Data Access and Reproducibility in Privacy Sensitive eScience Domains

Stefan Pröll
SBA Research
Vienna, Austria
sproell@sba-research.org

Rudolf Mayer
SBA Research
Vienna, Austria
rmayer@sba-research.org

Andreas Rauber
Technical University of Vienna
Vienna, Austria
rauber@ifs.tuwien.ac.at

***Abstract*—In privacy sensitive eScience domains, the data forming the basis for investigations is attributable for example to individuals. However, the disclosure of such data is often not allowed or advised if it contains sensitive data about the individual. Thus special attention needs to be paid when conducting eScience experiments, so that such data is not accessed in unauthorised ways. This affects the data in original or transformed forms, if the latter still allows deduction of information on individuals. Such concerns are opposing interests of repeatability and reproducibility, where the input data and traces of experiment executions form an important aspect to enable such goals. In this paper, we present a use case in the area of health policy planning, where statistical and mathematical models are trained from routine data health data, which contains privacy sensitive information. We thus discuss requirements for protecting the privacy, with the goal of still enabling repeatability and reproducibility.**

I. Introduction and Motivation

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, for example utilising the Taverna Workflow Engine [1]. 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 [2]. Workflows structure the execution of processes and allow the automation of computational tasks to a high degree. It also facilitates automatic capturing of provenance data, as the data flows between the process steps are often explicitly defined and can thus be easily recorded and stored. This provenance data can be well utilised for verifying and analysing experiments.

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 often not covered sufficiently by workflow systems [3]. Finally, the data that is utilised in a specific execution of a workflow is vital to be able to re-execute an experiment. 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 dynamical 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 employed. This can for example be in medical diagnosis, or for health care planning purposes.

Privacy is of concern not only for the original input data, but for also the provenance traces generated during experiment runs, or even the experiment design and implementation. In this paper, we therefore take a comprehensive look at the lifecycle of an eScience experiment, starting from data provisioning, the execution of the experiment, up to the output data generated. We analyse in which phase measures to enhance repeatability and reproducibility are needed, and what their relation to privacy protecting mechanisms is. We therefore discuss in this paper 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. For some areas, we also outline possible approaches.

As a guidance during this analysis, we utilise our experience from a project we are currently running in the domain of health policy planning in Austria [1], where we have access to a real scenario, and are able to gather requirements and insights from various stakeholders, starting from health insurance and policy organisations, health data providers, to data scientists.

The remainder of this paper is organised as follows. Section III gives and overview of the related work in the area. Section II will then describe our framing use case from the eHealth domain. Then, Section IV describes how data access needs to be adapted for enabling privacy in sensitive applications. Section V then shows how data citation provides evidence for experiment runs, and which privacy concerns are to be considered there. Further, we discuss how data citation can alleviate same privacy concerns of data sharing. In Section VI, we discuss what issues are to be considered when providing meta-data on the execution platform, to enable the technical re-execution of an experiment. Section VII provides a discussion on provenance data, and then describes an adaptation of the data citation model to allow for privacy-aware citation of provenance data, without needing to release the data in the

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

open. The paper closes with a conclusion in Section VIII.

II. Use Case

The specific scenario we are investigating in this paper is in the field of eHealth, specifically in the domain of health policy and planning in Austria. The goal of research in this discipline is to enable informed decisions on the future directions of the health care system, by selecting the most appropriate treatments and technologies. Contrary to empirical studies that are limited in size, the aim is to use large volumes of routinely collected health care data.

The Austrian health care system is characterised by a mandatory health insurance, thus 98% of the population are covered [4]. The Austrian National Health GAP-DRG database is a database containing record of publicly reimbursed health care events, e.g. from general practitioners, in- and out-patients from hospitals, and pharmacies. As this data is mainly collected for other tasks than for research, secondary utilisation of data implies also disadvantages. Drawbacks involve mainly restricted number of variables, missing details, inadequate documentation, and data quality. For example, the data does include only information on the health care service performed, but not any disease or diagnoses information. Initially the database contains information of two years, totalling almost 2.2 billion records. In a second iteration of the database, data from the largest of Austria's province (accounting for roughly 22% of the population) has been provided over a time-period of 4 years.

In the course of the Austrian project DEXHELPP [2], 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 in 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 [4].

The data is then 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 of the data bases.

III. Related Work

In the recent years data driven science and in-silico experimentation have produced remarkable results and constituted e-Science as a completely new paradigm in many different disciplines [5]. With growing complexity of experiments it becomes increasingly difficult to reproduce the results published in scientific journals and papers [6], [7]. Nevertheless reproducibility is the most important metric for valid research [8] and requires thorough documentation of all steps [9], [10]. Different approaches exist in order to preserve research environments [11] and capturing whole scientific workflows including all software dependencies and additional contextual information of experiments [12], [13].

For understanding how an experiment processed research data, transformed the data within a workflow and eventually provides the results, we need to understand the lineage of the data including all computational steps which were dependent on it. Provenance data is metadata which describes the lineage or evolution of data. It is used for denoting the sequences of steps which have been proceeded and for providing additional information about execution details to a varying degree depending on the environment and its requirements. For this reason provenance data is an essential building block for reproducible experiments [14] and constitutes evidence for the execution of research workflows and their internal data exchange.

