BioHackrXiv

Data validation and schema interoperability

Leyla Garcia[1], Jerven Bolleman[2], Michel Dumontier[3], Simon Jupp[4], Jose Emilio Labra Gayo[5], Thomas Liener[6], Tazro Ohta[7], Núria Queralt-Rosinach[8], and Chunlei Wu[9]

1 ZB MED Information centre for life sciences, Gleueler Str. 60, 50931 Cologne, Germany **2** Swiss Institute of Bioinformatics, Quartier Sorge - Batiment Amphipole, 1015 Lausanne, Switzerland **3** Maastricht University, Minderbroedersberg 4-6, 6211 LK Maastricht, The Netherlands **4** European bioinformatics institute EMBL-EBI, Wellcome Genome Campus, CB10 1SD, Hinxton, United Kingdom **5** Universidad de Oviedo, C/Federico García Lorca, S/N, CP 33007, Oviedo, Spain **6** Thomas Liener Consultancy, www.linkedin.com/in/thomas-liener **7** Database Center for Life Science, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, Yata 1111, Mishima, Shizuoka, Japan **8** Harvard Medical School, Countway Library 10 Shattuck St, Boston, MA 02115, United States **9** The Scripps Research Institute, 10550 N Torrey Pines Rd, La Jolla, CA 92037, United States

BioHackathon series:
NBDC/DBCLS BioHackathon
Fukuoka, Japan, 2019
Schemas Working Group

Submitted: 07 Apr 2020

Published by BioHackrXiv.org

Background

Validating RDF data becomes necessary in order to ensure data compliance against the conceptualization model it follows, e.g., schema or ontology behind the data. Validation could also help with data consistency and completeness. There are different approaches to validate RDF data. For instance, JSON schema is particularly useful for data expressed in JSON-LD RDF serialization while Shape Expression (ShEx)(Baker & Prud'hommeaux, 2019) and Shapes Constraint Language (SHACL) (Knublauch & Kontokostas, 2017) can be used with other serialization as well. Currently, no validation approach is prevalent regarding others, depending on data characteristics and personal preferences one or the other can be used. In some cases, the approaches are interchangeable; however, that is not always the case, making it necessary to identify a subset among them that can be seamlessly translated from one to another.

During the NBDC/DBCLS BioHackathon 2019, we worked on a variety of topics related to RDF data validation, including (i) development of ShEx shapes for a number of datasets, (ii) development of a tool to semi-automatically create ShEx shapes, (iii) improvements to the RDFShape tool (Labra-Gayo, Fernández-Álvarez, & GArcía-González, 2018) and (iv) enabling validation schema conversion from one format to the other. In the following sections we detailed the work done on each front.

Hackathon results

Development of ShEx shapes

We created and updated ShEx shapes for different biomedical resources including Health Care Life Science (HCLS) dataset descriptions (Gray, Baran, Marshall, & Dumontier, 2015), Bioschemas (Gray, Goble, Jimenez, & community, 2017), and DisGeNET (Piñero et al., 2017). In order to make it easier for future updates, we developed some applications to automatically create ShEx shapes from HCLS datasets specification and Bioschemas profiles.

Bioschemas

Schema.org is a collaborative effort aiminingto create, mantain and promote schemas for strutcture data on the Internet ("Home - schema.org," n.d.). Bioschemas is a community-driven project aiming to support schema.org types for Life Sciences. It contributes to the community by adding life Science types to schema.org, defining profiles adjusted to community needs, and developing supporting tools. A Bioschemas profile is a type customization including property cardinality and requirement level. Bioschemas shapes currently focus on profiles corresponding to the Biotea project, particularly those related to bibliographic data. Biotea (A. Garcia et al., 2018) provides a model to express scholarly articles in RDF, including not only bibliographic data but also article structure and named entities recognized in the text.

