Semantic Schema Mapping for Interoperable Data-Exchange

Harshvardhan J. Pandit(✉), Damien Graux, Fabrizio Orlandi, Ademar Crotti Junior, Declan O'Sullivan, and Dave Lewis

ADAPT SFI Centre, Trinity College Dublin, Ireland
{pandith,grauxd,orlandif,crottija,declan.osullivan,dave.lewis}@tcd.ie

Abstract. GDPR's Right to Data Portability requires data to be provided in an interoperable, commonly used, and machine-readable format and facilitates its transfer between controllers. However, a major challenge for such data to be used between different services is agreement over common schemas to define the semantics of data. We present our vision of a holistic process for organisations to export and import data in an interoperable manner by using ontology matching and mapping techniques to identify a common model towards schema-negotiation. Our approach enables organisations to exchange data using a common base schema, thereby motivating greater innovation in the ecosystem through data reuse. To demonstrate our vision, we present a proof-of-concept application of ingesting data from Facebook into Twitter.

1 Introduction

Interoperability of data between services can facilitate innovation, collaboration, and competition to enable a richer ecosystem of services. The Right to Data Portability (RtDP) was designed and implemented with this as the motivation in Article 20 of the European General Data Protection Regulation[1] to provide a legal impetus for data to be exported out of silos and shared between services. RtDP requires organisations[2] to provide a copy of personal data they have collected from an individual in a structured, commonly used, and machine-readable format. RtDP also permits data to be transmitted directly to another organisation. In principle, this provides individuals as well as organisations the freedom to obtain and reuse existing data from different services and encourages greater competition and innovation between services by countering data silos and user monopolies.

As of August 2020, however, RtDP is yet to be effectively implemented, and there is a lack of consensus in structure and semantics of data which presents technical difficulties associated with interoperability and data sharing across services [11]. One of the major issues in implementing RtDP concerns the 'semantics' of data i.e. how to indicate the structure, context, and meaning of data

[1] http://data.europa.eu/eli/reg/2016/679/oj

[2] consider 'organisation', Data Controller (GDPR), and 'service' as synonyms in article

in an interoperable form. This issue is further compounded given that GDPR does not mandate use of semantics in provision of data. Therefore, data made under RtDP will either (a) have no schema; or (b) its schema is dictated by the service that exported it. In either case, an individual or organisation that wants to use this data must first understand the structure and contents of the data before building tools to use it – which may be feasible when there are a few services but difficult to scale within an ecosystem.

In this article, we present an overview of practical problems regarding implementation of data portability which skew the balance of power against new services and SMEs (small and medium sized enterprises). We then present our vision for a solution that aims to solve this problem using the notion of semantic interoperability where 'data models' or 'schemas' are a) developed within a community, b) embedded or associated with data to convey meaning, and c) aligned with other schemas to enable importing and exporting data between services – thus achieving the intended goals of RtDP.

The novelty of our approach is within the lack of consensus about semantics required between exporting and importing services through a registry of curated schemas that act as a base for interpretation and permit variations in use-cases and applications. To achieve this vision, we propose the use of ontology matching and alignment techniques as the 'bridge' for data interoperability between two services. Further, we discuss the application and role of ontology matching to produce mappings for exporting (downlift) and importing (uplift) data directly between services.

The rest of this article is structured as follows: Section 2 presents the legal requirements and existing implementations of RtDP, and discusses practical challenges with a focus on the feasibility of meaningful exchange of data and the role of semantics; Section 3 presents our vision of a solution and its application on a hypothetical scenario involving transfer of data from Facebook to Twitter; Section 4 concludes this article with a discussion on the practical considerations for implementing our solution and its potential for helping SMEs innovate in an existing dominant ecosystem.

2 RtDP in the Real-World

2.1 GDPR Requirements, Authoritative Opinions, and Guidelines

Article 20 and Recital 68 of the GDPR[3] stipulate data to be provided under RtDP to be structured, commonly used, machine-readable, and interoperable format. further introduces the requirement of interoperability and motivates creation of interoperable formats that enable data portability. They also provide for such data to be transferred (directly) from one Data Controller to another. The guidelines on RtDP provided by Article 29 Working Party (WP29) further

[3] This articles focuses only on the data formats and interoperability requirements for RtDP. Conditions where the right applies, obligations of an organisation, and its compliance is not relevant to this work.

clarify that the RtDP "does not place obligations on other data controllers to support these formats" [5].