The authors of [15] identify six key concepts fundamental for assessing research data: quality, provenance, data extraction and related errors, processing and related errors, traceability of results and curation. These properties are interconnected and influence each other. Provenance is at the core of these principles as it allows increasing the quality by being able to detect extraction and processing erros while providing the knowledge how each record was used during a workflow. In our work we will investigate how privacy concerns influence the metadata collection and their usage. Finally the curation concept enables peers to discover, access and interpret results including the metadata how a result has been obtained. As provenance data supports reproducible and verification of research results, it is contributing to the long term preservation of research experiments [16]. Provenance data can be captured at several levels [17], ranging from low level file sytems solutions to integrated solutions for sophisticated scientific workbench applications [18] and semantic web applications.

A comparison of three views on provenance data is given in [19] and a taxonomy of data provenance approaches in eScience and workflows can be found in [20]. A provenance bibliography has been compiled by [21].

As this data may include sensitive information about individuals, provenance data is considered a risk for disclosing confidential data. For this reason, privacy issues in provenance data have been studied extensively [22], [23]. For increasing the reproducibility and reuse of the experiments and their results, researchers need to share their workflows and their datasets. It is essential for scientists to identify their work for the later reference. This also includes datasets and even workflows. As research is an iterative approach, several versions of a dataset or a workflow may exist. Therefore, researchers require data citation methods which allow them assign persistent identifeirs to workflows and attach specific subsets of data which are needed for verifying an experiment.

Data has gained more and more importance for organizations and therefore data leaks must be preventable and

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

detectable. Two methods are available which allows identifying data and therefore detect the source of a leak: fingerprinting and watermarking. Both methods are currently mainly used for detecting pirated multimedia content [24]. Watermarks are used for identifying the content owner, whereas fingerprints are individualized watermarks [25]. As the importance and value of data grew, both methods have also been applied to relational databases [26], [27].

Strategies for privacy in i2b2 is analysed in [28]. A set of privacy levels, which allow different detailed access to the data sources, is advocated. Users are, based on personal trust, granted access to a certain level.

IV. Privacy Aware Data Handling

While privacy laws are very different depending on the country where the data is hosted, privacy is a fundamental right e.g. in the United Nations "Universal Declaration of Human Rights" (Article 12). A basic necessity, applicable in most settings, is the need for removing the possibility for identification of individuals. Further, this might be extended to also include legal entities. Thus, a fundamental requirement also for eScience research using privacy sensitive data is that individuals can not be identified from the data utilised in the experiments.

Anonymisation deals with either encryption or removal of personally identifiable information, to hinder unintended disclosure of information on individuals. Pseudonymisation this provides a compromise between full anonymisation and handling raw private data. As In contrast to anonymisation, identifiers are not removed but replaced with a pseudonym, which is an artificial identifier. Quasi identifiers [29], 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 [29]–[31].

In this section therefore discuss basic requirements for enabling privacy in our eHealth use case scenario, from requirements for the actual data access, as well as concerns of anonymity and watermarking.

A. Data Access

In our scenario, various different tasks are investigated by the researchers, and thus, there is no single data set that is once created and distributed to them. Instead, there is the need for many customised and ad-hoc datasets that can specifically target the research question. Further, the data base is continuously expanded by fresh transactional data covering new periods of time. Further, as much as possible, the data shall remain on the location of the central repository, and exports shall not be distributed to users. Thus, it becomes rather infeasible to provide periodic snapshots to the researchers. Instead, ad-hoc queries to the data are a requirement. It has to be noted that important information in the database is already pseudonymised.

One can conceive settings where there is a data archivist that performs data access on behalf of the researchers, translating their information needs into queries to the data storage. This way, the access to individual records is limited to a certain group of people, which in the same time could perform checks that the data delivered to the researchers doesn't allow any deduction of privacy infringing information, by the methods discussed below.

However, such settings are likely to be costly, and instead, the researchers would be allowed to perform the data access themselves. In such a scenario, it is thus important to ensure basic requirements for keeping sensitive data protected. Instead of directly granting researchers access to the data sources, which in our example are in the form of a relational database, the researcher shall 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 access directly query individual records.

In our application scenario, we realise the data access approach by utilising the I2B2 software platform[3] [32], [33] (Informatics for Integrating Biology and the Bedside). I2B2 is a data access system built specifically for the integration of clinical data, and released as open source software. It is in wide-spread use for research on clinical health data, as shown e.g. in the SHRINE system [34], which aggregates queries over the federation of several data sources. It allows the users to build queries with assistance of a query builder. Plugins also provide data-warehouse like functionality.

B. K-Anonymity

Even after removing information that uniquely identifies individuals from a data set, de-anonymisation approaches, which try to cross-reference information from multiple data sets, might be successfully reveal the identify of individual data records. For example, [30] showed that even after removing attributes that uniquely identify persons (e.g., the social security number) from medical data, it is possible to identify 87 percent of all Americans based on combining quasi-identifiers [29], such as date of birth, ZIP-code, sex, and combinations of quasi identifiers with external data such as voter records [35]. Thus, to prevent the identification of individuals, [30] introduced a new concept called k-anonymity, which is a widely adopted anonymisation technique in research nowadays.

