Lessons Learned and Open Challenges Regarding Support for Data Management Plans and Research Data Management

Heike Görzig
University of Hagen,
Faculty for Mathematics and Computer Science
Universitätsstrasse 1
D-58097 Hagen, Germany
+49-2331-987-304
Heike.Goerzig@fernuni-hagen.de

Felix Engel
University of Hagen,
Faculty for Mathematics and Computer Science
Universitätsstrasse 1
D-58097 Hagen, Germany
+49-2331-987-304
Felix.Engel@fernuni-hagen.de

Holger Brocks
InConTec GmbH
Kirschenalle 7
D-96152 Burghasslach, Germany
+49-9552 931494
Holger.Brocks@incontec.de

Matthias L. Hemmje
University of Hagen,
Faculty for Mathematics and Computer Science
Chair of Multimedia and Internet Applications
Universitätsstrasse 1
D-58097 Hagen, Germany
+49-2331-987-304
Matthias.Hemmje@fernuni-hagen.de

ABSTRACT

This paper outlines an approach for developing tools and services that support automated generation, management, evolution and execution of data management plans (DMPs) by generating rules derived from the DMPs which can be applied to the data to be archived. The approach is based on existing models and tools that were developed in successive research projects SHAMAN, APARSEN, and SCIDIP-ES. The models include the Curation Lifecycle Model from the DCC, the OAIS Information Model and the Extended Information Model to support processes, domains, and organizations. An approach for deriving rules from policies is outlined to support using iRODS. OAIS and Context Information related to a data object is supported in a serialization using the OAI-ORE format.

General Terms

Frameworks for digital preservation

Keywords

SHAMAN, APARSEN, SCIDIP-ES, DMP, RDM, OAIS, OAI-ORE, data curation, automation, data management policies

1. Introduction and Motivation

In the Integrated Project SHAMAN that was funded by the European Commission in its Framework Program 7 (FP7) an *Archive-centric Information Lifecycle Model* (ACILM) had been introduced which conceptually supports pre-ingest and post-access activities by adding additional context information to Information Packages [1].

Building on this model and on related technical results of SHAMAN as well as on conceptual results from another FP7 project called APARSEN[2] a set of software tools had been developed in SCIDIP-ES[3]. The tools can assemble the required context and can package this context as provenance information together with the digital object itself as Information Packages, ready for submission, ingest, and archiving. Nevertheless, some remaining challenges regarding assembling provenance and authenticity information have been identified in one of the final reports of the project [4]. For example, higher usability of the preserved data can be ensured by establishing *Data Management Plans* (DMPs). These and related preservation policy processes ideally need to be defined at the beginning of a research project. In this way, preservation policies can be created much earlier than at production, assembly, and ingest time [4]. These preservation policies are then either created in isolation or in the context of an overall DMP. In many projects such DMPs are formally required which is, e.g., more and more the case in almost all public-funded research projects.

Funding agencies very often are requesting to make the research data generated in funded projects available for re-use in the future and therefore are demanding to elaborate DMPs already at proposal or at least at research-fund contracting time. To comply with this pre-requisite the DMPs have to include the archiving and preservation policy of the produced data of the project.

In order to maintain the archived data in an intelligible and interpretable way over a long period of time after the end of each such project, the generated data needs to be continuously enhanced with information about its production and usage context. The context to be preserved includes all known properties of the digital object and all operations carried out on it [5]. This includes the phases before ingesting the digital object to the archive and after accessing it. Within the preparation of a research project an initial production and use context can be foreseen and planned but during its execution the research process bears risks and uncertainties that can only be handled in a dynamic way when they appear. Therefore, on the one hand, DMPs describe the initial concepts in which the digital objects and their context need to be archived and preserved but on the other hand the DMP has to further evolve during the execution of the project. Therefore, also the initial production and use contexts and their related concepts have to evolve within the corresponding DMP.

Part of the production and use context is contained in the knowledge base of the designated community which can also change very quickly and unexpectedly [4]. Therefore, such context has to be identified, represented, added and maintained by three main actors in DMP context management. Typically, these actors are data producers, data managers, and data consumers. Adding DMP Context Information to data provenance information is usually a time-consuming, intellectual, i.e., human and manual process which is normally performed by the data managers. While working on this task, the data managers are also responsible for ensuring that the DMP's overall requirements are met. In large-scale projects and after the end of research projects the manual curation of this data might therefore become too costly or even impossible. Therefore, in order to achieve a more sustainable situation and working environment, the role of the data manager will have to be supported by appropriate tools and management processes. This means, that automating this work wherever possible has to become a goal of prime importance. To achieve this, automated curation and corresponding DMP support would have to incorporate all facets of context of the data object and respectively the evolvement of the context within activities of the data objects usage.

A DMP provides the concepts for archiving and preserving digital objects and also for preparing their potential re-use. It can be utilized to support automatic or at least semi-automatic contextualization. To support automation by means of applying Semantic Web technologies in this area of automation, any DMP needs to be supported by machine-readable semantic representations which are governed by an appropriate domain ontology. Within the context of our earlier work it has also been shown that preservation policy generation and DMP should be decoupled from the necessity to have knowledge of OAIS in order to support researchers in concentrating on the data in their field of expertise and scientific discipline. In this way, researchers should become free from the burden of having to know OAIS [4]. Therefore, a DMP can be seen as a dynamic document during a research projects life-time, it is evolving and needs to be adapted to changing needs.

In the following, we will first outline and analyze the requirements and challenges of the DMP domain in more detail in order to better explain the requirements and challenges of such an automated DMP support approach.

