating and operating VREs. In fact, D4Science was designed to conceptually play the role of a central hub offering seamless access to its own resources (datasets, services, computing and storage capacity) as well as to services and computing capacity offered by other infrastructures and service providers. All the resources aggregated by the federated service providers are registered into a unifying information system, monitored, and exploited on demand to contribute to the creation and operation of the various VREs.

Catering for co-creation is paramount to guarantee community uptake and the incremental evolution of the VRE to meet the designated community's changing needs. Communities of practice have evolving needs and often refine their requirements when using the provided working environments and services. VREs cannot be static environments; they must evolve, making available new tools and datasets to meet emerging needs. D4Science supports integration patterns [2] to complement the offering and bring new resources into VREs by facilitating the incorporation of community-specific existing applications, analytics methods and workflows, datasets and other resources for discovery and access. This co-creation mechanism counts on the presence of a working version of the VRE where community resources are "hot-plugged" without stopping or shutting down the environment.

Open Science is here to stay, yet it must be supported by an open access approach "as early as convenient". It implies collaboration and sharing, reproducibility and transparency to as wide and great an extent as possible. Scientific communities willing to operate in line with this approach have found in the D4Science VREs concrete support for flexibly meeting these properties and implementing Open Science practices with the needed shades of openness. D4Science VREs are equipped with basic services supporting collaboration and cooperation among their users. They also continuously and transparently capture research activities, authors and contributors, as well as every by-product resulting from every phase of a typical research life cycle, thus offering a solid base for addressing Open Science practices like, for example, reproducibility, research assessment, communication and collaboration, and transparency.

At the current stage, we can state that this solution has many advantages, as demonstrated by its high uptake. The VREs as-a-Service represents, for many communities (especially communities of practice in long-tail science), the ideal solution for solving the need for their collaborative activities, especially when these are data-driven and computationally intensive and go beyond the boundaries of institutions and regions. Indeed, they largely reduce the time for a community of practice to become operational and the need for skilled personnel dedicated to technology development.

Links:
[L1] https://www.d4science.org
[L2] https://services.d4science.org/thematic-gateways

References:

[1] M. Assante et al., "Enacting open science by D4Science", in Future Gener Comput Syst, vol. 101, pp. 555-563, 2019, doi:10.1016/j.future.2019.05.063.
[2] M. Assante et al., "Virtual research environments co-creation: The D4Science experience", in Concurrency Computat Pract Exper, e6925, 2022, doi:10.1002/cpe.6925.

Please contact:

Leonardo Candela, CNR-ISTI, Italy
leonardo.candela@isti.cnr.it

ARIADNE: A Data Infrastructure for the Archaeological Research Community

by Alessia Bardi, Massimiliano Assante and Francesco Mangiacrapa (CNR-ISTI)

The ARIADNE infrastructure provides tools and services for researchers to address archaeological grand challenges that require discovery and analysis of information scattered across different thematic and geographically distributed sources.

ARIADNE (Advanced Research Infrastructure for Archaeological Dataset Networking in Europe) is an infrastructure for the archaeological research community funded by the European Commission via the projects ARIADNE (from 2013 to 2017) and ARIADNEplus (from 2019 to 2022). [L1] The infrastructure offers a suite of services and tools to support research in archaeology in different phases of the research life cycle: tools for data visualisation (e.g. 3D modelling, geoportal); virtual research environments (VREs) for data analysis, data sharing and human-centred collaboration; and a knowledge base of archaeological resources, accessible as Linked Open Data and searchable via a web portal.

At the end of the ARIADNEplus project (December 2022), the 41 partners [L2] successfully extended the scope of the first ARIADNE project, delivering a digital infrastructure based on the D4Science e-infrastructure [3] developed and maintained by CNR-ISTI, with: (i) a knowledge base of 3.5 million archaeological resources including archaeological reports, findings, inscriptions, archaeological sites and monuments from archives and repositories in Europe and beyond (e.g. Argentina, Japan); and (ii) a set of Virtual research Environments (VREs) where archaeological researchers can find different tools for data exploration and analysis.

The knowledge base was built via an aggregation system, capable of collecting detailed descriptions of archaeological resources (metadata records) in different formats, transforming the records according to a common data model based on the CIDOC-CRM ontology, enriching them with dating information via PeriodO [L3] and subject terms via Getty AAT [L4]. For each dataset, an automated workflow has been devised via the D-NET Software Toolkit, a software for the realisation of aggregative metadata infrastructures developed by CNR-ISTI [1], and configured to apply a mapping defined with the 3M Editor developed by FORTH-ICS [2]. As a result of the mapping, each input metadata record is transformed into an RDF (Resource Description Framework) record suitable for ingestion into the ARIADNE knowledge base, an instance of GraphDB [L5]. For the enrichment with dating information, providers were invited to curate an authority on PeriodO, while for the enrichment with Getty AAT, they defined a mapping between their local terms and Getty AAT terms with the Vocabulary Matching Tool developed by University of South-Wales [L6]. SPARQL queries were then used to exploit such information to enrich the records in the knowledge base. The final step of the automated workflow is the publishing on the ARIADNE portal [L7], developed by the Swedish National Data Service, where the resources can be searched and filtered by different criteria (e.g. by location, by historical period, by subject, by contributor).

The VREs provided by ARIADNE are designed to meet the needs of its target community and scenarios. These environments are dedicated to exploring and developing various solutions for specific research questions and serving specific communities of practitioners in the archaeology domain. For instance, the Analytics Lab VRE [L8] offers archaeologists and scholars a virtual laboratory and set of tools for analysing and manipulating data to answer their research questions. The VRE allows users to import their own data files and those from the ARIADNE knowledge base, rearrange and analyse them, and generate statistical reports based on typology, geographical distribution and other factors. The results can also be visualised in the form of complex graphs and thematic maps. The D4GNA VRE [L9] (Dataset for the National Archaeology Geoportal) instead collects data intended to converge into the Italian National Archaeology Geoportal (GNA) under the auspices of the Italian Ministry of Culture. As Figure 1 shows, it allows any user to freely access [L10] and consult documentation related to archaeological excavations in Italy, and authorised users to collaboratively manage these contents. Users can

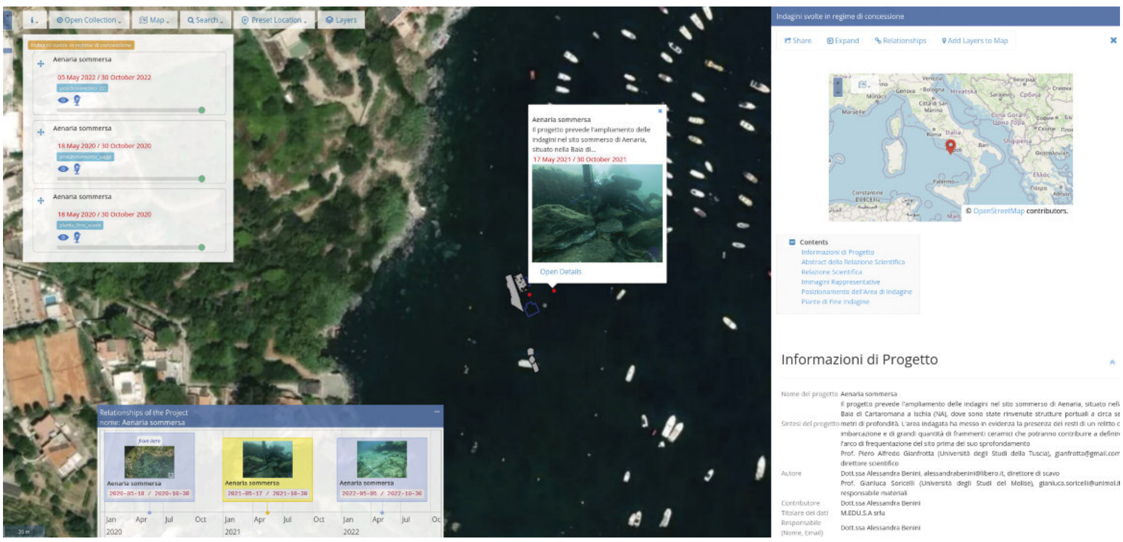

Figure 1: The geoportal viewer of D4GNA VRE.

discover the products either by interacting with the web-map or by searching for products by keywords, author, date and other metadata, and preview the products in a variety of formats, including maps, images and documents. This VRE is equipped with a dedicated service (geoportal) acting as the underlying storage and management layer for the products and their metadata, responsible for ensuring the persistence and accessibility of the data, and for enforcing the access control policies defined by the system administrators. Its data model is rich and flexible, designed to support the discovery and sharing of the products, and to allow users to easily link and connect products based on their metadata and contents.

In conclusion, the ARIADNE infrastructure provides a comprehensive solution for archaeological research, including a range of services and tools to help researchers throughout the research life cycle. From creating data management plans using templates, to visualising data using 3D modelling and geoportal tools, this offer is complemented with VREs designed to be user-friendly and accessible to users of all levels, whether they are experts or novice researchers. This allows researchers to focus on their work, rather than struggling with complex technology. Researchers can store their data securely and share it with their collaborators with ease, knowing that their data is protected by the highest levels of security and privacy. This helps to foster collaboration and innovation in the research community, as researchers can work together to tackle complex problems and make new discoveries. The knowledge base of archaeological resources is accessible as Linked Open Data and can be easily searched through the web portal, ensuring researchers have access to the resources they need to succeed.

Links:
[L1] https://ariadne-infrastructure.eu/
[L2] https://ariadne-infrastructure.eu/partners/
[L3] https://perio.do/en/
[L4] https://kwz.me/hwW
[L5] https://graphdb.ontotext.com/
[L6] https://vmt.ariadne.d4science.org/vmt/vmt-app.html
[L7] https://ariadne-portal.d4science.org/
[L8] https://ariadne-infrastructure.eu/the-ariadneplus-lab-vre/
[L9] https://gna.d4science.org/
[L10] https://gna.d4science.org/d4gna-data-viewer

References:

[1] P. Manghi, et al., "The D-NET Software Toolkit: A framework for the realization, maintenance, and operation of aggregative infrastructures", Program Electron Lib, vol. 48, no. 4, pp. 322–354, 2014. https://doi.org/10.1108/PROG-08-2013-0045
[2] Y. Marketakis, et al., "X3ML mapping framework for information integration in cultural heritage and beyond", Int. J. Digit. Libr., vol. 18, pp. 301–319, 2017. https://doi.org/10.1007/s00799-016-0179-1
[3] M. Assante, et al., "Enacting open science by D4Science", Future Generation Computer Systems, vol. 101, pp. 555–563, 2019. https://doi.org/10.1016/j.future.2019.05.063

Please contact:

Alessia Bardi, CNR-ISTI, Italy
alessia.bardi@isti.cnr.it

Boosting Open Science in the IPERION HS Research Infrastructure with OpenAIRE

by Alessia Bardi (CNR-ISTI) and Laura Benassi (CNR-INO)

IPERION HS is a research infrastructure that supports researchers in the field of heritage science, an interdisciplinary domain studying cultural and natural tangible heritage. This article describes how IPERION HS makes the research outputs open and accessible to the community and monitors its impact thanks to the services offered by the OpenAIRE infrastructure.

IPERION HS [L1] is a European infrastructure project working in the interdisciplinary domain of heritage science to understand, study and protect natural and cultural heritage. It contributes to establishing a pan-European research infrastructure on heritage science (E-RIHS). Led by the Italian National Research Council (National Institute of Optics), the consortium is composed of 24 national nodes across Europe and the US and collaborates with ICCROM, the intergovernmental organisation dedicated to the preservation of cultural heritage worldwide.

IPERION HS offers training and access to a wide range of high-level scientific instruments, methodologies, data and tools for advancing knowledge and innovation in heritage science. The research is actually focused on a better interoperability to data, sample materials, methods and instruments. Together with the users, the IPERION HS community is collecting an incredible amount of data and samples and working on the development of new protocols and instruments.

Making heritage science data open, findable and accessible is one of the main objectives of the IPERION HS infrastructure, whose Open Science strategy does not only include open access to publications, but also FAIR (findable, accessible, interoperable and reusable) and open data and methodologies. Tracking and monitoring the success of the strategy calls for technical tools capable of gathering together the research outputs of the IPERION researchers, which are published and deposited across different venues and repositories, and analysing the gathered data according to different perspectives, such as the types of outputs, their access rights, the policies of the journals they are published in, the availability of persistent identifiers and rich descriptive metadata, and the affiliations of the authors.

In addition, IPERION HS aims to support and ease the discovery of heritage science resources. Due to the high multidisciplinary aspect of the field, heritage science research resources are scattered in different repositories, archives, journals across the world and across many disciplines. Heritage science researchers, therefore, do not have a single web portal for searching for heritage science resources. On the contrary, they have to use many different thematic portals or the portals of large domain-agnostic aggregators. In the first case, researchers are likely to miss relevant sources; in the second they are likely to be overwhelmed by resources that are not interesting for them.

To address those challenges with a holistic approach, a new collaboration with OpenAIRE was launched in June 2022. OpenAIRE [L2] is a European infrastructure that promotes and supports the adoption of Open Science principles. It is composed of a distributed network of Open Science support desks and of a digital infrastructure with services that help institutions, research initiatives, communities and single researchers at embedding Open Science practices in all the different phases of the research life cycle. In particular, OpenAIRE is providing to IPERION HS three main services to strengthen, track and monitor the Open Science strategy of IPERION HS: Zenodo [L3], Connect [L4] [1] and Monitor [L5]. With Zenodo, IPERION HS offers to its researchers a trusted location where any type of research resources can be stored in a FAIR way. With Connect, IPERION HS is setting up an Open Research Gateway on Heritage Science [L6], where all research outputs of the domain are made discoverable, enabling searching and browsing across different sources (see Fig. 1). In addition, researchers can help grow the record of their community by adding links among research products and other entities of the research life cycle (e.g. this publication is supplemented by this dataset, the research described in that journal article has been funded by that project grant). The content available in the gateway is automatically updated by OpenAIRE based on a set of configuration criteria defined by gateway curators, who are members of IPERION and experts of the heritage science domain. In particular, gateway curators can select:
- domain-specific projects
- thematic data sources (e.g. journals and repositories)
- thematic Zenodo communities
- subject classification terms (free text keywords or terms from controlled vocabularies)
- affiliations.

OpenAIRE further assigns research products to the community based on the existing semantic relationships. For example, if a journal article is relevant for the community and is supplemented by a dataset, then the dataset is also added to the community. In order to add products that are relevant for the community but cannot be automatically detected as such, users can manually add them to the gateway with the Link functionality.

With a similar approach, OpenAIRE can identify the research outputs that have been produced thanks to the IPERION HS infrastructure and analyse the metadata to provide statistics and indicators about the uptake of Open Science, the compliance of its researchers to the Open Science policies, and the impact of the infrastructure on the research landscape.

The pilot collaboration between IPERION HS and OpenAIRE in the field of heritage science and the tools set up for practising Open Science will be the legacy to be left to E-RIHS (the European Research Infrastructure for Heritage Science), that will be established in the next few years with the headquarters in Florence, Italy. The heritage science community is facing the challenges of interoperability of research data and is building a reliable approach to describe, store, preserve and share the results globally. The operative support of OpenAIRE is becoming crucial for consolidating it.

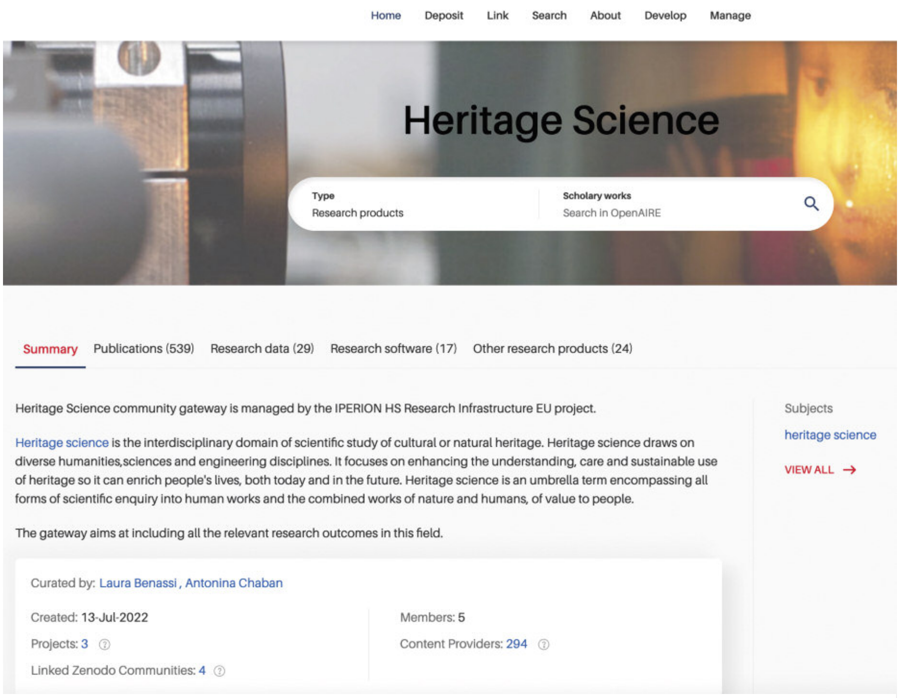

Figure 1: Home page of the Open Research Gateway on Heritage Science.

Links:
[L1] https://www.iperionhs.eu/
[L2] www.openaire.eu
[L3] https://zenodo.org/communities/871034
[L4] https://connect.openaire.eu
[L5] https://monitor.openaire.eu
[L6] https://heritage-science.openaire.eu/

References:
[1] M. Baglioni, et al., "The OpenAIRE Research Community Dashboard: On Blending Scientific Workflows and Scientific Publishing", in Proc. TPDL 2019 in LNCS vol. 11799, pp. 56–69, Springer, Ed. https://doi.org/10.1007/978-3-030-30760-8_5

Please contact:
Alessia Bardi , CNR-ISTI, Italy
alessia.bardi@isti.cnr.it
Laura Benassi , CNR-INO, Italy
laura.benassi@cnr.it

The Data Portal of the European Plate Observing System – A New Tool for the Solid-earth Sciences in Europe

by Vasco Avramo (EPOS-ERIC), Daniele Bailo (EPOS-ERIC, INGV) and Rossana Paciello (EPOS-ERIC, INGV)

EPOS, the European Plate Observing System, is currently the only research infrastructure for solid-earth sciences in Europe. It integrates a multitude of different data, metadata, software and services through one single portal, the EPOS Data Portal, released on January 1st 2023.

During the past decade, data infrastructures have become irreplaceable tools for a great variety of stakeholders, such as the scientific community, governments and the private sector. The availability of high-quality, standardised and secure data covering a great diversity of domains is providing a new perspective on the capability of human society to understand and manage the world around us. Solid-earth science data represent the knowledge base to unravel the physical and chemical processes that control earthquakes, tsunamis, volcanic eruptions and all processes driving tectonics dynamics. The Earth is an interconnected system and understanding how it works is crucial for modern society safety and wellness. EPOS is currently the only research infrastructure for solid-earth sciences in Europe that integrates diverse European Research Infrastructures under a common federated framework (Fig. 1).

The conception of EPOS started two decades ago with the collaboration of a great number of countries, national earth science institutions and projects. Until the 2018, EPOS has gone through the Preparatory Phase and the Implementation Phase, supported by European fundings. The main goal was to bring the infrastructure to a governance and technical maturity required to develop thematic and integrated services. On October 30th 2018, the European Commission granted the legal status of ERIC (the European Research Infrastructure Consortium) and, after a Pilot Operation Phase (EPOS POP), EPOS is officially entering its Operational Phase, with the public release of all services and functionality [1].

The key challenge of this system is ensuring sustainable and universal use and re-use of multidisciplinary solid-earth science data and products and fostering state-of-the-art research, innovation and human progress. All the infrastructure is built around the FAIR (findable, accessible, interoperable, reusable) principles, with the aim of harmonising, integrating and making accessible all the available data and metadata concerning solid-earth science. This concept, despite strong support from the international scientific community, is not yet implemented by all scientific institutions and infrastructure, and one of the EPOS objectives is to promote adoption of FAIR by the data providers [2].

EPOS architecture is a complex enterprise and has been planned to work as a single, but distributed, sustainable research infrastructure. The distinctive feature of the architecture is that it is designed by assembling complementary components (Fig. 2): the National Research Infrastructures (NRIs), where data are generated, processed, analysed, and archived, represent the foundation that sustains the EPOS Research Infrastructure; the Thematic Core Services (TCS), the community-specific integration component ensuring high-quality, standardised data and service, and the Integrated Core Services (Central Hub, ICS-C and Distributed, ICS-D), the place where actual integration of data and services occurs. Currently, EPOS has 10 TCS (Fig. 3), representing a wide spectrum of disciplines within the solid-earth domain.

Such an architecture enabled EPOS to provide integrated access to a multitude of different data, data products, software and services through one single portal, the EPOS Data Portal [3]. It is based on the Integrated Core Services system that, by pursuing a threefold approach based on metadata, semantic

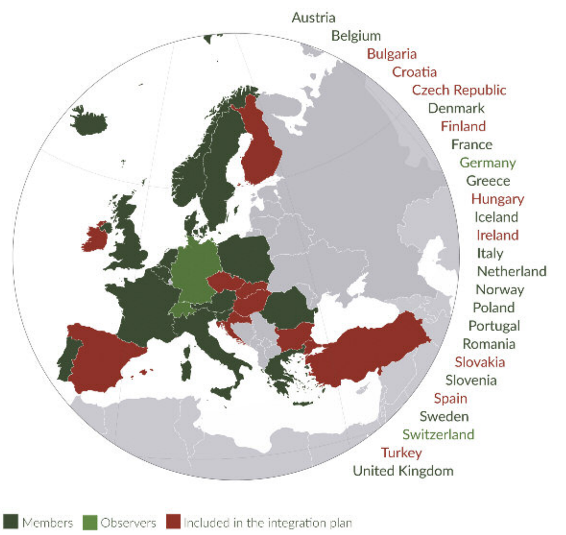

Figure 1: EPOS members (dark green), observers (light green), and countries included in the integration plan but not in EPOS-ERIC (red).

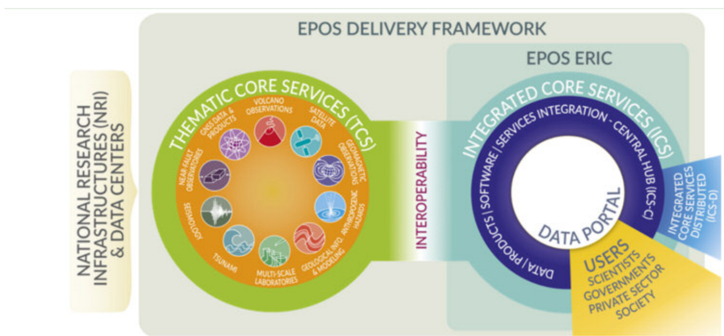

Figure 2: EPOS architecture.

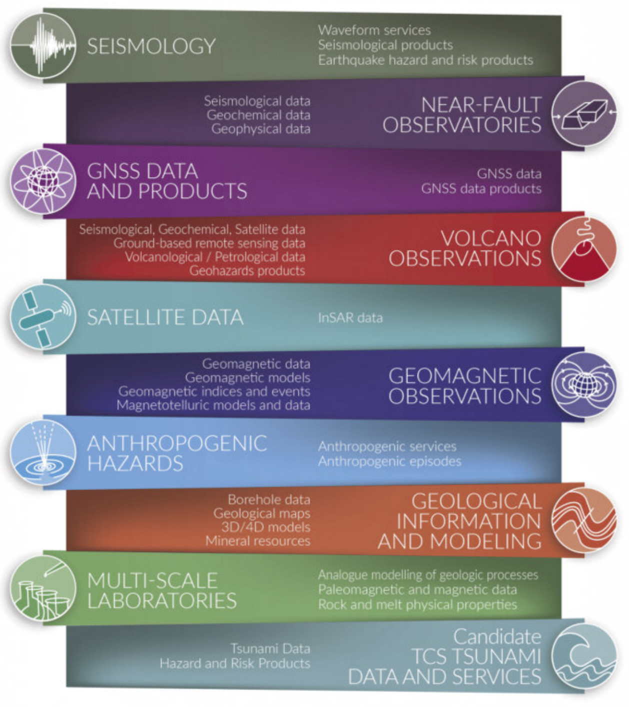

Figure 3: Thematic Core Services (TCS).

and service-based architecture, provides homogeneous access to heterogeneous resources. One of the key elements for the EPOS Data Portal provision, operation and maintenance beyond the technical dimension, is community engagement and management dimension. In this framework several activities and events are regularly carried out to guarantee the effective engagement of the communities that have the required skills for delivering the data and service provision, while ensuring the coherency and the effectiveness of the legal, governance and technical framework in EPOS.

On January 1st 2023, the data portal with 250 fully operational data services was released [L1] (Fig. 4), 20 years after its conception. This provides an insight on the complexity of the work required to set up EPOS Research Infrastructure as the unique research infrastructure for solid-earth science in Europe. During these years, indeed, more than 600 experts worked together with 100 IT developers, and decision makers from all the involved countries to integrate and harmonise more than 60 different data types. The results of such huge community effort will be presented for the first time at the European Geosciences Union General Assembly from the 24 to the 28 of April in Vienna [L2]. At this event, EPOS will be present with a dedicated booth, where it will be possible to physically test the portal and follow many different educational and disseminating initiatives. We consider this appointment an important opportunity to make the scientific community aware, together with institutional and private stakeholders, of this leading-edge tool supporting the solid-earth community in tackling future scientific challenges.

Links:

[L1] https://www.epos-eu.org/dataportal
[L2] https://www.egu23.eu/

References:

[1] M. Cocco, et al., "The EPOS Research Infrastructure: a federated approach to integrate solid Earth science data and services", Annals of Geophysics, vol. 65, no. 2, 2022.
[2] D. Bailo, et al., "Perspectives on the Implementation of FAIR Principles in Solid Earth Research Infrastructures", Frontiers in Earth Science, vol. 8, 2020.
[3] D. Bailo,, et al. "Data integration and FAIR data management in Solid Earth Science." Annals of Geophysics, 2022.

