Validata: A tool for testing profile conformance

Alasdair J G Gray

Department of Computer Science, Heriot-Watt University, Edinburgh, UK

Abstract. Validata is an online web application for validating a dataset description expressed in RDF against a community profile expressed as a Shape Expression (ShEx). Additionally it provides an API for programmatic access to the validator. Validata is capable of being used for multiple community agreed standards, e.g. DCAT, the HCLS community profile, or the Open PHACTS guidelines, and there are currently deployments to support each of these. Validata can be easily repurposed for different deployments by providing it with a new ShEx schema. The Validata code is available from `https://github.com/HW-SWeL/Validata`.

Keywords: Validation, Conformance, Metadata, Dataset Descriptions

1 Introduction

The Validata tool [1] is a web application that enables users to easily check the conformance of a dataset description, expressed in RDF, against a metadata profile expressed as a Shape Expression (ShEx) schema [2,3]. To support the multiple metadata profiles that have been created, e.g. DCAT [4], the W3C Health Care and Life Sciences Community (HCLS) Profile for Dataset Descriptions [5,6], or the Open PHACTS specification [7], the Validata tool can be easily reconfigured by supplying a ShEx schema corresponding to the metadata profile. This ease of configuration is in contrast to the bespoke validators that have been developed for the DCAT[1] and Open PHACTS[2] standards.

Validata provides client-side, browser-based validation of RDF documents, although it is also possible to use it through an API so that it can be included as part of a data publishing pipeline. The validation provides support for multiple RDF serialisations as well as different requirement levels, i.e. stating whether a property is required, desirable or entirely optional. Validata was developed to check dataset descriptions (RDF documents) against the W3C Health Care and Life Sciences Community (HCLS) Profile for Dataset Descriptions [5,6]. Due to its use of ShEx, it is capable of being re-purposed for other metadata standards (e.g. DCAT [4]) or validation use cases that occur in data exchange scenarios.

[1] System: `http://www.dcat.be:8080/validator/`
Code: `https://github.com/oSoc15/dcat-validator`
Blog: `http://open.summerofcode.be/2015/07/10/dcat/` all accessed Nov 2015

[2] Deployment: `http://openphacts.cs.man.ac.uk:9095/QueryExpander/validate`
Code: `https://github.com/openphacts/Validator` all accessed Sept 2016

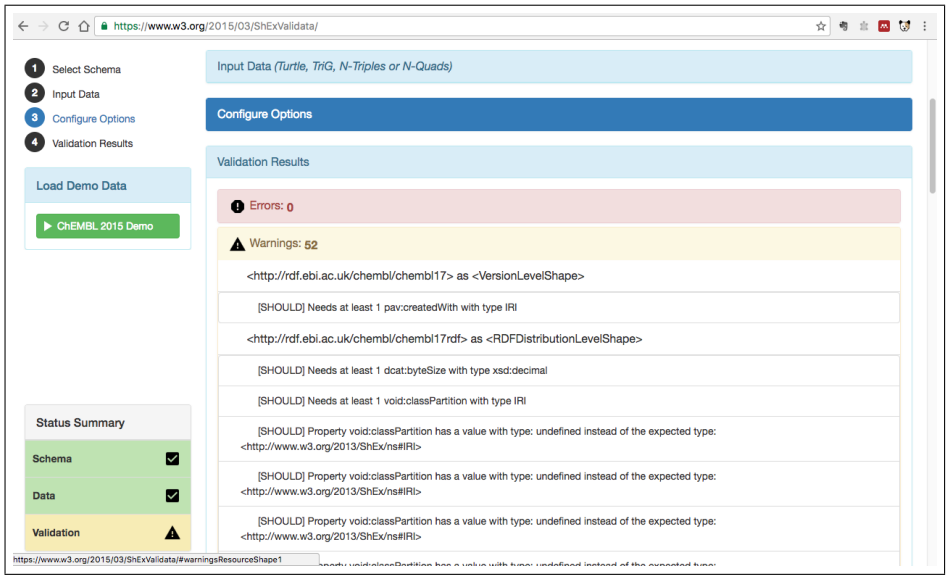

Fig. 1. Screenshot of the Validata tool.

2 Validata UI

Figure 1 shows a screenshot of the Validata web application. The user interface consists of several panels. The first is the Input Data panel, that allows the entering of the RDF representing the dataset description. File upload or direct entry are both supported, as are several RDF serialisations. Each line is syntactically checked and line numbers provided. The next panel is used to configure the resources to validate and the type of description to validate, e.g. in DCAT you can choose to validate a resource as either a dataset or a distribution. The final panel reports the validation results. There is a status panel at the bottom left of the application that provides a summary of the validation process, e.g. the screenshot shows that the description is valid RDF but does not provide all the optional properties of the description profile. In the screenshot, the input and configuration panels have been collapsed to highlight the error reporting. Each error report is hyperlinked back to the relevant lines of the input RDF.

