Deliverable from the COST Action CA19134 “Distributed Knowledge Graphs”

Use-cases and Requirements

Due date: 30 April 2022

Edited by: Tobias Käfer (DE), Michel Dumontier (NL), Andreas Harth (DE), Antoine Zimmermann (FR), Olaf Hartig (SE)

EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Authors:

(in alphabetical order)

Maribel Acosta, Rob Brennan, Kārlis Čerāns, Marcos Da Silveira, Mariana Damova, Anastasia Dimou, Michel Dumontier, Vincent Emonet, Özge Erten, Christian Esposito, Giorgos Flouris, Michael Freund, Kimberly Garcia, Steva Gostojić, Julian Grümmer, Andreas Harth, Katja Hose, Valentina Janev, Tobias Käfer, Sabrina Kirrane, Frédérique Laforest, Matteo Lissandrini, Simon Mayer, András Micsik, Cátia Pesquita, Axel Polleres, Ganesh Ramanathan, María Roldán, Tomer Sagi, Hannah Schuster, Danai Vachtsevanou, Jing Wu, Arif Yilmaz, Antoine Zimmermann

This deliverable is based upon work from COST Action "Distributed Knowledge Graphs", supported by COST (European Cooperation in Science and Technology).

COST (European Cooperation in Science and Technology) is a funding agency for research and innovation networks. Our Actions help connect research initiatives across Europe and enable scientists to grow their ideas by sharing them with their peers. This boosts their research, career and innovation.

www.cost.eu

COST EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the European Union

Preface

this document is the first in the series of deliverables of the COST Action CA19134 “Distributed Knowledge Graphs” (DKG). This COST Action is a research and innovation network to connect research initiatives around the topic of Distributed Knowledge Graphs. Using this deliverable, we want to showcase use-cases of Distributed Knowledge Graphs and corresponding requirements, which participants of the Actions deem relevant, are interested in, and working on. Moreover, we contribute to the capacity building aims of the Action by compiling this set of use-cases, as work of reference how participants of the Action improve on the state of the art, how participants of the Action reach out to other disciplines who apply Distributed Knowledge Graphs, and as basis for discussion.

The use-cases have been developed by Action participants, moderated by the Working Group leads in a virtual workshop of the Action in January 2022, and in subsequent refinement during additional virtual meetings. From those use-cases, potential requirements have been distilled by the authors and the Working Group leads.

The basic structure of the descriptions is author and use-case description, next to potential requirements that follow from the use-case, and target users. Some descriptions additionally come with references for further reading and sources. Depending on the Working Group’s focus on parts of a Distributed Knowledge Graph ecosystem, further information is added, e.g. on necessary DKG provisioning means for a DKG consuming case.

The application areas are from a diverse spectrum that contains bioinformatics, manufacturing, health, space, sustainability, open (government) data, law, financial services, supply chain, energy, smart cities, and water management.

In terms of the inclusiveness targets of COST, we note that the authors of the use-case descriptions come from 17 COST Member countries, of which 6 are inclusiveness target countries. The document has 34 authors, of which 13 are of the underrepresented gender.

With this deliverable, we want to show the breadth of use-cases for Distributed Knowledge Graphs and hope our collection of requirements is useful to shape future research directions in the field of Distributed Knowledge Graphs.

Karlsruhe (DE), Maastricht (NL), Nuremberg (DE), St Étienne (FR), and Linköping (SE)

Tobias Käfer
Michel Dumontier
Andreas Harth
Antoine Zimmermann
Olaf Hartig

WG1 – Producers

UC 1.1: Knowledge Graphs for Computational Drug Discovery

Use-case sponsors: Michel Dumontier (Maastricht University, Netherlands), Arif Yilmaz (Maastricht University, Netherlands)
Use-case description: Modern computational drug discovery, which uses machine and deep learning methods, depends on ready access to high quality public biomedical data. This includes access to information concerning drugs, diseases, drug targets, pathways, clinical trials, drug side effects, experimental and approved drug uses. Much of this data can be obtained from publicly available databases, but some valuable data are also contained in restricted data sources that either require monetary subscriptions or exclude commercial reuse. An overview of possible data sources is difficult to obtain, largely because data collections are not FAIR - they lack the necessary to facilitate their discovery. Another key challenge lies in integrating these data into a coherent knowledge graph, which can be used to create informative reports on entities of interest, to answer specific questions across multiple data collections, and build models to predict new functions (e.g. new drug uses or drug side effects). Each data source provides its data using different and largely incompatible formats, ontologies and terminologies, and there can be substantial divergence or disagreement in what is known and what the evidence is for this biological knowledge. Concerted efforts are needed to bring diverse stakeholders together to discuss and adopt common data standards, thereby facilitating the downstream use of multiple resources by scientists in both academic and industry environments.
Target users: Drug discovery researchers, research engineers, researchers.
Potential requirements:

- Standard documented and indexed Application Programming Interfaces (API)s and GUIs, with a particular attention to query APIs such as GraphQL or SPARQL, to retrieve particular subsets of data from an integrated knowledge graph
- Knowledge Graph metadata that indicates what kinds of entities and relationships are available from each distributed knowledge graph, which can be used in formulating distributed queries
- Ways to make the products of machine learning (e.g. ML models, embeddings) available to others in standard formats and with appropropriate metadata.

UC 1.2: Knowledge Graphs for Machine Learning in Manufacturing

Use-case sponsor: Anastasia Dimou (Belgium)
Use-case description: Building good Machine Learning (ML) pipelines is a long, complex and expensive task, performed in an iterative manner with trial and error. Besides building the ML pipelines, identifying the relevant data is a labour intensive task; not only it is difficult to find relevant data but it is even more rare to find data in the format and structure that the ML model expects. Knowledge graphs can describe the content of the data and allow the ML algorithms to find relevant data without relying on the structure of the data but on their

content. To achieve this, we need to have the knowledge graphs of potential data to be used by different algorithms and semantically enhanced descriptions of machine algorithms to match them with the knowledge graphs.
Target users: data providers who want to make their data available for ML algorithms and data consumers who want to discover relevant data for their (ML) algorithms.
Potential requirements:
- automated KG generation, provenance and versioning,
- alignment of different ontologies and languages
- matching of knowledge graphs and machine learning algorithms

UC 1.3: Knowledge Graphs for Remote Sensing Applications

Use-case sponsor: Maria Roldan (University of Malaga, Spain)
Use-case description: Earth Observation's satellite systems are continuously generating a great quantity of data, which are nowadays essential for applications in diverse areas, such as: climate change monitoring, precision agriculture, smart urban design, and many others. The development of general knowledge-driven approaches constitutes an open challenge in remote sensing, besides they provide human experts with domain knowledge representation,
support for data standardisation and semantic integration of multiple sources, such as multi-spectral (and hyper-spectral) data from various satellites and linked open data (meteorological, plant phenotype, etc.).
Possible applications are time series analysis, multiple satellite data product consolidation (Sentinel 2 and Landsat 8), data integration for analysis enrichment, and semantic reasoning for land-cover classification.
Target users: people interested in analysing vegetation state or monitoring important elements in remote sensing images, i.e. farmers.
Potential requirements:
- KG generation integrating data for different sources
- Ontologies development for formalising the domain, integrating data from different sources, providing semantics to the KG, allowing formal reasoning.
- Ontology alignment to build a common ontology.
- integration of KG with analysis algorithms.

UC 1.4: Biomedical Knowledge Graphs and Link Prediction for Evolving Knowledge Graphs

Use-case sponsor: Ozge Erten (Maastricht University, Netherlands), Vincent Emonet (Maastricht University, Netherlands)
Use-case description: A vast amount of scientific knowledge is available in public biomedical datasets, but significant challenges remain in finding and reusing them. These data are stored in different formats, using different data models, and are hardly interoperable without substantial work. Moreover, datasets change with time, but it is difficult to understand how they have changed and how this collective information can be analysed to reflect past and current understanding of important biomedical phenomena. Understanding the evolution

of biomedical knowledge graphs can prove useful in finding relevant connections (e.g. predicting new uses for existing drugs), and can also inform sociological and historical investigations in relation to how the scientific community targets future research. While there exist some proposals to catalogue changes in graphs, these are either trivial in their outcomes or too technically challenging to achieve in practice. New frameworks and implementations are needed to catalogue changes and provide insight into evolving knowledge graphs.

Target users: Biomedical researchers, Historians, AI researchers.

Potential requirements:

- Workflows and processes to integrate structured data to a distributed KG, and maintain this integration over time. We should reproduce a KG from prior data.
- Standard APIs such as OpenAPI, GraphQL or SPARQL to query the converted datasets, their previous versions, and information about their evolution (e.g. entities/edges added or removed)
- An up-to-date (dynamic) knowledge graph with information from popular public biomedical datasets (e.g. DrugBank)

UC 1.5: Distributed Knowledge Graphs for Privacy-aware Medical Imaging Tasks

Use-case sponsor: Nuno Garcia (University of Lisbon, Portugal)

Use-case description: Medical data is usually hard to obtain due to a number of issues such as quality of the data, legal issues, costly ground-truth annotations, or rarity of events of interest. Large medical imaging datasets could be more feasible to collect by merging small datasets spread around different institutions, but it raises privacy issues regarding ownership of data and others. Therefore, it becomes important to implement protocols and algorithms able to deal with the distributed nature of data while simultaneously protecting data privacy.

Recent machine learning methods learn to produce vector representations of knowledge graphs, also called graph embeddings. These vector representations are usually optimised representations with regard to a specific task. Once graph embeddings are obtained, learning algorithms can use them as inputs instead of raw data. One advantage is that the learning algorithms do not have to explicitly deal with the different types of data and other specific features from a particular source. One can leverage already existing algorithms to produce graph embedding vectors and go from there to solve a certain task.

Privacy-aware DKG algorithms may serve as an instrument to translate the graph embeddings from different sources to a common domain, and use this common embedding domain to learn the task. The common embedding domain may be achieved through a neural network that does the translation for each pair of sources, for example.

Target users: Health institutions / researchers

Potential requirements:

- Distributed datasets of medical data.
- Documentation regarding: data distribution statistics for each source; the preprocessing pipeline of data for each source; the graph embedding algorithm used

(typically a neural network of some sort) to produce the graph embedding; the learning algorithm used to work on top of the graph embeddings;

Privacy of data should not be compromised by these requirements.

