PAPER • OPEN ACCESS

Advancing data management and analysis in different scientific disciplines

To cite this article: M Fischer *et al* 2017 *J. Phys.: Conf. Ser.* **898** 082026

View the article online for updates and enhancements.

Advancing data management and analysis in different scientific disciplines

M Fischer[1], M Gasthuber[2], A Giesler[3], M Hardt[1], J Meyer[1], A Prabhune[1], F Rigoll[1], K Schwarz[4] and A Streit[1]

[1] Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
[2] Deutsches Elektronen-Synchrotron (DESY), Hamburg, Germany
[3] Forschungszentrum Jülich, Jülich, Germany
[4] GSI Helmholtz Centre for Heavy Ion Research, Darmstadt, Germany

E-mail: max.fischer@kit.edu, martin.gasthuber@desy.de, a.giesler@fz-juelich.de, marcus.hardt@kit.edu, joerg.meyer2@kit.edu, ajinkya.prabhune@kit.edu, fabian.rigoll@kit.edu, k.schwarz@gsi.de,achim.streit@kit.edu

Abstract. Over the past several years, rapid growth of data has affected many fields of science. This has often resulted in the need for overhauling or exchanging the tools and approaches in the disciplines' data life cycles. However, this allows the application of new data analysis methods and facilitates improved data sharing.

The project Large-Scale Data Management and Analysis (LSDMA) of the German Helmholtz Association has been addressing both specific and generic requirements in its data life cycle successfully since 2012. Its data scientists work together with researchers from the fields such as climatology, energy and neuroscience to improve the community-specific data life cycles, in several cases even all stages of the data life cycle, i.e. from data acquisition to data archival. LSDMA scientists also study methods and tools that are of importance to many communities, e.g. data repositories and authentication and authorization infrastructure.

1. Introduction
The project Large-Scale Data Management and Analysis (LSDMA) [1, 2] aims at joint research and development by data scientists and domain scientists of various fields. The organizational structure of LSDMA with five domain specific Data Life Cycle Labs (DLCL) and the Data Services Integration Team (DSIT) allows to address both community-specific and generic aspects of data management and analysis.
In this note, we discuss selected highlights of LSDMAs research and development activities. Specifically, we will address how the results have advanced the user communities in their data-driven research. We will conclude with the lessons we have learned in the past few years.

2. DLCL Climatology
Filtering and visualization of large amounts of data are important aspects of data analysis. In the DLCL Climatology, a server-client application was designed and developed that allows to visualize data from earth-observing satellites [3]. The distributed system consists of three building blocks. Semi-structured data are stored in a NoSQL database serving as horizontally scalable storage back-end. The visualization application runs as web application inside a browser,

i.e. without any need to install additional software. The web client benefits from the 3D-visualization capabilities of WebGL that is supported by all modern browsers and that supports GPU accelerated image processing. The storage back-end and the web client are connected via a middle layer called Node Scala [4]. Node Scala provides a REST interface for clients to request a specific selection of data. Node Scala does not only fetch data from the storage back-end, but performs parallel predefined pre-processing tasks. This design results in fast response times while minimizing the data transfer rates to the client and the client-side CPU demands. Climate researchers benefit from a easy-to-use tool for an interactive analysis of large amounts of data.

3. DLCL Energy

In the DLCL Energy, a concept for an energy management system has been refined and adapted [5]. To that end, a detailed analysis was conducted: In a first phase, a technical data life cycle analysis was performed and in a second step, a privacy analysis from a user's perspective was carried out. The resulting requirements were used to improve the existing design. Furthermore, the concept has been implemented in form of a demonstrator: the *Data Custodian Service* (DCS). The DCS is the only energy data sink in a household. In order to potentially gain access to that data, interested parties have to file a data request at the DCS. This request is evaluated with regard to privacy implications for the user. The result is presented to the respective data owners and depending on their decision, the data request is granted or declined. Thus, users are able to fully understand the consequences of sharing data and can make an informed decision regarding the circulation of their data. On the technical side, the demonstrator uses an SQL-based storage for the meta data and an HDF5 storage for the time series data. This enables efficient searching and filtering on the meta data as well as efficient access to large amounts of energy data.

