Generic Metadata Handling in Scientific Data Life Cycles

Kurzfassung

zur Erlangung des akademischen Grades Doktoringenieur (Dr.-Ing.)

vorgelegt an der
Technischen Universität Dresden
Fakultät Informatik

eingereicht von

Diplom-Informatiker Richard Grunzke
geboren am 14. Januar 1983 in Hagenow

Gutachter:
Prof. Dr. rer. nat. Wolgang E. Nagel, Technische Universität Dresden
Prof. Dr. rer. nat. Achim Streit, Karlsruhe Institute of Technology

Dresden, 15. Dezember 2015

Contents

1 Managing Complexity in Scientific Data Life Cycles

In science data is the essential focal point in todays computational and quantitative approaches to scientific knowledge gain. Computational simulations enable far reaching explorations of modeled realities while quantitative methods gather data to improve the understanding of observed phenomena. These methods are increasingly viable only via high-end storage and large-scale High Performance Computing resources with individual requirements dramatically rising. Data throughputs involve gigabytes per second continuously, volumes are of petabyte magnitude, continuous files per second rates are in the double-digit range, and a vast universe of complex data representations exists. The great potential of such data is evident by the current trend of Big Data in science that aims at large-scale information extraction to foster scientific discoveries. This is fundamentally enabled by intelligently handling data and by combining a large variety of information technology methods to so-called data life cycles. In principle, these consist of data sources, systems to manage data as well as compute resources, methods for access rights management, utilization interfaces and data sinks (see Figure 1). Scientists are naturally focused on their particular research. Thus, metadata is an essential step forward in the efficiency of use as it enables managing data based on its content instead of location. Via specific data life cycles scientists are freed from the necessity to extensively deal with IT infrastructures while still utilizing them to drive their research by handling their extensive data and computing demands. In this complex technological environment, a plethora of significant challenges presents itself that hinders the advancement of the state-of-the-art in data-driven knowledge gain.

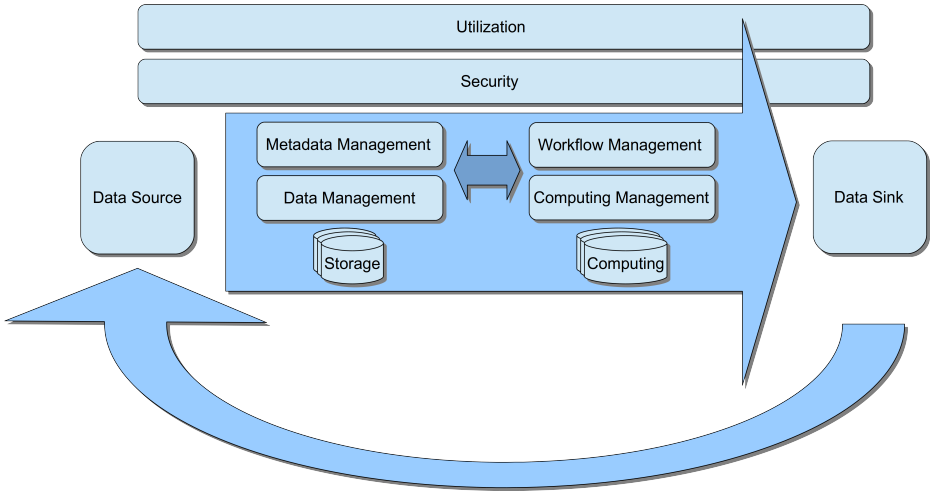

Figure 1: Principal data life cycle management component categories are depicted with data source, storage and computing pillars, data sink, and layers for security and utilization. Each category consists of a multitude of technologies to manage the inherent complexity.

