Integrating the Management of Personal Data Protection and Open Science with Research Ethics

David Lewis
ADAPT Centre
Trinity College Dublin
Ireland
dave.lewis
@adaptcentre.ie

Joss Moorkens
ADAPT Centre/SALIS
Dublin City University
Ireland
joss.moorkens
@dcu.ie

Kaniz Fatema
ADAPT Centre
Trinity College Dublin
Ireland
kaniz.fatema
@scss.tcd.ie

Abstract

This paper examines the impact of the EU General Data Protection Regulation, in the context of the requirement from many research funders to provide open access research data, on current practices in Language Technology Research. We analyse the challenges that arise and the opportunities to address many of them through the use of existing open data practices for sharing language research data. We discuss the impact of this also on current practice in academic and industrial research ethics.

1 Introduction

Language Technology (LT) research is facing an unprecedented confluence of issues in the management of experimental data. The EU's adoption of the General Data Protection Regulation (GDPR) (European Parliament and Council of the European Union, 2016) imposes new requirements for tracking informed consent for the usage of personal data that may impact all European LT research significantly. National guidelines now need to be established on how GDPR applies to scientific data, and given the large penalties involved, this uncertainty presents significant institutional risk for those undertaking research with the unanonymised or unanonymisable data often needed in LT research.

In addition, the European Commission (EC) and other research funding bodies increasingly encourage open science practices. The aim is to publish research data alongside research papers in order to reduce the cost of obtaining research data and improve the repeatability, replicability and reproducibility of research. While this is a positive move for the quality and integrity of LT research, it must respect the needs of data protection legislation, including different EU member states' implementation of GDPR, and the data protection regimes in jurisdictions outside the EU. These may greatly complicate and delay the benefits of open science policies. This paper reviews these trends and aims to distil the issues that researcher institutes as well as national and transnational research bodies need to face in the coming years to effectively manage research data amid these parallel and sometime conflicting needs. In particular, we highlight the interdependency between these issues and how those who manage research ethics will need to react.

2 GDPR and LT Research Data

As Hovy and Spruit (2016) point out, language data contains latent characteristics of the person producing it, and language technology therefore has the inherent potential to expose personal characteristics of the individual. Coulthard (2000) notes that identification of authors is very difficult from linguistic data alone, but has been successful when accompanied by metadata "information which massively restricts the number of possible authors". This presents a distinct data protection challenge for the sharing and reuse of language resources as they are difficult to reliably anonymise and in some cases can already be used as a biometric.

As the sharing of language resources is an established feature of LT research internationally we must carefully examine the provisions coming into force in the EU with the introduction of GDPR. As an example we can consider research conducted into the productivity changes to translator practice resulting from the use of LT. Translation memory (TM) data is often used for MT training, although identifying metadata is almost always re-

Proceedings of the First Workshop on Ethics in Natural Language Processing, pages 60–65,
Valencia, Spain, April 4th, 2017.

moved beforehand. Measures to retain the metadata in order to strengthen copyright claims in respect of translators, as suggested by Moorkens et al. (2016), would create a risk of data breach under the terms of GDPR. This means that one possibility for extending human translator earnings will almost definitely become an impossibility for creators of MT systems. Machine translation (MT) is another popular LT technique use in translation practice. The impact of MT is being increasingly assessed through detailed analyses of keystroke logs of translators making corrections to such translations. These logs may also be published to accompany such studies (Carl, 2012), but are known at the level of keystroke timings to possess biometric signals that can identify the translator. Another growing practice is translation dictation using automated speech translation. Here, repeatable studies may involve the sharing of recordings and transcripts of spoken translation, where again speech recording could be used to identify the speaker.

2.1 What is the GDPR?

The GDPR is an EU Regulation, adopted in April 2016 and due to come into force in May 2018. It addresses protection of people with regard to the processing and free movement of personal data, replacing the 1995 Data Protection Directive. The GDPR (Article 4) defines Personal Data as "information relating to an identified or identifiable natural person", who it refers to as a Data Subject. An identifier for the Data Subject may be a name, an identification number, location data, an online identifier or "one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity" of the Data Subject. It is these latter factors that perhaps lie latent in language resources and data that is increasingly subject to analysis by language technologies, such as samples of utterances from specific data subjects.