2. Overall Requirements and Challenges of Data Management Planning

As a basis for this identification of overall requirements and challenges, we will review the initial DMP of a very large research project that is funded by the European Commission: the so-called *Realizing an Applied Gaming Ecosystem* (RAGE) project that has just been kicked off and has made its DMP available to us for this initial analysis.

In *Research and Development* (R&D) projects like, e.g., RAGE, three roles or user stereotypes that are involved in *Research Data Management* (RDM) can be identified. These stereotypes span three dimensions that the DMP has to address. There is the *Formal Dimension* with project administration, the *Managerial Dimension* with project management and the *Operative Dimension* with project implementation and execution.

The Formal Dimension of DMP is spanned by the funding agencies' grant agreements (GA), corresponding laws and policies. The GAs usually provide the contractual framework for the DMP, specifying what the DMP has to accomplish and to comply with. Corresponding laws and regulations provide the legal, regulatory and policy-building framework. Alongside these contractual and legal specifications and requirements, corresponding DMPs have to be elaborated in compliance with all of them. In the case of our exemplar EU-funded project they have to follow the Horizon 2020 policies [6][7].

In the GAs, funding agencies mostly state that the DMP will, e.g., also have to comply with ethical guidelines, establish institutional and local procedures, specify the instruments for data collection, etc. The GA usually also refers to laws and regulations that will have to be fulfilled.

Project administrators in the back office usually study all these GA documents and corresponding requirements and challenges of the DMP specifications and have to extract a set of corresponding requirements and challenges and a corresponding representation schema of related constraints, targets, and activities which the project has to accomplish. For R&D project data access rights, duration of archiving, purpose of archival, sharing, and preservation policies according to the GA, policies and laws are formulated and specified. The Managerial Dimension uses the requirements and challenges schema to create the initial DMP.

To comply with the requirements and challenges created by the analysis of the formal DMP dimension, a RDM work plan is developed in the Managerial Dimension of the DMP. The RDM work plan describes the RDM scenario that has to be created to comply with the DMP requirements and challenges and their corresponding representation schema set up by the analysis of the Formal Dimension. This RDM work plan includes strategic and organizational aspects, concrete activities, and deliverables. In the RDM work plan sequences of activities and their dependencies are formulated. The implementation of the DMP is based on this RDM work plan. In the Managerial Dimension, quite often user stereotypes of a project coordinator, work package leader and task leader can be found.

The research project's R&D work plan is usually divided into work packages and is spread over various working groups. The work packages have organizational dependencies between each other; these can be dependencies on developed knowledge, results, deliverables, and experiences that will have to be shared between the working groups. These dependencies will be reflected in the work plan and will have to be defined in the DMP. Therefore, the creation of the DMP, e.g., needs to foresee communication and exchange strategies between the work package leaders. In analogy to the dependencies between work packages, there are also lower-level organizational dependencies on the level of tasks and activities within work packages. These tasks and activities will be carried out in working groups or other organizational entities within these working groups. In the R&D work plan the activities will have a time span assigned. In order to create the RDM work plan as part of the DMP, the project coordination has to work closely with the work package leaders, who are working together with the task leaders and so on. In each organizational layer of the R&D work plan activity that has to be performed, compliance with the GA and its corresponding DMP has to be achieved dynamically at the corresponding level of detail.

The creation of the DMP and its execution with the RDM work plan is a collaborative task. Between the work packages a consensus about dependencies, data management services and activities, needed sharing services and capacities will have to be achieved. Furthermore, the corresponding RDM will have to manage the *Intellectual Property Rights* (IPR) and corresponding access rights to project results and background and as well as data sharing policies in compliance with the constraints provided in the Formal Dimension. These IPR dependencies, access rights and sharing policies will have to be defined in the DMP and will have to be applied during RDM work plan execution when the data is finally generated, managed, archived and preserved.

The data are finally generated in the Operative Dimension, from this dimension the finest granularity of Context Information about the data to be generated will originate and will find its way into the RDM corresponding to the initial DMP.

The data producers who are, e.g., software developers and researchers in the project, form the Operative Dimension of DMP. Tasks and activities listed in the work plan are executed by them and thereby produce and use the data to be archived and preserved.

Staff working in this Operative Dimension contribute their specific input to the DMP and corresponding RDM activities. They will have the most concrete and operational information about the data to be produced and used and will be able to provide information about where the data is stored, data types, archive and file sizes, formats etc. Data generators will also be able to provide information about dependencies and relations between generated data. Information about relations between source code, binary code and application is also retrievable in this dimension.

In this dimension the produced and used data will have to be connected to its descriptive information in relationship to the specific knowledge of the R&D domain it has been produced for and used in. Therefore, produced and used data depends on the research domain, but also on other potentially related information already listed in the DMP. Deriving this knowledge from the input, e.g., researchers or developers have been providing in the planning phase will have to be added as descriptive information to the produced and used data.

After the digital object has been submitted, archived, and preserved, other users might later want to access and re-use the data and may add additional re-use Context Information to the digital object.

Information needed from the above described dimensions will have to be collected, managed, and finally packaged, ingested, archived, stored, and preserved. Therefore, respective tools will need to be developed. In order to introduce and analyze this overall set of problems in more detail, the related scientific challenges and technical requirements for these tools will now be described.

3. Scientific Challenges and Technical Requirements for DMP and RDM Support

The user interfaces needed for such support tools depend on the DMP dimension as well as on the user stereotypes, roles, and the type of activity users are performing.