2 Complex Challenges

Vital challenges in managing data life cycles are manifold. Federated authentication and authorization infrastructures need to be integrated while being mindful of the overall resilience of increasingly complex data life cycles. The increasing numbers of files and data amounts need to be managed by Big Data

systems. These in turn need to be efficiently integrated with High Performance Computing resources for analysis which signifies the need for advanced interoperability. Besides automated pre- and postprocessing, the user-friendly creation, and execution of workflows to encapsulate complex analysis procedures need to be supported. Integrated scientific environments need to be provided that hide the underlying complexity while enabling that use. Essential is also the building of trust that an infrastructure delivers what it promises. Closely connected is moving from a fixed-term build up phase to a sustainable operation phase. As these goals are partly opposing to each other, a effective balance between them needs to be developed for each data life cycle. The dissertation focuses on the major challenge of the organization of large numbers of files in the million range using information about data, so-called metadata. Currently, solutions are often either use case specific or lacking completely, thus, preventing easy access and re-use. Without metadata, users have to remember where an individual file is located. With a large number of files this is inefficient if not impossible. This especially holds true for Big Data use cases with a large number of files with complex content and stored in distributed locations. Currently, significant efforts need to be made to implement even narrowly applicable and pragmatic metadata handling solutions for every new scientific experiment.

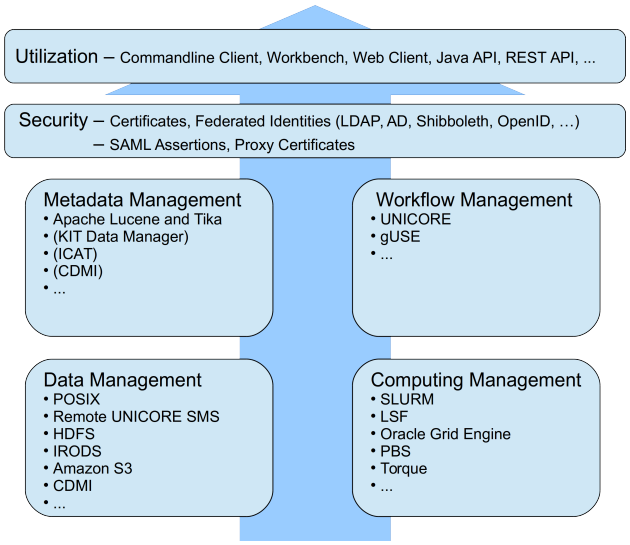

Figure 2: Abstraction from metadata systems and various other levels of abstraction from data and computing management up to the security and utilization layers are shown. Specific example technologies that the concept supports via the UNICORE middleware are noted.

3 Contributions and Impact

The contributions described in this thesis to meet the major challenge and the resulting impact are described in the following. In all cited references the author was active as main author or co-author.

First, to facilitate a thorough understanding of the context and challenges within the highly complex situation of managing distributed data life cycles, a comprehensive and overarching analysis and classification of all major data life cycle elements was created [GKG+15]. This novel and extensive analysis encompasses the complete data life cycle from creation over data handling and processing to archiving (see Thesis Sections 2.1 to 2.5). MoSGrid, the basis for the concept example implementation, is an ad-

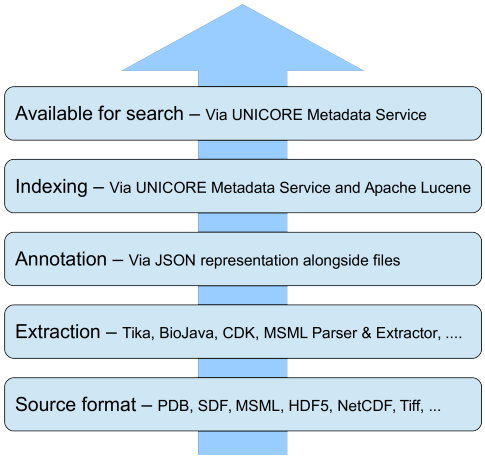

Figure 3: The figure depicts various abstraction layers reaching from data and metadata formats at the base to search availability in ascending layers. Example formats and technologies are noted.

vanced data life cycle for enabling complex and compute-intensive molecular simulations by integrating data resources with HPC and workflow management in a secure and efficiently usable way (see Thesis Section 2.6) [KGG+14, GBB+12, GKG+14]. Further data life cycles are described to give a broad context (see Thesis Section 2.7). Non-existent or highly use case specific metadata management approaches were identified, besides others, as major challenges in data life cycles across scientific communities (see Thesis Section 3.1). The necessity of generic overarching metadata approaches is elucidated via an evaluation of important metadata approaches [GHS+14] and the need for higher abstraction levels towards exascale [JMPK+15].

