Maximilian Eibl, Martin Gaedke. (Hrsg.): INFORMATIK 2017,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 1381

Selecting, Packaging, and Granting Access for Sharing Study Data

Experiences and Recent Software Developments in the LIFE Study

Toralf Kirsten[1 2] Jonas Wagner[1] Alexander Kiel[1] Mathias Rühle[1] Markus Löffler[1 2 3]

Abstract: Data in medical studies and research projects are captured, curated and analyzed, often, with a substantial personal and financial effort. Such study data are typically managed by institutions and groups who are involved in these studies and projects. Often, they are refrained by these institutions and, thus, not shared with other scientists who are interested in similar medical topics or hypotheses. Open the data for other scientists will speed up medical insights, enable analyzes which currently lacks data amount either by enlarge the set size of study objects and by finding suitable controls, and allow to validate published results taking data from other studies into account. In this paper, we introduce the data sharing approach we use at the LIFE Research Center for Civilization Diseases, University Leipzig. Our approach is influenced by the OAIS reference model for archiving and distributing data to a designated community. We highlight several aspects of this approach, sketch the process and describe the supporting IT infrastructure. In particular, we outline the LIFE Data Portal and the LIFE Proposal Manager allowing to find, access, and reuse metadata and study data for dedicated analysis projects.

Keywords: Data Sharing; Accessing Study Data; LIFE

1 Introduction

Medical scientific data are captured, cleaned, and analyzed with a substantial effort in clinical trials, epidemiological studies and other health-related surveys. Often, they are, however, refrained by the data owners (study consortia). Hence, data are not freely accessible and usable for the scientific community which has at least two negative effects. First, published results are typically not reproducible taking the captured and prepared study data into account, such that scientific questions are re-funded by new research projects which have been already sufficiently scientifically answered. Secondly, methodical imbalances or questions according to the analysis procedure and their impact on the interpretation are hard to discover. In case the results are too questionable but results are for general interest, the study is typically repeated. This "Waste in Clinical Research" has been described in a series of five papers in Lancet Journal 2014 [Ch14a, Ch14b, Sa14, Io14, Gl14]. In consequence, highly ranked medical journals including New England Journal of Medicine, Deutsches Ärzteblatt (German Medical Journal), Journal of Medical Association and others [Ta16]

[1] LIFE Research Center for Civilization Diseases, Univ. Leipzig, Philipp-Rosenthal-Str. 27, 04103 Leipzig
[2] Interdisciplinary Center for Bioinformatics, Univ. Leipzig, Härtelstr. 16-18, 04107 Leipzig
[3] Institute for Medical Informatics, Statistics and Epidemiology, Univ. Leipzig, Härtelstr. 16-18, 04107 Leipzig
{tkirsten, akiel, mruehle, jwagner}@life.uni-leipzig.de, markus.loeffler@imise.uni-leipzig.de

doi:10.18420/in2017_138

vote for publishing research insights together with information about accessing relevant study data.

Often, there are different stages in that data are created, filtered, and manipulated. Raw data are directly the result of data capturing processes. Curated data are generated when raw data are cleaned. Publication data are specifically selected curated data according to the analysis goal or scientific hypothesis for that the publication describes the result. They often include additional scores or other derivatives and are accumulated with external data, such as previous results and publicly available annotations, which enriches curated data in order to extract specific analysis results that are then published. Such publication data are directly included into a publication or are, sometimes additionally, made available as supplementary material in publicly accessible online resources, such as web pages of authors, institutions, and journals. More and more data citation platforms are commonly used, e.g., Havard Dataverse [4], Dryad [5] or other domain-specific data repositories hosting such data according to each publication.

In contrast to publication data, raw and curated study data are managed by study centers or research institutions where the data have been produced or which are basically primarily responsible for their management. Data are shared within an institution and consortia or are shared based on special relationships. Therefore, a general approach has been developed in recent years allowing sharing study data in Germany systematically. This sharing approach has been repeatedly applied and adopted in several studies including KORA [MPL16], SHIP [Jo01], GANI_MED [Gr14], NAKO [Co14], and LIFE [Lö15, Po17]. These studies are hosted by different research institutions. While the sharing approach is similar across these institutions, their support by innovative software tools is different. Some institutions select data manually from databases and utilize spreadsheets to manage all sharing processes whereas other institutions apply web-based software systems.