The Functional Dimension has to create a validation schema against which the DMP can be validated. Administrators in the back office have to be enabled to formulate, e.g., IPR, access rights, storage requirements, archival, preservation, and sharing policies for data to be produced and used by the project. This schema is based on the GA as well as on corresponding laws, regulations, and policies.

Later, when the DMP is created in the Managerial Dimension, its validation has to be possible using the created RDM work plan as a schema and validation errors must be made visible. For creating the DMP in the Managerial Dimension, a first RDM work plan has to be developed. As in the R&D work plan sequences of activities and their dependencies are formulated in the DMP and its RDM work plan. Therefore, a set of interfaces is needed to support RDM activity creation and interlinking. The RDM work plan finally shall result in a valid, i.e., formally fully complying implementation of the DMP, resulting in depending on the project GA, different schemas for the RDM application and in different user interfaces supporting these processes.

In the Organizational Dimension different DMP and RDM user interfaces depending on the research domain will have to be created. The Organizational Dimension will also need access to the DMP and RDM interfaces where activities are created and edited. These DMP and RDM interfaces should allow the linkage of R&D domain data to R&D activities.

R&D data users and producers might have to add additional metadata to the digital objects. In addition, R&D data which has already been produced and used in another working group has to be accessible to potential "re-users".

As the creation of the RDM work plan that is complying with the DMP is a collaborative task, corresponding user interfaces for collaborative DMP and RDM activities are needed. R&D data producers have to coordinate with the R&D data users, when, who and what exactly has to be delivered. R&D as well as RDM tasks will have to be submitted to the data producers.

To describe the time schedule documented in the R&D, as well as in the RDM work plan, a sequential workflow needs to be modeled where the work packages are producing digital objects. A digital R&D data object is produced in a certain activity/task in a work package. This digital R&D data object might be needed as a resource in another activity/task. The digital R&D data object which will be a resource in an activity/task has to be produced in a preceding R&D or RDM activity/task, thereby creating dependencies between activities/tasks. As a consequence, the sequence of R&D and corresponding RDM activities/tasks, and as well as the dependencies between these activities/tasks will need to be expressed.

The activities/tasks in which the digital R&D data objects are created or used, will be performed by resources. These resources are part of an organizational structure. This organizational structure will be another part of the digital R&D data object's context information.

Finally, the activities/tasks in which the digital R&D data objects are produced as well as the digital R&D data objects themselves are specific to a certain R&D domain. In order to describe an activity/tasks and a digital R&D data object, R&D domain-specific vocabulary will be needed.

These different types of information will have to be combined in a way that the DMP and corresponding RDM can be adapted and maintained from this information. Furthermore, it needs to ensure that the digital R&D data object which was produced and used can be archived together with its production and usage context as provenance information. This has to be achieved in a sustainable way which allows automating future access and re-use activities.

4. Architecture, Data Modeling and System Distribution Challenges

Users, creating the DMP and the RDM work plan and producing and using R&D data are usually based in different locations within different organizations but they all need access to commonly produced, used, and shared R&D data. Part of the R&D data will be stable and will not change very much during the duration of the R&D project but especially in the planning phase of the R&D data production and usage collaborative work is needed and R&D as well as planning data will have to be interchanged very frequently. Depending on the user profile and roles, different DMP and RDM services and related data types and distribution models are needed.

There will also have to be different R&D, DMP, and RDM data types to be stored which are the digital R&D data objects and their R&D, DMP, and RDM context data. This data will have to be accessed by the DMP and RDM support tools. Some of the data will have to be stored in a central place but there are also others types of data that have to be submitted from a local system and later stored in the central system when they are ready to be uploaded.

The architecture of the system to support the creation and realization of DMPs and corresponding RDM work plans, needs to address the above mentioned challenges. For expressing the knowledge in DMPs and RDM work plans, an ontology and its vocabulary will have to be developed, as well as a schema that can support the creation of Information Packages based on this DMP ontology. As the development of a DMP and RDM involves actors of the three Dimensions, a structure for collaborative development and execution needs to be created, for example defining who can decide what in a DMP and how decisions are made.

Building on existing ontologies that represent activities in processes, domains and organizations, an ontology will have to be developed that combines these ontologies with the *Open Archival Information System* (OAIS, ISO 14721)[8] Information model for *Long Term Archival* (LTA) and *Digital Preservation* (DP).

On the basis of these DP models the respective user interfaces can be created. The system architecture will have to be created respecting the distributed and collaborative work, offering the mentioned features as a service. In terms of storage a local storage for active work and a centralized storage for archiving will have to be considered.

Policies described in the DMP will have to be formulated in a formal way to support the overall automating of the application of these policies within RDM activities.

5. Scientific and Technical State of the Art

Many funding agencies require the development of a DMP. The DMPs are very often part of the GA [9] [10]. The DMP aims to help organize the created data, by preparing storage so that created data can be submitted according to a planned procedure in order to find them when needed and can later be referenced. A DMP helps to maintain data integrity and avoid creating duplicates. DMPs also include archiving of information, which makes digital objects understandable and retrievable [9] [10].