Biotea-Bioschemas ShEx shapes are created via a Jupyter notebook from the YAML Bioschemas profile files. Schema.org datatypes are transformed to XML Schema Definition (XSD) while supporting shapes are created for any combination of schema.org types used as ranges. In addition, three main shapes are created for any Bioschemas profile, corresponding to the three property requirement levels, i.e., minimum, recommended and optional. Profile information, i.e., profile name, schema.org type and YAML file location, are encoded in a comma separated value (CSV) file, making it easy to use the code to generate shapes for any other Bioschemas profile. More information is available at the GitHub repository for this project (Garcia, 2019).

DisGeNET

DisGeNET is a comprehensive gene-disease association knowledge base in the Life Sciences. It is widely used by the biomedical community and its Linked Data representation has been selected as an Elixir Europe ("ELIXIR A distributed infrastructure for life-science information," n.d.) interoperability resource. However, it is still lacking a way to easily query this vast amount of information and explore this knowledge across other domains through its SPARQL endpoint.

During the BioHackathon we implemented the DisGeNET-RDF ShEx shape (Queralt-Rosinach, 2019). In order to do so, we used RDFShape (University of Oviedo, n.d.) and the suite of generation and validation tools it comes with. We detected some disagreements between the DisGeNET schema illustrated on its website and the actual underlying data. We actively discussed around how to best tackle the development of the ShEx shapes in an automatic and data-driven way so we can continue working on it after the BioHackathon.

HCLS

The HCLS Community Profile for Dataset Descriptions offers a concrete guideline to specify dataset metadata as RDF including elements of data description, versioning, and provenance so as to support discovery, exchange, query, and retrieval of dataset metadata. As part of their work, the HCLS Community created Validata (Beveridge et al., 2015), a web application to check the compliance of RDF documents to the guideline specifications. Validata used a non-standard extension of ShEx to check various compliance levels.

We created a ShEx compliant document by processing the HCLS guideline using a PHP script (Dumontier, 2019). The result is several ShEx documents that can be used to check compliance at various levels (MUST, SHOULD, MAY, SHOULD NOT, MUST NOT). We validated our work against the exemplar documents that are provided as part of the guideline, and have also used it to detect errors in HCLS metadata from UniProt. Our work revealed errors in UniProt metadata and the RDFShape tool.

Rare disease catalogs and registries

Data on rare disease is currently fragmented across various databases and online resources making efficient and timely use of this data in rare disease research challenging. Several data catalogs exist that collect data from biobanks and patient registries but these data are neither comprehensive or readily interoperable across catalogs. There is now an international effort to improve the discovery, linkage and sharing of rare disease data through the development of standards and the adoption of FAIR data principles. One component of this process is the development of common metadata models for describing and sharing data across resources using standard vocabularies and ontologies.

During the hackathon we explored the use of both JSON schema and ShEx for validating data that conforms to schemas developed as part of the European Joint Programme on Rare Diseases (EJP RD) ("EJP RD – European Joint Programme on Rare Diseases," n.d.). The EJP RD schemas are expressed using JSON Schema, and are accompanied by an additional JSON-LD context file that enables instance JSON data from data providers to be transformed into RDF. At the hackathon we developed a set of new ShEx shapes that could validate the resulting RDF. This required mapping validation rules, such as required properties and cardinality/value type constraints, from JSON schema to an equivalent constraint in ShEx. We were able to demonstrate how more complex types of validation were also possible using ShEx when additional RDF based resources are available. For example, we can express that the `dcat:theme` of rare disease dataset must be a URI and that this URI should be any subclass of the root disease class in the Orphanet rare disease ontology. The resulting EJP RD schemas and accompanying ShEx files are all available on GitHub (Jupp, Cornet, Rajaram, Holub, & PhilipvD, 2019).

ShEx creator

While ShEx is very useful as demonstrated to validate RDF data, the syntax to actually write a ShEx expression can be hard for new users and is time-consuming also for experienced people. Therefore, a prototype of a ShEx creator was proposed for the BioHackathon. This tool should help users to write correct ShEx expressions faster. The prototype is implemented as a javascript tool, supporting the user through e.g. dropdown menus to create a correct ShEx structure and it uses the RDFShape API in the background to validate the created ShEx expression. The prototype can be found at the corresponding GitHub repository (Liener, 2019).

