Published online: 2019-09-11

Pan-European Data Harmonization for Biobanks in ADOPT BBMRI-ERIC

Sebastian Mate[1] Marvin Kampf[1] Wolfgang Rödle[2] Stefan Kraus[2] Rumyana Proynova[3] Kaisa Silander[4] Lars Ebert[5] Martin Lablans[5] Christina Schüttler[2] Christian Knell[1] Niina Eklund[4] Michael Hummel[6,7] Petr Holub[7] Hans-Ulrich Prokosch[1,2]

[1] Medical Centre for Information and Communication Technology, Universitätsklinikum Erlangen, Erlangen, Germany
[2] Chair of Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany
[3] Medical Informatics in Translational Oncology, German Cancer Research Center, Heidelberg, Germany
[4] Genomics and Biobank Unit, Finnish National Institute for Health and Welfare, Helsinki, Finland
[5] Federated Information Systems, German Cancer Research Center, Heidelberg, Germany
[6] Institute of Pathology, Charité-Universitätsmedizin Berlin, Berlin, Germany
[7] Biobanking and BioMolecular Resources Research Infrastructure (BBMRI-ERIC), Graz, Austria

Address for correspondence Sebastian Mate, Medical Centre for Information and Communication Technology, Universitätsklinikum Erlangen, Krankenhausstraße 12, DE-91054 Erlangen, Germany (e-mail: sebastian.mate@uk-erlangen.de).

Appl Clin Inform 2019;10:679–692.

Abstract

Background High-quality clinical data and biological specimens are key for medical research and personalized medicine. The Biobanking and Biomolecular Resources Research Infrastructure-European Research Infrastructure Consortium (BBMRI-ERIC) aims to facilitate access to such biological resources. The accompanying ADOPT BBMRI-ERIC project kick-started BBMRI-ERIC by collecting colorectal cancer data from European biobanks.

Objectives To transform these data into a common representation, a uniform approach for data integration and harmonization had to be developed. This article describes the design and the implementation of a toolset for this task.

Methods Based on the semantics of a metadata repository, we developed a lexical bag-of-words matcher, capable of semiautomatically mapping local biobank terms to the central ADOPT BBMRI-ERIC terminology. Its algorithm supports fuzzy matching, utilization of synonyms, and sentiment tagging. To process the anonymized instance data based on these mappings, we also developed a data transformation application.

Results The implementation was used to process the data from 10 European biobanks. The lexical matcher automatically and correctly mapped 78.48% of the 1,492 local biobank terms, and human experts were able to complete the remaining mappings. We used the expert-curated mappings to successfully process 147,608 data records from 3,415 patients.

Conclusion A generic harmonization approach was created and successfully used for cross-institutional data harmonization across 10 European biobanks. The software tools were made available as open source.

Keywords
- data curation
- metadata
- common data elements
- health information interoperability
- biological specimen banks

received
February 13, 2019
accepted after revision
July 12, 2019

DOI https://doi.org/10.1055/s-0039-1695793.
ISSN 1869-0327.

Background and Significance

An important task in the medical research of the postgenomic era is the investigation of diseases by linking "-omics" data with phenotypes.[1] Widespread conditions, such as cancer, diabetes, or Alzheimer's disease, are influenced by a variety of small, often cumulative effects[2] that result from a combination of genes,[3] lifestyle,[4] and environmental factors.[5] The complexity of molecular disease patterns requires many forms of therapeutic intervention, tailored to the individual characteristics of a particular patient.[6] Clinical data, associated with well-characterized biomaterials, both of the highest quality, are key for the development of new drugs and diagnostic tests and are expected to be the building blocks of personalized medicine in the near future.[7]

The Biobanking and Biomolecular Resources Research Infrastructure - European Research Infrastructure Consortium (BBMRI-ERIC) aims to facilitate access to biological resources.[8] Via its *Common Service Information Technology* (CS-IT) platform, a distributed computing environment, BBMRI-ERIC is working on establishing a pan-European search engine for biomaterial samples to support biomolecular and biomedical research. As described in more detail in Proynova et al,[9] it is planned that the CS-IT platform follows the principles of the *decentralized search approach*,[10] developed by the German Cancer Consortium.[11] In contrast to *centralized search approaches* (as implemented, e.g., in Schröder et al[12]), data never leaves the biobanks, which enables them to retain data sovereignty and protect their patients' privacy.

To support the implementation of the CS-IT platform, the *Implementation and Operation of the Gateway for Health into BBMRI-ERIC* (ADOPT BBMRI-ERIC) project was launched. One of its tasks was the collection and aggregation of colorectal cancer data across European biobanks.[13] Colorectal cancer was selected as it is one of the most prevalent neoplasms,[14] and most BBMRI-ERIC national nodes already had expertise in this area through ongoing research programs. The national nodes are the national biobank networks in each European country participating in the BBMRI-ERIC network.[15] Besides sample data (such as material type or preservation mode), the data collection covered patient information regarding diagnostic and surgical procedures, histopathology, molecular markers, demographics, planned and performed therapy (e.g., radiation or chemotherapy), and outcome. The goal was to make these anonymous data available on the BBMRI-ERIC platform after collection and validation. This process was known as the *Colon Cancer Data Collection* (CCDC) within ADOPT BBMRI-ERIC.[16]

Objectives

Consolidating the heterogeneous biobank data into the common format required the data to be processed in a systematic way. One of the ADOPT BBMRI-ERIC work packages was focusing on the development of a data integration and harmonization toolkit to support this extract-transform-load process (ETL, also see refs. 17–19). It had to be able to convert the biobank data semiautomatically into a format ready for import into the central CCDC database.

In this article, we describe this data harmonization approach and report on the experiences gained while integrating the data from 10 European biobanks. In particular, we analyze the quality of the semiautomatic mapping approach and report on the results of the ETL process.

Methods

Overview

The ADOPT BBMRI-ERIC project used an installation of the software *Open Source Registry System for Rare Diseases* (OSSE, *Open-Source-Registersystem für Seltene Erkrankungen*)[20] as collection system for the CCDC data. OSSE is a case report form (CRF) system that generates forms for data entry by using standardized data element definitions from the *Samply.MDR* metadata repository (MDR).[21] Both applications are part of the *Samply* software environment,[22] for which expertise was available within the BBMRI-ERIC community and which is also used in other projects.[11,23–25] It allows for manual data entry via the CRFs to enable the participation of those biobanks, which are only able to provide data manually, but also provides an application programmable interface (API) for the automatic import of XML-encoded data.

The latter method was preferred as it helps avoiding errors that may occur when copying data manually. The ETL process for this task should fulfill the following requirements:

1. The overall ETL process has to provide a generic input interface for the data originating from the biobanks to be compatible with the wide range of data formats.
2. It has to support the semantic translation of data elements and value sets from the biobank data into the target CCDC terminology as defined in the MDR.
3. It has to convert the biobank data into the XML import format as required by OSSE. Since biobanks might provide data in a completely different format, complex syntactic conversions need to be supported.

To address the first requirement, we specified two easy-to-generate input file formats for the ETL process, in which the biobanks could provide their data. The first one is a comma separated value file format, which follows the principles of the Entity-Attribute-Value (EAV) data modeling approach, as shown in ►Table 1 and described in more detail in refs. 26 and 27. It enables the provision of multiple instances of individual data elements, such as multiple locations of metastases. The second one is a flat file format. As shown in ►Table 2, it uses one row per patient and encodes multiple instances within single cells, separated by semicolons (see highlighting).

The goal of the second and third requirement was to convert the biobank data into the target XML format, while adhering to the semantics of the target terminology with a mapping and transformation process that requires minimal human intervention. As already mentioned, the MDR contained a detailed, machine-readable description of the CCDC data elements (see ref. 28), which not only serves the automatic OSSE form generation, but also defines the contents and data types that may be utilized in an OSSE import XML file. We asked the biobanks to provide the same standardization by using the

Table 1 Example for the EAV table format (mock-up data)

Patient ID	Concept	Value	Instance
1	TNM-T	T1	1
1	TNM-N	N0	1
1	TNM-M	M1	1
1	Age	45	1
1	Metastases	**Hepatic**	**1**
1	Metastases	**Osseus**	**2**
1	Gender	Male	1
1	Date of Surgery	21.07.2013	1
2	TNM-T	T4	1
2	TNM-N	N0	1
...	...	...	...

Abbreviation: EAV, Entity-Attribute-Value.

MDR for the description of their data items while adhering to the MDR data types. This approach helped us to achieve our third requirement by facilitating the comparison and translation between the two data sets, semantically and syntactically: A semiautomatic mapping process could help find correlations (semantically) between both metadata definitions, and a data transformation process could use this mapping information to transform the original biobank data (syntactically) into the target OSSE XML representation.

