A multi-modal and temporal antibiotic resistance knowledge graph

Brieuc Quemeneur[1,2], Audrey Bihouée[1], Samuel Chaffron[3], Claudine Médigue[2,4], Hervé Ménager[2,5] and Alban Gaignard[1,2]

[1]Nantes Université, CNRS, INSERM, l'institut du thorax, F-44000 Nantes, France

[2]IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 91057 Evry, France

[3]Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

[4]CNRS UMR8030, Université Evry-Val-d'Essone, CEA, Genoscope, LABGeM, 91000, Evry, France

[5]Institut Pasteur, Université Paris Cité, Bioinformatics of Biostatistics Hub, F-75015 Paris, France

Abstract

Understanding how antibiotic resistance genes spread is essential for protecting human, animal, and environmental health. It requires collaboration across multiple fields and expertise under One Health initiatives, emphasizing the pressing need to consolidate diverse antibiotic data from human, animal, and environmental samples. In this paper, we propose a domain-specific Knowledge Graph leveraging the SOSA ontology to uniformly represent multi-modal data and their analysis while allowing the description of provenance metadata covering both time and geographical locations. This work is driven by a national consortium of antibiotic resistance experts (ABRomics). As experimental results, we show how this domain knowledge can be used to answer a specific expert question as well as increasing the FAIRness of antibiotic resistance data.

Keywords

Knowledge graphs, Ontologies, Antibiotic Resistance, SOSA

1. Introduction

Antibiotic microbial resistance (AMR) arises from the extensive use of antibiotics in agri-food systems, human and veterinary medicine, leading to resistant bacteria: it is one of the biggest threats to human health, significantly undermining the effectiveness of existing treatments for bacterial infections. This issue is exacerbated by the evolutionary capabilities of bacteria that have developed complex resistance mechanisms over time in response to different biological and environmental factors. Indeed, AMR requires a comprehensive response through a One Health approach which integrates human, animal, and environmental ecosystems, addressing the dynamics of antibiotic resistance comprehensively [1]. Consequently, answering research questions about this issue calls for the integration of multiple knowledge and data resources [2].

The French national bioinformatics infrastructure, with a large inter-disciplinary consortium, is coordinating the development of ABRomics [1], a national platform that aims at gathering, analysing, and sharing multi-modal ABR data. This platform aims at facilitating ABR experts to answer typical

SWAT4HCLS 2025: 16th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences 2025
brieuc.quemeneur@univ-nantes.fr (B. Quemeneur)
0009-0006-7574-2254 (B. Quemeneur); 0000-0002-8689-2083 (A. Bihouée); 0000-0001-5903-617X (S. Chaffron); 0000-0002-3905-1054 (C. Médigue); 0000-0002-7552-1009 (H. Ménager); 0000-0002-3597-8557 (A. Gaignard)

[1]https://abromics.fr

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-04-22

questions such as "What are the most represented antibiotic resistance genes in a specific geographical region of interest ?" (CQ). Scientists face the challenge of integrating heterogeneous and possibly massive data covering both the precise description of biological samples, their location, the context of their acquisition, as well as the description of the results of their analysis though bioinformatics pipelines.

The open science movement, with the recent adoption of the FAIR principles led to the creation of public, interoperable Life Science knowledge graphs, empowering data scientists to query and integrate large, diverse, and decentralized knowledge repositories [3, 4, 5, 6]. One of the benefits of these public knowledge graphs lies in the standardization effort and the reuse of their schemas and ontologies. Some of them can be domain-specific, such as ARO [7], aimed at covering the field of ABR, and other ones can be more generic and reused in many application scenarios. For instance, the SOSA ontology has been introduced to facilitate the exchange of data within a sensor network. It provides a structured vocabulary for representing measures performed by sensors on specific objects of interest. SOSA is generic enough to be used in different domains such as in geosciences [8].

In this work, we investigate the following research question: *can a knowledge graph based on both the ARO and SOSA ontologies help biologists answer the ABR question CQ ?* Our main contribution is a Knowledge Graph allowing to represent, store and share FAIR bio-sample metadata as well as the results of a resistance gene detection workflow.

