PREPRINT
Final article published in: Sharing of Scientific Data in Enabling Open Semantic Science
(SemSci 2017), co-located with ISWC 2017
proceedings: http://ceur-ws.org/Vol-1931/paper-08.pdf

Linked Data Contracts to Support Data Protection and Data Ethics in the Sharing of Scientific Data

Ensar Hadziselimovic, Kaniz Fatema, Harshvardhan Pandit, David Lewis

Adapt Centre, Trinity College Dublin, Ireland
ensar@adaptcentre.ie

Abstract. In light of the new EU General Data Protection Regulation (GDPR) [1] there are certain challenges in relation to the sharing of scientific data. For a data controller in a research institute, the requirement to monitor, implement and demonstrate conformance to the provisions of data subjects' rights represents a major upshift in the complexity of research data management. We are trying to address the issues by analysing the details of data subject rights in GDPR followed by extending existing linked open data vocabularies. We are proposing a concept of machine- readable data protection rights contract through introducing Data Protection Rights Language (DPRL).

Keywords: Data Protection, GDPR, Open Scientific Data, ODRL, DataID.

1 Introduction

Research in Europe that involves data from individuals is impacted by dual challenges of GDPR and increasing demands from research funders and publishers to adopt open data practices that facilitate replicability of results and bolster research integrity. Research involving individuals could include survey data, personal behaviour observation, recording of communicative acts (e.g. for social media, speech or gesture analysis) or bio data.

Current rules and practices on academic research ethics tend to vary from country to country [2], but with the broad intention of protecting participants and researchers by making clear the purpose of data collection, and requesting explicit consent to use personal data. Good practice in research ethics requires that informed consent is gathered from individuals before any data is collected and may confer subsequent rights on the individual, including the right to withdraw their data from the study or restricting its use for possible other purposes.

Researchers may have to make an undertaking regarding data protection, but there is rarely any follow-up to ascertain whether the data has been stored or destroyed as promised. Overall, the focus of institutional research ethics guidelines has been on obtaining informed consent from experimental subject rather than the later monitoring and enforcing of experimental data handling as consistent with the terms under which that consent is provided.

2 Background

2.1 GDPR

The EU's adoption of GDPR imposes new requirements for tracking informed consent for the usage of any form of personal data. GDPR addresses the processing and movement of personal data, replacing the 1995 Data Protection Directive. It defines *personal data* as "information relating to an identified or identifiable natural person", who it refers to as a *data subject* (Art. 4). When applied to research data, GDPR potentially imposes more rigorous and legally enforceable requirements on the collection and processing of data from individuals than current research ethics practices. There is a strong requirement on *data controller*, who is responsible for handling of personal data, to be able to demonstrate to regulators that any data subject's personal data has been correctly processed, being consistent with the GDPR and the terms of the informed consent under which the data was obtained. This includes the sharing and use of personal data by third parties. GDPR potentially imposes regulatory requirements of tracking data usage that are similar to the rights offered to experimental subjects via informed consent, but that are rarely enforced in practice. Art. 89(2) of GDPR provides for derogations of data subject rights to *access*, *rectify*, *erase*, *restrict*, *port* or *object* to data processing, provided appropriate safeguards are in place. Significantly, the precise nature of the safeguards for any derogation are left for EU member states to legislate on [3] and the uncertainty about their precise nature around gathering personal data for scientific purposes is a major potential organisational risk in GDPR compliance. This may be especially complicated for research conducted across multiple jurisdictions or in collaboration with industry, where national derogations that may emerge will not apply.

2.2 Open Scientific Data

Determining strategies to address this risk is potentially further magnified by the requirement for open access to research data. Publication of the results of publicly funded research has become common practice in recent years. However, the central importance of data in all empirical research, in addition to the growth of big data research approaches, has heightened the call for common policies on publishing and sharing research data associated with a publication [4]. Major research funders, including the EC, have widened their guidelines on open science to now address open research data [5]. The aim in doing so is to make it easier for researchers to: build on previous research and improve the quality of research results; collaborate and avoid duplication of effort to improve the efficiency of publicly funded research; accelerate progress to market to realise economic and social benefits; and to involve citizens and society.