Second, improving upon this situation, a generic and comprehensive metadata concept was designed (see Thesis Section 3.2.1). The concept is widely applicable yet enables scalability and efficiency of use [GGJN14, GBG+14]. With respect to other approaches (see Thesis Section 3.2.2), it improves upon the state-of-the-art. On the one hand, the concept provides the following characteristics of generic and well-balanced metadata handling (see Thesis Section 3.3). Abstraction from various kinds of technologies is important in order to enable efficient integration and adaptability to new high performance Big Data life cycles (see Figure 2). Metadata and data formats need to be generically handled from the source data format over storing to being searchable in the end (see Figure 3). This enables users to efficiently and transparently organize their data. Steps such as metadata extraction, annotation, and indexing must be fully automated for management of large numbers of files. Metadata management needs to be seamlessly integrated with underlying infrastructures including systems from data, computing, security, and utilization categories. Together with the possibility to directly execute computing tasks based on metadata search results this facilitates a high usability. On the other hand, the concept provides a design guide to metadata in data life cycles (see Thesis Section 3.4). As scientific data life cycles are highly complex, overall design aspects are described in detail to facilitate an appreciation and understanding of this complexity [GDP+15]. Then, metadata design aspects are thoroughly discussed. For example, re-creating of metadata capabilities are avoided by facilitating the integration of standard components. This enables quick adaptations to different types of use cases and overall integration paths into new scientific data life cycles. Technology recommendations include an analysis of proven technologies that directly enable metadata management and beyond in data life cycles (see Figure 2 and 3). The concept advances the

state-of-the-art in data life cycle management. When implemented, it enables scientists to focus on their core research while utilizing Big Data and High Performance Computing resources.

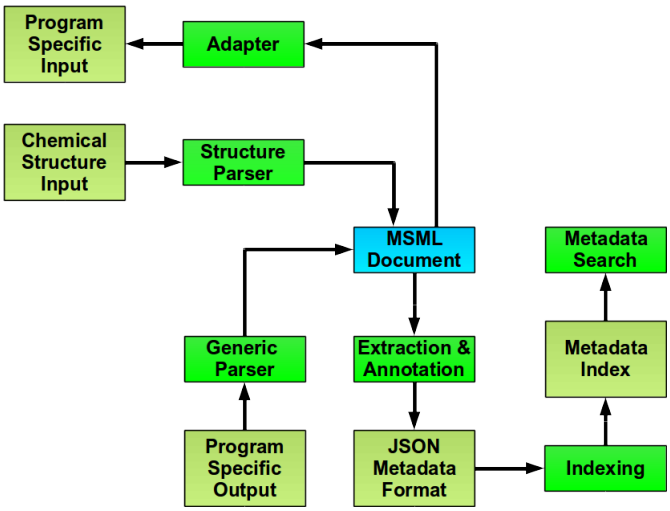

Figure 4: As part of the concept implementation within the complex MoSGrid data life cycle, the complete component chain for extracting, annotating, and indexing metadata based on the MSML format, the MoSGrid information hub, is depicted.

