Application of the RDF framework to integrate heterogenous experimental data of a large chemo- and biodiverse collection from a research collaborative project

Frederic Burdet[1], Pierre-Marie Allard[2,3,5], Louis-Felix Nothias[2,3,4], Olivier Kirchhoffer[2,3], Arnaud Gaudry[2,3], Sébastien Moretti[1], Robin Engler[1], Luis-Manuel Quiros-Guerrero[2,3], Emerson Ferreira Queiroz[2,3], Jahn Nitschke[7], Nabil Hanna[7], Chunyan Wu[8], Antonio Grondin[6], Bruno David[6], Thierry Soldati[7], Christian Wolfrum[8], Erick Carreira[9], Jean-Luc Wolfender[2,3], Marco Pagni[1], Florence Mehl[1]

[1] Vital-IT, SIB Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
[2] Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, 1211 Geneva 4, Switzerland
[3] School of Pharmaceutical Sciences, University of Geneva, 1211 Geneva 4, Switzerland
[4] Université Côte d'Azur, Institut de Chimie de Nice, Campus Valrose, Nice, France
[5] Department of Biology, University of Fribourg, 1700 Fribourg, Switzerland
[6] Green Mission Pierre Fabre, Institut de Recherche Pierre Fabre, 31562 Toulouse, France
[7] Department of Biochemistry, Faculty of Science, University of Geneva, 1211 Geneva, Switzerland
[8] Department of Health Sciences and Technology, ETHZ, 8092 Zürich, Switzerland
[9] Department of Chemistry and Applied Biosciences, ETHZ, 8093 Zürich, Switzerland

Abstract

Plants have a complex chemo-diversity and represent a reservoir of potential new therapeutic agents. Within a Swiss research project, six scientific research groups from different disciplines are collaborating to investigate a collection of more than 17'000 unique dried plant extracts. It aims to find new bioactive molecules and their modes of action, with for example anti-infective or pro-metabolic activities.

One of the main challenges of this enterprise is the management, integration and sharing of the highly heterogeneous data that are produced by the different research groups. Among these we find (i) massive high-resolution mass spectrometry data, (ii) the numerical results of innovative chemo-informatics methods, (iii) bioassay results from experimental models of tuberculosis and obesity, and (iv) organic synthetic chemistry. Additionally, requirements for data management plan and open-source science with the FAIR principles must be met.

We have established an agile pipeline to capture and structure this heterogeneous data into an RDF graph. The data content's gradual expansion and evolution throughout the project presented considerable challenges, particularly in terms of data modeling. Additionally, despite many collaborators not being RDF experts, most were technically adept at producing RDF triples relevant to their contributions.

We have deployed multiple instances of a triplestore and developed an in-house custom tool (*i.e.* kgsteward) to synchronize their content, based on a configuration file, which is centrally managed and version-controlled using Git. This strategy gave us the flexibility required to address global project challenges in common data management effectively.

Keywords

Plant extracts, RDF framework, Heterogeneous experimental data, collaborative research project, Bioactive molecules, FAIR principles

Introduction

Plants have a complex chemo-diversity and represent a reservoir of potential new therapeutic agents. Within a "Sinergia" research project supported by the Swiss National Science Foundation, six research groups from different disciplines are collaborating to investigate a collection of more than 17'000 unique dried plant extracts. It aims to find new bioactive molecules and their modes of action, with for example anti-infective or pro-metabolic activities.

One of the main challenges of this enterprise is the management, integration and sharing of the highly heterogeneous data that are produced by the different research groups. Additionally, requirements for the data management plan (DMP) and open-source science with the FAIR principles (findable, accessible, interoperable, reusable) must be met. We decided to use Resource Description Framework (RDF) and SPARQL to progressively assemble a knowledge graph (KG) containing relevant data and metadata produced by all collaborators. This paper summarizes our experiences in this still ongoing project.