UC 1.6: Distributed Knowledge Graphs Query for the Space Resources Utilisation

Use-case sponsor: Marcos Da Silveira (Luxembourg Institute of Science and Technology, Luxembourg)

Use-case description: Space resources have different facets and require the integration of information coming from many sources and from different formats. The use-case focuses on collecting the information from these sources and extracting some metadata that is able to describe the content and the contributors. The goal is to improve the interoperability between the sources and allow users to query them based on the available metadata.

The metadata includes the provenance, the licences, the access rights, the main domains, the format and language. The metadata will evolve according to user needs and this evolution will also be described in the metadata.

Target users: People interested in space resources utilisation.

Potential requirements:

- Text analysis and annotation
- Graph analysis and integration
- Automated metadata extraction.

UC 1.7: Knowledge Graphs for Product Life Cycle Assessment

Use-case sponsor: Katja Hose (Aalborg University, Denmark), Matteo Lissandrini (Aalborg University, Denmark)

Use-case description: Life Cycle Sustainability Analysis (LCSA) studies the complex processes describing product life cycles and their impact on the environment, economy, and society. Effective and transparent sustainability assessment requires access to data from a variety of heterogeneous sources across countries, scientific and economic sectors, and institutions. Given their important role for governments and policy makers, the results of many different steps of this analysis are made freely available, alongside the information about how they have been computed in order to ensure accountability. Semantic Web technologies in general and knowledge graphs in particular are key concepts to achieving this goal by enabling transparent sharing and integration of a diverse range of heterogeneous datasets. Sharing the result as interlinked knowledge graphs enables connecting this information to external sources and making it accessible to a broad range of users. Encoding provenance can then also enable domain experts to track the provenance of particular pieces of information that are crucial in higher-level analyses.

Target users: Experts and researchers in life cycle assessment

Potential requirements: Data extraction, data cleansing, data integration, data provenance, querying

References:

- Emil Riis Hansen, Matteo Lissandrini, Agneta Ghose, Søren Løkke, Christian Thomsen, Katja Hose. Transparent Integration and Sharing of Life Cycle Sustainability Data with Provenance. ISWC 2020, pp. 378-394
- Agneta Ghose, Katja Hose, Matteo Lissandrini, Bo Pedersen Weidema. An Open Source Dataset and Ontology for Product Footprinting. ESWC (Satellite Events) 2019, pp. 75-79

WG2 – Consumers

UC 2.1: Interlinking Open Government Data Catalogues

Use-case sponsor: Axel Polleres (WU Vienna, Austria)
Use-case description: Interlinking the information of OGD Portals by means of KGs has been partially investigated in several prior works, such as linking tabular data to Linked Data in several challenges, but the use case could be generalised to industrial settings where business partners are trying to align their enterprise data catalogues. Interlinkage could involve:

- Reference column detection
- Reference table detection
- Georeferencing datasets and records within datasets
- Interlinking and uniquely identifying publishers, organisations

Target users: OGD providers/publishers (mostly public institutions), but also citizens who could benefit from improved search interfaces.
Potential requirements: Entity linking and automatic annotation of non-semantic meta-data attributes, Knowledge Graph construction, resolving multi-linguality, etc.
References and sources:

- Sebastian Neumaier and Axel Polleres. Enabling spatio-temporal search in open data. *Journal of Web Semantics (JWS)*, 55:21--36, March 2019. [DOI | http]
- Jan Portisch, Omaima Fallatah, Sebastian Neumaier, and Axel Polleres. Challenges of linking organisational information in open government data to knowledge graphs. In *22nd International Conference on Knowledge Engineering and Knowledge Management (EKAW 2020)*, volume 12387 of *Lecture Notes in Computer Science (LNCS)*, pages 271--286, Bozen-Bolzano, Italy, September 2020. Springer. [DOI | http]
- SemTab Challenge: https://www.cs.ox.ac.uk/isg/challenges/sem-tab/

UC 2.2: Interlinking Semantic Digital Twins in Manufacturing

Use-case sponsor: Michael Freund (Fraunhofer Institute for Integrated Circuits IIS, Germany)
Use-case description: In Industry 4.0 environments, physical machines can have digital twins (DT). The DT consists of the mirrored sensor and actuator data of the cyber-physical system (CPS). If the data is represented in a graph structure modelled with RDF, it is also referred to as a semantic digital twin. These virtual replicas of CPSs are mainly used for monitoring, but can also be used to control machines via Read-Write Linked Data APIs.
By combining multiple machines and their graphs, it is possible to monitor and control not only a single process step, but an entire process involving multiple machines and factory floors. For this purpose, interoperability between the graph data of various CPSs has to be established, and the different knowledge graphs have to be made available.
The main challenge is the layout of the communication network between the various machines and CPSs.

The machine data and information can be exchanged in a decentralised way, where each CPS interacts with all of the other available systems or the system can be centralised where a central control and monitoring unit interacts with all the distributed knowledge graphs.
By accessing all sensor and actuator data of an entire process, it is possible to react dynamically to changes in the physical world and to adapt the further course of the process. The implementation of this use case will mainly benefit the fault resistance of production lines.
Target users: Manufacturing industry
Potential requirements: Using data from multiple CPS; adding external data sources; regular queries on distributed data sources; convenient way to exchange data between multiple factory floors

UC 2.3: Consuming Linked Legal Data

Use-case sponsor: Sabrina Kirrane (WU Vienna, Austria)
Use-case description: In the EU legal data (legislation, legal cases, etc) resides at both an EU level and also at a member state level (legal cases from small courts around the countries). Linking this data is beneficial for legal professionals that need to source legal data from different member states (multilingual).
Although data is primarily available in unstructured format (documents), the European Union is working towards making legal data more accessible across all member states. This goal is supported via standards such as the European Law Identifier (ELI) and the European Case Law Identifier (ECLI), which provide technical specifications for web identifiers and suggestions for vocabularies to be used to describe metadata pertaining to legal documents in a machine readable format. These ECLI and ELI metadata standards adhere to the RDF data format which forms the basis of Linked Data, and therefore have the potential to realise a pan-European legal Knowledge Graph.
To date much of the focus has been on extracting entities and temporal expressions from legal text and representing them in a knowledge graph, however there are still a number of open challenges concerning the consumption of linked legal data: (i) interfaces that facilitate search across distributed knowledge graphs; (ii) visualisations that are capable of representing facts relating to cases, chains of custody, legal inquiry processes, the links between legislations, etc. and (iii) access to legal information that is currently hindered by restrictive licensing and strict access policies, or the lack thereof.
Target data: Legislation metadata**;** court case metadata**;** facts relating to cases; chains of custody; legal inquiry processes; and links between legislations.
Target users: Legal professionals and legal scholars.
Potential requirements: Dealing with multilingual data; supporting search, queries and visualisation over distributed legal data sources; dealing with access and usage policies.
Relation to other WGs: WG1 (extraction)
External resources:
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52012XG1026(01)
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52011XG0429%2801
https://link.springer.com/article/10.1007/s10506-021-09282-8
https://aic.ai.wu.ac.at/~polleres/supervised_theses/Erwin_Filtz_PhD_2021.pdf
https://mnavasloro.github.io/PhDContributions/PhDThesis_mnavas.pdf

Recent events:
https://2021-eu.semantics.cc/sites/2021-eu.semantics.cc/files/Special%20Track%20Call_%20LegalTech%20-%20Semantics%202021_0.pdf
http://ai.di.uoa.gr/#iswc20-workshop

UC 2.4: Understanding Differences and Commonalities in Health (Covid) Knowledge Graphs

Use-case sponsor: András Micsik (SZTAKI, Hungary)
Use-case description: The COVID-19 pandemic gave momentum to the creation of several knowledge graphs collecting various information about the pandemic. However, these knowledge graphs were built from a limited set of raw data using different aggregation methods, thus the graphs have significant overlaps. The process to compare these knowledge graphs in terms of content is quite cumbersome and requires complex technical skills. There is a need to visualise the commonalities and differences for knowledge graphs in an easy to understand way. It is important to see the common classes and instance numbers between these knowledge graphs, as these show the potential for connecting graphs. This kind of investigation may also reveal that some graphs are very similar or subset of another with respect to the contained information. On the other hand classes that exist in a single graph only show the specificity of that graph and may be a sign of hidden 'treasures'. This kind of visualisation should work from a birds' eye viewpoint but could also allow diving into some details of the comparison. It can be a truly exploratory interface on the level of schema or ontology, but it is also important to detect different classes for the same concept, which is connected with ongoing research in ontology alignment.
Target users: Researchers in medicine, biology, social sciences, etc.
Potential requirements: Visualisation tools, automated ontology alignment
Relation to other WGs: WG4 (analytical tools)
Further reading:

- https://aic.ai.wu.ac.at/~polleres/supervised_theses/Felix_Helmreich_BSc2021.pdf

UC 2.5: Distributed Querying in Health (Covid) Knowledge Graphs

Use-case sponsor: András Micsik (SZTAKI, Hungary)
Use-case description: The COVID-19 pandemic gave momentum to the creation of several knowledge graphs collecting various information about the pandemic. A previous use case provides a method to identify classes common in selected knowledge graphs. Instances of these classes provide natural points for joining the knowledge represented in several graphs. The use case's need is to compose and execute queries on several knowledge graphs and to provide a unified result. Federated SPARQL querying has been available since long ago, but in practice we seldom see it working. Many huge KGs exist on different technical platforms, without having a SPARQL query endpoint. Labelled property graph databases support different query languages such as Cypher or Gremlin. A possible solution could be completely client-side, with interactive visual (drag-and-drop) query composition. Such

solutions would enable iterative and federated querying as a key function for consuming distributed knowledge..
Target data: Covid data, or research data, or research publications and citations
Target users: Researchers in medicine, biology, social sciences, journalists, policy makers, etc.
Potential requirements: Dynamic user interface tools, graph querying tools
Relation to other working groups: WG4 (query tools)
See also: SPARQL 1.1 Federated Query