There are different categorizations of the contents of DMPs. *Data Archiving and Network Services* (DANS)[7] identifies five [7]:

- Administration Information
- Data description
- Standards and metadata and everything else that is required to find and use the data
- Ethics and laws
- Storage and archiving

Information about time of collection and changes to the data also will have to be added. It might be necessary to justify the decision for a certain format, especially if it is a proprietary format, as, e.g., open access is in many funding agencies DMPs and corresponding policies required. It might also be expected to describe the relation and added value to existing data [9]. The sharing of the data might be restricted due to IPR, privacy concerns, or embargos. These restrictions will have to be outlined for the created data. For sharing and reuse of the data, information about which data will be shared with whom, who might be potential data users, it has to be stated when, how and where the data will be available and how the data will be licensed. Two aspects of data storage should be explained: short-term data storage, mostly locally, within the institution of the research project and long-term storage. For the later it needs to be explained, which data will be preserved, how the data will be preserved, including formats and technologies used. Budget and security issues might also be specified in the DMPs [11].

Many research institutions and funders are offering guidelines and templates for developing DMPs. More detailed help can be found in institutions that specialize in the development of DMPs. Some of these institutions do also offer some support tools for creating DMPs.

There are funding agencies that require periodical creation of DMPs, while others only request a DMP once [12]. Some funders ask for the DMP before the project starts, while others require the plan during project runtime. A DMP also includes information about how data will be managed and about policies to be applied. This will be discussed in the next sections.

The OAIS reference model is a widely accepted model for archiving digital objects. It consists of a functional model explaining needed functional entities to perform LTA and support DP. Furthermore, it provides an environment model describing involved actors which are data producers, consumers, and management, and it provides an information model for the structure of an Information Package that contains all data necessary to find, access, provide authenticity and the representation information to understand the archived data [8].

In OAIS, a digital object is interpreted using its representation information, by the so-called *Designated Community* (DC). The representation information itself is an information object and thus subject to representation information, the assignment of representation information is regressive until the assumed level of knowledge of the DC is reached. Over time the knowledge base of a DC can change, putting thereby the interpretability of a digital object at risk [8].

Parts of OAIS' functional model are the preservation planning functional entity and the access functional entity. The preservation planning functional entity supports recommendations and provides preservation plans to make sure that the information stored in the OAIS remains accessible and understandable over a long time to the DC [8]. The access functional entity provides services and functionalities that support users to discover, find and access digital objects.

Brocks et al. criticize the OAIS for leaving all responsibilities to what happens before digital objects enter an archive and after it leaves the archive to abstract stereotypes as producers and consumers. Important Context Information is not considered such as, for example, reviewing criteria in the process of scientific publishing [5].

The *Archive-centric Information Lifecycle Model* (ACILM) (Figure 1) developed in the project *Sustaining Heritage Access through Multivalent ArchiviNg* (SHAMAN) [1] can overcome this constraint and support the activities executed on a digital object during its life-span including the phases before and after archiving. The phases are creation, assembling, archiving, adoption, and reuse, where creation and assembling comprise the pre-ingest phase and adoption and re-use the post-access phase.

The creation phase involves a multitude of information describing, e.g., among other information the background of the data creation. In this phase so-called *Context Information* (CI) can be added to the digital object. The second phase when context is added to the digital object is the adoption phase, where the digital object can be re-contextualized; adding, for example, consumer information [5].

The creation of the digital object is based on the R&D work plan, the DMP and the RDM. In the assembly phase all information to meet the presumed needs of the designated community is assembled. In the archival phase policies concerning ingest, preservation and access are applied [1]. In the adoption phase the digital object received as an Information Package will be enhanced with process information as, e.g., representing examination, adaptation, and integration to enable understanding and re-use. The re-use of an object implies the dissemination and exploitation of an object and eventually transforms it or creates a new object. Adoption and re-use of a digital object can be subject to a research project's work plan and therefore underlay a set of research policies and rules. The OAIS information model has thus been extended.

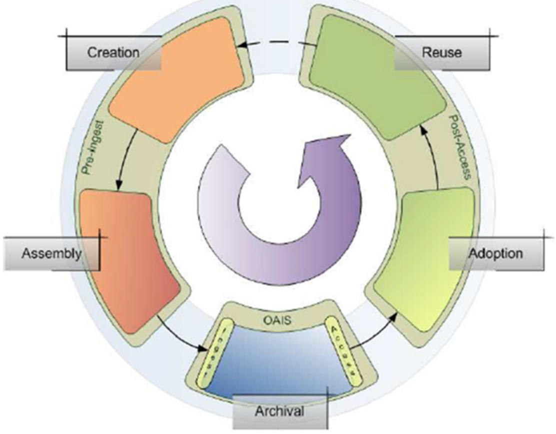


Figure 1 Information Life Cycle Phases[5]

The context of a digital object to be preserved over time comprises the representation of all known properties associated with it and of all operations that have been carried out on it. This implies the information needed to decode the data stream and to restore the original content, information about its creation environment, including the actors and resources involved, and information about the organizational and technical processes associated with the production, preservation, access and reuse of the digital object [5].

The context has been integrated into the OAIS Information Model without altering the concepts of its original information model [5].

The so-called *Extended Information Model* (EIM, see Figure 2) consists of the so-called *Context Information Package* (CIP) and the OAIS Information Package, sharing packaging information and package description. Additionally references exist to provenance, context and representation information.

Separating the context from the OAIS Information Package will allow for modeling the changes of concepts and terminology over time, characterizing production and (potential) reuse environments, and facilitates transfer to different communities by providing mappings of the underlying structured representations of concepts and relations [5].