4. DLCL Key Technologies

Light microscopy is a routine imaging technique in biological and medical research and diagnosis. Localization microscopy, especially Spectral Position Determination Microscopy, can scale the optical resolution down almost to the electron microscopy level in the 10 nm range, which is important for biological and medical research and diagnosis [6, 7]. But these techniques produce image data in the range of GB/s and require the handling, processing and evaluation of image stacks of up to thousands of frames per single cell. These data have to be stored and made accessible for the research community.

To this end, the DLCL Key technologies has designed a Localization Microscopy Open Reference Data Repository (LMORDR). Theses developments are a step towards data management and curation with long-term perspectives under the aspects of sustainability and potential of re-use for different analyses. LMORDR is a system for transmitting the data, adding metadata for characterization and retrieval and storing the data. It offers programs and processing procedures for fast evaluation, either on individual machines, or in clusters. A Generic Client Service API [8] for connecting disparate services is designed and implemented seamlessly, integrating with the KIT Data Manager [9] and the Large Scale Data Storage. A structured metadata model based on Core Scientific Metadata Model (CSMD) is established for describing the extremely large datasets of localization microscopy research. Standardized descriptions of the workflow steps with an automated execution of the workflow, based on extended image analysis programs, are achieved by a workflow management system. A provenance manager collects, models and stores the entire provenance information generated during the execution of a localization microscopy workflow [10]. Provenance information is stored in the W3C ProvONE standard format using a graph database.

Due to the interdisciplinary collaboration of computer scientists, biophysicists and biomedical users, constructive approaches led to an implementation that fulfills all requirements of the

domain scientists.

5. DLCL Structure of Matter

5.1. HiDRA Petra III and Flash

The development of detectors at 3rd generation light sources are currently outpacing experimental methods and data acquisition. Single clients will produce 0.5 GBytes/sec and the next generation is already pushing for 6 GBytes/sec. For 30 beamlines the expected aggregated rate is of 50 to 80 GBytes/sec, depending on detector deployments. Measurements last from a few hours to a few days resulting in many data sets of up to tens of TBs each. From next generation detectors we also expect multi GBytes/sec spread over many 10GE connections. The requirements will vary a lot due to the very dynamic experimental setup with inherent burst nature and a very heterogeneous environment regarding technology and social context. In order to support better data control and shorter turnaround cycles for analyses, the new system has to allow high-speed data access within seconds after data have been generated by the detector, within a few minutes (shorter is better) for full scale data analysis using multiple CPUs and within hours to be archived to tape media and to be available for external (remote) access.

The next generation of experiments will require controlled and fast access (bandwidth and latency) to the most current generated data to allow immediate experiment control [11]. Scientists at DESY have developed experimental setups where samples are constantly flowing in a liquid or gaseous jet across a pulsed X-ray source which has a repetition rate of up to 120 Hz. Significant amounts of sample are consumed in a very short time, and the data generated by the instruments requires a large amount of storage space. Furthermore, experimental parameters, such as the degree of molecular alignment in controlled imaging experiments, or the hit rate and resolution in an SFX[1] experiment, must be kept within acceptable bounds. By monitoring experimental conditions in close to real time, the experiment may be maintained in optimal alignment, or alternatively, one may pause the experiment to correct unfavorable conditions, thereby preventing the collection of unfavorable data while preserving valuable sample.