Guidelines by WP29 and various Data Protection Authorities on data formats includes use of XML, JSON, and CSV which are widely adopted and used for interoperability. WP29 states that such data formats should be accompanied "with useful metadata at the best possible level of granularity, while maintaining a high level of abstraction ... in order to accurately describe the meaning of exchanged information" [5]. ICO, which is the Data Protection Authority for UK, explicitly suggests RDF[4] as a standardised data format for interoperability. Thus, although the GDPR motivates data sharing between services, it only suggests semantic interoperability[5] with RDF being a practical solution.

Currently, EU's Next Generation Internet initiative is funding projects through the Data Portability and Services Incubator (DAPSI[6]) which lists directions for possible solutions as common shared formats, vocabularies and ontologies for domains, and methods for (semi-)automatically converting data including semantic mapping. The ISO/IEC 19941:2017[7] standard for cloud interoperability outlines the requirements for semantic interoperability, and the practical use of semantic web standards towards shared understanding. An early paper from 2008 presented reuse of semantic web vocabularies for data interoperability within social networks [1]. This shows that the semantic web domain has been a known direction for a solution towards effective implementation of RtDP and achieving semantic interoperability.

2.2 Real-world Implementations

RtDP has been implemented in a wide range of services given its nature as a legal obligation. Several organisations have developed dedicated tools for RtDP such as Google's 'Takeout', Facebook's 'Download Your Information', and Twitter's 'Your Twitter Data'. An example of data portability directly between services is transferring photos from Facebook to Google Photos[8]. The Data Transfer Project[9] (DTP) is a combined initiative consisting of IT behemoths Apple, Facebook, Google, Microsoft, Twitter - to develop an open-source, service-to-service data portability platform. To this end the project is developing[10] 'Data Models' as a common interoperable schema between services.

While these examples are optimistic, the reality is that RtDP has not seen its full impact, and has not been sufficiently implemented by any service or organi-

[4] https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-to-data-portability/

[5] Semantic interoperability was an explicit aim in earlier drafts of WP29 guidelines but was reduced to just 'interoperability' in the final published version [3]

[6] https://dapsi.ngi.eu/

[7] https://www.iso.org/standard/66639.html

[8] https://about.fb.com/news/2019/12/data-portability-photo-transfer-tool/

[9] https://datatransferproject.dev/

[10] https://github.com/google/data-transfer-project/

sation. A survey of data formats used in RtDP [10] shows variation in responses, non-conformance with GDPR requirements, and a lack of semantics. The Data Transfer Project, though it has been running for over 2 years (2018-2020), has not produced any usable results to achieve its aims despite involving the worlds largest IT organisations. An article by De Hert et al. [3] outlines the challenges in implementing RtDP with two potential approaches: (i) minimalist approach - which requires organisations to minimally comply with the GDPR; and (ii) empowering approach - where semantic interoperability provides a stimulus of choice and freedom to the user along with encouraging competition and innovation amongst services. It is the nature of free-market capitalism that established players prefer (i) whilst users and new entrants would prefer (ii) - each for their own benefit. Our vision thus rests on making possible the empowering approach within an ecosystem without additional obligations on organisations that only want to implement the minimal approach for compliance.

2.3 Challenges in implementing Right to Data Portability

Semantic interoperability, in its role as a solution for data portability, depends on the establishment and sharing of schemas along with the data. *schema.org*[11] is a good example of shared and interoperable schema development across services and use-cases based on its continued development and use at web-scale. Another example is Shape Repo[12] which reuses existing vocabularies (such as WikiData[13]) to declare schemas for use in SOLID[14] application development. Similar to these, we base our approach on establishment of common schemas for semantic interoperability through community engagement and maintenance. In this section, we discuss some challenges present within the ecosystem which justify our approach of a community-driven common schema.

(1) When exported data contains no schema: Unless there is an explicit legal requirement that mandates the association of schemas in a specific manner with exported datasets, this situation is likely to continue. So the question arises over who should develop and maintain the schemas? A dominant organisation has interest in maintaining control over its data and reducing its usefulness to other organisations who might be potential competitors. At the same time, these other organisations (and individuals) would be interested in reusing the exported data to enrich or enhance their own features and offerings. Therefore, it is in the natural interest of the community at large to produce schemas to enrich its data-based services to drive innovation and competition. The existing ecosystem based on services offering APIs presents validation of this argument.