►Fig. 1 outlines the architecture of our ETL approach and its integration with the Samply.MDR (on top) and OSSE (on the right). As indicated with the two big arrows, it processes either an EAV or a flat file and generates an XML import file, compatible with the OSSE system. The smaller arrows represent the data flows processed by our implemented software tools: *MDRExtractor*, *TablePreprocessor*, *MDRMatcher*, *MappingGUI*, and *ETLHelper*. The following section describes these programs and their role in the ETL process.

The BBMRI-ERIC ETL Approach

Metadata and Data Extraction

As a first step, we implemented the *MDRExtractor* program, which recursively extracts data elements, data types, value sets, and hierarchical information from MDR project namespaces via the MDR's REST API. The extracted information is stored in what we call *metadata definition files*. These are cached representations of the MDR's content and are faster to process by the tools compared with using the REST API of the MDR. Each namespace is either associated with a biobank's local terminology, or the central CCDC terminology. For the biobank namespaces, the MDRExtractor generates *local metadata definition files*, whereas for the latter, it produces a *central metadata definition file* (see ►Fig. 1).

We introduced support for the flat file format when it became clear that many biobanks would not be able to provide their data in the EAV format. To allow for the integration of such data, we implemented the program *TablePreprocessor*. It analyzes the flat files and converts their contents into the EAV format. It also generates a simplified local metadata definition file by extracting the information from the file's column headings. These simplified metadata definition files served as a fallback solution in such cases where the biobanks would not enter their local metadata into the MDR.

Lexical Matching

To create the semantic bridge between the biobanks' local and the CCDC's central data elements, it is necessary to create mappings between the two terminologies. To assist the user in this task, we implemented *MDRMatcher*, a tool that matches the local and central metadata definition files to find similar data elements. This results in a *mapping file*. This file contains automatically generated, initial proposals, which we call *matches*. For each term in a biobank's source terminology, MDRMatcher creates a list of possible matches to the target CCDC terminology. If MDRMatcher or a human expert considers one of these matches to be correct, we call it a *mapping*. To put it in other words, a match is a mapping candidate.

The implementation of MDRMatcher follows the assumption that if two metadata items contain the same or similar words, the two must be related. Consequently, the program compares *all* items from the source with *all* items from the target terminology. One such comparison is shown in ►Fig. 2. After *selecting two metadata items* (1), one from the local and one from the central metadata definition file, MDRMatcher performs a *string normalization* (2), where the program reformats the strings to uppercase. It also removes nonalphanumeric characters and duplicate words.

MDRMatcher then performs what we call *semantic expansion* (3). By inserting additional words into the strings to be compared, we augment the metadata entries and support the program in comparing abbreviations or differently named concepts. For example, based on the synonym definition "CT = Computer Tomography," the string "CT DONE" is transformed into "CT COMPUTER TOMOGRAPHY DONE." Other synonyms

Table 2 Example for the flat file format (mock-up data)

Patient	TNM-T	TNM-N	TNM-M	Age	Metastases	Gender	Date of surgery
1	T1	N0	M1	45	**Hepatic; Osseus**	Male	21.07.2013
2	T4	N0	M0	34		Female	23.11.2018
3	T2	N0	M1	21	Osseus	Male	04.04.2017
4	T1	N1	M1	43	Brain; Osseus	Male	19.03.2012
5	T2	N2	M1	76	Hepatic	Female	10.09.2013

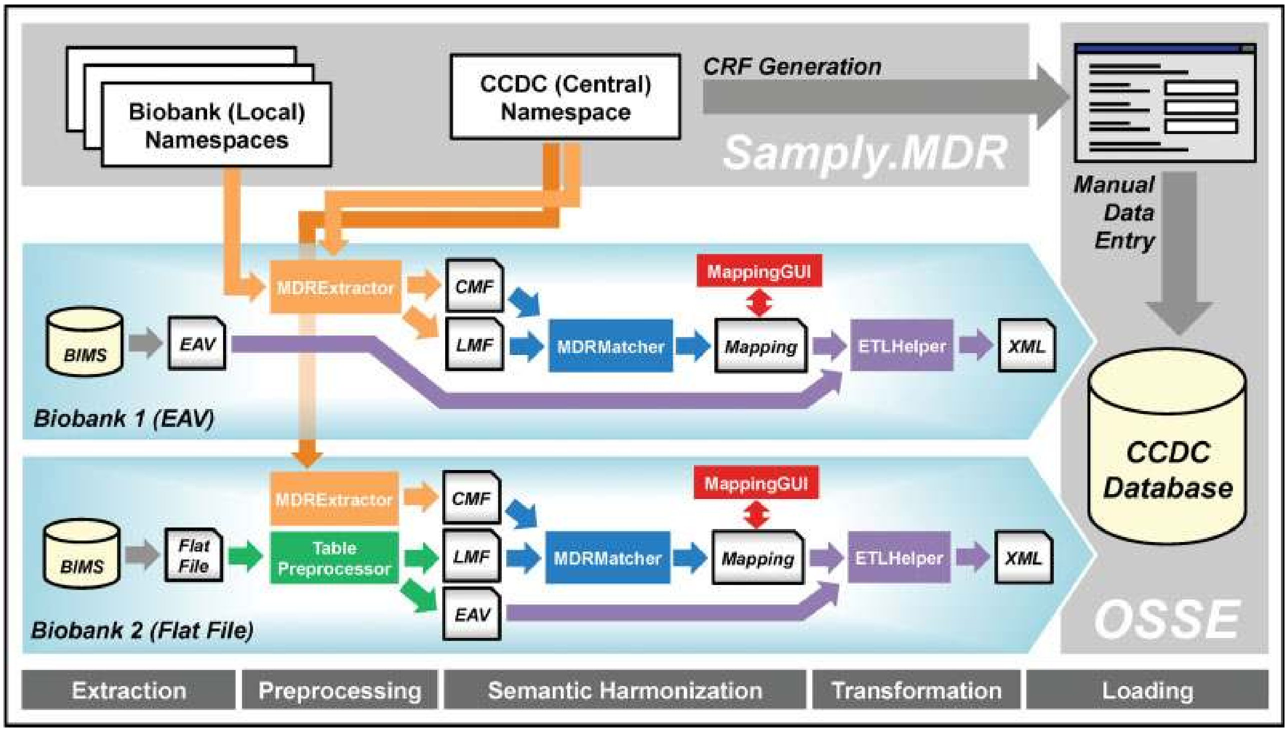


Fig. 1 The extract-transform-load (ETL) pipeline as configured for the ADOPT BBMRI-ERIC project, shown for two exemplary biobanks, one contributing an Entity-Attribute-Value (EAV) and the other a flat file. The files are extracted from the *Biobank Information Management Systems* (BIMS, left), processed by our tools into an XML file, and finally loaded into the OSSE system (right). CMF, central metadata definition file; LMF, local metadata definition file.

defined are, for example, "Sex = Gender," "Osseous = Bone = Skeleton," or "Pulmonary = Lung." We based the small database on the CCDC terminology, but also integrated other common medical abbreviations.[29]

The same technique is also used for tagging metadata items with positive or negative sentiments. Theoretically, this improves matches among different positive (e.g., "Yes," "True") or negative value sets (e.g., "No," "Not," "False"). As an example, consider the match between the source data element "Patient is alive" with the value set {"No," "Yes"} and the target data element "Patient is living" with the value set {"False," "True"}. Even though both value sets are different, the sentiment tagging may support the lexical matcher in finding the correct mappings, for example, where "Patient is alive = No **Neg**" is mapped to "Patient is living = False **Neg**" and "Patient is alive = Yes **Pos**" is mapped to "Patient is living = True **Pos**." In the example in ►Fig. 2, the word "NOT" was augmented with "NEG."

MDRMatcher then performs the actual *lexical matching* (4). The algorithm used is based on the bag-of-words approach.[30,31] It considers sequences of n successive words, called *n-grams*.[31] For this, it first determines g_{min}, which is the lower word count of the two normalized strings being compared (in the example, 4). It extracts the sets with all n-grams with $n \leq g_{min}$ from both sides and puts these into two "bags," one for the source and one for the target item. In the next step, the algorithm compares both bags via exact and fuzzy partial matches (with the latter not shown in ►Fig. 2). Each of these matches contributes to a final similarity score. Our approach is based on the assumption that longer words and in particular word sequences carry more information than short, individual words; this is why we calculate the scores using the product of the number of characters (red) and the gram count (blue). Fuzzy matches are computed only for single words ($n = 1$) by calculating the normalized Levenshtein distance[32] (0...1, with 0 = no similarity and 1 = same string), which is then multiplied with the half-length of the longer term. As a result, fuzzy matches are weighted less than exact matches.

