Provenance-Enabled Stewardship of Human Data in the GDPR era

Pinar Alper[0000−0002−2224−0780], Regina Becker, Venkata Satagopam, Christophe Trefois, Valentin Grouès, Jacek Lebioda, and Yohan Jarosz

Luxembourg Centre for Systems Biomedicine, Esch-sur-Alzette L-4362, Luxembourg
{firstname.lastname}@uni.lu
https://wwwen.uni.lu/lcsb

Abstract. Within life-science research the upcoming EU General Data Protection Regulation has a significant operational impact on organisations that use and exchange controlled-access Human Data. One implication of the GDPR is data bookkeeping. In this poster we describe a software tool, the Data Information System (DAISY), designed to record data protection relevant provenance of Human Data held and exchanged by research organisations.

Keywords: GDPR · Human Data · Provenance.

1 Background

1.1 EU General Data Protection Regulation

Today, personal data breach incidents are not only front-page news items, they are events with highly adverse impact on individuals and the society. In this regard, a new EU-level legislation, the General Data Protection Regulation (GDPR) [3], could not have been more timely. GDPR brings increased regulation for organisations utilising personal data. Specifically:

- Organisations are now required to **keep inventory** on the personal data they hold: from where, how and under what legal basis the data was obtained, with whom it has been shared and the nature data use. This data provenance will then serve as the starting point for audits performed by the national Data Protection Authorities.
- Individuals have more rights on their data, such as the right to access, right to deletion and the right to restriction of the use of their data. GDPR also requires that requests for rectification, erasure etc. are passed on to the recipients of the data, which means organisations must have a **fine-grained (subject-level) traceability of the sharing personal data**.
- Organisations are expected to **take data privacy measures at systems' design time**. These include data confidentiality, integrity and availability; data minimisation so that only necessary data attributes are used; storage duration limitations so that data is not kept longer than necessary. Furthermore, GDPR expects **documentation** of security such measures as well as

documentation on systematic assessments of data processing setups in terms of data privacy risks (aka "Data Processing Impact Assessment DPIA").

1.2 Stewardship of Human Data

Scientific Data Stewardship refers to "activities for the long-term care of data" to support scientific reproducibility or to enable data sharing [5]. In the context of life-science research, data collected from living human subjects (aka "Human Data") falls under the scope of GDPR as "sensitive personal data". Often, Human Data is solely collected for research and is kept in a pseudonymized fashion (detached from identifying attributes such as name or address). To ensure data-protection Human Data is typically shared via "controlled-access" data catalogues [4]. The common catalogue workflow starts by a study owner submitting a dataset along with descriptive metadata. The second step is the provisioning of a Data Access Committee (DAC) that will be responsible for assessing requests for this dataset in terms of compliance with ethical standards and legal requirements. Data is then advertised in the catalog. Scientists that seeks controlled-access data are required to make a formal application describing the planned study and data use. In addition to the requirements listed in Section 1, the stewardship of Human Data further brings the following requirements:

- GDPR allows EU countries to have their own legislative provisions. This leads to the requirement to know if the requested type of data processing is allowed in the country of the requester. Also subjects may disallow the transfer of their data outside the country of collection, or outside the EU. Currently catalogues do not model data use restrictions in detail, instead they rely on DACs to match restrictions against requests. Under GDPR, however, data catalogues will be accountable for granted accesses, therefore detailed consent modelling and conflict detection is necessary.
- Catalogues are typically maintained by life-science institutes that also run their own studies, which may involve Human Data. From the perspective of GDPR, all Human Data needs to be accounted for, regardless of it being a frozen data snapshot in the catalogue or an active dataset in the process of being generated. Henceforth common abstractions and tools are needed to keep inventory of Human Data.

Motivated by these observations, we are developing the Data Information System at ELIXIR Luxembourg.