Please contact:

Vasco Avramo, EPOS-ERIC, Italy
vasco.avramo@epos-eric.eu

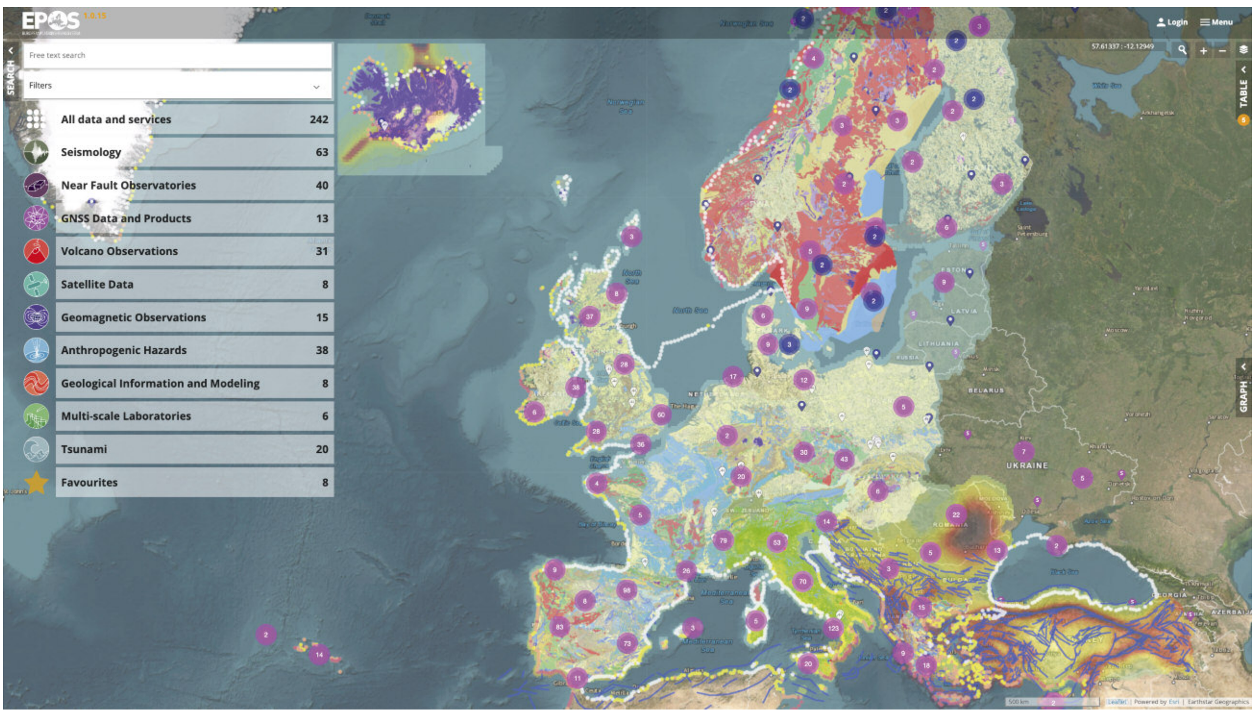

Figure 4: EPOS Data Portal functions and services.

RISER: The First All-European RISC-V Cloud Server Infrastructure

by Manolis Marazakis and Stelios Louloudakis (ICS-FORTH)

The European cloud market is estimated to grow from EUR 32 billion to EUR 127 billion from 2020 to 2026, playing a key role in the European economic growth and social development. RISER will develop the first all-European RISC-V cloud server infrastructure, significantly enhancing Europe's strategic autonomy in open-source technologies.

Cloud technologies with an important growth and economic impact include multi-core accelerators and high-speed communication infrastructures, with an expected Compound Annual Growth Rate (CAGR) of 24.7% from 2022 to 2027 [1]. Cloud infrastructures are expected to serve more than 50 billion users by 2025. Europe clearly needs an advanced energy-efficient EU-based cloud solution driven by EU technology and infrastructures. Thus, the development and long-term evolution of general-purpose and accelerator processors, together with their associated set of hardware and systems software support components, has been widely recognised as a crucial element of open strategic autonomy for Europe [2]. Expanding and validating open hardware interfaces, coupled with making available a fully featured operating system environment and runtime system, enables a pathway of innovation in several domains, including cloud services, which are so widely adopted that they have become an essential part of everyday life. All of the aspects described above, led to the conceptualisation of the RISER project [L1].

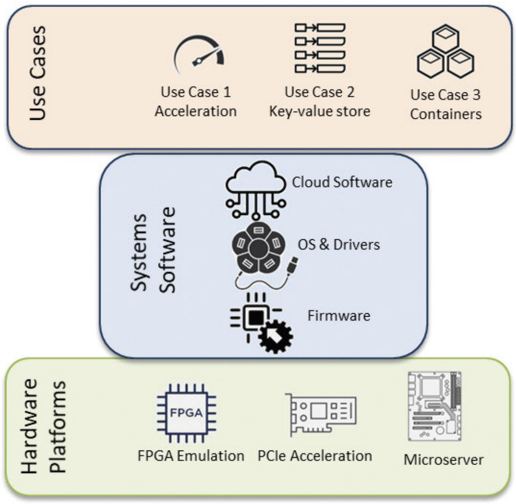

Figure 1: RISER Architecture logical layers.

RISER will develop the first all-European RISC-V cloud server infrastructure, significantly enhancing Europe's strategic autonomy in open-source technologies.

RISER will leverage and validate open hardware high-speed interfaces combined with a fully featured operating system environment and runtime system, enabling the integration of low-power components, including RISC-V processor chips from the EPI [L2] and EUPILOT [L3] projects, in a novel energy-efficient cloud architecture. RISER brings together seven partners from industry and academia to jointly develop and

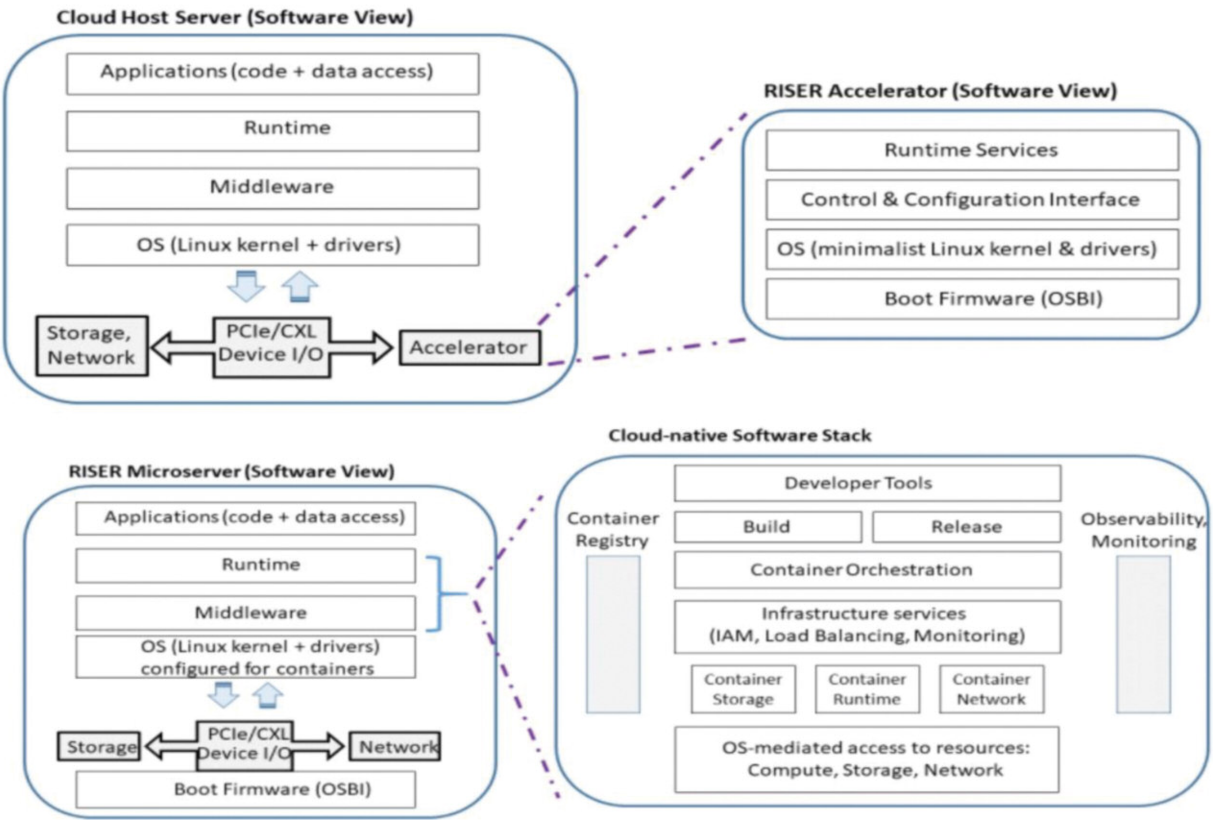

Figure 2: System software for the two RISER platforms: accelerator interacting with host, and microserver with cloud software stack

validate open-source designs for standardised form-factor system platforms, suitable for supporting cloud services. Specifically, RISER will build two cloud-focused platforms:

1. An accelerator platform, which includes the Arm RHEA SoC from EPI and a PCIe acceleration board to be developed within the project, which will integrate up to four RISC-V based chips from EUPILOT.
2. A microserver platform, interconnecting up to ten microserver boards all developed by the project, each one supporting up to four RISC-V chips coupled with high-speed storage and networking. Embracing hyperconvergence, our microserver architecture allows for distributed storage and memory to be used by any processor in the system with low overhead. The open-source system board designs of RISER will be accompanied by open-source low-level firmware and systems software, and a representative Linux-based software stack to support cloud services, facilitating uptake and enhancing the commercialisation path of project results.

Three use cases will be developed to evaluate and demonstrate the capabilities of RISER platforms:
a) Acceleration of compute workloads
b) Networked object and key-value storage
c) Containerised execution as part of a provider-managed IaaS environment.

RISER has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101092993. The project is funded under the European Health and Digital Executive Agency (HaDEA) call on "Digital and emerging technologies for competitiveness and fit for the green deal" [L4].

Links:
[L1] https://www.riser-project.eu
[L2] https://www.european-processor-initiative.eu
[L3] https://eupilot.eu
[L4] https://bit.ly/HaDEA_call_digital-Green-Deal

References:
[1] Data Center Accelerator Market by Processor Type (CPU, GPU, FPGA, ASIC), Type (HPC Data Center, Cloud Data Center), Application (Deep Learning Training, Public Cloud Interface, Enterprise Interface) and Region - Global Forecast to 2027 (Report by GlobeNewsWire Inc, November 2022). https://bit.ly/401obA5
[2] EU strategic autonomy 2013-2023: From concept to capacity (European Parliament briefing, July 2022). https://bit.ly/43w3hME

Please contact:
Manolis Marazakis, ICS-FORTH, Greece
maraz@ics.forth.gr

Stelios Louloudakis, ICS-FORTH, Greece
slouloudak@ics.forth.gr

Open and Shared Infrastructure for Software Research

by Jurgen Vinju (CWI and TU Eindhoven)

Research software engineers, scientific programmers, PhD students and postdocs alike spend their time and energy on engineering the software for their research infrastructure. Sometimes this can be reused in follow-up research projects, but most often valuable software output falls like trees in the forest, without an audience, and without users. Surprisingly, even in the academic field of software engineering, research software infrastructure is not sustained, leaving a wide gap of opportunity for more reuse and more impact.

Software has penetrated every aspect of society to the point where it has become critical to its day-to-day functioning. We must learn to better understand software: how to construct it, maintain it, check it and control it. Based on an increased understanding we can learn to better control the risk factors of software (from financial risk to personal safety risk) and we can learn to innovate with higher quality and more agility.

Understanding the complexity of software requires excellent observational instruments that enable top-quality empirical research methods. Since 2009 the Rascal metaprogramming language [L1] has become an infrastructure for empirical research in software engineering. By providing both language-agnostic and language-specific intermediate data representations of source code and data around software processes (versions, issues, discussion), new research was enabled that covers programming languages such as Lua, PHP, Java, C and C++ in a familiar and consistent environment. The development of Rascal and its core supporting language front-ends was done in the context of both national European collaborations such as FP7 OSSMETER and H2020 CROSSMINER and NWO MERITS.

However, there is ample room for growth and improvement. The field is still hampered by a lack of up-to-date, easy-to-use, and easy-to-combine (integrated) instruments for collecting data about software and the software development process. The existing instruments that do exist are scattered, isolated and incompatible. Therefore, we are extending the Rascal platform (and community) for software analysis and transformation with many new necessary high-quality data sources and accurate instruments for software data acquisition and analysis. By design, these instruments will be easy to integrate in order to link data about software in new and unforeseen ways. For example, this year the Ada-air front-end was added to enable the analysis of high-tech components programmed in the Ada language.

On the one hand, all these instruments are comparable to radio telescopes: they acquire the raw data which are essential for knowledge discovery. On the other hand, dissecting software is more like microbiology: every little piece has a different shape and size and requires specialised instruments. Even

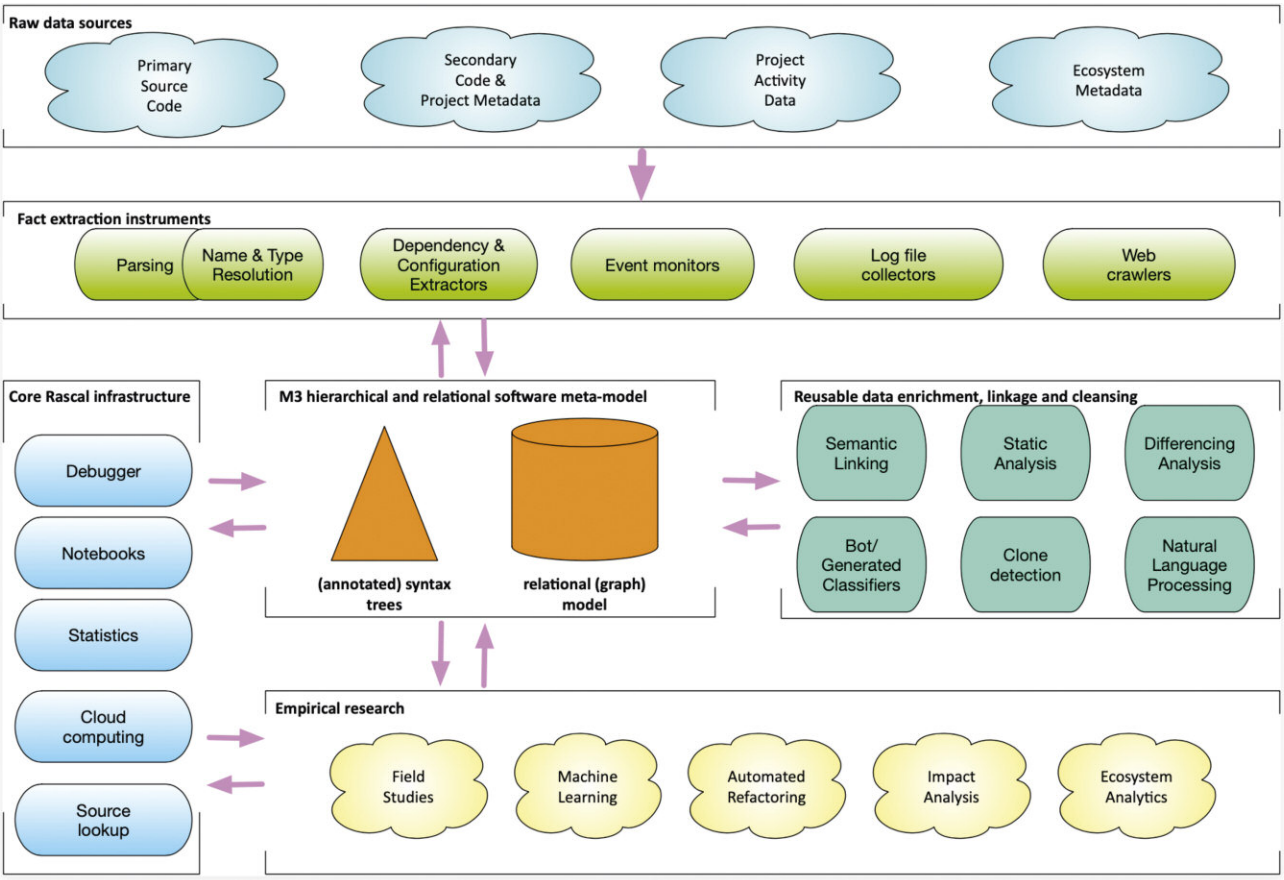

Figure 1: The Rascal Lab.

though these science analogies fail to explain the whole story, the need to learn from accurate observations of software is no less urgent. Software has grown exponentially in the last decades, making manual analysis impossible.

The automated extraction of data from source code and information about the software development process such as version control systems, discussion forums and design diagrams is burdened by countless heterogeneous details. Furthermore, such data acquisition instruments must avoid error and bias, and the information from different data sources has also to be linked. Relative to the downstream research, validated and well-integrated instruments are very expensive to obtain. Moreover, new programming languages, frameworks, and collaboration tools emerge at an alarming rate. Researchers can easily spend more than 50% of their time building and connecting their research instruments. Making high-quality software data acquisition and linking instruments reusable for a wide audience of researchers presents an opportunity to increase the impact of software research.

The "Rascal Lab", as described above and depicted in Figure 1, is a vision of a more complete and sustainable laboratory. It is based on the existing Rascal metaprogramming community [1] and supported by a broad consortium of researchers from all Dutch universities. The lab's vision consists of the construction of many more reusable and accurate data acquisition instruments, the acquisition of several (curated) corpora of software data using those instruments, and the instruments to link, integrate, analyse and report on the extracted data. This is an ongoing and growing effort, and we are welcoming new participants to the community regularly.

If you are interested in participating, please contact the author or connect on GitHub [L2].

Link:
[L1] http://www.rascal-mpl.org
[L2] https://github.com/usethesource/rascal

Reference:
[1] P. Klint, T. van der Storm and J. J. Vinju, "Rascal, 10 years later," in IEEE International Working Conference on Source Code Analysis and Manipulation, 2019.

Please contact:
Jurgen Vinju, CWI, The Netherlands
jurgen.vinju@cwi.nl

Creating a Privacy-aware Framework for Fine-grained Data Access

by Eliot Salant (IBM Israel Research Lab)

We have created a framework to enable secure, policy-driven data exchange, applying sophisticated, fine-grained access control to data stores. Incorporated into the EU Horizon2020 HEIR project, we have demonstrated the flexibility and power of our Privacy-aware Framework (PAF) in a number of use cases, including automatic anonymisation of data on export to a third party, and redaction of data based on user role and organisation affiliation.

The ability to exchange digital data between data owners and data requesters is critical to today's society. However, as the amounts and types of data proliferate, so do the requirements for data governance to protect the rights of the data subject – the person who can be identified by the data. Whereas in the past access to data was typically controlled by Access Control Lists (ACL) or Role Based Access Control (RBAC), the rise of new legislation such as GDPR (General Data Protection Regulation) requires a policy-driven model, such as Attribute Based Access Control (ABAC), to enforce broader conditions such as geographic or purpose of use constraints.

HEIR [L1] is a three-year EU H2020 project with fifteen partner organisations, which started in 01/09/2020. HEIR's focus is on the world of healthcare and aims both to provide cyber protection for hospital/medical centres, and a framework for securely sharing medical data.

There were many requirements behind the design of the PAF in HEIR. We needed a framework that could work with data stores such as FHIR [1] servers or healthcare registries. Fine-grained redaction at the attribute level within healthcare FHIR records was required – the FHIR standard only dictates an access classification at the resource level. Passwords to access data stores need to be securely stored, and not distributed to application developers. The transfer of potentially large amounts of data needs to be done efficiently. The data path must be securely locked down to prevent data leakage.

To help meet these goals, the HEIR PAF built on of Fybrik ([L2]), an open-source framework being developed by IBM Research. The Fybrik framework receives declarative, human-readable files to configure a secured path between the data requester and the data source, known as a Data Plane. These files would typically be created by different actors in the healthcare centre environment as shown in Figure 1.

The Data Plane constitutes a workflow, extracting data from the source, performing any policy-mandated redaction actions, and potentially transforming the data into a form required by the data requestor.

How does this work?

Before Fybrik can bring up the data path, a data access policy must be defined, data source(s) catalogued, and PII fields in the schema tagged. In our use cases, the data access policy will be based on user information. To this end, a data requestor needs to log in to an Identity and Access Management system to obtain a token (JSON Web Token), which authenticates the user and encodes information such as the user role and organisation affiliation. This token will need to accompany every request for data from the PAF.

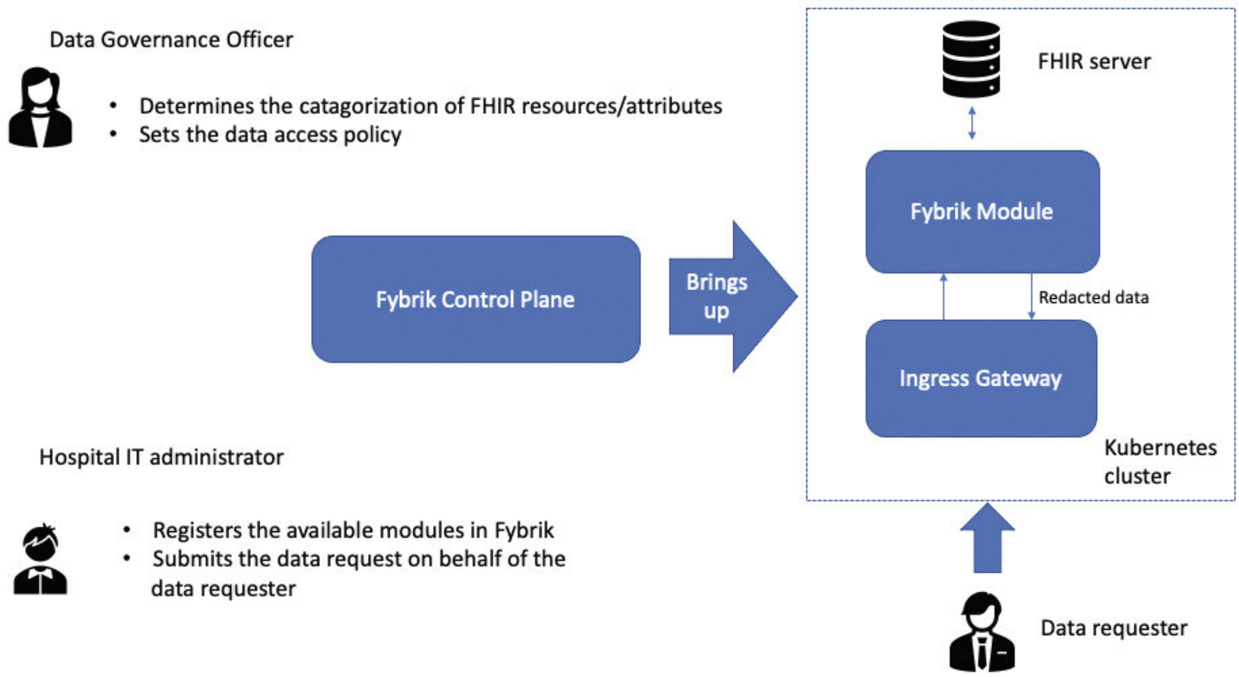

Figure 1: Use of Fybrik to create a PAF.

The PAF securely stores the access credentials to the FHIR server, keeping them hidden from application programmes. This guarantees that all access requests FHIR data must pass through the PAF, and therefore are subject to governance rules.

Data policy definition and evaluation is handled by use of the Open Policy Agent (OPA) [L3] in the Fybrik Policy Manager. OPA policies are defined in a language called Rego, and allow for evaluation of parameters supplied at runtime, such as the role of the requestor. Policy rules that match evaluation time conditions return an action, such as "RedactColumn <column list>" or "DeleteColumn <column>" which will be enforced before the data is returned to the requestor. Evaluation of policy allows the Fybrik Module to decide whether to allow an incoming request to pass to the FHIR server backend, and it then decides if the returned data needs to be transformed to support data redaction or anonymisation requirements.

In our healthcare use cases, FHIR resources are our data sources. While Fybrik can support commercial data catalogues such as IBM's Watson Knowledge Catalog, our HEIR implementation uses an internal catalogue that holds metadata to assign FHIR resource attributes with labels, such as PII (Personal Identifiable Information). Consequently, we are able to define policy rules that are specific to the attributes of the requestor (e.g. role, affiliation) that redact on the FHIR attribute level. For example, we can have a rule that will anonymise all PII information from Observation records being returned to a researcher, whereas another rule will allow patients to view all their information without redaction.

Conclusion

The PAF allows much finer-grained access to FHIR data than envisioned by the FHIR standard, significantly improving the ability to share data while conforming to policy. Additionally, policies can be dynamically determined and can factor in a wide variety of parameters beyond those associated with typical Access Based Control, such as geographic and time period constraints.

While extremely powerful in the world of healthcare, this framework can be used to protect data in virtually any other sector.

Links:
[L1] https://www.heir2020.eu/
[L2] https://fybrik.io/
[L3] https://www.openpolicyagent.org

Reference:
[1] HL7 FHIR. (n.d.). HL7 International. Retrieved from https://hl7.org/fhir/.

Please contact:
Eliot Salant, IBM Research, Israel
salant@il.ibm.com
+972-48296121

Outsourced Computations Maintaining Confidentiality and Authenticity

by Stephan Krenn, Thomas Lorünser, Sebastian Ramacher, Florian Wohner (AIT Austrian Institute of Technology)

Performing joint computations on data coming from different data owners offers great potential for more efficient resource use, cost reductions, a reduction in certain industry's carbon footprint and improved societal benefits. However, these synergies are often not exploited because stakeholders have high confidentiality requirements for business-sensitive data and are distrusting of each other. To address this gap, in a series of projects, we have developed provably privacy-preserving and verifiable protocols based on multi-party computation.

The ubiquitous availability of connectivity and cloud services offers a wide range of novel opportunities. One innovation area involves joint computations on data collected from various sources, ranging from sensors to business internal processes, in order to gain novel insights or to optimise existing processes. The potential sensitivity of the data involved in such applications makes their protection of utmost importance to successfully incentivise data owners to participate in joint computations, and thus to leverage tremendous potential.