In this paper, we report about the sharing approach we use at the LIFE Research Center for Civilization Diseases, University Leipzig. This sharing approach necessitates organizational structures and an IT infrastructure allowing us to select, package, and granting access to data for specific analysis projects. Therefore, we start the paper (Section 2) with a description of the sharing process and the required organizational structures. In the second part (Sections 3 and 4) we outline our IT-infrastructure for sharing curated study data. Section 5 concludes the paper.

2 Data Sharing at a Glance

Subsequently, we give an overview highlighting general aspects and outline organizational structures that we created and use in LIFE for sharing data before we outline the sharing process.

[4] http://scholar.harvard.edu/mercecrosas/publications/dataverse-4-defining-data-publishing
[5] http://datadryad.org

2.1 Overview

General data sharing aspects are covered by the Open Archiving Information System (OAIS) [Th12]. OAIS is a ISO reference model (ISO 14721:2012) for archiving and distributing data to a designated community. The reference model describes general sharing aspects, i.e., data are generated by one or more data producers, are then hosted and administered by a data management organization, and finally, provided to data consumers. The reference model describes such different authorities, their roles and functions, and necessary interfaces on a higher level. It is, therefore, a template for data sharing processes that can be applied in medical sciences, too, but need specializations in this application domain. In LIFE, we adopted the reference model

Sharing medical study data requires to concern legal and ethical aspects. Typically, such data are captured by and in interviews, questionnaires, and physical examinations in which patients or probands (participants of a study) answer to predefined questions and are examined by nurses and physicians. Additionally, laboratory data are generated when bio-samples including blood, urine, hair etc. taken from participants are analyzed in a wet lab. In all these cases, captured medical data refer to single participants. Therefore, privacy aspects need to be concerned when medical data will be shared.

2.2 Consent & Privacy Aspects

A first aspect pertains to participant's consent. In medical studies and other research projects, data are only captured from persons who have been consented that all captured data are managed by a named institution and are used for medical research. Sometimes, the given intended use is more specific than the term "medical research"; the consent is then given for a specific medical hypothesis. However, data with such a specific consent (called as informed consent) can only be used in the intended direction; their reuse to analyze other hypotheses is then not allowed and necessitates a renew consent of the participant. The consent is typically captured by a specific form that the participant needs to sign.

Privacy constraints and Good Clinical Practices (as used in clinical trials) necessitates the separation of identification and study data. Identification data include participant's name, address data, id card no. etc. These data are managed by an information system that generates and links a pseudonym to each identification representation; in LIFE the participant pseudonym is called "Subject Identifier Code" (SIC). The SIC is then used to capture study data, i.e., data entry systems never manages study data together with identification data. In LIFE, we also generate pseudonyms for bio-samples and physical examinations allowing to specifically track bio-samples and generated data files. Both pseudonyms are linked with the participant pseudonym SIC. Moreover, a further new pseudonym, the PSIC ("Proposal-specific Subject Identifier Code"), replaces the SIC pseudonym in shared data. Therefore, shared data never contain pseudonyms used in capturing processes.

2.3 Access Management

To respect participant's privacy, captured medical data are in general not freely available but are provided on request. This way, a scientist, a group or consortia need to apply for data using a *Data Usage Proposal* in order to receive data from LIFE studies. With each proposal, the scientist exactly describes what data are required for the intended analysis according to a medical hypothesis the scientist is willing to answer. This data specification contains both, relevant study items (i.e., variables) that should be included in the analysis and relevant cases (i.e., data of participants). The latter results in study events (e.g., visits) of participants. Inclusion and exclusion criteria are typically used to specify and, later, to retrieve relevant data. All proposals need to be maintained and, thus, need organizational and IT support.