K-anonymity 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 [31].

Generalisation is achieved by replacing a value with a more general value that is still semantically correct. This is achieved by defining a generalisation hierarchy. For the example of a ZIP code, this might be by removing the last (least significant) digit, for other numerical values it can be e.g. via a data binning. Suppression of a value can be modelled by introducing a new maximal element in the hierarchy. For a specific relation, a number of potential generalisations exist, the k-minimal

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

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. Finding this optimal k-anonymity generalisation is a NP-hard problem, but several heuristics exists. A good overview can be found in [36]-

Besides k-anonymity, l-diversity [37] is an important aspect. Even if k-anonymity is achieved for a group of records, i.e. there are at least k records that have the same values in the quasi-identifiers, if these records then all have the same values in another attribute, e.g. a prescribed pharmaceutical, still information is revealed about a certain person (without the need to identify the actual record). L-diversity aims at solving that, by ensuring that there is enough diversity in these attribute values.

These steps for protecting privacy have to be taken before the researchers obtains the actual data export. This export is generally a smaller subset of the original database, thus the problem of finding the optimal k-anonymity generalisation is easier. K-anonymity generalisation on subsets is thus ideally integrated in the data access platform, such as I2B2. 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 access and export step.

C. Watermarking

Data is crucial for many organizations as it constitutes an intellectual asset of unique value. This is especially true when sensitive data is processed and the disclosure of datasets can have serious consequences for individuals. Therefore data leaks must be detectable and the source of a leakage must be identifiable and be made accountable. A watermark introduces controlled but meaningless change into a database and therefore allows those in possession of this knowledge to detect data leaks. Naturally, there exists a trade-off between the strength of the watermark and the quality of data [38]. The more records need to be changed, the higher is the probability of the watermark to be detectable.

Watermarking provides several benefits if applied to the data stored in the relational database system. A watermark for relational databases has the following applications: claim authorship, fingerprinting and content validation. [26] introduced a watermarking scheme based on bit patterns to numerical database tuples. The concept has been developed further and several approaches have been identified in [25]. Based on these schemes the dataset can be detected, but not all schemes are resilient to a change in ordering or sub-set generation. An overview of database watermarking and the available schemes is given in [39].

In our scenario, we need to watermark individual subsets for detecting data leaks and have proof that the data was illegally retrieved from a malicious source. The eHealth database we consider in this work utilises various data types, mainly numerical and categorical data. As mentioned before, watermarking changes existing data in order to hide the detection information. For numerical data, the deviation the changes introduce to numbers can be controlled and remain within the specified boundaries. Introducing changes into textual data is more difficult, as the meaning will be completely lost if single letters are flipped. The same is true for categorical data, where flipping one value from a given domain can falsify results. The authors of [38] introduce a scheme where the watermark gets inserted by exchanging non-numerical values from a valid domain. Thus the scheme does not introduce meaningless random data but from the domain of the record.

Fingerprints are an extension to watermarking [40] and several approaches exist [41]. A fingerprint can be generated individually for users downloading a dataset via the application interface. Therefore the change which is introduced into a dataset is again minimal, but can be varied in order to distinguish between the different users and fingerprints respectively. In practice, a hash function is applied which can be recomputed. Therefore the introduced changes is meaningful, which contrasts the fingerprinting approach from watermarks [25].

Obviously the application of watermarking and fingerprinting schemes to sensitive eHealth data which is used for statistical analysis is not a trivial task. Care has to be taken which specific columns of a record may be utilized for applying a categorical watermark, depending on the questions the researcher aims to answer. As the users only may work with the application interface, the data selections can be analysed in advance. Domain experts can define which columns do not allow alteration and which columns can be used for applying the watermarking scheme. Therefore the system may decide on demand which data can be utilized for watermarking in a dataset. We are currently in the process of developing a modification to the I2B2 system that allows for inserting individual fingerprints into the exported data sets.

V. Dynamic Data Citation for Reproducible Research

As the name suggests, data citation is primarily used for referencing research data and provide long term access via persistent identifiers and institutionalized data retention. As research has become data driven in recent years, data sources require the same attention as publication references did so far. Data citation follows the century old tradition of making the sources of knowledge known to peers and therefore allow constructive critique. This does not necessarily imply that the data has to be open and accessible for all users [42], although there are efforts for promoting open access in many scientific areas. Disciplines which handle sensitive data have obviously different requirements which will be described later in this section.

Data citation has evolved by passing three stages [43]: provide descriptions, support access and enable verification and reproducibility. Being able to identify results in an unambiguous way, data citation offers several methods how datasets can be accessed. Researchers would upload one specific dataset to a repository and link the dataset in their publication with a persistent identifier. Peers can then resolve the link and usually are referred to a landing page, which contains metadata about the dataset and the possibility to retrieve the data for inspection. At its core, data citation provides access to evidence by utilizing persistent and unique identifiers supported by verification and attribution metadata[4].