OnDA (Online Data Analysis) is a fast online feedback framework which provides the possibility to decide in near realtime about the quality of the data produced in serial X-ray diffraction and scattering experiments. It is designed on a highly modular basis and provides stable and efficient real-time monitors for most common types of experiments. Recent beamtimes completed at the Petra III facility show a smooth integration with the new storage system supporting all required criteria. This integration work is ongoing, expecting more experiments with similar demands. Building a generic solution, supporting all types of data flow control and dispatch, meeting all performance criteria, is the base goal for the ongoing development effort. The developed HiDRA software package introduce a generic layer to allow a flexible data flow configuration between detector and the first 'touch down' of the data in the GPFS storage system for any type of online *synchronous* data analysis. [12]

5.2. FAIR

The experiments at the Facility for Antiproton- and Ion Research (FAIR) in Darmstadt have large storage requirements of several 10 PB/year. To support the large I/O demands of the experiment data analysis, the GSI Helmholtzzentrum für Schwerionenforschung operates for the FAIR experiments a large-scale, high-performance Lustre shared-file system. However, the local and WAN data access and management from the experimental frameworks are based on the XRootD protocol. In the context of LSDMA, work has been done to couple both worlds. Many aspects of this work have already been used in production in the context of the ALICE Tier 2 center operated at GSI [13, 14].

---

[1] Serial femtosecond X-ray crystallography

The storage resources pledged at GSI to the global ALICE community are provided via a Grid Storage Element which consists of a set of XrootD daemons running on top of the Lustre file system. The compute jobs of the ALICE Tier 2 centre are submitted to GSI's HPC cluster, which is considered as an isolated environment, where direct connections between the cluster's worker nodes and the internet are partially or fully restricted. Therefore, an XrootD forward proxy has been set up, which enables the site admin to allow worker nodes to read input files from, and write output files to, remote sites, while adhering to the aforementioned restriction.

All clients using XrootD to access ALICE Grid data request it through XrootD data servers. This means that I/O traffic needs to go through the limited link that one data server can provide and that Lustre's full I/O speed cannot be utilized directly.

The proposed solution for this is to use the XrootD client plug-in API to redirect underlying access to data on Lustre directly, bypassing the need to read indirectly via the XrootD data servers if the data is locally available. The XrootD client plug-in API comes with the possibility to change XrootD's underlying I/O operations, so that higher level software's performance (e.g. ROOT, xrdcp) is improved by the plugin transparently.
The current tests show that such a plug-in can be used to adjust XrootD to specific needs as described above.

6. DLCL Neuroscience
The three-dimensional Polarized Light Imaging (3D-PLI) is a neuroimaging technique used at the Institute of Neuroscience and Medicine (INM-1), Forschungszentrum Juelich, to reconstruct the three-dimensional nerve fiber architecture in postmortem mouse, rat, and human brains at the micrometer scale [15] [16]. The examination of a human brain with 3D-PLI generates about 2,500 histological sections, which are digitized at 1.3 μm pixel size resulting into image sizes per section of about 70,000 x 100,000 pixel with a color depth of 32-bit.
The subsequent post processing and the extraction of fiber orientations from the microscope images require a complex chain of tools. These tools have been integrated in a UNICORE (Uniform Interface to Computing Resources) [17] workflow towards a fully automated and parallelized image processing, utilizing advanced supercomputing infrastructure efficiently. This reduces significantly the processing time from days to hours, which is a relevant factor considering the thousands of sections to be analyzed in a whole human brain study. UNICORE turned out to be a valuable tool serving both the software developer by integrating their image processing tools, as well as the scientific user lacking in deep knowledge of how to use a supercomputer infrastructure. Neuroscientists were able to perform complex data analysis and routine data production, without knowing all details about the different data inputs, calls and requirements of individual software packages of the workflow. This setup clearly minimized operation failures as compared to manual processing of individual software packages. Furthermore, the workflow could be used as performance measurement tool for the utilized supercomputers. As a result, the advantage of specific features of different supercomputers (e.g., GPU vs. faster CPU) could be addressed by the workflow.