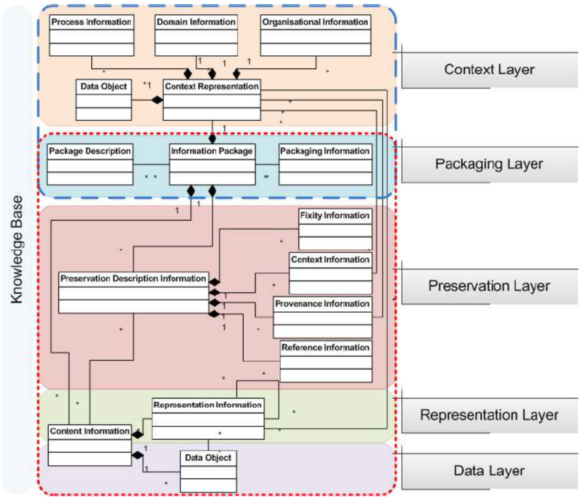


Figure 2 Extended Information Model [5]

The context representation consists of:

i) Process information
ii) Domain information
iii) Organization information

A context model has been created which can represent the information needed to describe the context of a digital object.

This model is based on the use of ontologies. The above introduced context dimensions i) can be represented by the use of domain ontologies, for ii) enterprise ontologies can be used and iii) can be described by process ontologies, where processes are divided in sequences of activities. The domain ontology defines the concepts and topics, but also their relations which are relevant for a particular application domain designated community. The enterprise ontology models the structural layout of organizational environments, such as affiliations, persons, or roles for describing a set of relevant concepts. The semantic classification of processes and activities as their building blocks requires their formal, hierarchical representation and description within an ontological structure [5]. Using the domain and the enterprise ontology rules can be specified as there are pre- and post-conditions, roles and interdependencies [13].

Brocks et.al explain the possibility of using OAI-ORE to develop ontologies that extend the OAIS information model in order to take into account Pre-Ingest and Post-Access processes much more than the OAIS suggests [5].

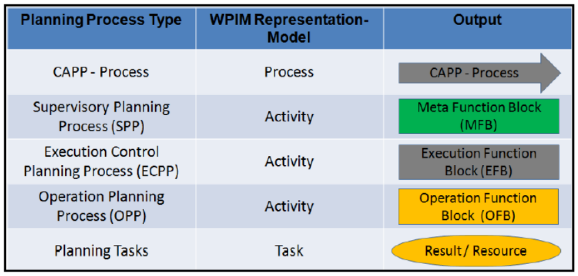
The *Extended Process Model* (EPM) integrates domain, enterprise and process ontologies into a conceptual unified process model [14]. This process model is meant to be applied in knowledge intensive processes with weakly structured activities, where the environment is dynamic and the process behavior and the entity concepts involved are unpredictable at design-time [14]**.** In this case traditional business process models with nearly static processes where the sequence of activities does not change frequently fail. The EPM is meant to enable flexible creation of processes, where a valid sequence of activities can be created by establishing rules for the activities by associating roles, objects, pre- and post-conditions and interdependencies [14]**.** The domain ontology comprises concept and topic information, the enterprise ontology can be used to describe roles and organizational structures and with the process ontology the dynamic aspects can be described [14].

To apply preservation management policies on digital objects, the policies will have to be described in a formal way. Therefore, the management policies will have to be refined in detailed policies which describe processes. For implementing these processes, procedures will have to be developed and described in workflows. These workflows can be formally represented in business process models/rules. For each refined policy each statement is described step by step by high-level rules in order to create a formalized description of the policy [1]. These high-level rules can later be transformed to operational rules, e.g., utilizing the *Integrated Rule-Oriented Data System* (iRODS) [15] for implementation. Using iRODS, small well-defined micro-services can be executed.

iRODS is open source distributed software to address key elements of data management. Rules derived from policies enable automation of data workflows, with a rule engine that permits any action to be initiated by any trigger on any server or client in the grid [15] and supports plug-ins for micro-services. iRODS micro-services can be executed based on these rules. The rules can e.g. initiate packaging operations using the Packaging-Service to create OAIS Information Packages for archiving or distributing access rights. iRODS can work in a distributed environment using a variety of storage locations and resource types. With an API it is possible to retrieve Data Objects from other storage applications [16].

The concept of *Knowledge-based and Process-oriented Innovation Management* (German: *Wissenbasiertes Prozess-orientiertes Innovationsmanagement, WPIM*) was developed to support capturing and usage of knowledge around innovation processes [17]. It assumes that innovation has both a knowledge and a process perspective which need to be used in combination. Activities of a process can be annotated with resources, such as experts and documents [17].

Gernhardt et al describe in [17] how WPIM and *Distributed Process Planning* (DPP) are used for supporting *Collaborative Production Process Planning* (CAPP) (Figure 3).



Planning Process Type	WPIM Representation-Model	Output
CAPP - Process	Process	CAPP - Process
Supervisory Planning Process (SPP)	Activity	Meta Function Block (MFB)
Execution Control Planning Process (ECPP)	Activity	Execution Function Block (EFB)
Operation Planning Process (OPP)	Activity	Operation Function Block (OFB)
Planning Tasks	Task	Result / Resource

Figure 3 CAPP Ontology based on WPIM Models and DPP Process Types and Resources/Results [17]