The paper is organized as follows. Section 2 introduces how we used the SOSA and ARO ontologies to represent multi-modal observations with time and geo-location metadata. Section 3 demonstrates how this ABR KG can be used to answer typical questions related to antibiotic resistance. Finally, section 4 summarizes our approach and proposes future works.

2. Representing multi-modal observations with SOSA

ABR research requires to sequence and analyze genetic material of multiple bacteria from biological samples. These analysis are generally performed through computationally intensive bioinformatics workflows, which process data to evaluate measurements on typical features of interest. These processes semantically align with the concepts and relations covered by the SOSA ontology, initially proposed for the semantic interoperability of sensor network data.

Representing bio-samples and their provenance metadata. Figure 1 illustrates the modeling choices made to semantically describe biological samples. They are represented as instances of the *sosa:FeatureOfInterest* class, so that each one can be an observable entity in the context of SOSA. For interoperability purpose, we also use the *prov:Entity* and *sio:Sample* classes from the PROV-O [9] and SIO [10] ontologies respectively. Reference Life Science ontologies such as NCBITAXON or NCIT were also used to link samples to their host species, microorganisms or source. Since biological samples are defined as *prov:Entity*, we use the provenance *prov:atLocation* property to refer to its country WikiData identifier. The properties *prov:wasGeneratedBy* and *prov:wasAttributedTo* allow to go even further by adding the *sosa:Sensor* that generated the data and some *foaf:Person* associated to a sample. A timestamp corresponding to the sample collection is accessible through the *prov:generatedAtTime* property.

Representing and linking data analysis results. ABR Studies generally need for biological samples to be sequenced, and their resulting large-scale omic data to be analysed through complex bioinformatics workflows. Figure 2 shows how we use the *sosa:Observation* concept to represent each measure performed by a data analysis workflow. Then, we rely on the *sosa:ObservableProperty* and the *sosa:FeatureOfInterest* properties to integrate multi-modal data. Multiple observation modalities, such as metagenomics, can be associated to the same biological sample (*sio:Sample*) through instances of *sosa:ObservableProperty*. Multiple *sosa:FeatureOfInterest* can be provided to indicate what is being

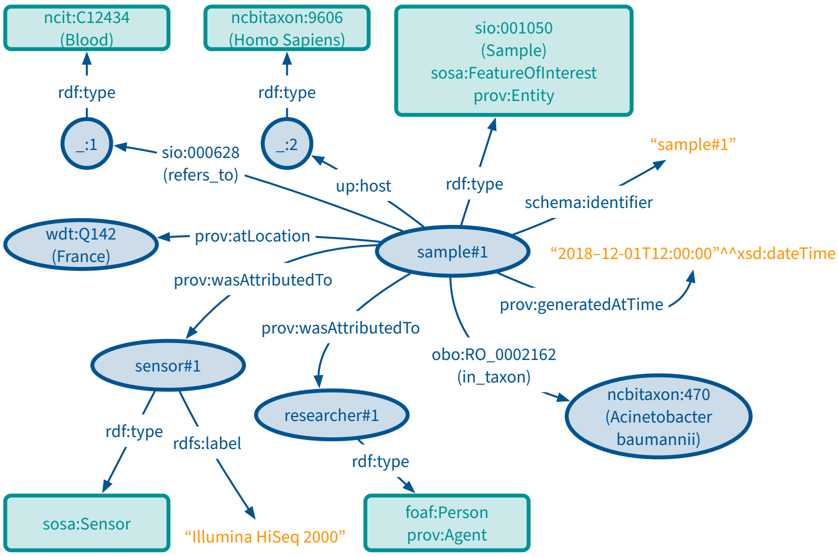

Figure 1: Example RDF instances for the sample metadata.