2 ELIXIR-LU Data Information System

ELIXIR [2] is a pan-European infrastructure for life-science data. ELIXIR-LU is the Luxembourgish node of ELIXIR based at the Luxembourg Centre for Systems Biomedicine (LCSB). ELIXIR-LU hosts a Translational Medicine data repository as well as a cloud platform and tools to support data integration, analytics and visualisation. ELIXIR-LU Data Information System (DAISY) is

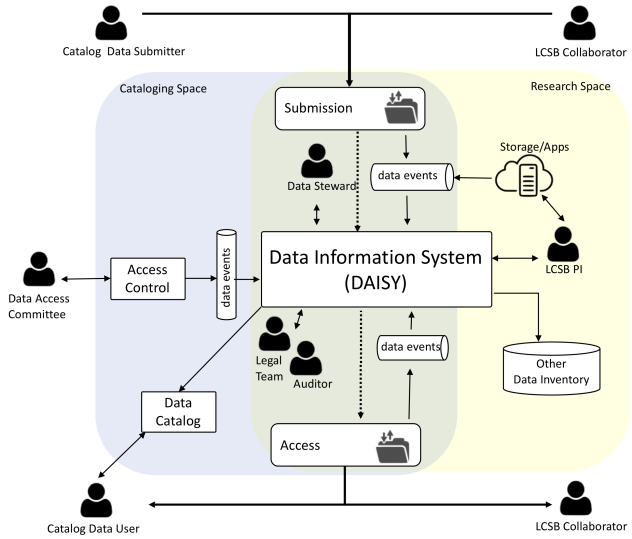

Fig. 1. Data Information System Overview.

a web application that is designed to collect rich provenance on the Human Data held in LCSB for both local research and for the ELIXIR-LU Catalog. Information gets accumulated in DAISY by (1) different stakeholders' manual input and (2) data-events that are generated from loosely coupled applications (depicted in Figure 1).

- Prior to data's physical arrival to LCSB, we ask submitters to fill in a "Data Information Sheet", which collects essential data protection metadata, such as data use restrictions, data's de-identification method and the legal basis for its processing.
- Data submitted for the catalog undergoes further processing, such as re-pseudonymization, where subject identifiers in data are replaced by catalogue accession numbers, also curation may be performed. These alterations and data storage endpoints are recorded in DAISY by the data steward/curator.
- Access control to data in the catalogue is mediated by an application [1] that facilitates the DAC decision. This application is monitored for information on who has been granted access, and for which duration, all captured in DAISY. This information is complemented with logs of transfers of data to authorised catalog users.
- Access to local-research data is controlled via application/file-system level permissions, through monitoring components DAISY can generate a report on who has access to local data at any given time.
- Documents that guarantee data legality, such as Ethics Approvals, Data Sharing agreements, Consent Templates may be renewed or revised. Such updates are facilitated by manual input of the Legal Team into DAISY.

Also data privacy measures can be recorded by IT Administrators as tags on datasets and any technical documentation, such as DPIA results, can be linked via the content management system.
– The provenance stored in DAISY will be exportable in a standards compliant form. This may be upon request by data subject, by auditor or for transfer of information to other inventories.

In addition to **recording provenance** identified above, DAISY will provide the following features:

– Detection and flagging of conflicts between datasets restrictions and the requests made on those datasets.
– Generation of notifications based data use restrictions. For example notifying the end of data storage durations or legal contracts nearing their end/renewal date, or notifications to data recipients to remind their obligations e.g. co-authorship on publications with data.

3 Future Directions

DAISY is currently under development and is scheduled for an alpha release in July 2018. We are establishing a GDPR working group in ELIXIR. Through this group we hope to refine requirements for DAISY and identify its functions that can be re-used by the ELIXIR community. We plan to make DAISY available as an open source tool.

Acknowledgments. This work was (partially) funded through the contribution of the Luxembourg Ministry of Higher Education and Research towards the Luxembourg ELIXIR Node.

References

1. Brandizi, M., Melnichuk, O., et al.: Orchestrating differential data access for translational research: a pilot implementation. BMC Med. Inf. & Decision Making **17**(1), 30:1–30:14 (2017)
2. Crosswell, L.C., Thornton, J.M.: Elixir: a distributed infrastructure for european biological data. Trends in Biotechnology **30**(5), 241 – 242 (2012). https://doi.org/https://doi.org/10.1016/j.tibtech.2012.02.002, http://www.sciencedirect.com/science/article/pii/S0167779912000170
3. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation). Official Journal of the European Union **L119**, 1–88 (May 2016), http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2016:119:TOC
4. Lappalainen, I., Almeida-King, J., et al.: The european genome-phenome archive of human data consented for biomedical research. Nature Genetics **47**(7), 692–695 (2015)
5. Wilkinson, M.D., Dumontier, M., et al.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data **3**, 160018 EP – (03 2016), http://dx.doi.org/10.1038/sdata.2016.18