Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230247

Concept for a Basic ISO 14721 Archive Information Package for Clinical Studies

Matthias LÖBE[a,1], Alessandra KUNTZ[b], Christian HENKE[b], Frank MEINKE[a], Ulrich SAX[b,c] and Alfred WINTER[a]
[a] *Institute for Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig, Germany*
[b] *Department of Medical Informatics at the University Medical Center Göttingen, Germany*
[c] *Campus-Institute of Data Science (CIDAS), Göttingen, Germany*
ORCiD ID: Matthias Löbe 0000-0002-2344-0426, Alessandra Kuntz 0000-0002-8259-2577, Christian Henke 0000-0002-4541-4018, Frank Meineke 0000-0002-9256-7543, Ulrich Sax 0000-0002-8188-3495

Abstract. Secondary use of medical data for research is desirable for intrinsic, ethical and financial reasons. In this context, the question becomes relevant as to how such datasets are to be made accessible to a larger target group in the long term. Typically, datasets are not extracted ad hoc from the primary systems, because they are processed qualitatively (FAIR data). Special data repositories are currently being built for this purpose. This paper examines the requirements for the reuse of clinical trial data in a data repository utilizing the Open Archiving Information System (OAIS) reference model. In particular, a concept for an Archive Information Package (AIP) is developed with the central focus on a cost-effective trade-off between the effort of creation for the data producer and the comprehensibility of the data for the data consumer.

Keywords. Data sharing, Research Data Management, OAIS, ISO 14721, Clinical trials, Archiving, Data repositories, Data reuse, FAIR

1. Introduction

Data sharing, the reuse of data once collected for further research purposes, has moved from an altruistic attitude of individual researchers to an obligation of many funders or scientific publishers in recent years. Data sharing makes sense for scientific, ethical, and resource reasons. Nevertheless, it is not widely practiced because, first, it involves a significant effort for the data provider that is currently not compensated, second, there is a lack of data repositories for permanent storage and legally secure access (particularly for sensitive personal health data), and third, there are no widely accepted ideas about which data structures are essential for data recipients in complex research projects.

In the field of medical research, local research data centers are currently being set up in Germany that will have the technical (hardware, storage space, network

[1] Corresponding Author: Matthias Löbe, Institut für Medizinische Informatik, Statistik und Epidemiologie, Universität Leipzig, Härtelstraße 16-18, 04107 Leipzig, Germany; E-mail: matthias.loebe@imise.uni-leipzig.de.

connection), content-related (data body, metadata), legal (framework agreement, usage regulations, consent, pseudonymization) and organizational (data access board, trust office, ethics committee) prerequisites. In projects such as the Medical Informatics Initiative (MII) [1] or the National Research Data Infrastructure for Personal Health Data (NFDI4Health) [2], staff positions have been created for data management. Thus, (semi-)permanent structures will exist for the storage and retrieval of health data.

On the other hand, these data repositories differ from institutionally funded libraries with a statutory mandate. They have less financial resources and are not primarily intended to archive health data in their entirety for all time, but rather to provide target group-oriented data packages that can be reused for secondary research projects. A number of special requirements arise for the new research data centers:

1. Managing data in a repository creates a third party of data stewards between data producers and data consumers. Data stewards have generally not been involved in the data collection process and cannot answer queries from potential consumers themselves. Therefore, standards must be created for data dissemination packages that make the data they contain *discoverable* and *reusable* by target audiences.
2. In many cases, there will be a time gap between when the data is delivered to the repository and when consumers first request it. It is possible that the data-producing projects have already been completed and no permanent contacts are available anymore. Therefore, standards need to be created for data dissemination packages that make the data *interpretable* and include descriptive, semantic, and provenance metadata.
3. Few current research projects have a financial line item for longer-term data sharing. The agreed standards must therefore also be guided by what is feasible for data producers when handing over the data and for data managers in routine operations, and what risks exist in terms of data protection and data security. Therefore, standards need to be created for cost-effective data dissemination packages that can be assembled by artifacts already existing in routine operation, automatic conversation and limited manual enrichment.

2. Methods

The Open Archival Information System (OAIS) [3] is a reference model for a digital archive, which has also been published as ISO standard 14721 (2012). It describes the components of an archival system for long-term preservation. From a data perspective, OAIS is based on *information packages* that contain descriptive information in addition to the actual content to be preserved. The most important type of information package is the *Archive Information Package* (AIP) (see Figure 1), which represents the underlying archived artifact. In addition to the actual data object, it contains detailed metadata on the structure and interpretation of the content (*Representation Information*) as well as on identifications, context, provenance and access options (*Preservation Description Information*). For the findability of the AIP, additional descriptive information is provided [4].

