International Journal of
Geo-Information

MDPI

Article

Requirements on Long-Term Accessibility and Preservation of Research Results with Particular Regard to Their Provenance

Andreas Weber * and Claudia Piesche *

IT Service Centre, University of Bayreuth, Bayreuth 95447, Germany
* Correspondence: andreas.weber@uni-bayreuth.de (A.W.); claudia.piesche@uni-bayreuth.de (C.P.);
Tel.: +49-921-55-5851 (A.W.); +49-921-55-5855 (C.P.)

Academic Editors: Constanze Curdt, Christian Willmes, Georg Bareth and Wolfgang Kainz
Received: 29 January 2016; Accepted: 28 March 2016; Published: 11 April 2016

Abstract: Since important national and international funders of research projects require statements on the long-term accessibility of research results, many new solutions appeared to fulfil these demands. The solutions are implemented on various scopes, starting from specific solutions for one research group up to solutions with a national focus (*i.e.*, the RADAR project). While portals for globally standardized research data (e.g., climate data) are available, there is currently no provision for the large amount of data resulting from specialized research in individual research foci, the so called long-tail of sciences. In this article we describe the considerations regarding the implementation of a local research data repository for the Collaborative Research Centre (CRC) 840. The main focus will be on the examination of requirements for, and an agenda of, a possible technical implementation. Requirements were derived from a more theoretical examination of similar projects and relevant literature, diverse discussions with researchers and project leaders, by analysis of existing publication data, and finally the prototypical implementation with refining iterations. Notably, the discussions with the researchers lead to new features going beyond the challenges of the mere long-term preservation of research data. Besides the need for an infrastructure that permits long-term preservation and retrieval of research data, our system will allow the reconstruction of the complete provenance of published research results. This requirement is a serious diversification of the problem, because it creates the need to qualify additional transformation data, describing the transformation process from primary research data to research results.

Keywords: research data management; publications; long-term preservation; metadata

1. Introduction and Motivation

The importance of long-term accessibility and public access to research data grew significantly with the publication of the "OECD Principles and Guidelines for Access to Research Data from Public Funding" in 2007 [1]. The main idea of the guidelines is to encourage the publication of research data with open access rights to allow the reuse of the results by other research groups. Thus, research processes in general can be accelerated by avoiding duplicated work. At the same time, significant journals (e.g., Nature) began to request the accessibility of the research data used in the publications. The published results should become reproducible for the reviewers and the readers. These basic ideas were adopted by major funders (e.g., in Germany by the Deutsche Forschungsgemeinschaft - DFG) [2]. Therefore, the description of the long-term accessibility of research data has to be a part of research proposals and is nowadays a precondition for the funding of large projects.

Although the demand for long-term preservation of research data has existed for a while, infrastructures for the storage of research data have developed slowly, especially in the areas where research data are primarily individual. Globally accessible repositories for the preservation and the reuse of research results are established in only few research areas. The main reason for the lack of transregional repositories is that the cooperative collection of research data and quality assurance is only possible for highly standardized data. Good examples where the standardization was achieved are the Human Genome Project (HGP) [3,4], climate research, or biological diversity (GRBIO) [5]. In most of the research fields experiments are highly specialized and sometimes even singular. The general description of these very specific experiments, calculations, or simulations and their results is very time-consuming and researchers often avoid spending time on the detailed description of their work. Existing solutions for the long-term storage of data of that kind are specialized and usually not designed for public use, much less for reuse.

The situation at universities actually is still diverse. The support for the preservation of research data is mostly limited to the provision of high-available disk storage and appropriate backup solutions. Collaboration is in many cases limited to the use of shared folders. Tools or portals to support the search of metadata are very rare. The institutions that could play an important role, like libraries or IT centers, hesitate to build up solutions, because policies for the treatment of research results are not yet installed by the administration. Thus, activities result usually from single projects leading to a number of independent and incompatible repositories. These very specialized solutions are not designed to be extended to be used in other research foci, and therefore the number of repositories is increasing.

An interesting approach arises from the installation of so called "Fachinformationsdienste (FIDs)" at German libraries [6]. In answer to an evaluation of researchers' expectations conducted in 2010-2011, the DFG restructured the nationwide special subject collections system. The main focus for the new installed FIDs will be the necessary supply of data and information regarding the special needs of each research area. Some libraries having applied for such FIDs to start including solutions for the long-term preservation of research data in the portals for the specific interest groups [7].

Whereas the funders and publishers generally demand long-term accessibility and the open access to research data [8], the discussions with researchers show that they have different views on that topic. The improvement of the long-term accessibility and the collaboration with other research groups is highly welcome, in general. Their stance on access to research data, however, is that it should be strictly limited to specific persons during the research process; even after the publication of the results, open access to primary research data is not wanted, because it is believed that the reuse of the data should be limited to the original research group, at least for some time. With this in mind, even the access of reviewers to the data is seen skeptically. Therefore, any solution to manage research data must implement a very sophisticated rights management system, coping with the demand of protecting data from unauthorized access on one side, and granting public access to data on the other side.

In the special case of the Collaborative Research Centre (CRC) 840, the coordinators of the sub-projects joined in as additional important stakeholders. They have demanded an easy way to understand the provenance of research results. More precisely, they want to get information on the origin, transformation, and interpretation of research data, if necessary. The question arises from the fact that researchers may leave the CRC during the funding period. When questions on research results arise it will be very difficult to reproduce the results just from the primary data or the results. Therefore, not only results and primary research data should be stored, but the whole process of the transformation from primary data to research results. The challenges evolving from this demand are discussed later in detail.

Due to these heterogeneous demands, we decided to evaluate the requirements for a data management repository on several levels. First of all, we made a desk-study of already existing projects, their requirements, and their implementation progress. After that, some detailed interviews and discussions with researchers within the CRC lead to the first concept for a repository, which was subsequently implemented as prototype. This first implementation was refined in some iterating steps to include the suggestions from the researchers. This attempt was focused on the processing of primary research data. To complete the concept in respect to the representation of the data transformations, initially we tried to get an idea of what kind and amount of data we would have to deal with. For that reason, some publications within the CRC were and will be analyzed to trace provenance from raw data to published results.

In a nutshell, the main focus of this manuscript is the description of the requirements origination process for a research data repository for the CRC 840, the requirements itself, and the concept of implementation resulting therefrom.

2. Requirements Resulting from a Theoretical Background

Due to the broad interpretation of research data and research data management, in general, and the heterogeneous demands from involved stakeholders, a wide range of related topics have to be taken into account. To not lose our way, we first declared three principle questions as an orientation through the various publications:

(1) What are research data, in general, and how do we define the term regarding the CRC 840?
(2) What does long-term preservation and data management mean?
(3) What can we learn from promising approaches in already existing projects?

For a start, we took a look at the academic view of long-term preservation, data management, and the resulting requirements for an appropriate concept for data curation, preservation, and publishing. In connection with the definition of useful metadata and the usage of reasonable metadata schemata, one can find various research foci. At first, one aspect concerns structural and semantic information on published data, and secondly, enhanced publications based on this information and finally, infrastructures, or concepts supporting e-Research approaches. Three projects were evaluated that represent three different foci concerning data management infrastructure. The project "RADAR" is presented as an implementation of an overall research data repository. The projects "Prospect" and "Driver I + II" were taken into account, because of the approach to enhance publications with necessary information concerning structure (structural information) and meaning (semantic information).

