Aus dem Institut für Medizinische Informatik
(Prof. Dr. rer. nat. D. Krefting)
der Medizinischen Fakultät der Universität Göttingen

Automated Metadata Transformation in a Medical Data Integration Center: Implementation of an Algorithm and Standardized Quality Analysis

INAUGURAL-DISSERTATION

zur Erlangung des Doktorgrades
der Humanwissenschaften in der Medizin
der Georg-August-Universität zu Göttingen

vorgelegt von

Caroline Marieken Bönisch, geb. Thoms

aus

Bergen auf Rügen

Göttingen 2023

Dekan:	Prof. Dr. med. W. Brück

Betreuungsausschuss

Betreuer*in:	Prof. Dr. med. T. Kesztyüs
Ko-Betreuer*in:	Prof. Dr. med. J. Brockmöller
Ko-Betreuer*in:	Prof. Dr. med. L. Saager

Prüfungskommission

Referent*in:	Prof. Dr. med. T. Kesztyüs
Ko-Referent*in:	Prof. Dr. med. J. Brockmöller
Dritt-Referent*innen:	Prof. Dr. med. L. Saager Prof. Dr. rer. nat. U. Sax Prof. Dr. L. Kolbe PD Dr. med. C. Wolff-Menzler

Tag der mündlichen Prüfung: 13.10.2023

Hiermit erkläre ich, die Dissertation mit dem Titel "Automated Metadata Transformation in a Medical Data Integration Center: Implementation of an Algorithm and Standardized Quality Analysis" eigenständig angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben. Weiterhin erkläre ich, dass anderweitig keine entsprechende Promotion beantragt wurde und hierbei die eingereichte Dissertation oder Teile daraus vorgelegt worden sind.

Göttingen, den 10.08.2023

(Unterschrift)

PUBLICATIONS

Die Daten, auf denen die vorliegende Arbeit basiert, wurden [teilweise] publiziert:

Parciak M, Suhr M, Schmidt C, **Bönisch C**, Löhnhardt B, Kesztyüs D, Kesztyüs T (2023): FAIRness through automation: development of an automated medical data integration infrastructure for FAIR health data in a maximum care university hospital. BMC Med Inform Decis Mak 23, 94

Impact Factor: 3.298

Bönisch C, Kesztyüs D, Kesztyüs T (2022): Harvesting metadata in clinical care: a crosswalk between FHIR, OMOP, CDISC and openEHR metadata. Sci Data 9, 659

Impact Factor: 8.501

Bönisch C, Kesztyüs D, Kesztyüs T (2023): FAIR+R: Making Clinical Data Reliable through Qualitative Metadata. Stud. Health Technol. Inform. 310, 110

Impact Factor: 0.285

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

API	Application Programming Interface
CDISC	Clinical Data Interchange Standards Consortium
CDISC ODM	Clinical Data Interchange Standards Consortium Operational Data Model
CDSTAR	Package-oriented data-management framework
CouchDB	CouchDatabase by Apache
DWH	Data Warehouse
EHR	Electronic Health Record
ETL	Extract-Transform-Load
FAIR	Findable, Accessible, Interoperable, Reusable
FDA	Food and Drug Administration
FHIR	Fast Healthcare Interoperability Resource
IDE	Integrated Development Environment
IT	Information Technology
JSON	JavaScript Object Notation
JSON-LD	JavaScript Object Notation for Linked Data
MariaDB	MariaDatabase
OMOP	Observational Medical Outcomes Partnership
openEHR	open Electronic Health Record
REST	Representational State Transfer
RWD	Real World Data
RQ	Research Question
SMART	Specific, Measurable, Achievable, Relevant and Time-bound
UMG	University Medical Center Göttingen
UMG-MeDIC	University Medical Center Göttingen Medical Dataintegration Center
UML	Unified Modeling Language
URI	Uniform Resource Identifier

CHAPTER 1: INTRODUCTION

Evidence-based medicine is the driver to improve quality of diagnosis and therapy. It relies on the connection of the experience of the clinicians and sound medical knowledge from research of multi-institutional collaborations (Bonomi and Jiang 2018). This research knowledge can only be as good as the data it is based on. As far as the usefulness of the clinical data collected is concerned, it depends to a large extent on its quality (Gliklich and Dreyer 2010; Weiskopf and Weng 2013). Metadata is the primary cue and oftentimes the first information one gets of the dedicated medical data. Furthermore, if the metadata are well defined and formalized, the process of finding the right data for research questions can be simplified. Therefore, data and their corresponding metadata have to be reliable in order to provide the qualitative fundament of clinical research.

1.1 Background

With the postulation of the FAIR Guiding Principles by Wilkinson et al. (2016), data owners worldwide received guidelines, which ensure that the data collected is findable, accessible, interoperable, and reusable. Metadata as a component of data is addressed throughout all the FAIR Principles. This underlines the high significance of metadata in the field of data processing, as the second principle states that "data are described with rich metadata" (Wilkinson et al. 2016). Furthermore, requirements for good metadata management can be derived from the FAIR principles. Metadata have to be "assigned a globally unique and persistent identifier", have to be "registered or indexed in a searchable resource", they have to "include qualified references to other (meta)data" and "use a formal, accessible, shared and broadly applicable language for knowledge representation" (Wilkinson et al. 2016).

Finally, metadata have to be still "accessible, even when the data are no longer available" (Wilkinson et al. 2016). Thus, metadata have to be omnipresent and stable throughout every part in the data process of medical data and beyond that.

The FAIR Principles are applied worldwide in the field of clinical research and healthcare (Meloni et al. 2021; Queralt-Rosinach 2022). No other field works with such highly sensitive data as clinical care and, consequently, clinical research (Dubovitskaja et al. 2017; Ecarot et al. 2021). Especially these data with their corresponding metadata have to meet high quality standards and need to be reusable in order to provide evidence for clinical research.

A unique aspect of the FAIR Principles includes that data have to be centralized in order to provide the basis of the FAIRification process. For this purpose, it is necessary to maintain a centralized data collection point within the clinical area (Sinaci et al. 2020).

Within the University Medical Center Göttingen, this task is performed by the Medical Data Integration Center (UMG-MeDIC). The UMG-MeDIC combines data from a maximum care hospital. Metadata and corresponding data from the clinical primary and secondary systems of the Medical Center are stored in a formalized manner within a data warehouse (DWH), using dedicated processes of extraction, transformation, and loading (ETL) (Parciak et al. 2023).

The UMG-MeDIC operates on data from controlled studies, as well as so-called real-world data (RWD). The definition of RWD was first postulated by the Food and Drug Administration (FDA) in 2016. RWD include all information that has been collected from routine clinical practice about healthcare services provided to the patient. They are thus reused beyond their actual purpose of collection.

1.2 Problem Statement

The Problem to be addressed in this thesis is the uncertain data quality of clinical data for the field of clinical research and the provision of the necessary information for researcher. Stausberg et al. (2022) highlighted that the FAIR Principles do not encompass the aspect of quality in (meta)data. Lack of data quality in the context of health care poses significant risks. Multiple identities, data gaps, or incorrectly assigned data are the more common errors among these.

Many consequences can result from poor data quality. Errors in patient treatment and poor data for subsequent research projects or studies are just some of the risks. “To achieve reliable and useful information from the large quantity of data and to make more effective and informed decisions, data should be clean and, as much as is possible, free of error. “(Ehsani-Moghaddam et al. 2021)

To address this described problem, more information about the clinical data is needed, which can only be provided by the respective metadata. Reliable data and metadata provide further knowledge about the quality of the collected data and can provide additional information to researchers. Therefore, this work emphasizes on the preparation and processing of metadata for the context of clinical research and evidence-based medicine.

An additional complicating factor is, that metadata, without preparation, can only be made available to researchers in one format, usually the format in which the metadata were collected or stored. This makes it difficult for researchers to easily access the information and get a first impression of the data, when the representation of the information is restricted to one format. This is where metadata play an important role and must be stored generically to the extent that it can be transferred to any data format using automated transformations and therefore provide the first and clear access to large storage of medical data.

1.3 Research Questions

In order to outline the thesis research objective, research questions (RQ) were developed. The research questions align with the research objective and were designed to be specific, focused, and answerable with the data collected during the research project. These RQ conform to the criteria of SMART, meaning that they are specific, measurable, achievable, relevant and time bound.

(RQ1) How can metadata be extracted from primary and secondary systems in clinical care and stored in a medical data warehouse?

(RQ2) How can metadata be prepared and made available for researchers to ensure the most insight into the corresponding data?

(RQ3) How can the reliability of metadata be demonstrated and what quality metrics must be met for this to be possible?

CHAPTER 2: METHODOLOGY

In this chapter the overall methodology as well as the methods of the cumulative parts of the thesis are delineated. The research approach and the data collection methods are thoroughly defined.

2.1 Research Type

Inferring from the objective of this thesis, this project is designed as top-down research, leading to an inductive approach. As this thesis relies on specific data and metadata and draws conclusions from the structure of the data and set up of the respective metadata, the research can be described as confirmative.

Building on this inductive approach, this research follows a mixed approach, combining quantitative and qualitative research. Data and Metadata of the UMG-MeDIC are exploited regarding the fit for specified data formats, used in healthcare and the relation of metadata and data quality in clinical care data. The open-ended research questions, listed under 1.3 strengthens the qualitative perspective of this research, while the numerical assignment and evaluation of the metadata refer to a quantitative approach.

2.2 Research Strategy

Before depicting the chosen research strategy, the criteria considered for this choice are listed below:

1. Following up from the research type, the research strategy should fit the objective of the research and be relevant for the research field.

2. The research strategy should be feasible within the research topic.

3. Reliability should be provided by the chosen research strategy, to make results of this thesis available for other researchers.

4. The appropriate method for the collection of information to conduct this research should be part of the research strategy.

Therefore, based on the key characteristics proposed by Benbasat et al. (1987), the proposed strategy of this research project can be represented as a case research. The research in this thesis relies on a single-case research design, in view of the fact that the described situation was previously inaccessible to scientific investigations and only possible with the establishment of a MeDIC at the University Medical Center Göttingen.

2.3 Data Collection

Data analysis in case research studies build on various methods of data collection. When conducting case research, one or more sources will be combined to generate generalisable findings.

As stated by Yin (1984), documentation, archival records, interviews, direct observation and physical artifacts are part of data collection methods in case research. Within this thesis the focus lies on documentation, direct observation, and physical artifacts in order to achieve a rich set of data surrounding the specific research issue. Based on this preceding strategy the aim is to answer the research questions in depth and supports the design of the research project within this thesis. As the thesis unfolds, revisions are anticipated according to unexpected observations, or limitations and opportunities. The subsequent use of the collected information and results of the research for other researchers is one of the highest requirements for data collection and provide inherent conditions for the selected data collection methods.

Additionally, literature searches accompany the data collection methods regarding the research objective.

2.4 Methodology used within Publications

The cumulative elements in form of peer-reviewed publications, as part of this thesis, refer to the methodology described above. The division into separate parts allowed for the differentiated use of various research methods (mixed-method approach) within the overarching case research.

2.4.1 FAIRness through automation: development of an automated medical data integration infrastructure for FAIR health data in a maximum care university hospital

The methodology of the first part of the research project within the thesis included a literature review, to provide an overview of findings from other research with a similar approach to establish an automated medical data integration infrastructure. The results of this review than formed the base for the development of the automated medical data integration infrastructure.

Starting from the literature review a requirement analysis was performed, including concepts and standards found via the literature review. As the requirements were defined, open-source software and informational workflows for the implementation of the infrastructure were chosen and tested for the setup of an automated medical data integration infrastructure. The final implementation of the found workflows and software tools concluded the methodology of the first phase of the research objective described in this thesis.

2.4.2 *Harvesting metadata in clinical care: a crosswalk between FHIR, OMOP, CDISC and openEHR metadata*

As for the metadata harvesting and establishing of a metadata crosswalk, a literature review was conducted to evaluate the existing metadata standards in the field.

The process of metadata harvesting involved gathering metadata from various data sources, documentation, or repositories and storing them in a centralized database schema. In this study, the harvested data was derived from the UMG-MeDIC. The UMG-MeDIC integrates medical information and its corresponding metadata from hospital information systems and clinical research databases, including data from studies, registries, case report forms, patient-reported outcomes, and research findings. The UMG-MeDIC employs a DWH, as denoted in 2.4.1, that combines datasets with different types of metadata and facilitates longitudinal data collection and integration. The medical data from the UMG are pseudonymized and transformed into the UMG-MeDIC's internal standardized data format.

The UMG-MeDIC's DWH aims to connect all available data sources from the UMG as part of an ongoing process. To make the medical data collected in the data integration center accessible to researchers, the metadata need to be made available in commonly used data formats within healthcare. Currently, the UMG-MeDIC supports data formats such as OMOP, openEHR, FHIR, and CDISC. This enables researchers to choose their desired target data format. To facilitate this selection, a metadata crosswalk was developed.

A metadata crosswalk is a chart or map that depicts elements and identifies metadata items that are specifically important to the UMG-MeDIC. These items are determined based on the aforementioned literature research and data format specifications. The identified items that meet the UMG-MeDIC's requirements are categorized into three groups. Category 1 includes items of utmost importance for the operation of the data integration center and data provision. Category 2 consists of items of moderate importance that are essential for data privacy and consent. Category 3 includes metadata with the lowest priority, which provide additional context information and language specifications. The urgency of the items is not considered; only their importance for the UMG-MeDIC is considered in the priority dimension (Table 1).

After prioritization, scores are calculated for each data format to determine which formats cover items from priority categories 1 and 2 most comprehensively, to decide which data format would be best suitable for the UMG-MeDIC.

Table 1. Essential metadata required in the MeDIC and respective priority level

Metadata Item	Priority	Description
MetadataID	1	Unique and persistent identifier of the metadata
MetadataDate	1	Date of the metadata creation
AffiliateDatasetID	1	Globally unique identifier of the data, which the metadata is associated with
MetadataVersion	1	Version of the Metadata
ReferenceData	1	References to other data via name or description
ReferenceMetadata	1	References to other metadata via name or descritption
DataLifecycleState	2	State of the data during its lifecycle (creation, processing, analysis, preservation, access, reuse)
UsageLicense/ Copyright	1	clear and accessible data usage license
UsageContext	3	Context in which the data should be used
SourceSystemName	1	Explicit name of the Source System
SourceSystemVersion	1	Version of the Source System when recording the (meta)data
SourceInformation	3	Additional information about the source of the

		data
SourceOriginal Contributor	2	Contributor of the Source data
VestingPeriod	3	Availability of data to other researcher outside the study during the time of the study
ConsentType	2	Type of patient consent (i.e. broad consent, study specific consent)
ConsentValidation	1	Validity period of the consent
ConsentVerification	1	Physical signature of the patient and start of the validity period
ConsentModule	2	Exact parts of the consent, to which the patient consented to
DataItemLanguage	3	Language of the data items

***Note*.**Priority level 1 = high, 2 = medium, 3 = low

2.4.3 *FAIR+R: Making clinical data reliable through qualitative metadata*

The initial step involved conducting a literature search to examine existing evaluation schemes and methods pertaining to the quality of metadata and data. The findings of this review indicated the existence of various approaches for assessing metadata or data quality. The authors further focused and investigated quality factors for data and incorporated them, whenever applicable, into the metadata domain. This was done to create a comprehensive set of quality factors for evaluating metadata.

Based on the literature review, identified quality factors for data quality were assessed and audited to see if they can be embedded in the assessment of metadata quality. In result metadata quality factors could then be obtained. These quality factors were then integrated into the ETL processes of the UMG-MeDIC, providing a starting point to check the quality and respective reliability of the metadata.

2.5 Methodology of the UMG-MeDIC's Metadata Management

Based on the previous methodology and findings presented in 2.4.1 - 2.4.3 a literature review to explore previous research related to metadata management across various scientific fields, formed the basis of the final part of the thesis research. Developing new approaches and solutions that focus on transforming metadata elements between different data structures was the main objective of this research part.

At the UMG-MeDIC , metadata are obtained as part of the Extract-Transform-Load (ETL) processes within the data integration infrastructure. Previously, the data format and schema utilized is JSON for Linking Data (JSON-LD), which enables storing metadata in machine-readable JSON documents and organizing them as datasets, in compliance with the requirements of schema.org. The stored information include the metadata's identifier (with the full Uniform Resource Identifier - URI), source system name, workflow name, sources, start and end modification dates, version, extracted metadata (brief information about the content), reference to the data table within the source system, and error messages if applicable.

In the context of this final part, a decision was made to store the metadata in a relational database, specifically a MariaDB. This choice was necessary because the actual data that the metadata refers to were already stored in the same relational database. As part of the UMG-MeDIC restructuring process, it was deemed fitting to consolidate the metadata and medical data in a shared database, rather than storing them apart. Consequently, the metadata, which was previously stored in JSON-LD documents, needed to be transformed into a relational data structure.

To facilitate the desired transformation of the metadata, all relevant information from the provided data format specifications were extracted. These specifications served as the starting point for mapping the metadata elements between the different data structures.

Additionally, the granularity of the metadata extracted from the primary source systems was taken into consideration. Metadata in primary source systems can vary in terms of granularity, such as having metadata dates represented as day-month-year hour-minute-second or month-year. These differences needed to be acknowledged and handled appropriately during the processing and inclusion of the metadata.

To carry out this transfer, all metadata information must be migrated from the previous storage location in CouchDB to the newly created relational database tables.
Once the mapping process was finished, an algorithm was established to automatically transfer the metadata from the previous JSON-LD format to the relational data tables. PyCharm, an integrated development environment (IDE), was used for programming the transformation with Python (see Appendix for the complete code).

It was opted for an approach that utilized a generic metadata structure which is depicted in Figure 1, ensuring a comprehensive method of storage.

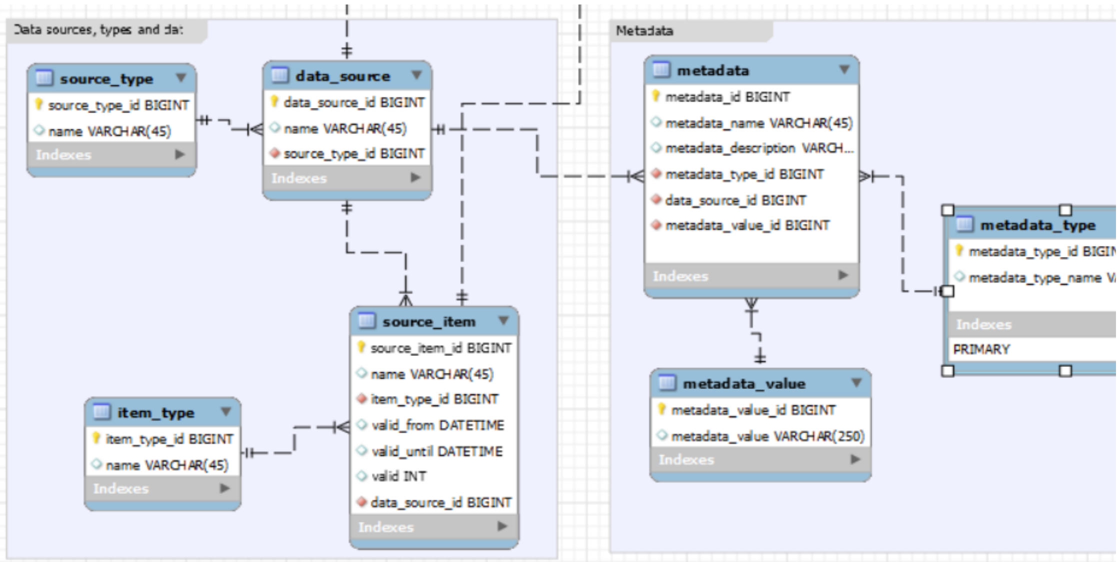


Figure 1: Relational metadata table structure within the UMG-MeDIC. Linkage to the data source tables are shown on the left. The relational data model was created via MySQL Workbench.