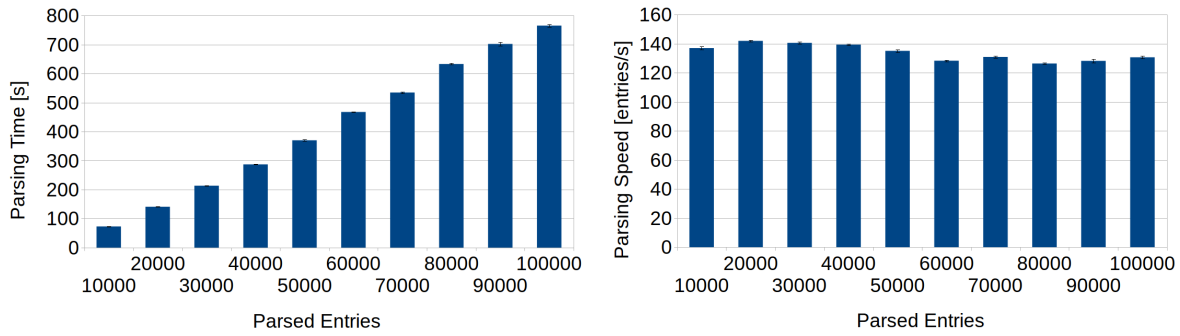

Figure 5: A performance evaluation of the extraction and annotation step from the central MoSGrid information hub MSML to the JSON metadata format is presented. Depicted is the parsing time in seconds on the left side and the corresponding parsing speed in processed entries per second on the right side. 10 datasets with 10000 to 100000 entries in steps of 10000 are shown. Each measurement was repeated 10 times and the respective average and error bars are shown. Even at this extreme benchmarking scale, that will not be reached in production use cases in the foreseeable future, the imposed overhead by the metadata extraction and annotation is insignificant in contrast to typical workflow runtimes of hours or even days.

Third, based on the generic concept an example implementation for the MoSGrid data life cycle was created within the MoSGrid collaboration (see Thesis Section 4). On the one hand, MoSGrid concept implementation adopts the concept characteristics by utilizing abstraction layers. It transparently integrates with the complex underlying High Performance Computing and data infrastructures and with the single sign-on concept and implementation [GGK+12]. Based on the MoSGrid data description format, the extraction, annotation and indexing is fully automatic [GBG+14] (see Figure 4). The seamless integration of the metadata capabilities are the basis for the implemented and integrated search interface. It

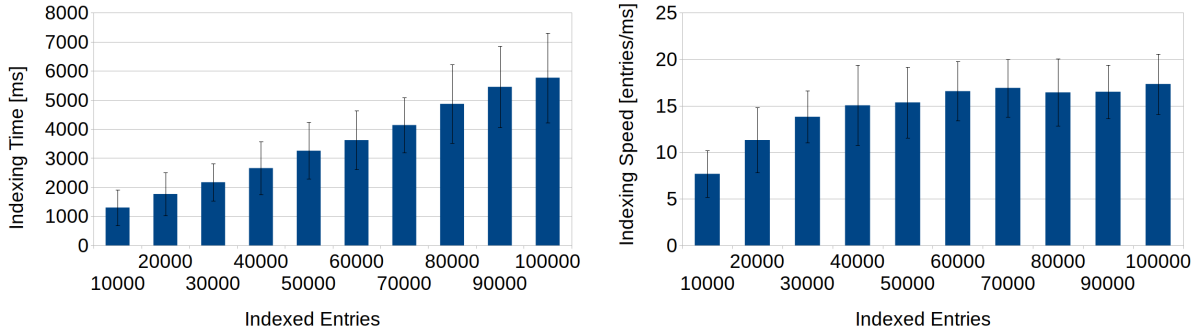

Figure 6: Here, the performance of the central step that indexes the metadata in JSON format in order to expose it to serach queries by users is evaluated. Depicted is the indexing time in seconds on the left side and the corresponding indexing speed in processed entries per second on the right side. 10 datasets with 10k to 100k entries in steps of 10k are shown. Each measurement was repeated 10 times and the respective average and error bars are shown. Even for the 100k entries, an extreme number that will not be reached in production use cases in the foreseeable future, the overall overhead is only about 6 seconds and, thus, insignificant in contrast to common workflow runtimes in the range of hours to days.

facilitates the finding of data and enables the seamless use of results for further workflows. On the other hand, the metadata aspects of the concept's design guide (see Thesis Section 3.4.2) played in central role in the MoSGrid implementation as well as the technology recommendations (see Thesis Section 3.4.3) that were closely followed. The implementation results in a major advancement of the MoSGrid data life cycle by extending its capabilities in handling large amounts of complex data.

