Towards a Metadata-driven Multi-community Research Data Management Service

Richard Grunzke*, Wolfgang E. Nagel
Center for Information Services and High Performance Computing
Technische Universität Dresden
Dresden, Germany
richard.grunzke@tu-dresden.de

Volker Hartmann, Thomas Jejkal,
Ajinkya Prabhune, Rainer Stotzka
Institute for Data Processing and Electronics
Karlsruhe Institute of Technology
Karlsruhe, Germany

Alexander Hoffmann, Sonja Herres-Pawlis
Institut für Anorganische Chemie
Rheinisch-Westfälische Technische
Hochschule Aachen
Aachen, Germany

Aline Deicke, Torsten Schrade
Digitale Akademie
Akademie der Wissenschaften und
Literatur Mainz
Mainz, Germany

Hendrik Herold, Gotthard Meinel
Monitoring of Settlement and
Open Space Development
Institute of Ecological and
Regional Development
Dresden, Germany

***Abstract*—Nowadays, the daily work of many research communities is characterized by an increasing amount and complexity of data. This makes it increasingly difficult to manage, access and utilize to ultimately gain scientific insights based on it. At the same time, domain scientists want to focus on their science instead of IT. The solution is research data management in order to store data in a structured way to enable easy discovery for future reference. An integral part is the use of metadata. With it, data becomes accessible by its content instead of only its name and location. The use of metadata shall be as automatic and seamless as possible in order to foster a high usability.**

Here we present the architecture and initial steps of the MASi project with its aim to build a comprehensive research data management service. First, it extends the existing KIT Data Manager framework by a generic programming interface and by a generic graphical web interface. Advanced additional features includes the integration of provenance metadata and persistent identifiers. The MASi service aims at being easily adaptable for arbitrary communities with limited effort. The requirements for the initial use cases within geography, chemistry and digital humanities are elucidated. The MASi research data management service is currently being built up to satisfy these complex and varying requirements in an efficient way.

***Keywords*—Metadata, Communities, Research Data Management**

I. Introduction

Today's research landscape is characterized by steadily increasing amounts of data that is caused by the use of improved data recording, increasingly complex simulation and by the correlation of numerous, often heterogeneous data sources. This increase of the data basis promises a higher amount of scientific insights. When amount and complexity of data is increasing, the requirements in regard to the data structure are also increasing. A suitable and specific data description becomes paramount. Present data processing methods are often reaching their capacity limit. Novel management methods for newer and more complex data become essential. Especially important are improved data descriptions, sustainable storages, findability, pre-processing for further use and the exploitation of existing data.

An established method to describe complex data structures is the use of metadata. This encapsulates in aggregated form the substance of a data set. Metadata ("data about data") plays a central role in making data available for the long-term. It is essential for the comprehension and storage of data, its preservation, curation and discovery for future re-use. Metadata allows for the easier applying of complex and costly tasks such as searching for data based on metadata. Aside from a better discovery, other data management aspects, such as managing and utilizing similarities between data sets, are fostered.

In diverse scientific communities partly very different metadata standards exist that each incorporate community specific data characteristics. This limited portability to new use cases causes established methods in a scientific field to be of limited use in other fields. Also, the number of standardized tools to extract metadata from heterogeneous data is limited.

In the MASi (Metadata Management for Applied Sciences) project [1] of the DFG (German Research Foundation) we are developing a generic data management service for scientific data. Along heterogeneous scientific use cases we are demonstrating its applicability. The kind and extent of the data of the participating communities is largely domain specific. Likewise, the use of metadata is not uniform across them so that a suitable overarching research data management service is a fundamental requirement.

II. Background

The MASi research data management service is being built using the KIT Data Manager repository framework (see Section II-A). Utilizing and extending the KIT DM enables MASi to offer elaborate metadata functionality with a large degree of automation and flexibility. Such metadata management capabilities are a higher level abstraction based on basic storage

devices and data management systems (e. g. iRODS [2]) in the data life cycle hierarchy [3]. Other systems including a delimitation in regard to the KIT DM are described in Section II-B.