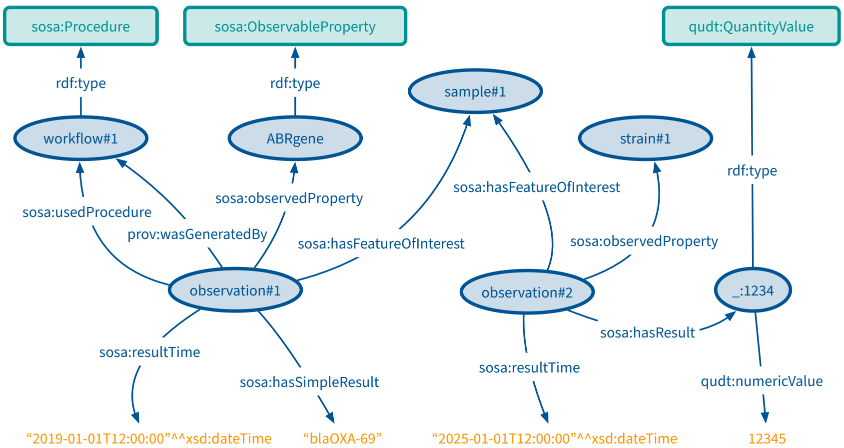

Figure 2: Example RDF instances for the data analysis results.

observed, such as antibiotic resistance genes (*go:Gene*) or a bacterial strain (*aro:Strain*). Finally we associate to each observation the data analysis pipeline (*sosa:Procedure*) through the *sosa:usedProcedure* and *prov:wasGeneratedBy* provenance properties.

3. Querying ABR-KG for antibiotic resistance research

Based on the ontologies described above, we have created a knowledge graph from public data on *Acinetobacter baumannii* strains used in an analysis of natural transformation in bacterial species [11]. The genomic sequences and metadata of 40 *A. baumannii* strains of human, animal and environmental origins have been integrated and processed into the ABRomics platform. The resulting 120 analysis reports gather sample metadata as well as antibiotic resistance genes detected with the ABRomics bioinformatics workflows.

CQ: **What are the most represented antibiotic resistance genes in a specific geographical region of interest ?**

Table 1
Used ontology classes

Class	Instances	Description
sosa:Observation	2519	the observations made by the workflows
sosa:FeatureOfInterest	241	the observed entities (genes, strains, samples)
aro:Strain	121	the detected bacterial strains
prov:Entity	120	the samples
sio:001050	120	the samples
sosa:ObservableProperty	84	the measured properties
go:Gene	73	the detected genes in the samples
sosa:Sensor	2	the used genomic sequencers
prov:Agent	1	the associated researcher
prov:Person	1	the associated researcher

Figure 3 shows the typical graph patterns that can be used to query the ABR knowledge graph. Lines 2-4 filter all biological samples and their identifiers. Then, lines 5-9 retrieve all linked instances of *sosa:Observation* looking for resistance genes (*abromics:ABRGene*). The result of each observation is a gene name that is counted (SELECT clause). Finally lines 12-17 shows how to leverage the Wikidata public knowledge graph to filter samples collected in France. More technical details are available through our GitHub repository [2].

```
1  SELECT ?gene_name ?location_name (COUNT(?gene_name) as ?count) WHERE {
2     ?sample rdf:type sio:001050 ;
3              schema:identifier ?sample_id ;
4              prov:atLocation ?location .
5      ?observations rdf:type sosa:Observation ;
6                    sosa:observedProperty <abromics:ABRgene> ;
7          sosa:hasFeatureOfInterest ?sample ;
8             sosa:hasSimpleResult ?gene_name .
9     ?gene rdf:type go:Gene ;
10             rdfs:label ?gene_name .
11     # fetch the id corresponding to the targeted location
12     SERVICE <https://query.wikidata.org/sparql> {
13             ?location wdt:P31 wd:Q6256 .
14             ?location rdfs:label ?location_name .
15             FILTER(?location_name = "France"@en)
16     }
17 }
18 GROUP BY ?gene_name ?location_name
19 ORDER BY DESC(?count)
```

Figure 3: A federated SPARQL query to identify resistance gene for a specific location.

4. Conclusion and future works

In this paper we show how Knowledge Graphs can be useful in the domain of antibiotic resistance, and more generally in One Health approaches. Our main findings are that SOSA is particularly suited for describing rich sample metadata, as well as the results of bioinformatics analysis. In addition, SOSA allows to easily combine multi-modal observations for biological samples collected at multiple time-points. More generally, adopting these semantic models is of particular interest towards FAIR sharing of ABR data. In future works, we plan to integrate GeoSPARQL to facilitate the definition of regions of interests (ROIs) in the queries with GPS coordinates associated, for example, to water or soil samples. More generally, by integrating dedicated domain ontologies such as EnvO [12], we plan to not only support human health, but a broader environmental community through ABRomics.