Fourth, the generic metadata management approach and its example implementation for the MoSGrid data life cycle were evaluated [GKJ+ed, GBG+14] on the basis of criteria that permeate the concept (see Thesis Section 5). The generic metadata approach is enabling important synergy effects in supporting new data life cycles (see Thesis Section 5.1). One is the ability to quickly integrate metadata capabilities. Another is the increased efficiency based on enabling an easy and seamless integration with High Performance Computing and Big Data infrastructures (see Thesis Section 5.5). A performance evaluation of the extraction, annotation, and indexing key components shows favorable characteristics (see Figure 5, Figure 6 and Thesis Section 5.2). Furthermore, sustainability (see Thesis Section 5.3) and resiliency aspects (see Figure 7 and Thesis Section 5.4) are evaluated.

This thesis is a step towards widely enabling metadata management accross scientific disciplines. The uptake of the concept and its implementation in the MASi research infrastructure (see Section 5) as well as its utilization within the MoSGrid data life cycle ensures a broad and lasting impact.

4 Conclusion

The thesis presents an overarching and generic metadata handling concept for scientific data. Utilizing metadata, the approach facilitates the move towards the next generation of data management within scientific data life cycles. Metadata management in general enables the organization of large amounts of data with file number in the millions. Instead of only being accessible via names in directory structures,

