Data Access and Reproducibility in Privacy Sensitive eScience Domains

Stefan Pröll, Rudolf Mayer
SBA Research, Vienna, Austria

Andreas Rauber
Vienna University of Technology, Austria

***Abstract*—In privacy sensitive eScience domains the disclosure of data is often not allowed or advised if it contains sensitive data about the individual. Applying data protection methods oppose interests of repeatability and reproducibility, as the data which serves as input and output for processing steps in experiments needs to be altered in order to preserve privacy. We thus discuss pre-requisites and methods for protecting the privacy, with the goal of still enabling reproducibility and allowing comparative studies to be performed. We align our discussion on with the experience from a use case in the area of health policy planning, where statistical and mathematical models are trained from health data.**

I. Introduction and Use Case

Repeatability and reproducibility are corner stones of sound research in science disciplines, including computational and eScience domains. A thorough and detailed description of the investigations performed is thus required to achieve these goals. This includes description of the experiment design, that is the computational steps that are performed to achieve the final result, specifically including also the order of steps, and how they are connected and invoked. Scientific workflows have shown to be a useful concept to this end. Workflows facilitate automatic capturing of provenance data, as the data flows between the process steps is often explicitly defined and can thus be easily recorded and stored. This data can be utilised for verifying and analysing experiments.

Further, descriptions building on top of workflows and augmenting the metadata on the experiments have been proposed, such as e.g. the Research Object Model [1]. The technical experiment setup, for example what software and hardware is utilised, including details on the configuration and dependencies, is also an important aspect, but usually not documented sufficiently by workflow systems [2]. Finally, the data that is utilised in a specific execution of a workflow is vital to be able to re-execute an experiment or to enable comparison of results if related analyses are performed. With ever increasing sizes of data sets, often stemming from a multitude of different sources, this is a challenge that is tackled by data citation, and more recently dynamic data citation approaches.

There are certain settings settings where all these efforts to enable repeatability are, however, opposing basic principles of privacy. This is the case when personally identifiable information (PII) is involved, which can frequently be the case especially in life sciences, but also other disciplines. One specifically prominent domain is eHealth, where records describing the medical history of patients are studied. This can for example be in medical diagnosis, or for health care planning purposes.

In this paper we analyse in which phase measures to enhance repeatability and reproducibility are needed, and what their relation to privacy protecting mechanisms is. We discuss the effects and constraints that privacy sensitive data has on data access, data citation, experiment execution, and provenance data generation. We describe both existing solutions, as well as illustrating areas where open research questions are still to be tackled and we outline possible approaches.

The use case we base our considerations on is in the field of eHealth, specifically in the domain of health policy and planning in Austria. The goal of research is to enable informed decisions on the future directions of the health care system by selecting the most appropriate treatments and technologies. Contrary to focused studies that are limited in size, the aim is to use large volumes of routinely collected data from the public health care system. The Austrian health care system is characterised by a mandatory health insurance, thus 98% of the population are covered. The Austrian National Health GAP-DRG database contains records of publicly reimbursed health care events, from general practitioners, in- and out-patients from hospitals, and pharmacies. The database holds information of two years, totalling almost 2.2 billion records, plus another four years of data from the largest of Austria's province (accounting for roughly 22% of the population). In the course of the nationally-funded Austrian project DEXHELPP[1], this data is combined with various data sources provided by other project partners, e.g. census data. Also, routine data from the health care system is periodically updated – additional data for new periods of time is provided, and also corrections of the data from previous periods might be provided. While the data is pseudonymised in most data sources, record linkage approaches can be utilised to identify matches between different data sources, as shown e.g. in [3].

The data is made available to project partners for investigating specific research questions – data access to the involved partners is based on the definition and approval of such a research project. However, not all data sources are going to be available to all partners, as some of them have conflicting interests and backgrounds. Especially access to the raw data is often prohibited. As such, the issues of data access and privacy are slightly different to settings found often in other research settings, where a static export of the data is made available to researchers. Here, we deal with a continuously increasing data set, and the data that is allowed to be used for each research project is potentially different, depending on the project partners and specific type of investigation. We thus rather face the scenario of ad-hoc needs for a specific subset

[1] http://www.dexhelpp.at/

of the databases, where the data still is evolving over the time.