UC 2.6: Transparency in Supply Chains

Use-case sponsor: Julian Gruemmer (Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany), Andreas Harth (Friedrich-Alexander-Universität Erlangen-Nürnberg and Fraunhofer Institute for Integrated Circuits IIS, Germany)
Use-case description: The new supply chain legislation, which will come into force in Germany in January 2023, is intended to bring more transparency to companies' supply chains. Some countries, such as the UK, Netherlands, Austria and Switzerland already have supply chain legislation. A mandatory European due diligence system is planned. The main goals are the protection and respecting of human rights and the access to remedy for victims of business-related abuses. In addition, companies might also analyse their supply chain in terms of other topics related to Environmental, Social, and (Corporate) Governance, e.g. sustainability. For this purpose, companies will be obligated to disclose their direct suppliers. By disclosing the upstream and downstream partners, data from the individual companies can be collected and their entire supply chain and suppliers' partners identified. The next step involves linking the individual datasets to a large knowledge graph, which will ultimately represent a network of the various companies in their roles as suppliers. Therefore, not only data is used that is disclosed by the companies within the scope of the supply chain legislation, but also data from publicly accessible databases, government publications and global supplier lists that are published due to the growing importance of the topic of sustainability. Via the collected data material and its processing in the Knowledge Graph, the supply chains of companies and also those of their suppliers can be traced more transparently. Data from different sources, which on their own are not very meaningful, can be connected, therefore enabling smarter processing and potentially generating new insights. The use of a knowledge graph that maps a network of real-world entities and their relationships. Such a perspective is especially valuable in the case of analysing supply chains with its many different components, which sometimes take on the role of the supplier, but also that of the delivering party.
The application of Knowledge Graphs provides several potentials. Among other things, companies can identify their weak members in the supply chain and thus avoid unexpected costs, delivery delays and sales losses and optimise their supply chain design. Furthermore, analysing using Knowledge Graphs can also enable disintermediation, resulting in more agile and flatter supply chains. The scarcity of semiconductors, which poses a major challenge for many companies, is also a topical issue.
While the visualisation and analysis of such Knowledge Graphs covering supply chains has some potential to offer, a critical view on certain application scenarios remains mandatory.

For example, large players can influence competition even more easily through their market power, as they can, for example, poach suppliers from other companies.
Ultimately, the rollout of a European-wide unified supply chain legislation leads to a number of challenges: What is the actual extent of the scope of consolidation? How far do indirect suppliers have to be declared? Should only the manufacturing industry be made responsible or do service providers also have to report? How can actors provide the required transparency while protecting their own interest?
Data:

- Databases:
  http://wwwen.ipe.org.cn/MapSCMBrand/BrandMap.aspx?q=6
  https://openapparel.org/facilities
- Global supplier list from companies:
  Adidas:
  https://www.adidas-group.com/en/sustainability/managing-sustainability/human-rights/supply-chain-structure/
  Apple:
  https://www.apple.com/supplier-responsibility/pdf/Apple-Supplier-List.pdf
  H&M:
  https://hmgroup.com/sustainability/leading-the-change/transparency/supply-chain/

Target users: Customers, members of the supply chain, legislation enforcers, non-profit organisations, (data) journalists
Potential requirements: Data conversion from tabular (CSV, TSV) data, data scraping from HTML pages and PDF documents, data integration including provenance, visualisation and data analysis for non-expert casual users

UC 2.7: Natural Disaster Management

Use-case sponsor: Axel Polleres (WU Vienna, Austria), Hannah Schuster (WU Vienna, Austria)
Use-case description: The increasing frequency of natural disasters, caused by severe weather phenomena and fueled by climate change, has become an undeniable reality. To make matters worse, the growing interconnectivity between countries as well as the transgressing nature of severe weather events further increase the challenges disaster and crisis management faces in managing such natural disasters. CRISP project (see https://www.crisp-project.org/) aims to address these challenges in a data-driven manner, enabling more effective crisis response and intervention, considering both the short-term management of disasters as well as long-term economic impact assessments, at fine-grained regional and temporal granularity.
To accomplish this, a unified Knowledge Graph shall be built, ingesting data from multiple heterogeneous sources. Data sources included in this pool are for instance disaster signals and perceptions from news and user-generated social media content as well as weather and climate observation data and data from municipalities and regional administrations, which should be structured into a distributed KG, since all of these informations come from different sources that update regularly, and partially can't share data publicly in an unlimited manner. Furthermore, structured and unstructured socio-economic data from Open Government Data

will be included. The result is a comprehensive and continuously updated knowledge graph, which represents a key asset for semantic modelling and impact forecasting.
Possible models include infrastructure models to simulate the impact of natural disasters on road networks and socio-economic models to simulate mid- and long-term effects of severe weather events on the economy of areas dependent on agriculture.
Target users: Emergency response organisations, economists, local communities, political decision makers, citizens
Potential requirements: Means to locate data sources, means to monetise data, means to combine open and closed data

UC 2.8: Data Quality / Public Open Data Quality

Use-case sponsor: Kārlis Čerāns (University of Latvia, Latvia)
Use-case description: Data is a key pillar for digital transformation because every interaction in the digital world generates data [1]. On the other hand, "[w]e need more authoritative and *filtered* data in our lives leading us to trustworthy sources so we can make fast, accurate decisions" [2].
European Open Data Directive [3] states that "Public sector information represents an extraordinary source of data that can contribute to improving the internal market and to the development of new applications for consumers and legal entities". It asks Member States to provide open access to high-value data falling into the following thematic categories: geospatial, earth observation and environment, meteorological, statistics, companies and company ownership, and mobility. The data to be opened "could, inter alia, cover postcodes, national and local maps (geospatial), energy consumption and satellite images (earth observation and environment), in situ data from instruments and weather forecasts (meteorological), demographic and economic indicators (statistics), business registers and registration identifiers (companies and company ownership), road signs and inland waterways (mobility)" [3].
The definition of the above categories is based on EU understanding of the data being "associated with important socioeconomic benefits having a particular high value for the economy and society" [3].
To enable the eventual users to rely on the data in taking the personal or business or public governance decisions, the data needs to be both accurate (the provided data need to match the reality) and complete (the data from the domain described by the data set need to be provided in full). Although estimating the full accuracy and completeness requirements may not be possible by looking at the data alone, some important aspects can be analysed on the basis of the structure of the data (e.g., missing attributes for existing data items and attribute values matching the respective data types). A deeper validation can be expected to be possible by exploring the internal semantic structure of the data, as well as its relations to other data sets.
Some existing use cases of open data are described in the Open data impact map [4] that lists cases from 90 countries and 1615 organisations (as of April 25, 2022).
Target users: Data set maintainers, interested parties in using the data (personal and business users, policy makers).

Potential requirements: Analyse the availability of the public open data in the designated categories. Provide means, metrics for accessing the data quality and means for quality improvement
Related WGs: WG1 Producers (Constructing Knowledge Graphs)
References:
[1] Data Is Essential To Digital Transformation
https://www.forbes.com/sites/forbestechcouncil/2020/12/03/data-is-essential-to-digital-transformation/
[2] Data Isn't The New Oil - Time Is
https://www.forbes.com/sites/theyec/2021/07/15/data-isnt-the-new-oil--time-is/
[3]
https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1561563110433&uri=CELEX:32019L1024
[4] https://opendataimpactmap.org/index

UC 2.9: Querying Heterogeneous Federations of Knowledge Graphs

Use-case sponsor: Maribel Acosta (Ruhr-Universität Bochum, Germany)
Use-case description: A federation of Knowledge Graphs (KGs) is composed of multiple, decentralised, and autonomous sources that provide access to individual KGs. In Federated Query Processing, the complete answer of a query typically requires contacting several sources to acquire partial results, which are then combined by a Federated Query Engine (FQE). Therefore, the role of FQEs is to provide users a unified view of the federation, i.e., users simply specify the query that should be answered by the system, and the engine is in charge of selecting the relevant sources from the federation and to retrieve the partial results from them.
Currently, most of the FQEs for querying KGs assume that the federations are homogeneous, e.g., all KGs are represented using the Resource Description Framework (RDF) and queryable through a SPARQL endpoint. Yet, with the emergence of new technologies for KGs, new forms of heterogeneity are starting to appear in federations of KGs. Heterogeneity of KGs can be present in different dimensions, which raises specific challenges. The focus of this use case is then to provide an overview of heterogeneous KG federations and the potential requirements for FQEs to efficiently query these federations.
Target users: Data Scientists who want to integrate knowledge available in decentralised, heterogeneous sources into a common representation.
Potential requirements: In the following, we describe some of the heterogeneity dimensions that can occur in decentralised KGs and the requirements for federated query engines to cope with the different types of heterogeneity:

- *Semantic Heterogeneity:* Due to the autonomous nature of federations, KGs are typically published using different ontologies/schemata or identifiers to represent knowledge. This can be observed even in KGs from the same knowledge domain. Therefore, semantic heterogeneity can be introduced at the ontological level or at the entity level. To cope with this type of heterogeneity, traditional integration systems provide the theoretical foundations to reconcile the usage of different labels to represent the same concepts or entities across different sources. The requirement for FQEs is to incorporate the capabilities of integration systems in order to provide a unified view of the federation to the user

- *Data Model Heterogeneity:* KGs can be represented using different data models, e.g., the Resource Description Framework (RDF), the Property Graph model, or the most recent proposal RDF*. The challenge for FQEs is to have the capacity of querying a federation of KGs independently from the underlying data model and to provide a unified view about the query results to the user.
- *Heterogeneous Performance:* Decentralised KGs can be served remotely by sources with different hardware specifications. The computational capacity of the server, the current workload of the server, and the network delays when contacting the KGs sources introduce unexpected performance behaviour during query processing. To mitigate these effects, FQEs can implement robust and adaptive techniques, which allow for adjusting the processing of the query according to the current conditions.
- *Heterogeneous Access Interfaces:* KGs available on the web can be accessed through different interfaces that use different protocols, query languages, and result serialisations. The Linked Data Fragments (LDF) framework provides a uniform way to describe the interfaces of KGs published with semantic web technologies. These interfaces range from low-expressive but highly available interfaces like Triple Pattern Fragments, to more expressive sources such as SPARQL endpoints. To query federations of KGs accessible through different interfaces, FQEs must submit requests to the sources that are understandable by the interface (to ensure correctness), while exploiting the capabilities of the sources (to improve the query performance). To this extent, the concepts of interface-compliant and interface-aware query processing techniques have been recently proposed in the context of FQEs. The requirement for FQEs is then to implement this type of techniques over a wide-range of LDF interfaces.

