Deliverable from the COST Action CA19134 “Distributed Knowledge Graphs”

Guidelines and Best Practices

Due date: 30 April 2023

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

cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

Authors:

Rob Brennan, Remzi Celebi, Mariana Damova, Michel Dumontier, Özge Erten, Michael Freund, Tobias Käfer, Pierre Maillot, András Micsik, Maryam Mohammadi, Catia Pesquita, Axel Polleres, Daniel Schraudner, Sebastian Schmid, Krzysztof Węcel, Jinzhou Yang, 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

Preface

This deliverable has been compiled by the network of the COST Action "Distributed Knowledge Graphs". This COST Action is a research and innovation network to connect research initiatives around the topic of Distributed Knowledge Graphs. The network gathered descriptions of funded research projects next to lessons learnt, from which we distilled guidelines and best practices.

The descriptions have been gathered in a virtual session in January 2023. In the following weeks, the project description sponsors improved the descriptions with the help of the editors. This phase has been followed by further editing by the editors. Lastly, the editors synthesised best practices from the lessons learnt.

In this deliverable, we present those best practices, next to the project descriptions.

The projects that have been provided by the network for this deliverable span a diverse set of application domains from 4 Horizon Europe categories: Health; Culture, Creativity, and Inclusive Society; Digital, Industry, and Space; Climate, Energy, and Mobility.

By sharing guidelines and best practices in this deliverable, the network wants to improve the state of affairs in how Distributed Knowledge Graph projects are planned and run.

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

Tobias Käfer
Michel Dumontier
Remzi Celebi
Andreas Harth
Antoine Zimmermann
Olaf Hartig

Best Practices

From our projects' lessons learnt, we distilled best practices along the dimensions: data modelling, system architecture, tools to build systems, and project management.

Modelling

Plan for ontology and vocabulary development efforts

All projects require terms for the knowledge graphs they work with. In some application areas, there are no suitable ontologies and vocabularies available yet (ARK-Virus, DigiBatMat), thus new ontologies and vocabularies need to be developed. Other areas, however, do have suitable ontologies and vocabularies, sometimes to an overwhelming diversity (KG-FAQ), where the matching and mixing of the ontologies and vocabularies requires effort (KATY). On top, the modelling prescribed by the existing ontologies may not be practical for the reasoning tasks at hand (CoSWoT, KATY), which requires mitigation efforts.

Determine required ontological expressivity and staff ontology engineering efforts accordingly

Different projects reported different requirements regarding ontological expressivity. These requirements can determine the semantic web training required to lead the modelling efforts. While in the medical domain, OWL ontologies are often the desired output (PMAGO) which require a lot of ontological expertise to get right (ARK-Virus), other projects mainly required an agreement on terminology, which could also get achieved by digitisation experts without any semantic web training (DigiBatMat). Project ARK-Virus followed a two-track strategy with OWL ontologies for core terms and SKOS taxonomies for other terms, which sped up their ontology development.

Architecture

Adopt solutions that mitigate established challenges due to decentralisation and distribution

Remote SPARQL endpoints can pose challenges when it comes to availability and data quality (KG-FAQ, DeKaloG).
Data catalogues are useful in larger projects (ARK-Virus, OMNIS)

Plan efforts to bridge gaps in solutions to address challenges due to decentralisation and distribution

Privacy and security are often required in decentralised and distributed settings, but the necessary technologies (as investigated in the Solid project or in workshops of this Action) are not there yet (ARK-Virus, CRISP) .

Small decentralised web resources are beneficial (MOSAIK), but reasoning on top of them using state-of-the-art techniques is not efficient (CoSWoT).

Tooling

Expect tool development efforts to bridge gaps between existing tools and systems

While nowadays there is a significant amount of mature semantic web tooling around, there are gaps between tools and approaches that projects need to fill: For instance, modelling using approaches of different ontological expressivity (ARK-Virus), interfaces from legacy sources (CRISP), scalability and integration of provenance for ETL workflows (KIWI), provenance authoring and SPARQL endpoint descriptions (DeKaloG), ETL pipeline (OMNIS), visualisations for decentralised or blank node-heavy data (MOSAIK, KATY).

Follow recent developments

Several projects highlighted the usefulness of recent standards and approaches such as SHACL and RML (COURAGE, AFA-KG, OMNIS, among others).

Project Management

Educate project partners early about semantic technologies

To get off the ground with the technical work, projects reported that knowledge about semantic technologies is often missing among project partners. To address this issue, corresponding teaching material can help the project get off the ground (MOSAIK, COURAGE).

Model the domain to develop a common understanding and structure of the domain of discourse

Not just for the technical work, but also to structure the ideas and to jumpstart the collaboration between partners, ontology development workshops to find a common terminology proved to be useful (KATY, DigiBatMat).

Lessons Learnt from Projects

DeKaloG: Decentralized Knowledge Graphs

Project description: DeKaloG follows the vision of a global decentralized knowledge graph that can be leveraged to answer questions at the scale of the web. For example, "*give me information about people who know Tim Berners-Lee*" or "*what is the number of famous scientists men and women per birth year?*". To face the LOD issues, DeKaloG promotes ATF (Accessibility, Transparency, and Findability) principles and a sustainable approach for implementing them. The contribution of the WIMMICS team to this project is to propose IndeGx, a framework to create a knowledge base used as an index of the descriptions of endpoints. IndeGx is a framework using a test suite described in RDF to send SPARQL queries to extract the description of KBs. It uses only standard semantic web technologies and each of its operations is tracable. It has been tested against 339 active endpoints over 6 months.

