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

Frederic Burdet[1], Luis-Manuel Quiros-Guerrero[2,3], Olivier Kirchhoffer[2,3], Jahn Nitschke[7], Pierre-Marie Allard[2,3,5], Louis-Felix Nothias[2,3,4], Arnaud Gaudry[2,3], Sébastien Moretti[1], Robin Engler[1], Emerson Ferreira Queiroz [2,3], 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] and 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*

1. Summary of the project

Plants possess an intricate chemo-diversity, serving as a rich reservoir for the discovery of potential therapeutic agents. In the framework of a Swiss research initiative called "Sinergia", six research groups from different disciplines collaborate to explore the potential of more than 17,000 unique dried plant extracts, to uncover novel bioactive molecules. To do so, heterogenous data from the different research groups was integrated in a Knowledge Graph (KG). This includes (i) highresolution mass spectrometry data, (ii) taxonomical information, (iii) chemo-informatics results, (iv) bioassay outcomes for tuberculosis, obesity, anticancer and antiviral models, as well as (v) data from organic synthetic chemistry.

We wanted to create a comprehensive framework to facilitate the alignment of multimodal data across chemical structures, biological activities, spectral features, and taxonomy, among others. At its core, it would include experimental data that is processed at the sample level, harmonized with external identifiers whenever feasible, semantically enriched and integrated into the KG.

The KG allows to enhance efficient data mining capabilities to address different scientific research questions related to the specific objectives of each participating group. For example, "Which bioactive molecules annotated in the extract of 'plant A' are present in other species of the same genus?", "Which taxonomic group presents a remarkable concentration of bioactive molecules for a specific assay?", or "Which non-toxic active extracts (for a specific assay) present a high probability of containing new molecules?".

The modelling of data from chemical analysis became a subproject in itself: ENPKG [1]. The recently published paper describing it exemplifies how this KG can be used to answer questions centered around the chemistry of incompletely described natural products.

The heterogeneity, expansion, and evolution of the data throughout time posed a major challenge for the project, centered on managing, integrating, modeling, and effectively sharing data generated by the different groups. Also, compliance with a data management plan and adherence to open-source science

SWAT4HCLS 2024: The 15th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences, February 26–29, 2024, Leiden, The Netherlands

principles, particularly the FAIR (Findable, Accessible, Interoperable, Reusable) principles, had to be ensured.

To overcome these challenges and handle this intricate data, a custom tool called KGSteward [2] was developed. This tool enables content synchronization based on a centrally managed versioncontrolled configuration file using Git [3]. This strategic approach provides the flexibility required to address the global project challenges in effectively managing shared data.

For managing the KG, the free version of GraphDB [4] was chosen, to which KGSteward is connected. KGSteward updates the KG through the Unix command line and relies on a YAML configuration file shared through Git, enabling users to host their local instances of the triplestore. The primary functions performed by KGSteward include: (i) creating the repository, (ii) uploading the TTL files listed in the configuration file, and (iii) executing UPDATE queries to clean and harmonize the KG.

Table 1 provides a snapshot of the different datasets integrated into the KG, emphasizing their origin. These datasets, presented as RDF graphs, are organized around the central concept of an "analysis run". This documents data executed by the same operator within the same laboratory on a specific date. Should the operator repeat the same assay on another date, it is regarded as a separate analysis run to ensure comprehensive traceability of integrated data.

Provenance information is encoded using the PROV Ontology [5], linking to raw data files, operators, protocols, and associated articles. This ensures the ‘Findable and Reusable’ part of the FAIRification. Additionally, for Interoperability, vocabularies like RDF, RDFS, OWL, are employed to describe data and metadata extensively, as described in [6]. We also created a customized vocabulary for project concepts that couldn’t be mapped on existing vocabularies. It is planned to progressively release the data in the public domain at the same time as the scientific results’ publications. Currently the KG contains about 200 million triples.

The use of RDF/SPARQL in our project has proven to be highly robust, particularly in handling the dynamic nature of our evolving datasets. These semantic technologies offer a notable level of flexibility in defining and adapting vocabularies according to our project’s specific requirements.

However, the delicate balance between flexibility and usability was a central concern during the deployment of our KG. The KG structure was modelled from a starting set of data, but due to the quick evolution of the cumulative heterogenous sets it was modified drastically for a better overall fit. The implementation of KGSteward in the evolution process was instrumental, as it bypasses the graphical interface, automatically detects the modified input TTL files, and does selective updates.

This strategic approach reflects the project’s dedication to effectively capture, organize, and adapt to the continuously generated data and concepts. The integration of RDF/SPARQL aligns with our project goals and reflects our commitment to navigate the complexities of contemporary data management challenges in interdisciplinary research endeavors.

Table 1
List of datasets integrated in the knowledge graph and their origin (non-exhaustive, as the project is still ongoing)

Dataset	Analysis description	RDF data workflow
CW - bioassays	Cytotoxicity test and lipid droplets analyses (ETHZ)	Experimental results were imported with R from multiple excel files, quality controlled and converted to TTL.
TPH - bioassays	Cytotoxicity and bioactivity tests on parasites (SwissTPH)	Same as above.
TS - bioassays	Anti-mycobacteria growth inhibition tests (UniGE)	Same as above.
Inventa – *in-silico* analysis	"Novelty score” from untargeted mass spectrometry data, spectral annotation, and literature reports (UniGE) [7]	*In-silico* results reformatted with R.
MZmine LC-MS2 data – *in-silico* analysis	Aligned LC-MS data from 1600 plants extracts [8]	Experimental results processed at MassIVE, reimported and converted to TTL with a Perl script.
LOTUS – Public database	One of the biggest and best annotated resources for natural products occurrences available free of charge and without any restriction [9]	Freeze of Wikidata available from Zenodo; public. SPARQL update was used to reshape and insert the data into the KG.
TAXO – Public database	A simplified and balanced taxonomy of plants	Data recompiled from Open Tree of Life taxonomy and Wikidata with Perl and SPARQL updates.
ENPKG – Public database	Experimental Natural Products Knowledge Graph of the 1600 plant extracts [1]	RDF imported directly from Zenodo.

2. 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 for the project (SNF N° CRSII5_189921/1).

3. References

References

[1] 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.

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

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

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.