Besides contractual agreements and legal obligation on data protection, cryptographic research on privacy-preserving federated computation has leaped forward over the last decades towards providing technical means with formal and provable security guarantees. Besides, e.g. fully homomorphic encryption, one key enabling technology is secure multi-party computation (MPC), which allows multiple parties to jointly evaluate an arbitrary function without revealing anything about the input data, beyond what can be derived from the computation result itself. In its most secure configuration, even confidentiality against computationally arbitrarily powerful attackers – including quantum attackers – can be achieved. As a result, MPC offers a provably secure alternative to trusted parties: for instance, in the case of sealed-bid auctions, instead of having to reveal secret bids depending on sensitive information about one's internal cost structure to a trusted notary, MPC allows the bidders to interactively identify the best offer. While only the tenderer obtains the final computation result, no information about the remaining bids is disclosed to any other party in the system.

One caveat of MPC is that its security relies on a non-collusion assumption: if more than a (parametrisable) number of parties collude, privacy and soundness of computation results can be jeopardised. Furthermore, the result of the computation in general cannot be verified by any party not participating in the computation, which may be necessary, e.g. for auditing or transparency reasons. These challenges are addressed by ongoing research activities on so-called verifiable MPC. There, besides the computation result, the MPC nodes also compute a cryptographic and publicly verifiable certificate in the form of a non-interactive zero-knowledge proof of knowledge, prov-

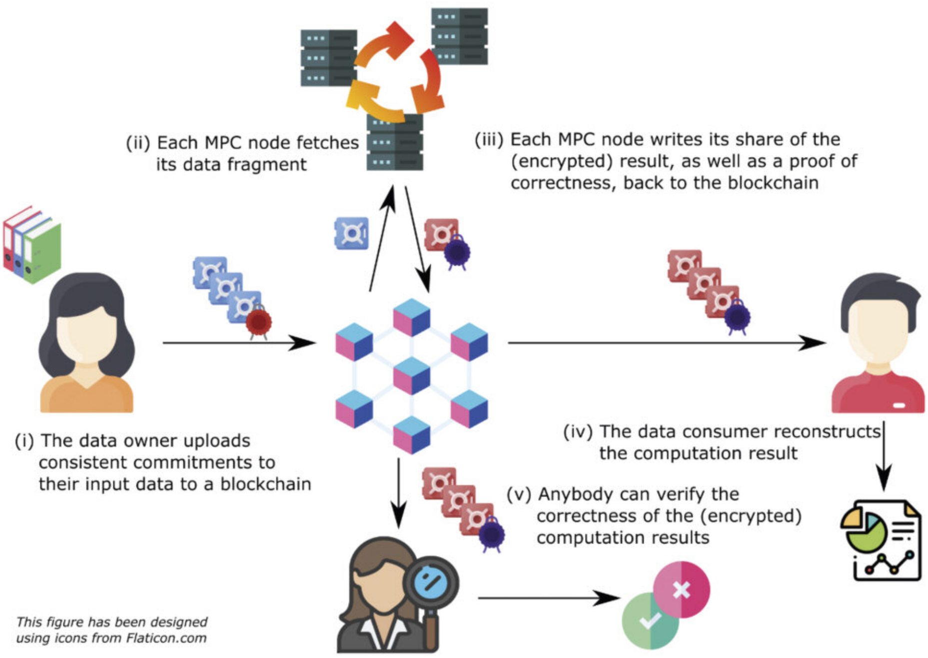

Figure 1: High-level architecture for verifiable multi-party computation.

ing that indeed the correct computation on valid and legit input data was carried out.

Unfortunately, the complexity of MPC and, in particular, its verifiable version is often considered too high for real-world adoption of the protocols and solutions, especially compared to their non-privacy-preserving counterparts. Within a series of national and European research and development projects, we have studied the feasibility of efficient verifiable MPC for a variety of application cases and found very favourable results.

FFG FlexProd considered the case of a sharing economy, where supply and demand of industrial capacities among potential competitors had to be matched, in order to increase the average load factor of production machines, increase competitiveness, and decrease costs, while at the same time protecting the individual bids of all involved stakeholders [1]. The H2020 SESAR JU SlotMachine project worked on optimising the assignment of airport slots in air traffic management across different airlines at major airports, aiming for minimising costs, delays, and carbon footprint of airlines, while avoiding privileging single players and while keeping business-critical information regarding the internal cost structure confidential [2]. Finally, H2020 KRAKEN developed a marketplace for personal data, allowing end users, data unions, and companies to commercialise their data in a privacy-enhancing fashion, while also ensuring the quality of the sold data products [3].

Interestingly, on a high level, we also identified a generic pattern for decentralised data markets that underlies all use cases, which is illustrated in the Figure 1. In the first step, the data owner commits to their input values without revealing them, and stores these commitments in an immutable manner, e.g. using a blockchain, together with the corresponding inputs for the MPC nodes and proofs that these data items are consistent. The MPC nodes then perform the requested computation, resulting in potentially encrypted fragments of the computation result as well as proofs of correctness. The data consumer may now reconstruct the final computation result, while everybody can verify the soundness of the computation.

As a result, we developed a generic architecture for verifiable MPC and demonstrated that it can be instantiated in an efficient manner if the protocols are tailored to the specific needs and requirements of each use case, e.g. in terms of real-time requirements or metadata protection. In particular, by replacing generic approaches to verifiability with problem-specific constructs, the alleged inefficiencies can be overcome for a broad range of computations such as auctions, optimisations and statistics. This proves that data sovereignty and authenticity are not mutually exclusive requirements that prohibit a sharing economy, but can be achieved simultaneously in collaborative environments and data spaces through the use of modern cryptography.

Links:
[L1] https://flexprod.at/de
[L2] https://kwz.me/hxo
[L3] https://www.krakenh2020.eu/

References:

[1] T. Lorünser, F. Wohner, and S. Krenn, "A privacy-preserving auction platform with public verifiability for smart manufacturing", in ICISSP, 2022, pp. 637–647.
[2] T. Lorünser, F. Wohner, and S. Krenn, "A verifiable multiparty computation solver for the linear assignment problem: and applications to air traffic management", in CCSW@CCS, 2022, pp. 41–51.
[3] K. Koch, et al., "KRAKEN: a privacy-preserving data market for authentic data", in DE@CoNEXT, 2022, pp. 15-20.

Please contact:

Stephan Krenn, AIT Austrian Institute of Technology, Austria
stephan.krenn@ait.ac.at

Thomas Lorünser, AIT Austrian Institute of Technology, Austria
thomas.loruenser@ait.ac.at

A Secure Trustworthy Privacy-preserving Framework for Cross-federation of Data

by Emmanouil Spanakis (ICS-FORTH), Evangelos Markakis (HMU), Dimitra Papatsaroucha (HMU), Ilias Polits (INQBIT) and Vangelis Sakkalis (ICS-FORTH)

Utilising self-sovereign technologies and with state-of-the-art homomorphic encryption, TRUSTEE offers a socially and environmentally aware framework for cross-discipline federation of data.

Nowadays, there is a general trend in the field of informatics towards the connection of multidisciplinary scientific fields such as health, space, automotive, education, cross-border and environment. The number of opportunities derived from this trend is impressive and includes the possibility to prevent future accidents/problems and security attacks, and also to tackle scientific problems where the social conditions and innovation form a complex matrix. Thus, it is evident that there is a need to measure how well existing technologies can be used to improve the availability and quality of multidisciplinary big-data sources, from various sectors, in a trustworthy, fair, responsible, and environmentally friendly manner, across the data life cycle to enable data sharing and manipulation, in compliance with prevailing and emerging legislation.

TRUSTEE EU project [L1] aims to combine technological and social innovations for secure and sustainable data operations in line with the principles of responsible/trustworthy AI by using a co-development approach as backbone methodology. TRUSTEE proposes a secure-by-design federated platform in line with EU data strategy (COM (2020) 66), and the main EU reference architectures (GAIA-X, EOSC, EGI) in the sector, able to ensure interoperability, enabling cross-border scenarios, and to scale varieties of AI-based applications, by using open APIs focusing on making the EU the world's most secure and trustful data hub. More than that, TRUSTEE acknowledges that today consumers must be very careful about sharing/accessing data, and national and international regulators/legislators step up privacy requirements. The industry needs to identify how investing in data protection and privacy can create a business advantage for the proper handling of data, consent, notice and regulatory obligations; how data is securely shared with third parties; how data is legally collected or stored; how data can be collected for processing; and regulatory restrictions such as GDPR, HIPAA, GLBA, or CCPA. Several effective actions have emerged seeking to address enhanced consumer-privacy and data-protection requirements. These span the life cycle of enterprise data, and expand the processes of operations, infrastructure and customer-facing practices, and are enabled by data mapping.

To achieve all these, we exploit an innovative homomorphic approach that guarantees user-friendly, safe, trustworthy, compliant, fair, transparent, accountable and sustainable collection, storage, processing, querying and delivery of data. The project factors three key areas: serverless computing, edge computing and secure clouds, to deliver an open-source, scalable, efficient and trusted solution, able to seamlessly operate on core and edge cloud infrastructures for time-critical, self-hosted applications, in a European, privacy-preserving, green and responsible data-centric model.

We utilise enhancements in self-sovereign technologies with the use of blockchain for the decentralised identifiers and homomorphic encryption for the cross-discipline federation of data. Data consumers will be able to operate over a framework that will not know anything about their identity, nor their equipment, which will allow them to perform complex search queries across the federated data repositories and communicate through robust authentication and authorisation mechanisms. Hence, the system can reveal only the necessary data for any given transaction or interaction. The goal is to create a framework aiming towards privacy preservation, green and responsible management, giving individuals or organisations full ownership of their digital and analogue identities, and control over how their personal data is shared and used, enabling search in the encrypted domain, while guaranteeing data privacy and data confidentiality as the data remain encrypted.

TRUSTEE promotes vital cross-functional collaboration and automation to build fast, trustworthy data pipelines. The overall goal is to create a data-driven platform that envisages cutting-edge solutions, and promote trustworthy green and re-

Main subsystems / layers of TRUSTEE architectural design (Figure 1)	
Front-end layer	Point of access for the end-user (consumers) or service APIs FOR devices, users, and federated service access layer
Accountable transaction layer	Responsible to record every single transaction of the users with the data while permitting private transactions
Homomorphic and self-sovereign framework	Trust and Security ensure that identity and privacy are managed across federated instances of the TRUSTEE framework.
Cloud continuum service layer	Cloud-based service continuum knowledge base for TRUSTEE
TRUSTEE core layer	Advanced multi-dimensional/sector approach to promote situational/context awareness of the EU data common spaces procedures for processing
Knowledge repository	Fuse data coming from those heterogeneous sources and will be integrated based on data homogenisation and semantic alignment toolkits
Common data space sources	The data collection sources are shown in the bottom layer where various (structured and unstructured data sources)

Table 1: Summarised architecture schematics of TRUSTEE.

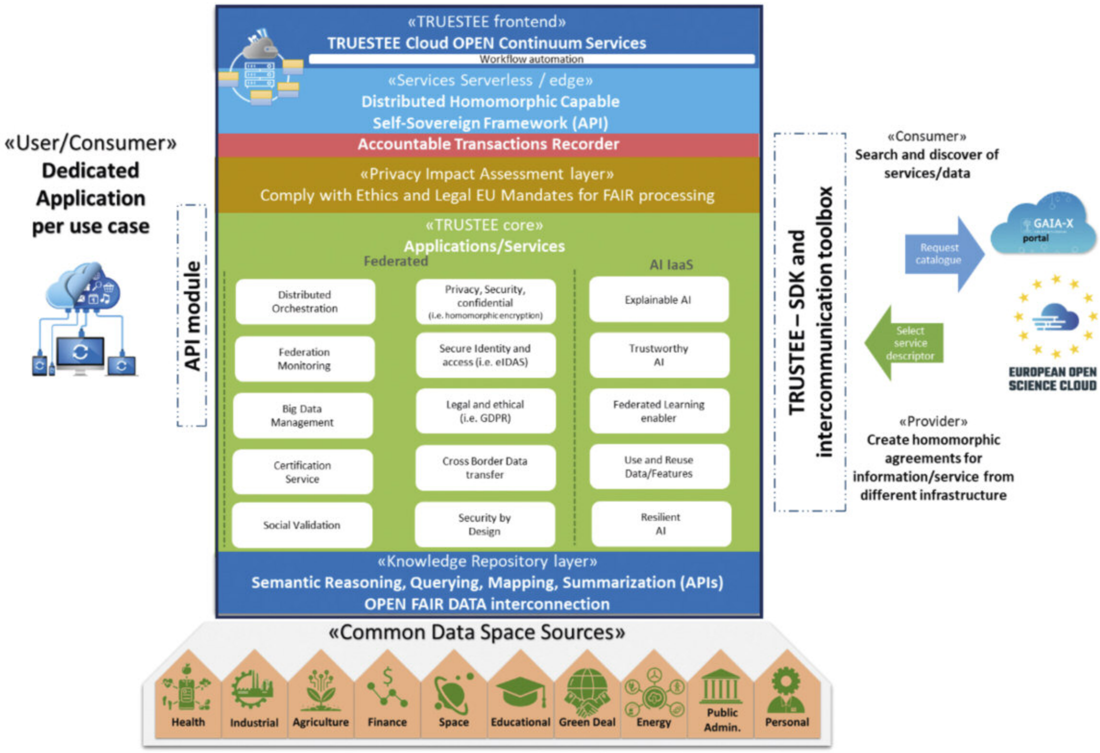

Figure 1: High level architecture of TRUSTEE Platform.

sponsible data management and processing, through value-based digital technologies and solutions for compliance, privacy homomorphic encryption and preservation. In Figure 1 we present our architectural designs aiming to satisfy the requirements of citizens, industry and research for federated data access through a private framework of public and private organisations. Table 1. presents the architecture schematics of TRUSTEE.

TRUSTEE emphasises social innovation and privacy impact assessment through the proposed technologies by optimising processing at the edge, resilience, transfer and storage, avoiding unnecessary manipulations in line with responsible/trustworthy AI principles. Our architecture supports FAIR communication with open data silos and platforms to exchange existing and future data originating from various EU data spaces are thus empowers stakeholders with user-friendly, safe, trustworthy, transparent, accountable and sustainable ICT services for collecting, storing, processing, querying and delivering data. TRUSTEE's fully encrypted solution will be validated through six different use cases supporting GAIA-X, EOSC, EGI, etc. demonstrating a multi-disciplinary, Pan-European federated FAIR and private data ecosystem. Figure 2 presents the six different areas for validation.

Link:
[L1] https://horizon-trustee.eu/

Please contact:
Emmanouil Spanakis, ICS-FORTH, Greece
spanakis@ics.forth.gr
+30 2810 391446

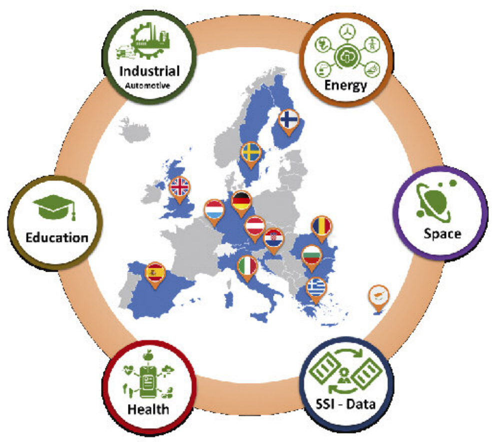

Figure 2: TRUSTE Use cases / Pilots.

Privacy-Preserving Data Sharing for Collaborative Analytics in Multi-Modal Transportation Systems

by Daniele Albanese, Giuseppe Crincoli, Marco De Vincenzi, Giacomo Iadarola, Fabio Martinelli, Ilaria Matteucci and Paolo Mori (CNR-IIT)

Allowing data producers to retain some control of the data they share is of paramount importance to encourage data sharing for collaborative analytics.

E-CORRIDOR [L1] is an EU-funded project aimed at defining a framework for multi-modal transport systems providing secure advanced services to passengers and transport operators. The framework implements collaborative privacy-aware edge-enabled information sharing, analysis and protection as a service. The actors of this framework, either travellers or transport companies, can be seen as information prosumers, i.e. producers and consumers of information. Information could be raw data as well as complex attack indicators that prosumers may wish to share in order to enhance security and safety, or just to get a better transportation service. Figure 1 shows how information analysis and sharing can be related to multiple aspects in multimodal transport. Users can share their travel preferences; cars can be seen as sensor platforms producing several kinds of data; cars can automatically recognise drivers and similarly, traffic jams. We plan to empower users with the capability to share the data they wish as they wish to enable the execution of collaborative data analysis.

The E-CORRIDOR framework is based on the concept of a data sharing agreement (DSA), which is an agreement among a set of parties that regulates the sharing of information among them. The framework provides an infrastructure enabling a DSA's enforcement when information is shared and analytics are executed on it. Preserving information privacy is a fundamental feature of the E-CORRIDOR framework, because it encourages producers to share their information with the other actors of the system since they still retain some control on the subsequent usage. Information can be analysed either globally (in the cloud) or locally (in edge devices). Local analysis increases privacy although global (with more information) analysis could be more accurate. Our framework has the following key components:

- Information sharing: share information (including security ones) in a controlled manner, ensuring confidentiality and integrity as well as regulation compliance, both at rest and while in transit
- Information analytics: advanced analytics functions for data analytics and correlation identifying threats that hide themselves in the massive usage of services and related number of logs
- Mixture of technologies: enable confidential and collaborative analysis of data, including homomorphic encryption
- Advanced seamless access: mechanisms that take advantage of the analytics and sharing infrastructure to provide continuous authentication and authorization as well as privacy-aware service as privacy-aware data usage control.

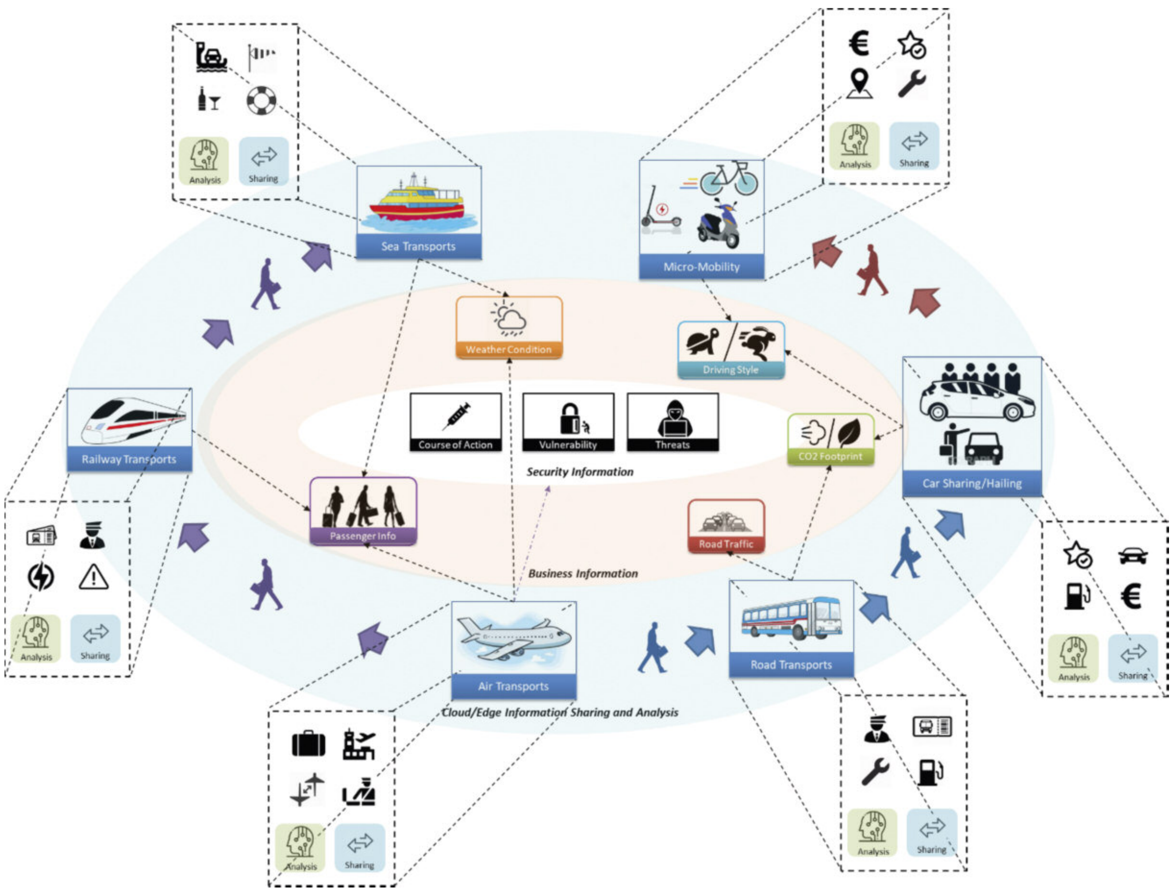

Figure 1: E-CORRIDOR operation concept.

Figure 2: Structure of Data Sharing Agreements.

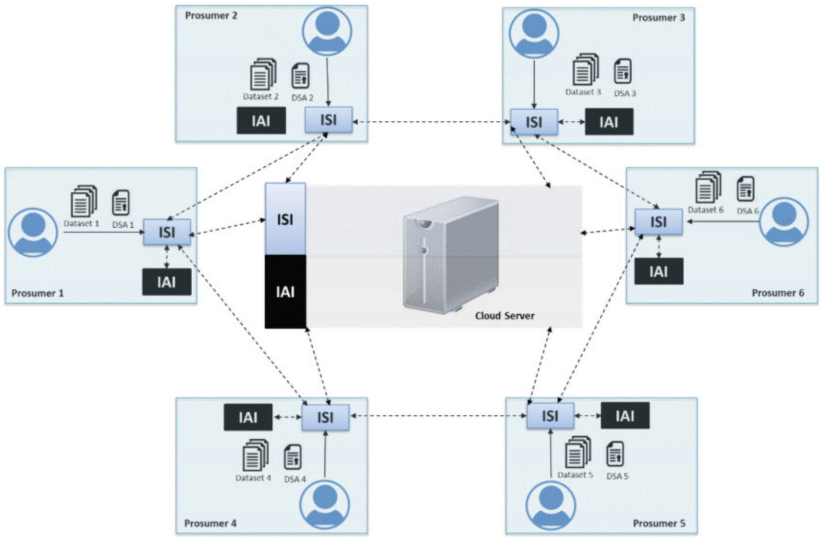

Figure 3: E-CORRIDOR Architecture.

Data Sharing Agreements

A DSA is a digital contract that defines a set of constraints to regulate the sharing of data among some parties. The DSA is at the base of the E-CORRIDOR data protection support, since it specifies which actions (e.g., analytics operations) can be performed on each piece of data, which subjects can execute these actions, and which other conditions should be satisfied in order the framework to authorise the execution of such actions. A DSA includes, among the other information, the Policy, which consists of a number of authorisation and prohibition rules expressing constraints concerning the attributes describing the subject and the data, to be enforced on data sharing. Policy rules also includes obligations, that are actions that must be executed (e.g., data anonymisation) before making the data available for being used. Concerning the enforcement time, rules can be of two kinds: pre or ongoing. The formers are evaluated when the E-CORRIDOR user requests the execution of an analytics and concur to decide whether the execution can be started or not. Ongoing rules are evaluated while analytics are in execution and determine whether the execution of such analytics can be continued or must be interrupted because of a policy violation. DSAs allow data producers to express rules expressing constraints on the other pieces of data that are involved in the collaborative analytics with the piece of data the DSA refers to. A graphical representation of a DSA is shown in Figure 2.

Architecture

The E-CORRIDOR architecture, shown in Figure 3, is meant at supporting data sharing and collaborative analytics execution while enforcing DSAs. E-CORRIDOR users act as data producers when they upload their data (paired with the related DSA) on the E-CORRIDOR framework, thus creating Data Bundles. E-CORRIDOR users act as data consumers when they request the execution of collaborative analytics by selecting a set of Data Bundles to be used.

Information Sharing Infrastructure (ISI)

The ISI is in charge of managing data, ensuring its secure storage as well as its privacy preserving sharing. To protect confidentiality at rest, when new data is uploaded by data producers, the ISI embeds it in cryptographic containers, called Data Bundles, before being stored on the storage system. Several storage systems can be supported, e.g., local storage or cloud based storage services. Instead, for protecting data privacy when they are shared to perform collaborative analytics, DSAs are embedded in Data Bundles and, every time that an E-CORRIDOR user requests to execute a collaborative analytics on a Data Bundle, the corresponding DSA is enforced to check whether such user actually holds the right to exploit such data to perform such analytics. Moreover, the ISI subsystem also allows the integration of customised privacy preserving operations (such as data specific anonymisation operations) following a plugin approach. These operations are executed on the data extracted from Data Bundles when required by the DSAs, before making data available for executing the analytics.

Information Analytics Infrastructure (IAI)

The IAI manages the execution of collaborative analytics on the data shared through the ISI subsystem. E-CORRIDOR users act as data consumers interacting with the IAI to request the execution of collaborative analytics on a given set of Data Bundles, which is typically defined using a query on the metadata paired with the Data Bundles themselves. The IAI invokes the ISI subsystem to search and retrieve such Data Bundles, which are used for executing the requested analytics. When the analytics execution has been completed, the IAI subsystem interacts with the ISI subsystem to create a new Data Bundle in order to make available to E-CORRIDOR users the results of the analytics execution.

Link:
[L1] https://e-corridor.eu/

Please contact:
Fabio Martinelli, CNR-IIT, Italy
fabio.martinelli@iit.cnr.it

Extensive and Secure Personal Data Management Systems

by Nicolas Anciaux and Luc Bouganim (Inria, UVSQ and Univ. Paris-Saclay)

Personal Data Management Systems (PDMS) are emerging to provide individuals with appropriate tools to collect, manage and allow applications to access their personal data. At the same time, the deployment of Trusted Execution Environments (TEE) both on the cloud (Intel SGX processors) and on smartphones (ARM TrustZone) opens new architectural perspectives for these PDMS. The PETRUS project [L1] proposes new solutions for PDMS, allowing third-party applications to exploit the individual's personal data while guaranteeing individuals the legitimacy of uses and privacy. A first proof-of-concept implementation of this new architecture has just been developed on Intel SGX processor.