To assess the reliability of UMG-MeDIC data, we introduced data quality as an additional metadata factor. Data quality was assigned a nominal value on a quality scale. This scale includes categories such as high-quality data (e.g., measured values from procedures and study-collected data), average-quality data (evaluations, billing data), and low-quality data (e.g., patient-initiated data, free-text entries).

The criteria used for this assignment were established based on preliminary work (see 2.4.3), focusing on identifying quality metrics for metadata. These criteria considered factors such as data completeness, consistency, correctness, correspondence to other data, relevance, semantic specificity, as well as timeliness, accessibility, and reproducibility of the data.

Data from various areas within UMG were assigned to corresponding entries on the data quality scale. Additionally, the authors manually verified the correct assignment of the data within the scope of this work.

CHAPTER 3: RESULTS

The chapter describes the results of the cumulative findings within the contributing publications and deducts the answers to the research questions defined at the beginning.

3.1 Findings of the publications

Each of the publications contributed to answering the research questions. The findings of each of these parts of research are presented and considered.

3.1.1 FAIRness through automation: development of an automated medical data integration infrastructure for FAIR health data in a maximum care university hospital

The UMG-MeDIC infrastructure operates based on the microservice paradigm, where each application functions independently. The process flows within these microservices are orchestrated by ActiveWorkflow, which is responsible for managing the workflows defined by UMG-MeDIC data engineers. ActiveWorkflow communicates with other services using agents, which implement the ActiveWorkflow REST API and operate autonomously. JSON documents are used for communication between ActiveWorkflow and the agents, carrying text-based payloads.

The goal is to automate the workflows for managing clinical care data and research data, including data transformation and data provenance. JSON-LD metadata templates are used to capture relevant information and create linked data compatible with provenance information models. Containerization is utilized for platform-independent execution and reproduction of data integration workflows. Data lake web services are employed to store copies of incoming source and intermediate data artifacts. All components communicate through RESTful web service APIs, ensuring compliance with IT-security policies. Custom ActiveWorkflow agents, such as the Docker Agent and Annotation Agent, have been implemented to support data integration workflows. The Docker Agent executes Docker images, often involving Python-based ETL processes, while the Annotation Agent collects process metadata and stores it in a CouchDB process metadata store.

The metadata documents include information about datasets, ETL processes, and their interconnections. The workflow specification can repeat the sequence of Docker Agent and Annotation Agent tasks as necessary. Metadata templates are used to standardize the capturing of relevant information, and the CDSTAR data lake provides the basis for defining archives and files. Processes are represented as CreateAction documents, linking input and output data, referencing implemented processing steps, and source code repositories. Manual documentation is created as workflow descriptions, enriched with free-text descriptions exported from ActiveWorkflow. A custom monitoring service and web-based user interface display process metadata, allowing visualization and validation of workflow runs. Statistical parameters of processed data are captured in process metadata, and validation functions are implemented to ensure data correctness and consistency. Validation results are stored in CouchDB as part of the process metadata.

3.1.2 Harvesting metadata in clinical care: a crosswalk between FHIR, OMOP, CDISC and openEHR metadata

This second part of the research focused on examining the data format specifications of CDISC, OMOP, openEHR, and FHIR. The corresponding metadata items for each data format were extracted from the documentation of each of the data formats and compared.

Once the metadata crosswalk was established, the next phase involved identifying metadata items that are highly relevant to the UMG-MeDIC.

Based on the findings from the literature research, including the FAIR Principles and UMG-MeDIC requirements, essential metadata items for the UMG-MeDIC were selected. Among the FAIR Principles, F1, F3, F4, A2, I3, and R1.1 were considered as they directly pertain to metadata. F1 emphasizes the need for globally unique and persistent identifiers for (meta)data. F3 involves metadata that clearly and explicitly describes the associated data. F4 highlights the importance of indexing metadata in a searchable resource. A2 requires metadata to be accessible even when the data itself is not available. I3 states that metadata should include qualified references to other metadata. Finally, R1.1 suggests that metadata should be accompanied by clear and accessible data usage licenses.

The requirements for the UMG-MeDIC were derived from a requirements analysis conducted during the initial development of the UMG-MeDIC in 2019. As evident from the scoring of the prioritization (as depicted in Figure 3), none of the data formats meet all the necessary UMG-MeDIC-specific criteria.

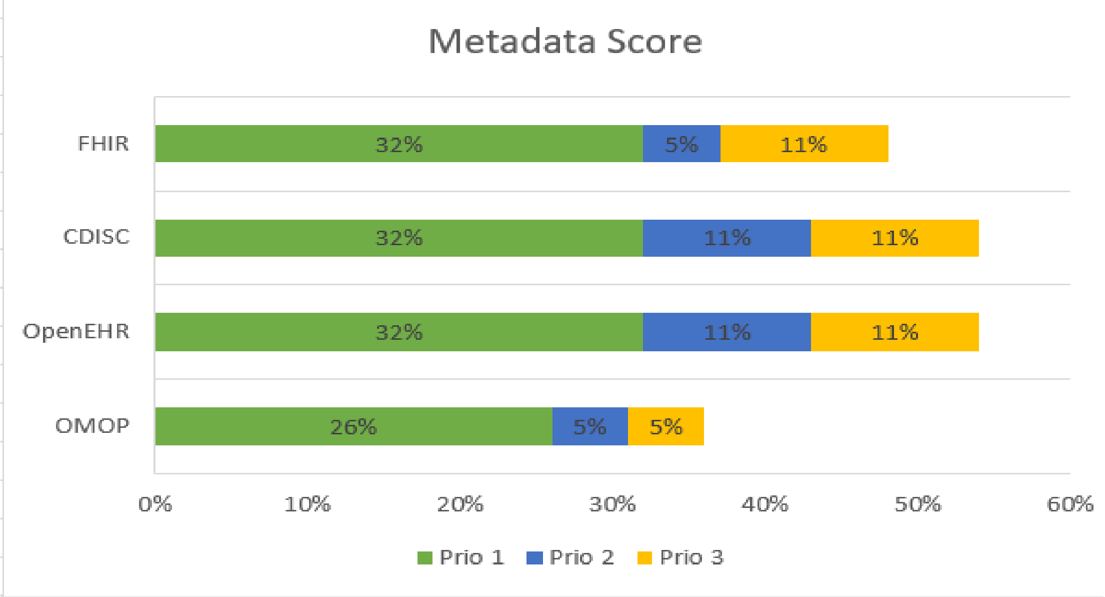


Figure 2: Qualitative priority scoring of metadata required by the UMG-MeDIC and quantitative coverage in the different data formats FHIR, CDISC, OpenEHR and OMOP.

Due to the different underlying premises of the individual formats, a complete transformation was also not feasible. For instance, CDISC extends the ODM format for study documentation in clinical care, while OpenEHR is designed for storing medical data in an EHR, and FHIR facilitates data exchange between institutions. OMOP, on the other hand, offers a common data format for unifying data from various databases. It was apparent that none of these data formats encompass all the required metadata for successful operation of the UMG-MeDIC in terms of reliable data management. Therefore, the proposed solution entailed a specific convergence format to overcome these differences.

The convergence format demonstrates how metadata items from different formats can be incorporated, preventing the loss of information by providing metadata items in the target format, even if they are not present in the source data format. The convergence format offers the best solution for maintaining the structure of the format by generating the required items during the transformation process and populating them with NULL values if the source format does not provide any input values.

3.1.3 FAIR+R: Making clinical data reliable through qualitative metadata

Based on the findings of the literature search, quality measures for metadata were derived. The quality measures are based on a compilation of research papers across diverse scientific domains and have been carefully chosen to meet the specific demands of the data complexity in clincal care.

Table 2 depicts the quality measures for metadata within the UMG-MeDIC.

Table 2. Assessment metrics for metadata in clinical care, based on the results of the literature review.

Measure	Description
Completeness	All mandatory data fields are filled with information
Consistency	Metadata should be conformed to existing standards and formats
Correctness	The information describes the metadata in an accurate and distinct way
Correspondence	Metadata that is linked or inter-dependent represents the same information through every instance
Relevance	The metadata corresponds to the requirement/expectations of the user
Semantic Specificity	Average specificity of a semantic concept in metadata information
Timeliness	Currency of the metadata information describing a resource information
Accessibility	The information of the metadata must be physically available and understandable either by human or machine
Reproducibility	Metadata quality scores should be reproducible and not lack clarity in terminology

3.2 Resulting UMG-MeDIC's Metadata Management

The results yielded from the final part of the research were elaborated as follows:

The metadata from the primary source systems of the UMG stored in CouchDB were extracted and moved to the new relational metadata structure in the DWH. This new structure (see Figure 1) aligns with the existing relational data tables in the DWH. The algorithm successfully transferred the existing metadata and included additional metadata types like data quality and data item language. The ETL processes were adjusted accordingly.

To ensure the reliability and integrity of the metadata, a transformation algorithm was developed and tested on a small amount of data in a test system. The algorithm performed as expected, extracting metadata from JSON-LD, matching it with relational data tables, and storing it correctly in MariaDB. The algorithm is reliable, producing consistent results with every run. The duration of the transformation and saving of the results in the relational database took 75,95 seconds for 3119 metadata documents on a 64-Bit Windows 10 Enterprise-22H2 with a 2.40GHz processor and 16GB RAM.

The use of hash values enabled effective audit trailing of the metadata in the DWH. Each transaction of the data, stored within the DWH, can be traced by checking the corresponding hash values of the metadata entries. This provides a secure process to detect unintentional manipulations, ensuring transparency for future use.

Based on this preliminary work, researchers can now filter the UMG-MeDIC data pool using a GUI and export the corresponding metadata for the filtered dataset as shown in Figure 4, including the information about the quality of the underlying data.

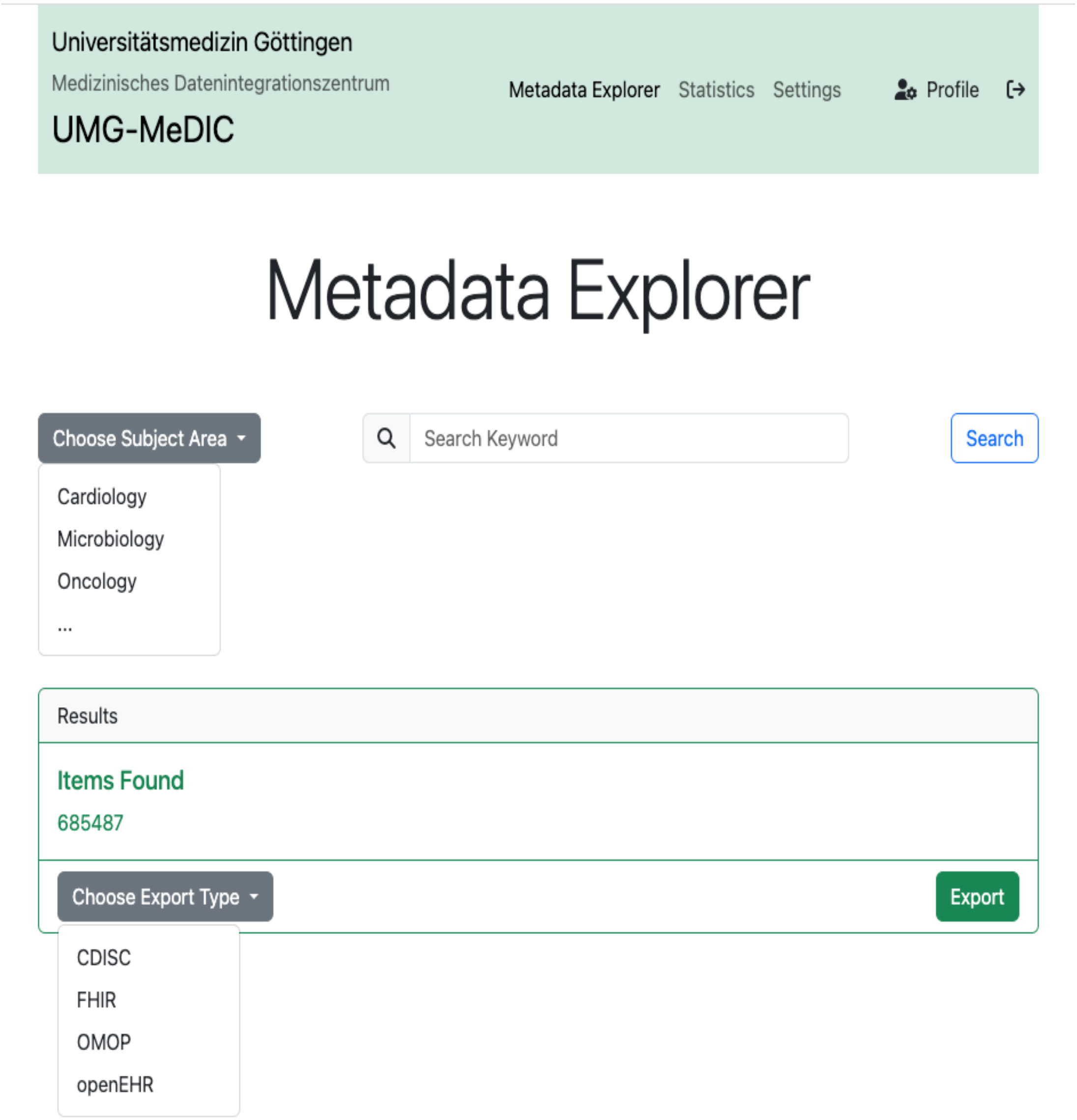


Figure 3: User Interface of the Metadata Export application with the selectable Subject area and search bar. Below are the results list and the export type drop-down menu to choose the data format. After the selected items are chosen, the export can be started via the green Export button.

As part of internal testing, the derivation of metadata in different selectable data formats for researchers was successfully conducted. Unit tests involving volunteers from the UMG confirmed that the metadata could be converted into the desired data format. The resulting data format file was made available for download directly onto the computer.

CHAPTER 4: SYNOPSIS

4.1 Results in relation to the Research Questions

To relate the results of each respective publication of the thesis to the research questions and demonstrate how the contributions of the work address these questions, the next step involves drawing the conclusions from the results to the research questions.

(RQ1) How can metadata be extracted from primary and secondary systems in clinical care and stored in a medical data warehouse?

In FAIRness *through automation: development of an automated medical data integration infrastructure for FAIR health data in a maximum care university hospital* it was shown that data and the associated metadata could be extracted from the primary sources within the UMG using customized ETL routes. This extraction could be implemented with open-source software and was based on the requirements of such a data infrastructure and the specifications by the site including UMG. The metadata was stored within the first design choice in a Couch DB as a JSON LD object. After a restructuring of the UMG-MeDIC, the already existing data was transferred from the Couch DB in a relational data structure. In doing so, the specifications from the requirements analysis of the UMG-MeDIC were again considered and changes were accounted for.

As a result, the metadata and the data could and still can be extracted in an automated manner from the UMG's clinical care systems and stored with consideration of secondary use.

(RQ2) How can metadata be prepared and made available for researchers to ensure the most insight into the corresponding data?

Firstly, metadata elements were identified within *Harvesting metadata in clinical care: a crosswalk between FHIR, OMOP, CDISC and openEHR metadata,* which are frequently used in clinical research. Based on this evaluation, information that are needed for the processing of the UMG-MeDIC such as data source, collection method, variables, units of measurement, data quality indicators, timestamps, and any other pertinent details were determined. The transformation and mapping applied to the original metadata items form the data formats considered was performed using a metadata crosswalk.

Subsequently a metadata schema, a so-called convergence format, was put in place, to ensure consistency and interoperability. Data formats like openEHR, OMOP, CDISC and FHIR provided metadata items to be considered. Additionally, a graphical user interface was developed, where researchers can input relevant information to gather insight into the underlying clinical data. This graphical interface aligns with the chosen metadata convergence format.

In order to promote discoverability, a unique identifier (a metadata id) was assigned to every metadata entry, providing detailed documentation that explains the metadata elements and their meanings.

As the metadataset evolves or new insights emerge, the need for updates of the metadata become evident. Therefore, a versioning system was established in manuscript three to keep track of metadata revisions.

(RQ3) How can the reliability of metadata be demonstrated and what quality metrics must be met for this to be possible?
The reliability of the metadata had to be investigated via a literature review. For this purpose, measures were derived from the field of data quality and applied to metadata (*FAIR+R: Making Clinical Data Reliable through Qualitative Metadata*). As shown in 3.1.3, 9 metrics were identified and used to calculate the quality of the metadata. Based on this, the metadata and associated data are classified into one of three categories. Meanwhile, the metadata quality is included within the ETL processes and associated with each record, providing an insight of the quality of the metadata for researchers.

4.2 Limitations

The limitations of the work are described in the following sections. The limiting factors are the research design, the data used and the convergence format.

4.2.1 Limitations from the research design

While a pure usage of quantitative or qualitative methods salvages the risk of limitations of the research (quantitative limited by the research question and method, qualitative being more subjective and prone to researcher bias) this thesis focused on a mixed-methods approach. The combination of the strengths from qualitative and quantitative methods are deployed to overcome these limitations.

On the downside mixed methods are more resourceful and can be complicated to archive.

Conducting a single-site case research is also limited to the organizational structure of the site. Furthermore, deriving general findings can be a limiting factor of single site case research because the results can be very specific to the respective site, when not focused on appropriate factors.

Conducting case research with a single site may provide restrictions and bias regarding the site. Although UMG is a maximum care hospital in Lower Saxony, it is subject to the constraints of the German hospital landscape and German legislation. The transferability of the results to other countries with different legislation and use of software can therefore not be guaranteed. The thesis uses generic approaches and open-source software, but the national situation must always be taken into account.

Furthermore, the replication of case research may be limited, here the collaboration with other researchers facing similar objectives can help strengthen the findings and results of this research thesis objective.

4.2.2 Limitations of the data used

For the development of the convergence format, a restriction was made to the four most frequently used data formats in clinical care. This restriction limits the maximum set of metadata, as the focus here is on the metadata items of the selected data formats.

Additionally, the results of the thesis are based on metadata and data from the UMG and are therefore subject to format specifications and possible bias in the number of different data elements (e.g., more metadata in relation to data origin, compared to other hospitals).

4.2.3 Limitations regarding the convergence format

The developed convergence format must be adapted repeatedly when changes in the chosen data formats occur. Since a static mapping was performed, changes in the original data formats can only be addressed with a manual change in the convergence format. These changes in the data formats must be monitored continuously.

4.3 Conclusion

Researchers seeking data provision from UMG-MeDIC can now request an initial overview of the data volume for their specific research question by querying the associated metadata in advance. To ensure simplicity and user-friendliness, researchers have the option to extract the metadata in different data formats such as OMOP, openEHR, and more. The comparison and contrast of the metadata items of the different data formats in healthcare is the only metadata crosswalk in the form of a mapping table of this kind so far. There is currently very little research in medical metadata and thus little information about possible advantages and disadvantages for the reusability of data depending on their metadata.

