Exploiting Distributed, Heterogeneous and Sensitive Data Stocks while Maintaining the Owner’s Data Sovereignty

M. Lablans; D. Kadioglu; M. Muscholl; F. Ückert
University Medical Center Mainz, Mainz, Germany

Keywords
Collaborative research, biobank, disease registries, federation, distributed search, data protection, data sovereignty, data integration

Summary
Background: To achieve statistical significance in medical research, biological or data samples from several bio- or databanks often need to be complemented by those of other institutions. For that purpose, IT-based search services have been established to locate datasets matching a given set of criteria in databases distributed across several institutions. However, previous approaches require data owners to disclose information about their samples, raising a barrier for their participation in the network.

Objective: To devise a method to search distributed databases for datasets matching a given set of criteria while fully maintaining their owner’s data sovereignty.

Methods: As a modification to traditional federated search services, we propose the decentralized search, which allows the data owner a high degree of control. Relevant data are loaded into local bridgeheads, each under their owner’s sovereignty. Researchers can formulate criteria sets along with a project proposal using a central search broker, which then notifies the bridgeheads. The criteria are, however, treated as an inquiry rather than a query: Instead of responding with results, bridgeheads notify their owner and wait for his/her decision regarding whether and what to answer based on the criteria set, the matching datasets and the specific project proposal. Without the owner’s explicit consent, no data leaves his/her institution.

Results: The decentralized search has been deployed in one of the six German Centers for Health Research, comprised of eleven university hospitals. In the process, compliance with German data protection regulations has been confirmed. The decentralized search also marks the centerpiece of an open source registry software toolbox aiming to build a national registry of rare diseases in Germany.

Conclusions: While the sacrifice of real-time answers impairs some use-cases, it leads to several beneficial side effects: improved data protection due to data parsimony, tolerance for incomplete data schema mappings and flexibility with regard to patient consent. Most importantly, as no datasets ever leave their institution, owners can reject projects without facing potential peer pressure. By its lower barrier for participation, a decentralized search service is likely to attract a larger number of partners and to bring a researcher into contact with the right potential partners.

1. Introduction

To achieve statistical significance in medical research, biological samples taken from one biobank often need to be complemented by those of other institutions, thereby creating the need for biobank networking [1–5]. Similarly, patient registries focusing on rare diseases need to network in order to combine their individually small number of cases [6]. Lastly, locating patients in multiple clinics can help to meet recruitment targets for clinical studies [7].

In all these three use-cases, research can benefit from a network of different institutions. In vast networks, numerous scientists from several medical faculties need to cooperate. Data and material from a large number of patients must be made available in a manner whereby their use presents no ethical, legal or data protection issues. Data, be it aggregated, multi-level pseudonymized or k-anonymized, must find their way to the respective research groups.

In their aim to open up the potential of distributed data and sample stocks for medical joint research efforts, all three use-cases share a common requirement: Bringing researchers with a project idea into contact with owners of relevant data collections or biospecimens and their associated sample and donor data. From the perspective of medical informatics, this means locating datasets matching a given set of criteria in databases that are distributed over several institutions.

A large number of national and international projects have established platforms that can be considered for such a project mediation:

- Public catalogues as provided by BBMRI [2], Orphanet [6] or the Ger-

Correspondence to:
Martin Lablans
University Medical Center Mainz
Obere Zahlbacher Straße 69
55131 Mainz
Germany
E-mail: lablans@uni-mainz.de
man Biobank Registry [8] are based on aggregated data usually entered manually by the owner using a web browser. Thus, catalogues are not intended for fine-grained search of matching samples or datasets, but offer a first, rough overview of existing biobanks or databases.

- Central search services as implemented in CRIP [3] or the NCI Specimen Resource Locator [4] receive periodic, usually automated data uploads to be stored in a central database. A website allows researchers to query this database using criteria sets that can, in contrast to aggregated catalogues, combine several desired properties and thus provide immediate and more meaningful results.