Lessons learnt:

What went well:

- Almost all SPARQL endpoints implement all of SPARQL 1.0, and most (85%) SPARQL 1.1 features supported;
- IndeGx managed to extract even complex statistics from most endpoints using pagination. Given enough time and with the right page size, our results indicate that is should be possible to extract statistics from any endpoint
- Readability: If there is at least one label in a base, one can expect 53% (±38%) of the resources to be labeled on average. On average, 89% (±21%) of the URIs are below 80 characters with no parameters.

What went wrong:

- Less than 10% of the indexed KBs have any kind of provenance metadata.
- This led to the creation of Metadatamatic to help to create basic dataset descriptions.
- Only 20% of the KBs contain language tags and only for 25% (±36%) of literals on average.
- By far, the main error received during indexation came from the unavailability of endpoints.
- We did not find a way to automatically discover new query endpoints.

Project consortium: Université Nantes (coordinator), Inria Sophia (WIMMICS), INSA Lyon
Duration: 3 years
Funding agency: Agence nationale de la recherche (France)
Sponsor: Pierre Maillot (pierre.maillot@inria.fr)
rdfs:seeAlso: https://dekalog.univ-nantes.fr/ http://prod-dekalog.inria.fr/ https://wimmics.github.io/voidmatic/ https://hal.science/hal-03946680 https://hal.science/hal-03652865

KIWI: AI-supported Value Stream Optimization

Project description: The objective of the KIWI research project is to make processes and value streams in logistical processes at an airport transparent and to identify requirements

and potential for improvement. The idea is to automatically record and analyse process models through the use of artificial intelligence, process mining and IoT technologies. Process data already available at the airport is analysed and necessary actions are derived. Process optimization is to be achieved through the use of continuous value stream management on the basis of existing data and newly collected data.
The existing processdata are distributed in relational databases on different servers owned by various stakeholders. In order to perform analyses on the basis of all data sets, we modelled the data by using ontologies and other semantic web technologies. To describe the data we used the Simple Event Model Ontology as a basis and extended it for our use case. We have mapped the existing data into the RDF format and stored it in a knowledge graph. The data in the knowledge graph is the foundation for the next steps in the project.
Project consortium: Flughafen München GmbH, Fraunhofer IIS, TrilogIQa
Duration: 10/2021 – 09/2024
Funding agency: German state of Bayern, IuK
Sponsor: Michael Freund
Category: 4. Digital, Industry, and Space
Lessons learnt:

- At the beginning of the project, many partners in the project were not familiar with linked data principles or semantic web technologies. Therefore, a joint workshop on the basics and advantages proved to be useful at the beginning of the project. This helps to clarify the planned approach and to answer any questions that may arise. The advantages can be illustrated using the FAIR principles.
- Data from different stakeholders is often stored in different data formats even though the same information is stored. An example is a timestamp, which can be stored using a 24-hour clock, a 12-hour clock or using the ISO 8601 standard. The data formats must be unified, which is easy to do using a general-purpose programming language. Typical preprocessing tasks we encountered were the transformation of date and time to ISO 8601 format or the generation of unique IDs. The data preprocessing scripts then become part of the mapping pipeline, making it more complicated and harder for others to reproduce.
- When mapping large amounts of data, computers quickly reach their memory limits. To prevent this, the data must be streamed dynamically into memory. This must be considered from the beginning when creating the scripts.
- In our use case, we are dealing with a logistics process with various events and associated timestamps. To make the process flow visible in the knowledge graph, we use a 'next' relation, which explicitly shows the sequence.
- If different data sources are loaded into a single graph, all queries can be made easily, but the source database is difficult to trace. Since different databases can contain different and contradictory information, it is essential to find out from which database the information originates. We therefore use subgraphs for the different source databases, which complicates the queries somewhat, but the origin of the information is always known via the subgraph mapping.

MOSAIK: Methodology for Self-organizing Aggregation of Interactive Components

Project description: Future information systems, e.g. the Industrial Internet, the Internet of Things and Smart X systems, are going to be more open, distributed, complex and networked than current operational IT. Therefore, MOSAIK is developing a system architecture that works without central elements. This is done with the aim of increasing fail-safety and avoiding expensive central infrastructures as well as power structures and monopolies or oligopolies. For this purpose, we ask ourselves the central question:
How can individual components communicate with each other in such a way that they show goal-oriented behaviour among and with each other?
The aim of MOSAIK is to research methods for self-organised aggregation of components into systems at development time and the adaptation of the aggregates at runtime. In the process, the aggregated system should fulfil predefined properties or produce defined phenomena and be resilient to perturbations. The basic idea is to take the intelligence of the actors and shift it into the environment, inspired by the principle of stigmergy. Based on this, actors should work through the affordances stored in the environment via simple control logic, whereby a predefined overall goal is achieved in a decentralised manner.
From the user‘s point of view, the researched methodology should contribute to the reduction of complexity in the development and operation of IT systems. The platform should not be operated by a single company, but consist of a set of technology standards and conventions that enable the digital participation of a large number of different companies.
In addition to researching the methodology, further goals of MOSAIK are the development of a runtime environment as open source and preliminary work on the standardisation of the methodology as well as its prototypical use in practice. The research topics addressed in MOSAIK concern fundamental concepts and approaches with the potential for broad application, for example in scenarios around the Internet of Things, Industry 4.0 as well as Smart X systems.
Project consortium: Friedrich-Alexander Universität Erlangen-Nürnberg, DFKI, Otto von Guericke Universität Magdeburg, Robert Bosch GmbH, NETSYNO
Sponsor: Daniel Schraudner, Sebastian Schmid
Duration: 03/2019 - 07/2022
Funding agency: German Ministry for Research and Education (BMBF)
Horizon Europe clusters: 4. Digital, Industry, and Space
Lessons Learned:

- We quickly learned that for our data it was very beneficial and conceptually easier to use different, smaller Web Resources than to have a big triple store with a SPARQL endpoint. Admittedly, the organisation of your data depends on the goal of your project, but in terms of flexibility and fast changes this came in handy during the life of the project as smaller resources were easier to adapt. We organised our data with Linked Data Platform, which helped to conceptually cluster data and also, motivated by our goal, have small modular programs (we called them "agents") work easily on your knowledge graph. Such big design decisions were based on real data and processes early in the project, as we actively negotiated with our partners to get such data. Adapting to different data later on may be difficult, especially when your project

results shall be applied to industrial brownfield systems or reflect the hierarchies of a real machine. To avoid this, we also defined uniform schemas for the data (e.g. in the form of ShEx shapes) early on. An example of the data organisation in MOSAIK, based on the manufacturing systems of one of our project partners, can be found below.
See: Charpenay. V. et al., "MOSAIK: A Formal Model for Self-Organizing Manufacturing Systems," in IEEE Pervasive Computing, vol. 20, no. 1, pp. 9-18, 1 Jan.-March 2021, doi: 10.1109/MPRV.2020.3035837.

- If your project includes the dynamic manipulation of Knowledge Graphs, e.g. via interaction of agents (as in our use case), the use of simulations can be a good benchmark for performance. As the results of manipulations may be reflected by effects in the underlying system, it is interesting to see if the overall performance is following the expected process, or if deviations occur. In a sense, this uses the idea of Rapid Prototyping. For the purpose of performance checks, we used agent-based simulations that mimicked the use of a knowledge graph for a system of transporters that used MOSAIK. This gave us a baseline estimation to see if a system using the knowledge graphs later on shows the self-organised behaviour we expect, if the discussed concepts (see above) are valid, and could be used as visualisation to explain the use case (see below), all without committing too early.
  See: Schmid, S., Schraudner, D., and Harth, A., "Performance Comparison of Simple Reflex Agents Using Stigmergy with Model-Based Agents in Self-Organizing Transportation," 2021 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C), DC, USA, 2021, pp. 93-98, doi: 10.1109/ACSOS-C52956.2021.00071.
- Get a common understanding of the technologies used across all involved parties. As you probably act as an expert in your project, it is also your duty to take care of partners in your project that may have less experience. Simple ways to increase understanding are, e.g. by giving other project partners small tutorials about RDF, knowledge graphs, the Semantic Web. If applicable, hackathons are a great way to present hands-on possibilities to get to know the techniques and to make step-by-step progress already very early in the project. By this, all partners, even non-experts, have at least basic knowledge and are kept in the development and research process. This can also be reinforced by the use of visualisations, see below, and also by publishing your code as Open Source on Github, which is also a straightforward way to generate outreach.
- Visualise your data by building GUIs for the Knowledge Graph. We noticed that a visualisation is a key element that helps in two ways: on the one hand, it helps developers and experts to easily debug the data and notice possible code defects; on the other hand, especially non-experts which may be associated in the project, can benefit greatly from a simple visualisation what has been done and how the underlying knowledge graph works – even though they do not have expert knowledge about the applied techniques, implemented use case, or are used to another style of working, e.g. only object-oriented. Still, be aware that visualisation may need additional time in your project and is usually regarded as something "on top" instead of a core requirement, thus keeping it simple but informative. An example of a simple visualisation we did to present the use of knowledge graphs and

an intralogistics use case can be found in Schraudner, D., Charpenay, V. (2020). An HTTP/RDF-Based Agent Infrastructure for Manufacturing Using Stigmergy. In: , et al. The Semantic Web: ESWC 2020 Satellite Events. ESWC 2020. Lecture Notes in Computer Science, vol 12124. Springer, Cham. https://doi.org/10.1007/978-3-030-62327-2_34

rdfs:seeAlso: https://mosaikprojekt.de/

DigiBatMat - Digital platform for battery material data, knowledge, and its interconnections (Tobias Käfer) - WG2

Project description: Project about digitising the workflow of battery scientists using a plattform with a semantic interface
Project consortium: Institute for New Materials Saarbrücken, University of Applied Sciences Aalen, Technical University Braunschweig, August-Wilhelm-Scheer-Institut, Karlsruhe Institute of Technology (KIT)
Duration: 2021/03/01 – 2024/02/28
Funding agency: German Ministry for Research and Education (BMBF) within MaterialDigital, ProZell clusters
Sponsor: Tobias Käfer
rdfs:seeAlso https://prozell-cluster.de/en/projects/digibatmat/ Slides
Horizon Europe Category: 5. Climate, Energy, and Mobility
Lessons learnt:

- If you are just after terminology and little formalisation, Ontology engineering works without ontology engineers!
- In this and many other projects, people try to model processes using semantic technologies, the output is almost always flakey

CRISP: Crisis Response and Intervention Supported by Semantic Data Pooling

Project description: The CRISP Project represents a data-driven approach to Crisis Response and Intervention. It considers both the short-term management of disasters as well as long-term economic impact assessments, at fine-grained regional and temporal granularity. For this purpose, CRISP will ingest data from multiple heterogeneous sources, shedding light on the impact of response and intervention strategies in close to real-time. The result is a comprehensive and continuously updated data pool, which represents a key asset for semantic modeling and impact forecasting. CRISP will not only increase the transparency of crisis response and intervention processes but also capture how the resulting outcomes are being perceived by citizens and professional stakeholders. The CRISP KG aims to provide a data backbone of Austrian infrastructure systems. By connecting data on population, medical services, weather, transport, and utilities, CRISP KG gives users a way to understand how interconnected systems react to crisis and shock situations.

Lessons Learnt: This project turned out so far really more about making data available and shareable between stakeholders (e.g. in stakeholder workshops with firefighters, emergency response services, etc.) and the necessary controlled data sharing platform, than KG technologies. We are now in the middle of the project and just starting to actually populate the KG actively and linking to existing KGs. The main technical/engineering challenge at the moment is establishing interoperable interfaces from the existing legacy sources, rather than RDF on top. As for the RDF on top, we think that agreeing on relevant metadata to keep and maintain to assess source data quality (e.g. fine-granular, but in fact interpolated weather data) will be crucial.
Some of the reference datasets could be reused, e.g., GIS data; but which one can work across different countries; dealing with public administration can take time, so workaround have to be made to advance in the project; need to take into account access control, policies, etc when data is not open; if we could have more showcase, it could convince partners to use more these technologies.
Sponsor: Axel Polleres
Project consortium: webLyzard technology (Austrian SME, coordinator), Complexity Science Hub Vienna, Austrian Central Institution for Meteorology and Geodynamics, nexyo (Vienna-based startup), KDZ Centre for Public Administration.
Duration: 12/2021 - 11/2024
Funding agency: FFG (Austrian Research Promotion Agency) **I**CT of the Future Program
rdfs:seeAlso: https://www.crisp-project.org/

KG-FAQ

Project description: The semantic web community has created numerous knowledge graphs (KGs), which are typically accompanied by basic or rich metadata to aid in their reuse. To query these KGs, users can use their provided SPARQL endpoints. However, formulating SPARQL queries, especially complex ones, can be challenging for users without knowledge about the structure of the KG and its domain. To address this issue, we propose a tool called KG-FAQ, which generates a set of frequently asked queries based on the query logs of a KG. In this study, we evaluate the usefulness of the selected query set for the Bio2RDF KG by conducting a user study to investigate its efficacy in formulating new queries.
Lessons learnt: The project allowed us to identify challenges to build knowledge graph metadata:

- Metadata is being researched by several research communities using different keywords such as dataset characterization, data summarization, dataset assessment or dataset profiling. There is now a confusion in the research community about the terminology in the area, further increased by the fact that certain terms are often used with different meanings in the relevant literature, denoting similar, but not identical research directions or concepts. This lack of terminology and classification hinders scientific development in this area.
- To address the challenge of evaluating the generated metadata, a user study and query formulation task have been done for evaluation purposes. However, it should be noted that this method of evaluation is time-consuming, costly and irreproducible.

- It is not advisable to rely heavily on remote endpoints for conducting experiments due to the high volume of queries and potential delays. Instead, it is recommended to download the datasets and query them from local endpoints, as this method may be more efficient.

Duration: Sep 2022 - ongoing (project KnowGraphs funding until September 2023)
Funding: H2020 ITN
Sponsor: Maryam Mohammadi
rdfs:seeAlso https://github.com/marmhm/kg-metadata-generation
rdfs:seeAlso https://knowgraphs.eu/

CoSWoT: Constrained Semantic Web of Things

Project description: CoSWoT objectives are to propose a distributed WoT-enabled software architecture embedded on constrained devices with two main characteristics:

- it will use ontologies to specify declaratively the application logic of devices and the semantics of the exchanged messages;
- it will add reasoning functionalities to devices, so as to distribute processing tasks among them. Doing so, the development of applications including devices of the WoT will be highly simplified: our platform will enable the development and execution of intelligent and decentralised smart WoT applications despite the heterogeneity of devices.