Improvements to RDFShape tool

The RDFShape tool (Labra-Gayo et al., 2018) comprises a set of tools to create and validate RDF data via ShEx and SHACL shapes. During the BioHackathon, it was used to create shapes and validate RDF data from different endpoints. Thanks to it, we identified some improvements for this tool such as the possibility to validate triples obtained from a mix including RDF data provided by the user and data already contained in a SPARQL endpoint. This feature was added to the new version developed during the BioHackathon.

We also explored and implemented new visualization features for ShEx. Our implementation resulted in the separation in several modules: - RDFShape client ("Weso/rdfshape-client," 2019) which consists of a javascript client based on the React framework. - RDFShape server ("Weso/rdfshape," 2019) contains the server part and is implemented in Scala using the http4s ("Http4s http4s," n.d.) library. - umlSHaclEx] ("Weso/umlShaclex," 2019) is a module that generates UML-like visualizations from Shapes schemas. The library can be used as a standalone command line tool. - SHaclEX ("Weso/shaclex," 2019) contains the main validation modules for ShEx and SHACL. - SRDF ("Weso/srdf," 2019) defines a simple RDF interface with the main features required by the validation library. The module contains

several implementations of that interface which enables the use of validation with Apache Jena ("Apache Jena -," n.d.) models, RDF4j (Guindon, n.d.), or SPARQL endpoints.

Schema conversion across validation approaches

As part of our work, during the BioHackathon we worked on identifying a common subset of ShEx that could be used as the basis for the generation of RDF data models documentation, which can later be converted to JSON schema, ShEx or SHACL. Although full interoperability between those languages is not feasible, we consider that a subset language could be defined that could handle the most common cases (Labra-Gayo, García-González, Fernández-Alvarez, & Prud'hommeaux, 2019).

Through CD2H's ("CD2H," n.d.) Data Discovery Engine ("CTSA Data Discovery Engine," n.d.) project, we previously developed a web-based tool called Schema Playground ("CTSA Data Discovery Engine Schema playground," n.d.) to facilitate the schema visualization, hosting and extension. It helps developers to publish their existing schemas as well as build new schemas by extending the existing ones. Schema Playground currently supports schema.org schemas defined in JSON-LD format and JSON-schema-based data validation. While JSON-schema is a good-fit for the underlying JSON-based data structure, ShEx and SHACL provide a more expressive way to describe validation rules when the underlying data are presented as triples. At the hackathon, we converted several JSON-Schema based validation rules to ShEx and performed the validation on the underlying data (e.g., dataset metadata). These exercises help us to identify the requirements to add support for ShEx in our BioThings schema playground.

Conclusion

We developed a formal description of the HCLS dataset metadata guidelines in a manner that is compliant with the latest version of ShEx. This work is important not only to the HCLS community that uses the guideline, but also can form a basis for automated computational validation of metadata descriptions, as per the FAIR (Findable, Accessible, Interoperable, Reusable) principles. In a similar vein, we prototyped a ShEx shapes (semi)automatic solution for Bioschemas which could be later extended to Bioschemas profiles other than those defined by Biotea. We also developed a prototype corresponding to the first formal description of the DisGeNET-RDF data model by using ShEx (Queralt-Rosinach, 2019). Our strategy to generate the DisGeNET-RDF ShEx shape comprised three steps: (i) manual building via the depicted schema on the web, then (ii) polishing via inference from some actual data instances, and (iii) validating against all the database via the SPARQL endpoint. The shapes created for DisGeNET will work as a basis to develop a more automated solution for this resource.