2.1. Fundamental Concepts

2.1.1. Long-Term Preservation and Data Management in General

Ludwig and Enke [9] describe the aims of the long-term preservation of data as the composition of bit stream preservation and the enabling of the subsequent use in content, as well as in technical purpose. To guarantee the reuse of research results, it is necessary to gather background and contextual knowledge regarding their provenance. For the gathering of preferably complete information on research data, Ludwig and Enke [9] developed a checklist on the basis of a lifecycle model and the "Curation Continuum" for the planning of useful research data management. The "Data Curation Continuum" [10] incorporates the fact that research data is generated, transformed, and published in different domains (research, collaboration, and public), where different stakeholders are involved, different metadata are needed, and differing access rights and control level take place.

Research data run from the private research domain, through the shared research domain of the internal group, to the public domain after publication of the results. Hence, the characteristics of the research data range between the start and endpoint of a continuum while processing this lifecycle. As a consequence, one cannot identify special information for each domain, but the same information with a different focus. For example, the concrete information about measuring context is a very important fact for the researcher actually doing the measurement and evaluation. This information takes a back seat in the scope of something like a CRC (shared domain). Here, the information about who's doing the experiment, when did it take place, and which hardware and software were involved are of greater interest. In comparison, the additional information for the same data object needed in the public domain are more abstract and simplified, because of the heterogeneous audience.

Because our approach of data management will support all phases of the data lifecycle, it will be necessary to consider the Data Curation Continuum (e.g., for the creation of adequate metadata schemata).

2.1.2. Semantic Information—“Scientific Publication Packages (SPPs)”

According to the definition of research data in the broader sense, the data has to be associated with useful metadata (*i.e.*, information describing its creation, transformation, and/or usage context). Furthermore, the increasing collaboration of researchers in recent years has evoked the demand to share primary research data as well as information concerning the transformation from raw data to published research results. The information needed for this goes beyond pure structural information, and might be defined as semantic information, enriching data with scientific context.

Hunter [11] describes a way of encapsulating research data, its metadata, and genesis within so-called “Scientific Publication Packages” (SPPs). These packages contain the complete structural and semantic information concerning individual research data, its relationship to other data, the genesis of research results from the primary data, linked publications, “and the associated contextual, provenance, and administrative metadata”. This aggregation of all associated information and data creates the opportunity to treat complex components as a single digital object. Insofar, SPPs allow the easy publication, sharing, and providing of complex research data.

The approach of SPPs allows for the enrichment of data with useful information, and therefore deserves further attention, even if it is only in a conceptual stage at the moment. Its main shortcoming is the current lack of a realistic implementation scenario. Thus, we were not able to use SPPs within our repository for enriching research objects, but we are encouraged to take a closer look at how to find a standardized way to enhance data objects.

2.2. Comprehensive Projects

2.2.1. An overall Research Data Repository—Project “RADAR” [12]

The main goal of the project “RADAR” [13], funded by the DFG, is the implementation of an overall research data repository as a service for research institutions. With its help, researchers will be able to preserve and publish research data [14]. On the basis of the ideas of the “Data Curation Continuum”, the project provides a two-stage process [15]. On the one hand, the repository will support a multidisciplinary approach by offering pure data preservation. On the other hand, there will be an advanced offer for the preservation and publication of research data.

At a closer look, the resulting system aims at so called long tail disciplines without standardized metadata schemes by defining one metadata schema for all.

“The (...) scheme aims to enhance the traceability and usability of research data by maintaining a discipline-agnostic character and simultaneously allowing a description of discipline-specific data. The RADAR Metadata Schema (...) includes nine mandatory fields which represent the general core of the scheme. (...). Additionally, 12 optional metadata parameters serve the purpose of describing discipline-specific data.” [16] (p. 6).

The discipline-specific data are treated as additional research data associated with the original data not considered in the search process for metadata. Accordingly, the RADAR solution implements a repository for long term storage of research data with very little focus on the consideration of individual metadata for different research foci. This fact indicates that an overall data management solution with individual support for various research disciplines has to be modular and involves manual adjustments (configuration). Because we consider it important for the CRC to implement manually configurable metadata information, we refrained from the requirement of one standardized metadata scheme. Therefore RADAR in the current state is not an adequate solution for the CRC.

2.2.2. Structural Information—Project "Prospect"

Every publication has structural components like semantic type, media type, media format, network location, and linked data, like additional information or other associated publications [17]. In addition to the manual editing of such information by the author, it would be very comfortable, and therefore desirable, to automatically gather this information during publication of research results. From a technical point of view, it is possible to easily extract information on the size, count, format, and hierarchical structure of data objects. Conversely, the automatic gathering of associated data objects is more complicated. One idea to meet this challenge is the enhancing of publications with the help of a so-called Markup Language (for example, by using XML). Authors must tag all relevant terms within their publications, so that similar or suitable data can be linked afterwards.

Currently, there are various Markup Languages for specific research areas such as Chemical Markup Language, Mathematics Markup Language, or Biology Markup Language.

The project "Prospect" [18], initiated by the Royal Society of Chemistry, offers an Ontology terms feature presenting drop-down boxes if an existing term of the associated Markup Language is used. Terms are searched within the Gene Ontology, the Sequence Ontology, or the Cell Ontology. Furthermore, it is possible to configure the highlighting of different terms in a publication. Thus, the reader is able to evaluate the text easily.

Both the implementation in the aforementioned project and the other Markup concepts lack the necessary abstraction level. An individual Markup Language is only useful within the specific academic field. Therefore, the concept is only useful for subject-specific repositories.

2.2.3. Enhanced Publications—Projects "Driver"/"Driver II"

Woutersen-Windhouwer and Brandsma [17] (p. 79) provide the definition that "An Enhanced Publication is a publication that is enhanced with research data as evidence of the research, extra materials to illustrate or to clarify or post-publication data like commentaries and ranking." The Driver projects are concerned with enhancing research results with additional and useful information before or during publication. To do this, existing standards, infrastructures, and concepts were evaluated. Driver deals with research results in the form of publications, and not with data produced during the whole research process. However, the idea of gaining information while uploading data in any form can be added to our list of requirements, because the automatic information search will be very handy for the user.

2.3. Lessons Learned and Resulting Requirements

To sum up the evaluation of related work, one can state the fact that there is no "Best Practice" to support research data management without technical and scientific consideration of the objectives. Every concept or solution has its own consequences for the implementation or usage of a data management repository. With this in mind, we had to provide and define the first attempts to answer the relevant questions from the beginning.

(1) Research data, in general, are data which were generated, observed, or measured to verify a scientific assumption [19]. That means that they are very heterogeneous in both amount and structure, and depend on the individual research area. Research data in a broader sense are associated with additional information regarding creation context, and transformation of data and structure, which means that these types of information have to be taken into account as well. Thus, the definition of research data in the special case of the CRC 840 must be made in cooperation with the associated research fields and their researchers. The concrete demands towards the support of diverse research data and metadata schemes were gained by interviews with some participating researchers and are described later. Nevertheless, we could already see the demand of a flexible and customizable data/metadata storage.