2.4 Organizational Structures in LIFE

Coordinated sharing of study data, preferably curated study data, necessitates organizational structures. Figure 1 shows an overview of such structures and their interplay in LIFE. Each group has specific responsibilities and functions. The Metadata and Data Transfer (DMT) group is, on the one hand, at the interface between scientists who conduct and scientifically attend the LIFE studies (left side in Figure 1). In this role, scientists are data producers. In LIFE, there are multiple studies including LIFE Adult, LIFE Heart, LIFE Child, LIFE Child Depression and LIFE Head & Neck Cancer. Each of these studies uses a set of participants who are examined by multiple assessments (interviews, questionnaires, ...). The DMT and study representatives discuss data entry, input forms and their containing study items as well as the resulting data structures; sometimes, the DMT supports the configuration of input forms in specialized data entry systems allowing to capture data. In LIFE, we use Lime Survey for electronic (online) and the Teleform system for paper-based data capturing. The latter allows to scan filled paper forms and to verify recognized input in a second step. On the other hand, the DMT group is at the interface to scientists who requests data by data usage proposals (right side in Figure 1). Therefore, the DMT group needs to have an overview about all assessments on data producing side, their application in any LIFE study and their usage in any data request by scientists on data consumer side.

The DMT group is in contact with IT group which is responsible for technical data management (see Figure 1). Data are delivered by study representatives, over the DMT group or are retrieved directly from data entry systems. In LIFE, all data are centrally managed. Structured data are integrated into the research database whereas unstructured data, i.e. files, are managed in file system. Metadata are usually used to describe captured data. Such metadata can be extracted from input forms since entry fields are labeled by questions and parameter names, e.g., for physical measurements, or need to be specified manually. In LIFE, metadata are managed by a Metadata Repository. Their use is threefold. Firstly, metadata are used for harmonization when data from different input forms and input systems need to be integrated. Secondly, they are used to search for study items of interest and, thus, answers questions regarding the availability and granularity of items (see below)

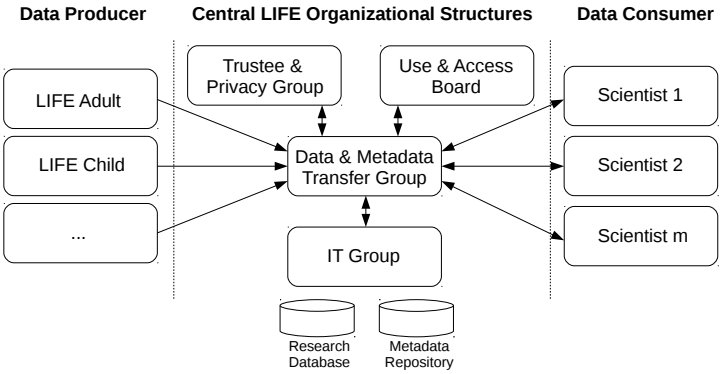

Fig. 1: Organizational Structures for Sharing Study Data in LIFE

for that data have been captured in LIFE studies. Finally, metadata are used to automatically build queries over data collections in the research database. In [Ki17] (data integration) and [UK15] (data retrieval) we describe both metadata-related aspects in more detail.

The Trustee and Privacy Management (TPM) group concerns data privacy aspects. Firstly, the TPM is responsible for managing consent information taken from each participant. The group manages the signed consent paper forms and withdraws (since both, consent and withdraws, are not yet electronically captured in LIFE). Moreover, the TPM group is also responsible for managing participant's identification data. While identification data are mostly used in processes before data have been captured, e.g., for participant invitations and appointment making, their relation to study data is only necessary in special and rare cases, e.g., when selected study results should be communicated on individual basis or when participants are selected based on study results and reinvited for a special follow up. The TPM group resolves the identification taking the relevant set of participant pseudonyms into account.

Finally, a Data Use & Access Board (DUA) reviews the Data Usage Proposals which have been submitted by scientists to request data from LIFE studies. As a result of the review process, the board accepts or declines a proposal. In LIFE, the DUA board is not an organizational unit but a group of scientists and administration staff from different groups including DMT, TPM, and IT. This board regularly meets every four weeks.