A. KIT Data Manager - A Repository Framework

The KIT Data Manager [4] is a generic, highly customizable open source software framework for building research data repository systems. Horizontally, it is organized into a number of well-defined high-level services providing functionalities for data and metadata management and sharing as well as administrative services for user and group management. Due to the focus on research data, KIT Data Manager also provides features in addition to typical repository systems, namely a flexible data transfer service literally supporting every data transfer protocol and a data workflow service allowing to locally or remotely trigger the automatic execution of data processing tasks. These as configured in the repository system and include data transfer to the processing environment, data ingest of the processing results and provenance tracking. High-level services can be accessed either via Java APIs, e.g. to implement Web-based user interfaces or to extend the basic framework by additional functionalities, or via RESTful service interfaces, e.g. to access KIT Data Manager based repository systems remotely using a programming language of choice.

Vertically, KIT Data Manager is organized into different layers where the upper layer is formed by the high-level services described before. For repository systems based on KIT Data Manager this upper layer provides reliable and well-defined extension points on the one hand and a high degree of abstraction from underlying technologies on the other hand. The lowest layer of the architecture interfaces these technologies by defining a basic set of functionalities that is provided by a corresponding technology, e.g. to store and restore a predefined hierarchical data structure in case of the interface that has to be implemented for integrating a data storage technology. This offers a high degree of sustainability as changing technologies only affects the lower layer whereas upper layers are unaffected by technology changes.

Currently, KIT Data Manager is used to implement repository systems for various scientific disciplines, namely biology, arts and humanities, and nano-science. Due to its extensibility the base framework can be tailored to fulfil the specific needs of each of these disciplines with a reasonable effort.

B. Other Systems and Delimitation

The ICAT system [5] aims at supporting data management for photon science facilities [6]. This includes supporting beamline proposals, access rights, experiments, studies and instruments that produce the actual data. This data is collected as datasets which can then be published. The attaching of metadata such as experiments parameters, instrument parameters, and sample descriptions is supported. This closely follows the physics requirements but at the same time makes it hard to adapt for other use cases. Technically, ICAT relies on a Java EE application server with Glassfish being the standard and the Oracle and MySQL databases are supported. It offers a web service interface to support, for example, the ICAT download manager TopCat. Authentication and authorization mechanisms are supported via LDAP, local data base or anonymous access with a plugin interface being available for extensions.

DSpace [7] is a mature and ready-to-use solution for institutional repositories and it is free and open source software. It is adaptable to fit the need of individual institutions and fosters open access to all kinds of content. It supports submission workflows and various ingest and export methods. Various file types, persistent IDs and PostgreSQL and Oracle databases are supported. Search capabilities via metadata (descriptive, administrative, structural) are provided that foster the long-term preservation and accessibility of data.

Fedora (Flexible Extensible Digital Object Repository Architecture) [8] provides a framework with individual basic components to build repositories. It is open source and aims to be robust and modular. The main use case is to provide specialized services that may be integrated with existing environments and technologies. A main goal is to foster digital content preservation for complex and large datasets. Metadata for data organization is supported as well as descriptions of relationships between and linking of datasets.

EUDAT [9] is a European project aiming to create a generically applicable infrastructure to manage, access, and preserve research data. The EUDAT services B2SHARE and B2FIND involve metadata. B2SHARE is for storing and sharing research data via a web portal. It is also the central mean to upload data. This has to be done via the web portal and metadata has to be entered manually with community specific profiles being definable. B2FIND enables to access data sets via their metadata and to annotate it with comments. A B2NOTE service is planned which aims a enabling an automatic annotation of metadata [10].

iRODS as a distributed data management systems is not focused on metadata management although it offers some basic metadata functionality. Metadata can be attached to data as attribute-value-unit triples on a per file basis which can be used for searching. Integrated capabilities for metadata extraction, annotation, provenance support is missing.