(2) This took us to the next question about long-term preservation and management of existing data. Long-term preservation in a more classical sense means the bit stream preservation, and aims at a subsequent use of data in content as well as in technical purpose. The technical realization is determined by the Open Archival Information System (OAIS) standard. This reference model goes far beyond highly-available storage and backup systems by providing preservation plans that even allow a conversion of data formats. Although the implementations of this model are rather complex, the standard is internationally accepted as the reference. Therefore, a solution for the long-term preservation of research data should be compliant to OAIS.A second important aspect of long-term storage is the access to these data. To grant distinct access to the research data, a unique identifier must be assigned. An appropriate system to identify digital objects is the Digital Object Identifier (DOI) system. As per the DOIs, the non-for-profit organization DataCite collects the information on research data with the aim to establish easy access to the data and make the data more visible.

The main aspect we learned from the evaluation of existing approaches was the fact that a higher standardization of data structure and metadata schemes leads to less individualization. So we had to decide whether to implement a repository with one or two standardized data management processes, or whether to allow for a configurable system that supports the creation of diverse metadata schemes fitting individual demands, with the drawback of a greater need for configuration.

3. Challenges, Requirements, and Concepts within CRC-Subproject INF Z2

Next, the "common" demands of research data management, and the individual demands of project leaders and researchers had to be taken into account for the implementation of an adequate research data repository within the CRC 840. The most important challenges which must be met for successful implementation are described below.

3.1. Description of the CRC 840 and the INF Z2 Sub-Project

The CRC 840 "From Particulate Nanosystems to Mesotechnology" seeks to connect molecular or particulate nanoscaled building units to complex functional building blocks with macroscopically utilizable effects. Establishing this essential interface between nano-sized objects and macroscopic systems is one of the grand challenges faced by the nanotechnology community today, and is referred to as mesotechnology. [20].

The sub-project "INF Z2", as part of the CRC 840, addresses the information infrastructure. As mentioned in [2], the purpose for this kind of funding module is twofold. First and foremost, there is the need to backup and preserve the particular research data within the special CRC. Secondly, a constructive collaboration of researchers within and outside the CRC requires suitable infrastructural support. "Thus, the project should serve all or most of the scientific projects. It should consider existing standards and get connected to existing repositories, data bases or the like. It is also required to maintain, manage, document, and backup the data in cooperation with a local library or computing center in a sustainable, permanent, and stable system." [2] (p. 36).

When the sub-project "INF Z2" started, the research data were stored on local devices or network storage systems of institutes. Therefore, the access was limited to single persons or small groups. The description of the data was done in traditional ways. The lifecycle of research data allowed broad access from outside only to the publications, while the research data was hidden in local storage systems, as shown in Figure 1.

The researchers are concerned with very specific problems and consequently only very few data objects are candidates to be stored in global repositories. Therefore, a local solution has to be designed which satisfies a variety of needs. Due to the great technical efforts involved, the IT-Service Centre (ITS) was asked to take over the responsibility for the project and the implementation of the data repository. The allocation within a central institution of the university allows a more general design of the system and thus a future use beyond the research fields of the CRC.

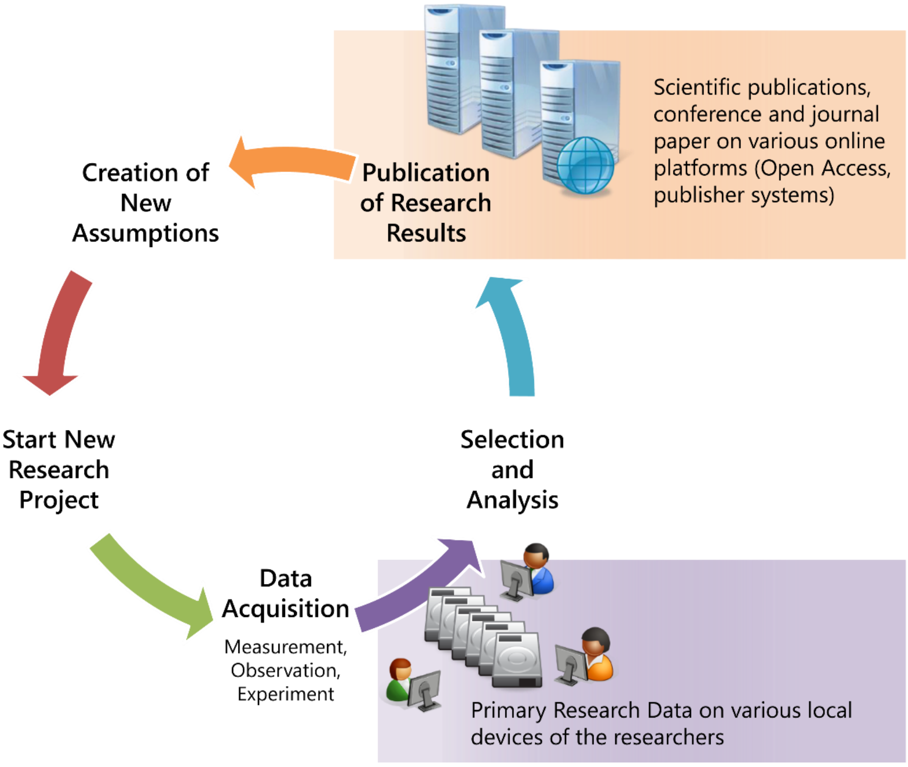

Creation of New Assumptions
Publication of Research Results
Scientific publications, conference and journal paper on various online platforms (Open Access, publisher systems)
Start New Research Project
Selection and Analysis
Data Acquisition
Measurement, Observation, Experiment
Primary Research Data on various local devices of the researchers

Figure 1. Traditional Lifecycle of Research Data.

3.2. Challenges

In [8] the DFG specifies the changing expectations of the researchers concerning the collaboration and publication of research data, and distributed scientific knowledge and research. The creation of integrated digital environments within research organizations and relevant nationwide policies are essential future issues which must be met by appropriate data management concepts and implementations. Furthermore, open access to all published research results must be guaranteed.

In addition to the challenges brought up by the desk-study of theoretical concepts and the requirements of the DFG, there are individual demands arising out of the scientific structure of the CRC 840. That is why project leaders and researchers were asked during structured interviews about their current work approach, future plans, and wishes. The questionnaire was based upon the checklist in Ludwig and Enke [9]. All in all, it contained 27 open questions, grouped by topic.

(1) *Provenance of Research Data*

This section considered questions regarding origin, kind of measurements, structure, and type of generated data, and the common processes on how to create and transform data.

(2) *Information Retrieval for Research Data*

The first aspect is related to restrictions in regard to the usage of research data, such as data privacy, copyright, or legal protection.

Furthermore, the section contains questions on how to gain information on researchers, associated projects, or project groups.

Finally, the section contains some questions aimed at information which must be collected to give the opportunity for an adequate subsequent usage of research data within the same team or for extern researchers. Are there metadata schemes that must be considered?