References:

- Acosta M., Hartig O., Sequeda J. (2018) Federated RDF Query Processing. In: Sakr S., Zomaya A. (eds) Encyclopedia of Big Data Technologies. Springer, Cham. https://doi.org/10.1007/978-3-319-63962-8_228-1
- Acosta M. (2022) Querying Heterogeneous Federations of Knowledge Graphs. Tutorial at the Marie Curie ITN-ETN KnowGraphs Winter School. https://doi.org/10.5281/zenodo.5972367
- Heling, L. and Acosta, M. Robust Query Processing for Linked Data Fragments. Semantic Web Journal. *To appear*. 2022. https://content.iospress.com/articles/semantic-web/sw212888
- Heling, L. and Acosta, M. (2022) Federated SPARQL Query Processing over Heterogeneous Linked Data Fragments. In Proceedings of the ACM Web Conference (WWW). *To appear.* https://doi.org/10.1145/3485447.3511947

UC 2.10: Knowledge Graphs in Medical Data Science

Use-case sponsor: Katja Hose (Aalborg University, Denmark), Tomer Sagi (Aalborg University, Denmark)

Use-case description: Medical applications are diverse in nature, there is an abundance of different types of data for example covering spreadsheets as well as textual annotations, time series (EKG, ECG, etc.), lab test results, and even genetic information. Hence, data science use cases apply machine learning to analyse these complex datasets and provide insights that can eventually help improve health care, e.g., detection of anomalies in medical images and other measurements, proposing drugs and treatments, etc. In this context, distributed knowledge graphs can not only help structure complex patient health care records in a powerful and robust way, they can also provide connections to a broad range of external information, for example, about diseases, drugs, demographics, or diagnosis code hierarchies, such as ICD-10, which can be used as additional information for machine learning tasks, such as patient risk or diagnosis prediction.
Target users: Data scientists working with medical data and use cases, experts looking for external information, data providers who want to share their data in relevant areas, data consumers who want to integrate the available knowledge in their data science tasks
Potential requirements: Automated KG alignment, mapping, and integration; provenance and versioning
Sources and references:

- Tomer Sagi, Emil Riis Hansen, Katja Hose, Gregory Y. H. Lip, Torben Bjerregaard Larsen, Flemming Skjøth: Towards Assigning Diagnosis Codes Using Medication History. AIME 2020, pp. 203-213
- Emil Riis Hansen, Tomer Sagi, Katja Hose, Gregory Y. H. Lip, Torben Bjerregaard Larsen, Flemming Skjøth: Assigning Diagnosis Codes using Medication History. Journal of Artificial Intelligence In Medicine, 2022, to appear

WG3 – Prosumers

UC 3.1: Knowledge Graph Sharing Platforms / Data Market

Use-case sponsor: Antoine Zimmermann (EMSE, France)
Use-case description: Generating data on specialised domains is often costly, so those who produce the data may want to get a return on investment by monetising datasets. For consumers, it can be more affordable to buy data from other parties rather than generate them on their own. A data market can then take place where data consumers and data producers negotiate to find mutual agreements in dealing data. In a large scale, open, and distributed setting such as the Web, connecting the right consumer to the right producer is difficult unless there exist data sharing platforms and brokers to enable simplified search and find functionalities, automatise financial agreements, host data and metadata, provide additional guarantees on the deals, etc.
A data sharing and data market platform must have a catalogue of datasets, with relevant metadata. The datasets may be hosted on the platform or just linked to third party websites. The metadata must provide information about the data themselves (size, frequency of update, format, etc.), provenance information (authors, location, original source, method, etc.), access control information (levels of privacy, access tokens, authentication protocols, etc.), as well as financial information (price, fees, warrantee, conditions, etc.). When the datasets are raw, unstructured or semi structured, the data platforms may provide transformed data that comply with Knowledge Graphs techniques and models, thereby enhancing the added value of the sharing platform. In addition, the data market can be connected to a market of services. Finally, multiple data platforms may transfer knowledge to one another. In an interoperable way. All together, the source datasets, the transformed datasets, the metadata sets, form a world-wide distributed knowledge graph with varying degrees of openness and targeted reusability.
The data market can also serve as a proxy for the exchange and sales of physical goods and services. Thanks to this, the match between consumer and producer (of services or goods) does not need to go to a real marketplace because the data marketplace makes the connection one-to-one.
Target users: organisations and companies in many sectors! (SMEs, any kind of partners in a B2B setting)
Potential requirements: multi-criteria search; access control; data cataloguing; fine-grained micropayment
Link with other WGs: this use case requires an ecosystem of datasets, platforms, utility tools, and services. It can be related to WG4.
Further reading: In the Horizon Europe, there is a call related to data sharing business and economy/data market: Technologies and solutions for data trading, monetizing, exchange and interoperability (AI, Data and Robotics Partnership) (IA) (HORIZON-CL4-2022-DATA-01-04).

UC 3.2: Distributed Knowledge Graphs for the Energy Sector

Use-case sponsor: Valentina Janev (Institute Mihailo Pupin, Serbia), Antoine Zimmermann (EMSE, France)

Use-case description: Energy management and the energy market is more and more distributed, and data needs to flow between energy actors and systems. In the EU, there will be different platforms for energy data, and different ways of generating, distributing, acquiring and selling energy. For the design of smart grid architecture, we need the sharing of metadata across platforms. Prosumers of energy (producing and consuming energy) are necessarily prosumers of metadata. This leads to new load balancing problems as well as new forms of economic exchanges regarding selling and buying energy. In order to make an informed decision on what amount of energy to buy from whom at what cost, or to sell, decision algorithms must use information that can be local (their own consumption and production), global (statistical data on seasonal household energy consumption), and possibly external to the grid (meteorological data). In such context, reliance on Web data from several sources adds real value to the decision process. The kind of data that has to be considered are, for the most part, highly fluctuating: weather for assessing heating needs, stock exchange for pricing appropriately, current and future supply and demand, etc. There is a need for a temporal model that covers historical data for statistical analysis, short term timestamped sensed data, and data about future predictions. The need for spatio-temporal information is even increased if the smart grid includes electric vehicles that can serve as energy producers when they are not consuming electricity for recharging.

Target users: mostly the energy producers, in the future, energy prosumers in general, actors from the energy value chain, SMEs that work with stakeholders from energy sector

Potential requirements:

- Manage real-time data and data streams;
- Publish and subscribe with high frequency updates;
- Ensure privacy, security and accuracy
- Portions of the distributed KGs must be legally binding

Sources and references:

- *Paper describing this use case:* Janev, V., Vidal, M. E., Endris, K., Pujić, D. Managing Knowledge in Energy Data Spaces. In *Companion of the Web Conference 2021*. https://project-lambda.org/sites/default/files/2021-08/PUPIN_TIB_BiDAW_2021.pdf
- See also Use Case 4.23 from the W3C Spatial Data on the Web Working Group: Optimizing energy consumption, production, sales and purchases in Smart Grids

UC 3.3: Knowledge Graphs in Law

Use-case sponsor: Stevan Gostojić (University of Novi Sad, Serbia)

Use-case description: Open access to legislation, regulation, and case law is a necessary condition to enable peace, justice, and strong institutions as one of United Nations' 17 Sustainable Development Goals (SDGs).

Public bodies (e.g. parliaments, courts of law, ministries and regulative agencies) and private bodies (e.g. private companies) could produce machine-readable legislation, regulation and case law as distributed knowledge graphs, reuse existing machine-readable legislation, regulation, and case law published by other public or private bodies both at the higher and at

the same hierarchical level as procumers, and private companies, NPOs/NGOs and citizens could consume those distributed knowledge graphs.
One legal entity can be both a producer and a consumer of legal knowledge graphs. It can produce legislation, regulation, and case law that is consumed by legal entities that are subordinate to it and it can consume legislation, regulation, and case law that is produced by legal entities that are superior to it. For example, second instance courts of law produce case law that is consumed by first instance courts of law and consume case law that is produced by supreme courts.
Having legislation, regulation, and case law in a machine-readable format has many advantages, including making it more accessible (e.g. by improving its retrieval and browsing) and promoting reuse.
There is a lot of work in the area both in academia (c.f. Artificial Intelligence and Law journal, ICAIL and JURIX conferences) and the private sector (a bunch of legal tech startups). Those knowledge graphs are either created by hand or are a result of (semi)automatic natural language processing techniques using textual corpora as inputs.
Target users: public bodies, public officials and civil servants, private companies as prosumers; and NPOs/NGOs and citizens as consumers
Potential requirements: privacy, provenance, temporal aspects, spatial aspects

UC 3.4: Open Knowledge Graph Catalogues and Meta-catalogues

Use-case sponsor: Antoine Zimmermann (EMSE, France)
Use-case description: Open data platforms can provide KGs of source datasets, but there are many of them focused on one region or country. Open data platforms can then be catalogued as well. Currently, there exist many existing data platforms that provide RDF data complying with the DCAT standard (e.g., all platforms deployed by OpenDataSoft). Cataloguing all these data platforms and catalogues requires consuming DCAT, and other types of metadata, and producing new metalevel KGs, possibly with additional statistics, quality measures, and discovery services on top. There can be domain-specific KG portals such as Covid-19 KG catalogues. One of the important benefits of a catalogue of KGs is in the interoperability and reusability of open data apps. If an app is designed to work with data conforming to a specific knowledge model, then the app should work with any other datasets conforming to the same model. Instead of building apps that work with hardcoded data sources, they could make use of KG catalogues that specify the knowledge model used (ontology) and work with any source conforming to the model. For instance, an app that offers statistics on Covid-19 infections in France should be able to provide the same indicators for Covid-19 in California.
Target users: the end users of open KG catalogues are mainly open data app developers
Potential requirements: provenance, quality metrics, temporal information, identification of ontologies/schema
Relation to other WGs: this use case is related to UC 2.1 and UC 2.9 from WG2.

