PROV4ITDaTa: Transparent and direct transfer of personal data to personal stores

Gertjan De Mulder, Ben De Meester, Pieter Heyvaert,
Ruben Taelman, Ruben Verborgh, Anastasia Dimou
Ghent University – imec – IDLab Department of Electronics and Information Systems
Ghent, Belgium
{firstname.lastname}@ugent.be

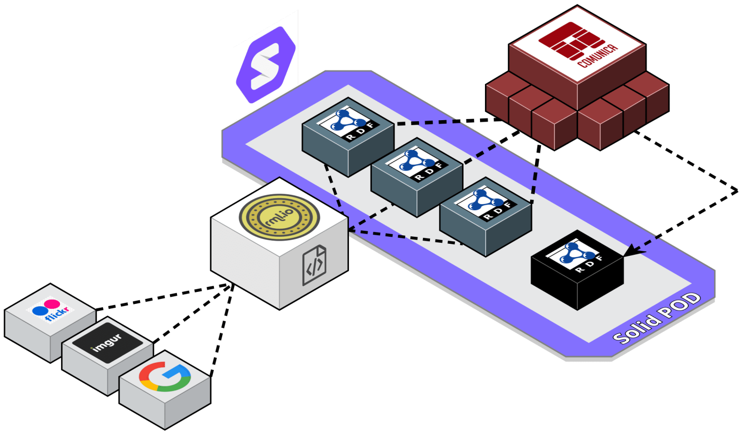


Figure 1: PROV4ITDaTa: Architecture overview

ABSTRACT

Data is scattered across service providers, heterogeneously structured in various formats. By lack of interoperability, data portability is hindered, and thus user control is inhibited. An interoperable data portability solution for transferring personal data is needed. We demo PROV4ITDaTa: a Web application, that allows users to transfer personal data into an interoperable format to their personal data store. PROV4ITDaTa leverages the open-source solutions RML.io, Comunica, and Solid: (i) the RML.io toolset to describe how to access data from service providers and generate interoperable datasets; (ii) Comunica to query these and more flexibly generate enriched datasets; and (iii) Solid Pods to store the generated data as Linked Data in personal data stores. As opposed to other (hard-coded) solutions, PROV4ITDaTa is fully transparent, where each component of the pipeline is fully configurable and automatically generates detailed provenance trails. Furthermore, transforming the personal data into RDF allows for an interopable solution. By maximizing the use of open-source tools and open standards, PROV4ITDaTa facilitates the shift towards a data ecosystem wherein users have control of their data, and providers can focus on their service instead of trying to adhere to interoperability requirements.

WWW '21 Companion, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8313-4/21/04.
https://doi.org/10.1145/3442442.3458608

KEYWORDS

Data portability, Interoperability, Transparency, Linked Data

ACM Reference Format:
Gertjan De Mulder, Ben De Meester, Pieter Heyvaert,, Ruben Taelman, Ruben Verborgh, Anastasia Dimou. 2021. PROV4ITDaTa: Transparent and direct transfer of personal data to personal stores. In *Companion Proceedings of the Web Conference 2021 (WWW '21 Companion), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3442442.3458608

1 INTRODUCTION

Data is scattered across different service providers (e.g. Google, Flickr, etc.), heterogeneously structured in various formats. Interoperability – the ability to be exchanged and unambiguously processed by machines – is lacking, and as a consequence, data portability is hindered. Service providers store data on their servers, but there is no trivial solution for porting this data elsewhere. Data which is siloed within different service providers negatively contributes to the vendor lock-in issue, where users become reliant on a single service provider. Not only does this confine the freedom of users and thus inhibits user control, it also burdens smaller providers by forcing them to gather data, rather than investing effort in the service they provide.

Existing initiatives, such as the Data Transfer Project[1] (DTP), attempt to address the aforementioned problems by providing an

[1] https://datatransferproject.dev/

open-source, service-to-service data portability platform that allows users to transfer their data. Unfortunately, such efforts – being hard-coded – may inhibit (i) *transparency* because inspecting the transfer would require reviewing its source code, as well as (ii) *interoperability* because service providers use custom data models.

On the one hand, users wishing transparency, i.e. assessing the trustworthiness of the transfer, would need to review the source code to determine the origin and applied transformations of the transferred data, justifying the need for incorporating data provenance [3]. On the other hand, complying to hard-coded implementations require development effort for both the creation and alignment of custom data models, underlining the need for a configurable approach [2].