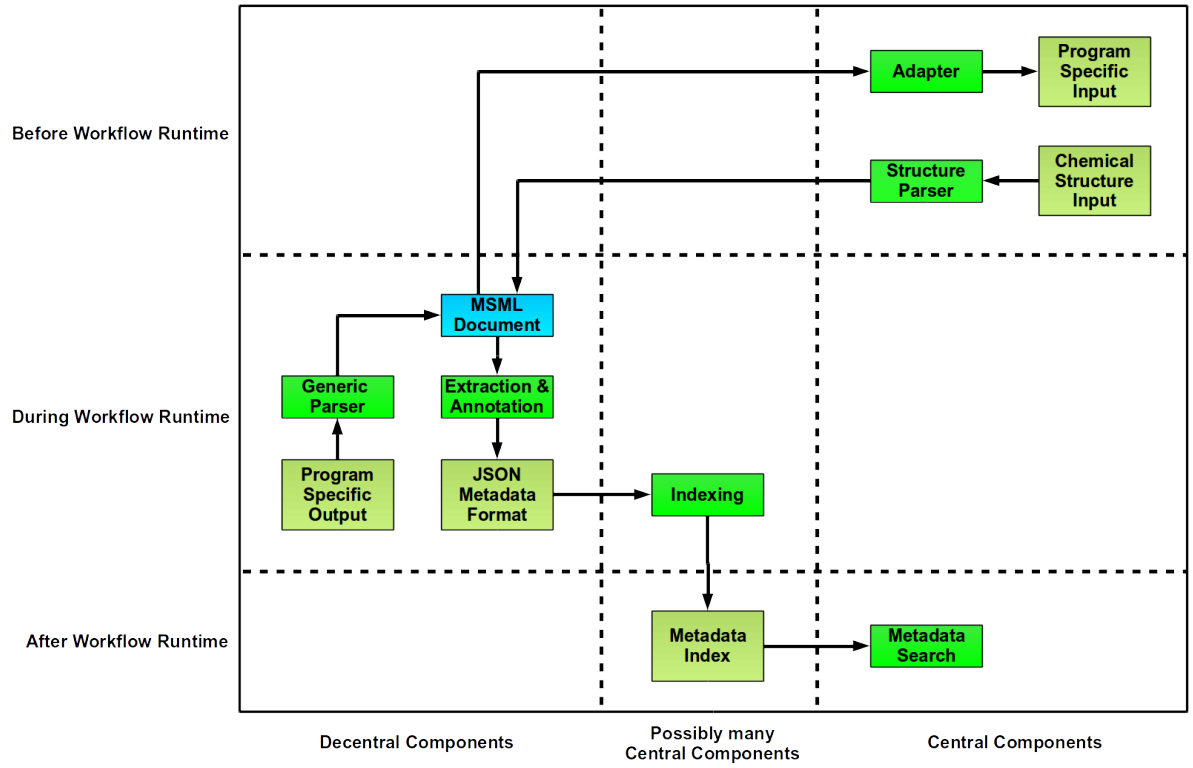

Figure 7: Different resilience classes are depicted. From top to bottom, the rows represent before, during and after a workflow is running. From left to right, the columns represent components that run decentralized, central but possibly many and completely centralized. The figure shows that during workflow runtime (middle left) within the MoSGrid data life cycle only at the end the central indexing component (middle) is used, but which may be duplicated for greater resilience. The created index (lower middle) can exist multiple times and is accessed via the central search interface of MoSGrid (lower right). Favorable resilience characteristics exist due to the principle of distributed when possible and central where necessary.

files can be accessed by their content. Despite the inherent complexity of data contents, the data can be seamlessly accessed via simple search queries and directly further utilized. The high complexity and large magnitude of scientific data life cycles is motivated. This subsequently necessitates sophisticated and integrated technologies for their management [GKG+15]. Based on such technologies, scientists are enabled to advance their respective state-of-the-art with the combined support of High Performance Computing and Big Data resources. A multitude of open challenges in this broad data life cycle context was identified. Within the challenges, a major one is that of completely missing or only narrowly applicable metadata management approaches.

The metadata concept [GGJN14, GBG+14] first specifies multiple advantageous characteristics. The abstraction of various technologies is heavily utilized to handle the high data life cycle complexity. Data and metadata formats are integrated in a generic way to enable advanced search capabilities for the efficient usage by scientists. The concept is inherently scalable within HPC-enabled and Big Data life cycles. Full automation is provided for extraction, annotation, and indexing of metadata. Second, the concept provides a design guide that facilitates an understanding of overall data life cycle design aspects [GDP+15] as well as aspects specific to metadata management. The guide includes recommendations of proven technologies. The overall concept is generic in the sense that it enables metadata management

to be more quickly and efficiently integrated in concrete data life cycles. The direct usage of metadata search results is enabled while underlying Big Data and High Performance Computing resources are seamlessly integrated. Users gain from new capabilities while the added necessary complexity is hidden.

The concept was implemented within the MoSGrid data life cycle [KGG+14, GBB+12, GKG+14]. MoSGrid enables highly complex molecular simulations within the three major computational chemistry domains. The MoSGrid implementation is HPC and workflow enabled, offers advanced data management capabilities with a sophisticated single sign-on architecture throughout all layers [GGK+12]. The implementation based on the generic approach was seamlessly integrated with this complex data life cycle. The metadata extraction, annotation, and indexing are performed in a fully automatic way [GBG+14]. A search interface enables finding data based on its content. Furthermore, search results can immediately be re-used as input within further workflows.

A thorough evaluation of the concept and its implementation was performed [GKJ+ed, GBG+14] with respect to adaptability, performance, sustainability, resilience, and efficiency of use. The existence of favorable properties was shown.

On a theoretical level, data life cycle management is advanced by facilitating higher level abstraction with metadata management. A practical impact is achieved by the implementation within the MoSGrid data life cycle and the uptake of the concept within the MASi research infrastructure.

5 Outlook

The dissertation contributes to the state-of-the-art in facilitating the handling of large amounts of complex data via metadata. Further advancements include the following.

Based on the contributions and the doctoral research collaborations, the author initiated and coordinated the proposal of the MASi (Metadata Management for Applied Sciences) research infrastructure project submitted to the German Research Foundation (DFG). MASi is now funded and led by the author. It is creating a distributed, generic, and sustainable service for the integrated management of large scale scientific data based on metadata. The service will be loosely coupled between data centers where the components are sustainably run locally for the respective communities. MASi aims at long-term sustainability via its involvement in the Research Data Alliance (RDA), the Large Scale Data Management and Analysis (LSDMA) project, and the Competence Center for Scalable Data Services and Solutions (ScaDS Dresden/Leipzig). MASi aims at supporting various advanced concepts as detailed in the following.