UC 3.5: Distributed Knowledge Graph Reasoning

Use-case sponsor: Antoine Zimmermann (EMSE, France), Frédérique Laforest (Univ Lyon, France)
Use-case description: A reasoner may exploit multiple KGs and materialise the results of inferences from these graphs. More specifically, in the case of a distributed network of sensors and constrained devices, information from multiple sensors must be aggregated to draw conclusions. Assuming all sensors provide data in RDF, we can generate new aggregated data. As a first example, assuming we have the temperature and CO2 concentration at specific locations, we want an overall value for a large room or a building. We may also have to combine the information with knowledge about the sensors themselves, or the area around the sensors (e.g., room topology), or user comfort preference. Reasoning can provide decisions on window opening, or heater commands. As a second example, sensors in agricultural fields provide soil humidity at different depths, weather stations provide wind, temperature and air humidity, external KG provides crop state. Reasoning must provide decisions for watering, and for machinery authorization to enter the field (must not be too muddy, there may be a maximum speed, etc.). In this use case, sensors do not provide their data to a central server, but to nodes of a fog architecture so as to save energy. Nodes are spread on the field borders, so they must be energy-frugal, cheap (can be accidentally broken or even stolen), and weather-proof. They are typically small devices like ESP32. Such nodes should reason as much as possible on the data they gather around them, collaborate with nodes in their vicinity to reason about their outputs and provide results to some mobile device (on machinery) once a day. This mobile device transports results to the end user app.
*Input KGs***:** any KG + some ontologies or rules
*Output KG***:** a KG that only contains statements that are logically following from the input one
Target users: people in the room / in the building / in the farm
Potential domains: life sciences (combining different KGs with genes, drugs, proteins to reason about specific cures); cybersecurity (combining vulnerabilities, weaknesses, and own infrastructure to reason about possible vectors of attacks);
Potential requirements: should be robust to non-trustworthy / buggy / incomplete data
Sources: Examples are taken from the coswot project https://coswot.gitlab.io/

UC 3.6 Distributed Knowledge Graphs for Federated Data Science

Use-case sponsor: Katja Hose (Aalborg University, Denmark)
Use-case description: Open Data initiatives have led to the advent of Open Data portals providing machine-readable and structured datasets on topics, such as health, education, transportation, agriculture, and food. Following these principles, the data science community has launched initiatives for sharing not only data but entire pipelines along with all their artefacts (data, code, intermediate results, documentation, etc.). The first challenge is therefore how to technically facilitate sharing, conversion, exchange, etc. Federated data science then goes one step further and does not only focus on "how" to exchange and share data science artefacts and their (meta)data but also provides methods to find out "what"

artefacts of different pipelines to combine across diverse platforms in a federated manner. To achieve this goal, distributed knowledge graphs are needed to semantically describe artefacts at different sources so that this information can later be used to find artefacts that are semantically related and that can be combined to achieve a certain goal.
Target users: Data scientist using different platforms, code, building pipelines, researchers sharing experimental pipelines, reproducibility
Potential requirements: Semantic representation of data science pipelines and artefacts, conversion between platforms, formats, languages, etc.
References: Essam Mansour, Kavitha Srinivas, Katja Hose: Federated Data Science to Break Down Silos. SIGMOD Record 2021 (50) 4, pp. 16-22

UC 3.7 Knowledge Graphs for Monitoring Industrial Entities

Use-case sponsor: Maribel Acosta (Ruhr-Universität Bochum, Germany)
Use-case description: In Industry 4.0, industrial entities perceive their environment for autonomous decision making. To realise this vision, industrial systems should be able to monitor other industry entities automatically. At their core, these systems rely on entity representations that can be efficiently accessed, integrated with other data sources, enriched with multi-modal data (in particular, images taken from on-site cameras) and temporal information. To address these challenges, Knowledge Graphs (KGs) can be applied to provide a semantic model that captures the different aspects of entities’ descriptions. This use case, therefore, focuses on devising temporal and multi-modal KGs to represent industrial entities, which allow for modelling the current state and the ageing process of the entities. These KGs can be used to re-identify industrial entities by exploiting their natural physical characteristics, without the need of additional identification methods, e.g., barcodes. In addition, capturing the state of industrial entities over time also allows for predicting the deterioration of entities to further support predictive maintenance techniques.
Target users: logistics analysts.
Potential requirements:

- Integration of external data sources into the KGs (object manufacturing data, object shipping, etc.) to enrich the description of industrial entities and provide context.
- Multi-modal KGs to represent industrial entities based on (semi-)structured data and visual data taken from cameras.
- Temporal models for KGs to capture the ageing process or physical deterioration of industrial entities.
- Provenance models to annotate how the monitoring process was carried out.

UC 3.8 Data Record Archives for Real Time Open Data for Mobility and Transportation

Use-case sponsor: Antoine Zimmermann (EMSE, France)
Use-case description: Real time open data is often made available through a web service that only provides the current or latest state of affairs, such as the availability of bicycles at a bicycle sharing station, the location or delay of a bus/tram/train/airplane. For various purposes such as data analytics, logistics, public transportation plans, it is useful to also

access the history of data over a period of time. Consuming real time data and publishing the historical records in a queryable form is then desirable.
We assume that, in addition to services offering real time data, there will be third party services continuously monitoring real time data sources and archiving a historical knowledge graph of evolving data.
Target users: data analysts
Potential requirements:

- Continuous queries over distributed KG sources
- Subscribing to KG change notifications
- Temporal (and possibly spatial) metadata (especially time series information)

WG4 – Systems

UC 4.1: Multi-Agent Systems in Smart Farming

Use-case sponsors: Ganesh Ramanathan (University of St.Gallen, Switzerland), Kimberly Garcia (University of St.Gallen, Switzerland), Simon Mayer (University of St.Gallen, Switzerland)
Use-case description: Classical farming practices require farmers to collect information about the land and rely on their knowledge to make decisions on the activities to perform and the equipment to use (e.g., fertilise or irrigate). To further the automation and autonomization of farming practices, we propose to make use of Distributed Knowledge Graphs (DKGs) to describe a farm, the equipment, and actions to perform to achieve diverse goals. Moreover, other non-farming distributed information sources could be linked (through the DKGs) such as weather stations, and supply prices. Such DKGs would then serve as the knowledge source for automated agents to coordinate, organise, and find the means to perform the needed work in a farm. In a farming multi-agent system, a farm manager agent will be in charge of monitoring the DKGs and contacting other agents to perform a task such as irrigation. Upon task assignment, an agent will then use the DKGs to find a plan that describes the step-by-step actions to perform, including the equipment to operate, and any other needed supplies (e.g., water, pesticide, fertiliser).
DKG producers: humans through user interfaces that allow farmers to communicate with the system; IoT-enabled equipment that produces data on the current state of the farm; agent software
DKG consumers: software agents; user interface module to communicate with farmers
Potential requirements:

- Platform for agents to interact with one another via DKGs
- Components for an automated monitoring of DKGs

Further reading:

- Sensor networks in smart farms share tuple spaces (local view of DKG): see TeenyLIME https://es-static.fbk.eu/people/murphy/Papers/midsens06.pdf

UC 4.2: Health Intervention from Nutrition Information Ecosystems

Use-case sponsors: Simon Mayer (University of St.Gallen, Switzerland), Kimberly Garcia (University of St.Gallen, Switzerland), Jing Wu (University of St.Gallen, Switzerland)
Use-case description: With the proliferation of digital payments, digital food shopping data is becoming increasingly promising to help understand peoples' actual diet and beyond. An example is digital receipts from loyalty cards of food chain supermarkets. The digital food shopping data is automated, scalable and free from recall bias. Thanks to the introduction of General Data Protection Regulation (GDPR), historic and up-to-date data could be accessed with users' consent in the near future. Nevertheless, the way different data controllers and processors share data with individuals is not standard. This makes it difficult to build smooth

and fast data flow pipelines, when collecting food shopping data from different sources, e.g., retailers.

To understand the nutritional content of food shopping, enriching the food shopping data with food composition information and defining food categories is critical. However, foods from different food categories have different characteristics. There is neither universal food composition standard nor uniform food category standard, particularly for certain foods.

Distributed Knowledge Graphs can help standardise an efficient collection of food-relevant information, including food shopping data, e.g., digital receipts, and food composition data. It can accelerate the data sharing and risk identification of food shopping data. To be more specific, the integration of distributed knowledge sources such as digital receipts and food composition data and a clear definition of food categories and threshold of high/low food/nutrient intake can help identify risks in peoples' shopping baskets efficiently. When combined with a human or non-human user interface, the risks and possible preventative measures can be communicated with end-users easily. This might nudge peoples' behaviour towards a healthier direction. Distributed Knowledge Graphs will constitute the basis of such health interventions.

DKG producers: digital receipts available through APIs of grocery stores (i.e., indirectly, end users through their shopping habits, grocery stores)

DKG consumers: mobile app for end users who are participating in nutrition interventional studies; mobile apps and web applications for medical professionals that could use food shopping data to provide patients with guidance

Potential requirements:

- Integration of APIs as producers of KGs
- Means to access DKGs from Web of mobile apps

UC 4.3: Scalable, Flexible, Open Industrial Manufacturing

Use-case sponsors: Simon Mayer (University of St.Gallen, Switzerland), Danai Vachtsevanou (University of St.Gallen, Switzerland)

Use-case description:

The fourth industrial revolution (a.k.a. Industry 4.0) promotes the networking and interoperability of industrial devices and the decentralisation of decision processes in manufacturing environments (Kagermann et al., 2013). Recent research demonstrates that a feasible approach to engineer Industry 4.0 systems is to integrate Web of Things systems and multi-agent systems (Ciortea et al., 2018). In this context, we envision hybrid communities of factory workers and autonomous (artificial) agents that manage and use industrial equipment to manufacture products on the fly; factory workers would focus on tasks that require creativity and leave to autonomous agents the tasks that can be automated, such as inferring production plans to achieve well-defined manufacturing goals.

Concretely, in this scenario, we envision at least two workspaces that both have people present who may support and troubleshoot. Human workers as well as devices and computational services feature semantic descriptions of their abilities. Based on input by a customer (i.e., a product specification), the system is able to determine how the available components can, together, manufacture the specified product. They are then able to execute the created plan while retaining the ability to react to run-time changes (e.g., services becoming unavailable). A system is desired that achieves vertical and horizontal scalability, can evolve at run time, and supports forward- and backward-compatibility of components. We propose that this can be achieved through the publication of (agreed-upon) interaction and regulation specifications in a hypermedia environment to enable the discovery and usage of these specifications by services at run time.