An expert-based approach was chosen to create a concept for an AIP specific to clinical trials. Unlike traditional archives or libraries, medical research data centers are usually sub-organizational units of IT departments that also perform other tasks such as data extraction from primary systems or operating Clinical Data Management Systems for data collection in clinical trials. For this reason, they are familiar with both the data-generating processes as well as the needs of the target group and can define the scope of an AIP in collaboration with data managers and investigators. Especially in the environment of regulated research, formal standard operating procedures (SOP) and working instructions exist for all substeps of a clinical trial. The documents and data files described there have been evaluated to determine a) the extent to which they are necessary or useful for understanding the clinical data and the research project, b) whether their content contains sensitive information in terms of data protection or intellectual property, and c) what effort would be involved in generating or transforming them to suit the target group. The last point aims at the fact that documents and data for controlled internal use have to meet lower requirements than externally published assets. Enrichment is often described under the keyword FAIRification Workflow [5] and can be a complex procedure. As previously mentioned, the focus of the work was on generating immediately understandable and cost-effective packages. High demands and mandatory standards of the regulatory authorities (CDISC, FDA, EMA) exist for the execution and analysis of drug approval studies. These were not the aim of the work.

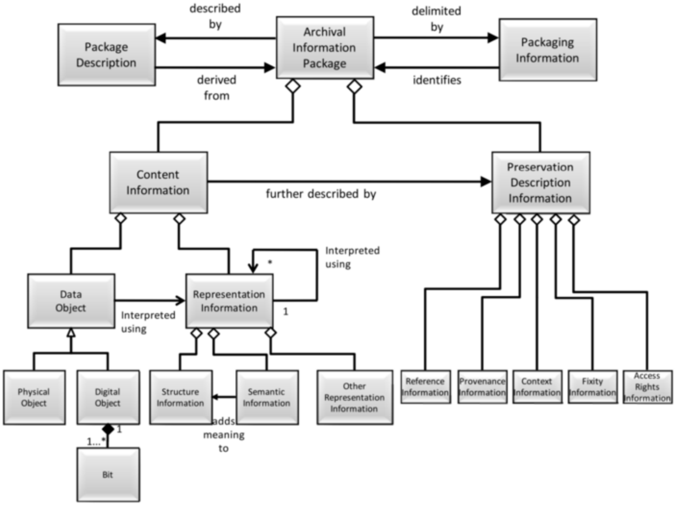

Figure 1. Schematic overview of the components of an archive information package [3]

The second step involved mapping the "metadata", i.e. all the logical classes referred to as "information" in Figure 1. This involved mapping the existing artifacts, that can be constructed from existing documents and files to the various types of information classes. Other relevant metadata may exist and will be added in later extensions of the concept.

3. Results

Based on existing experience and internal SOPs, we selected 22 types of clinical documents that were assessed as "possibly relevant". The assumed relevance resulted from demands from previous data transfers. The effort required to make them available as a published asset was estimated to be low in the majority of cases, since they already exist. However, it is questionable whether the information contained is sufficient for further processing (example Data Sharing Plan: intention to share "yes" is too unspecific). Internal processes, names of natural persons within documents, preliminary study results were considered to be significantly more problematic for publication. Even if accessibility can be restricted by defining rights and user roles, only low-problem resources should be published first. The remaining document types were examined for their expected usefulness for third party interpretation of the study. Ultimately, the following document types should be captured in an AIP (if available): a) data dictionary and/or (annotated) case report forms, b) schedule of visits and assessments, c) data quality rules, d) algorithms or scripts for transformation or evaluation, e) records in clinical registries, f) research data management plan, g) scientific publications and supplements, h) datasets.

The actual dataset (the OAIS Digital Object) is a special case. Although there cannot be a low risk here due to its nature, since it contains sensitive personal health data, its availability is the purpose of the project. The second step involved mapping the "metadata", i.e. all the logical classes referred to as "information" in Figure 1. This involved mapping the existing artifacts contained in the document (see Table 1). However, in practice, metadata is distributed in different source assets, which requires transformations from the Submission Information Package (SIP) to the AIP. Some metadata is present multiple times and in different granularity, some not at all. An example of this is an export of data in CDISC-ODM format with rich metadata compared to a CSV export without metadata.

Table 1. Exemplary mapping of metadata elements corresponding to types of information classes. Container *Content Information, Representation Information* and *Preservation Description Information* omitted.