(2) When exported data contains a schema: If a service already provides a schema with its exported dataset, it is easier to utilise this schema rather than develop a new one. However, in the longer run, an independent schema is

[11] https://schema.org/

[12] https://shaperepo.com/

[13] https://www.wikidata.org/

[14] https://solidproject.org/

more resilient to control by one provider and can also be managed more efficiently across use-cases. This is evident in the situation where the service changes its schema, thereby requiring every tool and service dependant on its schema to also change their implementations. Therefore, even where a data comes with a schema attached, it is beneficial to develop a common schema and super-impose the data's schema on it.

(3) Stakeholders beyond domains: Thus far, we have only considered situations where services directly compete with each other within the same domain. However, data can also be useful for integration into other services or for added features. An example of this is a service that offers recording 'daily logs' from a user's social media posts regardless of service. In such cases, it may be to the benefit of the service provider to encourage development of features dependant on its data. While the data providing service would want to restrict such services to only work with their data, the service itself would be inclined to support as many services as possible - an avenue for using common schema and tools based on it.

(4) Cost of development and Control: Larger organisations have more resources at their disposal and larger freedom to experiment. Small organisations (SMEs) are often resource-constrained and rely on innovation to compete. Therefore, a common and shared approach for managing intoperable data is of greater benefit to SMEs, which provides an incentive for them to pool their use-cases and resources together to collaborate and share the burden of competition.

3 Proposed solution

Our vision for implementing RtDP addresses the challenges discussed in Section 2.3 by proposing use of common schemas for 'semantic interoperability' in data exchange between services. This includes an interoperable data portability arrangement that benefits all stakeholders by permitting data exporters to continue using their own semantics and data importers understanding the embedded semantics in data. The common schema is used to abstract service-specific design patterns and to serve as a source for common data within a domain. The shared-community aspect of the approach enables sharing of tasks and reducing the effort required in reuse of data and establishing common schemas.

The role of semantic web in this process concerns acting as an interoperable semantic representation using the RDF, RDFS, and OWL standards. We propose utilising ontology matching and alignment to identify the correct schemas for data exported from service A to be transformed and imported into service B. We also propose utilising ontology matching to permit reuse of data based on common schemas without explicit agreement between an exporter and importer. Similarly, we also propose using uplift/downlift mappings between schemas as a technique to potentially perform this step without requiring transformation of data into RDF.

Ontology matching is "the process of generating an ontology alignment between a source and a target ontology" [4]. In the last 15 years, a number of surveys has been published in the area. They review the various techniques proposed for two main categories of approaches, focusing either on *simple correspondences* between concepts/resources [7][6] (*1:1* concept matching) or *complex matching* [9] (for *m:n* or more complex relations). Since ontology matching is one of the oldest and most relevant research areas in the Semantic Web community[15], it has produced a wide variety of techniques and tools ready to be used[16]. Popular implementations, such as the Alignment API[17] [2] or the NeOn Toolkit[18], assist practitioners in attempting to automatically align different schemas.

To explain and discuss the application of semantic web, ontology matching, and mappings in our approach in detail, consider the hypothetical use-case of an individual wishing to obtain posts exported from Facebook and import them to Twitter. This use-case can also be generalised for services both within and outside the social media domain looking to import and reuse some or all of the Facebook data - which furthers the usefulness of our approach.

3.1 Data Ingestion & Conversion

Currently, both Facebook and Twitter[19] export their data under RtDP as JSON[20] — a non-semantic format.

The first step in ingesting Facebook's JSON data is thus to understand its structure and its schema. Where services undertake this effort individually, each service has to duplicate the effort of understanding the structure and keeping its tool updated. By sharing this task, the community can maintain a documentation of the data's schema and structure. If and when Facebook changes the data structure or format, the community can update its documentation without duplication of effort. While it is Facebook's prerogative to structure its data and change it as it feels fit - an argument can be made that frequent and unreasonable changes are detrimental to the spirit of RtDP.

To minimise impact of such changes, a schema corresponding to Facebook's data is created in the common registry, and any tools ingesting Facebook's data utilise the schema instead. Minimal effort is required to 'transform' the data from its underlying structure to one corresponding with the established schema - such as through a python script to convert to CSV or through RDF mapping to convert to JSNO-LD - based on what the desired output format is.