So far, data citation has considered mostly static files which reside in a repository and can be identified for later reference. The growing size and complexity of datasets created a demand for a more flexible data citation approach. For this reason, the concept of dynamic data citation[5] was created which allows to reference user specific datasets and retrieve the data on demand at a later point in time.

A. *Citable Subsets of Dynamic Data*

In our earlier work we developed a data citation framework [44] based on relational database management systems (RDBMS) and demonstrated how it can be applied to existing infrastructures [45]. The framework we presented allows to attach persistent identifiers to the queries instead of the exported data set. We store the queries with additional metadata in the so called query store. Each query store record provides timing information about the query execution and hash keys for later result set verification. By applying versioning to the data by adding timestamp and event metadata, the result set valid at the time of the query execution can be retrieved at a later point in time. Therefore, there is no need to store each and every version of an exported dataset separately, as each query gets persisted and can be used for retrieving the data again. The approach is not limited to relational databases, but can also be applied to other data structures which fulfill requirements such as a query language [46]. Verification methods allow to ensure the completeness of the dataset and therefore increase the trust in research. In this work we show how we can expand our existing framework by privacy preserving data citation capabilities.

B. *Dynamic Data Citation for Sensitive Sources*

Recently data citation also constitutes to reproducibility, as it allows tracing dynamic datasets through their whole life cycle. Therefore data citation constitutes an important building block for reproducible research as it enable researchers to re-execute experiments with the very same data again. In this work, we utilize the concepts of data citation in order to support the long term availability of execution metadata. Additionally to describing datasets and larger data collections on landing pages, we focus on the provenance metadata which needs to be stored in order to provide the knowledge for detecting deviations in workflow executions. Additionally we maintain the privacy of sensitive data by enforcing secure protocols for subsets of data.

As described in Section IV-A, full data access is prohibited due to privacy considerations. Access is only grated via the interface, which gives extended control about the datasets to be retrieved by the users. In order to still enable reproducibility, the queries responsible for creating specific subsets get annotated metadata and a persisted identifier. The metadata which is needed for re-executing a k-anonymous query needs to be stored and linked to the query itself. The same is true for watermarking the result sets and the secret metadata to do so. Therefore a so called execution store is used which is described in VII-A. This approach collects sustainable metadata not only about the query itself, but also about the details required for maintaining the privacy of records inside the database.

We describe how the existing data citation framework [44] can be made compliant with the privacy requirements imposed by sensitive environments. In order to detect data leaks watermarking algorithms need to be applied, as described in Section IV-C. Each generated subset of a sensitive data source must comply to the requirements of the privacy policies (cf. IV-B). Both methods can be provisioned on demand and individually for each data set and for differential users. 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. Therefore the method introduces additional security layers for enforcing privacy and traceable datasets.

The necessary metadata for these operations needs to be preserved for enabling the dynamic creation of datasets on demand. The execution store is therefore an extension of the query store of our previous work. Additionally to the query metadata, anonymization, access and permission policies are stored and maintained and consulted upon re-execution. The anonymization and watermarking parameters are linked individually for each user respectively datasets and may be re-generated on demand. As the data is available at all times, permissions for retrieving data can be adapted to current requirements.

VI. Privacy in Process Preservation

Besides needing to document and preserve the data that is used for an investigation, there is further a need for means to allow repeatability and reproducibility of the experiment itself. This includes description of the experiment design, specifically 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. Further, a documentation and detailed description of the computational environment that was supporting the experiment, including the hardware and software setup. Further, configurations and parameters required for the experiment need to be made available to allow re-execution the same experiment.

Various approaches to achieve this goal have been presented. One step towards sharing the execution environment is via scientific workflows, which have shown to be a useful concept to this end, for example utilising the Taverna Workflow Engine [1]. However, a workflow definition file itself is often not sufficient, and execution often breaks, as has been shown e.g. in [47]. Therefore, descriptions building on top of workflows and augmenting the metadata on the experiments have been proposed, such as e.g. the Research Object model [2], which allows for semantically linking more resources that describe and compose the research workflow, to improve documentation and understandability. The Workflow4Ever project also provided functionality to monitor the evolution of the stability of a workflow, e.g. by checking whether external services are still available.

The aforementioned technical experiment setup, for example what software and hardware is utilised, including details on the configuration and dependencies, is also an important

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

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

aspect, but often not covered sufficiently by workflow systems [3] or the Research Objects. The Context Model presented in [48], along with the tools extracting this information from the system, can alleviate this issue.

There are, however, also potential privacy concerns with the meta-data gathered for enabling re-execution of the investigation. For example, [49], [50] argue 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. The authors introduce *module privacy* as the need to keep the functionality of a specific step in the workflow from unauthorised access, and *structural privacy* as the need to conceal the information that specific modules and data is used to obtain the output of a module or the whole workflow. As an example given, one might want to hide the fact that data from a publicly available repository is utilised, which might allow inference on private data in the workflow.

Besides the workflow definition, also other information captured in a Research Object or Context Model instance might contain privacy related information, e.g. on the users involved in the process.

VII. A Dataset Centric Model for Secure Provenance Data

Researchers use tools such as I2B2 to interact with the database and retrieve datasets which are subsequently used in their investigations. This abstraction allows us to collect and gather metadata which keeps the privacy of data intact and does still provide meaningful metadata for detecting errors or deviations in experiments. All data excerpts which are retrieved from the database must comply with privacy policies. Therefore each dataset only contains k-anonymised records, therefore no personally identifiable data is included as described in Section IV.