This thesis fills this gap and highlights the potential of highly curated metadata for clinical research. With the inclusion of real-world data to complement randomized clinical trial data for clinically evidenced research results, metadata quality measures come into play. Metadata, which are made available via a metadata explorer contain initial information that is relevant for an assessment of the quality of the medical data collected and form the basis of the secondary use of clinical data for research, before filing a (mostly needed) data usage agreement.

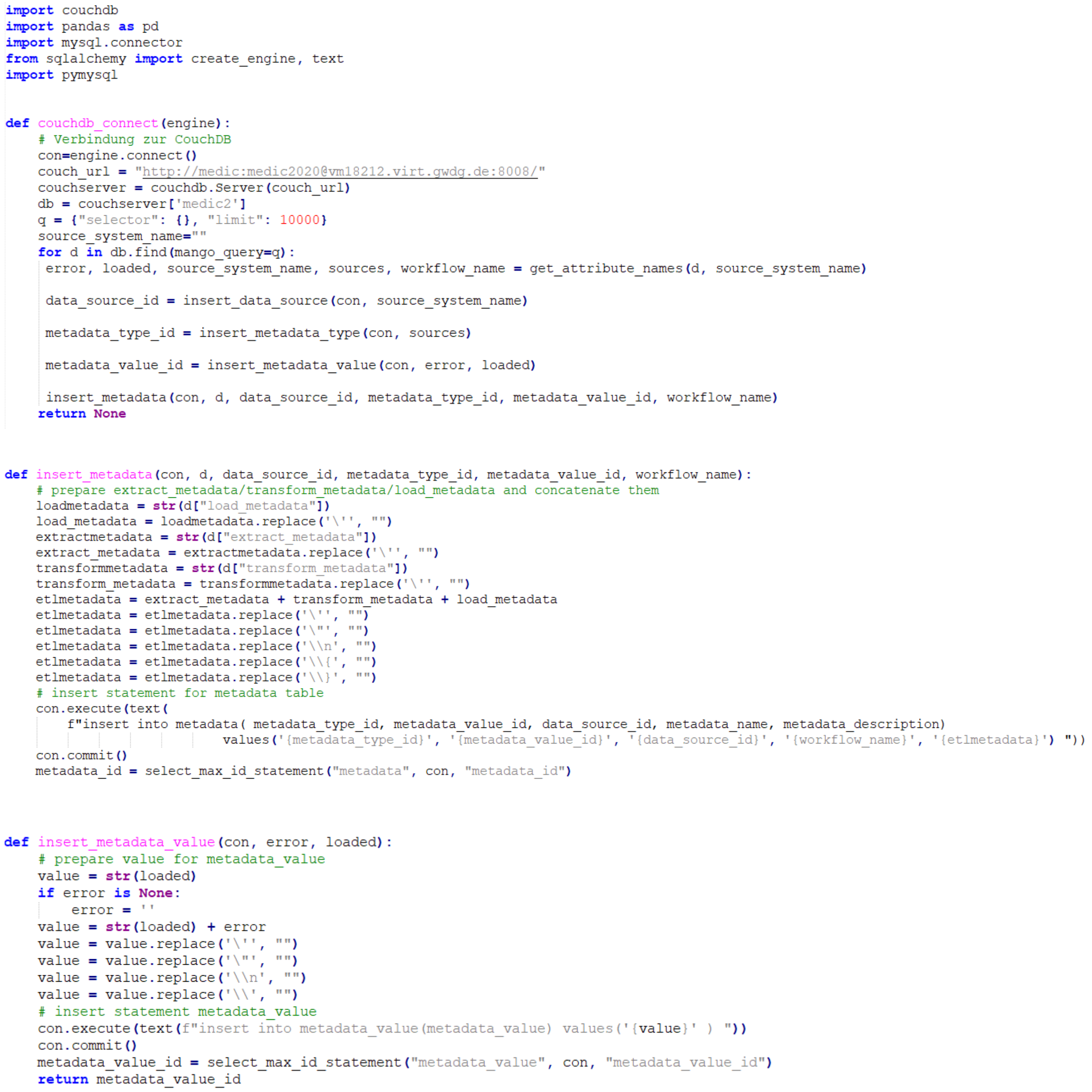
4.4 Recommendations for Future Research

A next step for future research should include further clear and comprehensive criteria for each of the nine quality metrics to enable a more fine-grained classification than currently used. In order to achieve this, however, the fundamental metrics must first be evaluated under real conditions. Additionally, the assignment of the quality is hard coded at the moment, a next step would be to use machine learning algorithms to assign the data quality on the fly, while the decision is based on trained data.

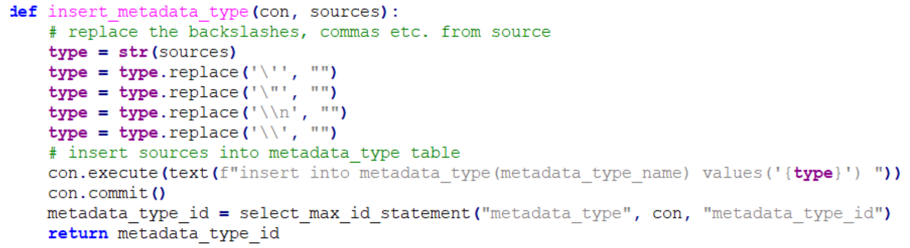
Furthermore, the established graphical interface and underlying provision of metadata should be accompanied with an in-depth documentation to guide researchers in evaluating the quality of the underlying data. It should address the specific requirements and challenges of the metadata set and domain.
Therefore, the assessing of the effectiveness of the fine-grained quality classification is part of further investigation.

Furthermore, collaborative research, like the collaboration with other researchers and practitioners in the field must be fostered. Collaborative efforts can help refine and validate the proposed approach and drive further advancements in the field.

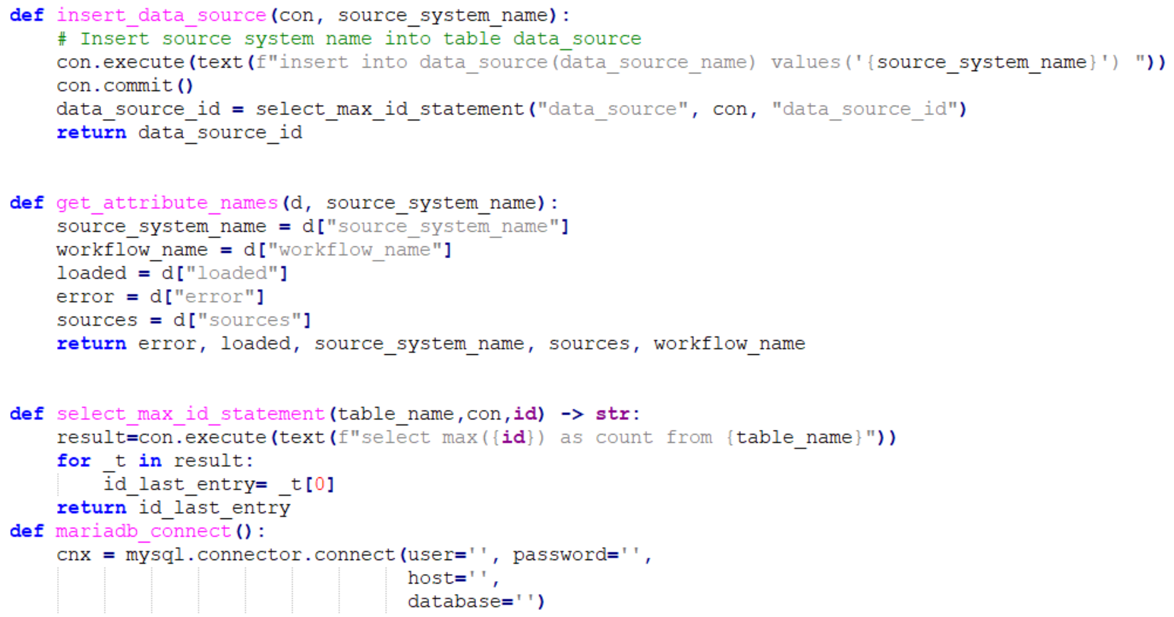
Appendix A – Program Code



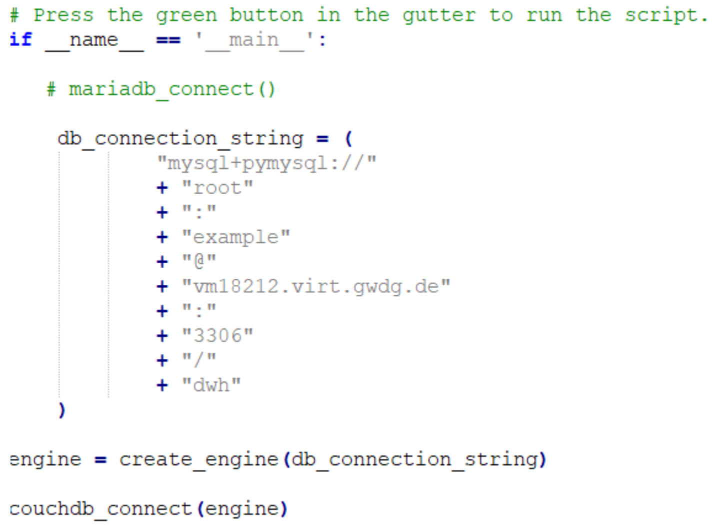
```
import couchdb
import pandas as pd
import mysql.connector
from sqlalchemy import create_engine, text
import pymysql


def couchdb_connect(engine):
    # Verbindung zur CouchDB
    con=engine.connect()
    couch_url = "http://medic:medic2020@vm18212.virt.gwdg.de:8008/"
    couchserver = couchdb.Server(couch_url)
    db = couchserver['medic2']
    q = {"selector": {}, "limit": 10000}
    source_system_name=""
    for d in db.find(mango_query=q):
     error, loaded, source_system_name, sources, workflow_name = get_attribute_names(d, source_system_name)

     data_source_id = insert_data_source(con, source_system_name)

     metadata_type_id = insert_metadata_type(con, sources)

     metadata_value_id = insert_metadata_value(con, error, loaded)

     insert_metadata(con, d, data_source_id, metadata_type_id, metadata_value_id, workflow_name)
    return None


def insert_metadata(con, d, data_source_id, metadata_type_id, metadata_value_id, workflow_name):
    # prepare extract_metadata/transform_metadata/load_metadata and concatenate them
    loadmetadata = str(d["load_metadata"])
    load_metadata = loadmetadata.replace('\'', "")
    extractmetadata = str(d["extract_metadata"])
    extract_metadata = extractmetadata.replace('\'', "")
    transformmetadata = str(d["transform_metadata"])
    transform_metadata = transformmetadata.replace('\'', "")
    etlmetadata = extract_metadata + transform_metadata + load_metadata
    etlmetadata = etlmetadata.replace('\'', "")
    etlmetadata = etlmetadata.replace('\"', "")
    etlmetadata = etlmetadata.replace('\\n', "")
    etlmetadata = etlmetadata.replace('\\{', "")
    etlmetadata = etlmetadata.replace('\\}', "")
    # insert statement for metadata table
    con.execute(text(
        f"insert into metadata( metadata_type_id, metadata_value_id, data_source_id, metadata_name, metadata_description)
                           values('{metadata_type_id}', '{metadata_value_id}', '{data_source_id}', '{workflow_name}', '{etlmetadata}') "))
    con.commit()
    metadata_id = select_max_id_statement("metadata", con, "metadata_id")



def insert_metadata_value(con, error, loaded):
    # prepare value for metadata_value
    value = str(loaded)
    if error is None:
        error = ''
    value = str(loaded) + error
    value = value.replace('\'', "")
    value = value.replace('\"', "")
    value = value.replace('\\n', "")
    value = value.replace('\\', "")
    # insert statement metadata_value
    con.execute(text(f"insert into metadata_value(metadata_value) values('{value}' ) "))
    con.commit()
    metadata_value_id = select_max_id_statement("metadata_value", con, "metadata_value_id")
    return metadata_value_id
```



```
def insert_metadata_type(con, sources):
    # replace the backslashes, commas etc. from source
    type = str(sources)
    type = type.replace('\'', "")
    type = type.replace('\"', "")
    type = type.replace('\\n', "")
    type = type.replace('\\', "")
    # insert sources into metadata_type table
    con.execute(text(f"insert into metadata_type(metadata_type_name) values('{type}') "))
    con.commit()
    metadata_type_id = select_max_id_statement("metadata_type", con, "metadata_type_id")
    return metadata_type_id
```



```
def insert_data_source(con, source_system_name):
     # Insert source system name into table data_source
     con.execute(text(f"insert into data_source(data_source_name) values('{source_system_name}') "))
     con.commit()
     data_source_id = select_max_id_statement("data_source", con, "data_source_id")
     return data_source_id


def get_attribute_names(d, source_system_name):
    source_system_name = d["source_system_name"]
    workflow_name = d["workflow_name"]
    loaded = d["loaded"]
    error = d["error"]
    sources = d["sources"]
    return error, loaded, source_system_name, sources, workflow_name


def select_max_id_statement(table_name,con,id) -> str:
    result=con.execute(text(f"select max({id}) as count from {table_name}"))
    for _t in result:
        id_last_entry= _t[0]
    return id_last_entry
def mariadb_connect():
    cnx = mysql.connector.connect(user='', password='',
                                  host='',
                                  database='')
```



```
# Press the green button in the gutter to run the script.
if __name__ == '__main__':

   # mariadb_connect()

    db_connection_string = (
            "mysql+pymysql://"
            + "root"
            + ":"
            + "example"
            + "@"
            + "vm18212.virt.gwdg.de"
            + ":"
            + "3306"
            + "/"
            + "dwh"
    )

engine = create_engine(db_connection_string)

couchdb_connect(engine)
```

Appendix B – Publications

The publications are embedded in chronological order, starting with the year 2022. The full author version of every publication is attached within the appendix.

scientific data

OPEN

ARTICLE

Harvesting metadata in clinical care: a crosswalk between FHIR, OMOP, CDISC and openEHR metadata

Caroline Bönisch ✉, Dorothea Kesztyüs & Tibor Kesztyüs

Metadata describe information about data source, type of creation, structure, status and semantics and are prerequisite for preservation and reuse of medical data. To overcome the hurdle of disparate data sources and repositories with heterogeneous data formats a metadata crosswalk was initiated, based on existing standards. FAIR Principles were included, as well as data format specifications. The metadata crosswalk is the foundation of data provision between a Medical Data Integration Center (MeDIC) and researchers, providing a selection of metadata information for research design and requests. Based on the crosswalk, metadata items were prioritized and categorized to demonstrate that not one single predefined standard meets all requirements of a MeDIC and only a maximum data set of metadata is suitable for use. The development of a convergence format including the maximum data set is the anticipated solution for an automated transformation of metadata in a MeDIC.

Introduction

Since humans began sorting and categorizing information and objects, metadata provided important alignment of information objects. Metadata are defined as data about data[1]. They describe information objects with regard to source, type of creation, structure, status, level and semantics. An information object can be either a data including a coded value or instance identifier, or a list of several dates, or an entire database with various dependencies[1]. Using metadata, related data can be reused, organized, described, validated, searched and queried. In the medical field, the provision, reuse and preservation of information is essential to ensure the best possible treatment of a patient as well as answering research questions. As Hegselmann *et al.* stated “Individuals with very specific characteristics could be identified, which is mandatory for personalized medicine as well as epidemiological and clinical studies, but also general big data applications would be possible”[2]. Retrospective acquired data, especially if largely available, provides opportunities not only to predict but detect e.g. novel risks and therapeutic options on an individual level (precision medicine)[3]. Consequently, the subordinate metadata are predominant to prepare the basis for combining and transforming the data by providing metainformation to enable linkage of information from different data sources. This goes to show in which way metadata benefits the medical sector.

The FAIR Principles, postulated in 2016[4], suggest that the reuse of (meta)data is of great importance in the context of medical research. The management of data with the corresponding application of metadata provides multiple opportunities for high-quality data analyses and subsequent high impact publications. Metadata for research data is also gaining importance in light of the increasing requirement of journals to make primary data from published research publicly available[5]. The FAIR Principles are divided into the categories Findable, Accessible, Interoperable and Re-Usable. Metadata are explicitly named in all four categories. Therefore metadata act as important building blocks for making information accessible and usable.

However, Dugas *et al.* acknowledge that most forms and item catalogs from healthcare research studies in Germany do not comply to these FAIR Principles and cannot be easily found and are therefore not re-usable[6]. This is due to the fact that forms are sometimes not allowed to be published, because of permission restrictions or they are not published based on interoperability points of view, e.g. without an identification number or

Medical Data Integration Center, Department of Medical Informatics, University Medical Center Göttingen, Robert-Koch-Str. 40, 37075, Göttingen, Germany. ✉e-mail: caroline.boenisch@med.uni-goettingen.de

accompanying metadata, and remain in a paper tomb. As stated in the article, it is important to publish metadata with the data, as this characterizes a first step towards open data[6].

Mainly in the biomedical field, it is noticeable that there were and still are implementations where the importance of metadata is highlighted[7]. Different consortia and working groups provide their approaches to utilize metadata with regards to re-usability, accessibility and findability[8,9]. Both aforementioned articles adopt the paradigm that qualitative metadata is helpful to retrieve, acquire and utilize metadata. The working group referenced in[8] proposes the potential of ontology concepts to annotate metadata making them easier to be found and semantic specific, resulting in a strong descriptor of the resource contents.

Gonçalves *et al.*[9] developed a software that pulls information from metadata records and analyses the information whether it is complete and correct according to given specifications (right format and legitimate content).

Data integration from heterogeneous source data systems is a major challenge, not only in the biomedical field. Initially, each system has very different metadata attributes that must be taken into account. The structures of the data used are specifically designed for the respective source data system. This can make it difficult to reuse data, which was collected within specific source systems, due to proprietary reasons.

Canham and Ohmann describe that metadata can be divided into two parts. On the one hand, there is intrinsic metadata, which is permanent and unchangeable[10]. For instance, metadata such as the date/timestamp and the performing clinician, as well as the status (active, postponed, complete) of a clinical examination, is considered intrinsic metadata.

On the other hand, they identify provenance metadata, which represents localization or history, like the data lifecycle state (creation, processing, analysis, preservation, access, reuse), the data custodian or the method of data collection. Provenance metadata is subject to change, because of its nature to provide information about non-static knowledge. Both intrinsic and provenance metadata are required for searching and uniquely identifying data. The variability of data in routine clinical practice makes the use of a unified metadata schema complicated, but nonetheless Canham and Ohmann proposed a common metadata scheme within the "protocol-driven clinical research", that would be applicable to any information system. As they note, it is more beneficial if the data and corresponding metadata remain in their original relational form and are converted to the desired target format FHIR, openEHR or OMOP using a parser[10]. An appropriate crosswalk between the individual metadata elements of the respective standards is of utmost importance to obtain the most fine-grained result possible with a maximum set of metadata elements.