3 Validata Deployments

Validata is a generic tool for validating RDF documents against a set of constraints captured as a ShEx schema. The tool can easily be deployed in different settings by editing the configuration file containing the ShEx schema. The deployment at `http://hw-swel.github.io/Validata/` allows the user to select between different schemas, and to supply their own schema written in ShEx.

We have also made three tailored deployments of Validata corresponding to the HCLS Community Profile, DCAT, and the Open PHACTS specification.

Validata was initially developed for the HCLS Community Profile and is deployed by the W3C at `http://www.w3.org/2015/03/ShExValidata/`. During the development of Validata, the HCLS Community Profile was still under discussion. The rapid reconfigurability of Validata proved to be invaluable to enable it to be kept up-to-date with the latest changes in the specification.

An instance capable of validating DCAT records [4] is available at `http://www.macs.hw.ac.uk/~ajg33/validata/`. This instance does not use requirement levels as these are not defined in the DCAT recommendation. Nor does the DCAT specification specify which properties are optional or mandatory, as such all properties have been specified as optional. The demonstration example uses the example in the DCAT specification. This example does not use closed world validation as it includes `rdfs:label` which is not specified anywhere else in the DCAT recommendation.

We have also deployed a version of Validata for validating dataset descriptions against the Open PHACTS specification [7]. This instance is available from `http://openphacts.cs.man.ac.uk/validata/` and uses a description for the DrugBank dataset for its example. Again requirement levels are used.

4 Conclusions

Validata provides a reconfigurable, online tool for validating RDF documents against a set of constraints defined in a ShEx schema, i.e. a representation of a metadata profile. Informative error and warning messages are presented to the user to enable them to refine their RDF to conform with the schema. The validation implementation extends the standard ShEx definition to support different requirement levels; with the levels being defined in the schema definition rather than prescribed by the ShEx language.

The tool has been deployed by the W3C HCLS Interest Group to validate dataset descriptions against their community profile. Additional deployments have been made for the DCAT specification and the Open PHACTS project.

Acknowledgements

I would like to acknowledge the Heriot-Watt MEng Software Engineering 2015 cohort for designing and implementing the Validata tool: Jacob Baungard Hansen, Andrew Beveridge, Roisin Farmer, Leif Gehrmann, Sunil Khutan, Tomas Robertson, and Johnny Val. We thank Eric Prud'hommeaux for the support and advice received with regard to ShEx throughout the development of Validata.

References

1. Hansen, J.B., Beveridge, A., Farmer, R., Gehrmann, L., Gray, A.J.G., Khutan, S., Robertson, T., Val, J.: Validata: An online tool for testing RDF data conformance.

In: Proceedings of the 8th Semantic Web Applications and Tools for Life Sciences International Conference, Cambridge UK, December 7-10, 2015. (2015) 157–166
2. Gayo, J.L., Prudhommeaux, E., Solbrig, H., Rodríguez, J.A.: Validating and describing linked data portals using RDF Shape Expressions. In: Workshop on Linked Data Quality. (2014) http://ceur-ws.org/Vol-1215/paper-06.pdf.
3. Prud'hommeaux, E., Labra Gayo, J.E., Solbrig, H.: Shape expressions: An RDF validation and transformation language. In: 10th International Conference on Semantic Systems. (2014) 32–40 doi:10.1145/2660517.2660523.
4. Maali, F., Erickson, J.: Data catalog vocabulary (DCAT). Recommendation, W3C (January 2014) http://www.w3.org/TR/vocab-dcat/.
5. Gray, A.J.G., Baran, J., Marshall, M.S., Dumontier, M.: Dataset descriptions: HCLS community profile. Interest group note, W3C (May 2015) http://www.w3.org/TR/hcls-dataset/.
6. Dumontier, M., Gray, A.J., Marshall, M.S., Alexiev, V., Ansell, P., Bader, G., Baran, J., Bolleman, J.T., Callahan, A., Cruz-Toledo, J., Gaudet, P., Gombocz, E.A., Gonzalez-Beltran, A.N., Groth, P., Haendel, M., Ito, M., Jupp, S., Juty, N., Katayama, T., Kobayashi, N., Krishnaswami, K., Laibe, C., Le Novère, N., Lin, S., Malone, J., Miller, M., Mungall, C.J., Rietveld, L., Wimalaratne, S.M., Yamaguchi, A.: The health care and life sciences community profile for dataset descriptions. PeerJ **4** (August 2016) e2331
7. Gray, A.J.G.: Dataset descriptions for the open pharmacological space. Working draft, Open PHACTS (September 2012) http://www.openphacts.org/specs/datadesc/.