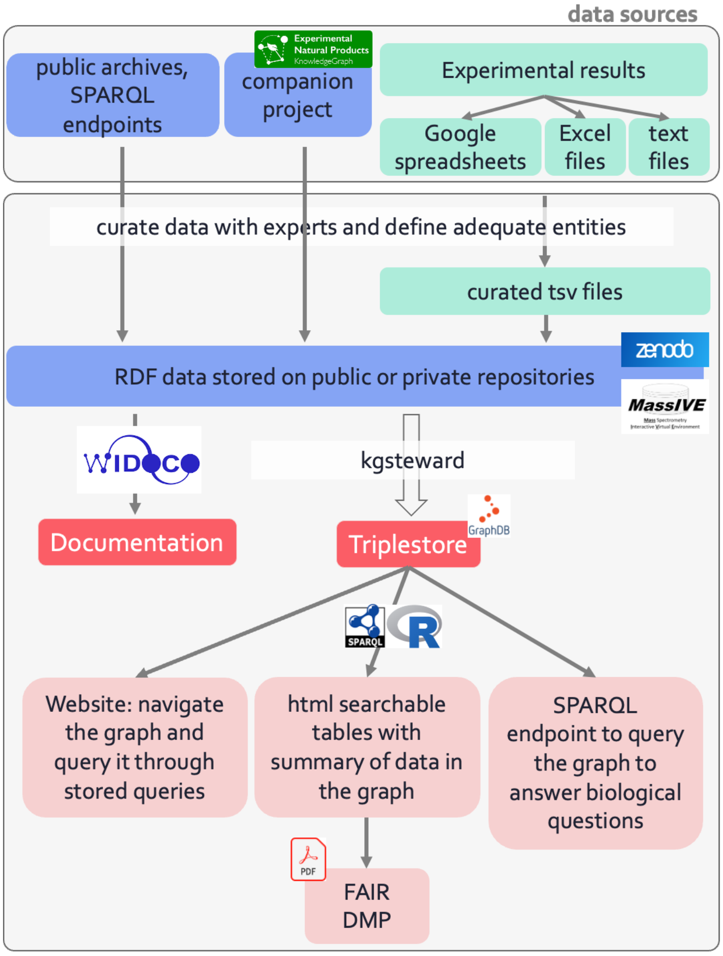
data sources
public archives, SPARQL endpoints
Experimental Natural Products KnowledgeGraph
companion project
Experimental results
Google spreadsheets
Excel files
text files
curate data with experts and define adequate entities
curated tsv files
zenodo
RDF data stored on public or private repositories
MassIVE
WIDOCO
kgsteward
Documentation
Triplestore
GraphDB
SPARQL
R
Website: navigate the graph and query it through stored queries
html searchable tables with summary of data in the graph
SPARQL endpoint to query the graph to answer biological questions
PDF
FAIR DMP

Figure 1: Workflow of integration of heterogenous data in a knowledge graph to manage, share and access data in a collaborative research project.

Methodology and Results

Figure 1 shows the general workflow used. A private Git [1] repository was established for sharing scripts, some (small) datasets, and the website of the project. Different *ad hoc* scripts are written to handle the diverse data formats. They are developed in Python, R/Rmd, Perl, and Bash, owing to the different skills of the contributors, and are not necessarily meant to be reused elsewhere.

We employed a combination of vocabularies including RDF, RDFS, OWL, XSD, SKOS, DCAT, VoID, PROV-O, FOAF, and WD, to extensively describe all useful data and metadata. We attempted to follow existing recommendations when available, e.g. [2], although a lack thereof seems to prevail in our domains of interest. Hence, we are developing our own vocabulary, prefixed “jlw:”, to encompass all the unique concepts encountered. In addition, we use the “enpkg:” prefix to denote entities from our 'companion project' ENPKG [3], which provides a first public release of the plant/chemistry data subset. Where appropriate, these jlw: and enpkg: entities are linked to their counterparts using the OWL property owl:SameAs for strict equivalence matching. This strategy allowed us to coordinate the two projects while preserving the necessary freedom to operate. A restricted-access website was set up, containing links to the SPARQL endpoint, vocabularies descriptions generated by Widoco [4] and summaries of the data as html tables.

Table 1
List of datasets integrated in the knowledge graph (not exhaustive).