Metadata harvesting describes a process to combine metadata from different data storages, archives or repositories and store them in a central database schema. The data that is harvested in this work is derived from the Medical Data Integration Center (MeDIC) of the University Medical Center Göttingen (UMG). The University Medical Center is a hospital of maximum care and extensive sources of medical data. The MeDIC joins medical information and their corresponding metadata from hospital information systems and clinical research data bases (which include, inter alia, data from studies and registries, such as case report forms, patient reported outcomes or findings) in a data warehouse. It involves data from datasets with different (meta)data types and longitudinal data collection, as well as data integration. The data and corresponding metadata are stored in a relational database, which underlies the data warehouse of the MeDIC. The metadata is kept in a distinct table separated from the data, connected via a primary/foreign key to the tables of data. Therefore, it is possible to store metadata in a n-dimensional repository in the same format The medical source data from the hospital and department information systems are pseudonymized and transformed into the internal harmonized data format of the MeDIC. During the process of harvesting metadata within the Extract-Transform-Load process, metadata is extracted and loaded via a MeDIC-specific data protocol, preventing duplicates. The data warehouse of the MeDIC anticipates to connect all available data sources of the UMG as part of an ongoing process.

In this article, we aim to provide a crosswalk between the formats mentioned above and try to convey them as accurate as possible.

Results

For the purpose of this research project, the specifications for the data formats CDISC, OMOP, openEHR and FHIR are examined. For every data format all corresponding metadata items are extracted and contrasted.

Following the conception of the metadata crosswalk, the next phase includes the identification of metadata items with high relevance for the MeDIC.

Taking into account the results from the literature research, e.g. the FAIR Principles and requirements (which are immanent to the MeDIC structure), essential metadata items are decided upon.

Within the FAIR Principles, the principles F1, F3, F4, A2, I3 and R1.1 have been taken into account, as they directly relate to metadata[4]. F1 postulates that (meta)data has to be assigned with a globally unique and persistent identifier. F3 involves metadata that includes identifier, which clearly and explicitly describes the corresponding data, whereas F4 claims metadata to be indexed in a searchable resource. According to A2 the metadata has to be accessible, even when the data are not obtainable. I3 contains that metadata must include a qualified reference to other metadata. Finally, R1.1 suggests that metadata has to be released with a clear and accessible data usage license[4]. The requirements for the MeDIC resulted from a requirements analysis, that was conducted at the beginning of the development of the MeDIC in 2019.

Table 4 shows the resulting matrix of mappings including the prioritization of the metadata items.

In Fig. 1, the scores regarding prioritization are calculated for the individual data formats and compared graphically in the following grouped bar chart.

As can be seen from the previous illustrations, none of the data formats fulfill all required MeDIC-inherent criteria.

No.	Search Step	Results
#1	"metadata"[MeSH Terms]	413
#2	"standard*"[Title/Abstract] OR "open standard*"[Title/Abstract]	1,415,975
#3	"data warehousing"[MeSH Terms] OR "health information interoperability/standards"[MeSH Terms] OR "health information exchange/standards"[MeSH Terms] OR information storage and retrieval/methods"[MeSH Terms]	17.042
#4	"metadata repository"[Title/Abstract] OR "data integration"[Title/Abstract]	4.071
#5	"Medical Records Systems, Computerized"[MeSH Terms]	45.478
#6	"open EHR"[Title/Abstract] OR "CDISC"[Title/Abstract] OR "FHIR"[Title/Abstract] OR "OMOP"[Title/Abstract]	587
#7	#1 AND (#2 OR #3 OR #4 OR #5 OR #6	205

Table 1. Search strategy in PubMed on 18.05.2022.

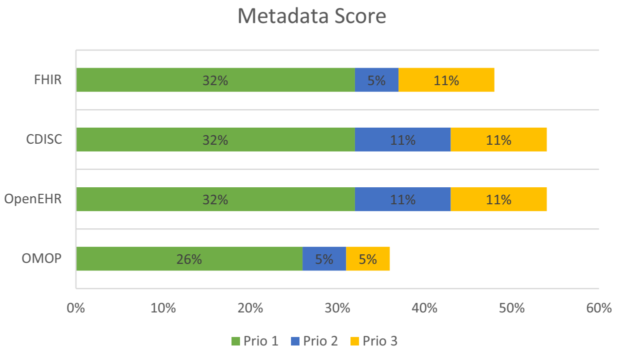


Fig. 1 Qualitative priority scoring of metadata required by the MeDIC and quantitative coverage in the different data formats FHIR, CDISC, OpenEHR and OMOP.

As stated, an entire transformation is not possible because of the different premises the individual formats are based on. CDISC, for example, has added structures, that build upon the ODM format and form in conjunction the basis for study documentation in clinical care. OpenEHR on the other side is designed for the storage of medical data in an EHR, while FHIR is intended for the exchange of data between different institutions. Whereas OMOP provides a common data format to unify data from different databases. It is shown, that none of the data formats include all metadata, which is required to successfully operate the MeDIC for the purpose of reliable data management. So we propose a specific convergence format, which bypasses the described challenges.

Figure 2 shows an example by providing an excerpt of two data formats OMOP and openEHR. It illustrates how the convergence format can incorporate metadata items of different formats and avoid loss of information by providing metadata items of the target format even if it is not part of the source data format. For example the metadata item *cdm_source_abbreviation* has no exact match in openEHR metadata. Without the convergence format, the information would have been lost, because it would be no longer represented in the openEHR metadata items after the transformation.

Additionally openEHR accepts metadata items such as resource_description:lifecycle_state and resource_description_item:language, which have no match within the OMOP metadata. Naturally these two metadata items would have not been created within the transformation process, because OMOP doesn't provide the equivalent structure. The convergence format is therefore the best solution to provide and maintain the format structure, by creating the items within the transformation process and filling them with NULL-values, if the source format doesn't provide any input values.

Discussion

The literature search revealed that the topic of metadata is of high importance in medical and biomedical informatics. A fundamental problem, however, is the definition of metadata. Ulrich *et al*. examines the literature on the definition and classification of metadata and points out the fact, that there is no clear explanation of the term "metadata". Furthermore, the article shows how the definition of matching, mapping and transformation of metadata also differs in the literature. Overall, the article points out possible problems that can result from the heterogeneous understanding of the term[11].

Some authors previously showed the possibility of transforming single data formats into each other. However, the transformation between more than two data formats within a metadata crosswalk has not yet been performed. The works of Doods, Neuhaus & Dugas[12], and Bruland & Dugas[13] for example, show the possibilities of a transformation of openEHR or FHIR into CDISC ODM. Within the MeDIC, however, structured as well as unstructured data and metadata are to be considered, which contradicts a mere use of CDISC ODM as target format.

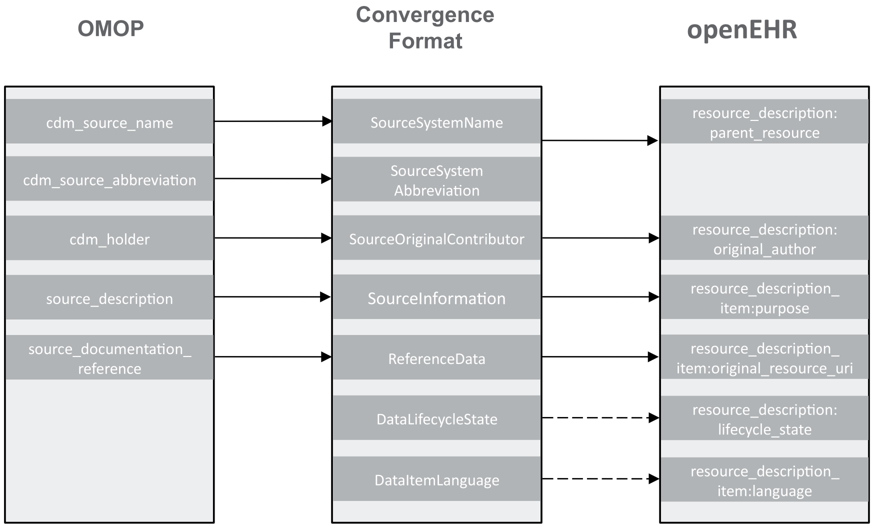

OMOP
Convergence Format
openEHR
cdm_source_name
cdm_source_abbreviation
cdm_holder
source_description
source_documentation_reference
SourceSystemName
SourceSystem Abbreviation
SourceOriginalContributor
SourceInformation
ReferenceData
DataLifecycleState
DataItemLanguage
resource_description: parent_resource
resource_description: original_author
resource_description_item:purpose
resource_description_item:original_resource_uri
resource_description: lifecycle_state
resource_description_item:language

Fig. 2 Example of an excerpt of metadata items from OMOP and openEHR, showing how the convergence format can avoid loss of information during the metadata transformation. Dotted arrows show data flow with NULL-values.

Meaning of local metadata items	OMOP	openEHR	FHIR	CDISC
Version of the metadata	metadata_concept_id	versionID	Meta.versionID	ODM/Study/MetaDataVersion
Identifier of the type of information	metadata_type_concept_id			
Name of the metadata version	name			ODM/Study/MetaDataVersion/Name
Metadata value as string	value_as_string			
Metadata value as concept	value_as_concept_id			
Date of the metadata creation	metadata_date		DataRequirement	ODM/AsOfDatetime
Datetime of the metadata creation	metadata_datetime		Meta.lastUpdated	ODM/AsOfDatetime
Full name of the source	cdm_source_name	resource_description:parent_resource	Meta.profile	def:Origin
Abbreviation of the source name, if applicable	cdm_source_abbreviation			def:Origin
contributor or publisher of the source data	cdm_holder	resource_description:original_author/resource_description:original_publisher	Contributor	def:Origin

Table 2. Comparison of metadata from different data formats frequently used in healthcare information systems and medical research. Note. OMOP Observational Medical Outcome Partnership, openEHR open Electronical Health Records, FHIR Fast Healthcare Interoperability Resources, CDISC Clinical Data Interchange Standards Consortium.

The use of different data formats and the associated metadata formats in health care results in heterogeneity of the metadata items. This leads to the fact that occasionally a complete match between the data fields is unachievable, resulting in inequivalence. Be it that the data formats support different application areas or that they allow different degrees of freedom in the development of extensions.

The findings presented in this article show that metadata items from different standard formats meet the requirements to be transformed into one another with few adaptations, because of the previously mentioned challenges of a metadata crosswalk.

In the next stage, the convergence format will be further developed and an automated crosswalk between the different data formats to and from the convergence format will be established. This convergence format comprises both MeDIC inherent metadata items and all items from the crosswalk of the four data formats depicted in Table 2.

This maximum set of metadata items will be the requisite to fulfill the gap between the metadata currently captured in hospital information systems and the derivatives of data needed in research, to be able to provide metadata to researchers in any data format. Additionally, the quality of the harvested metadata has to be evaluated. If providing metadata to researchers, they must also be assured that it is of high quality and allows safe evaluations. Therefore, a quality assessment schema will be developed. This assessment should lead to a visualization

Metadata Item	Priority	Description
MetadataID	1	Unique and persistent identifier of the metadata
MetadataDate	1	Date of the metadata creation
AffiliateDatasetID	1	Globally unique identifier of the data, which the metadata is associated with
MetadataVersion	1	Version of the Metadata
ReferenceData	1	References to other data via name or description
ReferenceMetadata	1	References to other metadata via name or description
DataLifecycleState	2	State of the data during its lifecycle (creation, processing, analysis, preservation, access, reuse)
UsageLicense/Copyright	1	clear and accessible data usage license
UsageContext	3	Context in which the data should be used
SourceSystemName	1	Explicit name of the Source System
SourceSystemVersion	1	Version of the Source System when recording the (meta)data
SourceInformation	3	Additional information about the source of the data
SourceOriginal Contributor	2	Contributor of the Source data
VestingPeriod	3	Availability of data to other researcher outside the study during the time of the study
ConsentType	2	Type of patient consent (i.e., broad consent, study specific consent)
ConsentValidation	1	Validity period of the consent
ConsentVerification	1	Physical signature of the patient and start of the validity period
ConsentModule	2	Exact parts of the consent, to which the patient consented to
DataItemLanguage	3	Language of the data items

Table 3. Essential metadata required in the MeDIC and respective priority level. Note Priority level 1 = high, 2 = medium, 3 = low.

of the metadata quality, which is then made available to the researchers. This visualization enables a researcher to easily recognize and evaluate the data quality and whether the data is suitable for this research purpose.

Methods

Literature research.

To assess the field of existing metadata standards a literature research was conducted using PubMed and Embase via Ovid.

Table 1. shows the search steps and partial results of the literature search exemplarily in PubMed. The results of this search were combined with a second search query using the same search strategy within Embase.

For the purpose of adequately selecting and evaluating the results of the literature search, articles proposing necessary metadata for data exchange and processing (search criterion one), as well as (meta)data format specification (search criterion two) and articles describing already existing metadata crosswalks (search criterion three) or approaches of transforming/mapping metadata formats into one another (search criterion four), where taken into consideration.

The search in both bibliographic databases yielded 517 results, of which 71 duplicates were removed automatically in Refworks, and 446 references remained. The relevance of every result was examined by scanning the associated title and abstract. After this examination, 60 articles with high importance were left, and full text of these was obtained. After reading the full text, 12 articles were deemed not to be suitable for the purpose of this research. The reference lists of all included articles were scanned for further important publications. Finally, the remaining articles were studied completely and evaluated in relation to the search criteria used to select, evaluate and prioritize the results of the literature search. The proposed FAIR Principles[4] and the documentation of the included data formats where the core answers to the first search criterion, while the documentation also answered the second search criterion and served as the preface for the development of the crosswalk.

Kock-Schoppenhauer *et al*. and Bruland and Dugas showed first elaborations of one-to-one transformations between different data formats, related to search criterion three[13,14]. while[15] and[16] supplied a founded overview over the procedure of a crosswalk.

Metadata crosswalk.

In order to make the medical data collected in the data integration center available to researchers, the metadata are supposed to be made accessible in data formats which are frequently used. Currently the data formats to be supported in the MeDIC include OMOP, openEHR, FHIR and CDISC. This allows researchers to be offered a choice of target data formats. To enable this selection a metadata crosswalk is built.

A metadata crosswalk entails a chart or map which depicts elements from different standards or formats and groups equivalent elements[15]. Crosswalks allow to transform elements from on format to another[16].

The excerpt of the developed metadata crosswalk for these four formats is depicted in Table 2. The complete crosswalk can be found in the Supplementary Table 1 within the Supplementary material of this manuscript.

During a transformation, data fields from the input format sometimes have to be split or merged in order to retain the semantic meaning of the metadata in the target format.

As a result of the above challenges, a loss of information can occur. In order to avoid this, a convergence format, has to be used that includes and stores a maximum set of metadata fields. To establish the convergence format a defined metadata crosswalk serves as the objective of this work.

Metadata Item (Priority level)	OMOP	openEHR	FHIR	CDISC
MetadataID (1)	metadata_ concept_id	versionID	Meta.versionID	ODM/Study/Metadata Version
MetadataDate (1)	metadata_date		DataRequirement	ODM/AsOfDatetime
AffiliateDatasetID (1)	source_ description_ reference	resource_ description_ item:original_resource	DataRequirement. Profile	def:Origin
MetadataVersion (1)	metadata_ concept_id	versionID	Meta.versionID	ODM/Study/Metadata Version
ReferenceData (1)		resource_ description: references	RelatedArtifact	def:LeafElement
ReferenceMetadata (1)				
DatalifecycleState (2)	resource_ description: lifecylce_ state	def:AnnotatedCRF		
UsageLicense/Copyright (1)	resource_ description: copyright			
UsageContext (3)	resource_ description_item:use	UsageContext	StudyDescription	
SourceSystem Name (1)	cdm_source_name	resource_ description: parent_ resource	Meta.profile	def:Origin
SourceSystem Version (1)				
SourceInformation (3)	source_description	resource_ description_ item:purpose	DataRequirement.type	def:Origin
SourceOriginalContributor (2)	cdm_holder	resource_ description: original_ author & resource_ description: original_publisher	Contributor	def:Origin
VestingPeriod (3)				
ConsentType (2)				
ConsentValidation (1)				
ConsentVerifikation (1)				
ConsentModule (2)				
DataItemLanguage (3)	resource_ description_ item:language			

Table 4. Matrix of mapping of priorities to metadata items. Note. OMOP Observational Medical Outcome Partnership, openEHR open Electronical Health Records, FHIR Fast Healthcare Interoperability Resources, CDISC Clinical Data Interchange Standards Consortium.

Prioritization. After the crosswalk has been executed, metadata items which are specifically important to the MeDIC are identified. These metadata are determined based on the literature review and data format specifications. Then, the items meeting the inherent requirements of the MeDIC are categorized. For this purpose, a split into three categories is chosen. Category 1 includes items that are of immense importance for the operation of the MeDIC and for the provision of data. Category 2 includes objects that span a medium importance but are requisite for data privacy and consent. Category 3 consists of metadata with the lowest priority, which are key for additional context information and language specification. An urgency of the corresponding items is not to be considered. For this reason, the priority dimension only includes the importance of the items for the MeDIC.

After prioritization, the scores for the individual data formats are calculated to show which data formats cover items of priority categories 1 and 2 as extensively as possible.

Table 3 shows the identified metadata items and the associated prioritization.

The prioritization is divided into categories 1, 2 and 3. Category 1 contains items with the highest priority, while category 3 contains the items with the lowest priority. 19 Metadata items are identified, which will contribute to the sustainability of the MeDIC in terms of data usage and exchange. Metadata items resulting from the FAIR Principles are assigned the highest priority because of its high importance to make data findable, accessible, interoperable and re-usable.

The mapping of priorities to metadata items is then applied to the metadata of the four different data formats.

Data availability

All data generated or analyzed during this study are included in this article and is referenced in[17].

Code availability

During this study no custom code was used.

Received: 17 June 2022; Accepted: 20 October 2022;
Published online: 28 October 2022

References

1. Baca, M. *Introduction To Metadata* 3rd edn (The Getty Research Institute Publications Program, 2016)
2. Hegselmann, S. *et al.* Automatic conversion of metadata from the study of health in pomerania to ODM. *Stud Health Technol Inform* **236**, 88–96 (2017).
3. Hulsen, T. *et al.* From big data to precision medicine. *Front Med* **6** (2019).
4. Wilkinson, M. *et al.* The FAIR guiding principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
5. Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., McGillivray, B. The citation advantage of linking publications to research data. *PloS ONE* **15**(4) (2020).

6. Dugas, M. *et al*. Memorandum "open metadata". *Methods Inf Med* **24** (2015).
7. Ohno-Machado, L. *et al*. Finding useful data across multiple biomedical data repositories using DataMed. *Nature Genetics* **49**(6), 816–819 (2017).
8. Ferreira, J. D., Inácio, B., Salek, R. M. & Couto, F. M. Assessing public metabolomics metadata, towards improving quality. *J Integr Bioinform* **14**, 20170054 (2017).
9. Gonçalves, R. & Musen, M. The variable quality of metadata about biological samples used in biomedical experiments. *Sci Data* **6**, 190021 (2019).
10. Canham, S. & Ohmann, C. A metadata schema for data objects in clinical research. *Trials* **17**, 557 (2016).
11. Ulrich, H. *et al*. Understanding the nature of metadata: systematic review. *J Med Internet Res* **24**(**1**) (2022).
12. Doods, J., Neuhaus, P. & Dugas, M. Converting odm metadata to fhir questionnaire resources. *Stud Health Technol Inform* **228**, 456–460 (2016).
13. Bruland, P. & Dugas, M. Transformations between cdisc odm and openehr archetypes. *Stud Health Technol Inform* **205**, 1225 (2014).
14. Kock-Schopenhauer, A. K. *et al*. Compatibility between metadata standards: import pipeline of cdisc odm to the samply.mdr. *Stud Health Technol Inform* **247**, 221–225 (2018).
15. Riley, J. *Understanding Metadata* (Bethesda, MD: NISO Press, National Information Standards Organization, 2004).
16. Baca, M., Gilliland, A. *Introduction to metadata: pathway to digital information* (Getty Education Institute for the Arts, 2000).
17. Bönisch, C., Kesztyüs, D. & Kesztyüs, T. Harvesting metadata in clinical care: a crosswalk between FHIR, OMOP, CDISC and openEHR metadata. *figshare* https://doi.org/10.6084/m9.figshare.21333042.v1 (2022).

