SAMPLE INDEX AND COLLECTIONLOCATOR

V. SHEKHOVTSOV[1], P. EISENKEIL[2], B. DEHARI[2], G. GOEBEL[1], J. EDER[2]
[1]Medical University of Innsbruck, Innsbruck, Austria, [2]University of Klagenfurt, Klagenfurt, Austria

INTRODUCTION

To provide material and data to researchers efficiently, biobanks have to implement means of searching for collections that contain the data needed for research.

The problem with such a search is that

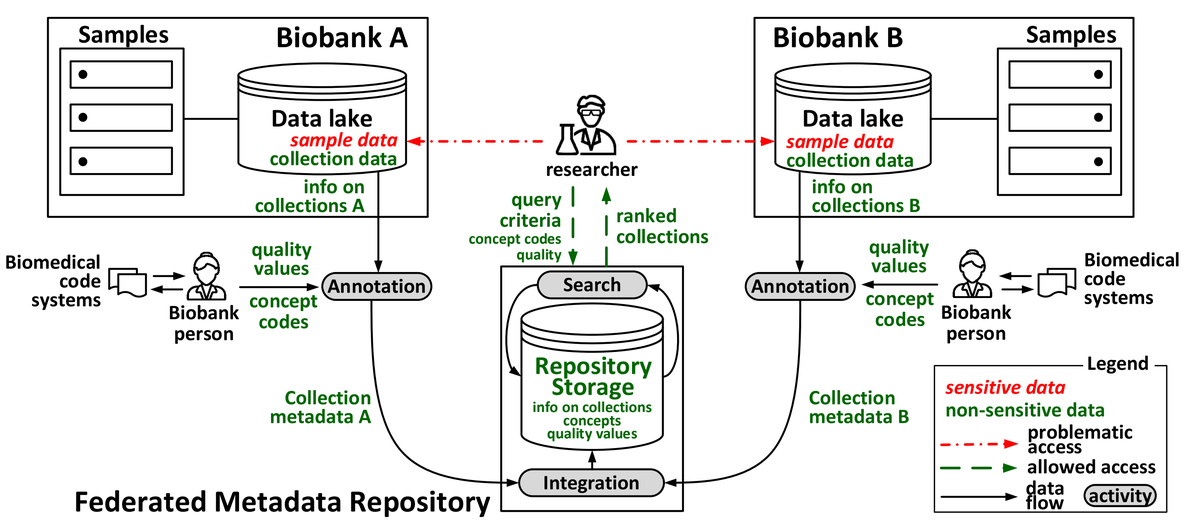
1. The sources of data in biobanks are heterogeneous - so no common schema is available
2. General direct access to the sample data impossible due to privacy restrictions.

The privacy problem leads to the following consequences:

- there is no publicly available central repository to perform biobank search taking into account privacy restrictions
- so the researchers could encounter problems with finding the resources they need to conduct their work.

CONCEPT

The proposed solution is to search in a **central repository of collection metadata** augmented with indexes. This search returns collections meeting certain metadata-based criteria.



COLLECTIONLOCATOR ARCHITECTURE

Goal: supporting the search for suitable collections within indexes consisting of

- **preprocessed non-sensitive data** (anonymized, aggregated data)
- **collection metadata** (quality values, concept identifiers).

The tool architecture contains:

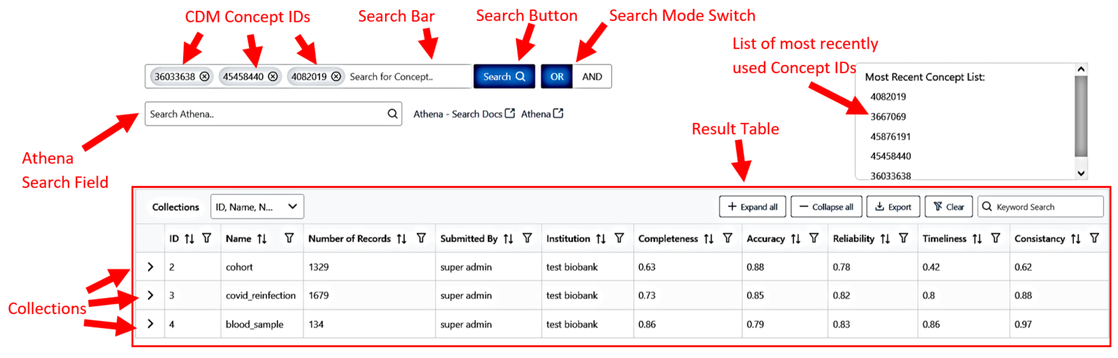
1. **The data anonymization component** converting the biobank data into a non-sensitive form
2. **The means for annotating collections** with semantic concepts and quality values
3. **The repository storage component** holding preprocessed non-sensitive data and collection metadata, namely semantic concepts and quality values
4. **The search component** running queries against the repository and returning requested collections.

COLLECTIONLOCATOR USER INTERFACE

Search Functionality: The tool user interface supports the following search modes:

- Search for collections annotated with certain LOINC (later generalized as OMOP CDM) concepts
- Search for collections possessing specific values for aggregated and anonymized data
- Search for collections possessing specific quality values

The solution was validated with data from the BBMRI Colorectal Cancer Cohort.



DISCUSSION

Compliance with the accessibility principle for biobank data:

- By offering the possibility to search within non-sensitive data and metadata when the original data is not accessible.

Novelty of the tool:

- Support for semantic data annotation together with its quality-based annotation, aggregation and anonymization
- As a result, the central repository can offer rather fine grain information about collections and their associated data sets without any potential compromise of the privacy of the donors.
- Complements BBMRI tools Directory and Sample Locator, Sample Finder

Contacts: volodymyr.shekhovtsov@i-med.ac.at, johann.eder@aau.at, georg.goebel@i-med.ac.at

REFERENCES

1. Eder, J., Shekhovtsov, V.A. Managing the quality of data and metadata for biobanks. In FDSE 2022. CCIS, vol. 1688, Springer, 2022, pp. 52-69.
2. Shekhovtsov, V.A., Eder, J. Metadata quality for biobanks. *Applied Sciences.* 12(19), 2022, 9578.
3. Shekhovtsov, V.A., Eder, J. Data item quality for biobanks. *Trans. on Large-Scale Data-and Knowledge-Centered Systems* L. LNCS. vol. 12930, 2021, pp. 77–115.
4. Eder, J., Shekhovtsov, V.A. Data quality for federated medical data lakes. *International Journal of Web Information Systems*. 17(5), 2021, 407-426.
5. Eder, J., Shekhovtsov, V.A. Data quality for medical data lakelands. In: FDSE 2020. LNCS, vol. 12466, Springer, 2020, pp. 28–43.
6. Eder, J., Dabringer, C., Schicho, M., & Stark, K. Information systems for federated biobanks. *TLDKS I*, LNCS, 2009, pp. 156-190.

Bundesministerium Bildung, Wissenschaft und Forschung

Funded by the Austrian Federal Ministry of Education, Science and Research
Grant numbers:
BMBWF-10.470/0010-V/3c/2018 (2018-2023)
2023-0.752.780 (2023-2028)