2.5 Feasibility Queries and Data Sharing Process

The BPMN diagram in Figure 2 outlines the most relevant steps of the sharing process we used in LIFE. Usually, a first step of a scientist with a medical hypothesis in mind is to look for relevant data. This question for information combines two aspects (activities 1 + 2). Firstly, a scientist is interested in looking for study items (variables) or document classes that are relevant for the intended hypothesis (activity 1). This allows the scientist to determine whether the necessary data is available and in the right granularity. Such querying

requires metadata about assessments and study items of structured and unstructured data. While the first include data that are typically generated by interviews and questionnaires and are in tabular form, the latter are usually different types of files, i.e., images, videos, and any other outcome of physical examinations using medical devices. Often the latter also includes raw data of genetic determinations ranging from a single large file to a large set of files per examination. Typically, a Metadata Repository manages metadata about structured and unstructured data and supports the intended querying and information retrieval (search).

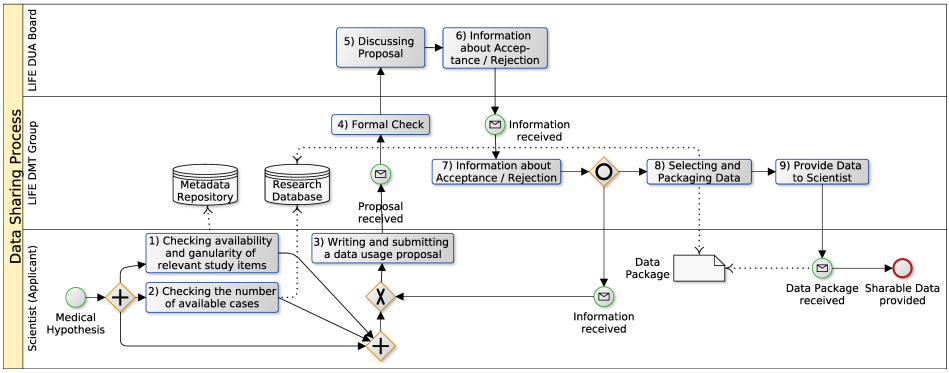

Fig. 2: Most relevant Activities of the Sharing Process in LIFE

Once a scientist found relevant data are structurally available, a next step (activity 2) is to count relevant cases (participants and their study events) according to the scientific hypothesis. The scientist can then analyze the number of cases to be included in the intended analysis before data is selected and provided by the DMT group. This saves time and resources for both sides, the scientist and the LIFE Center since each scientist can look for the availability of relevant data without any interaction with the LIFE Center (DMT group). Moreover, checking the case number allows to verify whether the analysis can be performed with a proper statistical power.

When relevant study items are available and the case number is sufficient, a scientist writes a Data Usage Proposal to request data for the intended analysis. In LIFE, we provide a template for writing such proposals that helps scientists and the LIFE Center to give / receive the right information. This information includes the scientific background, the medical hypothesis that should be analyzed, the analysis plan, and the data specification (see above). Moreover, the scientist need to inform the stakeholder, i.e., the PIs of the corresponding LIFE study from that data are requested; they need to sign the proposal, too. Finally, the scientist sends the proposal to the DMT group at the LIFE Center. The DMT group (activity 4) formally checks the proposal, e.g., whether all mandatory input fields are filled and that all relevant person have signed the proposal. Next, the DMT group provides the proposal to the DUA board checking (activity 5) the scientific goal and its consistency with the analysis plan and requested data. The DUA board decides whether the proposal is accepted and, thus, activated or need to be clarified in some points. In the latter case, the applicant can change

the proposal and resubmit it. A proposal is activated for an initial period of time but can be yearly prolongated at the end of each period. An accepted proposal creates immediately an analysis project for which data will be shared with the applicant. In section 3.2 (Figure 4), we show more precisely transition states a proposal (and analysis project) can have. At the end of each meeting, the DUA board informs the DMT group about decision regarding each discussed proposal that distributes this information to each applicant (activity 7).