The remainder of this paper is organised as follows: Section II gives an overview of the related work in the area. Section III describes how data access needs to be adapted for enabling privacy in sensitive applications. Section IV then investigates privacy concerns in data citation, and how data citation can alleviate same privacy concerns of data sharing. The paper closes with a conclusion in Section V.

II. Related Work

Data driven science and in-silico experimentation have emerged as a completely new paradigm in many different disciplines [4]. With growing complexity of experiments it becomes increasingly difficult to reproduce the results published in scientific journals and papers [5]. Nevertheless reproducibility is the most important metric for valid research [6] and requires thorough documentation of all steps [7]. Different approaches exist in order to preserve research environments [8] and capturing scientific workflows including software dependencies and additional contextual information of experiments [9].

In [10] six key concepts fundamental for assessing research data are identified: quality, provenance, data extraction and related errors, processing and related errors, traceability of results and curation. Provenance is at the core as it allows increasing the quality by being able to detect extraction and processing errors while providing the knowledge how each record was used during a workflow [11]. For scientific validation, collecting provenance metadata is not enough. What is needed is the actual re-execution of an experiment, which entails that the information about the utilised software and hardware components needs to be preserved [9]. There are, however, also privacy concerns with the metadata gathered for enabling re-execution of the investigation. [12] argues that also the formalised experiment structure and implementation of specific tasks in the workflow can be a threat to privacy, demonstrated in the example of a disease susceptibility workflow. For reproducibility and reuse of experiments and results, researchers need to share their workflows and their data sets, of which often several versions exist. Therefore, data citation methods that allow assigning persistent identifiers to workflows and specific subsets of data are required [13].

Two methods are available which allows identifying data and therefore detect the source of a leak: Fingerprinting and watermarking allow identifying a specific copy and therefore detect the source of a leak, and are currently mainly used for detecting pirated multimedia content [14]. Watermarks are used for identifying the content owner, whereas fingerprints are individualised watermarks [15]. Both methods have also been applied to relational databases [16]. Strategies for privacy in the data management tool i2b2 is analysed in [17]. A set of privacy levels, which allow access to the data is granted based on individual trust.

III. Privacy Aware Data Handling

Privacy is a fundamental requirement for eScience research, it needs to be ensured that individuals can not be identified from the any material used in or published on the experiments. Anonymisation deals with either encryption or removal of personally identifiable information, to hinder unintended disclosure of information on individuals. Pseudonymisation provides a compromise between full anonymisation and handling raw private data. In contrast to anonymisation, identifiers are not removed but replaced with a pseudonym, which is an artificial identifier. Quasi identifiers, which are pieces of information that by themselves are not uniquely identifying a record, but might do so if combined with other information, may still remain in the data set [18].

A. Data Access

In our use case, researcher need to be able to create individual subsets of the data based in their requirements in an ad-hoc manner. As the database is continuously growing and the data may not be distributed by law, creating and sharing snapshots is not a valid approach. Also producing the subsets which fulfil both the privacy requirements and the demands of the researcher is a challenge. Handling the sub-setting process manually by entrusted data experts is costly and does not scale well. For this reason researchers need means to create their data sets on their own.

Instead of directly granting researchers access to the data sources, which in our example are in the form of a relational database, the researcher are to be provided with a front-end that allows control over the results of the queries. This further reduces the complexity of interacting with the system. We can thus protect privacy sensitive individual data records from unintentional access by sanitising the query results in a way that the data does not reveal information on individuals, instead of allowing the researcher direct access to query individual records. In our application scenario, we realise the data access approach by utilising the i2b2 (Informatics for Integrating Biology and the Bedside) software platform[2] [19]. i2b2 is frequently used for research on clinical health data [20].

B. K-Anonymity

Even after removing information that uniquely identifies individuals from a data set, de-anonymisation approaches cross-referencing information from multiple data sets have shown to be successful in revealing the identity of individual data records [18]. Thus, to prevent the identification of individuals, [18] introduced the concept of k-anonymity, which ensures that for each subset taken from the database, each record shares the same attributes with at least $k-1$ other data samples. Thus, it becomes impossible to distinguish between these records, and linking them with other databases becomes more difficult. This requires a modification of the results of the query, by generalising attribute values to achieve the k-anonymity desired, or by suppressing the value altogether [21]. Generalisation is achieved by replacing a value with a more general value that is still semantically correct, along a defined a generalisation hierarchy. For a specific relation, a number of potential generalisations exist, the k-minimal generalisation being the one that is the least generalised. If multiple such generalisations exists, the minimal distortion of a relation can be chosen as a preference criterion.