(3) *Common Usage of Research Data*

This section dealt with the common handling of data usage. How are data used or transformed after creation? Which reuse approaches are imaginable, and which data will be free for reuse? Finally, what about "old" data?

(4) *Long-Term Storage/Publication of Research Data*

On the one hand, this section focused on default processes for storage and publication of research data in the CRC 840. On the other hand, questions regarding the goals of long-term storage, the authority to make decisions, and the selection of worthy data had to be answered.

(5) *Additional Wishes*

In the last section, the researchers had the opportunity to describe additional wishes for data management, long-term preservation, and meta-information retrieval within CRC 840.

The meaning of the questionnaire is twofold. First, it served as basis for the structured interviews with project leaders and researchers. Secondly, it will help during configuration of additional research fields (data objects) in the future.

3.2.1. Diversity

The range of research areas participating in the CRC 840 is very wide. Consequently, the research data accumulated in the individual sub-projects have diverging characteristics. On the one hand, the data stems from technical devices, such as a variety of spectrometers or microscopes, and is usually stored in vendor specific data formats. On the other hand, data is produced by computer programs, often programmed by the researchers themselves, producing arbitrary data formats. Usually, there are no supra-regional repositories available for the collection of data from these highly specialized research fields. A high level of abstraction in the description of research data through preferably generic metadata and the need to store a variety of data formats are crucial requirements for the local infrastructure. Therefore, the extension of the portals used to distribute publications could not be considered as a sustainable solution. The need for a specialized solution soon became clear. Moreover, the discussion with the researchers showed that data curation is a prerequisite for reasonable long-term preservation. As researchers have no experience in dealing with various data formats, they have to be supported in choosing the best data format and convenient descriptive data. Thus, the library became a partner in the project, due to its great experience in describing various objects with descriptive data. If a researcher wants to bring in data, the first step is to analyze the data format, assisted by the librarians. The aim is to determine the best format for both the long-term storage and the crucial descriptive data. This information is then transferred to the computer center, where the datasets for the descriptive data is technically implemented and, if possible, the automatic extraction of descriptive data from the primary data is implemented. It is obvious that this approach is very time-consuming and great efforts should be made to find generalized descriptions for similar data types. In addition, methods for the automated implementation of descriptive data are investigated.

3.2.2. Reproducibility

At the beginning of the project, we followed the approach to build a repository for the primary research data. Based upon the discussions with the project leaders, an extended set of requirements were identified. In addition to the primary research data itself, the published results, and their complete provenance should be preserved. From the point of view of a team or project leader, the portal should be a tool that allows the reconstruction of every published result, even when the researcher, who produced the result, has left the team. Therefore, the architecture must be chosen in a way that every step of the research process up to the published results can be described. Published results are furnished with a unique DOI. Based on the DOI, the complete genesis of the results may be reconstructed. This approach extends the need of data curation enormously, since for every processing step a set of descriptive data must be found and implemented in the system. This has to be done in close cooperation with the researcher, because a deep insight into the scientific processes is necessary. Moreover, it may be required to store programs that are used in the processing of the data. This leads to a very complex structure of linked components contained in the system.

This way of proceeding means that we enter the field of e-Research. Beyond the provision of a data store for the permanent storage of research data, we implement a tool to manage the provenance research data. We are now part of the lifecycle of research data, as shown in Figure 2. From this picture, one could think about the extension into the direction of the genesis of the data (*i.e.*, the implementation of an electronic laboratory journal). Various further ideas to support the researchers may arise from this consideration.

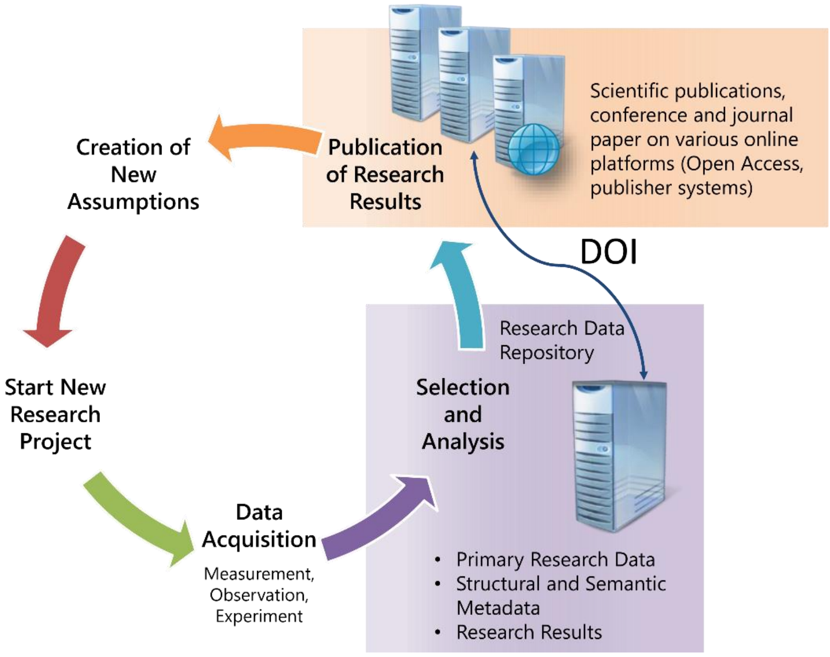

Creation of New Assumptions
Publication of Research Results
Scientific publications, conference and journal paper on various online platforms (Open Access, publisher systems)
DOI
Research Data Repository
Start New Research Project
Selection and Analysis
Data Acquisition
Measurement, Observation, Experiment
• Primary Research Data
• Structural and Semantic Metadata
• Research Results

Figure 2. Desired Lifecycle of Research Data in CRC 840.

3.2.3. Long-Term Preservation

In order to provide the reliable availability and sustainable usage of the data, the storage has to satisfy the norms of long-term preservation. Currently the most important standard is the Open Archival Information System (OAIS) reference model. The OAIS model specifies how digital assets can be preserved through subsequent preservation strategies. It is a high-level reference model, and therefore is not bound to specific technology. Although the model is complex, systems for the long-term storage of digital data will have to meet the requirements.

3.2.4. Authorization and Usability

Access to the data must be controlled by a flexible authorization in order to prevent unwarranted usage while still allowing reasonable collaborations. The storage of research data in a repository is often mixed with open access to the data. This is partially true only for published results. If the repository is seen as a tool that supports the researcher in the management of the research data and the reproducibility of the results, only a part of the data in the repository will be in open access. The rest of the data has to be protected, and only the researcher or a research group should be allowed access to it. To allow collaboration between researchers, the authorization must be organized in a very fine-grained way.

Finally, the solutions must be easy to use and bring advantages to the researcher. The process of data curation requires more caution in the managing of the research data. This effort must provide significant benefits for the researcher in terms of additional services, such as comfortably searching the data, collaboration, or tools for analysis. Moreover, the insertion of data must be convenient and the access rights administration must be easy to use.

4. Prototypical Implementation

Mainly based on the aforementioned challenges, a research infrastructure regarding scientific data within the CRC 840 was planned and implemented, as shown in Figure 3. In addition to these requirements, some further demands, namely in the form of some practical boundary conditions that arose, had to be taken into account.