It is anticipated that from 2018 onwards, EC-funded projects will transition from optional involvement in open data pilots to working under a stronger obligation to provide open access to research data. This however needs to be within the constraints of EU and national data regulations, which by that time will include GDPR. However, when sharing research data openly, techniques of pseudo-anonymisation may prove inadequate for protecting data subject identity if the third-party organisation receiving

the data attempts to combine the data with other data sources and potentially de-anonymise it. Important classes of experimental data in computer science and bio sciences (both at the forefront of scientific data sharing) have been shown as vulnerable to de-anonymization by third parties, including linguistic data [6], web search behaviour data [7] and genomic data [8].

The implication of this conflict between data protection and open scientific data requirement is that research institutes can no longer freely share research data containing any personal data. They must restrict sharing to third parties that make an undertaking: to ensure shared data is only used for the purposes agreed with data subject; to ensure cooperation in respecting data subject rights even after the data is shared; and to cooperate in any data protection compliance checks by the originating research institute. The legal basis for such inter-institutional undertakings, e.g. as a research data sharing contract, is out of scope of this paper, and ideally a subject for near term cooperation between national and EU-level scientific policy bodies. Instead, this paper addresses the data interoperability challenges involved in administering such data sharing contracts and supporting GRPR compliance. It proposes a machine-readable data protection rights contract that extends existing linked open data vocabularies and thereby aligns well with the existing directions of common technical interfaces and open data vocabularies for implementing open data science repositories. The paper starts by providing more details on the rights of data subjects that may be propagated between research institution sharing personal data about that subject. We then identify existing data management vocabularies that may form a suitable basis for large scale dataset cataloguing, sharing and data protection compliance checking.

3 Data Subject Rights

One of the issues that will affect all the organisations in all industrial and academic research is related to data subject rights. Although some rights presented in GDPR are already familiar from earlier documents and number of comparisons have been made [9], GDPR brings a lot of specific regulations data subject rights not granted previously. It is important that these are understood properly so the organisations react to the regulation by implementing appropriate data management processes and systems.

3.1 Details of Data Subject Rights in GDPR

The *right to information* states that there is an obligation on all data controllers to provide sufficient amount of **information** to the data subjects "in a concise, transparent, intelligible and easily accessible form, using clear and plain language" (Art. 12(1)). Data subject have the *right of subject access* - right to file a subject ***access*** request (SAR) and obtain a copy of their personal data from the data controller (Art. 15(1)). The *right to rectification*, as in previous regulations, gives data subjects rights to ***rectify*** their data in case of a need (Art. 16). Subjects have right to completely remove or ***erase*** their data in case the data is no longer needed for its original purpose, known as the *right to erasure* (the "right to be forgotten") (Art. 17(1a)). Also, if the data access was given based on consent, and the data subject removes that consent, the data must be

removed as well (Art. 17(1b)). It is up to every organisation to ensure they have facilities to truly and irreversibly remove the data (Art. (17(2)). GDPR also enables data subjects to get a copy of their data in machine-readable format and to transfer/***port*** such data to another service provider, known as the *right to data portability* (Art. 20(1)). Data subjects have the *right to* ***restrict*** access to the data and *right to* ***object*** to any aspect of processing their personal data in relation to their legal rights (Art. 21(1)). Further, data controller must prove and demonstrate that the data processing falls within the previous agreements, and failing to do so, must stop all the data processing activities, including data used in *scientific research* and statistical purposes (Art. 21(6)).

For a data controller in a research institute, the requirement to monitor, implement and demonstrate conformance to the provisions of these data subjects' rights represents a major upshift in the complexity of research data management. Whereas implementing the right to withdraw from studies granted to experimental subjects in conventional research ethics is typically left as a responsibility of the individual researcher or their school or department, under GDPR this becomes an institutional responsibility under a named data controller. From a data management perspective, this imposes new requirement to scale the cataloguing of datasets and the tracking of their use, particularly the provenance of data processing that leads to published results. This requires sophisticated data management and well-structured metadata schema for data sets. Further, as GDPR rights and the requirement to demonstrate compliance propagate to third party organisations in receipt of shared dataset containing personal data, bespoke or proprietary data management solutions will be inadequate, and solution based on interoperable meta-data will be required. Fortunately, open data vocabularies for managing the cataloguing, provenance, permission and obligations exist that may be applicable to this task as we examine in the next section.