[2]https://www.i2b2.org/software/index.html

C. Watermarking and Fingerprinting

Data constitutes an intellectual asset of unique value, especially when sensitive data is processed and the disclosure of which can have serious consequences for individuals. Therefore the source of a data leakage must be identifiable and be made accountable. A watermark introduces controlled but meaningless change into a database and therefore allows the detection of a leak. There is a trade-off between the strength of the watermark and the quality of data [22] – the more records need to be changed, the higher is the probability of the watermark to be detectable. An overview of database watermarking and the available schemes is given in [23]. In our use case, we need to watermark individual subsets for detecting data leaks and have evidence that the data was illegitimately retrieved from a malicious source. For numerical data, the deviation introduced by the watermark can be controlled and remain within the specified boundaries. For textual and categorical data, watermarking is more difficult, as the change can significantly falsify results. The authors of [22] introduce a scheme where the watermark only performs changes to values from a valid domain. Fingerprints are an extension to watermarking [24], and are generated unambiguously for each user obtaining a data export. Thus, it is possible to attribute the specific source of a data leakage.

The measures for protecting privacy have to be taken before the researchers obtains the actual data export.The k-anonymisation and fingerprinting are thus ideally integrated in the data access platform, such as i2b2; we are currently in the process of implementation. To prevent undesired artefacts in some of the attributes, domain experts will be able to restrict columns from being altered for fingerprint generation. One further requirement is that these procedure needs to be repeatable, in case the study needs to be evaluated and re-executed. Thus, the same procedures for enabling repeatability as with the actual experiment computation need to be applied in the data anonymisation and export step, i.e. documentation of the algorithms, parameters and software used.

IV. Privacy Aware Dynamic Data Citation

Data citation is primarily used for referencing research data and providing long term access via persistent identifiers and institutionalised data retention. Data citation provides access to evidence by utilising persistent and unique identifiers, supported by verification and attribution metadata[3]. Peers can resolve the identifier and are usually referred to a landing page, which contains metadata about the data set and the possibility to retrieve the data for inspection. Although there are efforts for promoting open access in many scientific areas, this does not necessarily imply that the data has to be open and accessible for all users [25]. This itself is already one step towards protecting privacy of personal information while enabling reproducibility – when there is a system that provides controlled access to the data that forms the basis of a study, access can be specifically granted to peers wanting to review or repeat the study, without requiring to publish the data in the open.

Until recently, data citation has considered mostly static files which reside in a repository. However, the growing size and complexity of data sets created a demand for a more flexible approach. Dynamic data citation[4] allows to reference user specific data sets or subsets, also in databases that have changes in content over time. The authors of [13] implemented a framework for relational database management systems, which attaches persistent identifiers to queries, instead of an exported data set. Versioning and timestamping the data allows retrieving the result set at a later point of time, as it was seen at the time of the original query execution. To this end, the so-called query store records queries and metadata such as timing information, as well as hash keys for result set verification for future re-executions. Thus, one can emulate the state of the database as it was for the original query, instead of having to keep a copy of each exported data set separately. The approach can also be applied to other data structures which fulfil certain requirements, such as a query language.

In this paper we discuss how the dynamic data citation framework can be expanded by privacy preserving capabilities. Integrating the k-anonymity and fingerprinting modifications to the result set directly, as optional functionality, into the data citation framework is the most obvious approach. We thus adapted the query store concept into an *execution store*, and extended its functionality for this purpose.

For supporting the reproducibility of privacy enabling measures, we need to store additional metadata that describes how these modifications controlled the final result set. Additionally, the re-execution of a query also requires to maintain the permission rights of the records, as access to a specific portion of the data may be granted or revoked. Thus the system can refuse access to previously available sources if policies change. The query metadata, anonymisation, access and permission policies are stored and maintained and consulted upon re-execution. The source database needs to be adapted and augmented with privacy relevant metadata. Therefore each table which contains potentially personally identifiable information needs to be marked as confidential, so that each query issued against such a table can fulfil the predefined privacy requirements.