In most papers on the topic of k-anonymity (cf. Section IV-B), datasets are exported only once and then made publicly available. Therefore most applications only require the application of the anonymisation algorithm one time. In our use case the data may change during the course of the project, thus ad-hoc methods are required. As a result we have to recompute the k-anonymous datasets after each update while still being able to generate previously issued datasets, including the very same equivalence classes. Therefore the knowledge about the construction of k-anonymity needs to be preserved and constitutes additional metadata for each processing step involving data retrieval from sensitive sources.

A. *The Execution Store*

We adapted the query store concept from [46] and recoined it for the persistent storage of execution metadata which goes beyond classical metadata queries. The execution store is used for storing the information of each workflow step which allows understanding *why* an agent (*who*) introduced *what* kind of data at which steps (*where*) at a given time (*when*) and *how* this step influenced the further processing. The source database needs to be adapted and augmented with privacy relevant metadata. Therefore each table which contains potentially personally identifiable information needs to get marked as confidential. Each query which gets issued against such a table needs to fulfil the predefined privacy requirements. This entails that the query potentially needs to get rewritten in order to return k-anonymised data as described in Section IV-B. All the metadata such as equivalence classes and further privacy enhancing metadata needs to get stored within the execution store for later reference.

Once the result set has been anonymised, it needs to be analysed for storing additional metadata such as result set sizes, hash keys and timing information. This data is relevant for detecting deviations and changes at a later point in time. It is important to stress that this metadata must not contain any information which could allow an attacker to derive information about the actual content. More details about the additional metadata is given in Section VII.

Additionally, the technical metadata describing how a step transformed and processed the data is preserved as well. Tools developed e.g. in [3] demonstrated that systems can be automatically described in a fine grained and accurate manner. Applying the same principles to the steps of a workflow provides the technical metadata needed for describing the required software and hardware components for each single step. This allows gathering the necessary metadata for describing the exchanged data, therefore enabling to reason about the expected qualitative properties of the data. A generic repository of common file formats and their significant properties can be built, which semantically describes properties such as expected file formats, sizes, data types, value ranges et cetera. Therefore the approach we present in this work goes much beyond the comparison of checksums such as SHA1 hashes for detecting deviations in exchanged data.

For achieving this goal, each execution step of a workflow requires a persistent identifier. This PID allows the unique identification of each step and allows the storage and retrieval of metadata needed for reproducing the behaviour of a component at any given point in time. Each execution step may subsume several datasets from various sources. As we described in Section IV-A users interact with the data via predefined interfaces. Therefore the parameters which they use for their queries can be recorded and analysed. The query is stored within the execution store and the metadata gets extracted. This includes data source identifiers such as database names and tables, the dataset specific query parameters such as sorting and filters and the user metadata.

Whenever the data needs to be aggregated or adapted for fulfilling privacy policies, the additional metadata is stored together with the query information and the user metadata. Thus the aggregation of results due to the application of the privacy preserving measures can be re-enacted and therefore the exact same dataset can be retrieved. This also includes watermarking metadata and anonymisation parameters, thus the very same dataset can be created by re-executing the query from the information stored in the query store. Figure 1 shows a schematic overview of the process.

Firstly, the user utilizes the application interfaces (for instance from I2B2) for selection the data needed for his experiment. When she submits the query, the system automatically intercepts and analyses the request. All properties of the query are stored together with a time-stamp and the user details

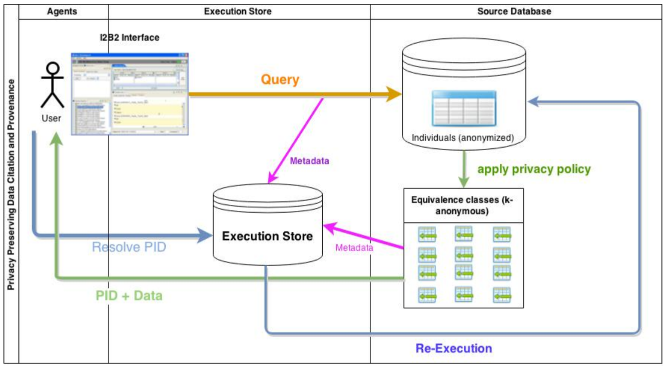

Fig. 1. The Execution Store

within the execution store. The system then normalizes the query and calculates a hash sum of the query parameters. Thus the system can detect whether or not the query has been sent before. If the query is new, a new PID is assigned. If the query is already known to the system, the data is checked for the most recent update. As the database is timestamped, it can be easily detected, whether or not there was an update between the execution of two queries. If no update was detected, the system immediately responds with the already known PID and trans mitts the anonymized data. The execution of the query gets recorded and the user details are being stored. If there was an update between two executions of a query, the system decides to assign a new persistent identifier for denoting the changed dataset. Hierarchical PIDs allow to create versioned trees of query executions and may therefore compare previous query executions.