Dataset	Description	Data workflow
CW - bioassays	Cytotoxicity test and lipid droplets analyses (ETHZ)	Imported with R from multiple Excel files; still unpublished
Inventa	*In-silico* "novelty score" from untargeted mass spectrometry data, spectral annotation, and literature reports (UniGE) [5]	csv file imported with R from MassIVE; public
TPH - bioassays	Cytotoxicity and bioactivity tests on parasites (SwissTPH)	Data still private imported with R from multiple Excel files
TS - bioassays	Anti-mycobacteria growth inhibition tests (UniGE)	Data still private imported with R from multiple Excel files
MZmine LC-MS^2 data	Aligned LC-MS data from 1600 plants extracts [6]	csv file imported with Perl from MassIVE; public
LOTUS	One of the biggest and best annotated resources for natural products occurrences available free of charge and without any restriction [7]	Freeze of Wikidata available from Zenodo; public
TAXO	A simplified and balanced taxonomy of plants	Data still private recompiled from Open Tree of Life taxonomy and Wikidata with Perl and SPARQL queries
ENPKG	Experimental Natural Products Knowledge Graph of the 1600 plant extracts [3]	RDF imported directly from Zenodo; public

The datasets described in Table 1 are integrated into the KG as RDF graphs, i.e. “context” in GraphDB/rdf4j terminology. To keep track of experimental workflows, the laboratory results have been organized around a central concept: the “analysis run”, representing the data captured in the same laboratory, at around the same time, by the same operator. The PROV Ontology [8] was used to encode the provenance of the laboratory data (protocols, operators, date of experiment, ...), as well as the persons involved in the project. The analysis run has several associated properties like the links to raw data files in public archive (using DCAT), the operator (linked by PROV-O), the protocol, and articles that were used as protocol, as well as the published articles reporting an analysis run.

As people contributing to this project were not RDF experts, they were educated on the way of producing their own RDF data. All predicates and relationships between entities were defined with them based on their domain expertise and their expectations to answer scientific questions using the KG. The SIB's role is central for maintaining the scheme *up-to-date* and coherent through the project.

To store and handle the KG, we decided to start with the free version of GraphDB [9]. The primary reason for this choice being the ease for novice users to install it on their working station. To populate the KG and manage its content with respect to many data sources, we developed a command line tool called kgsteward [10] that relies on a configuration file, currently in YAML format. This configuration file is shared through Git and permits the different users to host their own local instance of the triplestore, allowing each one to improve the schema/data at their own pace. This decentralized working approach was possible so far, as the KG remains relatively modest in size (~200 millions of triples). Additionally, kgsteward proficiently utilizes SPARQL updates to amend RDF imported from external sources and also features tools for quality control. A comprehensive DMP is required for the project. To facilitate its compilation, summaries including collections of samples, analysis runs and their metadata and number of measurements are retrieved via SPARQL queries from a Rmd script. The generated html report is put on the project website, and also provides an entry point for data retrieval. Dedicated graphical interfaces are being developed using visNetwork [11].

Discussion

The aim in this work is to capture in a KG, the data and concepts progressively produced by the researchers during a collaborative research project. One of the challenges is the continuous integration of heterogeneous unstructured and structured data from various sources, both public and private. Ensuring the FAIR management and sharing of data at the end of the project is another important challenge. In a sense, the purpose of our system sits in-between a Laboratory Information Management System (LIMS) and a static repository of project results, with the great additional features the KG can be queried to answer new scientific questions.

RDF/SPARQL proved to be robust from a technological point of view. The self-allowed flexibility in defining new vocabulary terms is required for the description of the incoming data was a great facilitator of data integration. This contrasts with other research data integration projects run at SIB with classical SQL database, which tend to be hampered by the rigidity of pre-defined table structures. Nevertheless, high flexibility comes at risk of dispersion and overfitting, which need to be counterbalanced by communication among all collaborators making sure that the progressively gathered knowledge is properly used to model the data. It cannot be overstated that interpersonal dynamics play a central role in the success of such an operation. This approach not only aids in accessing and sharing data but also facilitates seamless integration with other projects, exemplified by the ENPKG companion project [3].