This use case is relevant to DKG since it sees a distribution of knowledge across customers, integration engineers, component providers, and service providers - specifically regarding the semantic description of capabilities of individually provided components (e.g., machines) that are engineered to running systems by integration engineers, and the specification of products.

Relevant stakeholders: customers, engineers, component providers
DKG producers: cyber-physical systems and computational services that provide metadata about their functionality; customer interfaces that provide product specifications
DKG consumers: natural and artificial autonomous agents that receive functionality metadata and integrate it to manufacture specified products
Potential requirements:

- Evolution of system components at run time (e.g., changing APIs)
- Support of forward- and backward-compatibility of components (e.g., versioning)
- Vertical and horizontal scalability
- Possibly, monitoring and journaling of interactions between agents and artefacts
- Domain knowledge in the discrete manufacturing domain
- Access to relevant devices and services

References:

- Kagermann, H., Helbig, J., Hellinger, A., Wahlster, W. (2013). Recommendations for implementing the strategic initiative INDUSTRIE 4.0: Securing the future of German manufacturing industry; final report of the Industrie 4.0 Working Group. Forschungsunion, acatech.
- Ciortea, A., Mayer, S., & Michahelles, F. (2018). Repurposing manufacturing lines on the fly with multi-agent systems for the web of things. In Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems.

Source: ANR/SNF Project “HyperAgents” and EU H2020 NGIoT Project “IntellIoT”

UC 4.4: Knowledge Graphs for Decarbonisation

Use-case sponsor: Kimberly Garcia (University of St.Gallen, Switzerland), Simon Mayer (University of St.Gallen, Switzerland)
Use-case description: Life Cycle Assessments on GreenHouse Gas (GHG) emissions are commissioned by stakeholders, such as cities, production sites, and data centres, who have established decarbonization goals but need sustainability experts to help them achieve them. On their side, sustainability experts are not only experts on performing assessment methods but currently are also experts on the different information sources (databases) needed to create such assessments. Hence, sustainability experts spend large amounts of time understanding data sources, their schemas, and gathering data that is scattered and

accessible through diverse websites and services. Such a process is not only cumbersome, it is also lengthy and expensive. Hence, in a dynamic field such as sustainability, even when stakeholders have good intentions, their assessments might mismatch reality when they finally become available to them. In such a scenario, DKGs may be used to integrate information sources in order to compute GHG assessments for applications such as i) procurement for ICT equipment of cities, ii) GreenHouse Gas (GHG) emission assessment for production sites, and iii) decarbonization of data centres.
Relevant stakeholders: sustainability experts, companies and organisations that want to reduce their carbon footprint such as cities, production sites (factories), and data centres
DKG producers: tools to collect, analyse and monitor sustainability data of products and services (i.e., LCA software); sustainability data bases (e.g., ecoinvent) and other sustainability data sources produced by universities, non-profits, and standardisation bodies
DKG consumers: reporting tools, life cycle assessment tools for diverse sectors, such as cities, production sites, and data centres
Potential requirements:

- Domain knowledge on sustainability
- Access to relevant sustainability information sources and their schemas
- Access to stakeholders requirements

Source: 4 year swiss project WISER (funded by Innosuisse)

UC 4.5: Healthcare Risk/Safety Governance with Federated Knowledge Graphs

Use-case sponsor: Rob Brennan (Dublin City University, Ireland)
Use-case description: Governing safety and process improvement in health systems requires access to a wide range of evidence (e.g. data, analytics, reports). This is gathered for socio-technical analysis, resulting in organisational change management projects which must be monitored and verified via more evidence. The evidence is potentially sensitive data. Often it must be shared and analysed across multi-organisation chains of healthcare provision e.g. between health standards authorities, hospitals, ambulance services and community care services. Within each organisation different stakeholders require different perspectives and reporting in order to support risk governance. Traditional risk management and quality improvement depends on human-centric processes that focus on incident reporting and analysis, maintaining risk registers, compliance reporting and safety or cleanliness audits that are typically embedded in natural language reports to be read by people that change on daily, monthly or even an annual basis. These processes can be automated if more of this data is made machine-readable, for example in a knowledge graph. In parallel, advances in healthcare information systems and e-health mean a wealth of operational quantitative data and analytics are available to supplement the more qualitative, human timescale safety data. Also, open data publication on the web provides an ever more valuable source of additional evidence for healthcare risk governance.
Knowledge graphs are a natural way to bring together such diverse data sources due to their flexible schemas and through use of uplift to common ontologies, ontology alignment techniques, NLP-based knowledge extraction and metadata-based integration e.g. data catalogues. These integrative Knowledge Graphs can enable risk analysis across multiple

organisations, especially if expressed in or linked to standard ontologies for risk, organisations, culture and governance. In addition to risk analysis to determine corrective actions (risk mitigation projects) there is a need for synthesis of relevant evidence for safety best practice in the form of policies, training material or standards. Access control and data protection compliance are essential in this environment as some of the data collected in the graphs is sensitive personal data and all risk and safety data is potentially confidential due to the reputational or legal risks of publication. A key challenge is the uplift of safety information such as incident reports or safety reports that may be stored in traditional databases but in practice are semi-structured at best since most qualitative analysis lies in text fields designed to be read by human experts. Some relevant reports, policies and standards are purely in the form of natural language and must be subjected to knowledge extraction techniques to enable knowledge graph-based processing.

Target users: frontline clinical and operations staff; clinical specialist, unit manager, safety manager, healthcare operations manager, healthcare board member, healthcare organisation executive, health standards authority, patients and families

DKG producers: risk management platform, operational systems (via uplift) eg patient records, staff training, drug management, incident reporting tools, logistics, staff rosters, cleaning records, open data sources on the web, NLP-based knowledge extraction tools, data catalogues

DKG consumers: reporting tools, risk analysis tools, training tools, risk information circulation systems, project management tools

How do they interact: via SPARQL/the local KG, ARK CUBE Risk Ontology, ARK Risk Ontology, federated dataspace security controls, web of data lookup + interlinking, via standard vocabularies like DCAT, via data catalogues,

Potential requirements:

- Security (advanced access control policies to enable federated sharing)
- GDPR compliance (Privacy by Design, compliance reporting, …)
- KG distillation to extract sharable evidence that is not confidential/personal
- Mechanisms for Identification of evidence-based best practice
- Training material generation from KGs of evidence-based best practice
- Analytics governance mechanisms
- Privacy-aware data interlinking mechanisms
- Entity similarity measures
- Entity recommendation systems
- Standard ontologies for risk/safety, socio-technical system analysis, project management, organisational structures based on working relationships, …
- Automation of making decisions based on knowledge base of risks. E.g., answer set programming to select risks when there are multiple solutions or inconsistencies, hence compromises are required?

Source: Ongoing Irish project (ARK Platform, https://openark.adaptcentre.ie/)

References:

- McDonald, N.; McKenna, L.; Vining, R.; Doyle, B.; Liang, J.; Ward, M.E.; Ulfvengren, P.; Geary, U.; Guilfoyle, J.; Shuhaiber, A.; Hernandez, J.; Fogarty, M.; Healy, U.; Tallon, C.; Brennan, R. Evaluation of an Access-Risk-Knowledge (ARK) Platform for Governance of Risk and Change in Complex Socio-Technical Systems. Int. J.

Environ. Res. Public Health 2021, 18, 12572. https://doi.org/10.3390/ijerph182312572

- Julio Hernandez, Lucy McKenna, Rob Brennan, TIKD: A Trusted Integrated Knowledge Dataspace For Sensitive Healthcare Data Sharing, https://ieeexplore.ieee.org/document/9529379

UC 4.6: Distributed Knowledge Graphs in the Context of Smart Cities

Use-case sponsor: Christian Esposito (University of Salerno, Italy)
Use-case description: The Smart City paradigm mainly consists in the application of the ICT technologies to improve the main processes related to the optimal management of crucial aspects of the city governance. It also aims at creating novel and innovative services by connecting the physical infrastructures of the city with the citizens to offer them engagement within the community and empowerment . A smart city solution has a layered architecture where at the bottom we have different types of electronic sensors and/or digital services. The goal of those sensors and services is to collect several kinds of data, mainly related to pollution, traffic, citizen satisfaction, etc. Such a vast amount of data is further distributed towards the edge, fog and/or cloud level for various stages of processing, according to the paradigms of the Internet of Things (IoT). The edge level (see https://doi.org/10.1016/j.future.2019.02.050) consists of nodes closer to sensors and services where a first stage of processing may happen and can consist of IoT gateways but also nodes in the Cloud-RAN solutions. Additional nodes can be located in the core of the infrastructure, making the fog level (see https://doi.org/10.1016/j.jnca.2017.09.002), or outsourced to the cloud. Such processing is needed to infer new knowledge to be used to manage assets and resources efficiently, and to provide feedback on the optimality of the policies put in place within the city. The exploited sensors can be deployed at citizens' devices, public transportation vehicles, and assets with the intent of monitoring principal municipal business processes, such as traffic and transportation systems, power plants, water supply networks, waste management, crime detection, information systems, schools, libraries, hospitals, and other community services. This allows city officials and policy makers to interact directly with both community and city infrastructures, and monitor what is happening in the city and how certain phenomena are evolving over time. Such a novel process must allow real-time responses to detect possible issues and put countermeasures against them promptly.
A smart city solution is not realised from scratch by a single municipality that owns and manages the overall platform. Instead, it is the product of the integration of some existing infrastructures owned by multiple different organisations that collaborate to make the city smarter. For example, in a city, the public transportation company may monitor its buses and trains, while the various private companies related to bike, car, and/or scooter sharing have their ICT infrastructure to control their vehicles and customers. A municipality's central decision-making service can be fed with the monitoring data coming from these infrastructures, jointly with some traffic monitoring sensors deployed along the road by the municipality itself, to have a glimpse in real-time of the traffic state within the city and plan ahead for public transportation policies and transportation improvements. These monitoring

data may also be used by some private companies so as to improve their services, such as the online food ordering and delivery platforms as Uber Eats or Deliveroo can exploit them to improve the routes taken by their couriers and to reduce the waiting time of their customers.