7. Data Services Integration Team (DSIT)
In the context of LSDMA the Data Services Integration Team set out to develop solutions that are relevant to multiple DLCLs. Below, highlights of six fields are presented:

7.1. Authentication and Authorization Infrastructures (AAI)
The goal in this field is to allow globally operated infrastructures and global collaborations to access resources and share data in secure ways. DSIT provided a general roadmap to integrate

several different authentication mechanisms with one another. This includes token translation services, account linking and identity harmonization services. The goal is that users will be able to access or manage the data they possess, regardless of which technology they used to authenticate with a specific access protocol. This work builds on top of concepts found in todays WLCG grid middleware. Several ideas were introduced into successful EU proposals and are implemented in INDIGO Data-Cloud[2] and continuously improved within Authentication and Authorization for Research Communities 1, 2 (AARC/AARC2)[3] and EUDAT-2020[4].

In cooperation with AARC/AARC2 DSIT developed the Blueprint Architecture for a Pan-European AAI [18], a document that describes how authentication should work in future infrastructures. A prototype with AARC and EUDAT led to the integration of b2access (the EUDAT SP/IdP proxy) into services that allow shell access to unmodified ssh daemons. This allows logging in via a home-identity, or google or ORCID into an ssh host. Avoiding the need for source code modifications is one step to ensuring a sustainable solution.

In the context of INDIGO, the DSIT plan for TTS (on-site token translation service) was implemented. This tool provides an extensible web and REST service that returns access credentials to a user authenticated via OpenID-Connect (OIDC). OIDC [19] was chosen initially, since this is used in INDIGO, extension to the Security Assertion Markup Language (SAML [20]) is straightforward. As an example, users can now use an OIDC authentication token to obtain an ssh-key or an X.509 certificate should they wish to access ssh or gridftp, respectively. To ensure different credentials are mapped to the same user at a given site, identity harmonization is developed, which allows multiple accounts for the same user to access the same data. This is accomplished by setting all UIDs of users to a primary UID, as defined by the user at an external services. Such a service is IAM [21], which is developed within the INDIGO project by partners at INFN/CNAF.

In the context of the Human Brain Project [22] and of DSIT, the HPC grid middleware UNICORE [17, 23] was extended from SAML and the Simple Object Access Protocol (SOAP) to support federated authentication via OIDC. Also supporting REST interfaces on the whole UNICORE stack was a major achievement.

7.2. Federated storage
The high level objective of federating storage is the provisioning of tools for conveniently storing, federating, accessing and sharing huge quantities of data. The resulting toolbox is mainly targeting scientific communities, who are not willing or not able to develop their entire data management framework themselves. The selection of services and products within that toolbox is based on their use of open standards and their availability on the open source market. Even more importantly, significant focus has been put on the evaluation of the potential self-sustainability of components, due to an active user community or due to the commitment of the product teams to further maintain their products. Besides integrating well established and sustainable data management components, LSDMA evaluated gaps in existing data management procedures and, in response, either established working groups in international scientific organizations, like the Research Data Alliance (RDA)[5] or joined existing taskforces in industry, like Storage Networking Industry Association (SNIA)[6], on those topics. As those activities naturally require agreements on the European and possibly international level, DSIT partners successfully joined European projects, like the INDIGO-DataCloud or AARC, engaging a larger group of communities.

---

[2] https://indigo-datacloud.eu
[3] https://aarch-project.eu
[4] https://eudat.eu
[5] https://rd-alliance.org
[6] https://snia.org

7.3. Metadata

Managing metadata in a generic way is of essential importance in scientific data life cycles [24]. It needs to be both efficient and seamless. Such a concept was designed and implemented within the MoSGrid [25, 26] Science Gateway. The utilized UNICORE Metadata Management is a generic service within the UNICORE HPC middleware. The concept resulted in the DFG project MASi [27] that utilizes the repository framework KIT Data Manager to built up a generic metadata-driven research data management service. The initial use cases are situated in geography, chemistry, and digital humanities. Furthermore, metadata developments in general and provenance support specifically for the generic web processing framework birdhouse are introduced. This includes their applications within the earth sciences. We highlight the valuable contributions in generic and specific metadata management that were designed and implemented.