The PDMS is a database system that manages the personal data of a user, under the user's control, based on a paradigm inversion for the management of personal data: instead of externalising the personal data of the user to the services performing the processing, it is the processing code that comes to execute on the side of the PDMS close to the data in a controlled execution environment. To exemplify this trend, let's consider a scenario where companies incentivise their employees for environmentally friendly behaviour by offering a green bonus based on the number of commutes made by bike. GPS traces are collected from a reliable service and processed locally by the user's PDMS. The result is then delivered to the employer with proof of compliance. Several similar scenarios are realistic, e.g. patients providing statistics to hospitals, power meters etc. This scenario is challenging since it calls for extensiveness to develop and deploy ad hoc code, and security to assure PDMS users that their detailed personal data is not disclosed to third parties outside the sphere of control of their PDMS.

To solve this tension between extensiveness and security, we proposed in [1] a three-layer logical architecture where a minimal Secure Core implementing basic operations on personal data is extended with Isolated Data Tasks themselves accessed by Applications on which no security assumption is made (see Figure 1, left). The objective is to control the flow of raw personal data from the Core to the outside, such that only expected results are declassified to untrusted applications or third parties. The right part of Figure 1 gives details on the concrete implementation of an ES-PDMS, with some examples of data tasks to answer the above scenario. Properties 1 to 5 are generic security properties that must be properly orchestrated to strengthen the security of our architecture. Security properties 6 to 8 were designed for the specific ES-PDMS context and were introduced in [2] and demonstrated in [L2].

This is a first step, and many challenges remain, depending on the type of processing and the manipulated data that must be properly protected:

- Personal data of a single PDMS user. To properly address this use case, presented above, it is necessary to consider processing functions that handle large volumes of personal data, such as aggregation functions. These functions are defined by a third-party application called App, and their code is evaluated in the PDMS environment. However, the PDMS user cannot fully trust the App. Therefore, it is essential to focus on execution mechanisms that mitigate information leakage through the successive processing of these functions. As preliminary results, we introduce in [2] new execution strategies based on partitioning and replay of processing in different stateless Intel SGX enclaves, in order to guarantee an upper bound on data leakage through legitimate results, for certain categories of aggregation functions typical of PDMS use cases, and with a reasonable performance overhead.
- Personal data of a community of PDMS users. Individuals can cross data within large communities of users, e.g. to compute statistics for epidemiological studies or to train a machine learning model for recommendation systems or automatic classification of user data. While in the previous case we can reasonably assume that the user will not attempt to attack the confidentiality of their own data, this is no longer the case as soon as we assume decentralised computations on PDMSs that will manipulate the data of other users. It is therefore necessary to consider implementation strategies that limit data leakage in case of a breach of the (hardware) security of the PDMS. As a first possible scenario, we consider in [3] a threat model that we call the "Wolf in the sheepfold", in which (a small number of) PDMSs have been instrumented by their owner and thus

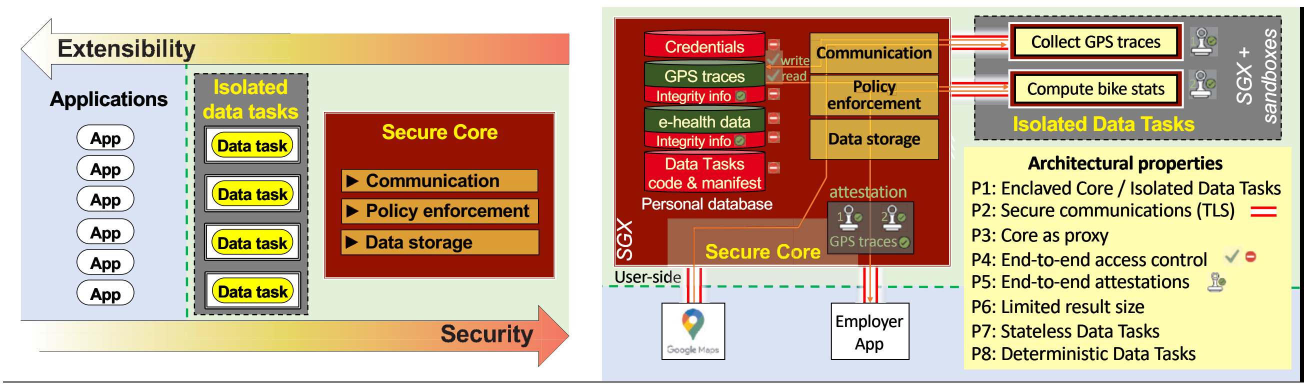

Figure 1: ES-PDMS logical architecture and a concrete implementation [L2].

operate in the so-called "sealed glass proof" mode, without confidentiality (but still as expected, with integrity guarantees). Other more difficult models are to be considered, where the communications between PDMS can be analysed and where the nodes could be corrupted also in integrity or in greater number.
- Personal data of third parties. Some cases of computation on PDMS require the exploitation of potentially sensitive third-party data (e.g. a machine learning model derived from a personal dataset or an IP-protected dataset). Integrating or storing third-party data in the application during deployment would, compared to the two cases above, raise additional security issues (the same set of sensitive data must be protected in all PDMSs running the application) with update and performance difficulties after deployment. Therefore, it is necessary to allow for the application to dynamically retrieve third-party data when running on the PDMS side, which represents a new potential leakage channel for PDMS data and requires new solutions to regulate PDMS data flows at runtime.

Since the initial paper [1] on the extensive and secure PDMS architecture, the PETRUS project has obtained interesting research results and realised some demonstrators [L2, L3] which underscores the potential of the PDMS approach, fostered by the development of TEEs.

Links:
[L1] https://team.inria.fr/petrus/
[L2] https://project.inria.fr/espdms/
[L3] https://project.inria.fr/edgeletdemo/

References:
[1] N. Anciaux et al., "Personal data management systems: The security and functionality standpoint," Information Systems, vol. 80, 2019.
[2] R. Carpentier, I. Sandu Popa, N. Anciaux, "Data leakage mitigation of user-defined functions on secure personal data management systems," SSDBM, 2022.
[3] L. Javet et al., "Edgelet computing: Pushing query processing and liability at the extreme edge of the network," CCGrid, 2022.

Please contact:
Nicolas Anciaux
Inria, UVSQ and Univ. Paris-Saclay, France
nicolas.anciaux@inria.fr

CyberSec4Europe - Securing and Preserving Privacy Sharing Health Data

by Raquel Cortés Carreras, Juan Carlos Pérez Baún and Miryam Villegas Jiménez (Atos Spain S.A.)

In the context of EU Horizon 2020 CyberSec4Europe project the medical data exchange demonstrator creates a trusted ecosystem for sharing medical data in a secure and privacy preserving manner.

The personal and sensitive data generated in the health domain including hospitals records, patient records or wellbeing wearable devices (data providers) is growing sharply. This huge amount of health data can be used by health organizations and research laboratories (data consumers) for improving the population's health. At the same time this information acquire economic value to pharmaceutical industry and insurance companies. These personal and sensitive data needs to be effectively managed in terms of quality, security, and privacy, assuring the integrity and reliability when this information is shared between parties. In the context of H2020 CyberSec4Europe project [L1, L2] the Medical Data Exchange demonstrator addresses these security and data protection challenges to ensure trust between two main actors: the data providers and the data consumers. The use of privacy preserving techniques has been adopted by this demonstrator for creating a trustworthy environment by developing a data anonymisation service (DANS) and a functional encryption service (FE2MED), increasing the user privacy and security when sharing data through a data exchange platform (Figure 1).

Anonymisation tool

The Data Anonymization Service (DANS) is an anonymization tool, developed by Atos, based on an open-source tool named ARX [1]. DANS makes it possible to mitigate tracking and user re-identification by anonymizing sensitive personal data, leveraging k-anonymity and l-diversity privacy models, which enable the application of some privacy criteria over a particular dataset, protecting biomedical data against data disclosure. Preserving data privacy applying anonymisation techniques, helps health stakeholders, such as health authorities, hospitals, research bodies and the private health sector, to be regulatory compliant, as fully anonymised data are considered out of the scope of the GDPR regulation (according to Recital 26 [L3]). Unfortunately, perfect anonymization is not possible, the use of anonymisation algorithms that maximise the utility and the privacy protection is quite difficult. and it is necessary to find a balance between privacy and utility. To this end privacy metrics and privacy risk features are envisage be included in the DANS service in the future.

DANS tool is offered in two flavours for being used by the data providers:
- A java library to be integrated in the data provider legacy systems. Also, this option allows the use of privacy-preserving techniques on the Internet of Things (IoT).

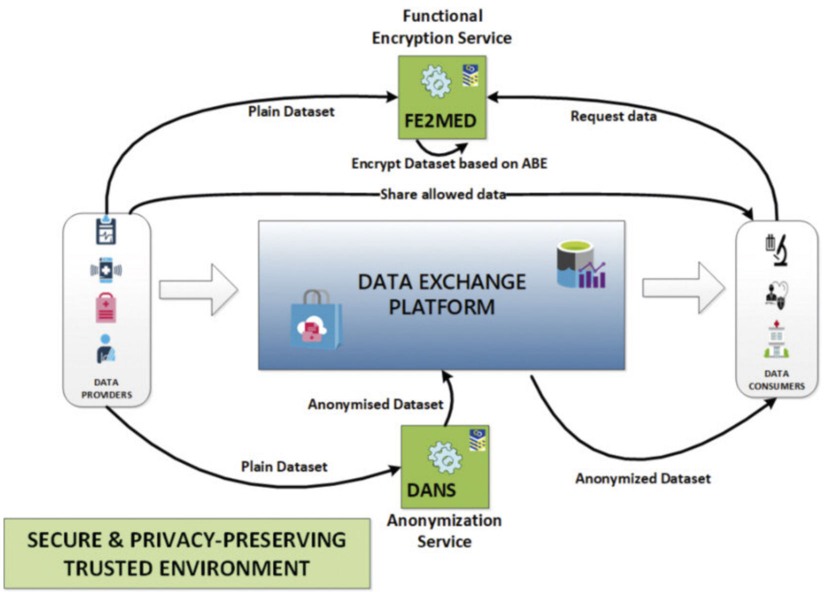

Figure 1: The privacy-preserving techniques anonymisation and functional encryption services create a trusted environment for sharing health records.

- An anonymization service to be deployed on the data provider premises or in a trusted third party, exposed as a RESTful API. In this case a graphical user interface is provided to the user to facilitate the use of the anonymisation tool and improving the user experience.

Functional encryption tool

The Functional Encryption to Medical Data (FE2MED) service is developed by Atos, based on Horizon 2020 FENTEC project crypto libraries [L4]. The FE2MED tool leverages libraries containing attribute-based encryption (ABE) and inner-product encryption schemes used for preserving privacy during the health data management. This technology allows selective share of the encrypted data in a fine-grained level. Basically, the encrypted data are labelled with sets of attributes and private keys are associated with access structures that control which encrypted data a user can decrypt [2]. This means that, only the allowed users are able to access the encrypted data or the results, ensuring the integrity and the confidentiality of shared data.

The FE2MED service is deployed in the data provider infrastructure, in order to have the full control on the generated key elements. Also, exposes a RESTful API to the users. Additionally, provides a graphical user interface to users for hiding the complexity of the encryption/decryption processes, improving the user experience. While the data provider is able to access to all the FE2MED functionalities, the data consumer only has access to certain features related to the decryption process.

The FE2MED service acts as a trusted entity managed by the data provider. It is in charge of generating cryptographic keys for encryption and for decryption. The two scenarios developed in the medical data exchange demonstrator are the following:

- Share encrypted datasets with different data consumers which can see only specific data. In this scenario the ABE schemes are applied, which give a more fine-grained control of the decryption capacities of third parties.
- Share encrypted results with data consumers which are able to see only the results but not the data. In this case the Inner-product schemes are applied for performing basic (Inner product-simple DDH scheme) or more complex (Inner product- Fullysec Damgard DDH scheme) statistic calculations.

Currently, the FE2MED tool is focused on encrypting health datasets to be shared with third parties, the protection of data generated in the health IoT infrastructure domain (e.g., connected medical devices and mobile devices) will be addressed in the future [3].

Links:

[L1] https://cybersec4europe.eu/
[L2] https://cordis.europa.eu/project/id/830929
[L3] https://gdpr-info.eu/recitals/no-26/
[L4] https://github.com/fentec-project/gofe

References:

[1] F. Kohlmayer et al., “Highly efficient optimal k-anonymity for biomedical datasets”, in: 25th IEEE International Symposium on Computer-Based Medical Systems (CBMS), 2012.
[2] V. Goyal et al., “Attribute-Based Encryption for Fine-Grained Access Control of Encrypted Data”, in Proc. of the 13th ACM Conf. on Computer and communications security, 2006.
[3] N. T. Dang et al., “Applying attribute-based encryption on mobile devices”, in Sci. Tech. Dev. J. – Engineering and Technology, 2020.

Please contact:

Juan Carlos Pérez Baun
Atos, Spain
juan.perezb@atos.net

On Reducing Complexity in AI Pipelines: Modularisation to Retain Control

by Dennis Hemker, Stefan Kreutter and Harald Mathis (Fraunhofer FIT)

Data is the key factor for solving problems with modern approaches like AI but only a part of the whole. To reduce complexity in the AI life cycle, we propose a modularised tooling approach and describe the data-related parts of it.

Recent developments like ChatGPT [L1] or Stable Diffusion [L2] have shown the high potential of data-driven approaches to tackle a wide range of problems across different domains.

With the involvement of Artificial Intelligence (AI), a typical data-focused process before training a model includes filtering, labelling, transforming and splitting. Orthogonal to these development steps of the AI itself, data needs to be acquired, stored, backed up and made accessible. Additionally, there is no guarantee that only people with deep technical backgrounds are part of such processes.

Current solutions for managing AI-related workflows and their data tend to be heavy-weighted, trying to offer all-in-one software. This often comes at the cost of complexity, either in usage (locked into one ecosystem), setup and maintenance (need for clusters like Kubernetes), pricing (high rates per member per month) or privacy (cloud only instead of on-premise). Within the project "progressivKI" (funded by the BMWK) [L3], a modular AI platform is developed. While this platform tries to solve different parts of a machine learning workflow like GPU-based training, containerisation and model performance comparison, it also decouples data handling aspects.

As described in [1], AI-code itself only takes a small part of the whole project, raising the need for supportive infrastructure tooling. The main idea behind our approach is to develop a platform following principles of the Unix philosophy [L4]. This means to solve problems with small and effective tools while retaining full control over all parts of the process.

Architecture

The main assumption within the developed platform is to treat every data as files. Each step in the pipeline (which can be a small script/executable in an arbitrary programming language) consumes files as inputs and produces files as outputs that can be consumed by later steps. We refer to these steps as "stages" in a pipeline.

The aforementioned stages form a direct acyclic graph (DAG). To handle their execution in the correct order and to avoid redundant calculations on unchanged stages, we employ Data Version Control (DVC) [L5]. In a Git-like fashion DVC can handle data sources as "remotes" and supports different storage locations like Amazon S3, Google Drive, SSH, HTTP, local file systems and more. This allows easy integration, updating, replacing and versioning of data sources. Additionally, there is no lock-in to any specific framework or (cloud) infrastructure. AI-code and DVC configuration files are versioned with Git.

In the current implementation we employ a S3-compatible storage server. It can make use of its own replication and backup mechanism and is managed separately. The project-specific bucket is split into three different parts:

1. Inbox. This place can be considered as data heap. New arbitrary data can be put here occasionally by external producers.
2. Cache. DVC-specific cache to store intermediate file artefacts, e.g. metrics, models and processed files. Fully managed by DVC.
3. Outbox. The place where trained models or other file artefacts like transformed data can be saved. Can be accessed by other consumers.

This architecture allows fine-grained configuration and rights management for each project and user. Data producers can be equipped with write-only access keys to the inbox while consumers of outbox artefacts can be granted read-only credentials. Developers of AI pipelines can access all three parts of the bucket. Furthermore, the division can also be spread across multiple buckets or even storage locations. By treating storage as a single component as DVC does, it can be designed per need. Either available at short notice and quickly or long-term, redundant and fail-safe.

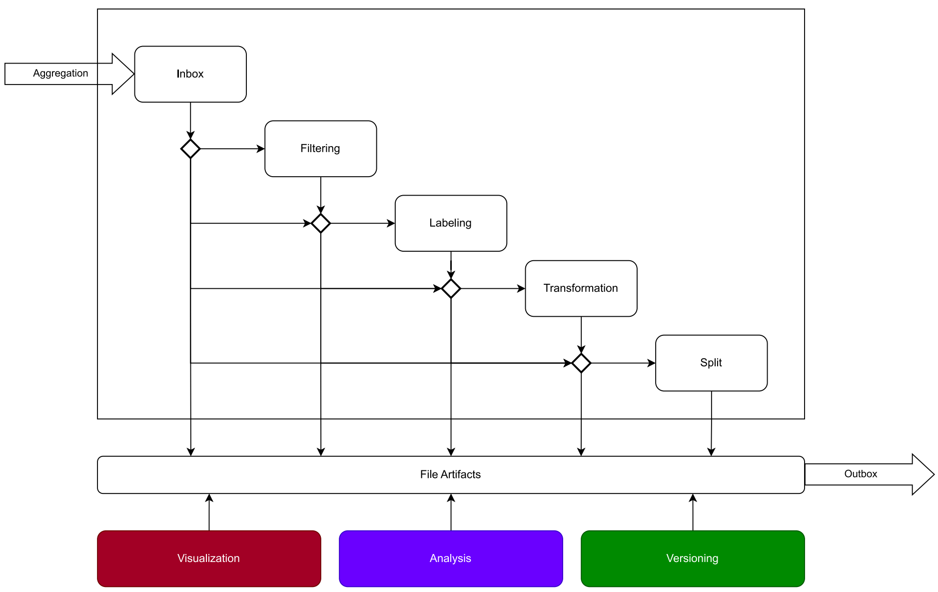

Figure 1: Example data preparation pipeline in an AI workflow.

An easy-to-use tool written in the Python programming language enables collaborators in such projects to simply upload data to specific inboxes. While avoiding complicated interfaces or processes, barriers are broken down for data producers with less technical backgrounds, allowing fast, secure and reliable exchange between them and developers. From respective inboxes, the data can be pulled effortlessly to local workspaces or remote servers for further processing or model training. The server hardware can be located anywhere, on-premise, cloud or in any possible combination.

Figure 1 depicts the process of a typical AI data preparation pipeline. The stages to the right of "Inbox" reside in the cache part of the bucket. Assets from both the inbox or the cache can be packaged and moved to the outbox for further consumption. As everything is treated as files, it is easy to use one's favourite tooling to visualise or analyse these files. By not forcing other interfaces, the structure remains flexible and customisable for attaching to other ecosystems of software.

Data validation and quality checks [2] can be integrated as stages, while data reduction algorithms [3] can be deployed as transparent services located in front of an inbox. This allows the building of pipelines tailored per need, across frameworks and hardware.

Future work includes adding a more sophisticated user and rights management, as access key generation and distribution is currently done manually. It is also planned to foster data privacy by utilising encryption techniques on the server side. Additionally, we want to investigate further deployment tools like [L6] and benchmark data throughput of the storage server.

Links:
[L1] https://openai.com/blog/chatgpt/
[L2] https://stability.ai/blog/stable-diffusion-public-release
[L3] https://kwz.me/hxm
[L4] http://www.catb.org/~esr/writings/taoup/html/ch01s06.html
[L5] https://dvc.org/
[L6] https://mlem.ai/

References:

[1] D. Sculley, et al., “Hidden technical debt in machine learning systems”, in NeurIPS, 2015.
[2] E. Breck, et al., “Data infrastructure for machine learning”, SysML conf., 2018.
[3] G. Karya, et al., “Basic knowledge construction technique to reduce the volume of low-dimensional big data”, ICIC, 2020.

Please contact:

Dennis Hemker
Fraunhofer Institute for Applied Information Technology FIT, Germany
dennis.hemker@fit.fraunhofer.de

SEDAR: A Semantic Data Reservoir for Integrating Heterogeneous Datasets and Machine Learning

by Sayed Hoseini and Christoph Quix (Hochschule Niederrhein University of Applied Sciences)

SEDAR is a comprehensive semantic data lake that includes support for data ingestion, storage, processing, analytics and governance. The key element of SEDAR is semantic metadata management, suitable for many use cases, e.g. provenance, versioning, lineage, dataset similarity or profiling. The generic ingestion interface can deal with any external data source ranging from files to databases and streams change and incorporates data capture with data versioning and automatic metadata extraction. Machine learning (ML) is integrated into the data lake as its artefacts (e.g., ML pipelines, notebooks, models) are stored in the data lake to allow a coherent development of data preparation and ML pipelines. As all these artefacts are related, their relationships and versions are maintained in the extended metadata repository.

Data lakes have been addressed intensively in research and practice in recent years as they provide the required data for data science and ML projects. Data lakes are scalable schemaless repositories to ingest raw data in its original format from heterogeneous data sources; thus, only a minimal effort is required for ingesting data in a data lake, which makes it an efficient tool to collect, store, link and transform datasets. While first implementations for data lakes aimed at processing “big data” efficiently using distributed, scalable systems like Hadoop, the need for proper management of metadata and data quality in data lakes has also been recognised [1].

Data lakes have many applications, one example is management of heterogeneous data in Industry 4.0 scenarios [2]. For efficient data management in this context, the FAIR principles (findable, accessible, interoperable, reusable) are important requirements. SEDAR is a comprehensive data lake that addresses these challenges within one uniform system. The key components of the system are the generic ingestion interface, the metadata repository (or data catalogue) and integrated machine learning (MLOps).

Our implementation of the data catalogue satisfies several technical requirements for comprehensive metadata management: 1. indexing, 2. data linkage, 3. data polymorphism / multiple zones (e.g. different formats or degree of processing), 4. data versioning, 5. usage tracking, 6. granularity levels, 7. similarity relationships, 8. semantic enrichment and 9. data quality features. Furthermore, the data catalogue must ensure ease-of-use for non-technical users and provide means to manage access to sensitive information, i.e. privacy-preserving measures. These requirements can only be met partially by existing data lake systems [3].

Conceptually, the architecture of SEDAR is divided into four functional areas: (a) ingestion, to continuously extract data from heterogeneous sources (e.g. files, databases, streams), (b) storage, to efficiently store the data in big data systems, (c) transformation, to have a scalable environment to preprocess, integrate and transform data and (d) interaction, to provide tools to the user to interact with the data.

The ingestion layer aims at generality in order to support a wide range of data formats and source systems. All configuration settings for data sources can be changed at run time; new systems can be integrated using an extensible ingestion interface. Metadata management is an important issue in the whole data lake; it is initiated by an extensive metadata extraction. Furthermore, data ingestion is not a one-time process; data sources are updated continuously and these updates need to be reflected in the data lake. Thus, versioning of datasets is supported, including change data capture between two versions of a dataset. By doing so, the system can incrementally process the updates instead of loading the dataset as a whole. On the other hand, the data versioning mechanism enables access to previous states of a dataset. The information about ingested datasets, their versions, updates, etc. are extracted automatically by integrated the Delta Lake API [L1] and stored in the data catalogue.

The storage layer provides efficient storage systems, based on big data systems, such as Hadoop, or NoSQL systems, such as MongoDB. According to the basic idea of data lakes, data is stored in its original format. Thus, we chose an appropriate storage mechanism for different data formats (e.g. JSON is stored in MongoDB). The data catalogue keeps track of all datasets and their storage locations.

Apache Spark provides uniform access to the storage systems and is used as the main component in the transformation layer. With its data access layer based on SparkSQL and the DataFrameAPI, Spark provides a very helpful tool for implementing data preprocessing and integration pipelines. We integrated Amazon's Deequ [L2] library built on top of Apache Spark for automating the verification of data quality at scale.

While the functions of the ingestion, storage and transformation layers are very common for data lakes, SEDAR provides a semantic metadata management and integrates MLOps (i.e. operational support for ML applications) in the interaction layer. Datasets can be annotated with and queried through semantic terms (e.g. using ontology-based data access). To keep track of the various inputs and outputs of an ML application (e.g. data versions, code and tuning parameters, reproducing results and production deployment), MLOps is proposed to streamline the ML lifecycle. SEDAR addresses the key challenges in ML applications: experimentation, reproducibility and model deployment.

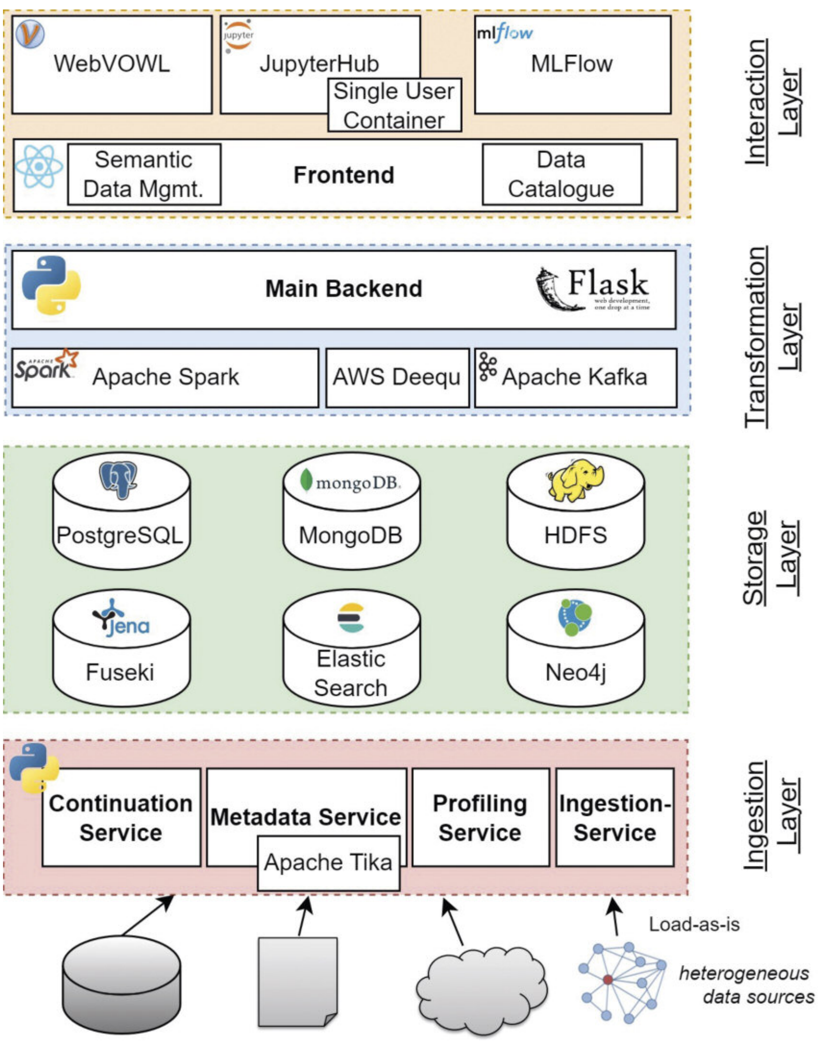

Figure 1: The SEDAR system architecture.