Using the data specifications of the Data Usage Proposal, the DMT group can select, prepare (or pre-compute if necessary) and package the requested data for sharing (activity 8). The data package can then be transferred to the applicant (activity 9).

3 LIFE IT Infrastructure for Data Sharing

Subsequently, we highlight aspects of the LIFE IT infrastructure for sharing data. After giving a high level overview about the infrastructure, we outline the proposal management.

3.1 Overview

Figure 3 shows an high level overview over the main parts of the LIFE IT infrastructure according to data capturing (left), to data integration / harmonization (center), and to querying and sharing study data (right); information systems for ambulance management (e.g., participant tracking and invitation) and managing bio-samples, i.e., laboratory information system are suppressed for simplicity. There are different kinds of information systems producing and are used to capture data. Electronic data capturing systems (EDC) provide online forms allowing to electronically capture study data. Paper based forms are used in situations and for participants who prefer paper instead of online surveys. In LIFE, we design such paper based forms using a software system allowing scanning and verifying filled paper forms afterwards. Simultaneously, there are examinations in that medical devices producing data files, directories of data files, and data within databases. Moreover, there are miscellaneous systems managing data in spreadsheets and desktop databases. Structured data of such sources are centrally integrated into a research database whereas unstructured data (e.g., images, image sequences, videos, etc.) are systematically managed in a central file system. Centrally managed data can be curated, e.g., excluding data from specific participants and manipulating data. All curations are specified by the Data Curation App that documents all curations for later reuse. A Metadata Repository collects metadata describing the structure and the matter of centrally managed data.

Centrally managed and integrated study data that are described by metadata are the basis for sharing in LIFE. We have recently developed two central information systems for sharing LIFE study data. The first is the LIFE Data Portal (LDP) allowing a scientist to execute feasibility queries. Such queries focus on metadata and study data available in LIFE. The LDP provides metadata on different levels. Firstly, there are metadata on study level, i.e., the objective and description of the study, principal investigators, contact persons, cohort population source and size etc. Secondly, metadata on assessment level describe the intention

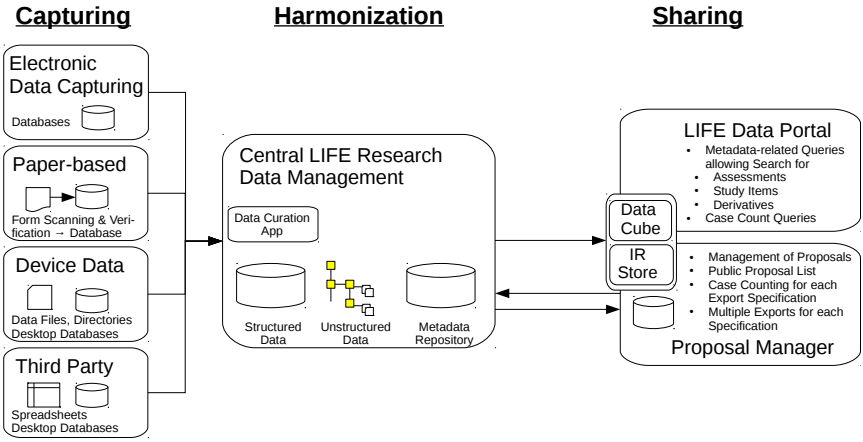

Fig. 3: Overview about LIFE Infrastructure for Capturing, Harmonization and Sharing Study Data

of each examination form, i.e., interview survey, questionnaire etc., and give further details about contact persons who were responsible and taking care data capturing. The latter is importantly for scientists in analysis projects, especially when intermediate or final results need to be explained and interpreted. On the third level, the LDP provides metadata about study items of each assessment, such as body height and weight of an anthropometry assessment. In LIFE, this metadata include item descriptions, the data type (text, number, date), the scale (metric, nominal, ordinal, ...) the item follows and the associated code list when a study item is of categorical scale. Using this metadata a scientist get insights into assessments of a study and can look for study items that are relevant for the intended analysis. The LDP provides search and browsing functions according to metadata. Moreover, the LDP allows case count queries, i.e., a scientist can use metadata to formulate queries about study data to find out how many visits and study participants are available. With both, searching / browsing in metadata and case count querying, the LDP offers more functionalities than similar software tools, such as i2b2 [Mu09, We09], that is mainly focused on querying a single or distributed sources.