The overall CAPP-Process can be divided into three sub-processes (called activities) as there are the so-called *Supervisory Planning Process* (SPP), *the Execution Control Planning Process* (ECPP) and the *Operation Planning Process* (OPP) and finally the operational Planning Tasks as sub-processes of the three planning sub-processes. This means while the CAPP-Process is represented as one overall WPIM Process each CAPP sub-process is mapped to a WPIM activity and its operationalization is finally resulting in a set of tasks which implement the low-level Planning tasks within the three types of WPIM Activities corresponding to the planning dimensions. In the SPP of a CAPP *Meta Function Blocks (MFB)* are produced which represent generic information of process planning as there are e.g. machining technology and constraints [18]. The *Execution Function Blocks (EFB)* are created in the ECPP and can be seen as an instantiation of a series of MFBs; it includes scheduling information and monitoring events. In the OPP *Operation Function Blocks (OFB)* are produced. The EFBs get assigned to resources by means of the OCPP activity which outputs corresponding OFBs. In the OPP the OFBs are defined. These OFBs are directly linked to resources that execute these OFBs. To achieve a representation of this kind of sub-process structure on the basis of WPIM, the process planning levels ECPP and OCPP have to be represented as additional underlying WPIM activities of the same Master Process. Therefore the resulting outputs EFBs and OFBs of these processes have to be represented as planning results and therefore as knowledge resources that are handed over between these three planning activities [17].

6. Related Technical and Scientific Work

The *Open Archives Initiative Object Reuse and Exchange* (OAI-ORE) format described in [19] defines standards for the description and exchange of aggregations of web resources. A resource can be seen as a set or collection of other resources. This resource is called an aggregation. The resource map describes the relation the aggregation has to its aggregated resources. In other words an aggregation aggregates resources and is described by a resource map. The resource map must contain the aggregation it describes, enumerate the aggregated resources and may contain relationships between aggregated resources. In OAI-ORE RDF triples of subjects, predicates, and objects are used to formulate statements. For implementing OAI-ORE serializations with Java frameworks like, e.g., Apache Jena [20] and Protégé [21] have been created. In the SHAMAN project the OAI-ORE format was used first for defining an OAIS Information Package which has later been implemented in the SCIDIP-ES project.

The Packaging Service is using the OAI-ORE format for packaging. It has its origins also in the SHAMAN project and was implemented in the SCIDIP-ES project. It could be extended for serializing the above mentioned extended Information Packages. The Packaging Service is a web service which can receive requests for packaging OAIS Information Packages in zip archives containing a manifest file describing the Information Package. The manifest file can be serialized among others in OAI-ORE [3]. The Packaging Service can therefore support the archival phase of ACILM (see Figure 1).

A promising approach to support automation has been identified by means of the linkage of data objects to be preserved with their representation information using the so-called *Preservation Assistant* (PA)[4]. This approach will be used as a base for linking digital objects to their context. The PA originates from the same projects as the so-called *Packaging Service* (PS)[3]. It had been implemented to support data creators and managers to link data objects to archives with relevant information. A form is presented to the users, which they have to complete. On basis of this form the data to be archived will be automatically connected with the respective representation information [4]. The PA can therefore support the assembly phase of the ACILM (Figure 1).

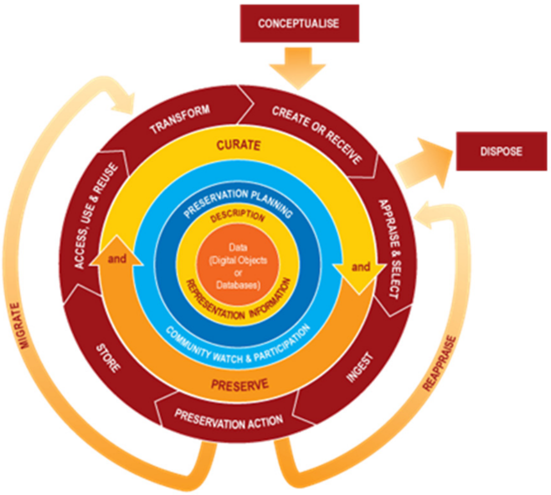


Figure 4 The DCC Curation Lifecycle Model [22]

In the *Digital Curation Centre* (DCC) the so-called *Curation Lifecycle Model* (CLM) was created, to provide a roadmap that ensures that all necessary steps in a curation lifecycle of RDM are covered [22]. As in the ACILM, the CLM of the DCC integrates activities before and after the preservation of the Data Object in its lifecycle model. While in the ACLIM the preservation of these activities is focused, in this lifecycle model the activities are specified for RDM.

Part of RDM is the curation of the created data. Digital curation involves maintaining, preserving, and adding value to digital research data throughout its lifecycle[23].

The RDM will have to interact during the different phases of a research project in the various steps of the lifecycle of a digital object. In the conceptualization phase, before the digital objects are created, capturing methods and storage will have to be planned. In this phase also requirements of the DMP will have to be incorporated, in order to comply with the funding agency's requirements. Assigning representation information, planning of preservation and curation will continue throughout the whole lifecycle of the digital object. Depending on the funder's requirements, DMPs will have to be created periodically throughout the lifecycle. The community will have to be watched and will have to participate, in order to *develop shared standards, tools and suitable software* [22]. So access, use and reuse of the digital object can be assured.

Two web-based approaches for establishing RDM the *Data Asset Framework* (DAF) and the *Collaborative Assessment of Research Data Infrastructures and Objectives* (CARDIO) have been developed by the UCC. The first is an interviewing tool covering main activities related to the curation lifecycle. The latter is a collaboration tool to find consensus by establishing RDM capabilities and finding gaps. The consensus is created by using ratings and comments [24]. Both tools are inspiring for creating user interfaces, but they themselves stay isolated in the RDM planning.

7. Modeling

In the same way as the CAPP-Process has been mapped to the WPIM-Process, the remainder of this paper will elaborate how WPIM and CAPP concepts can be applied in the next mapping step to the creation of DMPs. It should be noted that while CAPP was originally applied for planning processes in the manufacturing domain, it will now be applied to the RDM domain.