In CoSWoT, WoT applications will rely on a platform hosting the base services. Besides traditional services, it will host extensions that correspond to two scientific barriers:

- The use of ontologies as a generalised model for exchanges between heterogeneous devices. A joint statement from AIOTI WG3, IEEE P2413, oneM2M, W3C positions ontologies as key enablers for semantic interoperability on the WoT. However research questions remain concerning (i) the adequation of existing ontologies to the target application domains; (ii) the applicability of theoretical principles developed in a variety of protocols and standards, in the context of data streams; (iii) the discovery of heterogeneous devices, their services and how to solicit them.
- Distributed and embedded incremental reasoning. Devices become powerful enough to offer storage and processing; new architectures appear, based on edge computing including devices such as sensors and actuators. The data streams provided by sensors require to perform incremental reasoning tasks. Research questions remain on (i) how to embed reasoning in devices with various capacities, it requires specific optimisations; (ii) how to efficiently distribute reasoning tasks among devices. Smart agriculture is a typical application domain of such WoT architectures, where the surveillance of cultivated fields requires various sensors that push streaming data, which must be collected and reasoned upon to take decisions executed by actuators. Smart buildings is another such typical application domain where added-value application services involve other verticals such as energy management, e-health, or ageing well. We will define use cases and requirements for smart agriculture and smart buildings, run simulations, and then lead real experiments.

Lessons learnt: In order to make all components of the CoSWoT architecture semantically interoperable, our approach is to agree on the vocabularies/ontologies used. This requires identifying domain ontologies first, then extend these ontologies for each specific use case. However, rather than simply making an ontology module for each case, we found it best to agree on generic *ontology patterns* for extending the domain ontologies. The same pattern can then be reused across use cases and improves interoperabilty. It also helps the designers of the use-case-specific ontology modules to generate semi-automatically their extensions.
We also realised that the best *conceptual* structures that precisely and accurately describe a situation are not necessarily the best *practical* structures when instantiated for certain concrete cases. In particular, the accuracy of the conceptual models often comes with a complexity in structure, having many intermediary nodes that complexify reasoning. In order to accommodate efficient reasoning on constraint devices, we introduced "short cut" models where part of the information is lost but most of the practically useful inferences are preserved. For instance, a physical quantity that relates to a property of an object of interest is most accurately defined as the result of a measurement or evaluation of said property of said object. E.g., a table has a width which can be measured to provide a physical quantity. We have `<table> -> <tablewidth> -> <measurement> -> <quantity>`. Different methods of measurement may yield different quantities, and the quantity may change over time. But in the case of a table width, in most applications, it is convenient to assume that the value is fixed and related to the table itself.
Another interesting lesson is that reasoning over small knowledge graphs is usually not very efficient, in terms of memory footprint and execution time, with highly scalable reasoners. Optimisation techniques dedicated to small graphs, that do not necessarily scale, can outperform fast reasoners by several degrees of magnitude when the number of triples is small (see Bento et al. 2022).
Project consortium: INSA Lyon (coordinator), INRAe (national research institute in agriculture), Mines Saint-Étienne, Univ. Jean Monnet Saint-Étienne, Mondeca (private company)
Duration: 48 months
Funding agency: Agence Nationale de la Recherche (France)
Sponsor: Antoine Zimmermann
rdfs:seeAlso: Alexandre Bento, Lionel Médini, Kamal Singh, Frédérique Laforest, *"Do Arduinos dream of efficient reasoners?"*, ESWC 2022, Hersonissos, Greece, May 29 - June 2, 2022.
rdfs:seeAlso: https://coswot.gitlab.io/

KATY: Knowledge At the Tip of Your fingers: Clinical Knowledge for Humanity

Description: The KATY project aims to develop an AI-empowered personalized medicine system to assist medical professionals and researchers in diagnosing patients more accurately, making predictions about their future health, and recommending better treatments. KATY will tackle the challenge of translating AI-based suggestions into practical decision-making processes and treatment strategies that clinicians can understand and trust

by combining high performing blackbox machine learning approaches with a comprehensive knowledge graph. The KG will serve as input to AI methods as well as encode the AI outcomes themselves to create a shared semantic space for data, scientific context and predictions capable of supporting explanation methods. To build the KG we are: integrating 28 ontologies across a variety of domains (clinical features, genomics, proteomics, histopathology), developing a new ontology for a novel scientific domains: immunopeptidomics, and mapping scientific and clinical data to the ontologies. Tools we are using: GraphDB: to host the KG and support querying, VocBench: to manage the ontologies and vocabularies, Matcha: to produce the ontology alignments, RMLMapper/Streamer: to map the data to the ontologies

Project consortium: University of Vienna, The French Atomic and Alternative Energy Commission, Centre Hospitalier Universitaire de Grenoble, European Research and Project Office GmbH, Health Policy Institute, National and Kapodistrian University of Athens, Medical School, University of Rome "Tor Vergata", DS Tech, Fondazione IRCSS Istituto Nazionale dei Tumori, Personal Genomics, University of Gdańsk, FCiências.ID – Association for Research and Development in Sciences, Caretronic, Fundació Eurecat, University of Zaragoza, PredictBy Research and Consulting SL, Lund University, The National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", University of Edinburgh, University of St Andrews

Duration: 01/2021 - 12/2024

Funding agency: European Union's Horizon 2020 research and innovation programme

Sponsor: Catia Pesquita (clpesquita@ciencias.ulisboa.pt)

Horizon Europe Category: 1. Health

Lessons learned:

- Biomedical ontologies have specificities that are not common in other domains which can be problematic for existing tools and require tailoring or novel methods. Be aware of this, as it can result in misunderstandings and delay projects. Peculiarities include:
    - reuse via copy-paste → loss of semantics/versioning
    - extensive use of dbXRef to encode a variety of semantics: equivalentClass, subsumption, etc
    - very relevant lexical component, varying degrees of axiomatic richness, few to none instances
- Ontology Matching is far from "solved" - we are good at pairwise and equivalence matching but otherwise insufficient for more sophisticated applications: we needed to link several biomedical ontologies to integrate data across domains with more complex relations, where we needed to develop holistic matching approaches. We also need "complex mappings" without shared instances, which we are developing
- Existing tools for mapping data (RMLMapper/Streamer) are not entirely up to the task when it comes to large amounts of data with duplicates/coverage and usability for novices/non-experts: it would be great to have a more automated approach for a "first pass". Challenges such as SemTab are looking into this, but we are still having to do a lot of this mostly manually
- Visualizing the KG (part of it) is essential to support XAI, however the OWL to RDF conversion results in a lot of blank nodes, which hinder visualization in typical tools

(GraphDB, WebVOWL, etc) → the graph paradigm is a great mental model, but refining the graph for visualization is still not an "off-the-self" solution
- The semantic modelling of an emerging scientific domain is very helpful for scientists to formalize their new theories, make their own assumptions explicit and reach a consensus → quite different from the experience of "modelling after the fact"

rdfs:seeAlso: Silva, Marta Contreiras, Daniel Faria, and Catia Pesquita. "Matching multiple ontologies to build a knowledge graph for personalized medicine." *The Semantic Web: 19th International Conference, ESWC 2022, Hersonissos, Crete, Greece, May 29–June 2, 2022, Proceedings* . Cham: Springer International Publishing, 2022. https://2022.eswc-conferences.org/wp-content/uploads/2022/05/paper_115_Silva_et_al.pdf
rdfs:seeAlso: https://katy-project.eu/

OMNIS: integrated multi-search engine for resources of National Library of Poland

Project description: The project "OMNIS e-service" consists in the implementation of a nationwide IT system offering a single access point to the collections of Polish libraries and the current offer of the publishing market in Poland. As part of the project, three separate and integrated e-services were launched. The one in which we were involved was the 'Integrated OMNIS search engine', a service that allows us to search the resources of all Poland libraries from one place and search the full text content of books, magazines, and articles in digital form. Our task was to make resources (catalogue among others) of the National Library of Poland available as linked data. Not only was it necessary to coin URIs for local resources, but additionally the had to be linked to external sources used by other national libraries or common-sense knowledge bases. We thus distinguished and linked: works and their manifestations (e.g. books, articles), people (e.g. authors, editors), places (e.g. cities, countries, also historical), organisations, and subjects for classification. We linked with VIAF (Virtual International Authority File), where libraries are obliged to submit ORCID, Wikidata, Geonames, and for some cases also DNB (Deutsche Nationalbibliothek). For linking resources within this knowledge graph we used our own properties as well as acknowledge metadata used by libraries, especially RDA - Resource Description & Access, elements, guidelines, and instructions for creating library and cultural heritage resource metadata that are well-formed according to international models for user-focused linked data applications.
Funding: European Commission, The European Regional Development Fund (ERDF) "Digital Poland"
Sponsor: Krzysztof Węcel
Project Consortium: I2G and Inwebit
Horizon Europe Category: 2. Culture, Creativity, and Inclusive Society; preserving cultural heritage, offering broader access to knowledge, easier identification of source works
Lessons Learnt:
- We designed and implemented our own processing pipeline to integrate various data sources and produce RDF output. For the semantic lifting itself, we used RDF Mapping Language (RML), which offered nice abstraction from taking care of RDF

syntax and persistency. RML was used to extract triples from our target knowledge bases that were available mainly in XML and JSON. Results of the triplification were then uploaded to Virtuoso Open Source server and queried with SPARQL.

- Relying on RML was basically a good decision, although it required additional coding in some border cases (especially with iteration over some elements). There were some limitations to the mapping language that could not be solved out of the box. Fortunately, it was possible to implement and register our own classes in Java with required behaviour and add them to scripts in RML. Taking into account the development and standardisation of RML, we would follow the same path in future projects. What also went well is that we had a consistent processing pipeline, so that when the source files changed, we could process without hesitation.
- One of the challenges was the size of input data – the dumps from Wikidata, VIAF, and Geonames. On the one hand, processing time was significant, and on the other hand, there were a lot of entities to match, hence also a challenge of false positives. The quality of source data was the main concern during the construction of a knowledge graph, particularly that it consisted of many entities (counted in millions). Quality issues could be divided into two groups: identity and completeness. There were multiple occurrences of the same item within the same source knowledge base. For example, VIAF was surprisingly cluttered with multiple instances of the same people. Another challenge was the distinction (identity) of works where the only hint was the title of the work. Matching works from internal databases of the National Library with works described in Wikidata was error prone. The main problem of Wikidata was the completeness of the data, both with respect to people and places. Another notorious problem was disambiguation. The description of entities has not always allowed us to match them correctly.
- Looking from the perspective of time, we could better leverage dependencies between various identifiers in order to make sure that we unambiguously identify an entity. In particular, VIAF and Wikidata had additional identifiers for other data sources, such as ISNI, ORCID, DNB, and others. Creating and interpreting the 'network of identifiers', not just one relation at a time, would definitely allow one to spot the problematic points.