We provide and use the LIFE Proposal Manager (LPM) in LIFE for data sharing. The goal of LPM is to manage systematically all Data Usage Proposals that have been submitted to LIFE. The LPM annotates each proposals by a title, principal investigators, the current status, the date the proposal has been submitted, the analysis project has been started, and data have been shared etc. For these functions, the LPM is primarily used by the DMT group, internally. Moreover, we record all status of a proposal (c.f. next subsection 3.2), in particular, when the DUA board decides to activate a proposal and, therefore, to initiate a new analysis project. We also use the LPM at the interface to interested scientists by providing a list of all proposals (i.e., proposal titles), their principal investigators and the current state. The objective of this list is to show what medical hypotheses have been

worked on and are currently work in progress. Moreover, the list can also be used to find collaboration partners by taking principal investigators per topic into account.

Both, the LIFE Data Portal and the LIFE Proposal Manager, utilize a Data Cube and IR (Information Retrieval) Store for searching in metadata and executing feasibility queries. The data cube is mostly inspired by the data warehousing approach and contains structured study data in a multidimensional model. We use the Datomic database system [6] for managing the data cube that is nightly feeded with data from the central Research Database. Metadata are managed in both, the Datomic Database and an Elasticsearch system (IR Store) [7]. The first is used for browsing and for querying study data whereas the latter allows similarity based searching for assessments and study items taking their titles and descriptions into account.

3.2 Proposal Management

A central function of the LIFE Proposal Manager is the management of Data Usage Proposals. More than 400 of such proposals have been submitted in LIFE. The first has been activated in January 2012; the submission frequency per month raised from 2 in 2012 to 6 in 2017 (for four and a half months in 2017). Currently, most proposals are timed-out or closed, only one proposal has been rejected whereas 53 analysis projects are currently active (as of May, 2017). Therefore it is important to track the status of each proposal over its life time. Figure 4 shows the state transition graph the LPM follows. Currently, a new proposal is submitted as hard copy, yet, since each proposal is to sign by all partners of the intended analysis project as well as the principal investigators of the LIFE study from which data are requested. Simultaneously, the proposal is created within the LPM system; each new proposal is attached with the state "submitted". When the proposal arrives the DMT group, its state is changed to "in review". This holds for both, the formal check by the DMT group and the discussion round of the DUA board. As a result of a board meeting, the proposal is "activated" or need some clarifications. In the latter case, the proposal can be changed and resubmitted. In the meanwhile, the proposal is stated as "change requested". In very rare cases, a proposal is rejected, e.g., when the proposal violates ethical or legal constraints or is not in line with principal investigators. Activating a proposal creates an analysis project. Data are only shared for such projects and, thus, activated proposals (see Figure 2). Each analysis project is valid for a specific time. During this time, the medical hypothesis and, therefore, the analysis goal is reserved for the applicant and, thus, each new proposal with the same objective will be rejected. The initial time interval for an analysis project is one year but can can be prolongated before the project runs out of time. The "timed out" state is marked to reminding applicants to prolongate his/her projects or to "close" them. This is the case when the hypothesis is validated or proved as wrong. Similarly, a project can be "canceled" at any time.