The use case we described in Section II requires to detect data leaks from subsets of data. Hence watermarking requires special care in our scenario. As described in Section IV-C, we need to include individual watermarks for all participating institutions or even on user level. Again the creation of watermarks has to be repeated whenever the dataset gets updated. The previously inserted watermarks need to be preserved together with their metadata for the later retrieval. Thus each dataset and the user who created it can be recognized at a later point of time. Upon re-execution the system needs to detect whether or not the underlying data source changed.

Each run of an experiments generates new datasets. Linking the metadata of these datasets via the execution store with persistent identifiers allows retrieving the data again at any time. Therefore our metadata view is dataset centric, which entails that characteristics and significant properties are retrieved from the datasets. For each independent step within an investigation, we store additional metadata about the exchanged data between the components in order to detect deviations and errors from the expected standard behaviour. There are several aspects of datasets which can be considered and which do not disclose sensitive information.

- Different input and output data has different metadata characteristics
- Number of records / size in the resultset
- Number of attributes/features included (e.e. columns of a resultset)
- Expected values/units/datatypes per attribute (e.g. ranges, temporal, spatial, SI units, datatypes)
- Expected datatype/format (csv, table, binary file, ...)
- Other characteristics (comprehensiveness (e.g. null values) precision, encoding

The properties are also stored within the execution store and linked to each research investigation persistently. The metadata does not disclose any knowledge contained within the records themselves, as it has been anonymised previously. Still it constitutes to the provenance trail of an scientific experiment. Researchers may utilize the data for discovering alterations in the process and react accordingly.

B. Machine Actionability

The amount or provenance data can escalate quickly. Therefore the processing and interpretation of provenance data must be automated. Semantic provenance data can be used to add meaning to the collected metadata [51]. Therefore the data can be automatically interpreted by software and knowledge can be derived based on formal ontologies. So far semantic provenance data has been used mainly in workflow engines such as Taverna via standards as Janus [52]. Hardly any of the existing semantic data models as described in [53] are applied to data itself. Extracting dataset metadata and feeding it into formally described semantic models allows to interpret the data more efficiently. The relational database data model for instance can be retrieved from the DBMS in an automated way and allows describing selected columns and their properties such as data types, uniqueness, default values or constrains in an efficient way. Thus detailed knowledge about the datasets without having to disclose the actual content can be obtained and stored in a machine processable fashion. Together with the semantic metadata of the processing steps of an experiment, the experimental setup can be described automatically. Assigning identifiers to metadata, not only to data and processing steps, but to each experimental, cycle allows retrieving fine granular but secure metadata about complex research investigations.

C. Automated Deviation Detection

Machine readable and actionable semantic metadata about scientific investigations provides several advantages. In addition to the automated metadata collection during the executions of an experiment, the database of the metadata becomes a prime resource for analysing the experiment on a meta level. As the characteristics of each data exchange between steps are recorded, deviations can be detected. Changes in the characteristics of a dataset can be an indicator for activities which require further investigations.

D. Human Readability

In addition to the machine processable metadata information, human data consumption is an essential aspect for the acceptance of citable research investigations. Having printable documentation of the metadata generated during experiments increases the understanding of internal details and constitutes trust into the scientific process. In essence, landing pages need to be provided for each individual persistent identifier. As several different entities are to be resolved (e.g. metadata about datasets, process step descriptions or user identification data), several different anchors exist which can be browsed from agents but also from human beings.

VIII. Conclusions and Future Work

In this paper we present challenges introduced by the collaborative work of competing organisations in the eHealth domain. We motivate our findings with a concrete example from one of our current projects where researchers from organisations with potential conflicts of interest need to share and exchange data. Based on data excerpts from a very large and sensitive relational database, researchers conduct experiments and draw their results. The raw data cannot be shared as it contains pseudonymous but individual health care data from millions of Austrian citizens. As the combination of data from external sources would allow crafty attackers to deduce information even from anonymized data, privacy preserving methods have to be applied.

We identified reproducibility as a key factor of scientific endeavours. For this reason maintaining privacy has an adverse effect as it hides information and increases complexity in an already challenging task. The goal of our work was to find means how scientific investigations can be rendered reproducible while still maintaining the required privacy levels for all exchanged data. For this reason we focused on three areas which are crucial for understanding how an experiment was conducted: data access, data exchange and provenance.

We increased the privacy of the already pseudonymous data by introducing security preserving methods such as k-anonymization, which hinders the deduction of knowledge from the sensitive data by cross-joining several data sources and inferring facts by data linking. As an additional security layer we include automated and on-demand watermarking and fingerprinting of the data excerpts. We precisely describe the effects on data quality and introduced a persistent storage of the information which is needed for keeping not only the data, but also their fingerprinted and watermarked exports reproducible. Thus data leakage can be traced to individual organisations and even researchers. We enforce these security measures by only allowing indirect data access via a predefined application interface. Therefore all data extraction tasks are within the control of the data owner. This ensures that only privacy enabled subsets are retrieved from the database and that each subset can be linked to a user.