Multiple data and metadata sources are being integrated with MASi, so users can stage in data from external data sources and use it in a seamless way. Support for workflow provenance shall be offered as it is increasingly important to foster scientific reproducibility. The same holds true for persistent identifiers (PIDs) so data can be uniquely referenced worldwide. Different underlying metadata management systems will be transparently supported by MASi. This will be enabled by the CDMI (Cloud Data Management Interface) and OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) access standards as well as with support for the iRODS and ICAT systems. For example, the widely used

dCache distributed management systems plans to natively support metadata management and data access via CDMI.

MASi establishes metadata as the central source of information within a data life cycle. This needs to be established as a general design principle in data life cycles in general. With increasing data amounts, the computing power required for extraction is increasing as well. Metadata extraction methods have to be integrated in a scalable way as close to the data as possible. The dissertation approach includes this and it is exemplary implemented for the MoSGrid extraction methods. General extraction frameworks such as Apache Tika need to support scalable execution as well. In addition, pre-existing metadata, from whatever source available, needs to be seamlessly utilized. Future research also necessitates the integration of authentication, authorization, and single sign-on methods via nation-level federations such as DFN-AAI or on the European level via EduGain. This is essential to further lower barriers of entry by enabling the re-use of already existing login credentials.

As a central part of advanced data life cycles, metadata management should be resilient and fault tolerant to a high degree. Usually many distributed components depend on each other. Single points of failure have to be avoided by, for example, utilizing a combined metadata proximity approach as in the dissertation approach. On the one hand, it is decentralized as much as possible with loose coupling. On the other hand, the approach is centralized as much as necessary. Such an emphasis on resiliency enables a data life cycle to tolerate certain kinds of errors. When errors occur, users should get error messages in an understandable form while administrators should get all the details. In MASi, fault tolerance will additionally be supported by replicating metadata accross participating data centers.

A further research and development direction is to encapsulate even more functionality within scientific workflows. An example is to define a metadata search term as a workflow parameter. Then, during workflow enactment, the search query is executed and all files referenced via search results are utilized as input. A vital factor to enable such an approach is the interoperability and integration of High Performance Computing and Big Data systems with metadata management.

The Max Planck Institute of Molecular Cell Biology and Genetics in Dresden develops, builds, and utilizes high-throughput microscopes with common data rates of 0.85 GB/s and 10 files/s. These vast amounts of data alone are a significant challenge. An adaption estimation of the dissertation metadata approach and surrounding technologies for this use case is presented in Section 5.1.2. The author currently cooperates with the institute towards its implementation.

Exascale is on the horizon in both the computing and data domain. A multitude of challenges arises on various abstraction levels. These challenges necessitate advanced approaches in data life cycle management with metadata management as a vital component [JMPK+15]. The dissertation approach is a step in paving the way for the efficient management of increasingly large quantities of data towards exascale.

6 Publications

The results of the dissertation and intermediary connected research were published in various journals, book chapters, and proceedings of conferences and workshops. The following publications are closely

connected to the doctoral research and are referenced in Section 3. The complete list of the author's publications is attached as appendix in the thesis.

* These authors contributed equally to the respective work.

[GBB+12] **Richard Grunzke**, Georg Birkenheuer, Dirk Blunk, Sebastian Breuers, Andre Brinkmann, Sandra Gesing, Sonja Herres-Pawlis, Oliver Kohlbacher, Jens Krüger, Martin Kruse, Ralph Müller-Pfefferkorn, Patrick Schäfer, Bernd Schuller, Thomas Steinke, and Andreas Zink. A Data Driven Science Gateway for Computational Workflows. In UNICORE Summit 2012 Proceedings, volume 15 of IAS Series, pages 35–49, 2012.

[GBG+14] **Richard Grunzke**, Sebastian Breuers, Sandra Gesing, Sonja Herres-Pawlis, Martin Kruse, Dirk Blunk, Luis de la Garza, Lars Packschies, Patrick Schäfer, Charlotta Schärfe, Tobias Schlemmer, Thomas Steinke, Bernd Schuller, Ralph Müller-Pfefferkorn, René Jäkel, Wolfgang E. Nagel, Malcolm Atkinson, and Jens Krüger. Standards-based Metadata Management for Molecular Simulations. Concurrency and Computation: Practice and Experience,26(10):1744–1759, 2014.