The organisation that collects and uses personal information is the Data Controller, and bears the primary responsibility for implementing the provisions of GDPR. This role would be conducted by research institute which will be responsible for GDPR on many forms of personal data (e.g. student, staff and alumni records) beyond that generated by research. Other organisations in receipt of personal data from a controller is known as a Data Processor, and in LT research this would correspond to other research organisations receiving and reusing research data from the controller. The potential penalties for a data breach fall within two categories with differing maximum fines. A fine of up to 20 million or 4% of turnover (whichever is greater) may be imposed for failure to adhere to basic principles for processing, including conditions for consent (Articles 5, 6, 7 and 9), infringing on the rights of Data Subjects (Articles 20, 21, 22), or improper international data transfers (Articles 44-49). Other failures to comply, such as by failing to obtain proper consent for childrens' data, to keep sufficient records, or to apply proper safeguards, may result in a fine of up to 10 million or 2% of turnover (whichever is greater). Both Data Controllers and Data Processors may be considered liable for the security of personal data, and any data breach must be reported within 72 hours.

The GDPR explicitly encompasses pseudonymised data, which would require additional information (stored separately) to identify the Data Subject. This would include, for example, TM data with a translation unit ID that can be attributed to an individual. Personal data should be retained for a period no longer than is necessary to accomplish the purpose for which it was collected. However long-term archiving is permitted if this is in the public interest for scientific and historical research purposes, or statistical purposes, providing that there are some safeguards. These exemptions for research aim to reconcile privacy with data-driven innovation and the public good that may result. The GDPR states that the designation of scientific research should be "interpreted in a broad manner" including technological development, fundamental research, applied research and privately funded research. Importantly, therefore, consideration of GDPR exemptions for LT research may have widespread implications for industry as well as for academia. GDPR might not apply to data processing where the focus is not on "personal data, but aggregate data" and is statistical rather than referring to a particular individual. Where personal data is processed however, separate consents are required for different processing activities. However, providing that safeguards are implemented, secondary processing and processing of sensitive categories of data may be permitted for research purposes where data

has been collected lawfully. Article 89, which addresses exemptions for scientific research that, states these safeguards should include technical and organisational measures to protect data, following the principle of data minimisation, and may include pseudonymisation or anonymisation where possible or appropriate. As discussed above however, these mechanisms may not be adequate for protecting data subject identity in the sharing and reuse of language resources. Significantly, the precise nature of the safeguard required by Article 89 are left for EU member states to legislate on (Beth Thompson, 2016). So while this enables interpretation of GDPR that aligns with existing national standards for research data, different interpretations may impede efforts to share and reuse experimental data internationally if differing GDPR enforcement regimes emerge.

3 Requirements of Open Science

The requirement for open access research 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 (of European Research Universities, 2013).

Major research funders, including the EC, have widened their guidelines on open science to now address open research data (European Commission, 2016). 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 in order to realise economic and social benefits; and involve citizens and society. It is anticipated that 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 has to be provided within the constraints of EU and national data regulations, now including GDPR. Initiatives such as the Open Access Infrastructure for Research in Europe (OpenAIRE)[1] provide additional information and support on linking publications to underlying research data, and is developing open interfaces for exchange between research data repositories. However, for such open access to work at scale, improved level of interoperability will be required for the meta-data associated with data sets made availalbe through different institutional research data repositories. Such meta-data interoperability is needed to support the aggregation, indexing and searching of experimental data from different repositories so that researchers can find suitable data with less effort. Further, reflecting data protection and research ethics properties in such meta-data will also reduce the effort required to ensure that reusing experimental data from another source does not incur data protection compliance risks.

3.1 Open Data for Open Science

In parallel to other initiatives, Linked Open Data based on open data standards of the World Wide Web Consortium is being adopted as a common means for sharing all types of data between organisations, with strong uptake reported in the public sector. Linked Open Data is based upon the principle of interlinking resources and data with standardised Resource Description Framework (RDF) and Uniform Resource Identifiers (URIs) that can be read and queried by machines through powerful standardised querying mechanisms (Bizer et al., 2009).

Open RDF-based data vocabularies such as DCAT [2] help in expressing authorship of research data sets, while ODRL vocabulary[3] can express usage rights and licensing. The provenance of an experiment, in terms of which people and programmes performed which actions on which resources at what time, can be captured and modelled using the PROV[4] family of data vocabularies. Garijo et al. (2014) build on these standards to propose an open data format for recording both the sequence of experimental steps and the data resources passing between them. This would allow the publication and discovery of experimental descriptions with specific metadata (such as usage rights or data subject consent) associated with specific data elements.

Experiential description using these open vocabularies can be collected or aggregated to form linked repositories such as those supported by OpenAire and Linghub[5], which are being piloted

[1] https://www.openaire.eu/

[2] http://www.w3.org/TR/vocab-dcat/

[3] http://www.w3.org/TR/odrl/

[4] http://www.w3.org/TR/prov-0/

[5] http://linghub.lider-project.eu

for language resources. Existing research for these standard vocabularies has provided best practice for publishing data sets' metadata as linked open data (Brummer et al., 2014). The machine readable nature of metadata can make it easy for an automated system to verify the correctness of the data, or perform other operations such as checking of data formats, completeness of metadata and the provenance of data used (Freudenberg et al., 2016). This approach is amenable to extension with domain-specific experimental metadata, such as the machine learning metadata proposed in the MEX vocabulary (Esteves et al., 2015). The LT research community has already developed a schema, termed META-SHARE (Piperidis, 2012), for language resource metadata that shares many characteristics with the OpenAire scheme. The META-SHARE schema has also been mapped onto RDF with relevant attributes mapped to specific properties from the standard vocabularies previously mentioned, and is used by LingHub as an aggregation source (McCrae et al., 2015).