[15] The "OM" workshop has been continuously running at ISWC since 2006.

[16] OAEI, the Ontology Alignment Evaluation Initiative, has been running yearly since 2004, evaluating the latest ontology matching technologies: `http://oaei.ontologymatching.org/`

[17] `http://alignapi.gforge.inria.fr/`

[18] `http://neon-toolkit.org/`

[19] Information about Twitter's data may be out-of-date as its export tool has been non-operational as of August-15-2020.

[20] Facebook exports data as a JSON dump. Twitter exports data as a JavaScript file with JSON objects. Neither supply information about the schema or structure of their data.

3.2 Schema Description

The creation of a Facebook schema is based on first creating a common schema representing ‘a social media post’. The concepts in the Facebook schema are thus specialised variations of the common schema, representing Facebook as a specific type of social media. This abstraction permits a data importer to target data specifically from Facebook (through the Facebook schema) or any social media (through the common social media schema). The abstraction also works to encourage designing common tools to work on the data rather than specialised ones targeting individual services. Figure 1 depicts an example of a common schema targeting social media posts.

The creation of a common schema where none exists is difficult if a community agreement is necessary over its concepts and structure. Therefore, we suggest seeding the first common schema with concepts from dominant data providers in the domain and normalising it towards other existing providers. In the current use-case, this would mean first creating a schema from Facebook’s data, then creating a common schema based on Facebook’s schema, and updating Facebook’s schema to use the common one as its base. By this we mean sub-classing concepts in specialised schemas from common ones. Later, when creating Twitter’s schema, the existing common schema for social media can be used to guide the schema creation process.

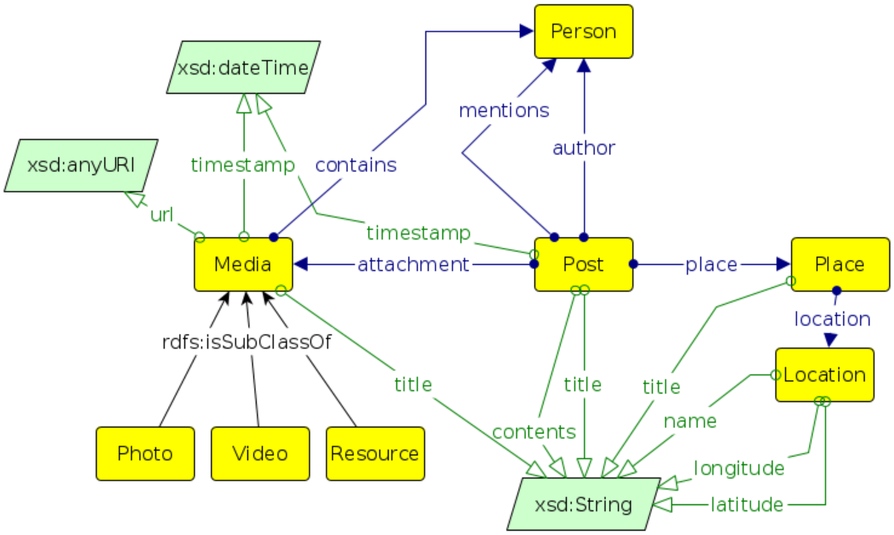

Fig. 1. Example of a common schema for social media post.

3.3 Schema Alignment

In the common and Facebook schemas, the generic terms 'post', 'media', 'timestamp' are suitable for use in both since Facebook does not have any specialised variation of these. However, concepts such as 'like' or 'reaction' may present problems in abstraction and generalisation as they may not be present in other service in the same context. For example, Twitter defines[21] the '♥' symbol to mean a 'like' whereas Facebook defines[22] its 'reactions' as an enumeration consisting of 'like, love, wow, haha, sorry, angry'. Aligning the two is difficult due to semantic differences in the two terms. One interpretation is that only a Facebook 'love' is equivalent to Twitter 'like', whereas another possible interpretation is that any Facebook reaction should be equivalent to Twitter 'like'.

We propose the use of ontology matching and alignment techniques to assist in the schema alignment and discovery process as well as to resolve equivalence between concepts. This can be an automated process, but we also emphasise its value in encouraging discussion amongst schema creators and maintainers through a human-in-the-loop process. The role of common schemas in this is to provide a measure of commonality in the identification and structuring of source and target schemas, as well as to ease the process of finding related and equivalent data patterns. For example, in the case of a Facebook post and Twitter 'tweet', the relationship is easy to establish based on their common super-classes.