Acknowledgements

The authors acknowledge funding, received from the German Federal Ministry of Education and Research for the project FAIRRMeDIC, grant number 01ZZ2012. This work was supported by the Medical Data Integration Center of the University Medical Center Göttingen.

Author contributions

T.K. coordinated and supervised this research project. C.B. performed the metadata crosswalk and contributed to the realization of the priority classification and the scoring, and wrote the initial manuscript. D.K. and T.K. contributed to the conceptualization of the research and were involved in the revision, editing and final approval of the manuscript. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-022-01792-7.

Correspondence and requests for materials should be addressed to C.B.

Parciak *et al.*
BMC Medical Informatics and Decision Making *(2023) 23:94*
https://doi.org/10.1186/s12911-023-02195-3

BMC Medical Informatics and Decision Making

DATABASE **Open Access**

FAIRness through automation: development of an automated medical data integration infrastructure for FAIR health data in a maximum care university hospital

Marcel Parciak[1,2,3†], Markus Suhr[1,4†], Christian Schmidt[1], Caroline Bönisch[1], Benjamin Löhnhardt[1], Dorothea Kesztyüs[1*] and Tibor Kesztyüs[1]

Abstract

Background Secondary use of routine medical data is key to large-scale clinical and health services research. In a maximum care hospital, the volume of data generated exceeds the limits of big data on a daily basis. This so-called "real world data" are essential to complement knowledge and results from clinical trials. Furthermore, big data may help in establishing precision medicine. However, manual data extraction and annotation workflows to transfer routine data into research data would be complex and inefficient. Generally, best practices for managing research data focus on data output rather than the entire data journey from primary sources to analysis. To eventually make routinely collected data usable and available for research, many hurdles have to be overcome. In this work, we present the implementation of an automated framework for timely processing of clinical care data including free texts and genetic data (non-structured data) and centralized storage as Findable, Accessible, Interoperable, Reusable (FAIR) research data in a maximum care university hospital.

Methods We identify data processing workflows necessary to operate a medical research data service unit in a maximum care hospital. We decompose structurally equal tasks into elementary sub-processes and propose a framework for general data processing. We base our processes on open-source software-components and, where necessary, custom-built generic tools.

Results We demonstrate the application of our proposed framework in practice by describing its use in our Medical Data Integration Center (MeDIC). Our microservices-based and fully open-source data processing automation framework incorporates a complete recording of data management and manipulation activities. The prototype implementation also includes a metadata schema for data provenance and a process validation concept. All requirements of a MeDIC are orchestrated within the proposed framework: Data input from many heterogeneous sources, pseudonymization and harmonization, integration in a data warehouse and finally possibilities for extraction or aggregation of data for research purposes according to data protection requirements.

[†]Marcel Parciak and Markus Suhr contributed equally to this work.

*Correspondence:
Dorothea Kesztyüs
dorothea.kesztyues@med.uni-goettingen.de
Full list of author information is available at the end of the article

Conclusion Though the framework is not a panacea for bringing routine-based research data into compliance with FAIR principles, it provides a much-needed possibility to process data in a fully automated, traceable, and reproducible manner.

Keywords Medical data reuse, Electronic health record, Medical data integration center, Automated medical data processing, Medical informatics, Maximum care hospital

Background

Cross-organizational secondary use of medical data is the key to large scale clinical and health services research and essentially important for establishing precision medicine. Reuse of routinely collected data offers extended sample sizes and follow-up times at lower costs and a more representative view of clinical practice in the real-world [1]. The "FAIR Principles for scientific data management and stewardship" were established to make data and the context of their generation Findable, Accessible, Interoperable, and Reusable. These principles summarize common data governance guidelines across multiple research domains with a special emphasis on the automatization of finding and using data [2]. Open data sharing platforms like DataONE or the meta-repositories like DataMed show the applicability of FAIR in real world examples [3, 4]. Both examples enable sharing of research datasets for reuse. However, the data platform presented here, collects and processes data generated during diagnostics and treatment of patients in clinical care at the university hospital and thus sets itself apart from the pre-processed and target-oriented research data, that DataONE and DataMed lean on. In the area of medical research, privacy concerns still remain about the open sharing health data, which oppose the publication of datasets in central repositories [5]. This holds especially true for real world data captured in routine healthcare, which contains patient identifying attributes and requires explicit legal clearing for secondary use in research. In Germany, the Medical Informatics Initiative (MI-I) funds development and implementation of medical data integration centers to create a technical and legal framework for cross-site secondary use of routine healthcare data [6]. As part of the HiGHmed consortium and the MI-I funding scheme, the University Medical Center Göttingen (UMG) implemented such a medical data integration center (UMG-MeDIC) [7]. Establishing data warehousing processes from scratch, we aim for high compliance with the FAIR Principles but face the challenge that data integration workflows are complex and inefficient when done manually [8]. As with any complex software engineering task, documentation is often neglected, not only for software artifacts themselves, but also for any executed workflow [9]. Although data about past workflow runs is highly useful, capturing this type of information is a difficult task [10].

Problem statement

Sharing research data necessitates an infrastructure that allows data to be found and accessed in an interoperable and reusable format [2]. As real world health data tends to stay in heterogeneous non-FAIR data silos, data engineers need to implement appropriate data integration workflows to make this data available [8, 11]. In contrast to other data-intensive research domains, the healthcare domain imposes additional requirements on data engineering [5, 12, 13]. Moreover, the UMG-MeDIC operates on a continuous flow of data instead of self-contained datasets. Current approaches for FAIR research data management tend to focus on manually "FAIRifying" individual datasets that are published in dedicated research data repositories, neglecting data processing steps prior to obtaining data [14]. Our data integration workflows have to be executed periodically and repeatedly, continuously moving data from multiple source systems to a central data warehouse component to again many target systems. This results in constantly evolving datasets, as illustrated in Fig. 1, that defy manual "FAIRification".

A trustworthy, fully validated data basis, which consists of structured and unstructured data, is therefore the minimum requirement for successful operation. All data processing and management tasks and comprehensive data provenance must be accounted for to enable meaningful scientific application.

Hence, in this article, we describe our implementation approach for a data processing automation framework in a medical data integration center of a maximum care hospital facing the challenges of multi-dimensional data warehousing and processing. Based on this framework, the UMG-MeDIC is dedicated to serve as a research service unit.

Methods

In order to obtain a comparative overview of proposed approaches from other projects with similar requirements, a comprehensive literature review must first be conducted. The results are then described and

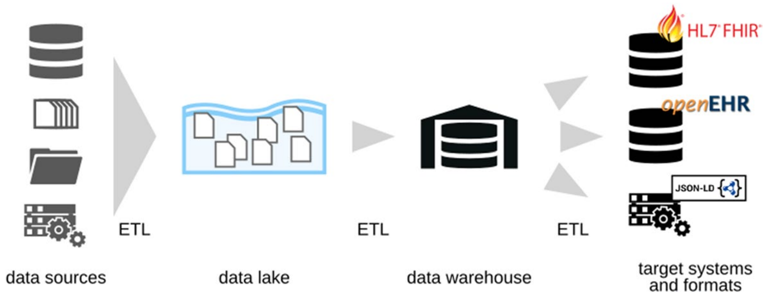


Fig. 1 High level view of the logical data flow at the UMG-MeDIC, depicting the different stages of the Extract-Transform-Load process. The information (i.e. healthcare data from the University Medical Center Göttingen) is extracted from the data sources and pooled in a data lake. Within the transformation and loading step the data is pushed to the data warehouse and is then provided to the target systems in the required format. UMG-MeDIC University Medical Center Göttingen-Medical Data Integration Center, ETL Extract-Transform-Load

discussed in terms of a possible solution to the specific requirements of our MeDIC project.

In a second step, the goals for the implementation of the framework in the MeDIC are defined. Additionally, frequently used concepts and standards are introduced.

The third step identifies workflow sub-processes and corresponding software tools for the realization.

The fourth and last step involves the implementation of the objectives into the framework of the UMG-MeDIC, using the appropriate open-source software components.

Literature review

PubMed, Embase via Ovid and Web of Science were searched using specific search terms and keywords. The search strategy for PubMed is depicted as an example in Table 1.

In addition to PubMed, the search in Web of Science resulted in 120 hits, the one in Embase delivered 98. Subsequently, a total of 356 references was imported into Refworks and 129 duplicates were removed automatically, leaving 219 references to be screened. Some of the concepts under consideration will be presented in the following.

The FAIR principles have gained a lot of momentum in the research community, resulting in various solutions and proof-of-concepts presenting FAIR data. Initiatives like DataMed and GO FAIR further imply that using Fast Healthcare Interoperability Resources (FHIR) means a central repository component is sufficient to enable persistent accessibility and this architecture would scale for large data volume [4, 12]. Usage of semantic modeling languages like FHIR or openEHR may contribute to overall FAIRness and especially reusability of data [15]. Contrary to the usage of FHIR within GO FAIR or DataMed, the UMG-MeDIC incorporates a data warehouse as central repository component, using a Structured Query Language (SQL)-based database, to store pseudonymized data. FHIR is, in the environment of the UMG-MeDIC, rather intended as an exchange format, than a central repository component.

The Emergency Department Catalog (EDCat) system was developed to improve the FAIRness of a project considering emergency department databases but still requires manual organization of datasets, which, in view of the volume of data, is not suitable for the operation of a MeDIC and further pursuit of this solution was discarded [14].

The SCALEUS-FD offers a semantic web tool that allows data integration and reuse in compliance with FAIR Data principles and was validated in the domain of rare diseases, where records are rather small, not comparable to the volume of data that has to be processed in a maximum-care hospital each day [16]. Since there was no experience with large numbers of data

Table 1 Search history in PubMed

Search Step	Search string	Number of hits
#1	"FAIR principles"[All Fields] OR ("FAIR"[ti] AND "principles"[ti])	103
#2	findab*"[All Fields] AND "access*"[All Fields] AND "interopera*"[All Fields] AND "reusab*"[All Fields]	248
#3	"data warehousing"[MeSH Terms] OR "databas*"[All Fields]	658,390
#4	(#1 OR #2) AND #3	138

points and records in the SCALEUS-FD, but the way to handle big data was mandatory for the UMG-MeDIC, this approach was not considered further.

On the contrary, the YOUth cohort study, a large-scale longitudinal cohort study with highly sensitive data, faced similar requirements concerning privacy, heterogeneous data and sources, and data quality checks [17]. The most important and decisive difference, however, lies in the way and objective of the data collection. While the data of the well-defined cohort are obtained a priori for research purposes, standardized and per protocol, the data of a MeDIC are primarily routinely collected, often also referred to as "real world data" [18]. The implicit difference in standardization and quality requires a correspondingly differentiated and more elaborate data management in order to finally make the latter usable for research.

The European Medical Information Framework (EMIF) created a catalog of data sources from research studies and routine clinical care to enable researchers to find, access, and reuse datasets while respecting privacy [19]. For this purpose, a four-layer concept was developed in which each layer can be authorized individually, thus enabling different degrees of data access. However, unlike the MeDIC, they do not primarily aim to integrate the data itself, but rather consolidate various data resources into an overarching biomedical marketplace based on FAIR principles [20].

In the biomedical environment, a vast amount of data for processing, in other words big data, concerns not only but above all omics-data. Hence, the tools and solutions applied can be similar to our concept, with the crucial difference that the diversity of data types is generally low in omics-data and very high in a maximum-care university hospital and therefore in the UMG-MeDIC.

In summary, there are many examples from different fields that aim to provide data according to FAIR principles to enable further research and offer the best possible patient-centered care or precision medicine. However, we could not find a solution that meets all our challenges of different data types, large amounts of data, high-velocity and timeliness, and processing of structured and non-structured clinical care data.

Goals

As no appropriate pre-existing solution could be identified, as portrayed in the prior subsection, we iteratively implement a prototype based on best of breed open source components. Primary goals for our implementation of a trusted medical research data integration system were:

- Operate an integrated data warehouse
- Ingest data from any number of source systems in any data format in batch or near real-time schedule
- Provide data into a varying number of target systems in data format
- Ability to scale data processing flexibly
- Orchestrate all data processing tasks across network security zone barriers
- Ability to monitor status and history of all data processing tasks
- Operate the entire system in a high level information security context with a special focus on data integrity and (long term) accessibility

In addition to the software used, which will be identified and presented in the next step, frequently used standards and concepts are introduced at this point for a better understanding of the approach. Table 2 provides an overview of these standards and concepts with a brief explanation of each.

Identification of workflow sub-processes and corresponding software tools

The required data integration workflows for operating the UMG-MeDIC (according to the goals defined above) can be divided into sub-processes and strategically split into modules, which are equal for many workflows. In our bottom-up description of the sub-processes, we start with the atomic data processing tasks, which can be summarized as follows: A data processing task takes input data and manipulates it in a well-defined manner to produce output data whenever triggered by the process control flow. Each processing task must be documented with process metadata containing all necessary information to recreate the exact parameters of its execution as well as the relevant runtime environment. The full process control flow includes the actual data processing task and the recording of process metadata (see Fig. 2). These atomic processing tasks may be concatenated by the control flow into an end-to-end data processing workflow that moves information across multiple storage systems and formats.

Based on the aforementioned sub-process tasks, we had to take into account an agnostic implementation of this fundamental framework. This includes the programming language or software tool used to implement actual data processing tasks, and information model used to document process metadata to be manufacturer-independent. We have chosen the following components: An orchestration system that manages process control flow, a compatible task execution engine, and a semantic

Table 2 Standards and concepts used in the development of the UMG-MeDIC framework

Term / acronym	Resolved	Short description
ACID	Atomic Consistent Isolated Durable	A set of standard properties that ensure reliable processing of database transactions
CSV	Comma Separated Values	Structure of a text file for storage or exchange of simply patterned data
Data lake		System or repository of structured or unstructured data, including raw copies of source system data and transformed data
DWH	Data warehouse	A central database optimized for analysis purposes that combines data from several, usually heterogeneous sources
DRG	Diagnoses Related Groups	Diagnosis-related grouping of patient cases with similar costs, used for medical billing
ETL	Extract Transform Load	An integrative process in which data is extracted from multiple sources, which may have different structures, processed, and merged into a target database
FAIR	Findable Accessible Interoperable Reusable	The principles were defined in 2016 by a consortium of scientists and organizations and emphasize machine-actionability with regard to the increase in volume, complexity, and creation speed of data
FHIR	Fast Healthcare Interoperability Resources	Fast Healthcare Interoperability Resources is a standard developed by HL7 that supports data exchange between healthcare software systems
HL7	Health Level 7	Health Level 7 is a non-profit, ANSI-accredited organization developing standards for the exchange of information between healthcare services
HTTP	Hypertext Transfer Protocol	Regulates the communication between browser and web server
IDE	Integrated Development Environment	Application development software that combines common developer tools in a central graphical interface
IRIs	Internationalized Resource Identifiers	internationalized form of the Uniform Resource Identifier (URI), an identifier consisting of a string of characters used to identify an abstract or physical resource
open EHR	open Electronic Health Record	An open standard health informatics specification for managing, storing, retrieving, and exchanging health data in electronic health records
LOINC	Logical Observation Identifiers Names and Codes	international standard of universally accepted names and identifiers of health measurements, observations and documents
REST API	Representational State Transfer Application Programming Interface	REST APIs communicate via HTTP requests to perform standard database functions such as creating, reading, updating, and deleting records within a resource
SQL	Structured Query Language	SQL is a database language for defining data structures and for editing and querying datasets in relational databases
TLS	Transport Layer Security	Encryption protocol for secure data transmission on the Internet
URL	Uniform Resource Locator	Identifies and locates a resource via the access method to be used and the location of the resource, e.g. web page via HTTP

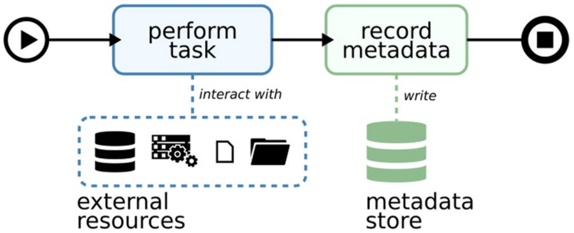


Fig. 2 Generic schema of an atomic Extract Transform Load task and metadata capture sub process. Process control flow: black lines left to right; data flow: blue outline; metadata flow: green outline. The process is started with a "perform task" which interacts with external resources and pulls data from different sources. Subsequently the process flow enables the recording of metadata that is written in a separate metadata storage

model, storage service and documentation engine for process metadata.

For our prototype implementation, we selected a set of tools published under open source licenses, described in Table 3, to implement these components.

Results

In this section, the implementation of our framework and the inherent workflow are described and an example is given to illustrate the function of the framework.

Currently, the following hospital department systems are connected to the UMG-MeDIC: Laboratory system (Opus::L), administration system (SAP: transaction data, billing data), microbiology and virology system (MLAB), clinical tumor documentation (Onkostar), cardiology system (CCW, including echocardiography, cardiac catheterization), sensor data intensive care (ICCA),

Table 3 Overview of tools implemented to perform the tasks of the UMG-MeDIC

Process Metadata				
Software	**JSON-LD**	**Schema.org**	**CouchDB**	
Description	A serialization technique for linked data (LD) objects using the JavaScript Object Notation (JSON) [21, 22]. This technique allows assigning unique identifiers using Internationalized Resource Identifiers (IRIs) for JSON objects and consequently to use these identifiers as references [23]	The collaborative and hierarchical vocabulary allows to create semantically standardized and linked metadata information [24]. It is serializable in different formats including JSON-LD. All elements from Schema.org are described in detail, allowing to define metadata understandable by humans and machines alike. Predefined types like Schema.org Dataset or DataDownload summarize relevant metadata elements that can be attached to data from the UMG-MeDIC [24]	Apache Cluster of unreliable commodity hardware Data Base (CouchDB) is a JSON based document database [25]. Data can be read, written, modified, or deleted using a Representational State Transfer Application Programming Interface (REST API). Built for large deployments, CouchDB allows to be quickly replicated to multiple servers while maintaining the ACID (atomic, consistent, isolated, durable) properties of the database. Solution to store JSON-LD process metadata documents	
Process Flow				
Software	**ActiveWorkflow**	**Docker**	**Celery**	**Data Storage CDSTAR**
Description	A web-based automation engine to orchestrate and monitor workflows [26]. A web-GUI allows to define and run a workflow consisting of individual agents. Each agent is an autonomous software service that communicates with the workflow engine via a standardized REST API. Workflows can be executed event based or on a predefined schedule	A runtime for software containers [27]. A software container is a lightweight and interoperable application bundle. These bundles include all requirements and can be run by the Docker engine, which manages networking, data storage and monitoring of run containers. Running containers with Docker is a lightweight, software-defined alternative to server virtualization. Runtime environment templates called Docker images enable fully reproducible process execution as well as encapsulation and preservation of the entire virtualized runtime environment	Highly scalable distributed message queue for task scheduling [28]. It works message-based, brokering messages to worker nodes and collecting results into a backend. Used to asynchronously execute atomic data processing tasks	Common Data Storage Architecture (CDSTAR) is a data storage abstraction layer [29]. It abstracts physical storage solutions and offers storing, reading and modifying data via a REST API. CDSTAR is organized into vaults, for which individual authorization can be set. A vault holds archives, uniquely identified by an archive ID. An archive contains individual files, identified in turn by internal ID or a filename. CDSTAR is used as a lightweight data lake solution

medication and substances (Meona), and emergency admission (E.Care). Soon to follow will be clinical trial software (secuTrial), image data (PACS), the pathology system (Nexus/IMS), surgical data (Win-OP), radiology (structured findings), treatment quality data (QS-Med), and dental data. Since the UMG-MeDIC is still in the process of being set up, not all departmental systems are connected yet. However, these will all follow and thus contribute the full range of data to be expected in a maximum care university hospital.