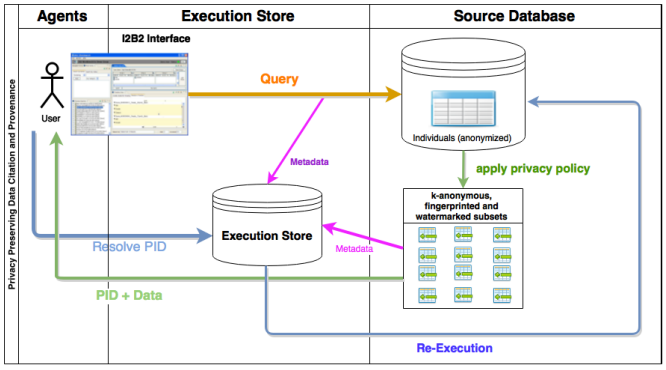

Fig. 1. The Execution Store

Figure 1 shows a schematic overview of the process. The user generates a query to select the data needed for an experiment, which is passed on to and analysed by the data citation module. All properties of the query are stored together with a timestamp and the user details within the execution store. The system then normalises the query and calculates a

[3]https://www.force11.org/datacitation

[4]https://www.rd-alliance.org/group/data-citation-wg.html

hash sum of the query parameters, to detect whether or not the query has been sent before. If the query is already known to the system, the data is checked for the most recent update, which is supported by the timestamps in the versioned database. If no update was detected, the system immediately responds with the already known PID, and executes the query and applies the same privacy protecting measures as before.

If an update was detected, or the query is new altogether, a new identifier is assigned. Hierarchical PIDs allow to map relationships between query executions and therefore describe the lineage of data sets. Once the query is executed, k-anonymisation and fingerprinting are applied, and the necessary information to repeat these steps on future queries is stored in the execution store. Finally, the execution of the query and the user details are being stored persistently in the system.

Data citation enables reproducibility without having to openly publish all the data. But even if the data is intended to be published, because it is in the most cases aggregated data, and has been treated to fulfil k-anonymity and fingerprinting requirements, publication is much less prone to privacy infringement threats than when releasing a raw data set.

V. Conclusions and Future Work

Reproducibility as a key factor of scientific endeavours partially contradicts privacy precautions, which hide information and thus increase complexity in an already challenging task. In this paper we thus discussed issues and threats for privacy in eScience experiments employing sensitive data. Driven by a concrete use case in eHealth, we considered holistically the research life cycle, and analysed where privacy and reproducibility are conflicting goals, and how these contrary aims can be reconciled. We focused on three areas which are crucial for understanding how an experiment was conducted: data access, data exchange and provenance. We thus designed to integrate measures to ensure anonymity, the ability to detect and attribute data leakages, with the dynamic data citation framework, thus enabling repeatable and reproducible access to subsets of the sensitive data. Future work will focus on implementing these designs, as an extension to the i2b2 health data management system.

Acknowledgement

Part of this work was co-funded by the research project DEXHELPP, supported by BMVIT, BMWFW and the state of Vienna, and COMET K1, FFG - Austrian Research Promotion Agency.

References

[1] K. Belhajjame, O. Corcho, D. Garijo, et. al, "Workflow-centric research objects: First class citizens in scholarly discourse," in *Proceedings of Workshop on the Semantic Publishing (SePublica 2012), 9th Extended Semantic Web Conference*, May 28 2012.

[2] R. Mayer, T. Miksa, and A. Rauber, "Ontologies for describing the context of scientific experiment processes," in *Proceedings of the 10th International Conference on e-Science*, Guarujá, Brazil, 2014.

[3] H. Katschnig, F. Endel, and G. Endel, "Depression and pathways of health services utilization in austria: A record linkage study for the total population," in *Proceedings of the 4th International Conference Exploiting existing data for health research*, 28–30 August 2013.

[4] T. Hey, S. Tansley, and K. Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.

[5] J. P. Mesirov, "Computer science: Accessible reproducible research," *Science*, vol. 327, no. 5964, 2010.

[6] J. Loscalzo, "Irreproducible experimental results causes,(mis) interpretations, and consequences," *Circulation*, vol. 125, no. 10, 2012.