4 Existing Data Management Vocabularies

ODRL [10] is Rights Expression Language (REL) maintained by W3C ODRL Standards Group. For this paper, we use ODRL version 2.1 from 2015. The ODRL classes can be hierarchically organised in subclasses so, for example, Privacy class has Policy class as a parent. Each Property has a Class as a domain so, for example, *prohibition* property has Policy class for its domain. Concepts can be better described as actions that can be performed on certain classes. They belong to *actions Concept Scheme* (currently the only concept scheme in ODRL), which in return relates to *Action* class.

DataID [11] is DBpedia project, with W3C member submission from 2016. DataID core ontology defines concepts and properties to describe simple and complex datasets in an interoperable way. It is an extension to the Data Catalog Vocabulary (**DCAT**) [12], adding features such as dataset hierarchies, permissions, distribution and machine-readable licensing information. DataID integrates provenance information from the PROV Ontology [13], enabling the tracking of provenance of the datasets, but also including versioning and inter-dataset relationship information. The main motivation behind DataID is to enhance DCAT's provenance information and to further explain relations between the datasets. DataID classes define and identify the subject of dataset

lifecycle. Object Properties describe how the data will be further manipulated, shared, authorised, as well as suggesting other related datasets that might be tied to the one in question. Finally, Dataset Usage Vocabulary (**DUV**) [14] is specifically used in tracking, sharing, and persisting dataset usage. Main purpose of DUV would be to keep a record of dataset sharing between the two data controllers.

5 Data Protection Rights Language

ODRL is very well suited to be used as a base for our further investigation into the subject. It handles permissions, prohibitions, obligations and assertions. It also has profiles support, enhancing the ODRL mode, enabling us to implement our own requirements "whilst providing a common semantic layer for interoperability" [15]. Some of the profiles include Creative Commons Profile [16] which offloads some of the terms used in ODRL to CC Ontology, and Linked Data Profile [17] extending ODRL with new sets of triples and constraints. There is also ready-made template for describing and publishing new ontology additions [18]. ODRL is currently being standardised by the Permissions and Obligations Working Group at the W3C, and we aim to align our use of ODRL to the resulting recommendation as it matures.

5.1 ODRL Template

The suggested approach is to use ODRL's templating facility to extend it through specific case template that will better be suited to GDPR document. Therefore, Data Protection rights Language (DPRL) is an extension to ODRL, by the means of templating.

We will focus on data subject rights mentioned earlier, to have the ODRL mapping put in context. It is worth mentioning that there are some technological obstacles in following the criteria set by GDPR, as there might be ambiguity in some articles. For example, in Art. 20 it is mentioned that the data should be "portable", but there is no common approach in the community at large when defining the format it should be in, as well as incompatibility of organisational systems between such providers. But the focus of this paper is not to define the techniques, but rather to establish simple workflow to follow the data from one data controller to another.

Common ground for all the mentioned data subject rights in GDPR are that the data subjects are entitled to view, edit, withdraw, delete, move their data, to complain about the procedures, to be informed about any changes in conditions and similar. When it comes to data subjects, that translates to following ODRL concepts: *copy*, *delete*, *read*, *give*, *modify*, *preview*, *grantUse*, *reviewPolicy*, *transfer*. For data controllers, relevant ODRL concepts are: *anonymize*, *archive*, *attribute*, *copy*, *digitize*, *display*, *grants*, *inform*, *present*. GDPR protects the rights of data subjects and set rules and obligations for data controllers. Data subject's role would mostly have certain "rights". In specialising ODRL for data protection, data controller's role would be more related to certain "duties" or "obligations". Data controllers also have rights to use the data as per agreements and consents. They can move, share, transfer, extract, watermark the data and exchange the data with data controllers in other organisations in line with the terms of

the data subject's informed consent. Although ODRL has these action concepts defined in its vocabulary, for data protection, actions to track the implementation of data subject rights are key to ODRL being useful in GDPR compliance. Currently we restrict our extension to these requirements. However, we recognise that the licensing-oriented action concepts on ODRL are insufficient to express the data processing semantic of all services subject to GDPR, and especially in a way that would be intelligible to the data subject. For example, ODRL concept *read* has class Action as its parent class, falling under *actions* concept scheme. Concept *read* in our case can be better classified as a right, rather than action, and should belong to the new *dpRight* class. ODRL concept *anonymize* also belongs to broad *Action* class and again falls under *actions* concept scheme. In our case, it would be better suited for *dpObligation* class.