The lists of the best matches with a similarity score above a configurable threshold are then stored in a mapping file. The correct choice of the threshold prevents, on the one hand, that too many unnecessary (wrong) proposals are shown (a threshold of 0 would propose *all* target terms, sorted by relevance) and, on the other hand, that a correct proposal is not shown. When the method was prototyped on the data supplied by the first biobank, it was found that a threshold of one-third of the highest similarity score provided the best results. This value was then used for processing the data of all other biobanks.

MDRMatcher also automatically suggests the best match as a mapping if the highest similarity score is higher than that of the second best match. If multiple top matches share the same similarity score, the program is unable to decide which one is better, and leaves the decision to the human expert.

Curation of Mappings by Human Experts

Human experts, typically the persons responsible for the data at the biobanks, have to complete, verify, and potentially correct the mapping file generated by MDRMatcher. To support this process, we developed a graphical user interface, *MappingGUI*, as shown in ►Fig. 3.

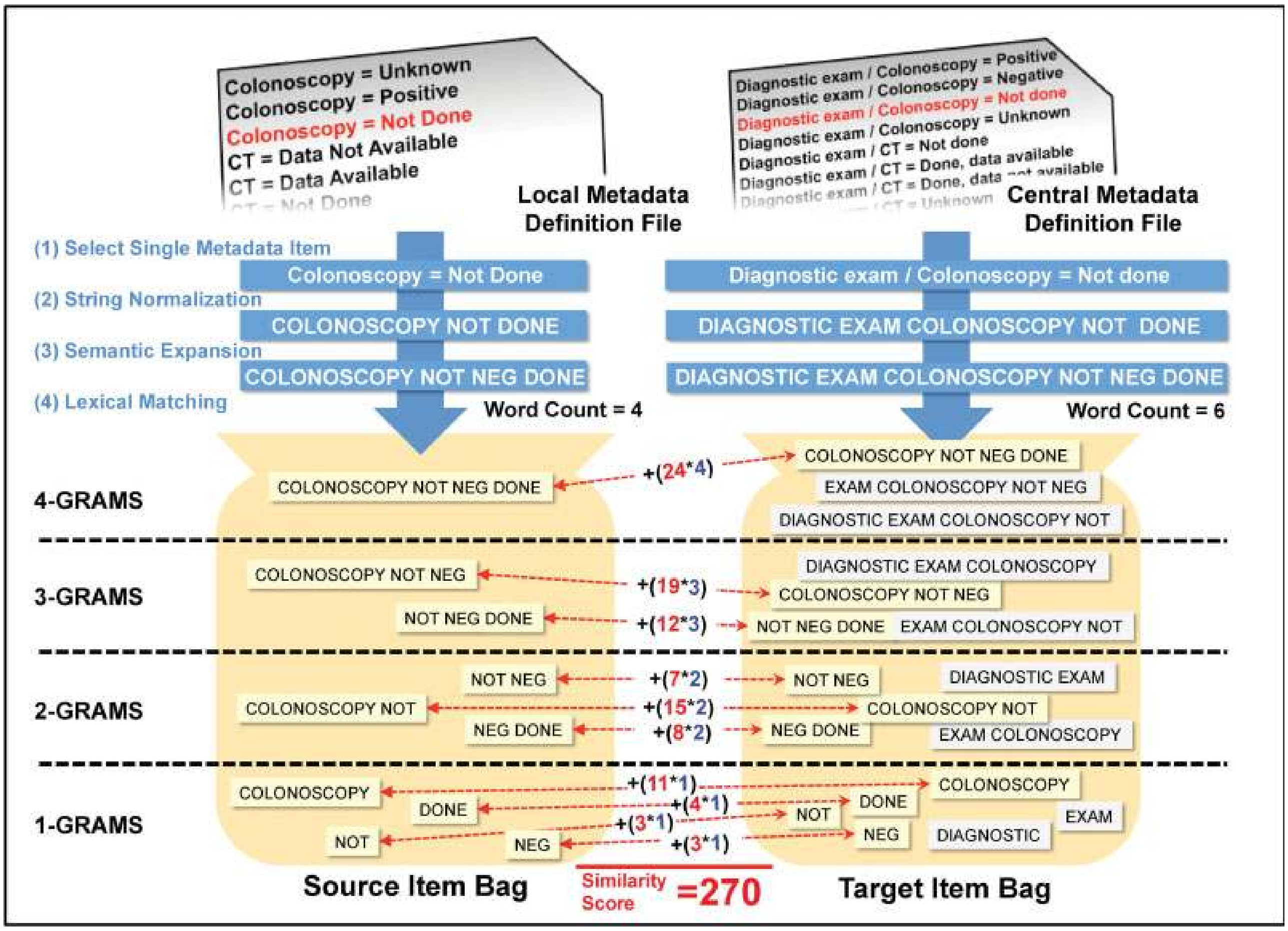


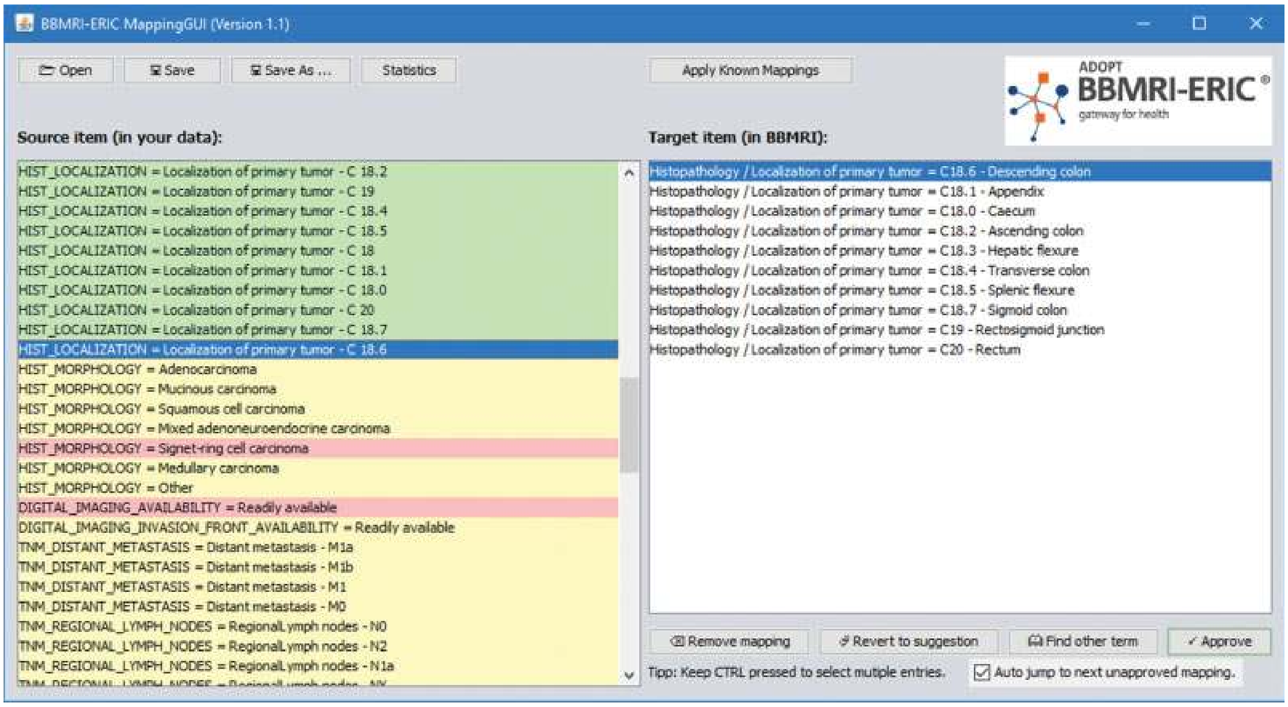
Fig. 2 Illustration of the bag-of-words algorithm in MDRMatcher. After normalization and semantic expansion, it compares all *n*-grams from the source and the target item and computes a similarity score.

The program displays the source terminology from the biobank on the left, and the mapping proposals from MDRMatcher on the right side of the window. The tool displays mappings curated by human experts in green, and mappings proposed by MDRMatcher in yellow. Otherwise, the tool highlights entries in red to indicate that only lexical matches from MDRMatcher are available. The expert can then either approve or correct a mapping by selecting another term from the right side. If the correct term was not among the proposals, the user can open a search window and search for the correct term. It is also possible to remove mappings via the "Remove mapping" button for cases where no mapping exists.

Transformation of Instance Data

The last program in the ETL pipeline is *ETLHelper*, which processes the mapping and EAV file containing the patients' instance data and then generates the OSSE XML import file. This step comprises replacing the source value sets with those from the target terminology, as well as converting between different data types according to the expert-curated mapping rules. The implementation of the ETLHelper uses an internal SQLite database[33] to efficiently sort, filter, and align the data. A single SQL join statement merges the data from the EAV file, the mapping file, and the metadata definition files, thereby creating the semantic bridge between the biobank's data and the CCDC's target terminology. Finally, the program steps through the SQL results, performs the data type conversions, and populates the OSSE XML file with the transformed biobank data. The generated XML file can then be imported into the OSSE system.