Technically, SEDAR uses a modular and flexible architecture based on microservices that can be extended easily with additional functionality, making the system suitable as a workbench for further research in data lakes. SEDAR is a web application with a modern, responsive user interface and several data management technologies in the backend exposed by a REST API.

Links:
[L1] https://github.com/delta-io/delta
[L2] https://github.com/awslabs/deequ

References:

[1] Farid, Mina, et al., "CLAMS: bringing quality to data lakes", in Proc. of the 2016 Int. Conf. on Management of Data, 2016, pp. 97-120.
[2] P. Brauner et. al., "A computer science perspective on digital transformation in production", ACM Trans. Internet Things, vol. 3, no. 2, pp. 15:1–15:32, 2022.
[3] Sawadogo, Pegdwendé, et.al, "On data lake architectures and metadata management", J. of Intelligent Information Systems, vol. 56, 2021.

Please contact:

Sayed Hoseini, Hochschule Niederrhein University of Applied Sciences, Germany
sayed.hoseini@hs-niederrhein.de

NET-IT: A Next Generation Intelligent Network Research Agenda

by Karin Anna Hummel (Johannes Kepler University Linz), Paul Smith (Austrian Institute of Technology), Markus Tauber (Research Studios Austria), Peter Dorfinger (Salzburg Research) and Olga Saukh (Graz University of Technology)

The NET-IT initiative at the Austrian Computer Society is making networks intelligent to ease adjustment to the digital evolution and societal needs: Let the network learn and act driven by goals!

Europe is experiencing an ongoing digital evolution, yet also faces several challenges for society such as the energy crisis, climate change and supply chain issues in a global economy. Supporting new digitised applications by low latency, high reliability, ubiquitous availability and high throughput networking technologies, while also being sustainable, is a challenge for future networks. Network intelligence is a high-level research topic that is expected to provide solutions to the pressing heterogeneous demands, which we investigate in the research and innovation initiative NET-IT [L1], a working group of the Austrian Computer Society. In this article, we discuss novel networked applications, network intelligence methods and major challenges of a future research agenda on network intelligence.

Intelligently Networked Applications

Innovation in networks and edge and cloud systems is driven by the emergence of novel application fields, which in turn reflect societal changes and needs. Selected examples are:

- *Digitisation of food production.* Environmentally friendly food production in Europe relies on novel approaches such as precision farming, enabled through digital technologies. Autonomous networked control systems [1,2] allow for large-scale wireless networked systems that sense and collect data, and reason about the status of food production such as the health of crops and irrigation. Figure 1 shows a sample system employing drones, see also [L2]. Easy set-up, secure communication, real-time control, and providing the system “as a service” according to farming goals is demanded.
- *Digitised missions.* Missions – such as search and rescue operations or inventories in warehouses – may involve humans, sensors, cloud/edge computing, and ground and aerial robots. The wireless network set-up should be fast and adaptive. Energy supply for the network, and best use and placement of vehicles and robots are key factors that result in a complex solution space for automated mission-based networks. Besides, missions are time-critical and often require low latency and timing guarantees.
- *Digital health and digital sports.* The digital evolution also changes the medical and sports sector. Shifting from quantitative metrics to qualitative metrics allows a new user experience. Providing, e.g. real-time feedback on the quality of the user’s steps while recovering from an injury, or on the turns while skiing may improve therapy results and skills. The supporting system has to place different services autonomously at different locations (device, edge, cloud) to fulfil the application’s specific timing, precision and privacy needs.

These examples demonstrate that addressing system, societal, and ecological demands altogether, results in a manifold of influential factors which challenge rule-based networking mechanisms and demand novel approaches.

Intelligence in Networks and Networked Systems

The vast amount of data available in today’s networks makes it possible to establish a novel portfolio of intelligent networking technologies that may employ the following methods:

- Detection and forecasting. Digital applications need to be aware of current and future connectivity. Hereby, network intelligence may help to predict the future communication network quality, e.g. at locations not visited before. Network monitoring and measurement tools are enablers of accurate classification and forecasting of network connection quality [3].
- Autonomy and adaptation. Changes in application demands lead to the need for network adaptation, which may be determined by learning from past observations. Different machine learning approaches may be applied such as automated reinforcement learning based on evaluating the network performance to generate a reward for the networking decisions taken, or, in case manifold features are available, deep learning. In an ideal setting, the network may then adapt to a current situation autonomously.
- *Selection.* Current networked applications may often choose between available networking solutions. In case of wireless networks, e.g. low-power wireless communication technologies could be preferred to reach a low environmental footprint. An

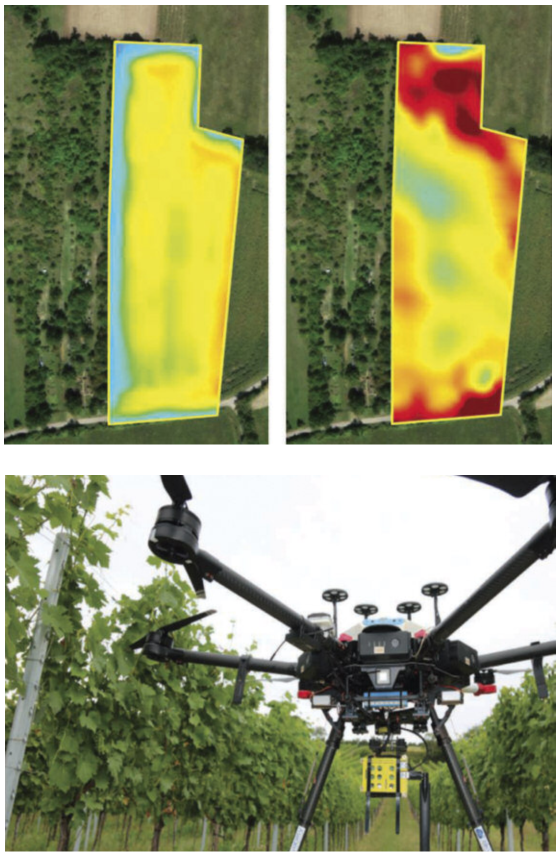

Figure 1: Precision farming system – (top) sensed micro-climate conditions visualiszed by different colours in a vineyard and (bottom) scanning drone [1].

intelligent network solution is expected to select the best technology among available ones, e.g. 5G and future 6G, Wi-Fi, BLE, ZigBee, ANT+, NRF24 or UWB. Placement of service provisioning (edge/cloud) is another autonomous selection task.
- *Context.* Not only data collected by network monitoring, but also data describing the wider context of networked devices, sensors, and vehicles can support networks to improve their performance. For example, location and mobility contexts are of interest for mobile networks, as are the energy demands of networking decisions and the current situation contexts.

Challenges for Network Intelligence
Network intelligence methods have to address traditional networking challenges such as scalability, reliability, interoperability, security, dynamic reconfiguration and programmability. In addition, novel embedded control systems have stringent real-time, dependability and safety requirements. Intelligent networking solutions thus need to provide guarantees from a holistic system perspective and to fulfil certification demands, which is an open issue.

Network intelligence also faces challenges inherent to AI and machine learning, which originate from the dependency on the data. Data bias and small datasets limit the generalisation of methods. Other limitations are due to energy costs and the computation time for (re-)training. Finally, network, edge and cloud providers need to understand the autonomous decisions taken by intelligent networking technologies, which calls for explainable network intelligence.

Summary
Network intelligence is a promising paradigm to address the heterogeneous demands on future networks that are, on the one hand, originating from employment in non-traditional sectors due to ongoing digitisation and automation, and on the growing concerns about energy resources and the demand to rethink the impact of technology on society. NET-IT [L1] is an open collaboration initiative that focuses on the particular challenges of network intelligence.

Links:
[L1] https://net-it.tk.jku.at/
[L2] https://www.youtube.com/watch?v=j6lTSSKLlAI

References:
[1] M. Tauber, et al., “Passive precision farming reshapes the agricultural sector”, IEEE Computer, Vol. 56, No. 01, 2023. doi: 10.1109/MC.2022.3218388.
[2] F. Papst, et al.: “Embracing opportunities of livestock big data integration with privacy constraints”, in Proc. of 9th Int. Conference on the Internet of Things, ACM, 2019. doi:10.1145/3365871.3365900
[3] S. Farthofer, et al. “An open mobile communications drive test data set and its use for machine learning”, IEEE Open Journal of the Communications Society, Vol. 3, 2022. doi: 10.1109/OJCOMS.2022.3210289

Please contact:
Karin Anna Hummel, Johannes Kepler University Linz, Austria, karin_anna.hummel@jku.at

Paul Smith, Austrian Institute of Technology, Austria
paul.smith@ait.ac.at

A Research Infrastructure where Artificial Intelligence Meets Persons and Society

by Roberto Trasarti, Katia Genovali and Beatrice Rapisarda (CNR-ISTI)

In our complex society, the ethical use and storage of data are essential for the scientific community and institutions to build trust in citizens. SoBigData is a pan-European and cross-disciplinary Research Infrastructure on social mining and data analytics, which bases its research activities on ethics and fairness. SoBigData doesn't apply science only to the most challenging societal issues; in fact, it provides data and facilities to researchers and services to firms and public administrations to develop innovative tools and respond to societal needs. Above all, it works to create an ecosystem for data research that respects the founding principles of Europe for the benefit of the whole community.

Every moment of our lives, we release data through our devices: computers, tablets and mobile phones, but also smartwatches, credit cards and internet access leave behind a wealth of data that records our daily activities. This data can reveal a lot about an individual's habits, movements and personal preferences. This is why it needs to be managed in the most ethical way possible.

Ethics is one of the key principles at the heart of SoBigData [L1], a Research Infrastructure (RI) dedicated to big data and social mining that spans the most pressing societal issues [1]: sustainable cities, demography and economics, online misinformation and migration. The social impact of artificial intelligence, sports data science and medicine can also be the subject of a SoBigData data scientist.

In 2017 SoBigData started including top research centres around Europe to create a technological common ground and provide all researchers from Europe access to them. From there, the consortium behind the (RI) grew by including more and more institutions and it became a multidisciplinary network of computer scientists and social scientists studying different aspects of society [2]. Today, SoBigData researchers, more than a hundred people from 29 data sites across Europe, are working together to answer challenging research questions about society. The aim is to create an ecosystem in big data and social mining research where values and standards of privacy, fairness, transparency and pluralism coexist (Figure 1).

The SoBigData RI includes numerous laboratories around Europe, e.g. the Institute of Information Science and Technologies “Alessandro Faedo” of the National Research Council (CNR-ISTI) in Italy, the project coordinator, creating the network of European nodes of the e-infrastructure. The infrastructure is based on the most advanced techniques of big data analysis combining artificial intelligence and social issues, a highly strategic sector in today's European and global economy.

The SoBigData RI platform, which serves more than 10 thousand users, enables researchers to perform large-scale social

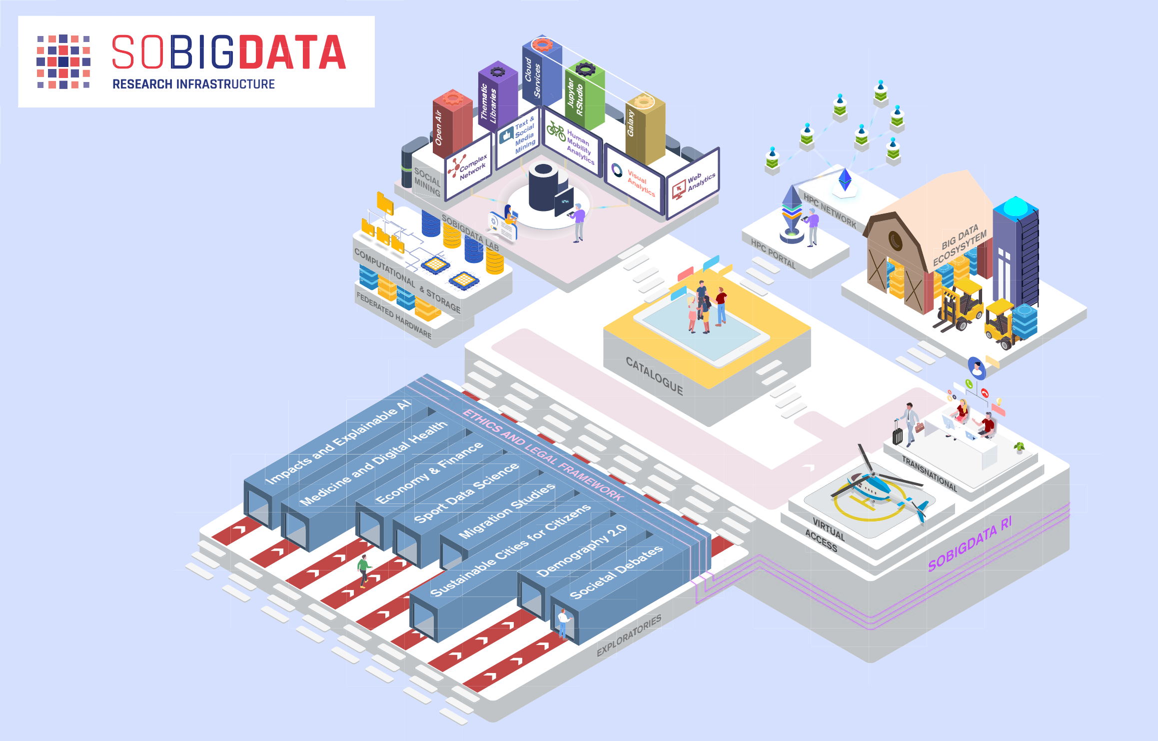

Figure 1: The SoBigData RI structure, centred on the Catalogue, allows the user to find the proper resources in the infrastructure. It includes the Exploratories, virtual environments where new research is developed (Demography, Economy and Finance 2.0, Migration Studies, Network Medicine, Social Impact of AI and explainable ML, Societal Debates and Misinformation, Sports Data Science and Sustainable Cities for Citizens), the SoBigData Lab, hosting resources available to run experiments on the cloud, and the High Performance Computing (HPC) Portal with useful information to access European facilities. SoBigData RI can be accessed virtually, through its website and gateway, [L1] or in person, thanks to its Transnational Access Programme visiting our nodes.

mining experiments in computational social sciences, digital humanities, urban planning, welfare, migration, and sport and health, within the legal and ethical framework of responsible data science. The RI offers innovative and free services for its users, such as the SoBigData Lab, where they can design and develop new algorithms in an interactive environment, with libraries ready to use, or deploy a method or run experiments on its cloud engine.

All the methodologies, datasets, scientific publications, and training materials available are collected in an omni-comprehensive navigable catalogue that aims at providing an easy way to search, find and reproduce innovative social mining experiments. In addition to online services, SoBigData offers a mobility programme under the European Transnational Access Programme, allowing users to visit and collaborate with top researchers at its nodes across Europe.

By its very nature, the infrastructure encourages and enables users to carry out novel interdisciplinary studies based on ethical and privacy principles. In fact, all services offered are subject to a rigorous ethical assessment in order to inform and promote European principles in social analysis. The ethical use of data and research results is the fundamental principle of SoBigData, whose developers and users respect the FAIR (Findable, Accessible, Interoperable, Reusable) and FACT (Fair, Accurate, Confidential and Transparent) principles and the ELSEC (Ethical, Legal, Social, Economic and Cultural) perspective to ensure the qualitative growth of RI.

But scientists are not the only beneficiaries of SoBigData; it is also a data-driven innovation accelerator that facilitates collaboration between researchers, industry and start-ups to develop pilot and proof-of-concept projects. In fact, companies are among the stakeholders of the SoBigData Infrastructure, as innovation in big data and artificial intelligence is one of the main challenges for the development of a true Industry 4.0 and an even more digitised and intelligent society.

SoBigData RI has been selected by the European Strategy Forum on Research Infrastructures (ESFRI), as part of the Roadmap 2021. This puts the RI in a long-term perspective, with the aim of becoming the European platform for social mining and a service provider for the European Open Science Cloud (EOSC) initiative. SoBigData RI can be an important player in the European scenario, creating the next generation of responsible data scientists and providing services to connect researchers, industry, and policymakers to improve the quality of life.

Link:
[L1] http://www.sobigdata.eu/

References:
[1] R. Trasarti, V. Grossi, M. Natilli, et al., "SoBigData RI: European integrated infrastructure for social mining and big data analytics"in SEBD, 2022, pp. 117–124.
[2] V. Grossi, F. Giannotti, D. Pedreschi, et al, "Data science: a game changer for science and innovation", Int. J. Data Sci. Anal., vol. 11(4), pp. 263–278 (2021).

Please contact:
Roberto Trasarti, CNR-ISTI, Italy
info@sobigdata.eu, roberto.trasarti@isti.cnr.it

Data Analytics and AI Testing Facility Based on Research and Technology Infrastructure

by Samuel Renault and Patrik Hitzelberger (Luxembourg Institute of Science and Technology)

Opening a high-performance data analytics and AI research and technology infrastructure to small companies will support digitisation and co-development of Industry 4.0 projects. A test-before-invest approach is supported by the local digital innovation hub.

In Luxembourg, the national strategy on AI defines as its two first focus areas, the first being a living lab for applied AI, and the second enabling access to data [L1]. To support this, LIST has invested in a hybrid high-performance data analytics and AI research and technology infrastructure with the support of the European Regional Development Fund.

The project started from a collection of data analytics and AI use cases identified within LIST's own project portfolio. From these use cases, a clustering was done to identify the main AI and analytics domains to be covered. This clustering and the use cases were used to describe the requirements for a high-performance data analytics and AI infrastructure to be used in research and innovation projects with industrial partners.

A call for tenders helped to select the technology providers that supported the implementation of the infrastructure. The tender focused on a hybrid infrastructure combining on-premises clusters with proprietary and open-source software stacks and cloud services.

The two-year implementation project led to the deployment of an on-premises infrastructure composed of three main interconnected computing clusters. The first computing cluster supports the data analytics software stack of IBM (i.e. IBM Cloud Pak for Data) with modules allowing users with limited skills in data analytics and AI to perform the data ingestion, cleaning and preparation steps before building their AI models with assistance modules. This offers users the required support for starting their AI and data analytics journey. The second computing cluster supports an equivalent open-source software stack from Apache (Hadoop, Spark, Hive, NiFi among others), which is meant to be used by technically advanced users. This second stack offers more flexibility and finer control in all the data pipeline steps (ingestion, storage, cleaning, preparation, analysis, model development, monitoring, deployment) and allows a larger deployment flexibility. A third cluster dedicated to data visualisation and graphical computing was set up together with a large visualisation facility: a 7m-by-2m multitouch visualisation wall. This allows for visualisation and analysis of large-scale or high-dimensional data sets at a glance. After its initial deployment in early 2022, the infrastructure was updated with containerisation capabilities. The open-source (Hadoop) cluster was split to accommodate a small Docker cluster. The figure shows the current overview of the infrastructure and a representation of the visualisation facility.

Since its launch, this infrastructure has been used in national- and European-funded projects in various fields: energy grids analytics and optimisation for a digital twin project, space images quality improvement [1], industrial process improvement [2] and product quality measurement, among others. Beginning 2023, the infrastructure is proposed as a testing facility in two specific cases.

One is the support to a Testing and Experimentation Facilities (TEF) project in relation to smart cities and electromobility. Here, the infrastructure will be used to jointly develop pilot cases in electromobility with private companies, allowing companies to access data sets, reuse and process them through the infrastructure to provide innovative mobility and charging solutions (e.g. optimal location of charging stations, physically and logically in the energy grid) to cities and communities.

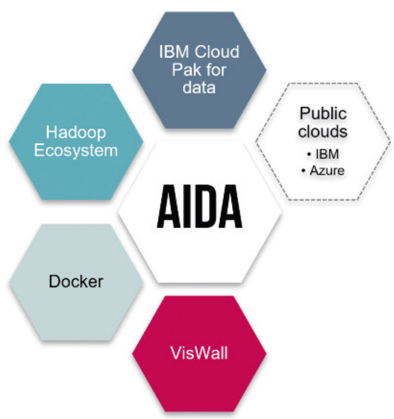

Figure 1: A use case of the visualisation facility: visualisation wall in the back and a tangible table in the front (left). Overview of the infrastructure's clusters and subscriptions (right).

This will also allow further development of interoperability standards such as OASC's minimal interoperability mechanisms [L2].

The other case is within Luxembourg's digital innovation hub where the technology infrastructure will be opened to industrial and manufacturing SMEs as a test-before-invest facility, with guidance and support through the infrastructure engineers and researchers. It will help companies to determine the worthiness of the technology for their specific business purpose, to tailor their artificial intelligence and data analytics projects and make informed investment decisions for their Industry 4.0 projects.

With numerous current developments both in data processing capabilities (e.g. AI models) and in data gathering, structuration and sharing (e.g. open data, data lakes, etc.), using dedicated data infrastructures will be a key enabler for innovation in private companies. Such infrastructures, like ours in Luxembourg, provide features to ensure data lineage, cleaning, description and access control, which are key to develop and test ideas before scaling up to the market. Where questions of sovereignty and locality of data, processing and expertise arise, technology infrastructures offered by research organisations should be taken into consideration.

Links:
[L1] https://gouvernement.lu/dam-assets/fr/publications/rapport-etude-analyse/minist-digitalisation/Artificial-Intelligence-a-strategic-vision-for-Luxembourg.pdf
[L2] https://oascities.org/minimal-interoperability-mechanisms/

References:
[1] O. Parisot, et al., "Improving accessibility for deep sky observation," ERCIM News, vol. 130, Jun. 2022.
[2] U. Iffat, E. Roseren, and M. Laib, "Dealing with high dimensional sequence data in manufacturing," Procedia CIRP, vol. 104, pp. 1298–1303, Nov. 2021, doi: 10.1016/j.procir.2021.11.218.

Please contact:
Samuel Renault, Luxembourg Institute of Science and Technology, Luxembourg
samuel.renault@list.lu

Curatorial Companionship: A New Framework for Managing High-quality Digital Cultural Content and Data

by Thomas Ballhausen (University Mozarteum Salzburg) and Stefan Gindl and Markus Tauber (Research Studios Austria FG)

In this paper, we present a novel approach for the intense management of data and creation of digital, cultural artefacts. We propose using interdisciplinary synergies between two new concepts, cultural companionship and ArtScrum. The former is an approach to combine skills and expertise from various, possibly disjunct domains; the latter is a structured, iterative concept adopted from digital engineering and specifically adapted to the needs of digital art creation.

Based on our experiences and findings in the context of technology-driven, arts-based research projects [L1] and the need for customised, curated and generated cultural content, i.e. data, in various fields of application [1] we introduce "curatorial companionship" as a new framework that focuses on the management of high-quality digital cultural content throughout its life cycle. This includes creation (i.e. digitalisation and creation of culture with and through digitalisation), preservation (including categorisation), consumption, etc. Based on a broad understanding of culture [L2], the current technical possibilities and the resulting need for digitally customised cultural content, we advocate for data management that is effective in all phases of the life cycle of the respective digital cultural object or product. Even though the topic is broader, meaning more research to do in the immediate future, we present here an initial investigation on the aspect of the creation of culture with and through digitalisation. Central ideas are not only the understanding of generative AI applications as tools that are used within human-machine assemblages, but also the examination of the creation of new cultural content as well as the possibility of synthetic reconstructions of incomplete, unrealised or even stylistically alternative artworks and cultural objects (i.e. synthetic fictions). The framework described in the following is meant as a tool to support involved stakeholders from interdisciplinary backgrounds.

Curatorial Companionship

Curatorial companionship brings together the domain knowledge of diverse fields, which are required for the creation and management of digital artwork. Furthermore, it adopts a set of archival practices and data management processes, including (i) the collection and preservation of existing digital artefacts, (ii) the generation of content and its refinement, as well as (iii) the selection and access of the final artefact.

This results in curatorial companionship requiring skills including data management, advanced digital skills, and, most importantly, contextual knowledge. The latter is specifically

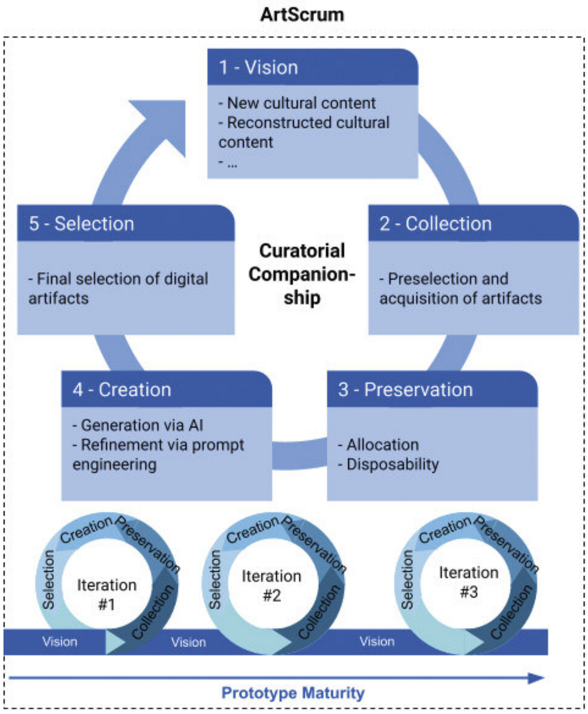

Figure 1: The iterative process of curatorial companionship in the methodology of ArtScrum.

relevant when dealing with cultural content. In a broader understanding of synthetic data that includes high-quality narratives and the generation of them, there is therefore not only the need for non-hierarchical principles of creatorship [2], but also the consideration of ethical principles to, e.g. avoid misuse [3]. An important aspect of how curatorial companionship creates cultural artefacts is described here as a process we define as ArtScrum.