PMAGO: Predicting missing annotations in Gene Ontology with Knowledge Graph Embeddings and True Path Rule

Project title: Predicting missing annotations in Gene Ontology with Knowledge Graph Embeddings and True Path Rule
Project description: In this study, we focus on Gene Ontology Annotation knowledge base, and address the gap in evidence between gene products and their annotations by making link predictions using Knowledge Graph Embedding (KGE) methods. Through the application of the True Path Rule (TPR) in the training stage of KGE, we were able to improve the performance of traditional KGE methods.
Duration: 19 September 2022 - 28 November 2022)
Sponsor: Özge Erten
Horizon Europe Category: 1. Health

Lessons learnt: Hereditary features (subclassof, partof, etc.) used in machine learning models can improve the accuracy performance.

ARK-Virus: Access Risk Knowledge platform for mindful governance of Virus infection prevention and control risk

Sponsor: Rob Brennan
Project Description:
ARK-Virus studied novel risk governance methods for infection prevention and control (IPC) by deploying the Access-Risk-Knowledge (ARK) knowledge graph platform in three healthcare providers. ARK guides users in a structured socio-technical analysis and governance of a clinical system: identifying and quantifying risks, developing and implementing risk mitigation projects while linking the analysis, risks and project progress to evidence. Users and stakeholders are supported by report generation, visualisations and flexible navigation of the linked risks, projects, analyses and evidence. ARK supports data governance best practice with extensive metadata and a data catalogue of evidence like datasets and publications. ARK uses privacy by design for controlled sharing of evidence between organisations. In parallel, a Community of Practice (CoP) has been developed, which connects ARK users from different healthcare organisations to facilitate inter-organisational learning and collaboration. The goal of ARK-Virus is to build a socio-technical infrastructure to support organisations in designing, implementing, and governing knowledge-based solutions to complex problems.
Knowledge graphs are the core data storage, organisation and interoperability technology used by the ARK Platform. It has a CKAN data catalogue with the DCAT plug-in. The core model of risks created and maintained by ARK is expressed as an OWL ontology. We rapidly created domain taxonomies to describe the health domain addressed in the project, Infection prevention and control (IPC). Our graph in the ARK platform was partitioned into separate named graphs based on the owning organisation as it was very important to not let sensitive data leak between the organisations. This required the design of a data security classification scheme and access control mechanism in the platform. Interlinking of platform concepts with external Linked Data resources was supported. The internal representation of risks and risk analysis supported the generation of reports for various stakeholders. Report generation as HTML/PDF and RDF was supported.
Project Consortium: Dublin City University, Trinity College Dublin, St James's Hospital, Beacon Renal, Dublin Fire Brigade (funded by Science Foundation Ireland)
Lessons Learned:

- Privacy by design and flexible security are key for modern DKG systems. We observed increasing demands in finding ways to define access, authority and security for knowledge graphs These are areas that still need some research, for example, on flexible access control mechanisms like SOLID. One data storage approach we applied was by observing that DCAT metadata is easier to share than detailed instance data, especially for sensitive data like hospital data (even if it is not about patients). This means that DCAT records are easy to share and enable more detailed sharing while requiring less protection.
  See: J Hernandez, L McKenna, R Brennan, TIKD: A Trusted Integrated Knowledge

Dataspace for Sensitive Data Sharing and Collaboration. Data Spaces: Design, Deployment and Future Directions, 265-291, https://link.springer.com/chapter/10.1007/978-3-030-98636-0_13

- Semantic web provides lots of standards to support modern data governance metadata models (for example, W3C DCAT, W3C PROV, W3C DQV, W3C SHACL, DPV) and these are general missing. But we do not make enough noise in e.g., the data governance community to tell them about the standards. Data governance platforms are being deployed in many organisations now due to increasing awareness of its importance and the complice issues raised by GDPR. In contrast the market for data governance platforms is dominated by a few, very expensive solutions aimed at highly regulated industries like finance and even these solutions do not use standards. Solutions that are good enough and standards-based are missing.
- See: LinkedDataOps:Quality Oriented End-to-end Geospatial Linked Data Production Governance, http://www.semantic-web-journal.net/content/linkeddataopsquality-oriented-end-end-geospatial-linked-data-production-governance
- Building DCAT data catalogues into your DKG system increases flexibility and provides standard user-oriented features e.g. dataset search, visualisation, metadata extraction. We built the catalogue into the workflow of data analysis tasks (that had to reference data). One way to increase the utilisation of data catalogues is if within an organisation you have have data stewards, with their names in the catalogue so the catalogue helps you find data and the people responsible for it. Data stewards person responsible for the dataset: the domain expert and the person you go to if you have problems with the data quality. CKAN is an open source system that supports DCAT. There is also Microsoft Purview or Google Data Catalog for cloud based solutions. Technology is important but an organisational shift is also necessary to deploy data governance.
  See: McDonald, N.; et al.. 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
- Knowledge modelling mixing SKOS and OWL can accelerate deployment. OWL models are very expressive but this means they are expensive to create and to get right. They typically require validation by multiple experts and evolve over time as your understanding of the domain evolves. SKOS models on the other hand are easily created by non-RDF experts. Even if not very rich, having a SKOS model ensures a consistent set of IRIs, definitions, and labels are available for these concepts in the system. It is possible to gradually migrate concepts to OWL as or if required.
- Note that different publication tool-chains are required for different model types and this increases the complexity of deployment of this solution e.g. Widoco for OWL, SKOS-Play for SKOS. In our example system we used 2 OWL ontologies and 4 SKOS terminologies, see https://openark.adaptcentre.ie/#ontologies.
- Build your system so it can accommodate non-RDF data sources as well. KGs give an excellent way to link together data resources that may reside outside the graph