Furthermore and for our data subject's rights usecase, we suggest adding the following concepts: *dpAccess*, *dpRectify*, *dpErase*, *dpPort*, *dpRestrict*, *dpObject*. Similarly as, for example, ODRL concepts *derive* and *digitize* are more specific terms of broader term *use*, the suggested concepts have their respective broader terms that are present in ODRL. Main reason for more specific concepts is to more clearly define certain scenarios that concern rights and where current ODRL model is not sufficient. Notice that all of the concepts pertain to data subjects. The suggested namespace would be *dprl*.

5.2 DataID Datasets

Proposed DPRL language/ontology would extend ODRL through its templating system, but would also extend some concepts from DataID that relate to the data usage and sharing tracking needed for data protection compliance under GDPR.

DataID has classes *Dataset*, *DatasetRelationship* and *Distribution*, that can be used in tracking the usage and sharing of data sets. *Dataset* and *Distribution* are subclasses of the similarly names classes in the DCAT vocabulary, and the *Entity* class of the PROV-O vocabulary, thereby integrating the fine-grained provenance tracking enabled by the latter vocabulary with the cataloguing metadata of the former. The *Distribution* class is a sharable form of the *Dataset* class and includes a license property (which can reference a machine-readable ODRL declaration). Based on that, there are some obvious advantages of bringing DataID into the mix to further explain and strengthen certain aspects of ODRL that are not ideal for describing and assigning datasets and their distributions. Distinguishing datasets and distributions also tackles the issue of data being in format that is machine-readable and portable (as required by GDPR), while the integration with PROV-O supports the tracking of informed consent, data sharing and data subject rights processing that must be logged to serve potential GDPR compliance checks. For logging purposes, DUV vocabulary would keep track of the usage through *Usage* class, again as per GDPR requirements.

5.3 Using DPRL for Sharing Scientific Data

Currently there is no established method commonly used between the academic institutions when sharing scientific data. Furthermore, data portability clause requires that

the data subject can access their data in common machine-readable format. GDPR requires not only the provenance tracking of data processing to support compliance with serving data subject rights, but that this tracking extends to the data controller of any other organisation with which the data is shared.

Data sharing architrave that we propose in this paper potentially addresses all the challenges, pending future work's improvements. Combining the knowledge on data subject rights, data controller's obligations and by means of extending the existing ontologies, we summarise the concept in the following figure.

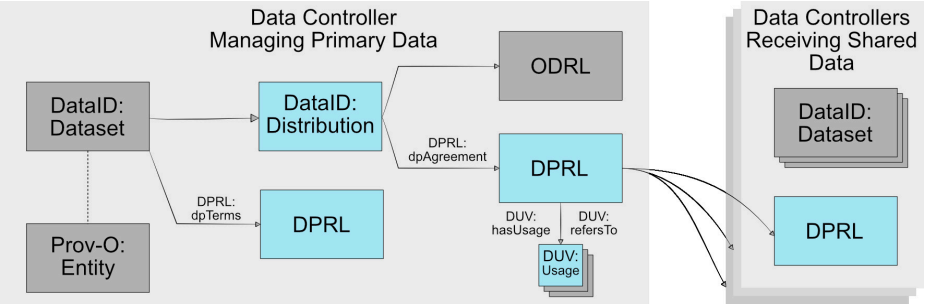

Fig. 1. Open data architecture for sharing the scientific data between academic institutions.