[6] http://www.datomic.com
[7] https://www.elastic.co

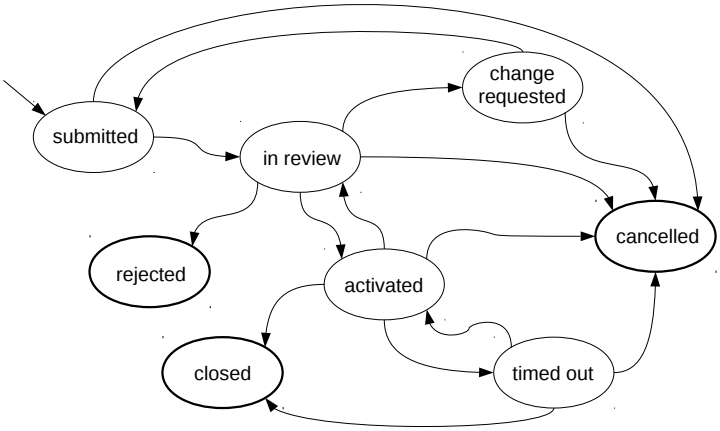

Fig. 4: States and their Transitions of Data Usage Proposals

4 Selecting and Packaging Data

Most scientists expect to receive data files containing the requested data; only few applicants wanted to get access on data using database API. Therefore, we export data in most cases to data files; for all other we create a user-specific database schema into which relevant data of the central Research Database are copied. Before data are exported or copied into the new database schema, all internal pseudonyms (SIC) are replaced by new proposal specific pseudonyms. With that, we impede merging data from two or more proposals by the same or multiple scientists and, thus, building shadow databases. Moreover, most scientists prefer a single data file containing data of different assessments in a joined manner. The intention is to directly load this single file into the analysis tools of their choice, such as R, SPSS etc., and can start the analysis. Each applicant receives data files and annotated case report forms (aCRF) explaining the meaning and representation (data type, code lists etc.) of assessments and study items. We automatically derive such aCRFs from the our Metadata Repository. Additionally, we create code books for R and SPSS, i.e., values of categorical items are directly associated with labels (e.g., gender: 1 - male, 2 - female) which can then be used in analyses and any visualizations instead of raw codes.

We started with an ontology based selection and retrieval of study data [UK15] which can then be exported into data files. This approach was implemented as Protegé Plugin [Mu15]. This tool is a desktop software and was mainly used by the DMT group only. However, as the total number of proposals and their frequency per month increase there is the need for a collaborative (web-based) software that is used by multiple users including all central groups (not only the DMT group), applicants and further interested scientists. Therefore, we are working on a new module extending the LIFE Proposal Manager allowing to specify data export specifications. Such specifications are the basis for data exports from databases into data files which are then transferred to the applicant. The goal is then, that the applicant can download the data directly from the LPM. Like the LIFE Data Portal, the LPM uses

queries for export specifications regarding to available metadata on assessment and on item level.

5 Conclusion and Outlook

In this paper, we reported about the data sharing approach we follow at the LIFE Research Center for Civilization Diseases. The sharing approach is implemented by a process and organizational structures, i.e.., groups, at the LIFE Center. Both, process and organizational structures are inspired by the OAIS reference model and are supported by a complex IT infrastructure. We provide the LIFE Data Portal for running feasibility queries taking metadata and study data into account. Each scientist of the Medical Faculty at the University Leipzig can execute such queries without any interaction with the LIFE Center to get insights what kind of data have been captured in LIFE or for finding relevant data that need to be included in an analysis project. Data are then requested by a Data Usage Proposal. The LIFE Proposal Manager is used to manage fast increasing number of proposals in LIFE. Currently, we are working on a new module allowing to export data from the LIFE Research Database collecting and integrating all captured data in LIFE.

Acknowledgment

This publication is supported by LIFE - Leipzig Research Center for Civilization Diseases, Universität Leipzig. LIFE is funded by means of the European Union, by the European Regional Development Fund (ERDF) and by means of the Free State of Saxony within the framework of the excellence initiative.

References

[Ch14a] Chalmers, I; Bracken, MB; Djulbegovic, B; Garattini, S; Grant, J; Gülmezoglu, AM et al.: How to increase value and reduce waste when research priorities are set. The Lancet, 383(9912):156–165, 2014.

[Ch14b] Chan, AW; Song, F; Vickers, A; Jefferson, T; Dickersin, K; Gøtzsche, PC; Krumholz, HM; Ghersi, D; Worp, HB: Increasing value and reducing waste: addressing inaccessible research. The Lancet, 383(9913):257–266, 2014.