- Federated search services as implemented in i2b2 SHRINE [9] and EHR4CR [10] offer similar search functionality as their central counterparts but do not rely on periodic uploads or central data storage. Instead, the participating databases are accessed using centrally initiated, federated queries.

Due to the restriction in fine-grained search mentioned for public catalogues, this article focuses on central and federated search services, which may operate on a non-aggregated sample level. Search services face a twofold challenge:

The first challenge lies in the field of data integration. The datasets required to carry out searches across multiple institutions need to be extracted from heterogeneous source systems and made comparable in their meaning (technical and semantic interoperability). In addition, since patient data are processed, strict data protection regulations apply.

The second challenge is of a human nature rather than technical. Until now, either the relevant datasets or their aggregated versions were provided either beforehand via regular uploads (central search services) or on-demand via distributed queries (federated search services). In both cases, third parties – whom the data or sample owners may neither know nor trust – gain insight into databases with highly sensitive content. While the level of this insight differs depending on the type (central or federated) of the search service and its implementation (e.g. non-aggregated or aggregated), this requirement for self-disclosure places data owners in a dilemma: On the one hand, they want to contribute to collaborative research projects that appear promising to result in successful publications. On the other hand, “registry custodians frequently hold proprietary views on their data or face legal limitations on data-sharing” [11] and want to carefully consider with whom to share their assets collected over years without facing peer pressure to justify rejecting a proposal.

2. Objective

In order to solve the dilemma outlined in the second challenge, we have devised a method to search distributed databases for datasets matching a given set of criteria, yet fully keep their owner’s data sovereignty.

3. Methods

We propose the decentral search – a modification of federated search services that exploits distributed, heterogeneous, highly sensitive datasets from heterogeneous systems for overarching research questions.

3.1 Local Data Integration

While we assume that the relevant patient and biomaterial data are stored in a structured format, its data schema as well as the used IT system may vary among the participating institutions. In order to create the technical and semantic interoperability required to search the heterogeneous databases, the decentral search relies on a set of software components we call a bridgehead: Using ETL processes, relevant data are regularly extracted from the primary system, transformed into a common format and loaded into each partner’s bridgehead.

Semantic interoperability between the bridgeheads is achieved by means of a central metadata repository that can import and re-use data elements from other ISO 11179-based implementations [12] such as NCI’s caDSR [13]. The metadata repository provides attributes according to the ISO 11179 [14] representative layer via a RESTful web interface. However, it is up to the consortium to define the required data elements and up to the data owner to map each attribute’s key to a data element in the metadata repository. The latter process may require the harmonization of different data models, which is a complex issue that may involve more than one-to-one mappings and simple data transformations [15, 16]. Differing permissible values, deviating semantics and missing data are examples of typical obstacles making data integration a “well-known grand challenge” [17] that easily exceeds the scope of this article. In the following, we thus assume that the data owner, probably assisted by a domain expert, manages to create an ETL process that transfers the information to the bridgehead.

As its name implies, the bridgehead runs locally under full control of the data owner rather than that of a third party running the search service. Therefore, although the technical approach is similar to the materialized data integration [18] employed in central search services, data in the bridgehead do not leave the institution, but stay under the owner’s sovereignty until he or she decides to share them as response to another researcher’s inquiry (see next section).

3.2 Inquiry vs. Query

Similar to other federated search services, the decentral search detects matching material in distributed data. However, their traditional query mechanism is replaced by a novel inquiry mechanism that involves the owner with a high degree of control: The owner can (each decentrally at their site using the user interface of their own bridgehead) decide if and what to answer based on a specific project proposal.

A decentral search involves six steps (Figure 1):

1. The relevant datasets are transferred into the bridgehead (see 3.1).
2. The bridgehead is registered with a so-called search broker, a software component operated by, for example, a research network (see 4.1). However, data about the contents of the bridgehead are
not transmitted (not even in aggregated form).