ArtScrum - A novel methodological approach

Given the strong proximity to digital through the usage of tools such as generative AI (but not exclusively), we explore the adoption of the Scrum framework, a process management tool rooted strongly in software development. Scrum emphasises an iterative approach for the creation of digital artefacts, where consecutive prototypes undergo stringent inspection and adaptation before the final goal, the product, is reached. This stands in contrast to traditional, hierarchical approaches, where the product is defined in detail at the beginning. Since the creation of art is a non-linear process, we believe that ArtScrum is the ideal framework for the creation of art by generative AI. It aligns the practices of curatorial companionship, i.e. from the vision of (data) collection, preservation, creation (by generative AI) to the final selection of an artefact on a chain of iterative recursions, where the output of each preceding step becomes the input for the subsequent iteration (see Figure 1 for an overview). The last selection step marks the end of the process, where the final digital artefact is the result.

A Use Case – Synthetic Fictions

A potential use case and experiment for the application of generative AI in a setting of curatorial companionship and ArtScrum is the creation of synthetic fiction. Based on the science-fiction novel Dune (1965) by Frank Herbert, which has received renewed widespread attention due to a cinematic adaptation in 2021, we will show two examples for our understanding of synthetic fictions. Referring to its reconstructive perspective, that is, the completion of incomplete, fragmented or even wholly unrealised works, we refer here to the failed film adaptation by avant-garde director Alejandro Jodorowsky in the 1970s. It is not surprising that this project, which despite its failure is relevant to art history, has already undergone a first visualisation using generative AI, albeit one that is limited in scope and aesthetic quality [L3]. A second example is the implementation of synthetic fiction in the sense of an alternative stylistic realisation, which clearly differs from the actual remake from 2021. In this case, the protagonists as well as the colossal landscapes created for the latest film adaptation were realised in the style of the equally nightmarish and aesthetically distinct works of H.R. Giger – one of the artists involved in Jodorowsky's failed project [L4]. Both examples, appealing in their different ways, not only point to the current developments in the design and production of individual cultural content, but also show us the necessities for an implementation of curatorial companionship, especially for large-scale projects and professional (re)use.

Links:
[L1] https://www.spotonmozart.at/
[L2] https://policytoolbox.iiep.unesco.org/glossary/cultures/
[L3] https://www.youtube.com/watch?v=qShRMiUggKE&t=2s
[L4] https://www.youtube.com/watch?v=KZd4pz8aEfs&t=249s

References:

[1] M. Tauber, et al., “Social inclusion, health and content”, ERCIM News, vol. 130, 2022. (online)
[2] K. El Emam, L. Mosquera and R. Hoptroff, Practical Synthetic Data Generation, O’Reilly, Sebastopol, CA, 2020
[3] L. Floridi, The Ethics of Information, Oxford UP, Oxford, 2013.

Please contact:

Thomas Ballhausen
Universität Mozarteum Salzburg
thomas.ballhausen@moz.ac.at

Data-aware Declarative Application Management in the Cloud-IoT Continuum

by Jacopo Massa (University of Pisa / CNR-ISTI), Stefano Forti, Patrizio Dazzi and Antonio Brogi (University of Pisa)

Nowadays billions of devices are connected to the Internet of Things and can reach computing facilities along the Cloud-IoT continuum to process the data they produce, leading to a dramatic increase in the number of deployed applications as well as in the amount of data they need to crunch. Following a continuous reasoning approach to speed up the decision-making process, our research proposes a declarative and data-aware solution to determine service-based application placements over the Cloud-IoT continuum while meeting functional and non-functional application requirements.

In recent years, the Internet of Things (IoT) has become increasingly integrated with our everyday lives, driving the definition of new computing models and paradigms. According to the latest reports, there will be more than 30 billion devices connected to the Internet by the end of 2023, leading to a massive amount of data being processed and acted upon, generated by several sources at the edge of the network, majorly IoT devices. There are several reasons why the cloud paradigm alone cannot always meet the constraints among end devices and cloud servers, not to mention data processing and transmission costs along physical links. Secondly, many applications require real-time interactions, i.e. quasi-real-time data exchange, which often cannot be offered due to the significant end-to-end delay (i.e. latency) between nodes. Last but not least, security is becoming more and more important since many applications work with sensitive data that should not traverse the Internet, e.g. for privacy reasons.

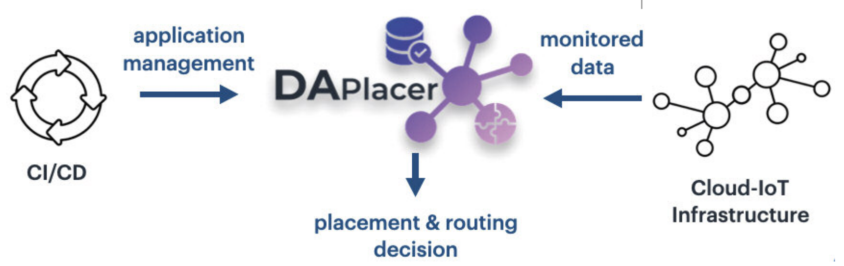

Figure 1: DA-Placer workflow.

To address these challenges, several Cloud-IoT paradigms – e.g. fog, edge, mist, and osmotic computing – have been proposed. They exploit heterogeneous computing capabilities along the Cloud-IoT continuum (e.g. smartphones, access points, gateways, datacentres) to process data close to their IoT sources. In between the cloud and the IoT, computational nodes act both as processing capabilities closer to the devices and as filters over data streams directed towards the cloud.

The previously mentioned paradigms require application services to be adequately placed along the available Cloud-IoT resources to meet all functional and non-functional application requirements. Hence, deciding where to deploy (and possibly later migrate) application services along the Cloud-IoT continuum has been largely studied in the literature. However, despite taming the data deluge and achieving data-awareness being among the main motivations of Cloud-IoT computing, to the best of our knowledge, the characteristics of the data (e.g. security needs, volume, velocity) processed by the application have been used only marginally to drive placement decisions.

To this end, we devised a Prolog open-source tool [L1], named DA-Placer [1], which inputs a declarative description of a multi-service application and its requirements, within a declarative description of a multi-layered infrastructure with dual capabilities. It exploits a declarative strategy that maps each application service to a compatible node, without exceeding nodes and links capabilities, and giving as output the set of placement and routing decisions (see Figure 1). Our declarative strategy takes into account requirements in terms of software, hardware and IoT, but also a taxonomy of security constraints and QoS (quality of service) requirements (latency and bandwidth). Since we are in a data-aware context, we also deal with data characteristics, such as their size, transmission rate and their sources and targets. Prolog has been chosen for the simplicity of its syntax and for managing and updating the code. Above all, the backtracking technique exploited by the Prolog reasoner to find a (possibly existing) solution, ensures an exhaustive search in the solution space.

To tame the EXPTIME complexity of the considered problem for prompter decision-making at runtime, DA-Placer exploits a continuous reasoning approach [2]. Once a deployment has been enacted according to a found placement and routing, continuous reasoning tries to reduce the size of the considered placement problem instances at runtime, by focusing on re-placing those services and data routings that cannot currently meet their requirements. This can mainly happen for two reasons:

- Due to changes in the monitored Cloud-IoT infrastructure (e.g. node crash or overloading, link QoS degradation) that prevent meeting application requirements
- Due to changes in the declared application (e.g. service removal/addition, requirements update, changes in the data types handled) that require (un)deploying services or migrating existing ones.

When possible, after identifying the deployment portion affected by the changes above, continuous reasoning attempts to determine a new placement and data routing only for such a portion.

The research that led to the prototyping of DA-Placer is only in its infancy and will continue to expand the knowledge barrier in several directions, including further management decisions (e.g. backup/replicas), multi-objective optimisation (e.g.

energy consumption, operational costs), or exploitation of local capabilities to guarantee resilience to churns in a distributed/decentralised context, as well as theoretical results to prove the correctness and completeness of the devised declarative approaches.

This work has been partly supported by the EU H2020 ACCORDION project, and by the project "Energy-aware management of software applications in Cloud-IoT ecosystems", funded with ESF REACT-EU resources by the "Italian Ministry of University and Research" through the "PON Ricerca e Innovazione 2014—20".

Links:
[L1] https://github.com/di-unipi-socc/daplacer

References:
[1] J.Massa et al., "Data-Aware Service Placement in The Cloud-IoT Continuum", in Service-Oriented Computing in CCIS,vol. 1603, p.139-158, 2022.
[2] Forti et al., "Declarative Continuous Reasoning in the Cloud-IoT Continuum" in J. of Logic and Computation, 2022.

Please contact:
Jacopo Massa, University of Pisa / CNR-ISTI, Italy
jacopo.massa@phd.unipi.it

IndeGx: An Index of Linked Open Datasets on the Web

by Pierre Maillot, Catherine Faron, Fabien Gandon and Franck Michel (Inria)

Where can we find the data we need for a given task? IndeGx aims to create an index of knowledge graphs published on the web, in the form of linked open datasets, for humans and machines alike. This framework provides descriptions of the knowledge graphs to draw a picture of their content, quality and compliance with the standards. A companion webpage is also provided to support data providers in the description of their datasets in compliance with the latest best practices.

In recent years, a large number of knowledge graphs have been built and published on the web in the form of RDF datasets, in fields as diverse as linguistics or life sciences, along with other general datasets such as DBpedia or Wikidata. The reliable exploitation of these datasets requires specific knowledge about their content, access points and commonalities. Yet, it is usually difficult to have such a clear and up-to-date description as most datasets have neither a machine-readable nor a human-readable description, and not all access points can handle the complex queries required to automatically generate such descriptions. The data providers are commonly regarded as responsible for describing the datasets they publish. However, this requires specific efforts, costs, and skills that some providers, who are not necessarily experts in semantic web technologies, do not have or cannot afford. In particular, these descriptions rely on a deep understanding and joint use of specialised vocabularies and there is no standard model or tool for generating and updating these descriptions.

IndeGx is a transparent, declarative, collaborative and extensible framework designed to generate the description of a knowledge graph solely based on information that can be extracted from a SPARQL endpoint that serves that knowledge graph. It is part of the ongoing effort to help humans and machines in the use of knowledge graphs on the web. It creates an open repository of descriptions to guide agents in selecting knowledge graphs, and supports a variety of use cases. For instance, they could be used to fuel the faceted search of a dataset catalogue meant for human agents, or a query federation engine could leverage the statistics on the usage of the classes and properties in a dataset to efficiently rewrite a query over multiple graphs.

IndeGx relies on generation rules expressed in standard RDF vocabularies and SPARQL to yield common metadata features such as provenance information and lists of classes and properties, as well as new kinds of metadata features not covered by previous approaches, such as quality indicators, lists of vocabularies and new statistics. Furthermore, the processing of a dataset at different points in time makes it possible to track its evolution. The descriptions generated by IndeGx are represented in RDF and published as a regular, public knowledge graph.

The rules that generate the descriptions are written with the same vocabulary as the one used by the W3C to describe test suites in RDF. When joined with the execution traces that are also kept and represented in RDF, the process of generating a dataset description is fully transparent and traceable. Moreover, the rules and the IndeGx application are available under open licences [L1], and anyone can replicate them, extend them or create their own set of rules, as long as they use the same vocabulary.

The KartoGraphI website [L2] relies on IndeGx to draw a picture of the open sources of the semantic web. At the time of writing, it provides different visualisations computed to monitor, over an 8-month period, 339 datasets with SPARQL endpoints retrieved from well-established dataset catalogues such as the Linked Data Cloud website and Wikidata. The details of

Figure 1: Geolocalisation of the endpoints described in IndeGx as shown on the KartoGraphI website.

the results of that experiment can be found in [1]. Despite the multiple technical constraints, we obtained detailed statistics on 54% of the datasets. Such statistics serve as the base of exploration, selection and federation methods. For instance, on average, the dataset's endpoints support 85% of the SPARQL 1.1 features, and knowing this allows for advanced usage of these datasets.

Figure 2: Graph of the endpoints (in green) and vocabularies (in blue) of the 339 datasets monitored.

Figure 1 shows the geolocalisation of the endpoints of the datasets based on their URL. Figure 2 shows the graph of endpoints and datasets. The three green nodes in the top-left part of the graph, connected to a group of vocabularies around them, are the three mirror sites of Linked Open Vocabularies. The vocabularies that surround them are listed in the Linked Open Vocabulary dataset but not used in any other. The blue nodes grouped at the centre of the graph are the most-used vocabularies; they are surrounded by the majority of the endpoints. Vocabularies used in a single dataset, or datasets using a single vocabulary, radiate from this middle group. Likewise, Figure 3 shows which endpoint uses which (meta)vocabulary (e.g. SKOS, SPIN, SHACL) thus providing an idea of the type of content that can be found in the corresponding dataset.

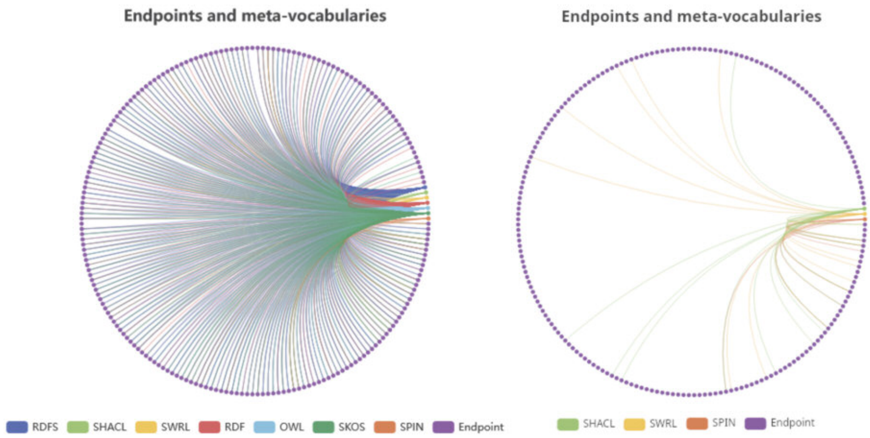

Figure 3. Graph of the endpoints connected to the different (meta-)vocabularies.

Only 10% of the analysed datasets contain provenance information. To encourage data providers to write provenance metadata in RDF and more generally rich descriptive metadata, we created the Metadatamatic webpage [L3]. It allows users to generate an RDF description by filling out an interactive form. The description is guaranteed to follow the latest best practices for dataset description, and part of the features can be automatically extracted from the dataset if it is served by a running endpoint.

Together, IndeGx, KartoGraphI, and Metadatamatic provide a set of tools to encourage, support and monitor the description of linked open datasets.

Links:
[L1] https://github.com/Wimmics/dekalog
[L2] http://prod-dekalog.inria.fr/
[L3] https://wimmics.github.io/voidmatic/

References:
[1] P. Maillot, O. Corby, C. Faron, F. Gandon and F. Michel, "IndeGx: A model and a framework for indexing RDF knowledge graphs with SPARQL-based test suits," Journal of Web Semantics, 2023, https://doi.org/10.1016/j.websem.2023.100775, https://hal.inria.fr/hal-03946680
[2] P. Maillot, O. Corby, C. Faron, F. Gandon and F. Michel, "KartoGraphI: Drawing a map of linked data," in ESWC 2022 - 19th European Semantic Web Conferences, May 2022, Hersonissos, Greece. https://hal.inria.fr/hal-03652865.

Please contact:
Pierre Maillot, University Côte d'Azur, Inria, CNRS, I3S, France
pierre.maillot@inria.fr

Promoting Federated and Trusted Sharing and Trading of Interoperable Data Assets and Data-Driven Intelligence – The PISTIS Platform

by Kyriakos Stefanidis (ISI/ATHENA), Sotiris Kousouris (SUITE5), Yury Glikman (Fraunhofer FOKUS) and Christos Alexakos (ISI/ATHENA).

PISTIS is a European research project that includes 29 partners (universities, research institutes, SMEs and industry) spanning eleven EU countries and brings forward a reference-federated data sharing/trading and monetisation platform for secure, trusted and controlled exchange and usage of proprietary data assets and data-driven intelligence (derivative data assets).

As part of the EU strategy for data, the prominence of data has led to the adoption of new legislative acts, i.e. the Data Governance Act and the Data Act, as well as to the emergence of centralised and/or local/edge data spaces that bring together relevant data infrastructures and governance frameworks to facilitate data pooling and sharing within and across different organisations, both in domain-specific settings and across sectors [1]. However, such data spaces are in an embryotic stage in certain industries, and organisations only have fragmented data at their disposal, while still operating in silos.

From a data provisioning/supply perspective, organisations across industries tend to focus more on the collection, access, and analysis of massive amounts of owned data for their own purposes, rather than transforming it, quantifying its business value and/or sharing it for the benefit of other stakeholders.

From a data-quality perspective, the different data providers often make their data available without continuously ensuring its requisite quality and reliability. The lack of globally agreed standards per industry (despite the proliferation of numerous, yet incompatible data models/schemas/ontologies and metadata standards) is responsible for the ever-present data interoperability challenges and further hinders any data re-use effort.

When organisations decide to share or trade their data, reaching consensus on the IPR (intellectual property rights) and licensing aspects is not a straightforward process and may require various iterations among their legal representatives regarding the modalities for data sharing/trading, under which terms and for which purposes data is to be re-used.

Although the data sharing contracts are often based on distributed ledger technologies and stored in an immutable manner, monitoring the actual data usage in compliance with geography-based data use regulations, as well as maintaining oversight that the data is provided and consumed as agreed in the contract terms, is very hard to automatically enforce on any data platform.

In response to the above challenges, PISTIS brings forward a reference-federated data sharing/trading and monetisation platform for secure, trusted and controlled exchange and usage of proprietary data assets and data-driven intelligence (derivative data assets). In its effort to unleash the actual and still-untapped data potential, PISTIS will advance the currently available techniques and technologies, such as federated data discovery and sharing, distributed ledger technologies (DLT), data non-fungible tokens (NFTs), AI-driven data quality assessment and monetisation, to build trust among data providers, data owners and data consumers.

PISTIS will establish the underlying methodological and technical foundations across different axes that will complementarily and interactively work together as follows. The Federated Data Management, Interoperability & Governance Axis aims at collecting, curating, securing and fully controlling the data made available through each organisation's data space. The Federated, Secure Data Sharing Axis concerns the effective management and on-chain storage of multi-party data contracts across their negotiation, settlement, remuneration, enforcement and monitoring life, as well as the secure peer-to-peer data transfer and usage monitoring mechanisms. The Data Valuation and Monetisation Axis systematically articulates and recommends an appropriate target value, indicatively taking into consideration the "cost" (total costs to generate, collect, store and maintain such data), the "income" (the value added by the underlying data itself), and the "market" (what other stakeholders pay for comparable assets).

As Figure 1 suggests, PISTIS can be seen on a macroscopic level as a set of services that are operating locally in each data space deployment, and are governed by a cloud-based control plane that is used for data discovery, transactions execution, enforcement and auditing. They are also used for horizontal aspects such as design and propagation of common information models and semantic information models, connection to real-world financial institutions and user authorisation based on global credential providers (such as eIDAS2.0). As such, no data leaves the premises of a data owner/producer during the data preparation, analysis and exchange preparation steps, and only metadata are stored in global, permissioned ledgers or metadata catalogues. In essence, data-exchange transactions take place on a peer-to-peer base, again refraining from using cloud repositories as the middleman. In addition, data brokers can perform limited data access and management operations for safeguarding data, maximising the value for the supply side (and lowering the fees for the demand side).

The data sharing-relevant operations will be executed without handing over control and ownership of data, following a secure and trusted paradigm where assets remain with the owner at all stages, exposing only metadata to facilitate the orchestration of the different services. At the same time, the demand side will be in a position to discover data assets and have at its disposal information regarding their quality, their lineage as well as the different options of acquisition, in order to be able to engage in a transaction that will be initiated, negotiated, settled and verified for acquiring an asset of interest and for tracking its subsequent usage based on agreed terms.

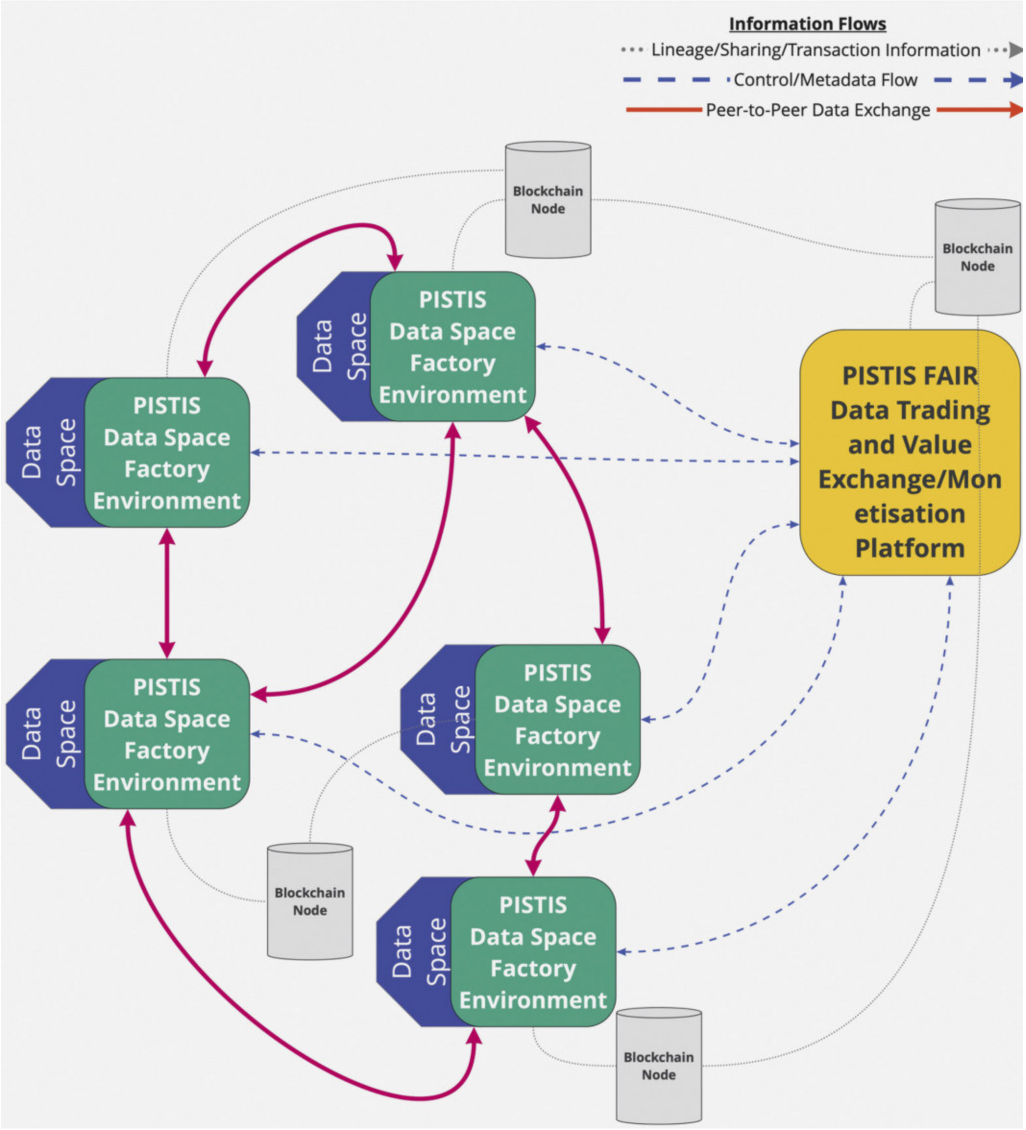

Figure 1: The PISTIS reference architecture, macroscopic view.

PISTIS will be tested in real-life settings in three large-scale demonstrators' hubs in Greece (mobility and urban planning), Spain (energy) and Austria & Germany (automotive).

The mobility and urban planning hub will focus on facilitating data trading and sharing amongst stakeholders in aviation, public transport and public administration. The hub includes AIA (Athens International Airport), GOLDAIR (Ground Handling Service Provider), OASA (Athens Public Transportation System, including bus and metro lines) and DAEM (City of Athens IT department), which act as data consumers and/or providers depending on the use case.

The energy demonstrator hub will focus on ensuring the resilient operation of the distribution grid through the utilisation of the flexibility capacity that can be offered by local prosumers and triggered by the aggregator. The hub involves CUERVA (DSO and owner of the Granada Living Lab infrastructure), OMIE (Market Operator) and BAMBOO (Aggregator).

The automotive demonstrator hub will use various data sources, e.g. connected vehicle data from car manufacturers, vehicle trip data, weather data, map data, and air quality data to feed data-hungry use cases, focusing on traffic quality assessment in urban areas and driving style and risk assessment. The hub includes CARUSO (Data provider - Marketplace), VIF (Virtual Vehicle Technologies), TRA (Traffic quality assessment) and UBIMET (weather data).

Link:
[L1] https://cordis.europa.eu/project/id/101093016

Reference:
[1] EC (2022), "Commission Staff Working Document on Common European Data Spaces," SWD(2022) 45 final. https://data.consilium.europa.eu/doc/document/ST-6532-2022-INIT/en/pdf

Please contact:
Kyriakos Stefanidis, ISI/ATHENA, Greece
stefanidis@isi.gr

Dealing with UAV-downwash Effects during Airborne Detection of Localised Gas Plumes

by Philip Taupe (AIT Austrian Institute of Technology), Martin Litzenberger (AIT Austrian Institute of Technology), Dieter Rothbacher (CBRN Protection GmbH) and David Monetti (Skyability GmbH)

Unmanned Aerial Vehicle (UAV)-borne gas detection and measurement can drastically reduce operational risks for the involved first responders in scenarios where hazardous gases may be present, e.g. accidents in industrial plants or gas leaks/accumulations in urban settings. However, measurement accuracy and detection abilities of gas sensors can be severely affected by the UAV's rotor downwash, thus leading to erroneously low gas-concentration readings. We have investigated the applicability and operational parameters of probing tubes extending from under the UAV to escape downwash effects and thereby improve measurement results.