[Co14] Consortium, German National Cohort (GNC): The German National Cohort: aims, study design and organization. European Journal of Epidemiology, 29(5):371–382, 2014.

[Gl14] Glasziou, P; Altman, DG; Bossuyt, PP; Boutron, I; Clarke, M; Julious, S; Michie, S; Moher, D; Wager, E: Reducing waste from incomplete or unusable reports of biomedical research. The Lancet, 383(9913):267–276, 2014.

[Gr14] Grabe, HJ; Assel, H; Bahls, T et al.: Cohort profile: Greifswald approach to individualized medicine (GANI_MED). Journal of Translational Medicine, 12:144, 2014.

[Io14] Ioannidis, JPA; Greenland, S; Hlatky, MA; Khoury, MJ; Macleod, MR; Moher, D; Schulz, KF; Tibshirani, R: Increasing value and reducing waste in research design, conduct, and analysis. The Lancet, 383(9912):166–175, 2014.

[Jo01] John, U; Greiner, B; Hensel, E et al.: Study of Health In Pomerania (SHIP): a health examination survey in an east German region: objectives and design. Sozial- Und Präventivmedizin, 46(3):186–194, 2001.

[Ki17] Kirsten, T; Kiel, A; Rühle, M; Wagner, J: Metadata Management for Data Integration in Medical Sciences - Experiences from the LIFE Study. In: Proc. 14th Conf. of Database Systems for Business, Technology and Web (BTW). 2017.

[Lö15] Löffler, M; Engel, C; Ahnert, P et al.: The LIFE-Adult-Study: objectives and design of a population-based cohort study with 10,000 deeply phenotyped adults in Germany. BMC Public Health, 15, 2015.

[MPL16] Meisinger, C; Peters, A; Linseisen, J: From the MONICA-project via KORA to the NAKO-study: Practical Utility of Epidemiological Studies in Augsburg Region. Gesundheitswesen (Bundesverband Der Ärzte Des Öffentlichen Gesundheitsdienstes (Germany)), 78(2):84–90, 2016.

[Mu09] Murphy, S; Churchill, S; Bry, L; Chueh, H; Weiss, S; Lazarus, R; Zeng, Q; Dubey, A; Gainer, V; Mendis, M et al.: Instrumenting the health care enterprise for discovery research in the genomic era. Genome research, 19(9):1675–1681, 2009.

[Mu15] Musen, MA: The Protégé project: A look back and a look forward. AI matters, 1(4):4–12, 2015.

[Po17] Poulain, T; Baber, R; Vogel, M; Pietzner, D; Kirsten, T; Jurkutat, A et al.: The LIFE Child study: a population-based perinatal and pediatric cohort in Germany. European Journal of Epidemiology, 32(20):145–158, 2017.

[Sa14] Salman, RA; Beller, E; Kagan, J; Hemminki, E; Phillips, RS; Savulescu, J; Macleod, M; Wisely, J; Chalmers, I: Increasing value and reducing waste in biomedical research regulation and management. The Lancet, 383(9912):176–185, 2014.

[Ta16] Taichman, DB; Backus, J; Baethge, C; Bauchner, H; de Leeuw, PW et al.: Sharing Clinical Trial Data — A Proposal from the International Committee of Medical Journal Editors. New England Journal of Medicine, 374(4):384–386, 2016.

[Th12] The Consultancy Committee for Space Data Systems: Recommended Practice, CCSDS 650.0-M-2 (Magenta Book) Issue 2. `http://public.ccsds.org/publications/archive/650x0m2.pdf`, Accessed: 10.05.2017.

[UK15] Uciteli, A; Kirsten, T: Ontology-based retrieval of scientific data in LIFE. In: Proc. Workshop on Data Management for Life Sciences, attached at 13th Conf. of Database Systems for Business, Technology and Web (BTW). 2015.

[We09] Weber, GM; Murphy, SN; McMurry, AJ; Macfadden, D; Nigrin, DJ; Churchill, S; Kohane, IS: The Shared Health Research Information Network (SHRINE): A prototype federated query tool for clinical data repositories. J Am Med Inform Assoc, 16:624–30, 2009.