The three planning levels of the CAPP-Process comprise similar functionalities as needed by the three dimensions of DMP as described above. Therefore, to address and support the Formal Dimension of DMP, it would be necessary to execute a planning process like the SPP in CAPP where a first DMP on a meta-level is created. The activities of this process are, e.g., the formulation of requirements, constraints, organizational resources as well as target outputs. These meta-level DMP planning results are passed to the Managerial Dimension of DMP in a representation similar to a MFB. In the Managerial Dimension a planning process similar to the ECPP is needed in order to be able to express the DMP activities of this dimension including its inputs and outputs. This means that using the MFB input of the SPP the ECPP in the Managerial Dimension will define the DMP activities on the level of the RDM work plan. In this second level of the DMP planning process, which is now called ECPP, the first version of the DMP will be instantiated and responsible work package- and task-leaders need to create a corresponding RDM work plan. Activities of this process include the formulation of concrete entities, as there is the Process Information with its workflow and corresponding activities, tasks and dependencies, the Domain Information where the outcome (deliverable, knowledge, experience, result) of an activity in the Operative Dimension is described and the Organizational Information where the involved organizational unit and infrastructure is described. The output of the ECPP is representations similar to EFBs which will be handed over to the third level of DMP planning in analogy to the OPP which needs to be implemented in the Operative Dimension of DMP. This means that on this level the responsible actors have to concretely formulate the RDM operations that implement the RDM work plan. In other words, these types of activities need to be represented on the WPIM task level.

An EFB can either be directly assigned to resources in the Operative Dimension of an OPP activity which is producing the result of an OFB or they can be dynamically assigned at execution time on the level of the OPP. The results of an OFB are deliverables, knowledge, experiences and results, representing the OAIS Content Information. The OFB contains the most concrete and detailed information about the created results. The resources in the Operative Dimension are described in the Organizational Information.

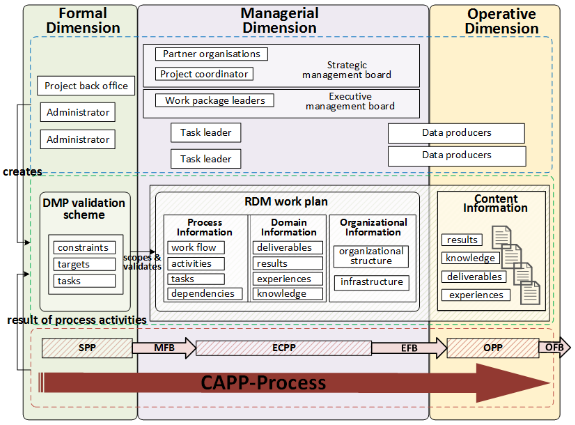


Figure 5 DMP Dimensions – CAPP-Process

Figure 5 displays a first draft for the design of the process models and information models in such a three-level DMP model that is inspired by WPIM and CAPP. The green ruled area represents the input for the OAIS Information Package with its extension to describe the context and provenance of a digital object in the Content Information, as described in ACILM and will be the information to be archived. The Process Information will be represented by WPIM-Processes which are structured in the CAPP-Process (Figure 5). The evolving concretized DMP will be extracted from the RDM work plan information.

For executing the DMP as mentioned above, iRODS rules could be used. They will have to be formulated by the Operative Dimension on the basis of an EFB/activity passed over by the Managerial Dimension. The formulated iRODS rules will have to be mapped against the policies formulated in the Formal Dimension. The policies are passed over to the Managerial Dimension in form of MFBs/activities. On the basis of these MFBs/activities, the Managerial Dimension has to be revised if the respective policy iRODS rules have already been defined. If this is the case the input data for executing the iRODS rules have to be selected and the EFB/activity can directly be passed to the iRODS rule engine for execution. Otherwise the rules will have to be mapped or formulated in the Operative Dimension. In this sense the iRODS rules can be seen as OAIS Content Information created by the Operative Dimension and will have to archived with its context.

8. Conclusions and Future Work

It has been outlined that for many of the remaining challenges starting points for research approaches do already exist. This includes modeling as well as technical challenges.

The CLM ACILM life-cycle models can guide activities and corresponding user-interfaces for creating OAIS-conformant information packages ready for ingest in an LTA.

The exiting tools and services for DMP creation and packaging are web-based to allow working in distributed environments. These web-based tools imply mostly the filling of forms that result from the DMPs in, e.g., a funding agencies' template. These tools mostly address the Managerial Dimension, needs for adding specific templates resulting from organizational backgrounds or research topic specific needs and thus affecting the Formal Dimension. The revised DMP planning tools normally do not allow for the delegation of work, as for example to the Operative Dimension for planning the concrete data to be created [25].

Modeling challenges for supporting OAIS can be approached using the EIM, which can be expressed using OAI-ORE and partly be serialized with the Packaging Service. Processes and organizations can be described and modeled using semantic models for enterprise resource planning and its application. Policies that give the context and explain the background of a digital objects creation, access and reuse can thus be formulated in a DMP in an ongoing research project.

In the modeling section an approach has been outlined using CAPP structure represented by WPIM to formulate DMPs respecting the three dimensions introduced at the beginning of this paper. As the analogies between CAPP and DMP have been shown, what remains is the formulation of an appropriate machine-readable representation of constraints as implied by laws, policies, regulations and contractual tasks. The function blocks of CAPP will have to be adapted to represent *Data Management Policy Rules* (*DMPR*) which will derive from the RDM activities represented by WPIM activities. Concrete instances of *Data Management Rules* (*DMRs*) could then be derived from the already rule-based DMPR representation in order to support an implementation using the *Integrated Rule-Oriented Data System (iRODS)* as an exemplar data management deployment infrastructure.