As our ETL approach is based on the foundation of Samply.MDR's semantics, we only had to consider the conversion between eight different data types, as shown in ►Table 3. Translations between simple value sets, represented in the MDR via the data type "Enumerated," are handled in the abovementioned SQL statement by replacing the original values with those of the target terminology according to the mapping rules in the mapping file. This is the case for example for the mapping "VITAL_STATUS = ALIVE => Vital status and survival information / Vital status = person is still alive." Here, the translation method between "Enumerated" and "Enumerated" from ►Table 3 applies, and ETLHelper replaces the original "ALIVE" cell entries in the biobank data with "person is still alive." When populating the XML file with the instance data, ETLHelper puts these entries into the respective location in the XML file, as defined in the CCDC namespace in the MDR (Vital status and survival information / Vital status). The translations between other data types are done following common sense, as shown in ►Table 3. If, for example, a float is to be converted into an integer, we round it to the nearest integer. To translate a



Source item (in your data):

HIST_LOCALIZATION = Localization of primary tumor - C 18.2
HIST_LOCALIZATION = Localization of primary tumor - C 19
HIST_LOCALIZATION = Localization of primary tumor - C 18.4
HIST_LOCALIZATION = Localization of primary tumor - C 18.5
HIST_LOCALIZATION = Localization of primary tumor - C 18
HIST_LOCALIZATION = Localization of primary tumor - C 18.1
HIST_LOCALIZATION = Localization of primary tumor - C 18.0
HIST_LOCALIZATION = Localization of primary tumor - C 20
HIST_LOCALIZATION = Localization of primary tumor - C 18.7
HIST_LOCALIZATION = Localization of primary tumor - C 18.6
HIST_MORPHOLOGY = Adenocarcinoma
HIST_MORPHOLOGY = Mucinous carcinoma
HIST_MORPHOLOGY = Squamous cell carcinoma
HIST_MORPHOLOGY = Mixed adenoneuroendocrine carcinoma
HIST_MORPHOLOGY = Signet-ring cell carcinoma
HIST_MORPHOLOGY = Medullary carcinoma
HIST_MORPHOLOGY = Other
DIGITAL_IMAGING_AVAILABILITY = Readily available
DIGITAL_IMAGING_INVASION_FRONT_AVAILABILITY = Readily available
TNM_DISTANT_METASTASIS = Distant metastasis - M1a
TNM_DISTANT_METASTASIS = Distant metastasis - M1b
TNM_DISTANT_METASTASIS = Distant metastasis - M1
TNM_DISTANT_METASTASIS = Distant metastasis - M0
TNM_REGIONAL_LYMPH_NODES = Regional Lymph nodes - N0
TNM_REGIONAL_LYMPH_NODES = Regional Lymph nodes - N2
TNM_REGIONAL_LYMPH_NODES = Regional Lymph nodes - N1a
TNM_REGIONAL_LYMPH_NODES = Regional Lymph nodes - NX

Target item (in BBMRI):

Histopathology / Localization of primary tumor = C18.6 - Descending colon
Histopathology / Localization of primary tumor = C18.1 - Appendix
Histopathology / Localization of primary tumor = C18.0 - Caecum
Histopathology / Localization of primary tumor = C18.2 - Ascending colon
Histopathology / Localization of primary tumor = C18.3 - Hepatic flexure
Histopathology / Localization of primary tumor = C18.4 - Transverse colon
Histopathology / Localization of primary tumor = C18.5 - Splenic flexure
Histopathology / Localization of primary tumor = C18.7 - Sigmoid colon
Histopathology / Localization of primary tumor = C19 - Rectosigmoid junction
Histopathology / Localization of primary tumor = C20 - Rectum

Tipp: Keep CTRL pressed to select mutiple entries. ☑ Auto jump to next unapproved mapping.

Fig. 3 The MappingGUI program, which is used to curate mappings between source and target terms and values.

Boolean to integer, we convert the Boolean "True" to 1 and the Boolean "False" to 0.

The implementation is robust in the sense that it can automatically parse dirty (but correct) data in different date or number formats. However, as indicated by the empty cells in ►Table 3, we do not support all types of transformations. For example, we do not attempt to recognize non-string data (such as time stamps) in strings, and therefore do not offer converting a string into another data type. Whenever the parsing of a data record fails or the mapping contains an unimplemented transformation rule, the user is notified with an error message.

Results

The ETL approach described above was used to implement the data harmonization for the CCDC in the ADOPT BBMRI-ERIC project. From 14 participating biobanks (as of November 2018), 10 were unable to generate the OSSE XML file themselves. Instead, they extracted and anonymized relevant data from

Table 3 Supported data type transformations, with source data types shown on the left and target data types on top

From/To	Enumerated	Integer	Float	Boolean (True)	Boolean (False)	String	Date	DateTime
Enumerated	Target value			**True**	**False**	*Cast to String*		
Integer		*Pass Through*	*Cast to Float*	*if = 1* **True**	*if = 0* **False**	*Cast to String*		
Float		*Round to n. Int.*	*Pass Through*	*if = 1.0* **True**	*if = 0.0* **False**	*Cast to String*		
Boolean (True)	Target value	**1**	**1.0**	*Pass Through*	**False**	*Cast to String*		
Boolean (False)	Target value	**0**	**0.0**	**True**	*Pass Through*	*Cast to String*		
String						*Pass Through*		
Date						*Cast to String*	*Pass Through*	*Date Only*
DateTime						*Cast to String*	*Date Only*	*Pass Through*

Note: Italics indicates the type of handling, bold the return value. Blank indicates no transformation rule.

their source systems and converted them into the flat file format. These files were then collected via a secured Web portal, the *CCDC Uploader*, and the ADOPT BBMRI-ERIC IT personnel was notified about new uploads. It downloaded the files and performed initial data quality checks to assess the overall suitability for the CCDC. They also resolved easily correctable problems, such as minor formatting or encoding errors; however, when major changes in data were necessary, the biobanks were consulted. The IT personnel then created the mappings in close collaboration with the biobank experts. In most cases, it was clear how the mappings were to be created; however, in other cases the biobank data experts had to be consulted. The final mappings, which we also used as the "gold standard" for the evaluation of MDRMatcher as described below, were finally approved by the biobanks. The biobanks were constantly informed about interim ETL results. The XML files produced by ETLHelper were XSD-validated and uploaded into the OSSE system. Finally, the biobanks were given the opportunity to review their data in the OSSE system.

Data Received

During this task, we observed different types of problems with the biobank data. Many of them were identified directly by the BBMRI-ERIC IT personnel upon receiving the data files, but most were detected later in the process through mechanisms implemented into the harmonization tools.

Shifted tabular data was the most common problem, which we attribute to copying and pasting of data. This could also have been the cause of other formatting problems, such as inconsistent upper and lower case or spaces at the beginning or end of cell entries. Character encoding errors, which were potentially caused by office software, also occurred frequently.

The 10 biobanks made great efforts to provide as much information as possible. Unfortunately, this also means that they were reluctant to leave table cells empty, even where this would have been appropriate. Entries, such as "Unknown," "Not available," or "N/A" were used frequently, but also more uncommon values, such as "ND" for "not done." The biobanks also used textual entries in numerical and date fields, which, as expected, led to data rejections in ETLHelper. This is the desired behavior of the software, because such entries do not contribute any relevant information.

Analysis of Lexical Matching

We used MDRMatcher to match the metadata items from the 10 biobanks to the target CCDC terminology. To assess the quality of this process and to be able to improve our approach in the future, we compared the program's output against the final, expert-curated mappings. We designed a four-axial classification scheme to categorize the behavior of MDRMatcher on a per-source-item basis. It is comprised of the four dimensions *conceptual*, *mapping*, *correctness*, and *matching*, as illustrated in ►Fig. 4.

The *conceptual dimension* describes whether a source data item has an equivalent in the target terminology or not. In other words, it describes on a conceptual level whether a mapping can be created. This can be easily determined in the final mapping file (the "gold standard") by checking whether the human experts left a mapping blank or checking whether the expert kept, modified, or removed a proposed mapping. If there was an equivalent item in the target terminology, it can be analyzed whether MDRMatcher has created a mapping or not, which we capture in the *mapping dimension*. The *correctness dimension* evaluates whether this mapping is correct or not by comparing the proposed mapping with the expert-curated mapping. The *matching dimension* determines whether the correct item from the target terminology was among the list of matches, or whether it has been cut off because its score was below the implemented cut-off threshold.