Multicopter UAVs are becoming an important tool for first responders in disaster situations to quickly develop an operational picture and improve situational awareness for decision makers. Especially in the case of (urban) search and rescue scenarios involving (partially) collapsed buildings, as well as in the case of emergency response in an industrial complex, UAVs are valuable assets. In such scenarios, the detection and measurement of the concentration of hazardous gases is sometimes also a critical requirement of the mission. If gas leaks and associated plumes occur only locally (i.e. the gas plume is relatively small in relation to both the UAV and the area of operation), UAV-borne gas sensors are severely affected by rotor downwash. The UAV's rotors create a strong airflow close to and especially below the vehicle, which can affect detection and measurements of such gas plumes in two ways: 1. Dilution, by mixing with ambient air, and 2. Shielding, by moving the gas plume away from the gas sensor. The boundaries between both effects are fluid; however, both lead to uncertain or invalid detection and measurement results.

Previous studies on the airflow effects on concentration measurements using a UAV-mounted gas monitor were conducted with the UAV hovering stationary in a chamber with a constant NOx (nitrogen oxide) concentration, but without considering highly local gas point sources. [1]. A MEMS metal-oxide (MOX) sensor has been evaluated for UAV-based gas concentration measurement, with a bottle of propylene alcohol as a tracer substance in an outdoor environment [2]. The experiment was done under uncontrolled environmental/wind conditions and therefore the effect of gas dilution by the UAV rotor downwash could not be evaluated. Indoor tests with ethanol as a tracer gas and a MOX gas detector mounted at the top of the fuselage of a small UAV have been presented in [3]. In this investigation, the airflow patterns around the UAV and with respect to the gas sensor have also been analysed using dry ice. Probing tubes have not been investigated as the MOX sensor

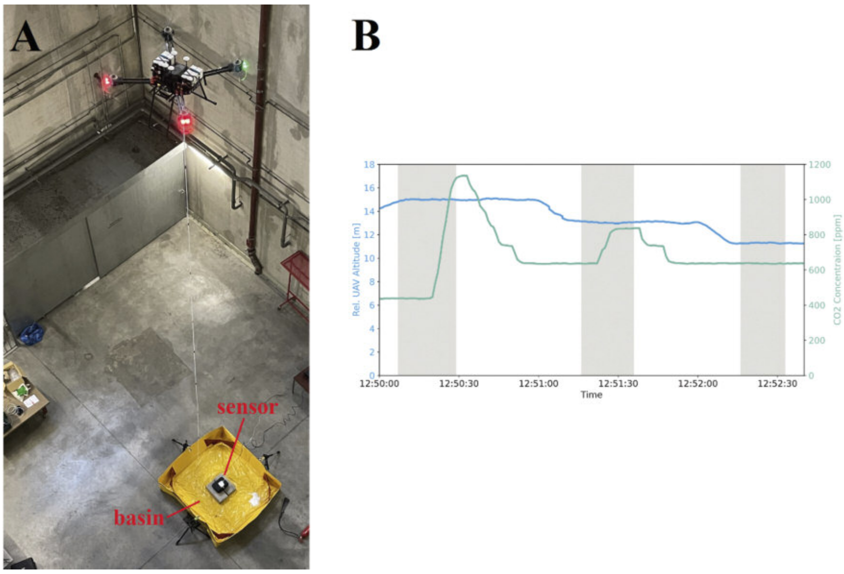

Figure 1: (A) Test setup with UAV hovering above gas basin. (B) Measured CO_2 concentration at varying hover altitude in response to CO2 bursts from a fire extinguisher. Blue line (lhs): UAV altitude; Green line (rhs): CO_2 concentration; Shaded areas: Duration of CO_2 bursts.

was mounted on top of the fuselage. McKinney et al. have installed an array of short sampling tubes for VOC under the frame of the professional-grade UAV [4]. Pressure distribution around the UAV has been investigated using computational fluid dynamics simulations to assess the impact of rotor downwash and airflow turbulence on the sampling efficiency. The impact of the downwash effect on gas detection and measurements is evident and one possible solution is to install a long probing tube to move the gas intake well below the downwash zone. The effectiveness of this solution has, however, not been addressed in any of the above-mentioned publications.

We have evaluated the use of probing tubes for the improvement of UAV-based gas detection and concentration measurements. The aim was to experimentally validate the approach and derive operational parameters for first responders in the field. We also investigated when downwash effects would not only impair quantitative concentration measurements, but also qualitative gas detections against an ambient background.

The experiments to detect CO_2 have been conducted inside a tall building to reduce the influence of environmental conditions, such as wind. Figure 1(A) shows the used professional-grade UAV hovering over the centre of the test basin. The basin (2m ×2 m footprint, 40cm in height) was used to retain some CO_2 close to the sensor (Sniffer4D v2, Soarability). We monitored the CO_2 concentration and the level of the basin's upper edge. The UAV started in a hover position 15m above the sensor and subsequently lowered its position in 2m intervals. At each hover position, the basin was flooded with a burst of gasous CO_2 using a fire extinguisher to simulate an active gas leak.

Upticks in the measured gas concentration relative to the ambient baseline have been recorded (Figure 1(B)). At altitudes of 15m and 13m, a clear response to the CO_2 burst is visible (subject to sensor response time), before the measured concentration falls back to the ambient level (note that the ambient level increases over time as more CO_2 is generally present). Response levels decreased with decreasing altitude, and at 11m no response was recorded. This indicates a clear boundary at which not even a qualitative assessment of the presence of hazardous gases is possible and misleading gas monitoring could pose a serious threat to first responders.

We have investigated gas plume dilution/shielding by UAV rotor downwash in a realistic test setup using a CO_2-filled basin and a professional-grade UAV together with a CO_2 sensor. Our method has shown that the ability to detect increased gas concentrations against an ambient background is severely impaired below a certain minimum hover altitude (in the range of 11 to 13m for the specific UAV used during the experiment), suggesting that a probing tube should always have a length that is considerably longer than this absolute minimum altitude.

This study was carried out as part of the research project "UAV-Rescue" which has been funded by the Austrian security research programme KIRAS of the Federal Ministry of Agriculture, Regions and Tourism (BMLRT). We would also like to express our gratitude towards the fire brigade of the city Vienna (Berufsfeuerwehr Wien), who kindly supplied facilities and materials used during the experiments.

References:

[1] J. L. Brinkman and C. E. Johnson, "Effects of downwash from a 6-rotor unmanned aerial vehicle (UAV) on gas monitor concentrations," Mining, Metallurgy & Exploration, Bd. 38, Nr. 4, S. 1789–1800, Aug. 2021, doi: 10.1007/s42461-021-00436-5.
[2] M. Rossi, et al., "Gas-drone: portable gas sensing system on UAVs for gas leakage localization," in Proc. IEEE SENSORS 2014, pp. 1431–1434. doi: 10.1109/ICSENS.2014.6985282.
[3] J. Burgués, et al., "Smelling nano aerial vehicle for gas source localization and mapping," Sensors, Bd. 19, Nr. 3, S. 478, Jan. 2019. doi: 10.3390/s19030478.
[4] K. A. McKinney et al., "A sampler for atmospheric volatile organic compounds by copter unmanned aerial vehicles," Atmos. Meas. Tech., Bd. 12, Nr. 6, S. 3123–3135, 2019. doi: 10.5194/amt-12-3123-2019.

Please contact:

Philip Taupe, AIT Austrian Institute of Technology
philip.taupe@ait.ac.at

Documenting Data and Knowledge about Feminisms and Politics in the Interwar Balkans with Synthesis-Core

by Pavlos Fafalios, Lida Charami (FORTH-ICS) and Katerina Dalakoura (University of Crete)

The project FE.P.I.B., hosted by the University of Crete and funded by the Hellenic Foundation for Research and Innovation (HFRI), explores the contacts, relationships and cooperation between the feminist/women's organisations of the Balkan and Central European countries, and particularly the entanglement of their emancipatory policies with the politics in the region during the interwar period. Researchers of FE.P.I.B. collaborate with the Centre for Cultural Informatics of FORTH-ICS for exploiting the information system Synthesis-Core as the project's data and knowledge management platform.

"Feminisms and Politics in the Interwar Balkans (1923-1939)" (FE.P.I.B.) is a research project in the field of transnational history of feminism, which aims to recover, fully document and analyse the collaborations among the interwar feminist/women's associations in the Balkan and Central European countries, and to explore the entanglement of their activities and feminist policies with the politics and diplomacy established or pursued in the region during the interwar period [L1]. More precisely, the research focuses on the study of the most representative case of collaboration among Balkan women's organisations, the Little Entente of Women (LEW) (1923-1939), involving feminist organisations from Greece, Bulgaria, Kingdom of Serbs, Croats and Slovenes /Yugoslavia (since 1929), Romania, and two Central European countries, Czechoslovakia and Poland [1]. The project also investigates the participation of feminist and women's associations of the Balkan countries in the Balkan Conferences (1930-1934) and their activities in this context, as a continuation of the LEW's activities and, most notably, as a direct involvement in government-planned policies in the Balkans and the wider region.

To cope with the project's research objectives, the FE.P.I.B. research team collects and studies a large and diverse set of archival material of different types related to feminisms and politics in the interwar Balkans and Central Europe, including announcements, printed and not-printed articles, biographical documents, certificates, correspondences, resolutions, diaries/notebooks, legislations, letters, meeting minutes, press releases, reports, photos, images and many others. The source of this archival material is mainly libraries and archives in Balkan and Central European countries, including Greece, Serbia, Bulgaria, Romania, Slovakia, Czech Republic and Poland.

The collected archival material needs to be digitally organised, documented, and then studied for extracting important evidence related to the project's objectives that can facilitate comprehension of entanglements, convergences and divergences, interpretations, possible causes, influences, and evolutionary trends. In particular, data management needs include: i) recording of metadata and content-based information about the different types of collected archival material, ii) documentation of rich information about the studied 'entities of interest', in particular about feminist associations (regional, national, international), political and socio-cultural associations, conferences, meetings, personages (feminist leaders/activists and other historical figures), iii) documentation of information about secondary entities related to the main entities, such as archives, libraries, collections, locations, historical events, and other, and iv) searching the data by exploiting the rich interconnections about the involved entities.

To cope with these data management needs, the FE.P.I.B. team collaborates with the Centre for Cultural Informatics [L2] of FORTH-ICS to create a configuration of the Synthesis-Core system [L3], especially adapted to be used as the project's main data management platform. Synthesis-Core is an information system for the scientific, administrative and collaborative documentation of cultural entities, which is configured based on specific use cases and needs. It is web-based, multilingual, and focuses on semantic interoperability. It enables data exchange between computer systems with unambiguous/shared meaning by making use of a standard model for data representation, namely the formal ontology CIDOC CRM (ISO 21127:2014) [L4]. The aim is the production of sustainable data of high value and long-term validity that can be reused beyond a particular research project. Technically, it utilises XML technology and a multi-layer architecture, offering flexibility in terms of versioning, work-

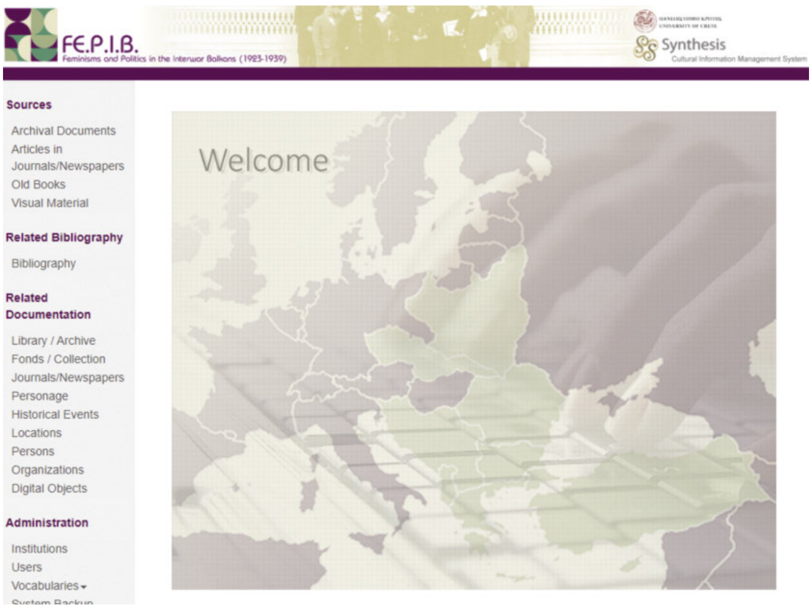

Figure 1: The home page of Synthesis as currently configured for the FE.P.I.B. project.

Figure 2: A part of the documentation card of an entity belonging to the entity type "archival document" (corresponding to a letter from Margery Corbett Ashby to Avra Theodoropoulou, 23/08/1927).

flow management and data model extension. The system has been widely used in different contexts, including cultural heritage / archaeology [2] and history of art [3].

In Synthesis-Core users create and document entities belonging to a set of pre-configured entity types. Each entity type has its own data structure (documentation schema), carefully designed to be fully compatible with CIDOC CRM (i.e. a direct mapping between a schema and CIDOC CRM can be defined). A documentation schema is XML-based and contains a set of fields organised in a hierarchical (tree-like) structure. The leaves in this tree-like structure are the documentation fields into which users enter data. Each documentation field is assigned to one of the following data types: link to other entity, link to a vocabulary term, link to a thesaurus term (exploiting the thesaurus management system THEMAS [L5]), unformatted free text, formatted free text, link to a web resource (URL), number, time expression (with support of historical time expressions, such as 'ca. 1920', 'early 16th century', etc.), location coordinates (by selecting a point or polygon on a map), location ID (TGN or Geonames), digital file(s).

In the context of FE.P.I.B, the following entity types have been currently configured that allow documenting information about the collected archival material and bibliography: i) archival document, ii) article in journals/newspapers, iii) old book, iv) visual material, v) bibliography, vi) library/archive, vii) fonds/collection, viii) journal/newspaper, ix) personage, x) location. Figure 1 shows the home page of Synthesis-Core, as configured for the case of FE.P.I.B, while Figure 2 shows a part of the documentation card of an entity belonging to the entity type archival document, which corresponds to a letter from Margery Corbett Ashby to Avra Theodoropoulou (of 23/08/1927).

Other system functionalities include map visualisation (for entities that have a relation to a location), versioning of the documented data, documentation in multiple languages, support of different user roles (system administrator, agency/group administrator, editor, guest), data export (in XML or RDF, as well as in PDF or DOCX based on predefined templates), and advanced entity search (within or across entity types).

The system configuration is still under development. Currently, we are in the process of designing the documentation schemas of the remaining entities of interest, in particular national organisations, regional-international associations, and conferences/meetings. In parallel, the research team (consisting of ten researchers) has started documenting information about the collected archival material, bibliography, and personage. The current number of documented entities per entity type is the following: 98 archival documents, 645 articles, 36 old books, 16 visual material items, 182 bibliographical item, 18 libraries/archives, 16 fonds/collections, 45 journals/newspapers, 106 personages, 14 locations.

Acknowledgements: This work has received funding by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "2nd Call for H.F.R.I. Research Projects to support Faculty Members & Researchers" (Project Number: 03050).

Collaborators: Konstantina Konsolaki (FORTH-ICS), Chrysoula Bekiari (FORTH-ICS), Dimitra Samiou, Vanessa Geragori (Univ. of Crete), Charitomeni Giasafaki (Univ. of Crete), Eleftheria Papastefanaki (Univ. of Crete), Eleni Fournaraki (Univ. of Crete), Maria Bucur (Indiana Univ.), Daskalova Krassimira (Univ. of Sofia), Gabriela Dudeková Kováčová (Slovak Academy of Sciences), Ivana Pantelić (Institute of Contemporary History, Belgrade).

Links:
[L1] http://feminisms-politics.ia.uoc.gr/
[L2] https://www.ics.forth.gr/isl/centre-cultural-informatics
[L3] https://www.ics.forth.gr/isl/synthesis-core
[L4] https://cidoc-crm.org/
[L5] https://www.ics.forth.gr/isl/themas-thesaurus-management-system

References:
[1] K. Dalakoura, "Feminisms and politics in the interwar period. The Little Entente of Women (1923-1939)," Aspasia, vol. 16, Berghahn Journals, 2022.
[2] C. Bekiari et al., "Building comprehensive management systems for cultural-historical information," CAA, Archaeopress, Oxford, 2014.
[3] P. Fafalios et al., "Towards semantic interoperability in historical research: documenting research data and knowledge with synthesis," ISWC, Springer, 2021.

Please contact:
Pavlos Fafalios, Centre for Cultural Informatics
FORTH-ICS, Heraklion, Greece
fafalios@ics.forth.gr

Katerina Dalakoura, Department of History and Archaeology, University of Crete, Rethymno, Greece
dalakoua@uoc.gr

Survival Skills in the Digital Era: Helping Teachers to Develop Digital Competences

by Angela Fessl and Katharina Maitz (TU Graz), Monica Divitini (NTNU)

Digital technologies have a great potential in supporting teaching practices. However, as also experienced during the COVID-19 pandemic, many teachers lack both the training and support they need to effectively use technologies in class or lecture halls. The DIGIVID project addresses this challenge with a system providing a comprehensive curriculum accessible through an innovative learning tool.

As are many other professionals, teachers are expected to use digital technologies in an effective way to improve and innovate their practices. In this context, one of the crucial challenges that teachers face is the acquisition of digital competences. Teachers must succeed as lifelong learners, both individually and collectively, to be able to adopt digital technologies and media to manage their teaching and promote learning. Teachers also need to help their students to develop adequate digital competences and 21st century skills. To address these challenges, various frameworks have been developed, nationally and internationally, to offer guidelines of relevant competencies and skills for citizens and, more specifically, for educators. See for example the DigCompEdu [1] framework.

Despite the recognised importance of digital competences, many teachers do not get access to adequate training and are struggling with developing the necessary digital competences. Even if a lot of material might be available, many teachers are not able to integrate the training in their busy workday. This challenge is addressed by the DIGIVID project [L1]. The project provides a curriculum accessed through a learning system that is teacher centred. The overall goal is to promote self-regulated and reflective learning that is sustainable for teachers. The curriculum, based on DigCompEdu [1], includes resources for learning about basic digital competences, examples of digital methods and tools that can be applied in class, knowledge about digital lifeworlds of youth, and tips for inclusion and accessibility. The curriculum is accessed through the Moodle-based DIGIVIDget. DIGIVIDget visualises the curriculum along competence-oriented learning goals, making the content easily accessible. DIGIVIDget supports teachers in going through the curriculum at their own pace, keeping track of their progress, and reflecting on how they can use what they learn in their practice.

DIGIVIDget is designed to provide a clear overview of what needs to be learned. At the higher level, content is organised into modules that help teachers to progress from basic to advanced competences. Each module includes different topics to organise the content into macro-areas. Within each topic, the content is organised around well-defined and competence-oriented learning goals. Learning goals are a key element for effective and successful teaching since they describe what a learner should be able to do after a specific learning experience [2]. Each learning goal consists of a set of 10–15 micro learning cards. In this way, teachers can go through small nuggets of content at a time, supporting flexible access to content. Teachers can navigate through micro learning cards sequentially or jumping directly to one specific card. To provide an engaging experience, content is visualised in different formats, including text, videos, and quizzes. The content is enhanced with reflective questions integrated in the middle and at the end of each learning goal. Reflective questions are designed as triggers to help teachers think about their learning process and how to integrate what they learn in their everyday practices. All the reflection notes that teachers take during their learning process are easily accessible at any time through a reflection diary.

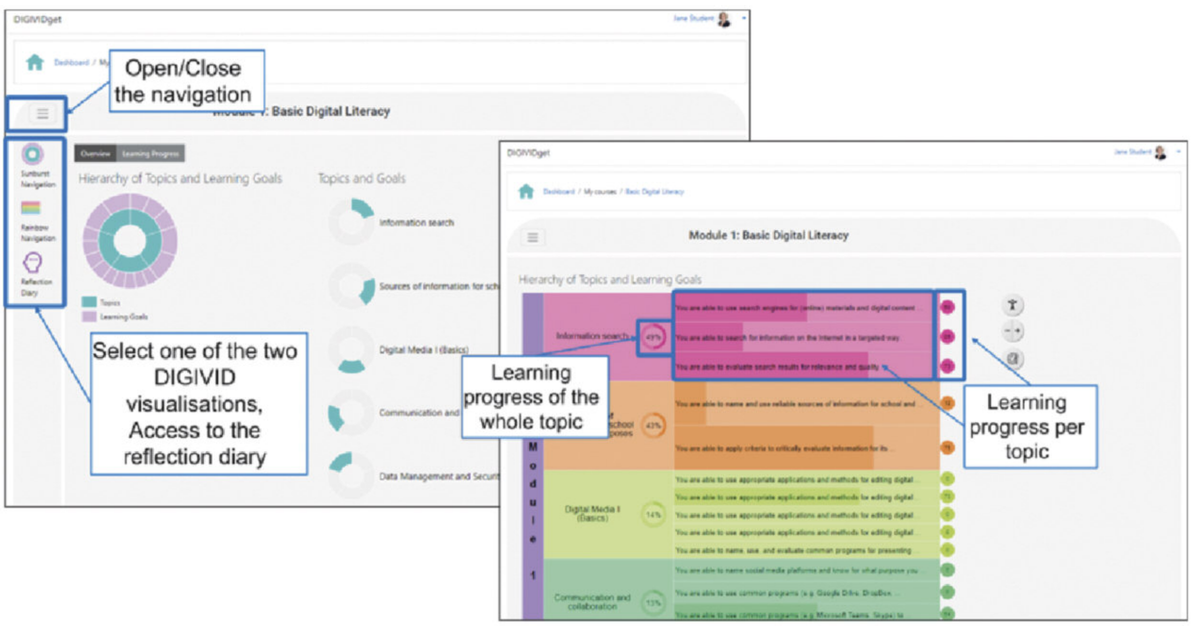

Figure 1: The DIGIVIDget visualisations, the Sunburst view (left) and the Rainbow view (right).

At the core of the DIGIVIDget are two different visualisations that allow DIGIVID users to easily navigate through the DIGIVID content. For each module, users can get an overview of the content, choosing between the Sunburst and the Rainbow views (Figure 1). The Sunburst view (Figure 1, left) visualises the main topics in the inner circle and, for each topic, learning goals in the outer circle. Clicking on a learning goal, the user can navigate directly to the corresponding learning content. The Rainbow view (Figure 1, right) structures the content in a table view. The first column indicates the module, the second the topic, and the third the learning goals. Each topic, and corresponding learning goals, are visualised in a different colour to make the grouping more visible. Both visualisations are designed to provide an overview of the structure of the content and easy access to it. Both visualisations offer the possibility to track one's own learning progress in an easy and intuitive way. To do so, users can select a percentage to estimate their learning progress. The visualisation of progress is intended to motivate the user to progress in their learning journey.

DIGIVIDget can be used by individual teachers or deployed in the context of more complex activities, like lectures, workshops, and courses. These extended scenarios of use allow grounding the learning in additional collaborative discussion and reflection. DIGIVIDget has been successfully evaluated in Austria, Germany, and Norway, proving its usefulness within different educational contexts.

You can try DIGIVIDget at [L1]. Also consider joining one of our upcoming training activities.

The DIGIVID Erasmus+ project (grant number: 2020-1-AT01-KA226-HE-092590) is co-funded by the Erasmus+ Programm of the European Union.

Links:
[L1] https://digivid.isds.tugraz.at/

References:

[1] C. Redecker et al., "European framework for the digital competence of educators: DigCompEdu," Joint Research Centre (Seville site), Tech. Rep., 2017.
[2] A. Fessl, et al., "The impact of explicating learning goals on teaching and learning in higher education: Evaluating a learning goal visualisation," in European Conference on Technology Enhanced Learning, Springer, 2021, pp. 1–15.

Please contact:

Angela Fessl (Project Coordinator), TU Graz, Austria
afessl@tugraz.at

Monica Divitini, NTNU, Norway
divitini@ntnu.no

Digital Battery Passports for a Circular Economy

by Veronika Siska (Austrian Institute of Technology) and Theresa Schredelseker (University of Duisburg-Essen)

The digital battery passport is an essential driver of sustainable production and circular economy by storing and tracking data for batteries throughout their life cycles. The BatWoMan project is paving the way towards carbon-neutral Li-ion battery cell production via new sustainable and cost-efficient methods, and by building a prototype for a digital battery passport.

Batteries are crucial for a sustainable energy supply. However, source materials often include hazardous or rare materials, whose mining and production processes cause a substantial environmental footprint. At their end-of-life, disposed batteries represent an additional burden, while reuse or recycling is often hampered due to a lack of information on the battery's materials and structure. European and global initiatives aim to mitigate this situation by promoting sustainable production processes and increased circularity along the battery life cycle. These include policy initiatives of the EU Green Deal (EU Battery Regulation [1], EU Ecodesign for Sustainable Products Regulation [2]) and the global public-private partnership Global Battery Alliance [L1].

An important part of these initiatives is a standardised digital battery passport [3]. This passport will include a detailed description of the properties and current state of the battery and will accompany each individual battery throughout and after its operation time. Data stored in the passport will provide means for regulatory compliance, support informed decision-making by users and companies and enable a circular economy by providing crucial information to collectors and recyclers.

The digital battery passport will record relevant data from the entire value chain, from raw materials and manufacturing to usage and end-of-life. Before bringing a ready-to-use battery to the market, the manufacturer will create a unique identifier, physically accessible on the product (e.g. as a QR code). During its operation, the battery passport can be constantly updated, i.e. with data on its current status or about repairs. When reaching its end-of-life, second-hand users or recyclers can retrieve these data to guide their decisions about recycling, repair or reuse. Access control is crucial at all stages: some data might be accessible to the public, or only to certain entities along the value chain, such as the authorities or auditors.