Overview

The UMG-MeDIC infrastructure follows the microservice paradigm. We operate each application as an autonomous service. ActiveWorkflow orchestrates process flows along these micro-services with workflow specifications being defined by UMG-MeDIC data engineers. Based on a given workflow definition, ActiveWorkflow communicates with the other services through so-called agents. An agent implements the ActiveWorkflow REST API and works autonomously and asynchronously of ActiveWorkflow itself. ActiveWorkflow communicates with the agents using JSON document in the HTTP message body. A JSON message can hold any (text-based) payload. Figure 3 shows an overview of the service ecosystem used to implement our task automation framework.

Against the background of the volume of data to be handled daily we aimed to automate the workflows for managing all kind of incoming clinical care data and the subsequent curating and provision processes of research data. Metadata about this automated processing have to be captured, collected, and published to enable traceability and reproducibility of data transformation as well as the publication of detailed data provenance records. By using JSON-LD metadata templates, we capture relevant information and create linked data compatible with provenance information models. Our implementation

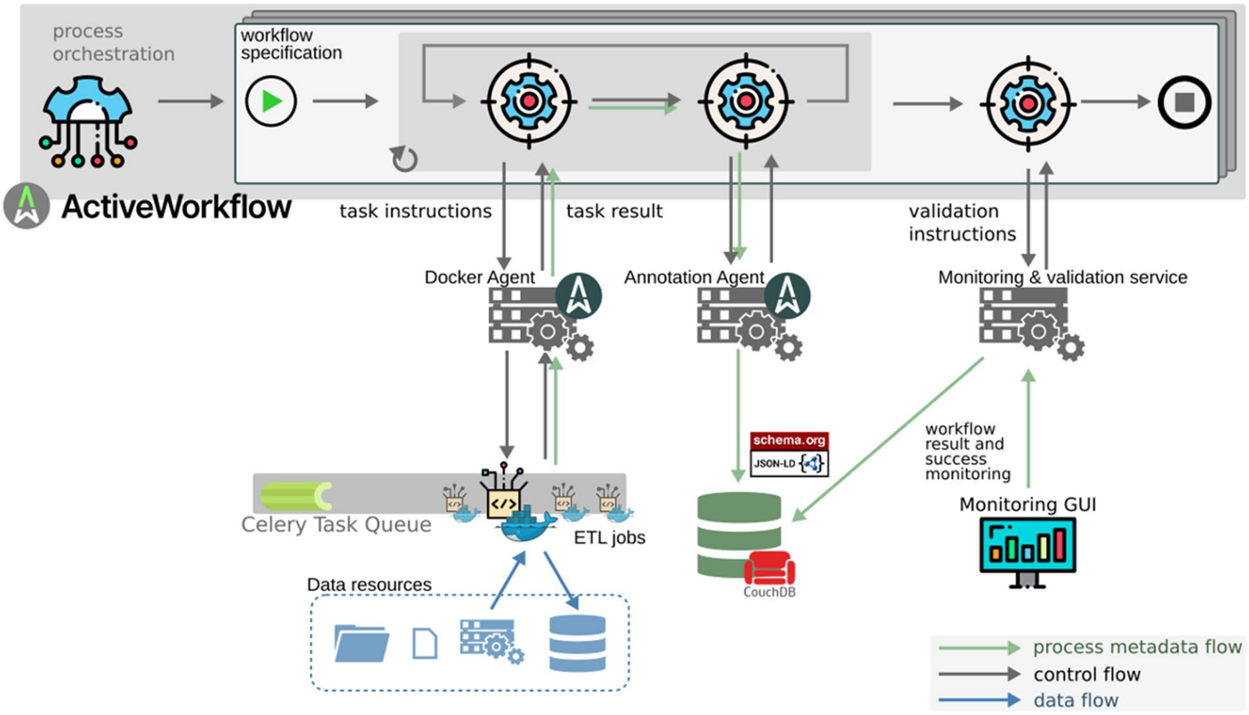


Fig. 3 System architecture of the implementation. The complete process is, as described in the text, divided in tasks, which are controlled by the ActiveWorkflow system. ETL Extract-Transform-Load, GUI Graphical User interface

uses containerization to allow platform-agnostic execution and reproduction of any data integration workflow. We employ data lake web services to persist copies of all incoming source and intermediate data artifacts. All components are autonomous and communicate through RESTful web service application programming interfaces (API), allowing to be operated in compliance with strong IT-security policies.

Metadata processing

We implemented a number of custom ActiveWorkflow agents to assist our data integration workflows. The Docker Agent allows executing generic Docker, and in our case, these Docker images are usually Extract Transform Load (ETL) jobs implemented in Python. The Annotation Agent collects process metadata and writes this data to our CouchDB process metadata store, which contains JSON-LD documents based on Schema.org metadata templates. These are two types of documents: metadata regarding datasets used as input or produced by ETL jobs, and metadata concerning the processes that manipulated a dataset. Figure 3 depicts all flows of metadata as green arrows.

The sequence of Docker Agent and Annotation Agent tasks in a workflow specification can be repeated as necessary until all logical steps of a desired ETL pipeline are completed. As a simple example, the "extract" part could be split from the "transform" and "load" parts of a pipeline to first store a persistent copy of the source data before applying further manipulation. A full example of data processing pipelines at UMG-MeDIC is described below.

Templates

We defined metadata templates in order to standardize capturing of relevant information during data integration workflows. Schema.org definitions for Dataset and DataDownload provide the basis for archives and files of our CDSTAR data lake, respectively [24]. Modeled after CDSTAR, a Dataset document has multiple parts (hasPart) of DataDownload documents. The inverse property isPartOf pointing from DataDownloads to a Dataset does also hold. Both metadata documents link to their CDSTAR counterparts. The data contained in the metadata documents consists of two parts: a redundant copy of the metadata provided by CDSTAR as well as manual descriptions written by the data engineer responsible for a data integration workflow. The manual descriptions are added through the workflow definition in ActiveWorkflow to every dataset processed.

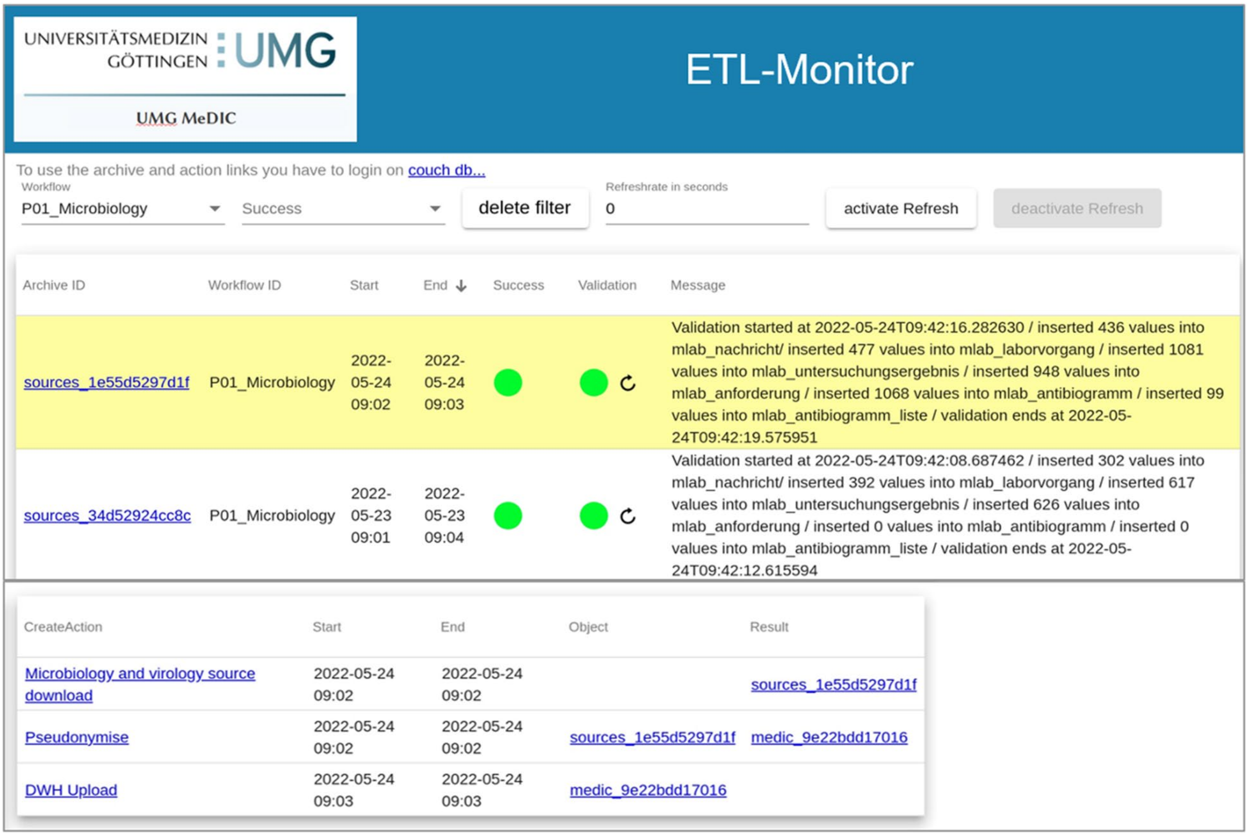
Processes are modeled as CreateAction documents. These documents contain an object and a result reference element to indicate input and output data respectively. An instrument element references an implemented data integration processing step represented as a SoftwareApplication. The SoftwareApplication object again references the source code repository in GitLab as “downloadUrl” to uniquely identify the code that ran. Moreover, supportingData refers to any configuration variables that may be supplied to the ETL process implementation influencing the code execution. Finally, each process metadata document contains an isPartOf reference to the workflow description. The workflow description is manually created as human-readable process documentation. We support this manual documentation step by exporting the workflow definition from ActiveWorkflow and enriching it with free-text descriptions.

Workflow monitoring and validation

We implemented a custom monitoring service and web-based user interface based on the collected process metadata. The monitoring service extracts process metadata documents from CouchDB and displays it in the interface from a workflow perspective. All processing steps that belong to a specific run of any defined workflow are displayed in sequential order. If any workflow run fails to reach its successful final state, a visual warning is displayed to the user. Figure 4 shows the ETL-Monitor running at the UMG MeDIC. To validate correct execution of each workflow run, a second status indicator is presented based on the actual data that was processed. During workflow execution, statistical parameters of the processed data are read and captured as part of the process metadata documents. At the end of each workflow run, a validation service checks whether the same statistical parameters can be re-calculated on the data loaded into the target system. Since data format, content, and expected transformations differ for each source or target system, custom validation functions are implemented per workflow. We start with a set of simple methods, like row counts, and develop more complex functions over time as needed. Validation results are again stored as part of the process metadata in CouchDB.

Example workflow

To better illustrate how the proposed data processing framework translates into practical application, we describe an exemplary workflow in more detail. Data



ETL-Monitor

To use the archive and action links you have to login on couch db...

Workflow: P01_Microbiology
Success
Refreshrate in seconds: 0

Archive ID	Workflow ID	Start	End ↓	Success	Validation	Message
sources_1e55d5297d1f	P01_Microbiology	2022-05-24 09:02	2022-05-24 09:03	●	● ↻	Validation started at 2022-05-24T09:42:16.282630 / inserted 436 values into mlab_nachricht/ inserted 477 values into mlab_laborvorgang / inserted 1081 values into mlab_untersuchungsergebnis / inserted 948 values into mlab_anforderung / inserted 1068 values into mlab_antibiogramm / inserted 99 values into mlab_antibiogramm_liste / validation ends at 2022-05-24T09:42:19.575951
sources_34d52924cc8c	P01_Microbiology	2022-05-23 09:01	2022-05-23 09:04	●	● ↻	Validation started at 2022-05-24T09:42:08.687462 / inserted 302 values into mlab_nachricht/ inserted 392 values into mlab_laborvorgang / inserted 617 values into mlab_untersuchungsergebnis / inserted 626 values into mlab_anforderung / inserted 0 values into mlab_antibiogramm / inserted 0 values into mlab_antibiogramm_liste / validation ends at 2022-05-24T09:42:12.615594

CreateAction	Start	End	Object	Result
Microbiology and virology source download	2022-05-24 09:02	2022-05-24 09:02		sources_1e55d5297d1f
Pseudonymise	2022-05-24 09:02	2022-05-24 09:02	sources_1e55d5297d1f	medic_9e22bdd17016
DWH Upload	2022-05-24 09:03	2022-05-24 09:03	medic_9e22bdd17016	

Fig. 4 ETL-Monitor: Extract-Transform-Load-processes are displayed with their respective status (success, fail). Here, for example, the work flow of importing microbiology data is displayed. Clicking on a specific process provides detailed information. If the process failed, the last step that was successful is displayed. UMG-MeDIC University Medical Center Göttingen-Medical Data Integration Center, DWH Data Warehouse

storage systems, the type of information to process, and the methods to achieve processing are subjective choices that fit our specific situation. The framework can also be applied with entirely different implementation choices.

The logic of the example workflow is as follows: laboratory results enriched with LOINC codes for all patients treated at the UMG are communicated to all hospital department systems in HL7 Version 2 standard "ORU" (HL7 Observation Result) messages, via a clinical communication server. The UMG-MeDIC is registered as a receiver of this message stream, which is the primary input for the workflow. A HL7 message contains identification data (IDAT) of the patient, like the patient id, the name and the date of birth. Additionally, the HL7 message includes the medical data (MDAT) such as lab values, LOINC codes etc. Information from these messages is to be extracted and pseudonymized in a process, where the IDAT is extracted, and a unique number is created by a special algorithm. The personal data, such as the name, is deleted and only the year is stored from the date of birth. This step replaces the IDAT with the pseudonym data (PDAT). The complete pseudonymization process takes place in a protected network segment called "patlan". After pseudonymization, the MDAT connected to the PDAT is ready to be transferred to the "medic" network segment. The information of the message is then transformed to a relational data model and stored in the central data warehouse system. From the relational database system the information is again extracted and transformed into a HL7 FHIR standard "Observation" resource (representing diagnostic and clinical data), and finally stored in a FHIR Server. Resources in the FHIR Server can ultimately be used for cross-organizational querying of medical data for research purposes within MI-I projects. Figure 5 is a graphical representation of this data flow logic. Each processing and annotation step is implemented using the framework described above.

We employ two distinct data lake instances based on CDSTAR as persistent object storage for all data integration processes. It assigns each dataset a unique identifier, the ArchiveID. Any agent is able to use the combination of CDSTAR URL and ArchiveID to identify and download any dataset in our data lake. CDSTAR enables versioning of datasets. Each modification of any dataset will result in a new version. In Figs. 3 and 5, flow of data artifacts is shown as blue arrows. Two instances are used to split raw data containing patient identifying information and pseudonymized data into different network security zones as required by the UMG-MeDIC information security policy.

We use MariaDB as a curated relational data warehouse. The data warehouse contains pseudonymized and transformed medical datasets. Relational database schema and table definition are defined in consultation of the respective source data custodians. The schema aims to cover as much information as possible provided from the sources and enables integrated queries across data from all sources. In addition, all information present in the data warehouse contains an attribute ArchiveID. The ArchiveID refers back to a dataset in the data lake, indicating the origin of any data in the data warehouse.

The exemplary lab result workflow concludes with an ETL step that extracts information from the data warehouse, creates HL7 FHIR Observation resources, and stores these in a FHIR server. This step is an example of how integrated data at UMG-MeDIC may be curated for researcher access and offered in different formats, according to the specific requirements of the research project. Transformation to FHIR format can be replaced

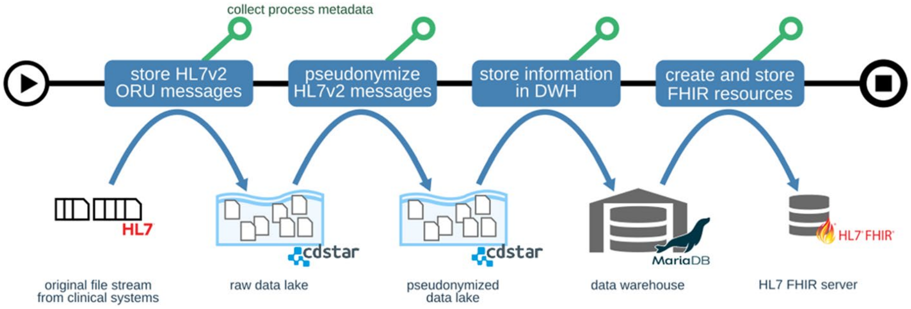


Fig. 5 Schematic representation of data flow between different storage systems. The example workflow shows transfer of laboratory result information through common stages of data processing and storage at UMG-MeDIC. The process starts with the HL7 file stream from the clinical systems, where the observation results are stored in a ORU message and the corresponding process metadata is collected. The information is pooled into a raw data lake. Subsequently the information is pseudonymized and transferred in a pseudonymized data lake. After preprocessing the information is stored in the data warehouse. In a final step FHIR resources based on the data are created and stored in a HL7 FHIR server. UMG-MeDIC University Medical Center Göttingen-Medical Data Integration Center, HL7 Health Level 7, ORU HL7 Observation Result, FHIR Fast Healthcare Interoperability Resources, DWH Data Warehouse

with any given semantic data model, e.g. openEHR or custom CSV. These output formats vary and will be regularly extended by the data engineering team. The generic nature of our data processing framework supports frequent addition of new output data pipelines.

Example ETL-process

Figure 6 illustrates the exemplary ETL-workflow of microbiology data from the source system (MLAB) to the DWH in several steps. First, the MLAB files are copied from the mount-folder to a working directory (1), then the files are loaded into the patlan CDSTAR (2) and metadata concerning (2) is written to CouchDB (3a). The HL7 file is pseudonymized (3b) and subsequently parsed, and the information it contains is inserted into a series of tables in the DWH (4a) and respective metadata are written to CochDB (4b).

These inserts include the patient and case number information stored in the HL7 file, the particular laboratory tests requested and their findings, and the sample material. The latter refers, for instance, to the body region from which the material was taken, and enables the storage of information on multiple samples that may occur in the HL7 files. Furthermore, the result of each examination and a reference to the corresponding laboratory test are stored. Antibiograms pertaining to the possible resistance of bacteria to antibiotics are also stored in separate tables to support multiple antibiograms with various bacteria and different antibiotics. Finally, metadata is written to CouchDB regarding the DWH-upload (5).