A portability tool is needed that allows transferring heterogeneously structured personal data from different service providers in an interoperable format, without bothering users with technicalities that would refrain them from doing so. This can help bring control back to the users. Furthermore, unlocking this data in an interoperable way levels the playing field for small and big service providers, allowing them to focus on their service's core functionalities.

Declaratively describing the transfer allows for (i) assessing its quality without requiring to inspect the source code or to execute the transfer, and (ii) automatic provenance generation. Fine-grained configuration of the transfer is required to increase flexibility in adhering to other data models, and reduces maintenance costs by minimizing implementation efforts.

In this demo paper, we present PROV4ITDaTa: an open-source Web application available online at https://prov4itdata.ilabt.imec.be/. The source code (MIT-licensed) is available at https://github.com/RMLio/prov4itdata-web-app. PROV4ITDaTa is a transparent data portability tool and allows for fine-grained configuration to improve interoperability of personal data.

2 PROV4ITDATA

In this demo paper, we present PROV4ITDaTa, an open-source Web application[2] that allows transferring personal data from multiple service providers to a personal data store. The PROV4ITDaTa platform exploits and advances the existing open-source solutions RML.io[3] [4], Comunica[4] [7], and Solid[5] [6], and combines them to yield a transparent transfer pipeline (Figure 1). In the following paragraphs we introduce the PROV4ITDaTa Web application, and unveil how (i) heterogenous data sources can be transformed to interoperable datasets through declarative fine-grained configuration and accompanied by automatically generated provenance statements, by leveraging RML.io; (ii) specific parts of the generated datasets can be selected and aligned with other data models and external datasets, using Comunica; and (iii) personal data can be stored in a decentralized environment that lets users control and interlink their data, by leveraging the Solid ecosystem.

PROV4ITDaTa Web Application. The user interface (Figure 2) contains multiple cards, relevant to the different transfer steps. During the first step, users can choose from an extensible list of curated RML rules, each describing the service provider to transfer data from, and how the original data will be transformed to interoperable data using RDF. As depicted by Figure 3, the RML rules are made available for either inspection or download. These RML rules are discussed in greater detail in the following paragraph. In the second step, the user initiates the transfer by clicking the "Execute"-button, and is subsequently guided through the authorization procedure which informs the user about the scope of personal data that will be accessed. The Web application supports multiple authorization schemes (e.g. OAuth1.0 and OAuth2.0), and is extensible to more, hence, improving portability. After successful execution, the interoperable RDF datasets (i.e. the generated dataset and automatically generated provenance statements) are persisted to the user's personal data store. These datasets are available for inspection and download through the user interface, similar to Figure 2.

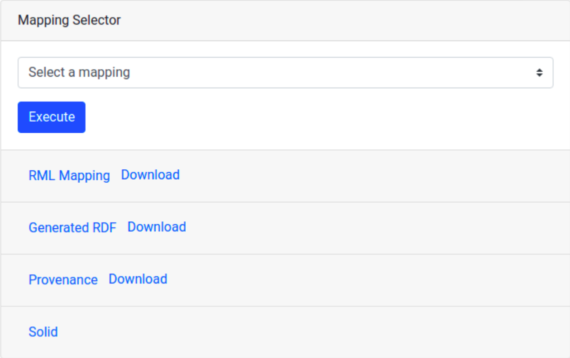
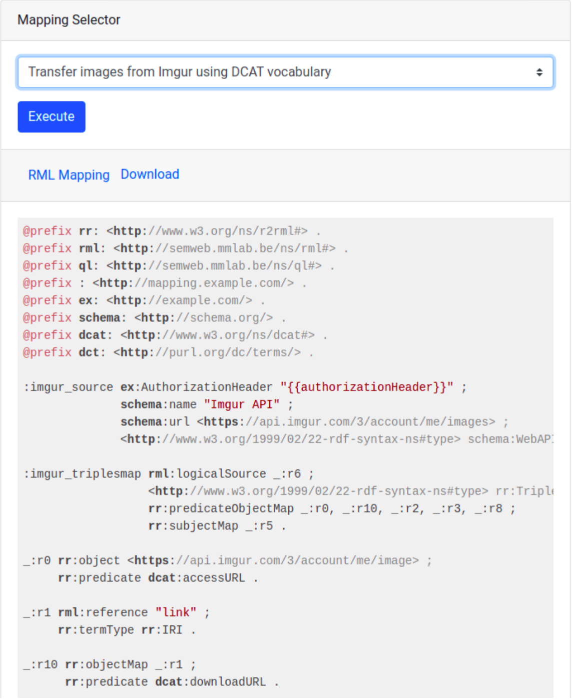