6 Conclusion

In this paper, we identify the potential challenges faced by research institutes using personal data from EU citizens in research studies amid the conflicting requirements of open scientific data policies and GDPR. We describe an initial design for a Data Protection Rights Language (DPRL), as a modest extension to the DCAT and ODRL vocabularies already assembled into the dataset management metadata schema DataID. DPRL aims to support the use of DataID to manage the sharing scientific datasets with support of a machine-readable contracts expressing the permissions and obligations the parties exchange, to satisfy both data sharing and data protection concerns. We hope the development of such an open data vocabulary may be useful in deliberations around the cost and complexity of handling these challenges by European science policy fora and EU member states in considering GDPR derogation under Art. 89(2). It may also promote vendor independence in the procurement of systems for managing the cataloguing, sharing and data protection compliance of scientific data, and provide a basis for agreeing data sharing interactions with collaborators, such as commercial companies, that are not covered by GDPR Art. 89(2) derogations.

In future work, we plan to further develop DPRL through fuller integration with data set processing provenance tracking using PROV-O and the use of SPARQL over DPRL and other DataID components in undertaking GDPR compliance checks. We hope this will inform the design of future open scientific data API and platforms, such as those previously developed for publication metadata in the OpenAire project [19].

7 Acknowledgements

Supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

8 References

1. European Parliament and Council of the European Union: Regulation (EU) 2016/679 of the European Parliament and of the Council (GDPR). Official Journal of the European Union, 119(1):1–88. General Data Protection Regulation, https://gdpr-info.eu (2016)
2. Godecharle S., Nemery B., Dierick K.: Guidance on Research Integrity: No Union in Europe", The Lancet, Volume 381, No. 9872, p1097–1098, 30 March 2013 (2013)
3. Thompson B.: Analysis: Research and the General Data Protection Regulation. Technical report, July (2016)
4. League of European Research Universities: Leru Roadmap for Research Data. Technical report, Dec (2013)
5. European Commission: Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020. Technical report, February (2016)
6. Hovy D., Spruit S. L.: The Social Impact of Natural Language Processing, ACL (2016)
7. Montjoye Y-A de, Hidalgo C. A., Verleysen M, Blondel V. D: Unique in the Crowd: The privacy bounds of human mobility, Scientific Reports 3, Article number: 1376 (2013)
8. Gymrek M., McGuire A. L., Golan D., Halperin E., Erlich Y.: Identifying Personal Genomes by Surname Inference, Science 18 Jan 2013, Vol. 339, Issue 6117, pp. 321-324 (2013)
9. The IT Law Community: Rights of Data Subjects under the GDPR, https://www.scl.org/articles/3575-rights-of-data-subjects-under-the-gdpr, last accessed 28/06/2017
10. Iannella R., Villata S., W3C: ODRL Information Model. 21 July 2016. W3C Working Draft. URL: https://www.w3.org/TR/odrl-model/ (2016)
11. Freudenberg M., Brümmer M.: DataID core Ontology, W3C Member Submission. URL: http://vmdbpedia.informatik.uni-leipzig.de/temporary/html/dataid-submission-pre.html
12. Maali F., Erickson J., Archer P.: Data Catalog Vocabulary (DCAT). W3C recommendation, The World Wide Web Consortium (2014)
13. Lebo T., Sahoo S., McGuinness D., W3C: PROV-O: The PROV Ontology. W3C Recommendation. URL: https://www.w3.org/TR/prov-o/ (2013)
14. Loscio B. F., Stephan E., Purohit S., W3C: Data on the Web Best Practices: Dataset Usage Vocabulary. W3C Note. URL: https://www.w3.org/TR/vocab-duv/ (2016)
15. W3C: ODRL Comm. Group, https://www.w3.org/community/odrl/, last acc. 28/06/2017
16. W3C: CC Profile, https://www.w3.org/community/odrl/work/cc/, last accessed 28/06/2017
17. W3C: ODRL Linked Data Profile, https://www.w3.org/community/odrl/wiki/ODRL_Linked_Data_Profile, last accessed 28/06/2017
18. W3C: Profile or Recommendation, https://www.w3.org/2012/09/odrl/archive/odrl.net/Profiles/prof-rec-template.html, last accessed 28/06/2017
19. Manghi P. et al: An Infrastructure for Managing EC Funded Research Output - The OpenAIRE Project. The Grey Journal: An International Journal on Grey Literature 6.1 (2010)

DPRL vocabulary, tables, figures, ODRL template and more information available at http://purl.org/adaptcentre/openscience/projects/CDMM/DPRL