Each of these four dimensions can be answered with "Yes" or "No" (see ►Fig. 4), and as such, a combination would result in $2^4 = 16$ classes. However, only seven of these are applicable

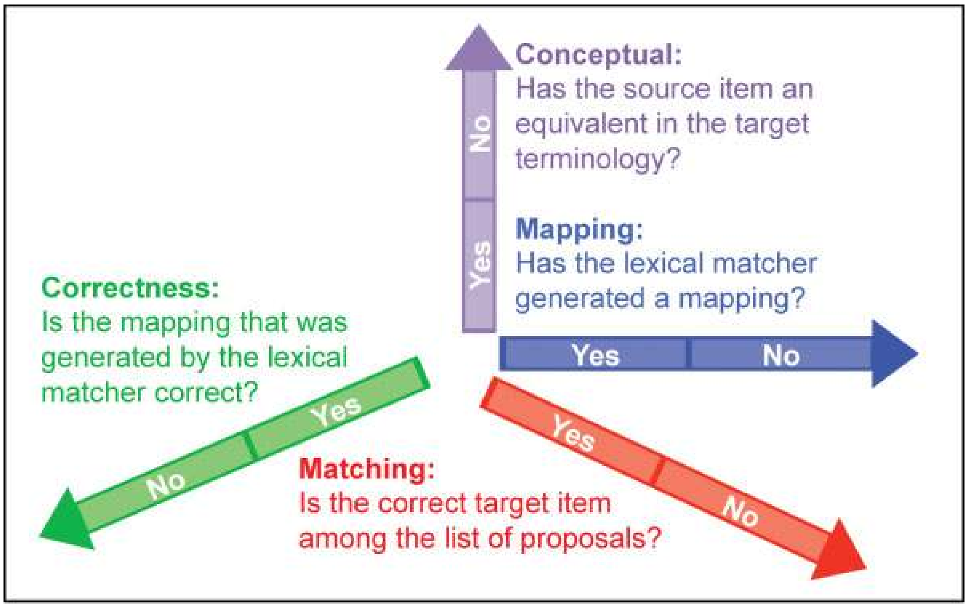

Conceptual:
Has the source item an equivalent in the target terminology?
No
Yes
Mapping:
Has the lexical matcher generated a mapping?
Yes
No
Correctness:
Is the mapping that was generated by the lexical matcher correct?
Yes
No
Yes
No
Matching:
Is the correct target item among the list of proposals?

Fig. 4 A four-axial classification scheme to assess the mapping quality of MDRMatcher.

in practice, as some of them do not make sense. For example, an automatically created mapping cannot be correct if there is no equivalent item in the target terminology. The seven valid classes are, ordered from "good" to "bad":

1. A mapping was created and was correct.
2. No mapping was created, which is correct because there is no equivalent item in the target terminology to map to.
3. No mapping was created, but should have been. At least the correct term was in the list of proposals, which later support the curation process by the human expert.
4. A mapping was created, but it was wrong. The correct item was in the list of proposals.
5. A mapping was created, but it was wrong. The correct item was not in the list of proposals.
6. A mapping was created, which is wrong, because there is no equivalent item in the target terminology.
7. No mapping was created, but should have been. The correct term was not in the list of proposals.

In the next step, we classified the mappings generated by MDRMatcher according to these seven classes. The MappingGUI program kept track of all modifications to the mapping files and thus allowed us to automatically compare the software-generated mappings with the final, expert-curated ones. The result is shown in ►Table 4. In total, we matched 1,492 biobank source items to the 226 target items of the CCDC target terminology. As shown, 77.21% + 1.27% = 78.48% of all automatically generated mappings by MDRMatcher were correct (classes 1 and 2). For 20.24%, the correct mapping was among the suggestions and could be selected easily in MappingGUI by the human experts (classes 3 and 4). In other words, 98.72% of the mappings were correct or easily correctable (classes 1–4) with our tools. In only 1.27% (classes 5–7) of all cases, the correct target item was not among the list of suggestions and had to be looked up manually.

Based on ►Table 4, it is possible to derive additional standard metrics to evaluate MDRMatcher's role as an automatic mapper. Based on a confusion matrix as shown in ►Table 5, it is possible to calculate the precision, recall, and F_1 score:

$$Precision = \frac{TP}{TP+FP} \approx 0.89,$$

$$Recall = \frac{TP}{TP+FN} \approx 0.87,$$

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \approx 0.88$$

We then analyzed why MDRMatcher did not succeed in finding the correct mappings for the terms in classes 3 to 7, with a deeper analysis of classes 3 to 5. The results for the classes 3 to 5 are summarized in ►Table 6.

The most common type of wrong behavior in MDRMatcher could be traced back to a lack of synonym definitions, for example, for "ND" ("Not Done") or "X" ("Lung Imaging," as in "X-ray"). Defining these commonly used words in the synonyms database would have resulted in correct automatic mappings.

Table 4 Assessment of the automatic mapping quality for metadata items received

Property		Class						
		1	2	3	4	5	6	7
Dimension	Conceptual *(Equivalent in Target?)*	Yes	No	Yes	Yes	Yes	No	Yes
	Mapping *(Created A Mapping?)*	Yes	No	No	Yes	Yes	Yes	No
	Correctness *(Correct Behavior?)*	Yes	Yes	No	No	No	No	No
	Matching *(Among Proposals?)*	Yes	No	Yes	Yes	No	No	No
Source Item Counts	Biobank 1	163	10	17	11	2	2	0
	Biobank 2	121	1	21	17	0	1	0
	Biobank 3	83	7	36	14	2	1	2
	Biobank 4	94	0	12	8	2	0	0
	Biobank 5	89	0	14	13	0	1	0
	Biobank 6	117	0	7	9	0	0	0
	Biobank 7	142	0	17	18	1	0	0
	Biobank 8	117	0	18	12	2	0	1
	Biobank 9	97	1	12	18	0	0	0
	Biobank 10	129	0	16	12	1	1	0
	Total	1152	19	170	132	10	6	3
	Percent	77.21	1.27	11.39	8.85	0.67	0.40	0.20

Note: The numbers are per source item, which is comprised of concept and value, e.g., "UICC_STAGE = Not known."

Table 5 Confusion matrix for the evaluation of MDRMatcher as an automatic mapper

	TP + FN **Should create mapping** **1171**	*FP + TN* **Should not create mapping** **321**
TP + FP **Created mapping** **1300**	*TP* Created mapping, correct 1152	*FP* Created mapping, wrong 132 + 10 + 6 = 148
FN + TP **Did not create mapping** **192**	*FN* Did not create mapping, wrong 170 + 3 = 173	*TN* Did not create mapping, correct 19

Table 6 Analysis of wrong behavior of MDRMatcher per target data element in classes 3 to 5 for metadata items

Type of problem	Class 3	Class 4	Class 5	Total
Missing synonyms	22			22
No down-ranking	21			21
N-gram matching across hierarchy/ concept/value	3	6	4	13
Wrong up-ranking due to bad fuzzy matches	3	8		11
Unfavorable removal of redundant words	3	8		11
Inability to compare Roman with Arabic number	1	8		9
Use of different languages	5			5
Spelling errors or inconsistent naming		5		5
Unable to map something to “Other”	2			2
Other	4	12		18

Similarly, the missing ability of down-ranking metadata items in the matching algorithm has often proven to be a problem. MDRMatcher uses only matches to increase (or up-rank) the similarity score, but it does not consider “non-matches”; that is, text passages that do *not* occur on both sides. For example, for the source item “MM_**MISMATCH_REPAIR**_GE = **EXPRESSION**,” the matcher could not decide if “Molecular markers / **Mismatch repair** gene expression = **Expression**” or “Molecular markers / **Mismatch repair** gene expression = Loss of **expression**” is the correct match. In both cases, the algorithm found the same partial matches (as indicated in bold) and thus calculated the same similarity score, but the additional information (“Loss of …”) was not considered.

The implemented *n*-gram-based matching approach theoretically allows hierarchical information to be taken into account, because all metadata items to be matched are built using strings in the form “Hierarchy Path / … / Concept = Value,” from generic to specific. In practice, however, this sometimes led to the exact opposite effect. As the slashes (/) and equal signs (=) were removed from the data during preprocessing, misleading matches across these components occurred. For example, “DIAG_**MRI_DONE** = Not done” was incorrectly matched to “Diagnostic exam / **MRI** = **Done**, data not available” instead of the correct “Diagnostic exam / MRI = Not done.”

Wrong up-rankings due to incorrect additional fuzzy matches were another common problem. For example, the International Classification of Diseases, Tenth Revision (ICD-10) code “C 19” was incorrectly mapped to “C 19.9” instead of the correct “C 19,” because in addition to the correct 2-gram match between “C 19” and “C 19,” the tool incorrectly created a second fuzzy match between “19” and “9.”