[7] M. Schwab, M. Karrenbach, and J. Claerbout, "Making scientific computations reproducible," *Computing in Science & Engineering*, vol. 2, no. 6, pp. 61–67, 2000.

[8] J. T. Dudley and A. J. Butte, "Reproducible in silico research in the era of cloud computing," *Nature biotechnology*, vol. 28, no. 11, 2010.

[9] R. Mayer, G. Antunes, A. Caetano, M. Bakhshandeh, A. Rauber, and J. Borbinha, "Using ontologies to capture the semantics of a (business) process for digital preservation," *International Journal of Digital Libraries (IJDL)*, vol. 15, pp. 129–152, April 2015.

[10] S. De Lusignan, S. Liaw, P. Krause, V. Curcin, M. Vicente, G. Michalakidis, L. Agreus, P. Leysen, N. Shaw, K. Mendis *et al.*, "Key concepts to assess the readiness of data for international research: Data quality, lineage and provenance, extraction and processing errors, traceability, and curation," *Yearb Med Inform*, vol. 6, no. 1, pp. 112–20, 2011.

[11] P. Missier, S. Woodman, H. Hiden, and P. Watson, "Provenance and data differencing for workflow reproducibility analysis," *CoRR*, vol. abs/1406.0905, 2014.

[12] S. B. Davidson, S. Khanna, V. Tannen, S. Roy, Y. Chen, T. Milo, and J. Stoyanovich, "Enabling privacy in provenance-aware workflow systems," in *5th Conference on Innovative Data Systems Research (CIDR)*, Asilomar, CA, USA, 2011.

[13] S. Pröll and A. Rauber, "Data Citation in Dynamic, Large Databases: Model and Reference Implementation," in *IEEE International Conference on Big Data 2013*, Santa Clara, CA, USA, October 2013.

[14] S. Craver, N. Memon, B.-L. Yeo, and M. M. Yeung, "Resolving rightful ownerships with invisible watermarking techniques: Limitations, attacks, and implications," *Selected Areas in Communications, IEEE Journal on*, vol. 16, no. 4, pp. 573–586, 1998.

[15] R. Halder, S. Pal, and A. Cortesi, "Watermarking techniques for relational databases: Survey, classification and comparison." *J. UCS*, vol. 16, no. 21, pp. 3164–3190, 2010.

[16] R. Agrawal and J. Kiernan, "Watermarking relational databases," in *Proceedings of the 28th international conference on Very Large Data Bases*. VLDB Endowment, 2002, pp. 155–166.

[17] S. N. Murphy, V. Gainer, M. Mendis, S. Churchill, and I. Kohane, "Strategies for maintaining patient privacy in i2b2," *Journal American Medical Informatics Association*, vol. 18, no. Supplement 1, 2011.

[18] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[19] S. N. Murphy, M. Mendis, K. Hackett, R. Kuttan, W. Pan, L. C. Phillips, V. Gainer, D. Berkowicz, J. P. Glaser, I. Kohane *et al.*, "Architecture of the open-source clinical research chart from informatics for integrating biology and the bedside," in *AMIA annual symposium proceedings*. American Medical Informatics Association, 2007, p. 548.

[20] G. M. Weber, S. N. Murphy, A. J. McMurry, D. MacFadden, D. J. Nigrin, S. Churchill, and I. S. Kohane, "The shared health research information network (shrine): a prototype federated query tool for clinical data repositories," *Journal of the American Medical Informatics Association*, vol. 16, no. 5, pp. 624–630, 2009.

[21] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 571–588, 2002.

[22] R. Sion, "Proving ownership over categorical data," in *Proceedings. 20th International Conference on Data Engineering*, March 2004.

[23] Y. Li, "Database watermarking: A systematic view," *Handbook of Database Security: Applications and Trends*, p. 329, 2007.

[24] Y. Li, V. Swarup, and S. Jajodia, "Constructing a virtual primary key for fingerprinting relational data," in *Proceedings of the 3rd ACM Workshop on Digital Rights Management*. New York, NY, USA: ACM, 2003.

[25] S. Callaghan, S. Donegan, S. Pepler, M. Thorley, N. Cunningham, P. Kirsch, L. Ault, P. Bell, R. Bowie, A. Leadbetter *et al.*, "Making data a first class scientific output: Data citation and publication by NERC's environmental data centres," *Int. Journal of Digital Curation*, 2012.