The technical architecture underlying battery passports would consist of a decentralised dataspace ("electronic exchange system" in the EU Battery Regulation), potentially combined with a centralised database. The centralised part may contain static information about battery models, or simply a searchable catalogue of available data in the decentralised dataspace. The decentralised dataspace would consist of data stored at their source (material supplier, manufacturer, etc.) and shared under pre-defined conditions. For example, suppliers could exchange data with manufacturers to facilitate optimisation of the production process, or authorities might have access to in-

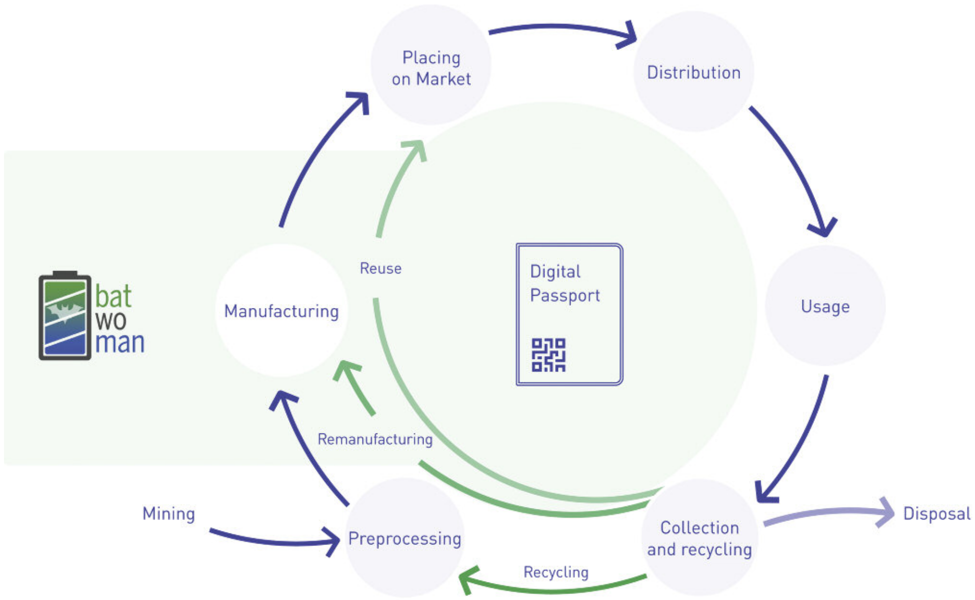

Figure 1: Battery life cycle, accompanied by a digital product passport. BatWoMan uses manufacturing data generated in the project and external/estimated data for other life cycle phases.

formation necessary to verify regulatory compliance. By making use of emerging standards for decentralised dataspaces (e.g. the European Gaia-X[L2] and IDSA[L3] initiatives), such data can also be integrated with other existing and upcoming data ecosystems.

The upcoming new Battery Regulation of the European Union [1] is expected to come into force in 2026 and will require that all electric vehicle and LMT (light means of transport, such as e-scooters or e-bikes) batteries, as well as industrial batteries with a capacity above 2 kW h, are equipped with a digital battery passport. The goal here should be to lay a good foundation with sensible standards, guidelines and architecture that focus on sustainability, but without a heavy bureaucratic burden on industry. We need to avoid excessively strict regulations that would impede European innovation or trigger companies transferring their activities outside of Europe, and support the European battery industry in shifting to more sustainable manufacturing and operation practices.

The project BatWoMan[L4], started in September 2022 and funded by the European Union's Horizon2020 research and innovation programme, aims to develop new sustainable and cost-efficient Li-ion battery cell production concepts, paving the way towards carbon-neutral cell production within the European Union. As part of the three-year-long project, data will be collected from materials sourcing, as well as from an optimised, data-driven manufacturing process. A full life cycle assessment, in which the battery's carbon footprint throughout its lifetime is estimated, is another essential goal of BatWoMan. These data serve as a testbed for a battery passport, considering all processes from material supply to end-of-life.

The battery passport demonstrator within BatWoMan[L4] will build heavily on cooperation with major European product passport initiatives. German-driven BatteryPass[L5] aims to develop cross-industry content and technical guidelines with participation from major industrial players and demonstrate them in a pilot project, while CIRPASS[L6] prepares the ground for digital product passports across industries, focussing on the battery, textile and electronics sectors. The BatWoMan passport builds on and validates the prepared guidelines and standards and showcases a battery passport based on real production data from pilot factories that are members of the BatWoMan consortium. Architecture design for the battery dataspace and passport is planned to be completed by the end of the first year, followed by an initial prototype in the second year and a functional battery passport demonstrator by the end of the project.

The BatWoMan battery passport and data-sharing ecosystem might serve as a validation testbed for battery research and production, to help prepare a suitable set of guidelines and promote an innovative, green battery life cycle.

Links:
[L1] https://www.globalbattery.org/
[L2] https://gaia-x.eu/
[L3] https://internationaldataspaces.org/
[L4] https://batwoman.eu/
[L5] https://thebatterypass.eu/
[L6] https://www.digitaleurope.org/digital-product-passport/

References:
[1] "Proposal for a regulation of the European Parliament and of the Council concerning batteries and waste batteries, repealing Directive 2006/66/EC and amending Regulation (EU) No 2019/1020." 2020. https://kwz.me/hwX (accessed Jan. 30, 2023).
[2] "Ecodesign for sustainable products." https://kwz.me/hwZ (accessed Feb. 09, 2023).
[3] Y. Bai, N. Muralidharan, Y.-K. Sun, S. Passerini, M. Stanley Whittingham and I. Belharouak, "Energy and environmental aspects in recycling lithium-ion batteries: Concept of Battery Identity Global Passport," Materials Today, vol. 41, pp. 304–315, Dec. 2020, doi: 10.1016/j.mattod.2020.09.001.

Please contact:
Veronika Siska
AIT Austrian Institute of Technology GmbH, Austria
veronika.siska@ait.ac.at
+43 664 825 1331

Remote Sensing: Detecting and Tracking Vehicles on Satellite Images and Videos

by Josiane Zerubia and and Louis Hauseux (Inria)

Remote sensing present various interests for civilian and military applications: automatic traffic monitoring, driving behavioral research, surveillance, etc. We work on images taken by satellite. This particular type of data has specific characteristics, quite different from those obtained by other aerial sensors (drones, planes, high-altitude balloons). The new Pléiades Neo satellites have high native resolution (i.e. without algorithmic processing) of 0.30 meter per pixel; and an impressive swath (photograph's width) of 14 kilometers. These technical characteristics make classical shape-based detection methods unsatisfactory.

This paper is composed of two parts: the first one focuses on vehicles' detection (see [1]), the second on the tracking of those vehicles (see [2]). The ideas mentioned in the first part for still images can be adapted to videos.

Geometric Priors for Vehicle's Detection

Satellite images or videos provided by constellations such as Pléiades (CNES – Airbus D&S – Thales Alenia Space) and Pléiades Neo (Airbus D&S) or, in the near future, CO3D (CNES – Airbus D&S) are taken several hundred kilometers above Earth level. Nevertheless, the high-resolution of these satellites enables to obtain detailed images: one pixel represents ~1 meter, up to 30 cm with Pléiades Neo satellites. However, precision remains much less than for usual images taken from airplanes, balloons or UAVs. Close objects may appear to merge into the image. Moreover, objects – let's say vehicles (that's our favorite objects of interest) – have a white-blob-looking appearance. The classical shape-based detection methods via Convolutional Neural Networks (CNN), which give excellent results for large objects, are therefore not reliable in this case.

Here is our first idea: Vehicles are not randomly arranged: the geometric configurations of some objects tell us something about the configuration of the other ones (in terms of shape or size, alignment, overlapping, etc.) Thus, we took back a previous CNN-method and associated it to another term considering geometry. These 'geometric priors' lend a hand to the program by encouraging it to promote certain configurations over others.

Let's take a closer look: a vehicle is represented by a rectangular box defined by its position (the x,y coordinates of its center), its length a, its width b and the angle α w.r.t. the horizontal. We then introduce priors on the possible configurations:

- *size prior:* a vehicle is a rectangle of particular shape. Thanks to the training-set, one can learn what type of rectangles a vehicle looks like. It prevents rectangles representing vehicles from being too small, too thin, too large, etc. Two parameters are jointly considered: the rectangle's area $a \times b$ and its ratio $r = b \div a$;
- *overlap prior:* to penalize two vehicles which would overlap;
- *alignment prior:* to reward a vehicle that would be aligned with its neighbors.

Mathematically, priors are some cost-functions seen as energy-functions. The total energy of a configuration is the weighted sum of these priors, plus the position-energy (obtained via CNNs), where the weights are learnt in a supervised way. The lower a configuration energy is, the more likely this configuration is.

So, seeking the best configuration (i.e. the most likely) leads to an energy-minimization problem. To do this, we use stochastic methods such as Monte-Carlo Markov Chains with reversible jumps.

Our algorithm has been trained and tested on the DOTA-image database. Figure 1 illustrates a difficult situation: a large building's shadow hides vehicles. (Night remote sensing does not show yet good results.) Alignment constraints make it possible to locate the long line of cars parked in the shadow.

Adaptive birth for Multiple Object Tracking

Some satellites (such as those of the future CO3D constellation) are equipped with matrix sensors that allow video capture (it is also possible to reconstruct short videos with linear sensors using image processing). These videos have a frame rate ~ 1 Herz (i.e. one image per second). As a point of comparison: the human eye can perceive up to 24 frames per second (FpS); that's why many cameras shoot at 24 FpS. Since time is discrete in videos (it is represented by the frame number), a

Figure 1: Cars parked in the shade are not well recognized by usual algorithms (left). Ours works fine (right).

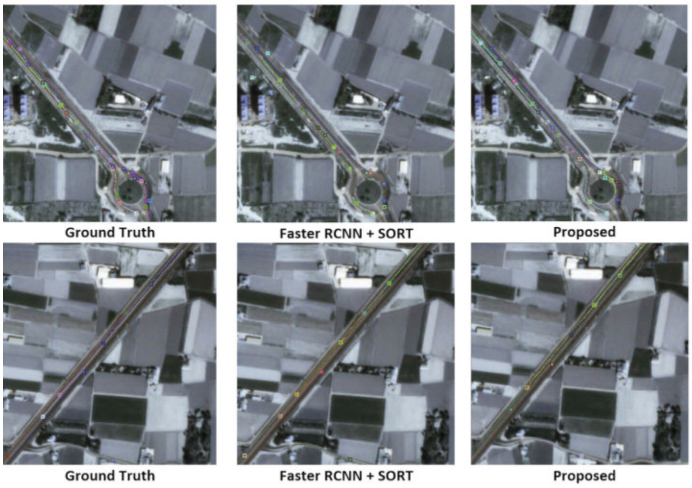

Figure 2: Difficult situations for tracking: vehicles are moving fast and there is a roundabout. Each colour represents a track.

low frame rate will produce the same effect as filming high-velocity objects: the same object can be distant between two frames.

Videos have an additional challenge to address compared to images: not only can we try to detect objects, but we can also try to track them until they disappear from the video.

Multiple Object Tracking (MOT) on videos is the process of jointly determining how many objects are present and what are their states (position, velocity, label) from noisy sets of measurements. In remote sensing, ground vehicles can appear from incoming roads or parking lots and disappear a few moment later. Additionally, noise sources such as illuminations changes, clouds, shadows, wind or rain generate numerous false positive detections. Fortunately, targets' dynamics in satellite images can be modeled well with linear motion. Therefore, deep-learning approaches used to track objects in highly cluttered and non-linear scenarios, with millions of parameters and astronomical calculations, seem unsuitable.

Conceptually, at each time step k (= at frame k), a set of measurements is received. Each measurement in this set is either a false alarm (= clutter) or is generated by an object. Furthermore, targets' measurements may be missed. The goal is to detect all the set of real objects. Notice that generally the number of measurements is different from the number of real objects, and both may vary with time. The sets of measurements and objects are thus seen as Random Finite Sets (of random length, varying at each iteration k).

Since a couple of years, frameworks working on these Random Finite Sets (RFS) has gained popularity due to their elegant Bayesian formulation. We have enhanced one RFS filter: The Generalized Labeled Multi-Bernoulli (GLMB) filter. The GLMB filter relies on strong assumptions such as prior knowledge of targets' initial state. We keep track of previous target states and use this information to sample the initial velocities of new-born targets. This addition significantly improves the performance of the GLMB in videos with low FpS (or, equivalently, with high velocity objects); this is particularly the case for satellite videos. (For another example of a RFS filter approximation by a GM-PHD filter, see [3].)

How to compare two algorithms? Several metrics exist in MOT. Diversity comes from the variety of clues that can be measured: false positives, false negatives, and, in the case of true trajectories detected, the partially tracked ones and the mostly tracked.

We tested our method with the WPAFB 2009 dataset. Adding our adaptive birth improves significantly the GLMB filter's results for usual metrics such as MOT-Precision, MOT-Accuracy and self-made metrics. Figure 2 shows very good results on a region which presents fundamental challenges for object trackers: objects move at high speeds, objects are often in crowded environments, and trajectories intersect regularly.

Conclusion

Remote sensing by satellite is both an important domain of computer vision and a very specific one. Algorithms must be designed for this precise use. This fact has been illustrated by two examples in vehicles' detection and tracking. The curious reader can find other examples on our web page.

This work has been conducted by Jules Mabon (PhD), Camilo Aguilar (Postdoc), Josiane Zerubia (Senior Researcher) and more recently Louis Hauseux (Research Engineer), in collaboration with Mathias Ortner (Senior Expert in data science and artificial intelligence at Airbus D&S).

Links:
[L1] https://team.inria.fr/ayana/
[L2] http://www-sop.inria.fr/members/Josiane.Zerubia/index-eng

References:
[1] J. Mabon, J. Zerubia and M. Ortner. Point Process and CNN for small object detection in satellite images. SPIE, Image and Signal Processing for Remote Sensing, XXVI-II (2022).
[2] C. Aguilar, M. Ortner, J. Zerubia. Adaptive birth for the GLMB filter for object tracking in satellite videos. MLSP – IEEE International workshop on Machine Learning for Signal Processing (MLSP) (2022).
[3] C. Aguilar, M. Ortner, J. Zerubia. Small moving target MOT tracking with GM-PHD filter and attention-based CNN. MLSP (2021).

Please contact:
Josiane Zerubia
Inria, Université Côte d'Azur, France
josiane.zerubia@inria.fr

Horizon Europe Project Management

A European project can be a richly rewarding tool for pushing your research or innovation activities to the state-of-the-art and beyond. Through ERCIM, our member institutes have participated in more than 100 projects funded by the European Commission in the ICT domain, by carrying out joint research activities while the ERCIM Office successfully manages the complexity of the project administration, finances and outreach.

Horizon Europe:How can you get involved?

The ERCIM Office has recognized expertise in a full range of services, including:
- Identification of funding opportunities
- Recruitment of project partners (within ERCIM and through ournetworks)
- Proposal writing and project negotiation
- Contractual and consortium management
- Communications and systems support
- Organization of attractive events, from team meetings to large-scale workshops and conferences
- Support for the dissemination of results.

Please contact:

Peter Kunz, ERCIM Office
peter.kunz@ercim.eu

The Institute for Software Engineering and Sortware Technology of the University of Malaga Joins ERCIM

The Institute of Software Engineering and Technology (ITIS, or ITIS Software) is a unit of the University of Malaga dedicated to scientific and technical research [L1]. The main objective of ITIS is to contribute to the development and application of software technologies through research, innovation, training, and technology transfer. Due to the transversal characteristics of software, the Institute is an interdisciplinary center in which specialists in different technological areas and application domains work together to advance these technologies. ITIS is oriented towards the foundations, tools, and applications of Software Technologies and Software Engineering for industry and society, with a special focus on five main areas: Automated Software Engineering, Data Science and Engineering and Artificial Intelligence, Cybersecurity, Smart Networks and Services, and Applications of Software Technologies and Software Engineering. The working language at the Institute is English.

ITIS research staff is composed by more than 50 doctors plus 50 PhD students and technicians. The unit has a large catalogue of open research infrastructures to support research on big data, security, fog computing, 5G & 6G and artificial intelligence.

ITIS is strongly committed to foster collaborations at the national and international levels both with companies and research institutions. The unit has R&D contracts with companies such as ABENGOA, ACERINOX, ADLINK, ATOS, DEKRA, EMERGIA, IBM, INDRA, KEYSIGT TECHNOLOGIES, ORANGE, TECNATOM, TELEFONICA, TELVENT, NEC LABS, NOKIA, and SIEMENS, among others. During the last five years the unit has been active in more than 20 European projects from H2020 and Horizon Europe, with special focus on 5G and 6G. ITIS is also active in international organizations like 5G IA &5G PPP, 6G IA & JU SNS, ETSI & 3GPP, Networld Europe, European Big Data Value Association (BDVA)/Robotics & Data, European Cybersecurity Organisation (ECSO), and now ERCIM. The average external funding of the Institute is 3,5M€ per year.

The Institute is located in Málaga. the capital of the Costa del Sol, in Andalusia, Spain [L2]. Málaga is a fantastic place to live and work, with an average temperature of 18ºC and a high quality of life. It is a cosmopolitan city, with an impressive history and an interesting present, both cultural and technological. Málaga is also home to the Malaga TechPark [L3], a technology park where more than 65 international companies in ICT have their headquarters and R&D centers and with which the Institute has a strong relationship and ongoing collaborations.

ITIS is always open to applications for different positions. The unit is currently offering six positions through the ERCIM Fellowship program.

Links:

[L1] https://itis.uma.es/en/home-2/
[L2] https://kwz.me/hxk
[L3] https://www.pta.es/en/malaga-techpark/

Please contact:

Javier Camara, ITIS/UMA, Spain
jcamara@uma.es

SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

Call for Proposals

Dagstuhl Seminars and Perspectives Workshops

Schloss Dagstuhl – Leibniz-Zentrum für Informatik is accepting proposals for scientific seminars/workshops in all areas of computer science, in particular also in connection with other fields.

If accepted, the event will be hosted in the seclusion of Dagstuhl's well known, own, dedicated facilities in Wadern on the western fringe of Germany. Moreover, the Dagstuhl office will assume most of the organisational/ administrative work, and the Dagstuhl scientific staff will support the organizers in preparing, running, and documenting the event. Thanks to subsidies the costs are very low for participants.

Dagstuhl events are typically proposed by a group of three to four outstanding researchers of different affiliations. This organizer team should represent a range of research communities and reflect Dagstuhl's international orientation. More information, in particular details about event form and setup, as well as the proposal form and the proposing process, can be found on

https://www.dagstuhl.de/dsproposal

Schloss Dagstuhl – Leibniz-Zentrum für Informatik is funded by the German federal and state government. It pursues a mission of furthering world class research in computer science by facilitating communication and interaction between researchers.

Important Dates
- *Next submission period:* October 15 to November 1, 2023
- *Seminar dates:* Between September 2024 and August 2025 (tentative).

ERCIM "Alain Bensoussan" Fellowship Programme

The ERCIM postdoctoral Fellowship Programme has been established as one of the premier activities of ERCIM. The programme is open to young researchers from all over the world. It focuses on a broad range of fields in Computer Science and Applied Mathematics.

The fellowship scheme also helps young scientists to improve their knowledge of European research structures and networks and to gain more insight into the working conditions of leading European research institutions. The fellowships are of 12 months duration (with a possible extension), spent in one of the ERCIM member institutes. Fellows can apply for second year in a different institute.

Where are the fellows hosted?
Only ERCIM members can host fellows. When an ERCIM member is a consortium the hosting institute might be any of the consortium's members. When an ERCIM Member is a funding organisation, the hosting institute might be any of their affiliates. Fellowships are proposed according to the needs of the member institutes and the available funding.

The fellows are appointed either by a stipend (an agreement for a research training programme) or a working contract. The type of contract and the monthly allowance/salary depends on the hosting institute.

ERCIM encourages both researchers from academic institutions and scientists working in industry to apply.

Equal Opportunities
ERCIM is committed to ensuring equal opportunities and promoting diversity. People seeking fellowship within the ERCIM consortium are not discriminated against because race, color, religion, gender, national origin, age, marital status or disability.

Conditions
Candidates must:
- have obtained a PhD degree during the last eight years (prior to the application year deadline) or be in the last year of the thesis work with an outstanding academic record. Before starting the grant, a proof of the PhD degree will be requested;
- be fluent in English.

Application deadlines
Deadlines for applications are currently 30 April and 30 September each year.

Since its inception in 1991, over 790 fellows have passed through the programme. In 2022, 40 young scientists commenced an ERCIM PhD fellowship and 69 fellows have been hosted during the year. The Fellowship Programme is named in honour of Alain Bensoussan, former president of Inria, one of the three ERCIM founding institutes.

http://fellowship.ercim.eu

"ERCIM has transformed my scientific career. Not only did I learn about a new field and expand my interdisciplinary skills, but I was also fortunate enough to receive a contract that will allow me to further develop my research. I can only recommend ERCIM to other researchers who are looking to broaden their horizons.

Simon DEL PIN
Former ERCIM Fellow

It is with great sadness that the Editorial Board of ERCIM News marks the passing of our esteemed colleague and friend Harry Rudin in October 2022.

Harry joined the Editorial Board in 2001 as representative of the Swiss ERCIM member SARIT, which later became SIRA, the Swiss Informatics Research Association. When SIRA withdrew from ERCIM in 2013, Harry continued to serve on the Editorial Board as our independent correspondent for Switzerland. Harry spent most of his professional career as a scientist at IBM Research Zurich. He was a member of IFIP (International Federation for Information Processing), where he focused on communication systems and related areas, and Editor-in-Chief for Computer Networks Special Issues at Elsevier. We will miss him!

CWI Spin-off DuckDB Labs Partners with MotherDuck

CWI spin-off company DuckDB Labs helped create startup MotherDuck, which aims to connect DuckDB to the cloud and which raised $47.5 million. The open-source DuckDB system was launched in 2019 by Hannes Mühleisen and Mark Raasveldt from the Database Architectures group of CWI. They designed and implemented an innovative lightweight database system that manages and analyzes large amounts of data in a highly efficient way, using state-of-the-art techniques, including CWI inventions such as vectorized query processing, and also introduced new techniques.

DuckDB is designed to work embedded within running processes – no separate server is needed. This new concept of 'embedded analytics' opens opportunities for adoption in many markets and use cases, for instance mobile apps or intelligent cars. Since its launch, DuckDB has gained widespread adoption from data scientists and tech companies alike for its ability to run efficiently anywhere, even inside a web browser. In 2021, the CWI spin-off company DuckDB Labs was created. DuckDB popularity keeps growing rapidly, with more than 800K downloads this month, a growth of more than 3x in the past year.

More information: https://www.cwi.nl and https://bit.ly/3mQfyuO

CWI and Inria Intensify their Collaboration

In the presence of Ministers Sylvie Retailleau and Robbert Dijkgraaf, Inria and CWI signed an agreement to extend their cooperation in the fields of Quantum Computing, Human interaction, Energy, Cryptography, Digital Health, Machine Learning and Software Engineering. New in this collaboration is the creation of European joint research teams. CWI and Inria, both founding members of ERCIM, have already collaborated successfully over the past decades and have come to know each other as reliable research partners. Both institutes have an excellent scientific reputation. By strengthening their cooperation through a partnership agreement, both parties join forces to support their research on a European scale and thus achieve important scientific results. This intensified collaboration will include a joint research and innovation agenda to strengthen networking, external partnership opportunities and funding. Scientific cooperation will be strengthened by creating joint projects as well as joint research teams.

Visit of the French Minister of Higher Education and Science Sylvie Retailleau to CWI. From left to right: Ton de Kok (general director CWI), Monique Laurent (CWI Fellow), Minister Sylvie Retailleau, Bruno Sportisse (chairman/CEO Inria) and Han la Poutré (Scientific Director cooperation Inria-CWI). Photo: CWI/Bas Kijzers.

The more intensive cooperation of both institutes to create a powerful alliance within Europe, as expressed by President Macron during the state visit, fits in with a joint ambition of both institutes to to join forces to cope with major scientific and societal challenges.

On 12 April 2023, French Minister Sylvie Retailleau and Inria Chairman/CEO (and President of ERCIM EEIG) Bruno Sportisse visited CWI in the context of the state visit of President Macron to the Netherlands.

ERCIM – the European Research Consortium for Informatics and Mathematics is an organisation dedicated to the advancement of European research and development in information technology and applied mathematics. Its member institutions aim to foster collaborative work within the European research community and to increase co-operation with European industry.

ERCIM is the European Host of the World Wide Web Consortium.

Consiglio Nazionale delle Ricerche
Area della Ricerca CNR di Pisa
Via G. Moruzzi 1, 56124 Pisa, Italy
www.iit.cnr.it

Norwegian University of Science and Technology
Faculty of Information Technology, Mathematics and Electrical Engineering, N 7491 Trondheim, Norway
http://www.ntnu.no/

Centrum Wiskunde & Informatica
Science Park 123,
NL-1098 XG Amsterdam, The Netherlands
www.cwi.nl

RISE SICS
Box 1263,
SE-164 29 Kista, Sweden
http://www.sics.se/

SBA Research gGmbH
Floragasse 7, 1040 Wien, Austria
www.sba-research.org/

Fonds National de la Recherche
6, rue Antoine de Saint-Exupéry, B.P. 1777
L-1017 Luxembourg-Kirchberg
www.fnr.lu

SIMULA
PO Box 134
1325 Lysaker, Norway
www.simula.no

Foundation for Research and Technology – Hellas
Institute of Computer Science
P.O. Box 1385, GR-71110 Heraklion, Crete, Greece
www.ics.forth.gr

Eötvös Loránd Research Network
Számítástechnikai és Automatizálási Kutató Intézet
P.O. Box 63, H-1518 Budapest, Hungary
www.sztaki.hu/

Fraunhofer ICT Group
Anna-Louisa-Karsch-Str. 2
10178 Berlin, Germany
www.iuk.fraunhofer.de

University of Cyprus
P.O. Box 20537
1678 Nicosia, Cyprus
www.cs.ucy.ac.cy/

INESC
c/o INESC Porto, Campus da FEUP,
Rua Dr. Roberto Frias, nº 378,
4200-465 Porto, Portugal
www.inesc.pt

Institute for Software Engineering and Software Technology
"Jose María Troya Linero", University of Malaga
Calle Arquitecto Francisco Peñalosa, 18, 29010 Málaga
https://gp.uma.es/itis

Institut National de Recherche en Informatique
et en Automatique
B.P. 105, F-78153 Le Chesnay, France
www.inria.fr

Universty of Warsaw
Faculty of Mathematics, Informatics and Mechanics
Banacha 2, 02-097 Warsaw, Poland
www.mimuw.edu.pl/

I.S.I. – Industrial Systems Institute
Patras Science Park building
Platani, Patras, Greece, GR-26504
www.isi.gr

VTT Technical Research Centre of Finland Ltd
PO Box 1000
FIN-02044 VTT, Finland
www.vttresearch.com