4 Discussion

Combining the emerging imperative of GDPR compliance and data science poses the following challenges for organisations undertaking LT research and concerned with research ethics. Firstly the encouragement by funders for researchers to provide open access to experimental data must be tempered by the overriding legal requirements of GDPR compliance. While GDPR offers derogation of certain rights when dealing with personal data for the purposes of scientific research, this does not remove the obligation for research performing organisations in the EU to demonstrate their conformance to GDPR, to conduct data protection impact assessments, and to ensure that the appropriate safeguards for the derogation are in place, especially in cases where anonymisation of experimental data is not possible.

In GDPR terms, a data processor receiving data from an LT experiment will need to know the terms of consent agreed to by the data subject in giving the primary data collector permission to use their personal data for a stated purpose. This will enable the receiving party to assess whether the purpose to which they now intend to put the data is compatible with that consent. Given that information, the receiving data processor would also need to give an undertaking that it will only perform processing of that data for purposes that are compatible with that consent.

This will mean therefore that exchange of LT research data with latent personal features cannot proceed without an appropriate contract on the usage of this data being signed and recorded for GDPR compliance purposes. It should also include an undertaking by the data processor not to attempt the identification of natural persons from the data, including through analysis in aggregation with other data. This goes beyond the standard form license agreements already in place for reuse of language resources, e.g. the META-SHARE licences [6], which focus mostly on issues of copyright ownership and usage conditions related to attribution or, in some cases, to compensation. To avoid GDPR unduly impeding the sharing and reuse of experimental data, we recommend that bodies such as the EC and ELRA develop standard form contract terms for the reuse of research data that the LT research community can use in documenting this aspect of GDPR compliance.

GDPR, in Recital 33, acknowledges that it "is often not possible to fully identify the purpose of personal data processing for scientific research purposes at the time of collection". It allows data subjects to provide consent to only specific parts of a research activity "when in keeping with recognised ethical standards for scientific research". This highlights the fact that good practice in research ethics *already* incorporates many features now formalised in GDPR, i.e. the need for a clear explanation of the data collection and processing purposes; the explicit gathering of informed consent and the option of the data subject to withdraw from any part of the research activity at any time. If EU proposals for ePrivacy (European Parliament and Council of the European Union, 2017) move on to become a regulation, the data subject will have more control over whether data may be repurposed by being offered "the option to prevent third parties from storing information on the terminal equipment of an end user or processing information already stored on that equipment" (Article 10). Key to data subjects exercising such control over the processing of personal data is their full understanding of the scientific research purposes to which their data will be subject.

Further research is required to assess the comprehensibility of plain language descriptions of

[6] http://wizard.elda.org/

purpose typically used by researchers for data subjects. The META-SHARE schema, for example, supports a 'purpose' attribute, but it is populated with names of different areas of LT research that are unlikely to be accessible to data subjects. Further, more applied research, perhaps conducted by industry, may be conducted with the known intention of supporting new service features, e.g. personalisation, targeted marketing, or differential pricing. As these are of direct concern to the data subject, such intentions should not be concealed by statements of purpose related to the broader generation of knowledge when seeking informed consent. From such research, the LT community and research institutes should seek to find classifications of purpose that are both accessible to data subjects and convey the differences in purpose of basic and applied research. Current rules and practices on academic research ethics tend to vary from institution to institution, with the intention of protecting participants and researchers by making clear the purpose of data collection, and requesting explicit consent to use personal data for that purpose. Researchers may have to make an undertaking with regard to data protection, but there is rarely any follow-up to ascertain whether the data has been stored or destroyed as promised. In contrast, GDPR compliance will require rigorous organisational and technical systems for record keeping and tracking the use to which data is put by data processors, including data transfer to processors in other institutes and other jurisdictions. Further, much LT research data processing involves secondary processing of industrial data, such as TMs or glossaries. As these are not collected from directly from experimental data subjects but via industrial processes, this data is collected, stored, retained, and shared without a reliable trace of research ethics clearance. Further, as LT research is increasingly undertaken by large companies with access to vast data-sets of customer information, the resulting experimental data is typically not subject to the *a priori* scrutiny of institutional review boards or ethics committees as is common with publicly funded research. This disparity between public and private norms for undertaking research ethics may create barriers to research collaboration and impede the progression of reproducable research results into the public domain. An opportunity therefore exists for the LT research community to better leverage open data standards tracking the transfer and use of personal data in a way that can support GDPR compliance checking. Use of open data standards that capture the detail of data processing workflows may be annotated to better record the processes by which: informed consent is gathered from individual data subjects; their individual objections to specific uses of personal data is handled and the purposes to which personal data is put is audited. Consent must be first collected, then stored and processed for checking compliance with data processing.