In contrast to these systems, the KIT Data Manager is more flexible. It can be specifically adapted in-depth to arbitrary target communities with a close integration into community workflows. It enables far reaching automations for high usage efficiency with ready-made capabilities to be adapted to specific communities.

III. MASi Research Data Management Service

A. Overarching Goals

The MASi service is building a generic and sustainable repository. It will be sustainably operated for the involved communities to fully handle their data management requirements by utilizing metadata. One part of the project is the development of a generic model as a concrete best practice

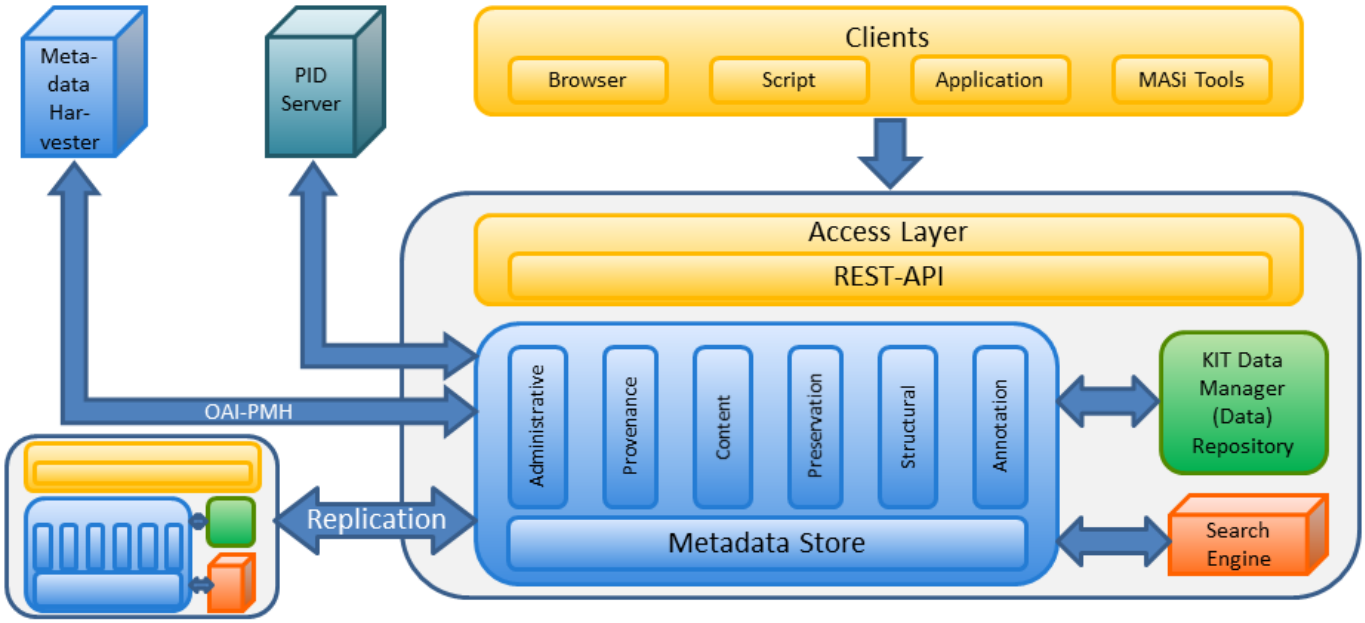

Figure 1. The MASi architecture for generic research data management.

implementation guide. It will enable to easily satisfy specific community data management requirements by using metadata. Along this guide further communities will be supported to build up their own MASi instances. The service is based on the KIT Data Manager repository framework (see Section II A) that we are currently extending. One the one hand, this includes generating a generic API to support further metadata models. On the other hand, we are implementing and will provide generic graphical interfaces to fundamentally lower the effort to adapt MASi to new use cases. Furthermore, we are closely collaborating within the Research Data Alliance (RDA) with other data researchers to develop recommendations and aim at implementing these within MASi. An example of such a RDA recommendation is the support for PIDs in conjunction with PID information types and a data type registry.