The development of ShEx shapes using the RDFShape tools resulted in a user testing exercise, where bugs were identified. This direct interaction with users allowed us quickly implement fixes and immediately testing them with users, giving place to a new version. In addition, from the creation of ShEx shapes and the transformation from one format to another, we identified a need to improve tools and technologies used to describe and validate RDF data. Such validation could facilitate machine-readable community agreements regarding metadata, thus leading to more Findable, Accessible, Interoperable and Reusable (FAIR) data as community-based validators could interoperate with the FAIR metrics evaluator (Wilkinson et al., 2018).

Future work

Regarding the generation of ShEx shapes, HCLS team plans to check the compliance of other HCLS dataset metadata documents on the web and report to the community our findings while Bioschemas will work on a validation platform that can later communicate with the FAIR evaluator. Regarding DisGeNET, the ShEx shapes will be finalized and move to a more automatic generation.

In order to overcome the necessity to learn yet another syntax, i.e., ShEx syntax, the work on ShEx tooling will continue. Currently, the ShEx creator is a rough prototype. Future work consists of (i) making the code more stable and potentially publish it as npm module and (ii) integrate the ShEx creator within the RDFShape tool website, so it could further be combined with existing functionality, e.g., ShEx visualization in RDFShape platform.

RDFShape will continue using user feedback to improve the services provided, taking into account scalability requirements of big SPARQL endpoints. Several issues appeared when validating those big data portals, such as the need to improve error messages, and to handle streaming validation for big RDF data. Regarding the visualization, we will work an a direction similar to the one carried out by the Japanese Life Science Database Integration portal ("Home - integbio.jp," n.d.). This portal uses data model representations drawn manually, combining instances and schemas. In such a way, they can show a visualization that users will follow more easily as they will observe real data rather than only the underlying model. Future work could extend the visualization capabilities of RDFShape to automatically generate those kind of visualizations. Other future works on the RDFShape platform include the development of Jupyter notebooks integrating and showcasing the different tools provided.

The BioThings team also plan to continue their work after the hackathon to allow publishing and visualizing ShEx schemas in Schema Playground, along with the support of schemas defined in schema.org and JSON-schema format. The ShEx parsing tools developed at RDFShape will be adopted to convert input ShEx schema into its JSON format for indexing purpose. And the visualization tool from RDFShape can also be used to generate the graph-representation of a ShEx schema.

Jupyter notebooks created

- Bioschemas ShEx shapes: https://github.com/biotea/validation-shapes-bioschemas

Acknowledgements

This work was done during the BioHackathon 2019 organized by NBDC/DBCLS in September 2019 in Fukuoka, Japan. We thank the organizers for the opportunity and the support via travel grants for some of the authors.

References

Apache Jena -. (n.d.). Retrieved from https://jena.apache.org/

Baker, T., & Prud'hommeaux, E. (2019). *Shape Expressions (ShEx) Primer*. Retrieved from https://shexspec.github.io/primer/

Beveridge, A., Baungard Hansen, J., Val, J., Gehrmann, L., Roisin, F., Khutan, S., & Robertson,

T. (2015, May). Validata: RDF Validator. *Validata: RDF Validator using Shape Expressions*. Retrieved from https://www.w3.org/2015/03/ShExValidata/

CD2H. (n.d.). Retrieved from https://ctsa.ncats.nih.gov/cd2h/

CTSA Data Discovery Engine. (n.d.).*http://discovery.biothings.io/*. Retrieved from http://discovery.biothings.io/

CTSA Data Discovery Engine Schema playground. (n.d.).*https://discovery.biothings.io/schema-playground*. Retrieved from https://discovery.biothings.io/schema-playground

Dumontier, M. (2019, October). Micheldumontier/hcls-shex. Retrieved from https://github.com/micheldumontier/hcls-shex

EJP RD – European Joint Programme on Rare Diseases. (n.d.). Retrieved from http://www.ejprarediseases.org/

ELIXIR A distributed infrastructure for life-science information. (n.d.). Retrieved from https://elixir-europe.org/