3. A researcher uses the search broker to describe the required material using a criteria set, i.e. pairs of keys according to data elements from the metadata repository and values in accordance with their value domain. He/she also provides a project proposal to advertise his/her project. Upon submission, the researcher does not receive an immediate result but is asked to wait to be contacted.

4. The search broker sends the inquiry (criteria and project proposal) forward to every registered bridgehead. More specifically, in order to deal with restrictive firewall policies, the bridgehead regularly polls the search broker to fetch new inquiries.

5. Whenever a bridgehead has received a new inquiry, it automatically runs the criteria against its local database. Rather than returning a result to the search broker, no data are transmitted externally. Instead, each bridgehead encounters one of these cases:

a. There are no datasets fitting the criteria set (bridgehead 3 in the example). The bridgehead discards the inquiry without notifying the owner.

b. Suitable datasets are found. The bridgehead sends an e-mail asking the owner to review the inquiry and its matching datasets. Clicking the link in the e-mail, the owner is presented with a webpage rendered by the bridgehead. The page contains a worklist with all inquiries and their respective status (e.g. "unanswered", "answered", etc.).
**4. Results**

A system participating in the decentral search consists of four software components (Figure 2):

- **The local data warehouse** is the materialized integration layer that stores data about available material and the respective patient annotation received from a source system as result of an ETL process (1).
- **These datasets** correspond to the data elements in the metadata repository (2). That way, data in bridgeheads at several institutions become comparable.
- **The inquiry mechanism itself** is implemented as a RESTful web service provided by the decentral search broker, which uses the metadata repository to create a criteria set (3a). The resulting inquiry is then downloaded by the decentral search client (3b) using the broker’s RESTful web service.

We have developed a reference implementation as a set of four Java servlets. However, participants are not required to use our software. Instead, we aim to publish the RESTful interfaces, enabling database implementations developed by third parties to participate as a bridgehead in the decentral search.

**4.1 Practical Deployment**

In order to demonstrate the practicality of the decentral search, we provide two examples which employ the decentral search in different use-cases: to network several clinics (cf. 4.1.1) and to network networks of clinics (cf. 4.1.2).

**4.1.1 Networking Clinics**

The decentral search has been deployed in the German Consortium for Translational Cancer Research (DKTK), one of the six German Centers for Health Research comprised of eleven university hospitals [19, 20]. The bridgeheads deployed at each site support two search mechanisms:

- For patients who have signed an informed consent, a 62-attribute report dataset is uploaded to a traditional, central search service. In a two-staged pseudonymization, identifying data are sent to a central “Mainzelliste” pseudonymization service [21] that allows datasets of patients treated in multiple centers of the consortium to be linked. However, the number of data elements in the report dataset is limited by data protection requirements: Too detailed data uploads might allow unauthorized re-identification of patients in the central database. Also, the data of most patients has been collected in the context of treatment prior to establishment of the DKTK. As patient consent is collected prospectively and no center has implemented the necessary organizational process for the collection of consent yet, consenting patients are expected to constitute only a small fraction compared to those documented prior to the establishment of the DKTK.

- The latter, retrospective cases are made accessible by the decentral search. In addition to the larger number of patients searched, the number of data elements in the criteria dataset is larger as well because it is not limited by issues of data protection or data sovereignty. In addition to using the 174-attribute “ADT dataset” proposed by Altman et al. [22–24], task forces are being established that will define further attributes to describe special tumor entities within the consortium.

The development of the necessary software components was completed at the end of 2014. At the end of the first funding period of the consortium’s IT project in February 2015, an internal interim report evaluated the practical rollout of the bridgeheads and their integration with the local clinical systems. A survey revealed that five of eleven sites had completed all the required steps to establish and use their local bridgeheads, i.e. installed the bridgehead server, implemented the necessary ETL process and obtained approval of both the data owners and their local data protection officers. In total, these five sites contribute about 267,000 patient datasets with routine data (based on the ADT dataset) to the consortium’s decentral search.