So far, the introduced steps increase privacy but have negative effects on reproducibility as the data looses precision. For overcoming this issue, we adapted existing data citation approaches which enable to identify each dataset individually as it is transformed along the processing chain. The execution store we introduced in this work allows not only to trace which user created as specific dataset, but also to understand how the source data had to be adapted for preserving privacy. This approach has several advantages compared with traditional data publication paradigms. First of all, data citation clearly improves reproducibility as each dataset can be retrieved at a later point in time. Secondly as the information of each dataset and how it was created is available. This entails that each dataset can be retrieved in its original, unchanged version from the source database, as long as the required security clearance levels can be granted. Hence the results can be reviewed with the highest available precision on demand, for instance by trusted reviewers, while the data does not have to be released publicly. Adding the generic research process metadata provides a holistic provenance data collection which allows understanding how a investigation was conducted.

Acknowledgment

This was co-funded by COMET K1, FFG - Austrian Research Promotion Agency

References

[1] P. Missier, S. Soiland-Reyes, S. Owen, W. Tan, A. Nenadic, I. Dunlop, A. Williams, T. Oinn, and C. Goble, “Taverna, reloaded,” in *Proceedings of the 22nd international conference on Scientific and Statistical Database Management*. Berlin, Heidelberg: Springer, June 2010, pp. 471–481.

[2] 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.

[3] 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, SP, Brazil, October 20–24 2014.

[4] 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.

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

[6] J. P. Mesirov, “Computer science. accessible reproducible research,” *Science (New York, NY)*, vol. 327, no. 5964, 2010.

[7] R. Mayer, S. Strodl, and A. Rauber, “On the complexity of process preservation: A case study on an e-science experiment,” in *Proceedings of the 9th International Conference on DigitalPreservation (iPres 2012)*, 9 2012.

[8] J. Loscalzo, “Irreproducible experimental results causes,(mis) interpretations, and consequences,” *Circulation*, vol. 125, no. 10, pp. 1211–1214, 2012.

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

[10] G. K. Sandve, A. Nekrutenko, J. Taylor, and E. Hovig, “Ten simple rules for reproducible computational research,” *PLoS computational biology*, vol. 9, no. 10, p. e1003285, 2013.

[11] J. T. Dudley and A. J. Butte, “Reproducible in silico research in the era of cloud computing,” *Nature biotechnology*, vol. 28, no. 11, p. 1181, 2010.

[12] S. Strodl, R. Mayer, D. Draws, A. Rauber, and G. Antunes, “Digital preservation of a process and its application to e-science experiments,” in *Proceedings of the 10th International Conference on Preservation of Digital Objects (IPRES 2013)*, 9 2013.

[13] G. Antunes, M. Bakhshandeh, R. Mayer, J. Borbinha, and A. Caetano, “Using ontologies for enterprise architecture integration and analysis,” *Complex Systems Informatics and Modeling Quarterly*, 4 2014.

[14] P. Missier, S. Woodman, H. Hiden, and P. Watson, “Provenance and data differencing for workflow reproducibility analysis,” *CoRR*, vol. abs/1406.0905, 2014. [Online]. Available: http://arxiv.org/abs/1406.0905

[15] 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.

[16] L. Moreau, "Provenance-based reproducibility in the semantic web," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 9, no. 2, pp. 202–221, 2011.

[17] K.-K. Muniswamy-Reddy, D. A. Holland, U. Braun, and M. I. Seltzer, "Provenance-aware storage systems." in *USENIX Annual Technical Conference, General Track*, 2006, pp. 43–56.

[18] S. Woodman, H. Hiden, P. Watson, and P. Missier, "Achieving reproducibility by combining provenance with service and workflow versioning," in *Proceedings of the 6th Workshop on Workflows in Support of Large-scale Science*, ser. WORKS '11. New York, NY, USA: ACM, 2011, pp. 127–136. [Online]. Available: http://doi.acm.org/10.1145/2110497.2110512

[19] J. Cheney, L. Chiticariu, and W.-C. Tan, *Provenance in databases: Why, how, and where*. Now Publishers Inc, 2009, vol. 4.

[20] Y. L. Simmhan, B. Plale, and D. Gannon, "A survey of data provenance in e-science," *ACM Sigmod Record*, vol. 34, no. 3, pp. 31–36, 2005.

[21] L. Moreau, "The foundations for provenance on the web," *Found. Trends Web Sci.*, vol. 2, no. 2&#8211;3, pp. 99–241, February 2010. [Online]. Available: http://dx.doi.org/10.1561/1800000010

[22] S. B. Davidson, S. Khanna, S. Roy, and S. C. Boulakia, "Privacy issues in scientific workflow provenance," in *Proceedings of the 1st International Workshop on Workflow Approaches to New Data-centric Science*, ser. Wands '10. New York, NY, USA: ACM, 2010, pp. 3:1–3:6. [Online]. Available: http://doi.acm.org/10.1145/1833398.1833401

[23] S. B. Davidson, S. Khanna, S. Roy, J. Stoyanovich, V. Tannen, and Y. Chen, "On provenance and privacy," in *Proceedings of the 14th International Conference on Database Theory*. ACM, 2011, pp. 3–10.

[24] 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.

[25] 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.

[26] 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.

[27] R. Sion, M. J. Atallah, and S. Prabhakar, "Rights protection for relational data," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 16, no. 12, pp. 1509–1525, 2004.