B. Generic Metadata Interface

The metadata groups within RDA compiled a set of principles regarding metadata. The well-known definition of metadata as "data about data" is the basis. Metadata differs in the mode of use and should be easily machine-understandable. It may cover the whole lifecycle of the data starting at the idea of a project, the acquisition to the publication which references the data. MASi will be a single point of access for all such kinds of metadata. The metadata is linked to the data via a unique identifier. The identifier may be a custom one as long the data is only managed internally. As soon the metadata is available for the public, a persistence identifier (PID) such as DOI (Digital Object Identifier) is used to make the data referencable. Each PID contains at least two attributes holding an URL to the metadata and to the data. The metadata is available as a METS (Metadata Encoding and Transmission Standard) [11] document which is structured in XML. In MASi it is used as the standard format for all interfaces. In the final stage there will be a registered MASi profile of METS which is valid in all configurations. METS defines seven sections for different purposes. The metadata handled by MASi itself is split in several packages (see Figure 1). Some of the packages are very similar with the sections used in METS. Others are allocated in a way so that they are most suitable for MASi. Each package is responsible for a special purpose (administrative, structural, content, bit preservation, provenance and annotation metadata). While not all packages are needed by every community the structure of the METS document may slightly differ. In the future, also new packages can be added without conflict.

To store these different kinds of formats in an efficient way, MASi offers a generic storage API to various kinds of underlying data management systems. To keep the maintenance effort manageable, we will focus on widely used standards. MASi offers a REST interface (see Figure 1) which allows for a high extensibility. This interface supports the CRUD (create, read, update, delete) operations for each package or the whole metadata sets. In case of published open access data, anyone can perform read operations on metadata without authentication. For all other operations the user has to be authenticated and authorized.

The following serves as an example regarding the provenance package functionality: There are many workflow engines each implementing its own format. But there are two standards which are supported by the majority, Open Provenance Model (OPM) [12] and ProvOne [13]. It is possible to transform OPM metadata to the ProvOne format without loss. ProvOne describes the provenance as a graph represented as

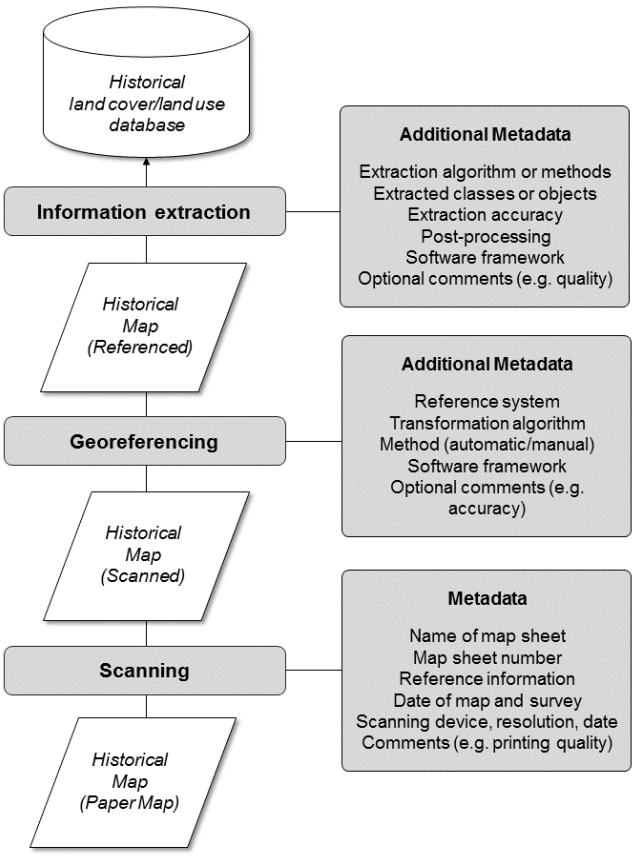

Figure 2. Workflow and metadata for historical maps.