The removal of redundant words during the preprocessing, which we believed prevented overfitting, also caused problems in practice. For example, MDRMatcher incorrectly proposed mappings between NRAS and KRAS, due to a fuzzy match between “NRAS” and “KRAS” and the removal of duplicate words in the hierarchy during text preprocessing. “NRAS” and “KRAS” were part of the hierarchical grouping (e.g., “Molecular markers / KRAS mutation status / NRAS exon 2 [codons 12 or 13] = Not mutated”) in the MDR. This combined type of error affected 9 biobanks and could have been prevented if the text preprocessing would not have removed duplicate words.

MDRMatcher proposed some wrong mappings because it failed to match Roman to Arabic numbers (e.g., for the UICC stages). Bad spelling or inconsistent naming (also within the CCDC target terminology) also led to incorrect mappings. For example, the ICD-10 code “C18.0” was mixed up for surgery and histopathology due to inconsistent writing of “caecum” (vs. “cecum”) in the CCDC target terminology. Similarly, the misspelled value “3th” was mapped to “4th,” because for the Levenshtein algorithm “3th” is closer to “4th” than “3th” is to “3rd.”

Finally, MDRMatcher was not able to map certain biobank terms to “Other” values in the target terminology. This was the case, for example, when a biobank specified metastasis localizations for which there were no equivalent entries in the target terminology. This is where the software simply lacks the intelligence to make the correct assignment to “Other.” As a result, the human expert must complete the mappings.

Class 6 errors (mapping despite no representation in target terminology) were incurred, because the lexical matcher always proposes the “best” match as a mapping if its score is higher than the second-best matching. For example, for “UICC_STAGE = Not known” it incorrectly proposed “Histopathology / UICC staging / UICC version = Not known.”

Class 7 did not contain enough mappings to discern any recurring pattern.

Table 7 ETL results for the actual facts data received from 10 biobanks participating in the CCDC pilot as reported by ETLHelper.

Property			**Biobank**										
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
Input data													
1	Patients		1,066	218	50	55	300	600	300	308	218	300	3415
2	Data records		50,020	8,756	2,067	2,477	11,984	22,617	13,225	13,729	9,160	13,573	147,608
3	Different concepts		53	47	49	47	47	43	49	52	40	56	483
4	Different source items		205	161	145	116	117	133	178	150	129	158	1,492
Mapping rules													
5	Mappings between source and target items		193	159	134	116	116	133	178	150	127	158	1,464
6	Data records that should have a mapping		50,020	8,756	2,067	2,477	11,984	22,617	11,325	13,729	9,160	13,573	145,708
7	Data records that do have a mapping		49,263	8,331	1,931	2,477	11,983	22,617	11,325	13,729	9,148	13,573	144,377
8	Percentage of data records that have a mapping		98.5%	95.1%	93.4%	100.0%	100.0%	100.0%	100.0%	100.0%	99.9%	100.0%	99.1%
9	Data records that don't have a mapping		**757**	**425**	**136**	0	**1**	0	0	0	**12**	0	**1331**
Data type transformations													
10	Total number of transformations		49,263	8,331	1,931	2,477	11,983	22,617	13,225	13,729	9,148	13,573	146,277
11	**Good**	*Enumerated ⇒ Enum.*	35,786	6,544	1,383	1,907	9,436	19,350	10,294	10,798	7,622	10,234	113,354
12		*Enumerated ⇒ Boolean*	1,119		100	28				308	218	91	1,864
13		*Integer ⇒ Integer*	7,093	815	230	300	995	2,568	2,140	1,627	436	1,851	18,055
14		*Float ⇒ Integer*		102			300				218		620
15		*String ⇒ String*	2,654	435	126	55	636	98	191	380	218	797	5,590
16		*Date ⇒ Date*	2,132	388	92	110	652	600	600	616	436	600	6,226
17		Total number of good transformations	48,784	8,284	1,931	2,400	11,983	22,616	13,225	13,729	9,148	13,573	145,673
18	**Bad**	*Integer ⇒ Integer*	**479**	0	0	**77**	0	**1**	0	0	0	0	**557**
19		*Date ⇒ Date*	0	**47**	0		0	0	0	0	0	0	**47**
20		Total number of bad transformations	**479**	**47**	0	**77**	0	**1**	0	0	0	0	**604**
21	Percentage of good transformations		99.0%	99.4%	100.0%	96.9%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	99.6%
Summary													
22	Percentage of input data records that could be mapped and transformed		97.5%	94.6%	93.4%	96.9%	100.0%	100.0%	100.0%	100.0%	99.9%	100.0%	98.7%

Abbreviations: CCDC, Colon Cancer Data Collection; ETL, extract-transform-load.
Note: Rejected data records are printed in bold.

Analysis of Data Transformation

We used the user-validated mappings to transform the biobank data into the XML import format. ►Table 7 summarizes the whole ETL process for the data of the 10 biobanks. In total, we were able to process 98.7% of 147,608 data records from 3,415 patients. As we will describe later, we consider the remaining 1.3% as correct rejections.

►Table 7 consists of four sections. The first one describes the input data. When loading the biobank data, ETLHelper first counts the number of patients (row 1) and the number of data records, the number of nonempty cells in the original flat file (row 2). The number of different concepts (row 3) corresponds to the number of columns in a flat file, whereas the source items count (row 4) is the number of concept-value combinations. The second section (rows 5–9) describes the availability of mapping-rules, where ETLHelper evaluates for how many of the data records (not metadata items) a mapping rule can be utilized to transform the actual patient data. A low percentage in row 8 indicates that a biobank's data did not match the target terminology well, or that the mapping needs to be revised. The third section (rows 10–21) summarizes the number and kind of data type validations and transformations utilized. Finally, row 22 outlines the total percentage of records that we were able to transform into the target CCDC representation.

Although the ETL process delivered overall good results, where most biobanks achieved a transformation rate of or close to 100%, we were interested in what happened in the remaining cases. We therefore analyzed two aspects: First, why the human expert did not map a source data item, and second, why certain data type castings resulted in errors.

As mentioned above, many biobanks (1, 2, 3, and 9) provided data with "unknown" values that did not contribute any useful information. Examples are entries such as "UICC_STAGE = Not known" or "WHO_GRADE = Not known." Additionally, biobank 1 provided data for a data element "Targeted Therapy Scheme" with nine enumerated values, which was not collected in the CCDC. Similarly, biobank 5 contributed entries with "Localization of primary tumor = C18.9," which were outside of the inclusion criteria of the colorectal cancer cohort. Such entries were deliberately unmapped by the human experts and thus appear in ►Table 7 as data records that do not have a mapping (row 9). ETLHelper then rejected these unmapped data.

Similarly, the root cause for the data type casting errors (rows 18–20), which affected "INTEGER" and "DATE" entries, was the utilization of string values for numeric data types, in most cases entries such as "Unknown." As these could not be parsed as numeric values or timestamps, ETLHelper correctly rejected them.

Discussion and Outlook

Explanation of the Results

ADOPT BBMRI-ERIC was facing the challenge of supporting a cross-intuitional data collection project, the CCDC. From the beginning it was not clear in which format the biobank data would be provided, therefore we defined two simple file formats. We sought to convert these data into a representation that could be imported into a central research database via a generic ETL approach. We successfully supported this process via a semiautomatic matching approach, as implemented in MDRMatcher. It correctly mapped 78.48% of the 1,492 local biobank terms from 10 European biobanks, achieving a precision of 0.89, a recall of 0.87, and an F_1 score of 0.88. With ETLHelper, the actual instance data transformation tool, we successfully processed 147,608 data records from 3,415 patients.

These numbers must be read with caution, however. For example, it has to be considered that the biobanks were informed in advance which data elements are collected in the CCDC. We do not know to what extent they had preprocessed their data, but in many cases the designators of the data elements and value sets were largely identical to those in the CCDC. Differences in data preprocessing could also explain the different mapping results for the individual biobanks. Similarly, the achieved data transformation rate of the instance data with ETLHelper is not surprising, as experts curated the mappings in an iterative process. Rather, these figures confirm that the implemented ETL approach succeeded in successfully transforming the biobank data into the OSSE import format.