Based on our regular consortium meetings, we expect the remaining sites to finish their local implementations in the next months. The delay was mainly caused by complex and thus time-consuming local processes involving several stakeholders, e.g. obtaining approvals from the local data owners and the local protection officer or implementing complex ETL processes. A first evaluation of the decentral search in clinical practice (including, for example, a user survey and a comparison of the number of generated inquiries and the number of mediated projects) is scheduled for 2017.

**4.1.2 Networking Networks of Clinics**

The decentral search also opens up new ways to further network whole consortia. In 2009, the European Union recommended that member states establish registries for rare diseases in compliance to the EUCERD recommendations [25]. The small number of cases makes interconnection of these collections especially important for the clinical research in the field [11, 25]. The OSSE project (Open Source Regis-
try System for Rare Diseases in the EU) [26] tries to fill this gap by pursuing two goals.

The first goal is to provide a toolkit that allows building remote data entry systems for clinical registries from scratch. One form of this toolkit is a CD image that can be inserted into a (physical or virtual) server to install a Linux-based operating system, a remote data entry solution with role-based access control, the “Mainzel-liste” pseudonymization service and a decentralized search client. A system administrator can make the centrally located database accessible via the web for different partners working at different locations, who then form a research network spanning several clinics.

The second goal is to network these networks even further and is made possible by the metadata repository in conjunction with the decentral search: Similar to other configurable web-based remote data entry systems [27, 28], OSSE allows users to create custom forms according to the specific requirements of the respective disease. However, OSSE requires all attributes in each clinical form to correspond to data elements in a central metadata repository, thus enforcing semantic interoperability (cf. 3.1). Building on this foundation, the built-in decentralized search client can be used to register with one or more decentralized search brokers and thus receive inquiries for datasets, thus creating a federation of OSSE-based registries.

The OSSE software components are released open-source under the GNU Affero General Public License (AGPL version 3 or later) [29]. European dissemination will be provided by the European Commission’s Joint Research Center [30]. More information on OSSE, including data protection concepts and manuals on functionality and technical details, is available on the project homepage (see section ‘Acknowledgments’).

5. Discussion

5.1 Related Work

Over the last years, several federated search services have been established. For instance, the i2b2 (Integrating biology and the bedside [31]) clinical data warehouse can be loaded by ETL processes and queried using a user-friendly search interface. Using i2b2 SHRINE (Shared Health Research Information Network [9]), such
searches can be extended to several i2b2 instances that may span several institutions. McMurry et al. present five established networks, the largest of which consists of 61 institutions [32]. Natter et al. present another network comprised of 56 sites [33].

A similar architecture is employed by EHR4CR (Electronic Health Records for Clinical Research [10,34]): Local data warehouses are filled using an ETL process and queried using a central workbench. Doods et al. demonstrate a recent pilot across eleven sites in five countries [7].

Lastly, Eder et al. propose a federated hub-and-spoke architecture as part of the BBMRI (Biobanking and Biomolecular Resources Research Infrastructure) preparatory phase [5]. However, unlike its catalogue [2] of over three hundred biobanks [35], BBMRI’s federated search service has, to our knowledge, never been deployed in a practical use-case comparable to those of SHRINE, EHR4CR or the decentral search.

While all federated search services greatly differ in detail, the difference most relevant to their comparison to the decentral search is their involvement of the data owner. In previous federated search services, queries are answered automatically by the respective local database, thus implementing a query rather than an inquiry. The instantaneous reply leads to several advantages as well as disadvantages, both of which are discussed in the next section.

5.2 Limitations and Advantages

As decentral search clients do not disclose any query results without prior permission of its owner, inquiries do not yield results immediately, and it may take several days to receive an answer. On one hand, this delay leads to several disadvantages:

- The inquiring researcher cannot issue multiple queries in succession to determine a criteria set that can be satisfied by the potential research partners’ material. Thus, it is impossible for her him to fine-tune her his inquiry’s criteria set without involving the data owner.
- The approach is infeasible for applications relying on prompt response times. For example, study personnel recruiting patients into clinical studies during their hospital visit as proposed by Trinczek et al. [36] would not receive notifications about matching patients in time.
- Similar to a bilateral negotiation, the decentral search guarantees neither completeness (as not all owners have to answer) nor correctness (as owners can modify their reply). Thus, statistical evaluations such as incidence counting require prior coordination among the owners.