XML. To allow for sophisticated queries the graph is stored in a graph database. Therefore, it is possible to query, e.g., for similar workflows and even more complex queries are possible. METS has a pre-defined section for provenance metadata. It uses digiProvMD which can be losslessly transformed to ProvOne and vice versa.

For each package there can be a specialized database storing the metadata. MASi will collect all metadata and compile it in a METS document using the MASi profile or in case of a data ingest split the metadata in its packages to store them accordingly. On client side there will be MASi tools supporting communities to compile a valid METS document matching the MASi profile. Subsequently, on server side a basic quality control is triggered during metadata ingest. It is based on the respective XML schema which is available for all packages except for content metadata as each community has its own specific content metadata. Such a schema needs to be created for each community. Registered schemata can be used for the quality control. If this control is required to be more sophisticated, a Java plugin can be implemented and easily activated in order so support any kind of quality control capability.

C. Generic Graphical Interface

The motivation to develop and provide a generic web interface for MASi is to significantly lower the time and effort required to adapt the MASi service to further use cases. Developers are then freed from the need to re-develop a user interface for each new use case. This saves time for developers familiar with the technology. However, it is fundamentally enabling for developers unfamiliar with it.

The generic web interface is partly built on the basis of the Liferay portal framework [14] that provides ready-made capabilities such as plugins, menus, groups, roles, separable areas and user authentication management with systems such as LDAP and Shibboleth. This enables the integration of federations such as eduGAIN [15] for the easy re-use of existing institute logins. Liferay is open-source, mature and widely used. For this to seamlessly work within MASi, we are currently developing a Liferay plugin to integrate Liferay with the KIT Data Manager repository framework. The plugin will ensure consistency between the user management systems of Liferay and KIT DM by automatically syncing new Liferay users to KIT DM. Adding users to KIT DM via another way will be disabled in this operation mode to ensure that all users exist in both systems. This integration will enable the KIT DM to transparently support authentication systems that are already supported by Liferay such as LDAP and Shibboleth. Also, the KIT DM admin interface will be integrated with Liferay. We will provide a detailed installation and configuration guide in order to further lower the barrier of adoption.

The second main part is the current development of a generic Liferay graphical interface portlet with common functionality. Initially, it will include basic upload, search and download capabilities. To fundamentally increase the impact of this development, we will create an extensive documentation to enable developers to easily adapt the portlet for their specific use case requirements. The documentation will include everything from code checkout, development project configuration, adaptation examples to compilation. A main goal of the documentation is to lower the training period as much as possible. All binaries, source code and documentation will be open source and will become part of the KIT Data Manager framework. In the course of MASi, the generic GUI portlet will be continuously extended with increasingly advanced generic capabilities. Consequently, the documentation will be appropriately extended in order to enable quick community adaptations.

IV. Initial Use Cases

A. Historical Maps

Historical topographic and cadastral maps are a valuable and often the only source for reconstructing land use changes over long periods of time. To access this information for large scale spatial analyses and change detection, advanced image analysis and pattern recognition algorithms have to be applied to the scanned map documents. The retrieved information can hence be used to “historize” existing land use and land cover databases [16].

The automatic information acquisition from historical maps generates and necessitates a variety of metadata. The process comprises three major components: firstly, the scanning of the paper maps (which are only partially available as digital images); secondly, the georeferencing of the scanned maps (which is only provided for the minority of digital available maps); and thirdly, the information extraction from the geo-referenced digital map images. Each of the three components generate at least four obligatory metadata entries. Figure 2 shows the workflow of the information acquisition process as well as the essential metadata that are generated during the process. The given metadata are essential for both the change detection process as well as the correct interpretation of the retrieved information by third users.