Comparison with Related Research

Supporting information integration via automation has a well-established background in research. In particular, schema and instance matching techniques have been thoroughly investigated in the past (e.g., see refs. 34–38). Some use advanced background knowledge to improve matching results, such as ontologies.[39] These approaches aim to provide generic frameworks and may therefore require the utilization of additional techniques, such as metadata discovery.[40,41] Similar techniques are used in the field of ontology matching.[42,43] According to Euzenat and Shvaiko,[43] MDRMatcher can be classified as a string-based matcher that incorporates some informal resources, such as the synonyms database. The evaluation results of MDRMatcher are similar to those of other matchers (see, e.g., refs. 44 and 45). However, as already stated above, these numbers must be interpreted with caution. To fully compare the results, we would have to evaluate MDRMatcher with a known data set, as done in, e.g., Kock et al[46] and Achichi et al.[44]

In the biobanking field, we are aware of only a few similar works. Pang et al presented two related methods for mapping biobank data.[45,47] Both are based on lexical matching and use synonyms. Compared with our work, the authors start supporting the data collection and harmonization process at an earlier stage by implementing an initial search of the desired research data elements in the source systems. The authors do this by matching from the target terminology to the source terminology (another reason why their metrics are not directly comparable to ours). Adopting this approach could also be useful for the future of BBMRI-ERIC, where larger data sets need to be recognized for future project purposes from biobank databases. However, it must be clarified whether and in which format biobanks could provide their complete metadata. Based on our experiences from the CCDC, this could prove to be very challenging.

Limitations and Future Work

The matching algorithm in MDRMatcher was not modified throughout the execution of the CCDC to generate consistent results. However, certain peculiarities in the behavior of the matching component were identified in our analysis, and these need to be taken into account in future revisions of the program. For example, we will have to extend the program's synonyms database, integrate functionality for down-ranking, and prevent matches across the components of metadata items (hierarchy, concept, and value). As identified in our analysis, simple improvements of the fuzzy matcher (e.g., preventing fuzzy matches across numbers) and deactivating the removal of redundant words should also improve the results further. With regard to missing synonyms, a potential solution could be to use BioPortal ontologies[48] as a source for synonym definitions, as implemented by Pang et al.

The analysis of MDRMatcher' results in the "Analysis of Lexical Matching" section is very detailed, but still has its limitations. For example, we did not analyze in detail the impact of synonyms and sentiment tagging—we only noticed afterwards that synonym definitions were missing. If we succeed in integrating further terminological resources as described above, such aspects should also be analyzed in future research.

Consideration must also be given to the data transformation part as implemented in ETLHelper. Although our approach was designed to process unstandardized data, it makes sense to investigate how it can efficiently handle data that is already standardized (e.g., ICD-10 or ICD-O-3 codes). Here, one could possibly reuse already existing mappings from ontologies (e.g., UMLS,[49] OxO[50]) and other mapping resources (such as between SNOMED CT and ICD-10[51]). This would avoid the lexical matching of already standardized, mapped terms.

ETLHelper does not yet support automatic unit conversion. This feature was not required in the CCDC, but support for it would be useful, for example, for supporting laboratory parameters. Because Samply.MDR supports storing units, conversion rules could potentially integrated into ETLHelper.

The ETL method presented in this article only supports 1:1, but no n:1 mappings. In other words, there is currently no possibility to merge information from multiple data elements in the source terminology into one data element of the target terminology. This was sufficient for the CCDC, because small calculations could be done directly in the flat files (such as the calculation of the patient age at diagnosis). The definition of machine-interpretable, complex mappings has already been investigated in the past (see, e.g., Mate et al[52]). The *German Biobank Node* and *German Biobank Alliance* projects,[10,23,53] which are working on establishing the German national node for BBMRI-ERIC and whose members are active contributors to the Samply software environment, are currently discussing how the MDR could be extended to store such transformation rules.

Generalizability of Our Method

The approach presented here is not limited to the biobanking domain. The individual steps of the ETL process were implemented using modular software components to enable the logical separation of the individual tasks. Each tool can be used independently from the others. As such, the MIRACUM project[25] is currently investigating whether MDRMatcher can be used to map local laboratory terms to the LOINC standard.[54] As part of this work, MDRMatcher has been extended with multithreading capabilities and a new matching algorithm that considers word frequencies to up-rank rare and down-rank frequent words. The implementation of ETLHelper is still geared toward generating an OSSE XML import file, but modifying the source code allows for generating other output formats as well. ETLHelper uses the EAV format internally to represent the instance data, and as such, the processed data can be transformed into any output format.

During the CCDC it became apparent that the creation of metadata in the MDR was a challenging task for the biobanks. We attribute this to the fact that the use of the MDR is initially complicated and time-consuming. Since it was hardly used by the biobanks (and if it was, then often incorrectly), we developed the workaround of extracting the reduced metadata from the flat files. Although this has worked well for the CCDC, it is unlikely to be sufficient in a future context. Due to the increased complexity and heterogeneity of upcoming research scenarios, semantically richer metadata will be required. An important finding is that organizational and administrative aspects must be given even more attention in advance. For BBMRI-ERIC this means that it will have to ensure that biobanks routinely maintain their metadata in the MDR. In addition, as target terminologies will continue to evolve, a workflow will be required to enable biobanks to keep their mappings up to date.

Conclusion

The reuse of biomedical data for research is a challenging task, especially when data from multiple sources are to be merged. These challenges arise, one the one hand, from syntactic and semantic differences in the source data, and on the one hand, from potential data quality issues.

We presented an ETL approach for the integration of heterogeneous data that provided good results in the CCDC use case. On the one hand, the lexical matching process in MDRMatcher did most of the work of identifying the correct mappings, with the bag-of-words algorithm achieving a fully automatic, correct mapping almost 80% of the time. On the other hand, the generic approach of data type conversion in ETLHelper proved to be suitable for performing the subsequent data transformations and for detecting errors in the biobank data. The approach has thus provided considerable support to the CCDC of ADOPT BBMRI-ERIC to date.

The source code of our tools, including full documentation and a small demo data set with artificial data, is available under the GPL3 license on GitHub: https://github.com/seb-mate/ADOPT-BBMRI-ERIC-ETL-Tools.

Clinical Relevance Statement

The collective analysis of patient data and biomaterials is becoming increasingly important. The use of uniform metadata simplifies this process, as the data can be uniformly checked and processed by only a few software components. In this article, we report on a working, largely automated approach for merging heterogeneous biobank data.

Multiple Choice Questions

1. When collecting the ADOPT CCDC data from the different European biobanks, what was the most common problem associated with the biobanks' data quality?
   a. Missing or erroneous data.
   b. Missing synonym definitions.
   c. Copy and paste errors.
   d. Encoding errors.

   Correct Answer: The correct answer is option c. Shifted tabular data was the most common problem, which we attribute to copying and pasting of data.

2. Sentiment tagging can improve the matching quality by
   a. Up-ranking positive and down-ranking negative terms.
   b. Up-ranking negative and down-ranking positive terms.
   c. Creating matches between different positive and negative terms.
   d. Creating matches between different positive or negative terms.

   Correct Answer: The correct answer is option d. Sentiment tagging "forces" the matcher to create matches only between two positive or between two negative terms. This reduces the likelihood of a match being created between a positive and a negative term.

Protection of Human and Animal Subjects

The experiments were performed using anonymized patient data. The authors therefore declare that this study was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects.

Funding

The present work has been co-funded by ADOPT BBMRI-ERIC supported by EU Horizon 2020, grant agreement no. 676550. It was performed in (partial) fulfillment of the requirements for obtaining the degree "Dr. rer. biol. hum." from the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) (SM).

Conflict of Interest

None declared.

Acknowledgments

The authors would like to thank Julian Gründner for proofreading the manuscript and for providing additional feedback.

References

1 Debnath M, Prasad GBKS, Bisen PS. Molecular Diagnosis in the Post Genomic and Proteomic Era. In: Molecular Diagnostics: Promises and Possibilities. Dordrecht Heidelberg London New York: Springer; 2010:520

2 Lin Y, Chen J, Shen B. Interactions between genetics, lifestyle, and environmental factors for healthcare. Adv Exp Med Biol 2017; 1005:167–191

3 Futreal PA, Coin L, Marshall M, et al. A census of human cancer genes. Nat Rev Cancer 2004;4(03):177–183

4 Reddy PH. Can diabetes be controlled by lifestyle activities? Curr Res Diabetes Obes J 2017;1(04):x

5 Yegambaram M, Manivannan B, Beach TG, Halden RU. Role of environmental contaminants in the etiology of Alzheimer's disease: a review. Curr Alzheimer Res 2015;12(02):116–146

6 Katsios C, Roukos DH. Individual genomes and personalized medicine: life diversity and complexity. Per Med 2010;7(04): 347–350

7 Kinkorová J. Biobanks in the era of personalized medicine: objectives, challenges, and innovation: overview. EPMA J 2016; 7:4

8 van Ommen G-JB, Törnwall O, Bréchot C, et al. BBMRI-ERIC as a resource for pharmaceutical and life science industries: the development of biobank-based Expert Centres. Eur J Hum Genet 2015;23(07):893–900

9 Proynova R, Alexandre D, Lablans M, et al. A decentralized IT architecture for locating and negotiating access to biobank samples. Stud Health Technol Inform 2017;243:75–79