[2] https://github.com/Phloemus/ABRomics-KG

Acknowledgments

This work is financially supported by the French Priority Research Programme on Antimicrobial Resistance (PPR antibioresistance), coordinated by Inserm and funded by the Secretaire General Pour L'investissement (SGPI) and by the French government grant by the Agence Nationale de la Recherche under France 2030 for structuring research facilities / EQUIPEX+, reference ANR-21-ESRE-0048. Part of this work was carried out within the IBD-NExT cluster and benefited from the "France 2030" plan and financial support from the Pays de la Loire Region and Nantes Métropole.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] A. O. Ajayi, A. T. Odeyemi, O. J. Akinjogunla, A. B. Adeyeye, I. Ayo-Ajayi, Review of antibiotic-resistant bacteria and antibiotic resistance genes within the one health framework, Infection Ecology & Epidemiology 14 (2024) 2312953.

[2] J. Youn, N. Rai, I. Tagkopoulos, Knowledge integration and decision support for accelerated discovery of antibiotic resistance genes, Nature communications 13 (2022) 2360.

[3] S. Jupp, J. Malone, J. Bolleman, M. Brandizi, M. Davies, L. Garcia, A. Gaulton, S. Gehant, C. Laibe, N. Redaschi, S. M. Wimalaratne, M. Martin, N. Le Novère, H. Parkinson, E. Birney, A. M. Jenkinson, The ebi rdf platform: linked open data for the life sciences, Bioinformatics 30 (2014) 1338–1339. doi:10.1093/bioinformatics/btt765.

[4] Y. Yamamoto, A. Yamaguchi, A. Splendiani, Yummydata: providing high-quality open life science data, Database: The Journal of Biological Databases and Curation 2018 (2018).

[5] M. R. Kamdar, M. A. Musen, An empirical meta-analysis of the life sciences linked open data on the web, Scientific Data 8 (2021) 24. URL: https://www.nature.com/articles/s41597-021-00797-y. doi:10.1038/s41597-021-00797-y, publisher: Nature Publishing Group.

[6] S. S. I. of Bioinformatics RDF Group Members, The SIB Swiss Institute of Bioinformatics Semantic Web of data, Nucleic Acids Research 52 (2023) D44–D51. doi:10.1093/nar/gkad902.

[7] B. Jia, A. R. Raphenya, B. Alcock, N. Waglechner, P. Guo, K. K. Tsang, B. A. Lago, B. M. Dave, S. Pereira, A. N. Sharma, et al., Card 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database, Nucleic acids research (2016) gkw1004.

[8] R. Zhu, S. Stephen, L. Zhou, C. Shimizu, L. Cai, G. Mai, K. Janowicz, P. Hitzler, M. Schildhauer, Environmental observations in knowledge graphs., in: DaMaLOS, 2021, pp. 1–11.

[9] T. Lebo, S. S. Sahoo, D. L. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, J. Zhao, Prov-o: The prov ontology, 2013.

[10] M. Dumontier, C. J. O. Baker, J. Baran, A. Callahan, L. L. Chepelev, J. Cruz-Toledo, N. R. D. Rio, G. Duck, L. I. Furlong, N. Keath, D. Klassen, J. McCusker, N. Queralt-Rosinach, M. Samwald, N. Villanueva-Rosales, M. D. Wilkinson, R. Hoehndorf, The semanticscience integrated ontology (sio) for biomedical research and knowledge discovery, Journal of Biomedical Semantics 5 (2014).

[11] F. Mazzamurro, J. B. Chirakadavil, I. Durieux, L. Poiré, J. Plantade, C. Ginevra, S. Jarraud, G. Wilharm, X. Charpentier, E. PC Rocha, Intragenomic conflicts with plasmids and chromosomal mobile genetic elements drive the evolution of natural transformation within species, PLoS biology 22 (2024) e3002814.

[12] P. L. Buttigieg, E. Pafilis, S. E. Lewis, M. P. Schildhauer, R. L. Walls, C. J. Mungall, The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation, Journal of biomedical semantics 7 (2016) 1–12.