B. Spectroscopy in Chemistry

In bioinorganic chemistry, a multitude of spectroscopic information can be obtained by experimental methods such as UV/Vis, IR, Raman, EPR and XAS spectroscopy. In most cases, these data are complemented by theoretical simulations which help to interpret the experimental data and obtain scientific insights. In a concrete case, we investigate metal complexes and their redox behavior with oxidants (electron-taking reagents) and reductants (electron-delivering reagents) by UV/Vis spectroscopic measurements. Here, for instance, a copper(I) complex is treated with a cobalt(II) complex yielding copper(II) and cobalt(II) complexes under exchange of an electron. The copper(I) spectroscopic features decay and those of copper(II) form. The speed of this development is monitored every 1.5 ms for some seconds producing a large amount of raw data. This raw data is reduced by the researcher, e.g. by choice of a suited wavelength and absorption time traces are generated, at the moment manually. From these time traces, the kinetic decay constants are determined. This analysis is performed for several ratios between oxidant and reductant to resolve the second-order kinetics of the electron transfer. This final data shall be stored together with the theoretical analyses of the electron transfer by density functional theory. Metadata annotation is important in all steps but yet an open issue: the original raw data need annotation of who measured which chemical system and which ratio, temperature, setup etc. This information is traditionally documented manually in laboratory notebooks which are stored in the working group. The reduced data is then stored electronically. Here, metadata can comprise all metadata of the raw data but additional information on data reduction steps must be added to the metadata. The theoretical data imply different metadata: the version of the code, functional, basis set, dispersion and solvent modelling as well as grid size should be noted in the corresponding workflow [17]. Figure 3 summarizes these different levels of data production for this example.

C. Church Windows

The „Corpus Vitrearum Deutschland“ [18] is part of the international “Corpus Vitrearum Medii Aevi” (CVMA). The main focus of this long-term research project, funded by the

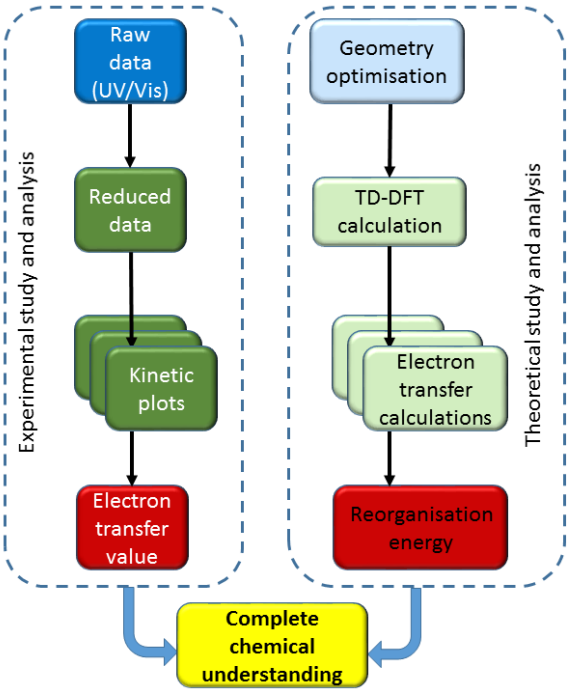

Figure 3. Metadata for the electron transfer (subgroup of the spectroscopic use case).

Academy of Sciences and Literature Mainz and the Berlin-Brandenburg Academy of Sciences and Humanities, lies in the analysis of medieval stained glass preserved in church windows, museums, galleries and other places all over Europe, the US and Canada (see Figure 4 for an example).

Due to its fragile nature, medieval stained glass is greatly affected by environmental impacts. In a first step during the research, all windowpanes are photographed and then documented in schematic drawings. With this documentation as a basis, the history of each window’s glazing and any changes or restoration activities that might have been carried out throughout the centuries are studied. Finally, the iconography and the religious context of each window within its ecclesiastical space are interpreted.