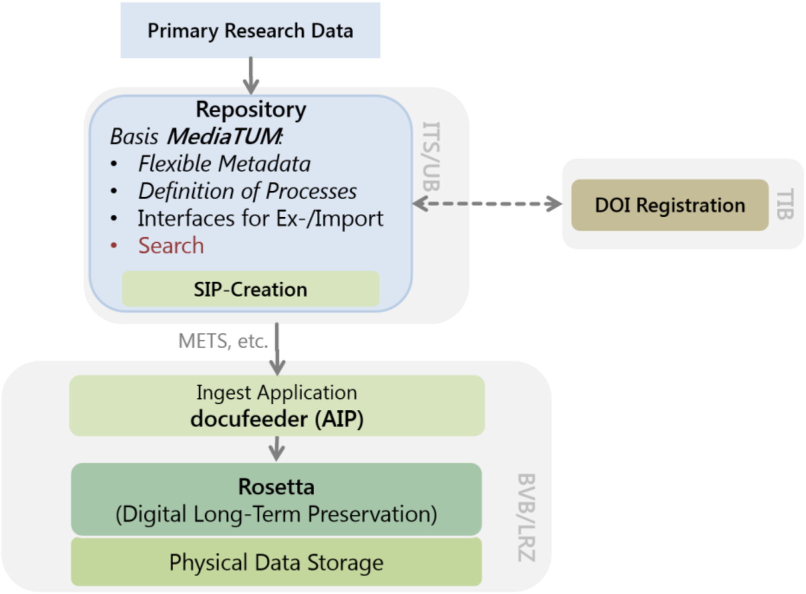


Figure 3. Implementation of Long-Term Preservation of Research Data in CRC 840.

Cooperation. The implementation is built substantially on already existing solutions. Therefore, the focus is on the cooperation between various partners. Regarding the strategies for the transfer of data into the long-term storage and the retrieval options for metadata, contacts were established with the TU Munich (use of their component MediaTUM) and the ETH Zurich (Application of Docuteam software). With regard to long-term preservation, an agreement exists with the Bavarian State Library (BSB) that permits the joint use of the installation of the "Rosetta" software by ExLibris.

Research Data Portal. The current approach describes all processing steps from the primary data to the published result as nodes with attached specific data and metadata. The framework MediaTUM is suited for this purpose as it allows a flexible definition of metadata schemata. The provenance of the research results (pictures, graphics, tables, and various data) has to be analyzed based on the primary data, with every step of the process needing to be abstracted as far as possible, and mapped as a node. The associated metadata schema and the data upload process have to be defined and implemented for each node.

Long-Term Retrieval. For each node of a published result, a DOI is assigned by the Technische Informationsbibliothek (TIB) Hannover, thus guaranteeing the global visibility of research results via DataCite. All nodes that are necessary for the result's reconstruction are automatically stored in Rosetta during the assignment of a DOI, ensuring both the long-term availability of the result and its provenance.

Data Protection. The protection from unauthorized access to the data is guaranteed by the authorization model of MediaTUM, which allows for a differentiated access control for each node. In addition to the login via Lightweight Directory Access Protocol (LDAP) at the local Identity Management (IdM), users might be set up in the system itself. A user might be assigned to flexible groups, allowing the mapping of cooperation between scientists.

4.1. Status Quo

The specific demands concerning the research process were evaluated by structured interviews and a questionnaire, and the answers were analyzed with regard to similarities and differences. Because of the wide range of research areas participating in the CRC, there is no uniform research process. Quite the contrary, researchers take diverging steps to gather research results and have various habits regarding the preservation and commenting of produced data, even within the same research area. For this reason, we are developing a prototype of the data repository for the CRC 840, based on the platform MediaTUM in the first instance. The current course of action for implementing this prototype contains the following steps:

(1) Analyzing existing publications
(2) Defining and implementing required object types
(3) Defining and implementing associated metadata schemata
(4) Testing upload for defined object types

4.1.1. Analyzing Existing Publications

In the first step, the results in the publications have to be located and classified (data, image, table, text, *etc.*). Then, their provenance must be traced back to the source—the associated primary research data sets. Figure 4 shows an overview of the analysis of a publication from the chemical sciences.

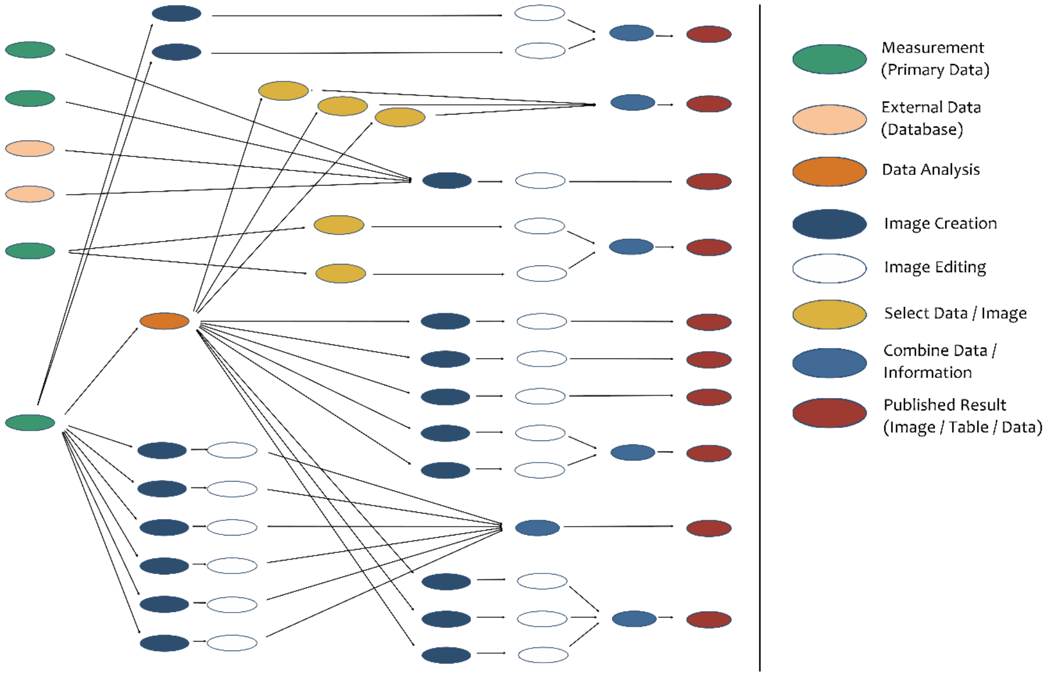

Measurement (Primary Data)
External Data (Database)
Data Analysis
Image Creation
Image Editing
Select Data / Image
Combine Data / Information
Published Result (Image / Table / Data)

Figure 4. Provenance of published research data for defined publication (Schematic Overview, You can see a complete version in Supplementary).

The details in Figures 5 and 6 show the genesis and relationships of some graphics used in the publication. Based on the measurement of the crystal structure of a specific substance with the so called "Einkristalldiffraktometer STOE IPDS", the researcher started the analysis of data with the help of the application "SHEIXTL 5.1" and gained a CIF-File containing all analysis information.