10 Lablans M, Kadioglu D, Mate S, Leb I, Prokosch H-U, Ückert F. Strategies for biobank networks. Classification of different approaches for locating samples and an outlook on the future within the BBMRI-ERIC [in German]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2016;59(03):373–378

11 Lablans M, Kadioglu D, Muscholl M, Ückert F. Exploiting distributed, heterogeneous and sensitive data stocks while maintaining the owner's data sovereignty. Methods Inf Med 2015;54(04): 346–352

12 Schröder C, Heidtke KR, Zacherl N, Zatloukal K, Taupitz J. Safeguarding donors' personal rights and biobank autonomy in biobank networks: the CRIP privacy regime. Cell Tissue Bank 2011;12 (03):233–240

13 Litton J-E. Launch of an infrastructure for health research: BBMRI-ERIC. Biopreserv Biobank 2018

14 Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA Cancer J Clin 2015;65(02): 87–108

15 Vuorio E. Networking Biobanks Throughout Europe: The Development of BBMRI-ERIC. In: Hainaut P, Vaught J, Zatloukal K, Pasterk M, eds. Biobanking of Human Biospecimens: Principles and Practice. Biobanking of Human Biospecimens: Principles and Practice. Cham: Springer; 2017:137–153

16 BBMRI-ERIC. BBMRI-ERIC Annual Report 2017. bbmri-eric.eu. 2017

17 Sellis TK, Simitsis A. ETL Workflows: From Formal Specification to Optimization. In: Ioannidis Y, Novikov B, Rachev B, eds. Advances in Databases and Information Systems. ADBIS 2007. Lecture Notes in Computer Science. Vol 4690. Berlin Heidelberg: Springer; 2007

18 Simitsis A, Vassiliadis P, Sellis T. Optimizing ETL processes in data warehouses. Proc Int Conf Data Eng 2005;•••:564–575

19 Kimball R, Ross M. The Data Warehouse Toolkit-The Complete Guide to Dimensional Modeling. 2nd ed. Hoboken, NJ, USA: John Wiley & Sons; 2002

20 Storf H, Schaaf J, Kadioglu D, Göbel J, Wagner TOF, Ückert F. Registries for rare diseases: OSSE - an open-source framework for technical implementation [in German]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2017;60(05):523–531

21 Kadioglu D, Weingardt P, Ückert F, Wagner T. Samply.MDR – Ein Open-Source-Metadaten-Repository. German Medical Science GMS Publishing House; 2016. Available at: http://www.egms.de/static/de/meetings/gmds2016/16gmds149.shtml. Accessed August 8, 2019

22 medinfo_mainz — Bitbucket [Internet]. bitbucket.org. Available at: https://bitbucket.org/medinfo_mainz/ Available at: September 11, 2018

23 Mate S, Kadioglu D, Majeed RW, et al. Proof-of-concept integration of heterogeneous biobank IT infrastructures into a hybrid biobanking network. Stud Health Technol Inform 2017; 243:100–104

24 Schlue D, Mate S, Haier J, Kadioglu D, Prokosch H-U, Breil B. From a content delivery portal to a knowledge management system for standardized cancer documentation. Stud Health Technol Inform 2017;243:180–184

25 Prokosch H-U, Acker T, Bernarding J, et al. MIRACUM: Medical Informatics in Research and Care in University Medicine. Methods Inf Med 2018;57(S 01):e82–e91

26 Prokosch H-U. Datenmodellierung und Datenbankdesign für relationale Datenbanken. Software Kurier für Mediziner und Psychologen. 1991;4:39–45

27 Nadkarni PM, Marenco L, Chen R, Skoufos E, Shepherd G, Miller P. Organization of heterogeneous scientific data using the EAV/CR representation. J Am Med Inform Assoc 1999;6(06):478–493

28 BBMRI-ERIC. ADOPT BBMRI-ERIC CCDC Terminology [Internet]. mdr.osse-register.de. Available at: https://mdr.osse-register.de/view.xhtml?namespace=ccdg. Accessed January 2019

29 List of medical abbreviations - Wikipedia [Internet]. en.wikipedia.org. Available at: https://en.wikipedia.org/wiki/List_of_medical_abbreviations. Accessed June 17, 2019

30 Brownlee J. A Gentle Introduction to the Bag-of-Words Model [Internet]. machinelearningmastery.com. 2017. Available at: https://machinelearningmastery.com/gentle-introduction-bag-words-model/. Accessed September 5, 2018

31 Jurafsky D, Martin JH. Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2nd ed. Upper Saddle River, NJ, USA: Pearson Prentice Hall; 2009

32 Levenshtein V. Binary codes capable of correcting deletions, insertions and reversals. Sov Phys Dokl 1966;10(08):707

33 Allen G, Owens M. The Definitive Guide to SQLite. 2nd ed. Berkely, CA, USA: Apress; 2010

34 Rahm E, Bernstein PA. A survey of approaches to automatic schema matching. VLDB J 2001;10(04):334–350

35 Bernstein PA, Melnik S, Churchill SE. Incremental Schema Matching. Proceedings of the 32nd International Conference on Very Large Data Bases. VLDB Endowment; 2006:1167–1170

36 Engmann D, Massmann S. Instance Matching with COMA++. BTW Workshops; 2007

37 Papotti P, Torlone R. Schema Exchange: Generic Mappings for Transforming Data and Metadata. Data Knowl Eng 2009;68(07): 665–682

38 Bernstein PA, Madhavan J, Rahm E. Generic Schema Matching, Ten Years Later. Proceedings of the VLDB Endowment 20114(11): 695–701

39 Aleksovski Z, Klein M, Kate ten W, van Harmelen F. Matching Unstructured Vocabularies Using a Background Ontology. In: Staab S, Svátek V, eds. Managing Knowledge in a World of Networks. EKAW 2006. Lecture Notes in Computer Science. Vol 4248. Berlin Heidelberg: Springer; 2006:182–197

40 Yu AC. Methods in biomedical ontology. J Biomed Inform 2006;39 (03):252–266

41 Zhang M, Hadjieleftheriou M, Ooi BC, Procopiuc CM, Srivastava D. Automatic discovery of attributes in relational databases. SIGMOD Conference; 2011

42 Otero-Cerdeira L, Rodríguez-Martínez FJ, Gómez-Rodríguez A. Ontology Matching: A Literature Review. Expert Syst Appl 2015;42(02):949–971

43 Euzenat J, Shvaiko P. Ontology Matching. 2nd ed. Berlin Heidelberg: Springer; 2013

44 Achichi M, Cheatham M, Dragisic Z, et al. Results of the Ontology Alignment Evaluation Initiative 2017. In: Proceedings of the 12th International Workshop on Ontology Matching co-located with the 16th International Semantic Web Conference (ISWC 2017); 2017:61–113

45 Pang C, Hendriksen D, Dijkstra M, et al. BiobankConnect: software to rapidly connect data elements for pooled analysis across biobanks using ontological and lexical indexing. J Am Med Inform Assoc 2015;22(01):65–75

46 Kock A-K, Bruland P, Kadioglu D. Mappathon - A Metadata Mapping Challenge for Secondary Use. GMDS 2018 [Internet]. 2018 August 27;1–2. Available at: https://www.egms.de/static/en/meetings/gmds2018/18gmds192.shtml. Accessed August 8, 2019

47 Pang C, Kelpin F, van Enckevort D, et al. BiobankUniverse: automatic matchmaking between datasets for biobank data discovery and integration. Bioinformatics 2017;33(22):3627–3634

48 Noy NF, Shah NH, Whetzel PL, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res 2009;37(Web Server issue):W170-3

49 Bodenreider O, Nelson SJ, Hole WT, Chang HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. Proc AMIA Symp 1998;•••:815–819

50 Jupp S, Liener T, Sarntivijai S, Vrousgou O, Burdett T, Parkinson HE OxO - A Gravy of Ontology Mapping Extracts. In: Proceedings of the 8th International Conference on Biomedical Ontology (ICBO 2017); 2017

51 U.S. National Library of Medicine. SNOMED CT to ICD-10-CM Map [Internet]. nlm.nih.gov. U.S. National Library of Medicine. Available at: https://www.nlm.nih.gov/research/umls/mapping_projects/snomedct_to_icd10cm.html. Accessed May 4, 2019

52 Mate S, Köpcke F, Toddenroth D, et al. Ontology-based data integration between clinical and research systems. PLoS One 2015;10(01):e0116656

53 Schüttler C, Buschhüter N, Döllinger C, et al. Requirements for a cross-location biobank IT infrastructure : Survey of stakeholder input on the establishment of a biobank network of the German Biobank Alliance (GBA) [in German]. Pathologe 2018;39(04): 289–296

54 McDonald CJ, Huff SM, Suico JG, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem 2003;49(04):624–633