The CVMA curates a digital image archive of the photographs taken. For each image, an extensive set of metadata is provided. The records are modeled according to the guidelines of the internationally acknowledged XMP metadata standard. All XMP information is directly embedded in the TIFF files. Due to this approach, an accidental separation between the file and its metadata becomes highly unlikely. In addition to XMP, the ICONCLASS vocabulary is used to describe and classify the contents of each image.

The MASi service will open up the CVMA image archive to further interested parties, e.g. providers of cultural heritage photography such as the Prometheus, Foto-Marburg or the Europeana online platforms. During the implementation of the service, an OAI-PMH interface will be created. Also, a proof-of-concept for the automatic matching of metadata records with other cultural heritage data repositories will be drafted.

Figure 4. Mauritius and his companions refuse the idolothyte. Act of the saints window in the Marktkirche Hannover.

Finally, a configurable web interface will be built that will allow the generic embedding of this automatically enriched metadata records into the CVMA image files.

V. Conclusion and Outlook

The MASi research data management service provides a solution for the highly relevant challenge of managing large amounts of complex data. It builds on substantial previous work that is further extended and broadened. The MASi service that is currently being built up, is easily able to seamlessly integrate with highly diverse use cases. In this capacity it plays an essential role in fulfilling the complex requirements while further use cases are currently being planned.

As future work, we are evaluating the integration of MASi both with the UNICORE HPC middleware [19] and science gateways such as MoSGrid [20]. We are also continuing to work within the RDA and contribute our own expertise in discussions to create RDA recommendations on how to best handle various aspects of research data management. We aim at implementing the resulting joint recommendations within MASi. This will contribute in the creation of MASi as a service that is efficient, future-proof and has a high user acceptance.

Acknowledgment

The authors would like to thank the DFG (German Research Foundation) for the opportunity to do research in the MASi project (NA711/9-1). Furthermore, financial support by the BMBF (German Federal Ministry of Education and Research) for the competence center for Big Data ScaDS Dresden/Leipzig is gratefully acknowledged. The research leading up to these results has been supported by the LSDMA project of the Helmholtz Association of German Research Centres.

References

[1] MASi, "Metadata Management for Applied Sciences," 2016. [Online]. Available: http://www.scientific-metadata.de/

[2] A. Rajasekar, R. Moore, C.-y. Hou, C. A. Lee, R. Marciano, A. de Torcy, M. Wan, W. Schroeder, S.-Y. Chen, L. Gilbert *et al.*, "iRODS primer: Integrated Rule-Oriented Data System," *Synthesis Lectures on Information Concepts, Retrieval, and Services*, vol. 2, no. 1, pp. 1–143, 2010.

[3] R. Grunzke, J. Krüger, S. Gesing, S. Herres-Pawlis, A. Hoffmann, A. Aguilera, and W. E. Nagel, "Managing complexity in distributed data life cycles enhancing scientific discovery," in *IEEE 11th International Conference on e-Science*, August 2015, pp. 371–380.

[4] T. Jejkal, A. Vondrous, A. Kopmann, R. Stotzka, and V. Hartmann, "KIT Data Manager: The Repository Architecture Enabling Cross-Disciplinary Research," in *Large-Scale Data Management and Analysis - Big Data in Science - 1st Edition*, 2014. [Online]. Available: http://digbib.ubka.uni-karlsruhe.de/volltexte/1000043270

[5] D. Flannery, B. Matthews, T. Griffin, J. Bicarregui, M. Gleaves, L. Lerusse, R. Downing, A. Ashton, S. Sufi, G. Drinkwater *et al.*, "ICAT: Integrating Data Infrastructure for Facilities Based Science," in *e-Science, 2009. e-Science'09. Fifth IEEE International Conference on*. IEEE, 2009, pp. 201–207.