What remains is to formulate concrete representations of the DMPRs and DMRs.

Our future work can be divided into two subsets of R&D activities. The division into two subsets follows the suggestion in the lessons learnt from SCIDIP-ES[4] where the information modeling related to the direct users environment is separated from the OAIS Information Package creation. This means that users only have to deal with information of their research domain and does not need knowledge of the OAIS standard. The first subset consists of creating a concept of user interfaces that results in the creation a) of the DMP and b) formulating the rules that derive from the DMP. The second subset would use these rules for automating OAIS Information Package creation with Context Information by applying the formulated policies.

9. Acknowledgements and Disclaimer

This publication has been produced in the context of the RAGE project. The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644187. However, this paper reflects only the authors' views, and the European Commission is not responsible for any use that may be made of the information it contains.

10. References

[1] SHAMAN Consortium, 2011, Automation of Preservation Management Policies.

[2] APARSEN, APA | Keeping digital resources accessible, understandable and easy to find. [Online]. Available: http://www.alliancepermanentaccess.org/. [Accessed: 19-Apr-2015].

[3] SCIDIP-ES, 2013, D32 . 2 Generic Services / Toolkits and Robustness Research Report and Plan.

[4] SCIDIP-ES, 2011, D33 . 3 European ES LTDP infrastructure interoperability , architecture and governance model report.

[5] Brocks, H., Kranstedt, A., Jäschke, G., et al., 2010, Modeling context for digital preservation, *Stud. Comput. Intell.*, vol. 260, 197–226.

[6] European Commission, 2013, Guidelines on Data Management in Horizon 2020, no. December, 6.

[7] DANS, 2015, Datamanagementplan voor wetenschappelijk onderzoek, Den Haag.

[8] CCSDS, 2002, Reference Model for an Open Archival Information System (OAIS), *Forsp. Data Syst.*, no. January, 1–148.

[9] Jones, S., 2011, How to Develop a Data Management and Sharing Plan, *DCC How-to Guid.*, no. Dcc.

[10] MIT, Why manage & share your data? | Data management, *Online*. [Online]. Available: http://libraries.mit.edu/data-management/plan/why/. [Accessed: 08-Apr-2015].

[11] DATAVERSE, Data Management Plans — Dataverse.org. [Online]. Available: http://best-practices.dataverse.org/data-management/index.html. [Accessed: 10-Apr-2015].

[12] Doorn, P., 2014, Data Archiving and Networked Services World Wide Data Management : chaos of harmonie ? [Online]. Available: https://wiki.surfnet.nl/download/attachments/46794177/world wide data management ede - doorn - 10092014 .pdf?version=1&modificationDate=1410790253419&api=v2.

[13] Bayer, K., Kempf, S., Brocks, H., et al., 2006, A Multi-Agent Environment for the Flexible Enactment of Knowledge-Intensive Processes, *Cybern. Syst.*, vol. 37, 653–672.

[14] Brocks, H., Meyer, H., Kamps, T., et al., 2006, The Extended Process Model - Transforming Process Specifications into Ontological Representations, *Cybern. Syst.*, vol. 37, 1–6.

[15] iRODS, iRODS (integrated Rule-Oriented Data System). [Online]. Available: https://irods.org/. [Accessed: 18-Apr-2015].

[16] iRODS, 2014, iRODS Technical Overview. [Online]. Available: http://irods.org/wp-content/uploads/2012/04/iRODS-Overview-November-2014.pdf. [Accessed: 01-Jul-2015].

[17] Gernhardt, B., Miltner, F., Vogel, T., et al., 2015, Semantic Representation for Process-Oriented Knowledge Management based on Functionblock Domain Models Supporting Distributed and Collaborative Production Planning.

[18] Wang, L., Adamson, G., Holm, M., et al., Jul. 2012, A review of function blocks for process planning and control of manufacturing equipment, *J. Manuf. Syst.*, vol. 31, no. 3, 269–279.

[19] OAI-ORE, 2008, ORE User Guide - HTTP Implementation. [Online]. Available: http://www.openarchives.org/ore/1.0/http. [Accessed: 14-Apr-2015].

[20] Apache Jena, 2015, Apache Jena - Home. [Online]. Available: https://jena.apache.org/. [Accessed: 08-Jul-2015].

[21] Stanford University, 2015, protégé. [Online]. Available: http://protege.stanford.edu/. [Accessed: 08-Jul-2015].

[22] DCC, DCC Curation Lifecycle Model | Digital Curation Centre. [Online]. Available: http://www.dcc.ac.uk/resources/curation-lifecycle-model. [Accessed: 12-Apr-2015].

[23] DCC, What is digital curation? | Digital Curation Centre. [Online]. Available: http://www.dcc.ac.uk/digital-curation/what-digital-curation. [Accessed: 12-Apr-2015].

[24] Jones, S., Pryor, G., and Whyte, A., 2013, How to Develop Research Data Management Services - a guide for HEIs How to Develop Research Data Management, *Digit. Curation Cent.*, no. March, 1–22.

[25] Schoots, F., 2014, Datamanagementplannen Research Data Management Rapportage. [Online]. Available: https://wiki.surfnet.nl/download/attachments/42697239/Rapportage_DMP_Open.pdf?version=1&modificationDate=1397649373751&api=v2.