Figure 2: A straight-forward user interface is used to interact with the PROV4ITDaTa Web application

RML.io. RML.io (Figure 4) is a set of open-source tools to generate RDF knowledge graphs, where access and transformations are described using RML rules [4]. These RML rules provide the means to generate semantically enriched RDF data from heterogenous and (semi-)structured sources, using a declarative set of rules. By using a declarative configuration, we provide a transparent transfer, independent of source code, with fine-grained configuration possibilities. Moreover, describing the access is a one-time effort per service provider and can be reused by other RML rules. We demonstrate the flexibility and reusability of this configurable approach by providing RML rules that use different vocabularies (e.g. Schema.org, DCAT) for the transformations. Furthermore, the RML rules support automatic generation of provenance statements concerning applied transformations during the transfer [1, 3]. The provenance information, structured using the W3C recommended standard PROV-O [5], allows further automatic processing to evaluate correctness by the user.

Comunica. Comunica provides a meta-query engine designed in a highly modular and configurable manner to deal with the heterogeneous nature of Linked Data on the Web. Comunica supports executing federated SPARQL queries over one or more interfaces, allowing us to query and combine data from the generated datasets and transfer it to new services. We extended Comunica to include a provenance trail, allowing us to know where resulting data came

[2] https://prov4itdata.ilabt.imec.be/
[3] https://rml.io/
[4] https://comunica.dev/
[5] https://solidproject.org/



Mapping Selector

Transfer images from Imgur using DCAT vocabulary

```
@prefix rr: <http://www.w3.org/ns/r2rml#> .
@prefix rml: <http://semweb.mmlab.be/ns/rml#> .
@prefix ql: <http://semweb.mmlab.be/ns/ql#> .
@prefix : <http://mapping.example.com/> .
@prefix ex: <http://example.com/> .
@prefix schema: <http://schema.org/> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dct: <http://purl.org/dc/terms/> .

:imgur_source ex:AuthorizationHeader "{{authorizationHeader}}" ;
              schema:name "Imgur API" ;
              schema:url <https://api.imgur.com/3/account/me/images> ;
              <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> schema:WebAP

:imgur_triplesmap rml:logicalSource _:r6 ;
                  <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> rr:Triple
                  rr:predicateObjectMap _:r0, _:r10, _:r2, _:r3, _:r8 ;
                  rr:subjectMap _:r5 .

_:r0 rr:object <https://api.imgur.com/3/account/me/image> ;
     rr:predicate dcat:accessURL .

_:r1 rml:reference "link" ;
     rr:termType rr:IRI .

_:r10 rr:objectMap _:r1 ;
      rr:predicate dcat:downloadURL .
```

Figure 3: Full transparency is provided, e.g. by allowing to inspect the RML dataset generation rules

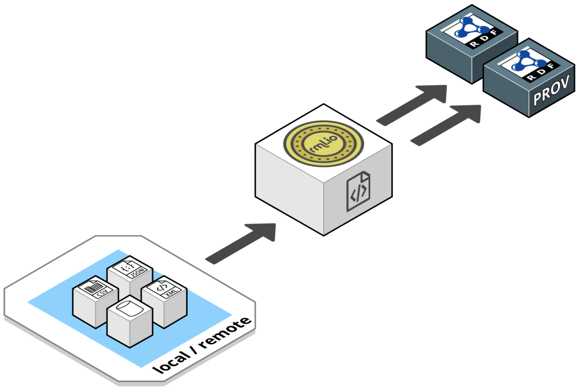


Figure 4: PROV4ITDaTa: RML.io

from. Using Comunica allows us to declaratively describe enriched datasets, that can integrate internal and external data.