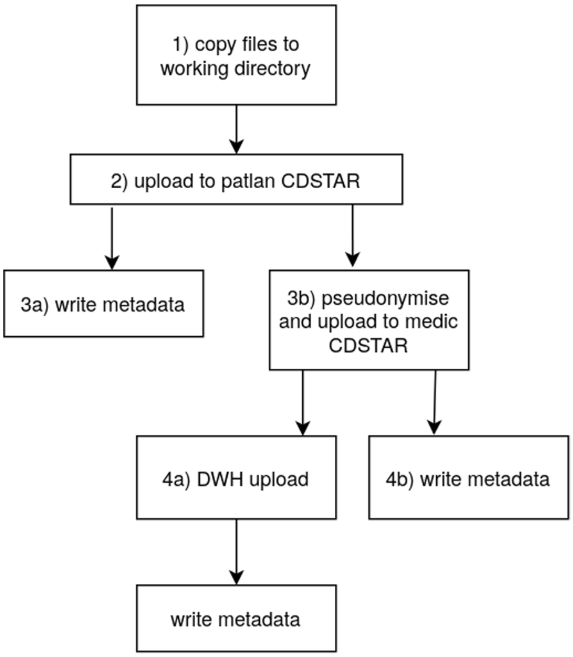

1) copy files to working directory
2) upload to patlan CDSTAR
3a) write metadata
3b) pseudonymise and upload to medic CDSTAR
4a) DWH upload
4b) write metadata
write metadata

Fig. 6 Graphical representation of a standard ETL Workflow. ETL Extract-Transform-Load, CDSTAR Common Data Storage Architecture

Discussion

Implementation of the framework

The proposed data processing automation framework meets the requirements for data management tasks and contextual constraints of the UMG-MeDIC as defined above. We successfully implemented the framework and operate on real-world data from many source systems. Source data is persistently stored in data lake services and transferred into an integrated relational data warehouse. Information from the data warehouse is collected into different subsets, transformed and stored into target systems for research use cases such as the Medical Informatics-Initiative or the HiGHmed project. The components are divided into multiple network security zones as required by the UMG-MeDIC information security policy. Communication across network zones is allowed along well-defined unidirectional HTTP routes while still enabling full workflow control through a single ActiveWorkflow GUI for the data engineering team. Capability for data processing can be increased by horizontally scaling the Celery task queue to include more compute nodes if necessary. Monitoring of workflow progression and success is implemented based on automatically captured process metadata and enables quick status checks for data engineers in day-to-day operation of many automated workflows. Maintaining long-term integrity and availability of the handled data is a challenge for the operational processes and therefore outside the scope of the proposed framework. The choices to permanently store input and intermediate datasets in data lake services and tracking all process metadata are fundamental in enabling long-term preservation, routine integrity checks, and (public) availability of metadata.

Data flow and metadata annotation are implemented as independent yet interwoven subsystems of microservices. It is possible to extend or exchange one subsystem without having to discard the other. This means that the information model of metadata capture or the storage engine may be altered at any point in time to reflect consolidation of domain or global standards. The same applies for the data lake storage layer, which is open for addition of further services, i.e. scaling out to (on-site) object storage cluster if needed. Data processing steps are encapsulated within Docker images, which again allows for a high degree of flexibility as data engineers are not forced to implement the required functionality in a given programming language but may choose whatever tooling fits the use case best. The system is open to the integration of any legacy transformation processes, which may be packaged and run as a Docker container. Finally, the orchestration engine itself is interchangeable. The concept is in principle open to be orchestrated by any controlling mechanism that is able to subsequently call RESTful web-services and pass data from one to the next. We chose ActiveWorkflow as our orchestration engine over larger open source projects like Apache Airflow or Luigi mainly because of accessibility and the fact that the plugin system is REST API based and thus again independent of any specific programming language.

Challenges and limitations

The high degree of flexibility described above is a double-edged sword. If the system architecture is so generic in principle, the people developing, operating, and maintaining the system need to have a clear understanding of common goals and tools. A strict set of guidelines must be agreed upon and followed to avoid unnecessary creolization of implementation methods and expansion of complexity. In our case, we decided to consolidate all programming efforts on Python. Legacy ETL jobs designed and implemented with Talend Open Studio for Data Integration have been incorporated to avoid premature refactoring. The data processing workflows themselves follow a consolidated structure, reducing implementation and documentation effort within the data engineering team.

One drawback to the REST-based extensibility of ActiveWorkflow is that there are no advanced methods to secure its REST service endpoints. The ActiveWorkflow Remote Agent protocol does not yet support authentication mechanisms like HTTP Basic Authentication or Token based methods like OAuth. Our custom Remote Agents are implemented with an "API key" parameter that must be specified in the agent configuration in the ActiveWorkflow user interface. Since that parameter is transmitted in plain text as part of the HTTP message body, encrypted TLS communication should always be used between ActiveWorkflow controller and Remote Agent services. This circumstance is not problematic for our setup, as all instances of the services are operated in dedicated network zones shielded against external requests or attacks. Exposing a Remote Agent service to the public internet is not recommended at the moment. Due to our custom Docker Agent connecting to the host server's Docker engine, it must be well protected and workflow admins need to be aware of the security implications of executing arbitrary images.

Lessons learned

During implementation of the framework, a major lesson learned was to embrace actual programming languages when developing ETL jobs, in our case Python. The internal best practice had been Talend Open Studio for Data Integration, a fully graphical IDE that allows users to create and execute no-code (or at least low-code) ETL jobs. While this tool had been successfully used for data integration processes in various research projects, we quickly hit walls when scaling across multiple users, network security zones, and the multitude of pipelines required for the UMG-MeDIC. Source code and the software engineering tool-chain are mature with regard to accessibility, versioning, multi-user interaction, review and refactoring workflows, change management and documentation processes, and are common knowledge among employees and candidates from the software engineering domain. We have experienced a major increase of transparency and productivity of the data engineering team since selecting Python and its data science libraries as primary ETL development tools.

Data warehouse architecture

Considering data warehousing architecture patterns, Armbrust et al. propose a third generation architecture that completely removes a curated relational data warehouse and performs all operation directly on a data lake component [30]. We have explicitly decided to opt for a second generation architecture that combines data lake component for ingested datasets and intermediate artifacts but still emphasizes the curated warehouse as the core component for data integration and semantic enrichment. Where required by technical constraints of the use cases, datasets will be directly pulled from data lake components into analysis processes, e.g. where medical imaging artifacts contain the relevant information. In these use cases, we pull relevant metadata about the available raw data items into our data warehouse and enable stratified search for source data collections based on the integrated medical facts from all clinical department systems indexed at UMG-MeDIC. As a dedicated service unit of a large university hospital, we do simply not face the challenges of scale that lead to the development of a third generation architecture. Benefits of a tightly curated model and expertise about all content indexed at the data warehouse are far more important for our primary goal of guaranteed quality of the information provided to our research partners.

Ongoing processes of further development

We consider our current data processing automation framework and its implementation an ongoing work, because of changing requirements that arise from the ongoing development in the clinical routine. Requirements change over time, new functionality might become necessary with new data sources or data formats we encounter. The modular nature of the framework will be key to constantly extend and refine components. One major area of expansion in the coming years will be metadata cataloging and governance. The current process of metadata capturing is merely the tip of the iceberg. Metadata about medical facts, as well as metadata about source system state, consent and access rights, schema, format, medical vocabulary and mapped ontology terms is already collected in certain places and retrieved by the systems. Here, however, an expansion of the information must already take place at the source system. The list can be almost indefinitely be continued. To document all these types of metadata and to link their semantic meaning back to the actual data items opens up a new load of functional requirements. The user base also extends beyond just the data engineering team and will include domain experts, data stewards, and project managers. While some of these metadata-related challenges are exclusive to medical data, a lot may be learned and repurposed from the existing and growing ecosystem of open source big data metadata management tools like the Egeria Project or Amundsen [31, 32].

Our process metadata schema will most definitely be adjust in the future as the field of data provenance research enters a consolidation phase [10]. Recent open source developments like the OpenLineage specification and its reference implementation, Marquez, do not yet provide harmonized methods for the full depth of process metadata we collect, but could become viable options down the line [33, 34]. Ultimately, process metadata should be publicly available for datasets that are used in published scientific works and the use of standards for provenance documentation will increase re-usability and value of the information [2]. The push for harmonized data provenance frameworks from the big data community outside academia implied by these recent developments might be an indicator that the idea of linked FAIR data objects reaches a stage where industry adoption increase maturity and usability of prototypes and tools from the scientific community [35, 36].

The current monitoring and validation mechanisms build into our framework fulfill the basic requirements but might see in-depth refactoring in later iterations. While both are generic and extendable in nature, adaptation of converging solutions from the open source community would be preferable in the end. The topics observability and metrics are gaining traction and fitting open specifications for information models and services are probably already emerging. Validation of data ingest, data in the warehouse, and data produced for specific research use cases is a necessity for a meaningful use of the UMG-MeDIC core services. Current validation only scratches the surface by proving correct behavior of internal data processing. Full content validation would require methods to compare information stored in the original clinical source systems with the information present in the data warehouse, as demonstrated by Denney et al. [37]. Data quality monitoring, development of domain specific data quality metrics and scores are other highly interesting topics that research from UMG-MeDIC will focus on in the future.

With the medical research data warehouse and the data processing framework fully established, focus for subsequent work shifts towards data publication and analysis. To fully comply with the FAIR guiding principles, the process for automatic registration of persistent identifiers and publication of metadata needs to be implemented. Automatic deployment of research data marts as proposed by Spengler et al. hold great potential to generate value for medical researchers with minimal barriers [38]. A comprehensive metadata management system as described above will enable on-the-fly data mart deployment as well as increasingly automatic deployment of ETL pipelines. Once data formats of input and output systems are modeled precisely in a metadata management system, generating code snippets or fully functional workflow definitions may again enable significant gains in efficiency for the data engineering team.

Conclusions

We analyzed the basic requirements for data processing at a medical research data service unit and proposed a generic architecture framework and prototype implementation that focuses on scalable automation of such tasks. Data extraction from many and heterogeneous sources, including structured and unstructured data, pseudonymization and harmonization, integration and aggregation can be orchestrated completely independent of data and metadata formats. Our implementation works with a custom database schema for initial data harmonization and incorporates Schema.org metadata information model to track data provenance. While highly powerful, extensible, and flexible by design, the prototype implementation is tailored towards operation in a secure internal network environment by privileged users and does not yet incorporate advanced measures to enforce information security in the wild. The framework itself—like any software available today—is not a silver bullet to make research data comply with the FAIR principles but provides much needed ability to process data in a fully automated, traceable, and reproducible manner. If applied comprehensively, the complexity of research data annotation can be largely transferred from the individual researcher into infrastructure. Quality and speed of the data acquisition process that drives scientific insight can be increased.

Abbreviations

ACID	Atomic, consistent, isolated, durable
CDSTAR	Common Data Storage Architecture
CSV	Comma separated values
DRG	Diagnoses Related Groups
DWH	Data Warehouse
EDCat	The Emergency Department Catalog
EMIF	European Medical Information Framework
ENCODE DNase	Encyclopedia of DNA Elements
EHR	Electronic Health Record
ETL	Extract Transform Load
FAIR	Findable, Accessible, Interoperable, Reusable
HL7	Health Level 7
HL7 FHIR	Fast Healthcare Interoperability Resources
HTTP	Hypertext Transfer Protocol
IDE	Integrated development environment
IRIs	Internationalized Resource Identifiers
LD	Linked data
LOINC	Logical Observation Identifiers Names and Codes
ORU	HL7 Observation Result
REST API	Representational State Transfer Application Programming Interface
SQL	Structured Query Language
TLS	Transport Layer Security
UMG-MeDIC	University Medical Center Göttingen (UMG) Medical Data Integration Center
URL	Uniform Resource Locator

Acknowledgements

We would like to thank the MeDIC staff for supporting the further development and organization of the MeDIC as a service facility for researchers and interested parties.

Authors' contributions

TK leads the development and expansion of the MeDIC and is responsible, among other things, for the structure and selection of the components used. MS and MP were involved in the construction as members of the technical team and created the first draft of the manuscript. CS and BL are also members of the technical team and extended the manuscript. CB is responsible for the selection and application of metadata in the MeDIC and edited the manuscript in this regard. DK performed the systematic literature search, coordinated the authors and supervised the scientific writing. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. The study was funded by the German Federal Ministry of Education and Research within the Medical Informatics Initiative (Grant ID 01ZZ1802B; HiGHmed). The funder had no involvement in the design of the study, the collection, analysis or interpretation of data, or the writing of the manuscript.

Availability of data and materials

All presented software components are freely available on the Internet.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Medical Informatics, University Medical Center Göttingen, Von-Siebold-Straße 3, 37075 Göttingen, Germany. [2]University MS Center, Biomedical Research Institute (BIOMED), Hasselt University, Agoralaan Building C, 3590 Diepenbeek, Belgium. [3]Data Science Institute (DSI), Hasselt University, Agoralaan Building D, 3590 Diepenbeek, Belgium. [4]NextLytics AG, Kapellenstrasse 37, 65719 Hofheim Am Taunus, Germany.

Received: 10 August 2022 Accepted: 9 May 2023
Published online: 15 May 2023

References

1. Martin-Sanchez FJ, Aguiar-Pulido V, Lopez-Campos GH, Peek N, Sacchi L. Secondary Use and Analysis of Big Data Collected for Patient Care. Yearb Med Inform. 2017;26(1):28–37.
2. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. Comment: The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:1–9.
3. Cao Y, Jones C, Cuevas-Vicenttín V, Jones MB, Ludäscher B, McPhillips T, et al. DataONE: A Data Federation with Provenance Support. In: Mattoso M, Glavic B, editors., et al., Provenance and Annotation of Data and Processes IPAW 2016 Lecture Notes in Computer Science. Springer Cham; 2016. p. 230–4.
4. Ohno-Machado L, Sansone SA, Alter G, Fore I, Grethe J, Xu H, et al. Finding useful data across multiple biomedical data repositories using DataMed. Nat Genet. 2017;49(6):816–9.
5. Holub P, Kohlmayer F, Prasser F, Mayrhofer MT, Schlünder I, Martin GM, et al. Enhancing Reuse of Data and Biological Material in Medical Research: From FAIR to FAIR-Health. Biopreserv Biobank. 2018;16(2):97–105.
6. Knaup P, Deserno T, Prokosch H-U, Sax U. Implementation of a National Framework to Promote Health Data Sharing. Yearb Med Inform. 2018;27(01):302–4.
7. Haarbrandt B, Schreiweis B, Rey S, Sax U, Scheithauer S, Rienhoff O, et al. HiGHmed - An Open Platform Approach to Enhance Care and Research across Institutional Boundaries. Methods Inf Med. 2018;57(Open 1):66–81.
8. Terrizzano I, Schwarz P, Roth M, Colino JE. Data wrangling: The challenging journey from the wild to the lake. In: CIDR 2015 - 7th Biennial Conference on Innovative Data Systems Research. 2015.
9. Aghajani E, Nagy C, Vega-Marquez OL, Linares-Vasquez M, Moreno L, Bavota G, et al. Software Documentation Issues Unveiled. Proc - Int Conf Softw Eng. 2019;2019:1199–210.
10. Parciak M, Bauer C, Bender T, Lodahl R, Schreiweis B, Tute E, et al. Provenance solutions for medical research in heterogeneous IT-infrastructure: An implementation roadmap. Stud Health Technol Inform. 2019;264:298–302.
11. Bauer CR, Umbach N, Baum B, Buckow K, Franke T, Grütz R, et al. Architecture of a biomedical informatics research data management pipeline. Stud Health Technol Inform. 2017;228:262–6.
12. Sinaci AA, Núñez-Benjumea FJ, Gencturk M, Jauer ML, Deserno T, Chronaki C, et al. From Raw Data to FAIR Data: The FAIRification Workflow for Health Research. Methods Inf Med. 2020;59(6):E21-32.
13. Löbe M, Matthies F, Stäubert S, Meineke FA, Winter A. Problems in fairifying medical datasets. Stud Health Technol Inform. 2020;270:392–6.
14. Bhatia K, Tanch J, Chen ES, Sarkar IN. Applying FAIR Principles to Improve Data Searchability of Emergency Department Datasets: A Case Study for HCUP-SEDD. Methods Inf Med. 2020;59(1):48–56.
15. Bönisch C, Sargeant A, Wulff A, Parciak M, Bauer CR, Sax U. FAIRness of openEHR archetypes and templates. CEUR Workshop Proc. 2019;2849:102–11.
16. Pereira A, Lopes RP, Oliveira JL. SCALEUS-FD: A FAIR Data Tool for Biomedical Applications. Biomed Res Int. 2020;2020.
17. Zondergeld JJ, Scholten RHH, Vreede BMI, Hessels RS, Pijl AG, Buizer-Voskamp JE, et al. FAIR, safe and high-quality data: The data infrastructure and accessibility of the YOUth cohort study. Dev Cogn Neurosci. 2020;45(August):100834.
18. U.S. Food and Drug Administration FDA. Real-World Evidence. 2022. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence. Accessed 02 Aug 2022.
19. imi innovative medicines initiative. EMIF European Medical Information Framework. 2018. https://www.imi.europa.eu/projects-results/project-factsheets/emif. Accessed 02 Aug 2022.

