ELIXIR biovalidator for semantic validation of life science metadata

Isuru Liyanage\(^1\), Tony Burdett\(^1\), Bert Drosbeke\(^2,3\), Karoly Erdos\(^1\), Rolando Fernandez\(^1\), Alasdair Gray\(^4\), Muhammad Haseeb\(^1\), Flavia Penim\(^1\), Cyril Pommier\(^5,6\), Philippe Rocca-Serra\(^7\), Mélanie Courtot\(^*\), and Frederik Coppens\(^2,3\)

\(^1\)European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, CB10 1SD, UK
\(^2\)Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 71, Ghent, 9052 Belgium
\(^3\)VIB Center for Plant Systems Biology, Technologiepark 71, Ghent, 9052 Belgium
\(^4\)Department of Computer Science, Heriot-Watt University, Edinburgh, UK
\(^5\)Université Paris-Saclay, INRAE, BioinfOmics, Plant bioinformatics facility, 78026 Versailles, France
\(^6\)Université Paris-Saclay, INRAE, URGI, 78026, Versailles, France
\(^7\)University of Oxford e-Research Centre, Department of Engineering Science, University of Oxford, 7 Keble Road, OX1 3QG, Oxford.

October 28, 2021

Abstract

To advance biomedical research, increasingly large amounts of complex data need to be discovered and integrated. This requires syntactic and semantic validation to ensure shared understanding of the entities we describe. This article describes the ELIXIR biovalidator, which extends the syntactic validation of the widely used AJV library with ontology-based validation of JSON documents.

Availability and implementation: https://github.com/elixir-europe/biovalidator, Release: v1.9.1, License: Apache License 2.0

I. Introduction

Today’s genomics data ecosystem has been described as a ”Tower of Babel”, due to an ever-increasing amount of data generated, using different technologies, in a widening number of domains, hosted in a constantly growing number of databases. This massive diversification makes data science an extremely labour intensive and thus a costly undertaking \([13,4]\). Data FAIRification \([17]\) aims at addressing those challenges by promoting adherence to a set of principles that facilitate data re-use and interoperability. Validation of metadata describing biomedical entities is a crucial part of this process. However, rules for validation are often hard coded in specific resources, and not shared efficiently. Moreover, checklists such as those used by archives \([6]\) can still lead to various interpretations and diverging implementations, resulting in data heterogeneity which

*To whom correspondence should be addressed: mcourtot@gmail.com*
prevents its efficient reuse [15]. Therefore, next to clear documentation of the practices, real-world implementations of tools enforcing shared validation processes are needed.

JavaScript Object Notation (JSON) is an IETF standard specifying a lightweight data interchange format [1]. JSON Schema is a vocabulary to specify the structure of a JSON document. Both JSON and JSON Schema are extensively used for data exchange, APIs and standard definitions. Whilst JSON Schema provides a comprehensive vocabulary to validate the structure and the syntax of a JSON document, it contributes little to checking semantics of the content. In life sciences, compliance to metadata schemas often mandates assessing if a value adheres to specified ontologies - for example check that the value of a ‘disease’ attribute is a subclass of a disease ontology term. To ensure high-quality metadata, such strict validation checks are required, specifically via queries based on the ontology structure itself. To address this, we have extended the JSON schema vocabulary with custom keywords that describe how a particular property constrained to an ontology term identifier should be validated. This paper describes how we deployed ELIXIR biovalidator and applied it to plant related use cases to enhance FAIRness of the data collected and submitted to public archives.

II. Implementation

We have developed the ELIXIR biovalidator, a tool for validating life sciences metadata, encoded as JSON documents, against declarative metadata standards that are encoded as JSON schema. The ELIXIR biovalidator is based on the widely used Ajv JSON Schema validator [12]. Through the addition of validation rules for user-defined keywords, we have augmented the validator with ontology-based constraints, such as `isValidTerm` to check if a given ontology term exists in the EMBL-EBI Ontology Lookup Service (OLS) [8]. At the time of writing, ELIXIR biovalidator supports four extended keywords for ontology and taxonomy validation [5]. For example, the extended keyword `graph_restriction`, used with a parent term ID and an ontology ID, allows us to express that a JSON property such as `disease_ontology_id` can only have terms that are from the Phenotype and Trait Ontology (PATO) [16] or Monarch Disease Ontology (MONDO) [10]. Furthermore, these terms must be a subclass of the disease classes `PATO:0000461` or `MONDO:0000001`.

The ELIXIR biovalidator is capable of running as a service or as a one-time script to validate a given JSON document against a schema [5]. When run as a service, users can validate using the web interface or an API, which is more suited for batch validations. A Docker image is available for testing in a local environment. The biovalidator is currently being used in EMBL-EBI BioSamples [3], where it ensures checklist compliance for over 18 million samples, and as part of the data ingest system for the Human Cell Atlas project.