The use of kgsteward, the data organization in many RDF subgraphs and the concept of analysis runs will facilitate the progressive release of the data in the public domain at the same time as the publications of the scientific results. The imperative for scalability and interoperability in environmental omics data speaks in favor of software tools which produce their output directly in RDF format [3].

Acknowledging the inherent risks of flexibility and simultaneous work on structure and data addition, we are actively developing queries and processes to automatically flag and correct inconsistencies. In the context of such a collaborative research project involving individuals from diverse scientific domains and varying levels of programming experience, working within the RDF framework has proven well-suited to achieve our objectives.

Acknowledgements

The authors are grateful to Green Mission Pierre Fabre, Pierre Fabre Research Institute, Toulouse, France, for establishing and sharing this unique library of extracts. The authors thank the Swiss National Science Foundation for received support the project (SNF N° CRSII5_189921/1).

References

[1] S. Chacon, B. Straub, Pro git, 2014. https://git-scm.com/.

[2] M. Dumontier, A.J.G. Gray, M.S. Marshall, V. Alexiev, P. Ansell, G. Bader, J. Baran, J.T. Bolleman, A. Callahan, J. Cruz-Toledo, P. Gaudet, E.A. Gombocz, A.N. Gonzalez-Beltran, P. Groth, M. Haendel, M. Ito, S. Jupp, N. Juty, T. Katayama, N. Kobayashi, K. Krishnaswami, C. Laibe, N. Le Novère, S. Lin, J. Malone, M. Miller, C.J. Mungall, L. Rietveld, S.M. Wimalaratne, A. Yamaguchi, The health care and life sciences community profile for dataset descriptions, PeerJ. 4 (2016) e2331. doi:10.7717/peerj.2331.

[3] A. Gaudry, M. Pagni, F. Mehl, S. Moretti, L.-M. Quiros-Guerrero, A. Rutz, M. Kaiser, L. Marcourt, E. Ferreira Queiroz, J.-R. Ioset, A. Grondin, B. David, J.-L. Wolfender, P.-M. Allard, A Sample-Centric and Knowledge-Driven Computational Framework for Natural Products Drug Discovery, Chemistry, 2023. doi:10.26434/chemrxiv-2023-sljbt.

[4] D. Garijo, WIDOCO: a wizard for documenting ontologies, içinde: International Semantic Web Conference, Springer, Cham, 2017: ss. 94-102. doi:10.1007/978-3-319-68204-4_9.

[5] L.-M. Quiros-Guerrero, L.-F. Nothias, A. Gaudry, L. Marcourt, P.-M. Allard, A. Rutz, B. David, E.F. Queiroz, J.-L. Wolfender, Inventa: A computational tool to discover structural novelty in natural extracts libraries, Frontiers in Molecular Biosciences. 9 (2022) 1028334. doi:10.3389/fmolb.2022.1028334.

[6] P.-M. Allard, A. Gaudry, L.-M. Quirós-Guerrero, A. Rutz, M. Dounoue-Kubo, T.W.N. Walker, E. Defossez, C. Long, A. Grondin, B. David, J.-L. Wolfender, Open and reusable annotated mass spectrometry dataset of a chemodiverse collection of 1,600 plant extracts, GigaScience. 12 (2022) giac124. doi:10.1093/gigascience/giac124.

[7] A. Rutz, M. Sorokina, J. Galgonek, D. Mietchen, E. Willighagen, A. Gaudry, J.G. Graham, R. Stephan, R. Page, J. Vondrášek, C. Steinbeck, G.F. Pauli, J.-L. Wolfender, J. Bisson, P.-M. Allard, The LOTUS initiative for open knowledge management in natural products research, ELife. 11 (2022) e70780. doi:10.7554/eLife.70780.

[8] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, J. Zhao, PROV-O: The PROV Ontology, World Wide Web Consortium, United States, 2013.

[9] Ontotext, GraphDB, 2023. https://www.ontotext.com/products/graphdb/.

[10] M. Pagni, KGSteward, 2023. https://github.com/sib-swiss/kgsteward.

[11] B.D. Almende, visNetwork, 2022. https://datastorm-open.github.io/visNetwork/.