Garcia, A., Lopez, F., Garcia, L., Giraldo, O., Bucheli, V., & Dumontier, M. (2018). Biotea: Semantics for Pubmed Central. *PeerJ*, *6*, e4201. doi:10.7717/peerj.4201

Garcia, L. (2019, December). Biotea/validation-shapes-bioschemas. biotea. Retrieved from https://github.com/biotea/validation-shapes-bioschemas

Gray, A. J. G., Baran, J., Marshall, M. S., & Dumontier, M. (2015). *Dataset descriptions: HCLS community profile*. Retrieved from https://www.w3.org/TR/hcls-dataset/

Gray, A. J. G., Goble, C., Jimenez, R. C., & community, T. B. (2017). Bioschemas: From Potato Salad to Protein Annotation. In *Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks co-located with 16th International Semantic Web Conference* (Vol. 1963, p. 4). Vienna, Austria.

Guindon, C. (n.d.). Eclipse RDF4J The Eclipse Foundation. *Eclipse rdf4j*. Retrieved from https://rdf4j.eclipse.org/

Home - integbio.jp. (n.d.). Retrieved from https://integbio.jp/en/

Home - schema.org. (n.d.). Retrieved from http://schema.org/

Http4s http4s. (n.d.). Retrieved from https://http4s.org/

Jupp, S., Cornet, R., Rajaram, Holub, P., & PhilipvD. (2019, September). Ejp-rd-vp/resource-metadata-schema. ejp-rd-vp. Retrieved from https://github.com/ejp-rd-vp/resource-metadata-schema

Knublauch, H., & Kontokostas, D. (2017). *Shapes Constraint Language (SHACL)*. Retrieved from https://www.w3.org/TR/shacl/

Labra-Gayo, J. E., Fernández-Álvarez, D., & GArcía-González, H. (2018). RDFShape: An RDF playground based on Shapes. In *Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks co-located with 17th International Semantic Web Conference* (p. 4). Monterey, USA.

Labra-Gayo, J. E., García-González, H., Fernández-Alvarez, D., & Prud'hommeaux, E. (2019). Challenges in rdf validation. In G. Alor-Hernández, J. L. Sánchez-Cervantes, A. Rodríguez-González, & R. Valencia-García (Eds.), *Current trends in semantic web technologies: Theory and practice* (pp. 121–151). Cham: Springer International Publishing. doi:10.1007/978-3-030-06149-4_6

Liener, T. (2019, September). LLTommy/RDFvalidation4humans. Retrieved from https://github.com/LLTommy/RDFvalidation4humans

Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E.,

García-García, J., et al. (2017). DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, *45*(D1), D833–D839. doi:10.1093/nar/gkw943

Queralt-Rosinach, N. (2019, September). NuriaQueralt/shex-shapes. Retrieved from https://github.com/NuriaQueralt/shex-shapes

University of Oviedo, W. semantics group at. (n.d.). RDFShape. Retrieved from http://rdfshape.weso.es/

Weso/rdfshape. (2019, December). Web Semantics Oviedo, University of Oviedo. Retrieved from https://github.com/weso/rdfshape

Weso/rdfshape-client. (2019, December). Web Semantics Oviedo, University of Oviedo. Retrieved from https://github.com/weso/rdfshape-client

Weso/shaclex. (2019, December). Web Semantics Oviedo, University of Oviedo. Retrieved from https://github.com/weso/shaclex

Weso/srdf. (2019, December). Web Semantics Oviedo, University of Oviedo. Retrieved from https://github.com/weso/srdf

Weso/umlShaclex. (2019, December). Web Semantics Oviedo, University of Oviedo. Retrieved from https://github.com/weso/umlShaclex

Wilkinson, M. D., Dumontier, M., Sansone, S.-A., Santos, L. O. B. da S., Prieto, M., McQuilton, P., Gautier, J., et al. (2018). Evaluating FAIR-Compliance Through an Objective, Automated, Community-Governed Framework. *bioRxiv*, 418376. doi:10.1101/418376