Smart cities are mainly an example of an ecosystem of multiple tenants producing, analysing and storing data, with the goal of making decisions by adjusting the city policies or sending feedback/advice to the citizens to optimise their behaviours. The services and infrastructures interact by means of a federated architecture that characterizes the current platforms for a smart city.

Various solutions provide the services required to implement such a vision, such as the well known FIWARE platform. The primary goal of these platforms is to provide interoperability at the syntactical and semantic level of the various data models specified and managed by the various organisations composing the smart city. Ontologies have been considered as the winning technology to model the data instances in the various parts of a smart city and to simplify their interconnections to support interoperability (see http://dx.doi.org/10.3390/su13105578). For this scope, the FIWARE consortium has standardized de facto the data models for the various applications that can be built on top of their platform by using an ontological approach: https://www.fiware.org/about-us/smart-cities/. Also, other organisations have provided various ontologies to model the data managed by smart cities, such as in https://www.preprints.org/manuscript/202108.0101/v1 or https://www.etsi.org/deliver/etsi_gs/CIM/001_099/006/01.01.01_60/gs_CIM006v010101p.pdf . Assuming these experiences as starting points and applying some of these ontologies to a set of individual data points available in the smart city data sets, we can create a knowledge graph that can be used to infer new connections between data points and generate novel knowledge and ease the issue of data portability and interoperability. Recently, some experiences of a knowledge graph for smart cities have been proposed such as https://n-bridges.com/, https://www.cares.cam.ac.uk/research/cities/ or https://doi.org/10.1109/IC3INA48034.2019.8949613. These solutions are still at their beginning and should be evolved so as to encompass other possible aspects to be modelled, such as the security and privacy-related policies, selective disclosure of personal data related to the users, representation of user consent or any other legal bases for processing personal data, and so on. Smart cities share a lot of similarities with other intelligent ambients such as smart factories, smart hospitals, smart schools, and so on. The major difference is related to the system scale, which is small or medium in these examples, but is extremely large within the context of smart cities.

Target users: citizens, municipalities, public transportation companies, municipal utilities, private companies

DKG producers: IoT nodes deployed on cars, buses, smart city elements (such as lighting elements, smart trash bins), monitoring services of city processes

DKG consumers: citizen personal devices, decision services at the stakeholders

How do they interact: smart city platforms such as FIWARE offers Web-based interaction patterns and protocols

Sources and references:

- A. Fensel, "Keynote: Building Smart Cities with Knowledge Graphs," 2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA), 2019, pp. 1-1, doi: 10.1109/IC3INA48034.2019.8949613.
- M. Chen, X. Wei, J. Chen, L. Wang and L. Zhou, "Integration and Provision for City Public Service in Smart City Cloud Union: Architecture and Analysis," in IEEE Wireless Communications, vol. 27, no. 2, pp. 148-154, April 2020, doi: 10.1109/MWC.001.1900264.
- De Nicola, A.; Villani, M. L. Smart City Ontologies and Their Applications: A Systematic Literature Review. Sustainability 2021, 13, 5578. https://doi.org/10.3390/su13105578

Potential requirements:
- Security - authentication and authorization for data publication and consumption, data confidentiality and integrity
- Privacy and GDPR compliance
- Knowledge Graph Matching and Mapping
- Modelling of legal bases for processing personal data
- Syntactical and Semantic Interoperability
- Multi-modal and multi-protocol data retrieval

UC 4.7: Intelligent Information Infrastructure for Monitoring Water Resources

Use-case sponsor: Mariana Damova (Mozaika, Ltd., Bulgaria)
Use-case description: The exploitation of hydropower reservoirs and rivers requires monitoring based on heterogeneous datasets, such as spatial information, digital measurements, meteorological forecasts. The hydropower reservoirs exist in an integrated natural environment and their functioning influences the water sources, e.g. watersheds and the like, that are contingent to them. That is why it is important to monitor hydropower reservoirs within their integrated natural environment by taking into account the impact of their existence and functioning for their contingent environment. To do this, it is necessary to use periodically collected information from different resources, including: detailed data about the condition of the distinct hydropower reservoirs, water economic data, meteorological data and forecasts, geographical information, information about the integrated nature environment that has to be provided in a convenient for analysis and evaluation visual form. Predicting sediment deposition potential in the rivers or the water balance of dams depends on a variety of meteorological and environmental factors like precipitation, snow cover, soil moisture, vegetation index, solar irradiance, turbidity, surface reflectance, velocity. Remote sensing and earth observation provide a wealth of meteorological data and imagery that can be instrumental for these analyses and monitoring tasks.

The creation of an information system that will successfully address the need of an effective daily monitoring of the hydropower reservoirs and rivers and informed decision making for maintenance, routine exploitation and emergency situations requires capabilities of federated and integrated representation of spatial and digital, symbolic data, images and metadata and their easy linking in a common open, easy to maintain, update and reliable information infrastructure. The data that are to be used are heterogeneous on the one hand, dispersed in different locations on the other hand, and with different access rights on the third. The challenges in this context are mainly related to the integration of spatial information of GIS systems, and the remote sensing information along with symbolic and numerical data coming from different sources with different levels of clearance to the benefit of real life exploitation and maintenance of the water reservoirs and rivers, and the management of their environment.

Relevant *stakeholders* in this context are, on the one hand, different groups of persons who contribute to producing the data and related knowledge graph artefacts. Such groups include knowledge engineers, modelling experts, software engineers, software architects, logicians, hydrologists, earth observation experts, and earth observation data providers. On the other hand, there are different types of end users of the envisioned information system, including water resources managers, operations officers, government officials, policy makers, planning bodies, river maintenance bodies. Further, from a software development perspective, consumers of the knowledge graphs are visualisation and analytics components, recommendation and alert generation components, data access layers and intelligent AI-based components.

Potential requirements:

- Semantic integration and interlinking of heterogeneous data
- Ontological models
- Data access protocols
- Proprietary and sensitive data protection
- Integration of logical, numeric and geospatial reasoning

References:

- Mariana Damova, Emil Stoyanov, Mihail Kopchev, Hermand Pessek, Martin Petrov, Stefano Natali. Linked Data Meet Deep Learning to Empower Water Resources Monitoring of Dams. Big Data From Space 2021 (BiDS'2021), Virtual Event, Originally planned to be hosted by University POLITEHNICA of Bucharest (UPB), May 2021.
- Mariana Damova, Emil Stoyanov, Mihail Kopchev, Hermand Pessek. Ontology-Based Semantic Search Over Linked Satellite, Geospatial, Numeric and Symbolic Data in the Water Management Domain. In: Proceedings of The ESA Earth Observation Φ-week EO Open Science and FutureEO, Frascati, Italy, 2019.
- Mariana Damova, Emil Stoyanov, Martin Petrov. Linked Data Infrastructure for Monitoring Large Hydropower Reservoirs. In: Proceedings of The ESA Earth Observation Φ-week EO Open Science and FutureEO, Frascati, Italy, 2018.

Sources:

- http://isme-hydro.com/ - ESA Contract No 4000122783/17/NL/SC
- http://ismosede.bg/ - ESA Contract No 4000133836/21/NL/SC (ISMoSeDe)

UC 4.8: Fine-grained Data Sharing Between Enterprises and their Financial Advisors

Use-case sponsors: Tobias Käfer (Karlsruhe Institute of Technology, Germany), Andreas Harth (Friedrich-Alexander University Erlangen-Nuremberg, Germany)
Use-case description: Data-sovereign small or medium-sized enterprises are cautious (e.g. due to perceived or real legal requirements) when it comes to maintaining sensitive data (e.g. raw data from which publicly disclosed financial statements have to be calculated) on cloud infrastructures, typically provided by (US) companies with comparatively low standards when it comes to data protection. To nevertheless engage in digital data exchange with other parties (e.g. providers of financial advice and service), such enterprises require decentralised solutions, where they can host their sensitive data on-premise or in data centres of their choice under a data-protecting jurisdiction. Hereby, we assume the raw data or at least the master data about the involved parties to be in RDF. The aim of the involved parties is to digitise the way of getting and providing financial advice. In this decentralised setting, the usual authentication and authorisation infrastructure cannot be used, which typically relies on centralised identity providers. Thus, businesses require new techniques in order to facilitate fine-grained standards-based data sharing with selected partners. Idea: build this based on Solid, i.e. decentralised web-accessible knowledge graphs, where the data is shared using Solid pods.
Potential requirements:

- Means for identities, credentials, authentication, authorisation, sharing, revoking
- Means for data processing of data from a multitude of decentralised sources

Source: MANDAT project (funded by the German federal ministry for education and research).

UC 4.9: Knowledge Graphs for Arguments and Public Debates

Use-case sponsors: Giorgos Flouris (FORTH ICS, Greece)
Use-case description: There is lots of argumentative information on the web, which largely goes unexploited, as standard search engines (or even query engines) cannot query the structure of a debate or an argument. Current state-of-the-art cannot answer queries like “give me the premises that feature in the most popular arguments in favour of the Pfizer vaccine”. Also, it is not possible to navigate through the arguments themselves (only in the pages that contain them). This particular use case deals with arguments appearing in Greek newspaper articles that have been published online, and the aim is to assist the professional (journalist), and also the everyday user, to be better informed regarding arguments and debates on interesting issues. Data producers in this setting are journalists working in newspapers, and the data that they produce are the articles they publish online. Consumers are also journalists (mainly), but also interested citizens. The information need of these consumers is to understand debates, be better informed on interesting aspects, understand both sides of a debate, take informed decisions, etc. The process in this use case is then as follows: first, original data (arguments in articles) is mined by the newspapers (websites), using argument mining techniques, and the data are converted to a KG (under an appropriate ontological schema). Thereafter, tools for reasoning with and analysing the

arguments are built upon the KG. User information needs are then expressed graphically through appropriate applications (like the “argument navigator”), and are answered by executing appropriate queries over the KG. Hence, the KG is essentially the mediator between the producers (newspapers) and the consumers (citizens, but mainly journalists) who want to analyse public debates.