20. Trifan A, Oliveira JL. A FAIR Marketplace for Biomedical Data Custodians and Clinical Researchers. Proc - IEEE Symp Comput Med Syst. 2018;2018:188–93.
21. World Wide Web Consortium W3C. JSON-LD 1.1 A JSON-based Serialization for Linked Data. 2020. https://www.w3.org/TR/json-ld/. Accessed 02 Aug 2022.
22. Internet Engineering Task Force (IETF). The JavaScript Object Notation (JSON) Data Interchange Format. 2017. https://www.rfc-editor.org/rfc/rfc8259. Accessed 02 Aug 2022.
23. Duerst M, Suignard M. Internationalized Resource Identifiers (IRIs). 2005. https://www.rfc-editor.org/rfc/rfc3987.html. Accessed 02 Aug 2022.
24. Schema.org. Welcome to Schema.org. 2021. https://schema.org/. Accessed 02 Aug 2022.
25. Apache CouchDB. CouchDB relax. Seamless multi-master sync, that scales from Big Data to Mobile, with an Intuitive HTTP/JSON API and designed for Reliability. 2021. https://couchdb.apache.org/. Accessed 02 Aug 2022.
26. Automatic Mode Labs. ActiveWorkflow. Turn complex requirements to workflows without leaving the comfort of your technology stack. 2021. https://www.activeworkflow.org/. Accessed 02 Aug 2022.
27. docker. Developers Love Docker. Businesses Trust It. Build safer, share wider, run faster: New updates to our product subscriptions. 2021. https://www.docker.com/. Accessed 02 Aug 2022.
28. Celery.org. Celery - Distributed Task Queue. 2021. https://docs.celeryproject.org/en/stable/. Accessed 02 Aug 2022.
29. Schmitt O, Siemon A, Schwardmann U, Hellkamp M. GWDG Object Storage and Search Solution for Research – Common Data Storage Architecture (CDSTAR). Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), editor. Göttingen: Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen. 2014.
30. Armbrust M, Ghodsi A, Xin R, Zaharia M, Berkeley U. Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. 11th Annual Conference on Innovative Data Systems Research (CIDR '21). 2021. https://www.cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf. Accessed 12 May 2023.
31. Amundsen. Overview Amundsen. 2022. https://www.amundsen.io/amundsen/. Accessed 19 Jan 2023.
32. EGERIA. Open metadata and governance for enterprises - automatically capturing, managing and exchanging metadata between tools and platforms, no matter the vendor. 2022. https://egeria-project.org/. Accessed 19 Jan 2023.
33. OpenLineage. OpenLineage. An open framework for data lineage collection and analysis. 2022. https://openlineage.io/. Accessed 02 Aug 2022.
34. LF AI & Data Foundation. Marquez. Collect, aggregate, and visualize a data ecosystem's metadata. 2022. https://marquezproject.github.io/marquez/. Accessed 02 Aug 2022.
35. Wittenburg P. Common Patterns in Revolutionary Infrastructures and Data. 2018. p. 1–13. Available from: https://b2share.eudat.eu/records/4e8ac36c0dd343da81fd9e83e72805a0
36. van Vlijmen H, Mons A, Waalkens A, Franke W, Baak A, Ruiter G, et al. The need of industry to go fair. Data Intell. 2020;2(1–2):276–84.
37. Denney MJ, Long DM, Armistead MG, Anderson JL, Conway BN. Validating the extract, transform, load process used to populate a large clinical research database. Int J Med Inform. 2016;94:271–4.
38. Spengler H, Lang C, Mahapatra T, Gatz I, Kuhn KA, Prasser F. Enabling agile clinical and translational data warehousing: Platform development and evaluation. JMIR Med Informatics. 2020;8(7):1–18.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

FAIR+R: Making Clinical Data Reliable through Qualitative Metadata

Caroline Bönisch[a,1] and Dorothea Kesztyüs[a] and Tibor Kesztyüs[a]
[a]*Medical Data Integration Center, Department of Medical Informatics, University Medical Center Göttingen, Robert-Koch-Str. 40, 37075 Göttingen, Germany*
ORCiD ID: Caroline Bönisch https://orcid.org/0000-0001-7169-6090
Dorothea Kesztyüs https://orcid.org/0000-0002-2166-846X
Tibor Kesztyüs https://orcid.org/0000-0003-0813-2393

Abstract. Metadata are often the first access to data repositories for researchers within secondary use. Through automatic metadata generation and metadata harvesting the amount of data about data has been growing ever since. In order to make data not only FAIR but also reliable, the aspect of metadata quality has to be considered. But as earlier assessments of metadata of different repositories showed, metadata quality still lacks behind its capability. Providing an extensive literature review the authors conclude nine measures to assess metadata in relation to clinical care repositories, such as Medical Data Integration Centers (MeDICs). Proceeding from these measures the authors propose an addition of the FAIR Guiding Principles by adding a fifth block for Reliability including three principles, that resulted from the measures presented. The results form the basis for the future work of an assessment of metadata, that is stored in a MeDIC.

Keywords. FAIR, metadata, data quality, reliable data

1. Introduction

Since the FAIR Principles were introduced in 2016 [1] the commitment to make data FAIR has increased in different scientific fields [2]. The FAIR principle advise data stewardship and make data Findable, Accessible, Interoperable and Reusable (FAIR), making it particularly applicable in the health area. Since the introduction of the Principles, several initiatives and work groups have formed in order to apply the FAIR Principles in the medical research area. This is necessary because medical data would not be reused for research, although it already exists but is not accessible or findable [3]. However, the FAIR Principles include not only data but also corresponding metadata. Considering the vastly growing collection of data in the field of clinical care, and the establishment of so-called Medical Data Integration Center (MeDIC) at different university hospitals in Germany, the data should not only be findable, accessible, interoperable, and reusable, but also reliable. Only reliable data can form the scientific base of data analysis and provide a potential to validate the results originating from these analysis.

It means in effect that the quality of the data has to be captured and assessed on a scientific premise. The stored data also has to be protected from unplanned changes,

[1] Corresponding Author: Caroline Bönisch, caroline.boenisch@med.uni-goettingen.de.

being they organizational, structural or content-related. If the data has to be changed, all adjustments must be transparent and stored along the data, comparable to audit trails for electronic medical records [4].

Metadata as accompanying information to the specific data inherit major aspects in providing reliability. In some repositories, data can only be accessed via their metadata and this information is a starting point in secondary use to give researchers a first impression of the data. By adding further details about the quality of the data and metadata being of good quality themselves, the reliability of the information is secured [5].

Metadata consist of intrinsic metadata, for example version number, title, authors, date of creation, and provenance metadata, meaning information about access rights concerning the data or organization hosting the data [6]. Previous literature shows that the quality of data is not fully available within metadata of clinical data [7]. While Ochoa et al. [8] show an overview of different metrics for metadata quality in repositories, they also provide insights of the difference between manual quality evaluation and simple statistical quality measurements and conclude that the quality of metadata should be measured automatically. However, the metrics are short of in the field of multimedia metadata [8].

2. Methods

In the first step, a literature search was executed in order to review already existing evaluation schemes and methods for (meta)data quality.

2.1. Literature Search

Embase via Ovid and PubMed were searched using appropriate search steps and keywords. Table 1 shows the search steps exemplary for the PubMed database.

Table 1. Search Steps of the literature review including count of results for each step in PubMed database.

Number	Search Step	Results
#1	"data accuracy"[MeSH Terms]	3,786
#2	"metadata"[MeSH Terms]	507
#3	"data curation"[MeSH Terms]	816
#4	"quality improvement"[MeSH Terms]	32,640
#5	#1 OR #2 OR #3 OR #4	37,478
#6	("quality"[Title/Abstract] OR "reliable"[Title/Abstract]) AND "metadata"[All Fields]	990
#7	#5 AND #6	136

The search process followed a deductive top-down approach, initially using very abstract search terms, which were then further refined. The PRISMA statement ("Preferred Reporting Items for Systematic Reviews and Meta-Analyses") [9] was used as a conceptual guide of the literature review and the foci of the literature search results were analyzed regarding the following criteria:

- data collected and stored within the clinical environment
- methods or evaluation schemes for estimating data quality
- analysis of (meta)data quality factors

- earlier approaches to make data reliable in other disciplines

Based on the criteria and including the results of the literature search (331 results), title/abstract screening was performed first. The screening resulted in the exclusion of 279 articles and the remainder of 52 articles were then studied completely and finally three were omitted within the full-text review. The publications included were reviewed with a focus on the research criteria of this paper.

2.2. Summarizing Metadata Quality Factors

The results of the literature review revealed different approaches to assess metadata or data quality. As Stausberg et al. [7] stated, the FAIR Principles lack the dimension of quality in (meta)data, therefore the authors additionally examined quality factors for data and adopted them, where possible, to the metadata domain, to provide a complete collection of metadata quality factors for assessment.

Inferred from the metadata quality factors found [10] and presented, the FAIR Principles are then extended with a block for reliable data (RL) and outlined with three principles for this block.

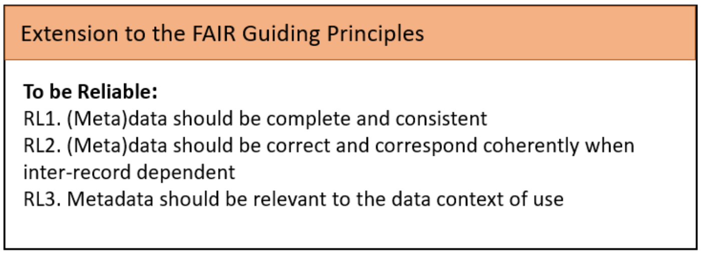
3. Results

Based on the results of the literature review and the related work found, the authors propose the quality measures listed in Table 2. The measures are consolidated from various research manuscripts in different scientific fields and selected concerning the requirements of the data complexity within clinical care.

Table 2. Assessment metrics for metadata in clinical care, based on the results of the literature review

Measure	Description
Completeness	All mandatory data fields are filled with information
Consistency	Metadata should be conform to existing standards and formats
Correctness	The information describes the metadata in an accurate and distinct way
Correspondence	Metadata that is linked or inter-dependent represents the same information through every instance
Relevance	The metadata corresponds to the requirement/expectations of the user
Semantic Specificity	Average specificity of a semantic concept in metadata information
Timeliness	Currency of the metadata information describing a resource information
Accessibility	The information of the metadata must be physically available and understandable either by human or machine
Reproducibility	Metadata quality scores should be reproducible and not lack clarity in terminology

In conclusion, the authors propose an extension of the FAIR Principles by adding the Principle of Reliability. Figure 1 depicts the three principles, which are added for (meta)data to be Reliable.



Extension to the FAIR Guiding Principles

To be Reliable:
RL1. (Meta)data should be complete and consistent
RL2. (Meta)data should be correct and correspond coherently when inter-record dependent
RL3. Metadata should be relevant to the data context of use

Figure 1. Proposed principle block of Reliable (meta)data

Using these proposed measures the respective metadata quality can then be calculated. The results of the calculations can subsequently be grouped using the categories *Reliable, Reliable with restriction* and *Not reliable,* to provide a straightforward classification of outcomes of a future automatic quality assessment.

4. Discussion

As seen by the literature review, the topic of metadata quality and transferability of qualitative metadata being a key component to reliable data, lacks further research. A highly accepted definition of metadata quality is still somewhat missing as of today, due to the fact that there exists some data quality metrics definition in several scientific areas like bibliography, but little literature results [7], [11] for metadata quality in clinical research could be obtained. Therefore, the literature search had to be extended, to access metrics within data quality, in hopes to apply them to metadata.

Most research regarding metadata, expressed metadata regarding completeness of the data as important quality factor. Nevertheless, it is not the only important quality factor, although maybe the easiest to be measured. Other factors like relevance, consistency and timeliness are also target components of qualitative metadata.

The metadata is on the one hand machine-generated, like date or version, but on the other hand humanly entered via different clinical professionals. Additionally, reusing human-generated data without questioning, when the creator is an expert in the field of research but not an expert in metadata creation, results in discrepancy, as stated by Masor [3].

It should be emphasized that, research of the value and quality of metadata still lags behind the metadata's possibilities. Masor showed that metadata are not being used to their full potential [3]. This is particularly concerning because they are often the first entry point for researchers who want to reuse data from a repository, such as Medical Data Integration Centers (MeDICs) .

As of today, metadata quality assurance is still seen as more of a casualty, and research on this topic is limited. However, as repositories grow, quality issues in metadata gain more visibility and influence the usage of repositories of clinical data.

5. Conclusion

The present article aims to provide an overlook of the existing literature of metadata and data quality concerning clinical care repositories and data integration centers. As the literature on this topic is sparse, further scientific areas were included in order to obtain a complete overview of quality factors of metadata, which are crucial for reliable data.

The quality measures include Completeness, Consistency, Correctness, Correspondence, Relevance, Semantic Specificity, Timeliness, Accessibility, and Reproducibility. Results of these measures can then be classified in reliable, reliable with restriction and not reliable categories, to aid researcher in judging metadata quality.

Based on this aggregation of factors the authors propose an extension of the FAIR Guiding Principles by adding the block of Reliability with additional principles in correspondence to the identified factors.

In accordance with this research, future work will include the automatisation of the metadata quality assessment. This assessment should then be performed on the clinical data collected within the MeDIC of the University Medical Center Göttingen (UMG) and give an overview on the metadata quality of this data collection.

6. Acknowledgements

Funded by the German Federal Ministry of Education and Research (FAIRRMeDIC, grant number 01ZZ2012).

References

[1] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. (2016);3:160018. doi: 10.1038/sdata.2016.18.

[2] Stall S, Robinson E, Wyborn L, Yarmey L R, Parsons M A, Lehnert K, Cutcher-Gershenfeld J, Nosek B, and Hanson B. Enabling FAIR data across the Earth and space sciences, *Eos, 98. (*2017), https://doi.org/10.1029/2017EO088425

[3] M. Dugas, K.H. Jöckel, T. Friede, O. Gefeller, M. Kieser, M. Marschollek, E. Ammenwerth, R. Röhrig, P. Knaup-Gregori, H.U. Prokosch, Memorandum "Open Metadata" - Open Access to Documentation Forms and Item Catalogs in Healthcare, 2015

[4] Masor, J. Electronic Medical Records and E-Discovery: With New Technology Come New Challenges, 5 Hastings Sci. & Tech. L.J. 245. (2013)

[5] Bruland P, Doods J, Storck M, Dugas M. What Information Does Your EHR Contain? Automatic Generation of a Clinical Metadata Warehouse (CMDW) to Support Identification and Data Access Within Distributed Clinical Research Networks. Stud Health Technol Inform. (2017); 245:313-317.

[6] Canham S, Ohmann C. A metadata schema for data objects in clinical research. *Trials* 17, 557 (2016). https://doi.org/10.1186/s13063-016-1686-5

[7] Stausberg J, Harkener S, Jenetzky E, Jersch P, Martin D, Rupp R, Schönthaler M. FAIR and Quality Assured Data - The Use Case of Trueness. Stud Health Technol Inform. (2022);289:25-28. doi: 10.3233/SHTI210850. PMID: 35062083

[8] Ochoa X, Duval E. Automatic evaluation of metadata quality in digital repositories. *Int J Digit Libr* **10**, 67–91(2009).https://doi.org/10.1007/s00799-009-0054-4

[9] Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. (2009);6(7):e1000097. doi: 10.1371/journal.pmed.1000097

[10] Bruce R, Hillmann D. The continuum of metadata quality: Defining, Expressing, Exploiting. National Science Digital Library(2004) Metadata in Practice, ALA Editions

[11] Shang N, Weng C, Hripcsak G. A conceptual framework for evaluating data suitability for observational studies. J Am Med Inform Assoc. 2018 Mar 1;25(3):248-258. doi: 10.1093/jamia/ocx095.

Address for Correspondence
Caroline Bönisch, M.Sc. (caroline.boenisch@med.uni-goettingen.de)
Medical Data Integration Center, Dep. of Medical Informatics, University Medical Center Göttingen
Von-Siebold-Straße 3, 37075 Göttingen, Germany

REFERENCES

Benbasat I, Goldstein DK, Mead M (1987): The case research strategy in studies of information systems. MIS quarterly, 369-386

Bonomi L, Jiang X (2018): Patient ranking with temporally annotated data. Journal of biomedical informatics 78, 43–53

Dubovitskaya A, Xu Z, Ryu S, Schumacher M, Wang F (2018): Secure and Trustable Electronic Medical Records Sharing using Blockchain. AMIA Annual Symposium proceedings. AMIA Symposium, 2017, 650–659

Ecarot T, Fraikin B, Lavoie L, McGilchrist M, Ethier JF (2021): A Sensitive Data Access Model in Support of Learning Health Systems. Computers 10(3):25

Ehsani-Moghaddam B, Martin K, Queenan J A (2021): Data quality in healthcare: A report of practical experience with the Canadian Primary Care Sentinel Surveillance Network data. Health information management: journal of the Health Information Management Association of Australia 50(1-2), 88–92

Gliklich R E, Dreyer NA, Leavy MB (2010): Registries for evaluating patient outcomes. A User's Guide 5, 19-29

Meloni V, Sulis A, Mascia C, Frexia F (2021): FAIRifying Clinical Studies Metadata: A Registry for the Biomedical Research. Stud Health Technol Inform.281,779-783

Parciak M, Suhr M, Schmidt C, Bönisch C, Löhnhardt B, Kesztyüs D, Kesztyüs T (2023): FAIRness through automation: development of an automated medical data integration infrastructure for FAIR health data in a maximum care university hospital. BMC Med Inform Decis Mak 23, 94

Stausberg J, Harkener S, Jenetzky E, Jersch P, Martin D, Rupp R, Schönthaler M (2022): FAIR and Quality Assured Data - The Use Case of Trueness. Stud Health Technol Inform289,25-28

Queralt-Rosinach N, Kaliyaperumal R, Bernabé CH, Long Q, Joosten S A, van der Wijk HJ, Flikkenschild ELA, Burger K, Jacobsen A, Mons B, Roos M (2022): Applying the FAIR principles to data in a hospital: challenges and opportunities in a pandemic. Journal of biomedical semantics 13(1), 12

Sinaci AA, Núñez-Benjumea F, Gencturk M, Jauer M, Deserno T, Chronaki C, Cangioli G, C B, Rodriguez JM, Perez M, et al. (2020): From Raw Data to FAIR Data: The FAIRification Workflow for Health Research. Methods of Information in Medicine. 59. e21-e32

Weiskopf NG, Weng C (2013): Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association: JAMIA 20(1), 144–151

Wilkinson M, Dumontier M, Aalbersberg I. et al. (2016): The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018

Yin RK: Case Study Research, Design and Methods, Sage Publications, Beverly Hills, California 1984

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my deepest gratitude and appreciation to all those who have been instrumental in the completion of this dissertation thesis.

First and foremost, I am immensely grateful to my supervisors Prof. Dr. med. Tibor Kesztyüs, Prof. Dr. med. Jürgen Brockmöller and Prof. Dr. med. Leif Saager: Prof. Dr. med. Tibor Kesztyüs, for his unwavering guidance, expertise, and encouragement throughout this research journey. His invaluable insights and constructive feedback have been key in shaping the direction of this work. Prof. Dr. med. Jürgen Brockmöller I would like to thank for his constructive feedback and his critical questions, that helped form the basis of the research work and his insights in medical data and Prof. Dr. med. Leif Saager for his ongoing support of this research work and the helpful remarks.

I extend my heartfelt thanks to the faculty members at the Medical Data Integration Center of the University Medical Center Göttingen namely PD Dr. Dorothea Kesztyüs, Christian Schmidt M.Sc. and Sarah Schnabel B.Sc. for providing a stimulating and intellectually enriching environment.

I would also like to acknowledge the ongoing support and encouragement of my husband, who stood by me during moments of doubt and celebrated all my achievements and my son, who showed me, why it is all worth it.

Lastly, I am grateful to all the researchers, authors, and scholars whose work and ideas have been an inspiration and foundation for this thesis. Their contributions to the field of medical informatics have been invaluable.

CURRICULUM VITAE

Caroline Marieken Bönisch, geborene Thoms, wurde am 22.10.1989 in Bergen auf Rügen geboren.

Nach dem Abitur, im Jahr 2008 an der Europaschule Rövershagen, studierte sie an der Hochschule Stralsund Medizininformatik und Biomedizintechnik im Bachelorstudium. Nach erfolgreichem Abschluss des Bachelorstudiums, im Jahr 2012, begann sie den konsekutiven Master Medizininformatik an der Hochschule Stralsund. Das Masterstudium schloss sie 2015, als Master of Science ab und begann anschließend am aQua-Institut in Göttingen, als Softwareentwicklerin zu arbeiten. 2016 wechselte sie an das Institut für Medizinische Informatik unter Prof. Otto Rienhoff und arbeitete als wissenschaftliche Mitarbeiterin an dem Verbundprojekt HiGHmed, der Medizininformatik-Initiative, gefördert durch das Ministerium für Bildung und Forschung, mit. Seit dem Sommersemester 2022 ist sie Studentin des Promotionsprogramms Humanwissenschaften in der Medizin der Georg-August-Universität Göttingen. Ihre Doktorarbeit erstellte sie, als wissenschaftliche Mitarbeiterin, ab 2022 innerhalb der Nachwuchsgruppe des UMG-MeDIC.

Seit 2018 ist Caroline Bönisch Mitglied der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS e.V.). Sie ist seit 2018 verheiratet und seit 2020 Mutter eines Sohnes.