Information classes (AIP)	**Metadata elements relevant for clinical trials**
Structure Information	• Datatype, format, precision of variables in the dataset
Semantic Information	• Representation of the logical structure of the dataset, e.g.forms that data elements belonged to or events there were collected • Annotations with concepts from medical terminologies like CDISC CDASH/SDTM, LOINC or SNOMED CT
Other Information	• Standards the data object depends on: XML, CDISC ODM, UFT8
Reference Information	• Authors, name, title, version, … • Identifiers like DOIs or internal LHA-IDs
Provenance Information	• Source systems of the dataset and others from the set of 19 data elements as defined in [6] • Audit trail (change log to dataset)
Context Information	• Funding, related projects • References to other parts for composite datasets • Documents helping to understand the experiment like publications or SOPs
Fixity Information	• Electronic signatures from investigators (from CDISC ODM) • Certificates or cryptographic keys for privacy-preserving distributed analysis (DataSHIELD in planning) or pseudonymization algorithms of patient-identifying data like names or addresses
Access Rights Information	• Rights and roles • Licence information • Intellectual property contracts such as publication moratoria or data use agreements
Packaging Information	• File structure in ISA/SEEK
Package Description	• Study acronym, conditions observed, sample size and 8 others from the NFDI4Health metadata schema [7]

An obvious challenge is metadata that is not included in the SIP but is important for data sharing, such as licenses or data usage agreements.

4. Discussion and Conclusions

Many projects constantly generate data that they offer in repositories for subsequent use. Once established, the question of optimized storage of data arises to support subsequent research questions and ensure reproducibility. The concept presented here is just currently being implemented in the Leipzig Health Atlas (LHA) [8]. One limitation is that the concept has not yet been evaluated by external partners and projects. Furthermore, no harmonized vocabularies exist for parts of the metadata, e.g., on informed consents. It will be expectedly challenging to get high quality metadata from data providers who need to do manual research on it, e.g., providence information. In the future, the LHA will serve as a template for "Local Data Hubs (LDH)" of NFDI4Health [9], which requires standards-compliant, machine-readable vocabularies and interfaces [10]. From a clinical perspective, it will be important to set the bar low for discoverability and access.

The OAIS standard is widely used and is currently being further elaborated [11]. OAIS is the dominant standard due to its age and influence, most alternatives are conceptually based and address additional facets such as trustworthiness (ISO 16363) and planning (Digital Curation Centre Curation Lifecycle Model) [12]. For clinical trials, there is an EMA "guideline on the content, management and archiving of the clinical trial master file", which refers solely to archiving and covers organizational and financial issues.

Acknowledgements: Research was supported by DFG grant WI 1605/10-2, SA 1009/3-2 and BMBF grant 031L0026.

References

[1] Semler SC, Wissing F, Heyder R (2018) German Medical Informatics Initiative. Methods Inf Med 57, e50-e56.

[2] Fluck J (2021) National Research Data Infrastructure for Personal Health Data (NFDI4Health) doi: 10.4126/FRL01-006430386.

[3] ISO 14721:2012: Space data and information transfer systems — Open archival information system (OAIS) — Reference model. International Organization for Standardization, Geneva, Switzerland. https://www.iso.org/standard/57284.html

[4] Schrimpf S (2014) Das OAIS-Modell für die Langzeitarchivierung: Anwendung der ISO 14721 in Bibliotheken und Archiven, Beuth, Berlin.isbn: 9783410239543

[5] Sinaci AA, Núñez-Benjumea FJ, Gencturk Met.al. (2020) From Raw Data to FAIR Data: The FAIRification Workflow for Health Research. Methods Inf Med 59, e21-e32.

[6] Henke C, Graf L, Kuntz AS et.al. (2022) The Way Data Flows: Current Provenance Options in Collaborative Research. 67. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS).

[7] Klopfenstein SAI, Golebiewski M, Schmidt CO et.al. (2022) NFDI4Health Task Force COVID-19 Metadata schema V2_0. doi: 10.4126/FRL01-006431357

[8] Meineke FA, Löbe M, Stäubert S (2018) Introducing Technical Aspects of Research Data Management in the Leipzig Health Atlas. Studies in health technology and informatics 247, 426–430.

[9] Kirsten T, Meineke F, Löffler-Wirth H et.al. (2022) The Leipzig Health Atlas - An open platform to present, archive and share bio-medical data, analyses and models online. Methods Inf Med.

[10] Löbe M, Ulrich H, Beger C et.al. (2022) Improving Findability of Digital Assets in Research Data Repositories Using the W3C DCAT Vocabulary. Studies in health technology and inf. 290, 61–65.

[11] David Giaretta, John Garrett, Mark Conrad, Eld Zierau, Terry Longstreth, John Steven Hughes, Matthias Hemmje, Felix Engel (2022) OAIS Version 3 Draft Updates, Open Science Framework. doi: 10.17605/OSF.IO/93AED

[12] Higgins S (2008) The DCC Curation Lifecycle Model. International Journal of Digital Curation 3, 134–140.