On the other hand, the sacrifice of real-time responsiveness guarantees that no data leave the institution unless explicitly permitted by the owner. This promise leads to several beneficial effects:

- Participation in the network does not impair the owner’s data sovereignty. The requirement for self-disclosure – and, thus, the owner’s dilemma outlined in the introduction – is resolved.
- The method provides a higher degree of data parsimony than central or query-based federated search services. For mere participation in the network, obtaining data protection approval becomes facilitated or even redundant. Similarly, informed patient consent is not always necessary for mere participation in the decentral search as data-sets are searched locally.
- The inquiry-based search service can tolerate incomplete schema mappings. For example, consider searching a biobank specialized in cervical cancer for samples of female donors. As this disease occurs only in females, the criterion is obviously fulfilled but not (redundantly) documented in the biobank software. A query-based system evaluating such criteria defined in the inquiry but not in a data stock may either include such a sample in the result set (as the attributes do not contradict the criteria) or omit it (as not all criteria are fulfilled). The decentral search provides a third option: Include the sample in the result set but ask the owner afterwards to decide whether to include the sample and even to add missing data.

Overall, the improved data protection due to data parsimony, the tolerance for incomplete data schema mappings and the flexibility with regard to patient consent lead to a decreased effort when the network is initially joined. While negotiations with ethics committees, data protection officers as well as data or sample owners cannot be omitted, they may be facilitated by the fact that data are loaded into a local bridgehead rather than made accessible to an external party. Of course, as with other search services, approval is still required before samples or datasets are actually shared.

Once processes for data sharing are defined, necessary approvals are obtained and trust among researchers is built (e.g. through successfully mediated research projects), the decentral search client can be configured to auto-accept certain incoming inquiries (e.g. originating from trusted colleagues and/or inquiring a certain set of attributes) and thus avoid the disadvantages introduced by the manual permission. However, in order not to weaken the system’s promise to the data owner for data sovereignty and data parsimony, the criteria for automatic replies should be defined by the owner himself herself and default to asking the owner. The owner can still interpret the results prior to making contact with the inquirer, e.g. by alternating the search question out of experience or by adding data or material not known to the system. As a result, a single bridgehead can serve cooperations at different levels of trust and/or service.

6. Conclusion

We have presented a novel method to search distributed, heterogeneous and highly sensitive databases for datasets matching a given set of criteria. In contrast to the query found in previous work, the inquiry does not require owners to disclose information and thus protects their data sovereignty. We have presented the decentral search concept (cf. 3) and components of a reference implementation (cf. 4) that relies on a RESTful web interface and a metadata repository. Lastly, we have presented two practical deployments (cf. 4.1): the first within one of the six German Centers for Health Research comprised of eleven university hospitals, and the second to interconnect rare disease registries.
While the lack of real-time answers is at first glance detrimental to the inquiring researcher’s user experience, the increased data sovereignty, enhanced data protection and tolerance for incomplete data schema mappings and missing patient consent lead to increased acceptance with the owner and lower the barrier for participation in the network. As a result, the network is likely to attract a larger number of participants and to bring researchers into contact with more potential research partners. Thus, although intended to serve the owner, the inquiry can benefit the inquiring researcher as well.

Acknowledgments

This work is part of a doctoral thesis mentored by Professors Frank Ückert and Srej Gorlatch, the latter from the University of Münster. The German Consortium for Translational Cancer Research (DKTK) is funded by the German Federal Ministry of Education and Research. For more information, see http://www.unimedizin-mainz.de/dktk. OSSE is funded by the German Federal Ministry of Health (IIA5–2513-FSB-511) as part of the “National Action Plan for Rare Diseases”. For more information, see http://osse-register.de.

References