[6] R. Grunzke, J. Hesser, J. Starek, N. Kepper, S. Gesing, M. Hardt, V. Hartmann, S. Kindermann, J. Potthoff, M. Hausmann, R. Müller-Pfefferkorn, and R. Jäkel, "Device-driven Metadata Management Solutions for Scientific Big Data Use Cases," in *22nd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP 2014)*, February 2014.

[7] M. Smith, M. Barton, M. Bass, M. Branschofsky, G. McClellan, D. Stuve, R. Tansley, and J. H. Walker, "DSpace: An Open Source Dynamic Digital Repository," 2003. [Online]. Available: hdl.handle.net/1721.1/29465

[8] FedoraCommons, "Fedora Commons Repository Software," 2016. [Online]. Available: http://fedora-commons.org/

[9] D. Lecarpentier, P. Wittenburg, W. Elbers, A. Michelini, R. Kanso, P. Coveney, and R. Baxter, "EUDAT: A New Cross-Disciplinary Data Infrastructure for Science," *International Journal of Digital Curation*, vol. 8, no. 1, pp. 279–287, 2013.

[10] EUDAT, "Eudat semantics working group," 2015. [Online]. Available: http://eudat.eu/semantics

[11] J. P. McDonough, "METS: Standardized Encoding for Digital Library Objects," *International journal on digital libraries*, vol. 6, no. 2, pp. 148–158, 2006.

[12] L. Moreau, J. Freire, J. Futrelle, R. E. McGrath, J. Myers, and P. Paulson, "The Open Provenance Model: An Overview," in *Provenance and Annotation of Data and Processes*. Springer, 2008, pp. 323–326.

[13] V. Cuevas-Vicenttín, B. Ludäscher, P. Missier, K. Belhajjame, F. Chirigati, Y. Wei, S. Dey, P. Kianmajd, D. Koop, S. Bowers, and I. Altintas, "ProvONE: A PROV Extension Data Model for Scientific Workflow Provenance," in *DataONE Provenance Working Group*, 2014.

[14] Liferay, "Enterprise Open Source Portal and Collaboration Software," 2016. [Online]. Available: http://www.liferay.com/

[15] Geant, "eduGAIN - Interconnecting Federations to Link Services and Users Worldwide," 2015. [Online]. Available: http://www.geant.net/service/eduGAIN/Pages/home.aspx

[16] H. Herold, G. Meinel, R. Hecht, and E. Csaplovics, "A GEOBIA Approach to Map Interpretation-Multitemporal Building Footprint Retrieval for High Resolution Monitoring of Spatial Urban Dynamics," in *International Conference on Geographic Object-Based Image Analysis*, 2012, pp. 252–256.

[17] S. Herres-Pawlis, A. Hoffmann, T. Rosener, J. KrÃŒger, R. Grunzke, and S. Gesing, "Multi-layer Meta-metaworkflows for the Evaluation of Solvent and Dispersion Effects in Transition Metal Systems Using the MoSGrid Science Gateways," in *Science Gateways (IWSG), 2015 7th International Workshop on*, June 2015, pp. 47–52.

[18] "Corpus Vitrearum Deutschland," *http://www.corpusvitrearum.de/*, 2016.

[19] K. Benedyczak, B. Schuller, M. Petrova, J. Rybicki, and R. Grunzke, "UNICORE 7 - Middleware Services for Distributed and Federated Computing," in *International Conference on High Performance Computing Simulation (HPCS)*, 2016, accepted.

[20] J. Krüger, R. Grunzke, S. Gesing, S. Breuers, A. Brinkmann, L. de la Garza, O. Kohlbacher, M. Kruse, W. E. Nagel, L. Packschies, R. Müller-Pfefferkorn, P. Schäfer, C. Schärfe, T. Steinke, T. Schlemmer, K. D. Warzecha, A. Zink, and S. Herres-Pawlis, "The MoSGrid Science Gateway - A Complete Solution for Molecular Simulations," *Journal of Chemical Theory and Computation*, vol. 10(6), pp. 2232–2245, 2014.