Facebook	**Common Schema**	**Twitter**	**Type of alignment**
Post	Post	Tweet	Simple
Contents	Contents	Contents	Simple
Timestamp	Timestamp	Timestamp	Simple
User	Person	Profile	Complex
Friend	Knows	Follows	Complex
Attachment	Media	Media	Simple

Ontology alignment techniques may also provide a way to integrate data where no possible contextual similarity is apparent. For example, Facebook's 'friend' concept and Twitter's 'follows' concept are different in their behaviour and discourse - yet they share similarity in their pattern of association with an individual. It is up to the importer then to determine whether they want to support and utilise such alignments or to discard them in favour of more semantically-refined ones.

Once the matching concepts have been found, the process of transferring data to the target service can take place. An explicit way to do this is to first transform the source data to RDF using its corresponding schema (in this case the Facebook schema), then creating an alignment table using the ontology matching process, and then to generate the dataset using the target schema (in this

[21] https://help.twitter.com/en/glossary

[22] https://developers.facebook.com/docs/graph-api/reference/v8.0/object/reactions

case the Twitter schema). To reduce the number of transformations required in this process, mappings can be potentially used to directly enable the importing service to ingest the source data without the intermediary transformations.

Uplift mapping is the process of converting a data into RDF, while downlift is its inverse. Considering that Facebook exports a JSON data dump, and that Twitter similarly will import[23] a JSON data dump - the process of transformations will involve: (i) uplift Facebook's JSON data into RDF using Facebook schema; (ii) transform RDF data from source schema into target schema using the ontology mapping process; (iii) downlift data into JSON for Twitter. Since the role of step (ii) is merely to find an alignment between the schemas of Facebook and Twitter, the actual transformation of data can take place directly from Facebook's JSON to Twitter's JSON format.

3.4 Using mappings to automate the process

An interesting research question thus arises out of this arrangement - "can we utilise the schema alignments and the mappings to create a tool that will convert the source data to target data?". We believe that it is reasonable to hypothesise that such a tool can indeed be created based on the fact that the structure (i.e. specific arrangement of data structures) of source and target data can itself be considered schemas, and therefore can be utilised to convert one to another. The question around implementing this is then concerned about the efficiency rather than sufficiency. A related area facing similar research challenges is the utilisation of GraphQL to retrieve data from a triple-store in the shape requested by rewriting the query in SPARQL [8].

The use-case we discussed concerned moving data from one social media service to another (Facebook to Twitter). However, RtDP makes it possible to reuse data across a larger plethora of services across domains. For example, Facebook's data contains information about locations the user has tagged their post with (or checked-in). This information could be relevant in any other service providing features that utilise location data - such as a visualisation service that shows all the locations an user has been to on a map. Such a service may want to broaden its data import feature to encourage users to submit *any* location data regardless of its source. Potential sources of such data include: explicit location data shared by user, location tagged in photos, location tagged in social media posts, location inferred from place names and descriptions, location associated with review of a restaurant, or location associated with monetary transactions of a card. Instead of developing separate tools for each of these sources, the service can instead target the underlying common location schema and utilise our approach to ingest the data from a variety of source without additional effort.

In order to identify the potential sources of data, the service can declare the schema for the data it intends to import. For example, this can be a location

[23] Twitter does not provide a data import service. So we reasonably assume its import tool will accept the same data format and structure as its export tool

concept with a label and co-ordinates. A label-based search for related schemas will only retrieve schemas that contain the concept location or its synonym such as ‘place’. However, ontology matching techniques can provide richer results by identifying similarly ‘shaped schemas’ that contain labels and co-ordinates. Further fine tuning is possible by focusing on co-ordinates and its variations while excluding labels. This thus provides an opportunity for utilising ontology matching techniques to identify relevant design patterns for schema discovery.