Consent can be modified. The modification can be initiated by the data subject or due to change of context the controller can re-solicit for consent that can lead to modification of consent. Consent can be revoked. After revocation, the data may be archived for the time necessary for research result verification and finally destroyed. Further research is needed on how to annotate open experimental workflow provenance records with details of consent management and its impact on the lifecycle management of the subject's data. A possible benefit of an open data approach, is that it may allow individual institutes to publish the attributes of their differing ethics review processes, allowing collection and analysis of variations that may assist in normalising standards. This will also allow those reusing others' data to be reassured that it was collected under ethical standards with which they are familiar. Ultimately this could result is a simple badge system, similar to that employed for creative commons, that could simplify the selection of LT research data according to the compatibility of research ethics and data protection protocols under which it was produced with those sought by the research hoping to use that data. Design of such a seal for reuse of experimental data could benefit from the work already underway in developing data protection seals [7] given the overlap between research ethics protocols and the informed consent requirements of GDPR.

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.

[7] https://ico.org.uk/for-organisations/improve-your-practices/privacy-seals/

References

Beth Thompson. 2016. Analysis: Research and the general data protection regulation. Technical report, July.

Chris Bizer, Tom Heath, and Martin Hepp. 2009. Special issue on linked data. *International Journal on Semantic Web and Information Systems*, 5(3):1–22.

Martin Brummer, Ciro Baron, Ivan Ermilov, Markus Freudenberg, Dimitris Kontokostas, and Sebastian Hellmann. 2014. Dataid: Towards semantically rich metadata for complex datasets. In *Proceedings of the 10th International Conference on Semantic Systems (SEM'14)*, New York, USA. ACM.

Michael Carl. 2012. Translog-ii: a program for recording user activity data for empirical reading and writing research. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Malcolm Coulthard. 2000. Whose text is it? on the linguistic investigation of authorship. In Srikant Sarangi and Malcolm Coulthard, editor, *Discourse and Social Life*, chapter 15, pages 270–288. Routledge, London.

Diego Esteves, Diego Moussallem, Ciro Baron Neto, Tommaso Soru, Ricardo Usbeck, Markus Ackermann, and Jens Lehmann. 2015. Mex vocabulary: A lightweight interchange format for machine learning experiments. In *Proceedings of the 11th International Conference on Semantic Systems*, SEMANTICS '15, pages 169–176, New York, NY, USA. ACM.

European Commission. 2016. Guidelines on open access to scientific publications and research data in horizon 2020. Technical report, February.

European Parliament and Council of the European Union. 2016. Regulation (eu) 2016/679 of the european parliament and of the council (GDPR). *Official Journal of the European Union*, 119(1):1–88.

European Parliament and Council of the European Union. 2017. Proposal for a regulation concerning the respect for private life and the protection of personal data in electronic communications.

Markus Freudenberg, Martin Brummer, Jessika Rucknagel, Robert Ulrich, Thomas Eckart, Dimitris Kontokostas, and Sebastian Hellmann. 2016. The metadata ecosystem of dataid. In *Special Track on Metadata & Semantics for Open Repositories at 10th International Conference on Metadata and Semantics Research*.

Daniel Garijo, Yolanda Gil, and Oscar Corcho. 2014. Ninth workshop on workflows in support of large-scale science (works). In *Ninth Workshop on Workflows in Support of Large-Scale Science (WORKS),*, New Orleans, USA, Nov. IEEE.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association of Computational Linguistics*, pages 591–598, August.

John P. McCrae, Penny Labropoulou, Jorge Gracia, Marta Villegas, Víctor Rodríguez-Doncel, and Philipp Cimiano, 2015. *One Ontology to Bind Them All: The META-SHARE OWL Ontology for the Interoperability of Linguistic Datasets on the Web*, pages 271–282. Springer International Publishing, Cham.

Joss Moorkens, David Lewis, Wessel Reijers, Eva Vanmassenhove, and Andy Way. 2016. Translation resources and translator disempowerment. In *ETHICA 2016: Workshop on ETHics In Corpus Collection, Annotation and Application*, Portoroz, Slovenia, May.

League of European Research Universities. 2013. Leru roadmap for research data. Technical report, Dec.

Stelios Piperidis. 2012. The meta-share language resources sharing infrastructure: Principles, challenges, solutions. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).