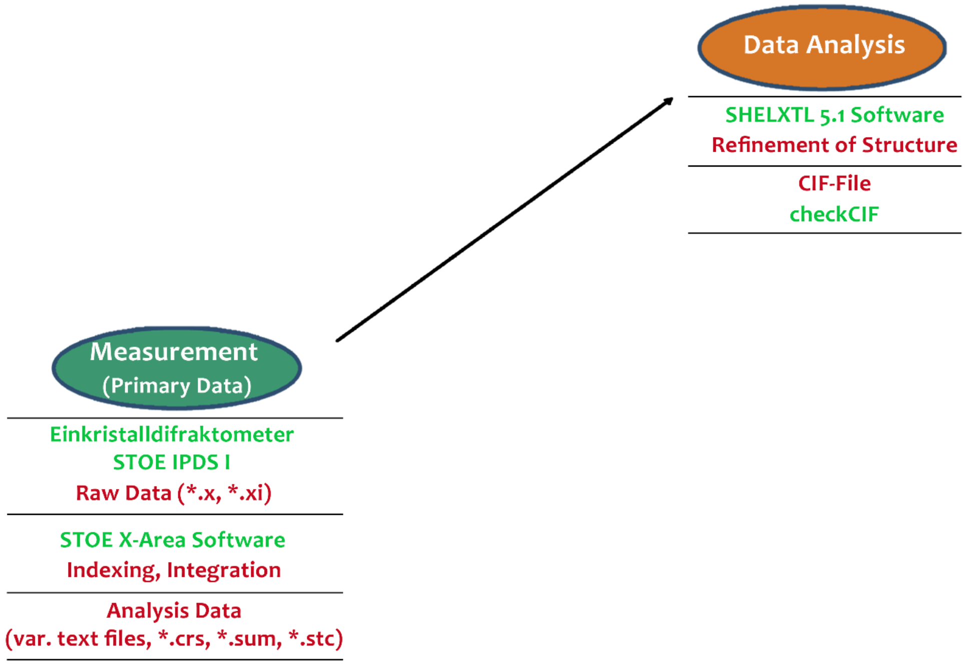


Figure 5. Provenance of published research data for defined publication (Detail: Step 1).

Subsequently, the application "Diamond v3" produced relevant graphics with the help of information from the CIF-File. Finally, these graphics had to be edited with a graphics tool and, as the case may be, combined to meet the publication's requirements.

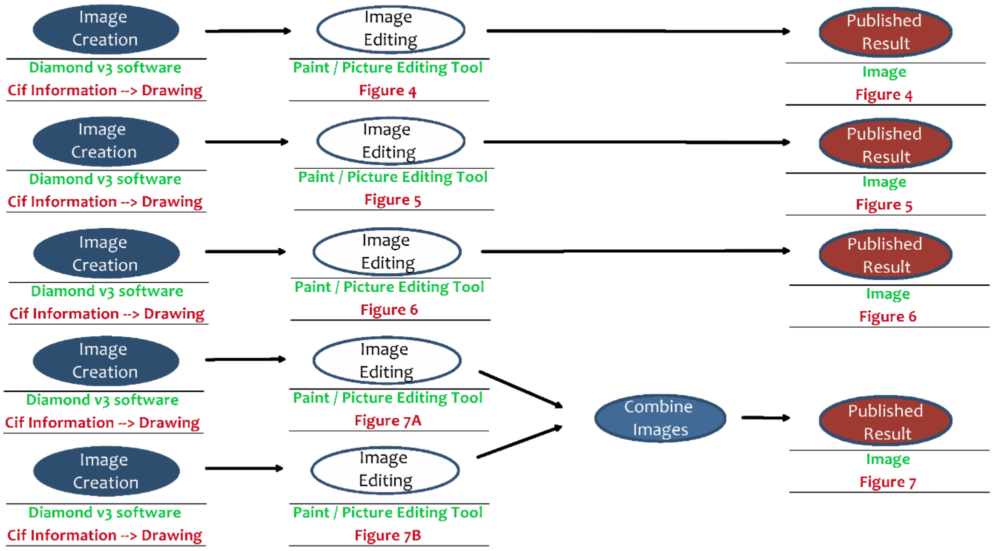


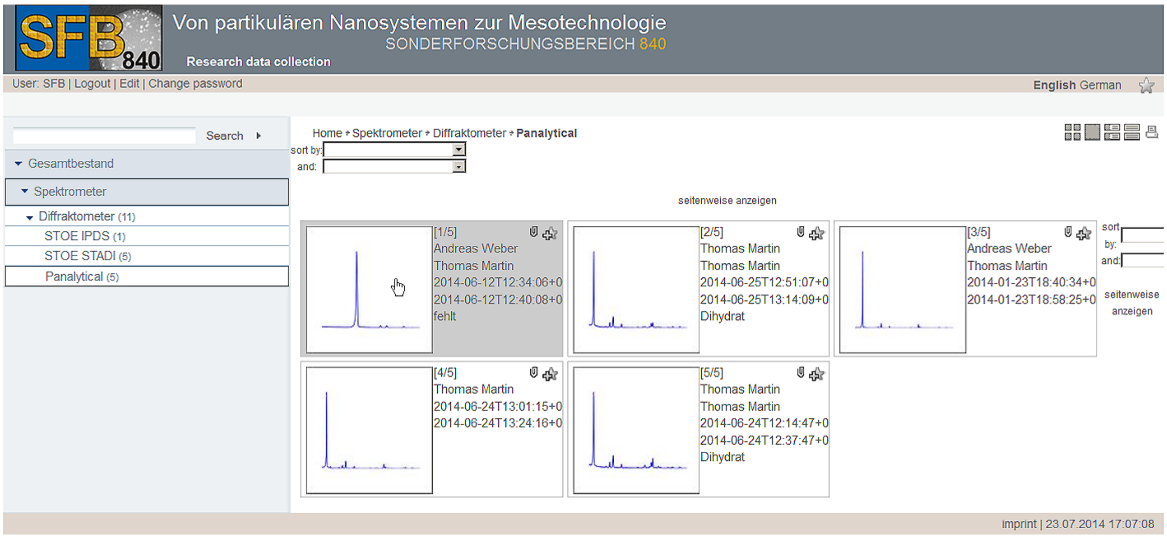
Figure 6. Provenance of published research data for defined publication (Detail: Step 2–5).

For the implementation of the prototype, three publications will be analyzed in this way to better understand the genesis of research results within associated research areas. As this course of action is very time consuming, it must be standardized in a productive environment. Therefore, the analysis of publications itself must be abstracted as much as possible.

4.1.2. Definition and Implementation of Required Object Types

The analysis of a publication reveals all associated nodes, which, subsequently, have to be stored in the data repository. For this reason, the nodes will be classified (primary data, transformation data, resulting data, *etc.*) into discrete types with the same or similar characteristics. On the basis of this classification, new object types are implemented in the data repository. The flexible structure of MediaTUM allows for the very easy creation of new object types. The creation includes the definition of standardized object attributes, the presentation of the object type, the kind of type (content, container), and the course of action while uploading data for this object type.