Potential requirements:

- Input is articles in Greek newspapers (conceivably, it could be in any language, given an appropriate argument mining technique; and in any medium, insofar it does not “confuse” the argument mining algorithms; but we haven’t checked any other setting, as it is out of the scope of the DebateLab project).
- In principle, other parties could also be producers of the above information; basically anyone publishing well-structured arguments on the web and/or websites allowing for public debates could also be a producer in this respect. Note that the DebateLab project does not address this (it is out of the scope of the project).
- In principle, other parties could also be consumers of the above information, including politicians and policy-makers. Note that the DebateLab project does not address this (it is out of the scope of the project).

Source: DebateLab project (“DebateLab will conduct research towards developing the theoretical infrastructure for representing, mining and reasoning with online arguments, while also delivering a suite of tools and services supporting the uptake and initial exploitation of the related technologies. This research will pave the way for a new Web paradigm, a modern agora, where the different types of arguments and human deliberation can be amenable to machine-interpretable representation and algorithmic processing.” https://debatelab.ics.forth.gr/)

UC 4.10: Vaccination Roll-Out

Use-case sponsors: Sabrina Kirrane (Vienna University of Economics and Business, Austria), joint work with Timotheus Kampik, Adnane Mansour, Olivier Boissier, Julian Padget, Terry R. Payne, Munindar P. Singh, Valentina Tamma, Antoine Zimmermann

Use-case description:

The synergy between autonomous agents that leverage Web technologies, and the resources (i.e., things, services, information) that they can exploit to achieve their goals can be illustrated through a motivating scenario that demonstrates the need for governance through the use of norms, policies, and preferences.

Consider a scenario whereby there is a vaccination roll-out (for example, for the COVID-19 pandemic), where patients who request vaccinations may have differing personal circumstances. For example, John, the patient, may ask to be vaccinated early as he is the caregiver for a vulnerable member of his family. As the demand for vaccines outstrips supply, policies exist that determine vaccination eligibility. Furthermore, as vaccines are available from different manufacturers (e.g., AstraZeneca and Pfizer-BioNTech) and can be of different types (e.g., mRNA or inactivated vaccines), these vaccination policies may vary depending on the recipient's personal health record and/or their preferences, as well as vaccine availability.

Patients may be registered to different clinics or health centres that follow local or national policies or guidance on health care. In this case, John is registered at a clinic in his country,

but has a preference for vaccination near his current residential address in State A. Each country or state can be seen as having an organisation of different health centres (clinics, hospitals, and vaccination centres), following their own national health policy that prescribe a specific specification/format for patient medical records, which may be held under disparate data models and access policies. Patient medical records are available (subject to appropriate authorisation) via web services using secure protocols across the web infrastructure and are encoded using established medical ontologies and vocabularies to facilitate record exchange within and across different national health organisations.
Vaccination centres store batches of vaccines within one or more temperature-controlled vaccine storage systems, where each storage system is responsible for both inventory management and the dispensation of the different COVID-19 vaccine batches from a specialised cold store via a robotic arm. The release and retrieval of vaccine batches is guarded by policies that must be satisfied to ensure appropriate use by authorised personnel (i.e., the vaccine guard). Once a batch of vaccines has been released, the vaccine doses should be used within a given time-frame to avoid spoilage and wastage, as they have a short shelf-life once thawed. Furthermore, a scheduling system determines which patients can be vaccinated in a given time-slot, based on vaccination demand and patient requirement (determined by the current vaccination policy that may change frequently).
This scheduling system should ensure that no vaccines are wasted, whilst ensuring that the policies determining which patients can receive which vaccines are adhered to.
Thus, the vaccination centre could be considered as an organisation that coordinates and exploits a variety of disparate information technology (IT) systems integrated through a Web infrastructure, including data management, scheduling, patient-facing services, and IoT-based physical assets such as the robot arm and the automated vaccine stores.
Typically, however, the task of orchestrating and using these different systems requires costly and time-consuming human intervention.
Finally, once a vaccine has been administered, the patient's medical records should be updated, and the patient should be able to prove their vaccination status if required (e.g., using a vaccine passport). The vaccine records should ideally be resilient to forgery whilst being privacy preserving and easy to administer; thus they may utilise a passport mechanism that itself exploits web-based resources such as verifiable credentials, decentralised data platforms, blockchains, etc.
This scenario underlines the need for systematic and scalable approaches for the governance of the different IT systems and IoT-based physical assets, taking into account the need to operate under different governance institutions, as well as interact across organisational boundaries (e.g., between countries). Such interactions must comply with applicable norms and policies encountered at different stages of the vaccination roll-out. For example, the European Commission recently proposed a Digital Green Certificate, recognised by all EU member states, that facilitates the safe free movement of citizens within the EU during the COVID-19 pandemic.
Given the intrinsic openness of the Web, coupled with the fact that autonomous agents can act on behalf of both patients and medical practitioners that need access to critical medical applications, the need for regulation, security, and privacy are of utmost importance.
Additionally, there is a need to facilitate coordination between stakeholders and ensure that relevant regulatory requirements are adhered to throughout.
Target users:

- Patients who request vaccinations.
- Health centre personnel responsible for procurement, scheduling, administration.
- Local and National governments who specify policies and oversee distribution.

DKG producers:

- The various governmental policies could be produced by local and national governments either in the form of a KG directly or alternatively in a manner that could be translated into KG based policies.
- Vaccine availability and scheduling data may be produced by local and national governments and/or the health centres.
- The patient's health record may be stored in a central health system or supplied by the patient or their physician.
- Devices such as the temperature-controlled vaccine storage systems store data in relation to inventory management.

DKG consumers:

- Local and national governments need statistical health information in relation to covid in order to make policies.
- Health centres need the vaccine availability and scheduling data.
- Patients need the vaccine availability and appointment data.
- Local and national governments need statistics in relation to vaccination to manage the supply chain.

Potential requirements:

- This scenario raises challenges due to the decentralised and dynamic characteristics of the involved organisations, policies, services, and stakeholders.
- Patients can request vaccination based on their interpretation of eligibility, which should then be validated by the vaccination centre.
- The handling of requests may require the collection of patient data from multiple sources and the mapping to a shared data model.
- The vaccination eligibility policy can change frequently due to, for example, the emergence of a new variant of concern, that may accelerate the need for vaccinating a specific population cohort or demographic.
- Changes to vaccination administration guidance may prioritise the use of certain types of vaccine over others for specific sub-groups (e.g., prioritising Pfizer-BioNTech over AstraZeneca, where possible, for certain patients based on medical risk assessments, or prohibiting certain vaccines for users where safety data is not available). Thus, the verification of vaccination eligibility for patients may rely on the aggregation of multiple policies, and on resolving inconsistencies between them.
- A further challenge involves ensuring that the process for adhering to the national prioritisation criteria is fair and transparent.
- Additional legal and ethical challenges arise when considering the complete socio-technical system, including electronic health record access and supply chains.
- Finally, vaccination scheduling needs to take into account patient availability (to avoid no-show cases and thus avoid vaccine wastage), as well as stock availability. Scheduling is therefore a collaborative process involving factors such as the vaccination centre capacity, vaccine availability, and patient availability. However, availability data may be distributed across multiple sources and, for privacy reasons, cannot be held centrally.

Existing resources:

- *Andrea Cimmino, Michael McCool, Farshid Tavakolizadeh, and Kunihiko Toumura. 2020. Web of Things (WoT) Discovery, W3C First Public Working Draft 24 November 2020. W3C First Public Working Draft. World Wide Web Consortium (W3C). http://www.w3.org/TR/2020/WD-wot-discovery-20201124/*
- *Sebastian Kaebisch, Takuki Kamiya, Michael McCool, Victor Charpenay, and Matthias Kovatsch. 2020. Web of Things (WoT) Thing Description, W3C Recommendation 9 April 2020. W3C Recommendation. World Wide Web Consortium (W3C). https://www.w3.org/TR/2020/REC-wot-thing-description-20200409/*

External resources:

- Governance of Autonomous Agents on the Web: Challenges and Opportunities; Timotheus Kampik, Adnane Mansour, Olivier Boissier, Sabrina Kirrane, Julian Padget, Terry R. Payne, Munindar P. Singh, Valentina Tamma, Antoine Zimmermann; ACM Transactions on Internet Technology Journal; *to appear 2022* [Preprint]

UC 4.11: Artificial Intelligence for Personalised Oncology

Use-case sponsor: Catia Pesquita (University of Lisbon, Portugal)

Use-case description:

Artificial intelligence (AI) offers vast potential for the future of personalised medicine. Promising tailor-made treatments for patients, AI could help win the fight against serious illnesses such as cancer. However, the introduction of AI-enabled personalised medicine also presents challenges. Chief among these is the translation of AI-based suggestions into practical decision-making processes and treatment strategies. The EU-funded KATY project (https://katy-project.eu/) will develop an AI-empowered personalised medicine system that will greatly assist medical professionals and researchers in utilising and interpreting AI data in their daily work. This next-generation technology will bridge the gap between AI data and medical application and thus become a powerful tool in diagnosing, treating and defeating serious illnesses.

The KATY project proposes an AI-empowered Personalized Medicine system that can bring medical "AI-empowered knowledge" to the tips of the fingers of clinicians and clinical researchers. The AI-empowered knowledge is a human interpretable knowledge that clinicians and clinical researchers can: understand, trust and effectively use in their everyday working routine. KATY is built around two main components: a Distributed Knowledge Graph and a pool of eXplainable Artificial Intelligence predictors. As a stress test and due to the lack of personalised clinical responses, KATY will be experimented in a low prevalence and complex cancer: Clear cell renal cell carcinoma (ccRCC).

DKG producers:

- Biomedical ontologies
- Biomedical and clinical data from public and open repositories
- Private biomedical and clinical data

DKG consumers:

- Machine learning methods that use the KG as a data source
- Intelligent interfaces that query the KG to explain therapy recommendations for patients to clinicians

- End users: Biomedical researchers/bioinformaticians who explore the KG to feed eXplainable AI approaches; Clinicians who interact (query/visualise) the KG to understand AI recommendations for their patients

Potential requirements:

- System for automated alignment and integration of ontologies from multiple domains
- Validation of alignments by expert consensus
- Integration of multi-modal data into the KG
- Integration of distributed private patient data (at each healthcare institution) with the public/shared Knowledge Graph based on public ontologies and data
- APIs to query private data in an integrated fashion with the public/shared KG
- Mechanisms to ensure updating and evolving the shared KG is also reflected at the private data level