(Semantic web technologies can offer a good integration layer). Some data may be more efficiently processed in their own native formats. FAIR and 5 star open data mean that the graph can add value by linking and describing these data resources. Sometimes you need to go outside the RDF: "think outside the graph!"

rdfs:seeAlso: https://openark.adaptcentre.ie
Relevant Horizon Europe clusters: Health

COURAGE: Cultural Opposition – Understanding the CultuRal HeritAGE of Dissent in the Former Socialist Countries

Project title: *Cultural Opposition – Understanding the CultuRal HeritAGE of Dissent in the Former Socialist Countries*
Project description: COURAGE created a comprehensive online knowledge graph (linked data registry) of existing but scattered collections on the histories and forms of cultural opposition in the former socialist countries and thereby made them more accessible. The knowledge graph powered several project outcomes: the webpage, the online exhibition, learning material, etc. It is browsable here: http://cultural-opposition.eu/registry/ and data snapshots are uploaded to Zenodo as well.
Project consortium: MTA BTK Research Centre for the Humanities, Hungarian Academy of Sciences, IFIS PAN Institute of Philosophy and Sociology, Polish Academy of Sciences; TCD Trinity College Dublin; IOS The Institute for East and Southeast European Studies, University of Regensburg; MTA TK Centre for Social Sciences, Hungarian Academy of Sciences; LII Lithuanian Institute of History; CUNI Charles University, Prague; UB University of Bucharest; HIP Croatian Institute of History, Zagreb; Comenius University, Bratislava; The University of Oxford; MTA SZTAKI Institute for Computer Science and Control, Hungarian Academy of Sciences
Duration: 1 February 2016 - 31 January 2019 (~3 years)
Funding agency: H2020
Sponsor: András Micsik
rdfs:seeAlso: http://cultural-opposition.eu/
Horizon Europe category: 2. Culture
lessons learnt: to cover the domain, ontologies were built from scratch, which takes time: more than one year; first we created the ontologies and then provided mappings to other ontologies afterwards; the website makes use of the KG using SPARQL queries; importance of having tools to validate inputs; there is a danger of allowing everyone to edit, must use access control; In hindsight, using SHACL could have been better for validation and integrity control; attempt to use Semantic Tech in humanities with people with no prior training; Access control could be useful already on the level of allowing the edition of a portion of the KG for groups of people based on their affiliations

AFA-KG: Automated FAIRness Assessment of Knowledge Graph

Project title: *Automated FAIRness Assessment of Knowledge Graph*

Project description: Applying FAIR data guidelines (Findable, Accessible, Interoperable, and Reusable) to knowledge graph (KG) can facilitate open access to the data, which could break down the barriers, of which the difficulty of accessing and reusing the existing information. The FAIR principles do not specify technical tests, however automated FAIRness evaluation tools need to be developed as they are useful to improve the state of affairs in data sharing. We test the outputs of different FAIRness evaluation tools for consistency.
Duration: from 01-2022 to 01-2023 , 1 year, finished
Funding agency: Maastricht University
Sponsor: Michel Dumontier & Remzi Celebi (+ presented by Jinzhou Yang)
Horizon Europe Category: 1. Health
Lessons learnt: Tools for evaluating FAIRness give completely different results; need to understand how the metrics are implemented by the tool (knowledge of the FAIR guidelines are not enough); automatic assessment only based on metadata
Additional automation could be done using SHACL on the data; question is how to discover the SHACL that must be used to validate the data

ISMoSeDe: Information System for Monitoring of Sediment Deposition

Project title: *ISMoSeDe: Information System for Monitoring of Sediment deposition*
Project description: The project consists in creation of a semantic information infrastructure mixing satellite data, in-situ measurements, geo-spatial information, domain knowledge in the water resources management domain implementing a distributed architecture with an underlying semantic data lake extended with intelligence that makes calculations, generates forecasts, formulates queries, throws alerts and drives the communication with the users, and GUI featuring a dashboard with synchronized table view, graph view and GIS visualization.
Project consortium: Mozaika, Ltd., IMDC, SISTEMA, GmbH
Duration: 02/2021 – 02/2023
Funding agency: ESA PECS Programme
Sponsor: Mariana Damova
rdfs:seeAlso http://ismosede.bg/
Lessons learnt: A semantic data lake where all critical information revolves around geospatial objects and geolocations proves to be a very suitable design, as it allows the use geospatial reasoning to describe or access information. Knowledge graphs based on geospatial information are very easy to maintain, and to query because all objects are described with respect to their geospatial positioning. All objects are represented with geospatial information. The queries then do the work of representing the geospatial relations between the objects in order to identify them. The availability of geospatial information lifts the need to explicitly describe relations between objects, as they become obsolete. We connected satellite data, in-situ measurements to geospatial objects along with the hydrometeorologic stations and the outputs of hydrodynamic models, that made possible to easily visualize the information on tables, graphs and maps at the same time.