[GDP+15] Sandra Gesing, Rion Dooley, Marlon Pierce, Jens Krüger, **Richard Grunzke**, Sonja Herres-Pawlis and Alexander Hoffmann. Science Gateways - Leveraging Modeling and Simulations in HPC Infrastructures via Increased Usability. High Performance Computing Simulation (HPCS), 2015 International Conference on, 2015, 19-26.

[GGJN14] **Richard Grunzke**, Sandra Gesing, René Jäkel, and Wolfgang E. Nagel. Towards Generic Metadata Management in Distributed Science Gateway Infrastructures. In IEEE/ACM CCGrid 2014 (14th International Symposium on Cluster, Cloud and Grid Computing), pages 566–570, Chicago, IL, US, May 2014.

[GGK+12] Sandra Gesing*, **Richard Grunzke***, Jens Krüger, Georg Birkenheuer, Martin Wewior, Patrick Schäfer, Bernd Schuller, Johannes Schuster, Sonja Herres-Pawlis, Sebastian Breuers, Ákos Balaskó, Miklos Kozlovszky, Anna Szikszay Fabri, Lars Packschies, Peter Kacsuk, Dirk Blunk, Thomas Steinke, Andre Brinkmann, Gregor Fels, Ralph Müller-Pfefferkorn, René Jäkel, and Oliver Kohlbacher. A Single Sign-On Infrastructure for Science Gateways on a Use Case for Structural Bioinformatics. Journal of Grid Computing, 10(4):769–790, 2012.

[GHS+14] **Richard Grunzke**, Jürgen Hesser, Jürgen Starek, Nick Kepper, Sandra Gesing, Marcus Hardt, Volker Hartmann, Stephan. Kindermann, Jan Potthoff, Michael Hausmann, Ralph Müller-Pfefferkorn, and René 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.

[GKG+14] Sandra Gesing, Jens Krüger, **Richard Grunzke**, Luis de la Garza, Sonja Herres-Pawlis, and Alexander Hoffmann. Molecular Simulation Grid (MosGrid): A Science Gateway Tailored to the Molecular Simulation Community. In Science Gateways for Distributed Computing Infrastructures, pages 151–165. Springer International Publishing, 2014.

[GKG+15] **Richard Grunzke**, Jens Krüger, Sandra Gesing, Sonja Herres-Pawlis, Alexander Hoffmann, Alvaro Aguilera, and Wolfgang E. Nagel. Managing Complexity in Distributed Data Life Cycles Enhancing Scientific Discovery. In e-Science (e-Science), 2015 IEEE 11th International Conference on, pages 371–380, August 2015.

[GKJ+ed] **Richard Grunzke**, Jens Krüger, René Jäkel, Wolfgang E. Nagel, Sonja Herres-Pawlis, and Alexander Hoffmann. Metadata Management in the MoSGrid Science Gateway for Quantum Chemistry. Journal of Grid Computing, accepted.

[JMPK+15] René Jäkel, Ralph Müller-Pfefferkorn, Michael Kluge, **Richard Grunzke**, and Wolfgang E. Nagel. Architectural Implications for Exascale based on Big Data Workflow Requirements. In Big Data and High Performance Computing, volume 26 of Advances in Parallel Computing, pages 101 – 113. IOS Press, 2015.

[KGG+14] Jens Krüger*, **Richard Grunzke***, Sandra Gesing*, Sebastian Breuers, André Brinkmann, Luis de la Garza, Oliver Kohlbacher, Martin Kruse, Wolfgang E. Nagel, Lars Packschies, Ralph Müller-Pfefferkorn, Patrick Schäfer, Charlotta Schärfe, Thomas Steinke, Tobias Schlemmer, Klaus Dieter Warzecha, Andreas Zink, and Sonja Herres-Pawlis. The MoSGrid Science Gateway - A Complete Solution for Molecular Simulations. Journal of Chemical Theory and Computation, 10(6):2232–2245, 2014.