Solid. The Solid ecosystem encapsulates a set of W3C standards and tools, based upon the Linked Data principles, aiming towards a sustainable Web and data-ecosystem [8]. Solid includes decentralized data stores called Solid Pods, that we use to store data resulting from the transfer: the generated datasets and provenance trails as well as the parts selected through querying over the generated data. By transferring the personal data to a Solid Pod, the user regains control of its personal data. Moreover, the personal data that was siloed within different service providers, can now be linked together.

3 CONCLUSION

We showcase how users can transparently transfer heterogeneous data from service providers to their own Solid Pod by leveraging and exploiting the open-source solutions RML.io, Comunica, and the Solid ecosystem. We showcase the applicability and extensibility for decentralized environments by applying it to the Solid ecosystem. We showcase the interoperability of our solution by maximizing the use of RDF: the configuration of the transfer pipeline, as well as the (generated) datasets and provenance data it generates are in RDF format. We showcase configuration and personalization abilities by providing multiple configurations for transferring personal data.

PROV4ITDaTa lowers the barrier for services to adhere to data portability requirements as no more development effort is needed, only configuration. Furthermore, the playing field for service providers is leveled: instead of competing in a data harvesting race, providers can focus on their application and use our platform to allow users transferring (parts of) their data into the new application. Aside from the stand-alone setting showcased in this paper, we envision our solution to also be integrated in existing portability solutions such as DTP.

Acknowledgements. The described research activities were funded by Ghent University, imec, Flanders Innovation & Entrepreneurship (VLAIO), and the European Union. Ruben Verborgh and Ruben Taelman are postdoctoral fellows of the Research Foundation – Flanders (FWO). This project has received funding from the European Union's H2020 research and innovation program "DAPSI" under Grant Agreement no 871498.

REFERENCES

[1] Ben De Meester, Anastasia Dimou, Ruben Verborgh, and E. Mannens. 2017. Detailed Provenance Capture of Data Processing. In *Proceedings of the First Workshop on Enabling Open Semantic Science (SemSci)*, Vol. 1931. CEUR. http://ceur-ws.org/Vol-1931/#paper-05

[2] Ben De Meester, Wouter Maroy, Anastasia Dimou, Ruben Verborgh, and Erik Mannens. 2017. Declarative Data Transformations for Linked Data Generation: the case of DBpedia. In *The Semantic Web: 14th International Conference, ESWC 2017, Portorož, Slovenia, May 28 – June 1, 2017, Proceedings, Part II*, Vol. 10250. https://doi.org/10.1007/978-3-319-58451-5_3

[3] Anastasia Dimou, Tom De Nies, Ruben Verborgh, Erik Mannens, and Rik Van de Walle. 2016. Automated Metadata Generation for Linked Data Generation and Publishing Workflows. In *Proceedings of the Workshop on Linked Data on the Web co-located with 25th International World Wide Web Conference (WWW2016)*, Vol. 1593. CEUR. http://events.linkeddata.org/ldow2016/papers/LDOW2016_paper_04.pdf

[4] Anastasia Dimou, Miel Vander Sande, Pieter Colpaert, Ruben Verborgh, Erik Mannens, and Rik Van de Walle. 2014. RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data. In *Proceedings of the 7th Workshop on Linked Data on the Web*, Vol. 1184. CEUR. http://ceur-ws.org/Vol-1184/ldow2014_paper_01.pdf

[5] Timothy Lebo, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. 2013. *PROV-O: The PROV Ontology*. Recommendation. World Wide Web Consortium (W3C). https://www.w3.org/TR/prov-o/

[6] A. Sambra, E. Mansour, Sandro Hawke, Maged Zereba, N. Greco, Abdurrahman Ghanem, Dmitri Zagidulin, A. Aboulnaga, and Tim Berners-Lee. 2016. Solid : A Platform for Decentralized Social Applications Based on Linked Data.

[7] Ruben Taelman, Joachim Van Herwegen, Miel Vander Sande, and Ruben Verborgh. 2018. Comunica: a Modular SPARQL Query Engine for the Web. In *The Semantic Web – ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II*, Vol. 11137. Springer, Cham. https://doi.org/10.1007/978-3-030-00668-6_15

[8] Ruben Verborgh. 2017. Paradigm shifts for the decentralized Web. https://ruben.verborgh.org/blog/2017/12/20/paradigm-shifts-for-the-decentralized-web/