[28] S. N. Murphy, V. Gainer, M. Mendis, S. Churchill, and I. Kohane, "Strategies for maintaining patient privacy in i2b2," *Journal of the American Medical Informatics Association*, vol. 18, no. Supplement 1, pp. i103–i108, 2011.

[29] T. Dalenius, "Finding a Needle In a Haystack or Identifying Anonymous Census Records," *Journal of Official Statistics*, vol. 2, no. 3, pp. 329–336, 1986.

[30] 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. [Online]. Available: http://dx.doi.org/10.1142/S0218488502001648

[31] ——, "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. [Online]. Available: http://dx.doi.org/10.1142/S021848850200165X

[32] 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*, vol. 2007. American Medical Informatics Association, 2007, p. 548.

[33] S. N. Murphy, G. Weber, M. Mendis, V. Gainer, H. C. Chueh, S. Churchill, and I. Kohane, "Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2)," *Journal of the American Medical Informatics Association*, vol. 17, no. 2, pp. 124–130, 2010.

[34] 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.

[35] L. Sweeney, "Identifiability of de-identified clinical trial data," Carnegie Mellon University, School of Computer Science, Data Privacy Laboratory,, Tech. Rep., 2009.

[36] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, "-anonymity," in *Secure Data Management in Decentralized Systems*, ser. Advances in Information Security, T. Yu and S. Jajodia, Eds. Springer US, 2007, vol. 33, pp. 323–353. [Online]. Available: http://dx.doi.org/10.1007/978-0-387-27696-0_10

[37] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, p. 3, 2007.

[38] R. Sion, "Proving ownership over categorical data," in *Data Engineering, 2004. Proceedings. 20th International Conference on*, March 2004, pp. 584–595.

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

[40] 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*, ser. DRM '03. New York, NY, USA: ACM, 2003, pp. 133–141. [Online]. Available: http://doi.acm.org/10.1145/947380.947398

[41] ——, "Fingerprinting relational databases: Schemes and specialties," *IEEE Transactions on Dependable and Secure Computing*, vol. 2, no. 1, pp. 34–45, 2005.

[42] 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 nercâ s environmental data centres," *International Journal of Digital Curation*, 2012.

[43] M. Altman and M. Crosas, "The evolution of data citation: From principles to implementation," *IASSIST Quarterly*, vol. 37, 2013.

[44] S. Pröll and A. Rauber, "Citable by Design - A Model for Making Data in Dynamic Environments Citable," in *2nd International Conference on Data Management Technologies and Applications (DATA2013)*, Reykjavik, Iceland, July 29-31 2013.

[45] ——, "Data Citation in Dynamic, Large Databases: Model and Reference Implementation," in *IEEE International Conference on Big Data 2013 (IEEE BigData 2013)*, Santa Clara, CA, USA, October 2013.

[46] ——, "A Scalable Framework for Dynamic Data Citation of Arbitrary Structured Data," in *3rd International Conference on Data Management Technologies and Applications (DATA2014)*, Vienna, Austria, August 29-31 2014.

[47] J. Zhao, J. M. Gómez-Pérez, K. Belhajjame, G. Klyne, E. García-Cuesta, A. Garrido, K. M. Hettne, M. Roos, D. D. Roure, and C. A. Goble, "Why workflows break - understanding and combating decay in taverna workflows," in *8th IEEE International Conference on E-Science (e-Science 2012), Chicago, IL, USA, October 8-12, 2012*. IEEE Computer Society, 2012, pp. 1–9.

[48] 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. [Online]. Available: http://www.springer.com/-/7/7e9c68c9a6ac468aaacd08a7827e82bf

[49] S. B. Davidson, S. Khanna, V. Tannen, S. Roy, Y. Chen, T. Milo, and J. Stoyanovich, "Enabling privacy in provenance-aware workflow systems," in *CIDR 2011, Fifth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 9-12, 2011, Online Proceedings*. www.cidrdb.org, 2011, pp. 215–218. [Online]. Available: http://www.cidrdb.org/cidr2011/Papers/CIDR11_Paper30.pdf

[50] S. Davidson, S. Khanna, and T. Milo, "To show or not to show in workflow provenance," in *In Search of Elegance in the Theory and Practice of Computation*, ser. Lecture Notes in Computer Science, V. Tannen, L. Wong, L. Libkin, W. Fan, W.-C. Tan, and M. Fourman, Eds. Springer Berlin Heidelberg, 2013, vol. 8000, pp. 217–226. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-41660-6_10

[51] S. S. Sahoo, A. Sheth, and C. Henson, "Semantic provenance for escience: Managing the deluge of scientific data," *Internet Computing, IEEE*, vol. 12, no. 4, pp. 46–54, 2008.

[52] P. Missier, S. S. Sahoo, J. Zhao, C. Goble, and A. Sheth, "Janus: from

workflows to semantic provenance and linked open data," in *Provenance and Annotation of Data and Processes*. Springer, 2010, pp. 129–141.

[53] J. Peckham and F. Maryanski, "Semantic data models," *ACM Computing Surveys (CSUR)*, vol. 20, no. 3, pp. 153–189, 1988.