Figure 7 shows, for example, the presentation of data for the object type "Panalytical" as a list in small preview mode. The presentation mode was defined during the creation of the object type and includes a visual preview of spectral data (created while uploading measurement data to the repository), selected metadata information, and information regarding the responsible researcher.



[1/5]
Andreas Weber
Thomas Martin
2014-06-12T12:34:06+0
2014-06-12T12:40:08+0
fehlt

[2/5]
Thomas Martin
Thomas Martin
2014-06-25T12:51:07+0
2014-06-25T13:14:09+0
Dihydrat

[3/5]
Andreas Weber
Thomas Martin
2014-01-23T18:40:34+0
2014-01-23T18:58:25+0

[4/5]
Thomas Martin
2014-06-24T13:01:15+0
2014-06-24T13:24:16+0

[5/5]
Thomas Martin
Thomas Martin
2014-06-24T12:14:47+0
2014-06-24T12:37:47+0
Dihydrat

Figure 7. Presentation of Object Type "Panalytical" in the data repository.

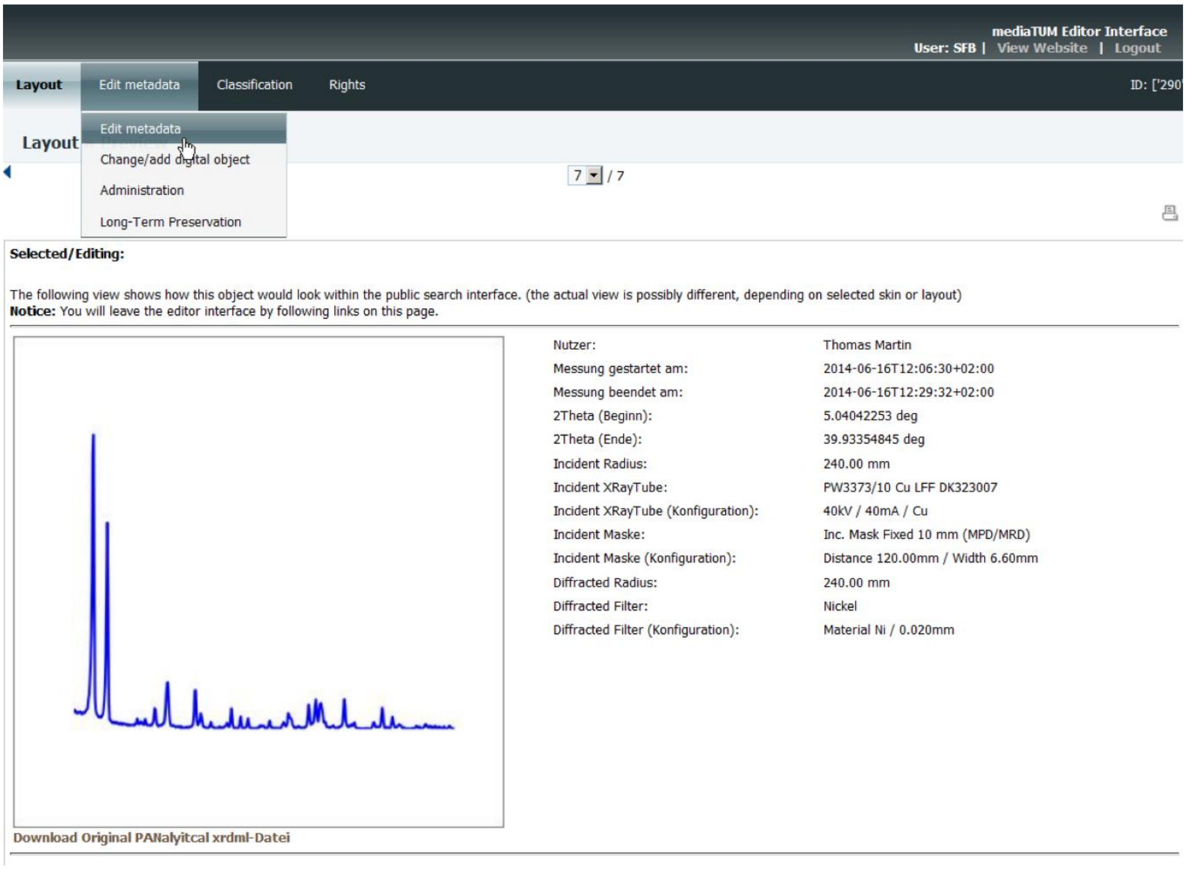
4.1.3. Definition and Implementation of Associated Metadata Schemata

A metadata schema is realized as a mask for displaying selected metadata values of a specific object type; whereas the actual metadata information is always completely stored in the database, the presentation of metadata can be adapted to the individual use case. For example, it is not required to show all metadata in an overview. However, when editing information, it is necessary to see all editable metadata fields. Thus, every object type has various metadata schemata associated with different use cases. Even the subsequent change of a metadata schema has an effect on existing data for the particular object type.

After the definition of new object types, the desired metadata schemata have to be created as follows:

(1) Defining all necessary metadata fields
(2) Creating desired metadata schemata
(3) Assigning metadata fields to the particular schema
(4) Definition of access rules/access rights for groups and/or user

Figure 8 shows the overview of metadata information for the object type "Panalytical" in the administration area of the data repository. If the user has the correct rights, it is possible to view and edit the metadata information of the current data object.



Selected/Editing:

The following view shows how this object would look within the public search interface. (the actual view is possibly different, depending on selected skin or layout)
Notice: You will leave the editor interface by following links on this page.

Nutzer:	Thomas Martin
Messung gestartet am:	2014-06-16T12:06:30+02:00
Messung beendet am:	2014-06-16T12:29:32+02:00
2Theta (Beginn):	5.04042253 deg
2Theta (Ende):	39.93354845 deg
Incident Radius:	240.00 mm
Incident XRayTube:	PW3373/10 Cu LFF DK323007
Incident XRayTube (Konfiguration):	40kV / 40mA / Cu
Incident Maske:	Inc. Mask Fixed 10 mm (MPD/MRD)
Incident Maske (Konfiguration):	Distance 120.00mm / Width 6.60mm
Diffracted Radius:	240.00 mm
Diffracted Filter:	Nickel
Diffracted Filter (Konfiguration):	Material Ni / 0.020mm

Download Original PANalyitcal xrdml-Datei

Figure 8. View and edit metadata in the data repository.

4.1.4. DOI Assignment

One of the core requirements of the sub-project "INF Z2" is to make data citable and re-usable by generating public access to defined data. This happens by setting a DOI for each data object that is worthy to be cited. With this persistent identifier, data objects are referenced non-ambiguously. The responsibility for the issuing of a DOI falls to the researcher and has to be done in particular consideration of the quality conditions for DOIs and the TIB DOI contract. As a consequence, the data object will afterwards be undeletable from the repository and write protected, while the DOI is existing. Necessary changes will subsequently only be allowed to administrative users. The process of the assignment of DOIs has to be manually started by a user of the system. It is not an automated process for every data object uploaded to the repository, whereas the update of changed meta-information for a DOI will be automated. To guarantee that the conditions of the DataCite metadata schema are met for every data object metadata schema, there will be a matching between metadata fields of the data object and the DataCite metadata schema. As described in the implementation chapter, the metadata schemes for new data objects and the assigned upload process have to be manually configured. Thus, we guarantee the flexible usability for various research fields in the future with the drawback of some manual configuration. The benefit is high individualization and adaptability.