4 Conclusion

In this paper, we proposed an approach leveraging ontology matching and alignment techniques to achieve data interoperability between online services dealing with personal data. Having GDPR’s Right to Data Portability (RtDP) in mind, we described a typical use-case where users of a social networking service (e.g. Facebook & Twitter) are willing to — and should be allowed to — export their own personal data in a machine-readable format and reuse it on a different service. We described how Semantic Web technologies and ontology matching could assist in the alignment with a common schema that is used as a ‘bridge’ between heterogeneous data schemas. The role of common schemas is to provide a measure of commonality in the structuring of source and target schemas. Finally, we showed how data mappings could be used, and shared via a community-driven repository, to automate the conversion processes. Actually, this last point opens the doors of efficient Data Portability to SMEs which have to allow this feature given the RtDP; in particular, SMEs will be able to minimise the cost of making user data more easily ported to another provider.

We envisage several advantages with the adoption of the proposed approach, both for end-users and companies. *First*, schemas and mappings are open and maintained by the community, lowering the costs for both parties in managing the data transformations. *Second*, maintenance costs are lowered and distributed to the community, removing possible bottlenecks or single points of failure, typical of ad-hoc data transformation pipelines. *Third*, a descriptive and machine-readable schema would not be required from the data exporters anymore, keeping the complexity low at the data sources. *Fourth*, reliability of data transformations would increase. For instance, when one data source changes, mappings updates are faster to perform compared to changes to many ad-hoc pipelines. *Fifth*, the automation potential would increase dramatically with improved, more accurate, ontology matching techniques.

As part of our future work, we plan to implement and test our solution in different use-cases and with different services. This would create a baseline that can be offered to the community and, ideally, adopted and expanded by the community itself. From a more scientific perspective, we will investigate the increased automation possibilities offered by complex ontology matching techniques. Other avenues of potential work include exploration of our approach for interoperability between services and APIs based on semantics, evaluating the efficiency and feasibility at large scales, and discussing the application of our approach within the broader areas of legal compliance and data protection.

Acknowledgments: This research was conducted with the financial support of the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreements No. 801522 and No. 713567 at the ADAPT SFI Research Centre at Trinity College Dublin. The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant #13/RC/2106.

References

1. Boja, U.: Social Network and Data Portability using Semantic Web Technologies. In: Social Aspects of the Web (SAW 2008), Advances in Accessing Deep Web (ADW 2008), E-Learning for Business Needs. p. 15 (May 2008)
2. David, J., Euzenat, J., Scharffe, F., dos Santos, C.T.: The alignment api 4.0. Semantic Web **2**, 3–10 (2011)
3. De Hert, P., Papakonstantinou, V., Malgieri, G., Beslay, L., Sanchez, I.: The right to data portability in the GDPR: Towards user-centric interoperability of digital services. Computer Law & Security Review **34**(2), 193–203 (Apr 2018). https://doi.org/10/gdtmx7
4. Euzenat, J., Shvaiko, P., et al.: Ontology matching, vol. 18. Springer (2007)
5. Guidelines on the right to data portability 16/EN WP 242 rev.01. Article 29 Data Protection Working Party (Dec 2016)
6. Otero-Cerdeira, L., Rodríguez-Martínez, F.J., Gómez-Rodríguez, A.: Ontology matching: A literature review. Expert Syst. Appl. **42**(2), 949–971 (2015), https://www.sciencedirect.com/science/article/pii/S0957417414005144
7. Shvaiko, P., Euzenat, J.: A survey of schema-based matching approaches. In: J. Data Semantics IV (2005)
8. Taelman, R., Vander Sande, M., Verborgh, R.: GraphQL-LD: Linked Data querying with GraphQL. In: Proceedings of the 17th International Semantic Web Conference: Posters and Demos (Oct 2018), https://comunica.github.io/Article-ISWC2018-Demo-GraphQlLD/
9. Thiéblin, E., Haemmerlé, O., Hernandez, N., Trojahn, C.: Survey on complex ontology matching. Semantic Web pp. 1–39 (Oct 2019). https://doi.org/10/gg6rd4
10. Wong, J., Henderson, T.: How Portable is Portable?: Exercising the GDPR's Right to Data Portability. In: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers. pp. 911–920. UbiComp '18, ACM, New York, NY, USA (2018). https://doi.org/10/gfsqrk
11. Zichichi, M., Rodrıguez-Doncel, V., Ferretti, S.: The use of Decentralized and Semantic Web Technologies for Personal Data Protection and Interoperability. In: GDPR Compliance - Theories, Techniques, Tools Workshop of Jurix 2019. p. 10 (2019)