III. Validation of Plant metadata

Plant research institutes across the globe have developed databases and tools to manage and store plant phenotyping data, tailored to their specific use cases. MIAPPE (Minimal Information about a Plant Phenotyping Experiment [11]) is an open, community driven metadata standard that adequately describes plant phenotyping experiments. The Breeding API (BrAPI [14]) was developed synergistically with MIAPPE to provide a common, programmatic interface ensuring databases and tools interoperability through the use of a common metadata representation; BrAPI is therefore a web service API implementation of MIAPPE. This standardized API enables the development of scripts that work on all plant phenotyping databases, as long as they have BrAPI implemented. One of these scripts is BrAPI2Biosamples which can be used to export JSON
ELIXIR biovalidator for semantic validation of life science metadata

Figure 1: Data validation for the plant use case. A data submitter uses an institutional data repository as a broker to submit Biosamples metadata through the API, which is validated against the MIAPPE JSON schema. This metadata from the plant phenotyping databases is exposed through the Breeding API (BrAPI) and formatted using the BrAPI2Biosamples script to JSON objects. These objects can be validated using the ELIXIR biovalidator against a MIAPPE JSON Schema checklist.

IV. Conclusion

The ELIXIR biovalidator is a tool for validating both the structure and the content of JSON documents by extending the existing JSON Schema syntax. The biovalidator is capable of validating ontology terms embedded in JSON instance documents against requirements. Enabling this quality control for community standards is crucial to develop semantic interoperability in a distributed ecosystem of FAIR digital objects, as envisioned in the European Open Science Cloud Interoperability Framework [2]. In the future, we plan to further extend the biovalidator by adding support for identifier cross-reference checking by integrating it with Identifiers.org [9]. This will enable the biovalidator to check the validity of accessions present in the JSON data.
ACKNOWLEDGEMENTS

The authors acknowledge Sarah Cohen Boulakia for initial contributions to the project and Guillaume Cornut and Jeremy Destin for contribution to the MIAPPE to Biosamples interoperability and Simon Jupp.

FUNDING

This work was funded by EMBL-EBI core funds, ELIXIR - the research infrastructure for life-science data and supported by Research Foundation - Flanders (FWO) for ELIXIR Belgium (I002819N).

REFERENCES

[1] Bray, T. et al. (2014). The javascript object notation (json) data interchange format.
[2] Corcho, O. et al. (2021). EOSC interoperability framework: Report from the EOSC Executive Board Working Groups FAIR and Architecture.
[3] Courtot, M. et al. (2019). Biosamples database: an updated sample metadata hub. Nucleic acids research, 47(D1):D1172–D1178.
[4] Directorate-General for Research and Innovation (European Commission), PwC EU Services (2019). Cost-benefit analysis for FAIR research data : cost of not having FAIR research data. Publications Office of the European Union.
[5] elixir-europe (2021). biovalidator/README.md at elixir-europe, biovalidator.
[6] Harrison, P. W. et al. (2021). The European Nucleotide Archive in 2020. Nucleic Acids Res., 49(D1):D82–D85.
[7] Johnson, D. et al. (2021). ISA API: An open platform for interoperable life science experimental metadata. GigaScience, 10(9):giab060.
[8] Jupp, S., Burdett, T., Leroy, C., and Parkinson, H. E. (2015). A new ontology lookup service at embl-ebi. SWAT4LS, 2:118–119.
[9] Juty, N. et al. (2012). Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. Nucleic Acids Res., 40(Database issue):D580–6.
[10] Mungall, C. J., McMurry, et al. (2017). The monarch initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic acids research, 45(D1):D712–D722.
[11] Papoutsoglou, E. A. et al. (2020). Enabling reusability of plant phenomic datasets with MIAPPE 1.1. New Phytologist, 227(1):260–273.
[12] Poberezkin, Evgeny (2021). Ajv - Another JSON Schema Validator. Accessed: 2021-10-25.
[13] Powell, K. (2021). The broken promise that undermines human genome research. Nature, 590(7845):198–201.
[14] Selby, P. et al. (2019). BrAPI—an application programming interface for plant breeding applications. Bioinformatics, 35(20):4147–4155.
[15] Silvester, N. et al. (2018). The European Nucleotide Archive in 2017. *Nucleic Acids Res.*, 46(D1).

[16] Tan, S., Mungall, C., and et al. (2021). pato-ontology/pato: 2021-09-09 release.

[17] Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, 3:160018.

[18] Wolstencroft, K. et al. (2015). SEEK: a systems biology data and model management platform.