The DOI consists of two parts: the identical prefix for the whole data center, and an individual suffix for the explicit data object. This individual suffix is generated by the web service and is unambiguous throughout the data center.

The implementation of the requested functions occurs as a web service application, which can be utilized by the research data repository when necessary. It offers interfaces for requesting new DOIs, updating DOIs, metadata, or media, and deleting DOIs. In the background, the web service uses the DataCite RESTful API for data centers [21].

4.2. Synergistic Effects

In addition to the long term preservation of research data, there exists the requirement to save digitalized data (such as scanned slides, images, audio or video records, or digitized originals) from archives and libraries. Due to the ability to create flexible object types and associated metadata schemata, our approach can easily be adopted to serve as repository for digitalized data. As a side effect, we are on the way to implement a platform for digitalized plant slides.

Furthermore, there are additional advantages stemming from the described project regarding the assignment of DOIs. Access to the implemented web service for requesting DOIs can be provided throughout the whole university, and can be utilized by other research projects or IT services.

5. Outlook

The traditional concept of long-term preservation of research data mainly covers the physical storage of either primary data or complete publications individually enhanced by descriptive information. Consequently, concepts for models and supporting infrastructure are divided regarding these two points of view.

In contrast, our approach goes beyond these divisions and generates an integrated system. The complete research process will be supported regarding the preservation of data and information. Here, our data repository offers the storage of primary data, the preservation of transforming activities, the publication of research results, and the enhancing of all data objects with describing information. Thus, the provenance of research results is traceable, and therefore the quality of research can be assured.

The next challenge in our current work is the implementation of the relations between different data objects. Furthermore, suitable ways for the representation of these relations have to be implemented in the user interface. Unfortunately, because of the design of the underlying system (MediaTUM) there are restrictions in representing and displaying different relations of nodes. Thus, a further challenge will be to examine adequate ways to integrate external applications with the repository user interface.

Acknowledgments: This research was supported by the Deutsche Forschungsgemeinschaft (DFG) in the scope of the sub-project "INF Z2" of the Collaborative Research Centre (CRC) 840.

Author Contributions: Andreas Weber and Claudia Piesche designed and implement the infrastructure that permits long-term preservation and retrieval of research data created within the CRC, and wrote this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. OECD: OECD Principles and Guidelines for Access to Research Data from Public Funding. Available online: http://www.oecd.org/science/sci-tech/38500813.pdf (accessed on 27 October 2014).
2. Effertz, E. The funder's perspective: Data management in coordinated programs of the German Research Foundation (DFG). In *Proceedings of the Data Management Workshop*; Curdt, C., Bareth, G., Eds.; University of Cologne: Cologne, Germany; pp. 35–38.
3. HGP: The Human Genome Project. Available online: http://www.genome.gov/10001772 (accessed on 27 October 2014).
4. The Human Genome Project (Archived Information). Available online: http://web.ornl.gov/sci/techresources/Human_Genome/index.shtml (accessed on 31 March 2016).
5. GRBIO: The Global Registry of Biorepositories. Available online: http://grbio.org (accessed on 27 October 2014).

6. Kümmel, C. Nach den Sondersammelgebieten: Fachinformationen als forschungsnaher Service. *Z. Bibl. Bibliogr.* **2013**, *60*, 5–15. [CrossRef]
7. Mittler, E. Nachhaltige Infrastruktur für die Literatur- und Informationsversorgung: Im digitalen Zeitalter ein überholtes Paradigma—Oder so wichtig wie noch nie? *Bibl. Forsch. Praxis* **2014**, *3*, 344–364. [CrossRef]
8. Deutsche Forschungsgemeinschaft (DFG). Scientific Library Services & Information Systems—Funding Priorities through 2015. Available online: http://dfg.de/download/pdf/foerderung/programme/lis/pos_papier_funding_priorities_2015_en.pdf (accessed on 28 October 2014).
9. Ludwig, J., Enke, H., Eds.; *Leitfaden zum Forschungsdaten-Management—Handreichungen aus dem WissGrid-Projekt*; Verlag Werner Hülsbusch: Glückstadt, Germany, 2013.
10. Treloar, A.; Harboe-Ree, C. Data Management and the Curation Continuum: How the Monash Experience is Informing Repository Relationships. Available online: http://valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf (accessed on 20 August 2014).
11. Hunter, J. Scientific publication packages—A selective approach to the communication and archival of scientific output. *Int. J. Digit. Curation* **2006**, *1*, 33–52. [CrossRef]
12. RADAR—Research Data Repository. Available online: https://www.radar-projekt.org/display/RE/Home (accessed on 31 March 2016).
13. RADAR—Research Data Repositorium. DFG-Antrag. Available online: http://www.radar-projekt.org/display/RD/Projektantrag (accessed on 29 October 2014).
14. Razum, M.; Neumann, J. Das RADAR Projekt: Datenarchivierung und -publikation als Dienstleistung—disziplinübergreifend, nachhaltig, kostendeckend. *Ver. Dtsch. Bibl. (VDB)* **2014**, *1*, 30–44.
15. Potthoff, J.; van Wezel, J.; Razum, M.; Walk, M. Anforderungen eines Nachhaltigen, Disziplinübergreifenden Forschungsdaten-Repositoriums. Available online: https://www.dfn.de/fileadmin/3Beratung/DFN-Forum7/konferenzband/02-Anforderungen_eines_nachhaltigen__disziplinuebergreifenden_Forschungsdaten-Repositoriums.pdf (accessed on 5 August 2014).
16. Kraft, A. RADAR—A Repository for Long Tail Data. Available online: http://docs.lib.purdue.edu/iatul/2015/mrd/1 (accessed on 31 March 2016).
17. Woutersen-Windhouwer, S.; Brandsma, R. Enhanced Publications, State of the Art. In *Enhanced Publications—Linking Publications and Research Data in Digital Repositories*; Vernooy-Gerritsen, M., Ed.; Amsterdam University Press: Amsterdam, The Netherlands, 2009; pp. 19–91.
18. Royal Society of Chemistry. Project “Prospect”—Linking Compounds and Concepts in Articles. Available online: http://www.rsc.org/Publishing/Journals/ProjectProspect/index.asp (accessed on 29 October 2014).
19. Büttner, S.; Hobohm, H.-C.; Müller, L. Research data management. In *Handbuch Forschungsdatenmanagement*; Büttner, S., Hobohm, H.-C., Müller, L., Eds.; BOCK + HERCHEN Verlag: Bad Honnef, Germany, 2011; pp. 13–25.
20. Collaborative Research Centre 840: “From Particulate Nanosystems to Mesotechnology“: Focus and Approach of the Collaborative Research Center SFB 840. Available online: http://www.sfb840.uni-bayreuth.de/en/index.html (accessed on 23 November 2015).
21. DataCite. DataCite API v2 for Datacentres, API Documentation. Available online: https://mds.datacite.org/static/apidoc (accessed on 31 March 2016).