7.4. Archives

Scientific and cultural organizations, international collaborations and projects have a need to preserve and maintain access to large volumes of digital data for several decades. Existing systems supporting these requirements span from simple databases at libraries to complex multi-tier software environments developed by scientific communities. All communities see an increasing volume of data that must be stored efficiently and economically, which today is usually a combination of storage on disk and tape. Development and integration of components to enable secure and reliable archival storage that make use of existing computer centre infrastructures is a long standing goal in LSDMA. The project brings together diverse communities and functions as pivot for generic solutions. At the same time requirements have been collected to support long term access to data for multiple scientific domains and international projects. The material was used in accompanying projects to implement an infrastructure for long time storage, develop easy access to archives and enable new user groups.

7.5. Performance analysis

Managing and analysing large amounts of data requires high performance storage systems that can keep up with the applications' I/O demands. Additionally, energy efficiency plays an important role as storage systems are often responsible for a significant part of the total cost of ownership. Within the Performance and Power Optimization work package of the Data Services Integration Team, we have focused on both of these aspects. Based on demands observed in real systems and applications, we have developed tools and solutions to improve both performance [28] and cost efficiency [29, 30].

7.6. Data Intensive Computing (DIC)

Besides data management, the analysis of research data is another important aspect covered by LSDMA. The overarching goal of all data efforts must be to start managing the scientist's data right after the data left the data acquisition device as this is the only way to capture gapless provenance information. This is inevitable for transparent and reproducible science. However, this means also that the research data is at the very beginning of its lifecycle. In order to obtain publishable results the data often has to go through several processing steps until the final results are available [31]. These steps can reach from basic scripts to complex scientific workflows consisting of several dependent processing steps. In order to achieve the aforementioned reproducibility capturing the data provenance, e.g. what happened with the data and led to which results, should be an essential part of every step.

8. Conclusions

After five years it is fair to claim that LSDMA has advanced selected scientific communities in their data management and analysis. Some selected highlights are presented and referenced in

this note. From our experience in the project we draw the following conclusions on lessons we learned:

The variety of topics presented here already show that the needs of communities vary immensely, even within research areas. The communities interest in new tools and methods is fueled by need and by new research potential. The automation of procedures and of workflows may still boost a scientists' efficiency in performing research significantly. One important contribution for interoperability is the design of an Authentication and Authorization Infrastructure. This work continues in the context of EU-projects. Policies (e.g. Open Data) and legal regulations (e.g. data privacy) are additional challenges.

The clear separation of domain specific methods and generic data methods was sometimes not that obvious as certain communities are the main drivers for generic methods or tools. However, the success of various community projects and the highlights presented here justify the dual approach with DLCLs and the DSIT.

Acknowledgements

References
[1] Jung C, Gasthuber M, Giesler A, Hardt M, Meyer J, Prabhune A, Rigoll F, Schwarz K and Streit A 2015 *Journal of Physics: Conference Series* **664** 032018 URL `http://stacks.iop.org/1742-6596/664/i=3/a=032018`

[2] Jung C, Gasthuber M, Giesler A, Hardt M, Meyer J, Rigoll F, Schwarz K, Stotzka R and Streit A 2014 *Journal of Physics: Conference Series* **513** 032047 URL `http://stacks.iop.org/1742-6596/513/i=3/a=032047`

[3] Szuba M, Ameri P, Grabowski U, Meyer J and Streit A 2016 A distributed system for storing and processing data from earth-observing satellites: System design and performance evaluation of the visualisation tool *2016 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)* pp 169–174

[4] Maatouki A, Meyer J, Szuba M and Streit A 2015 A horizontally-scalable multiprocessing platform based on node.js *2015 IEEE Trustcom/BigDataSE/ISPA* vol 3 pp 100–107

[5] Rigoll F and Schmeck H 2014 Konzeption eines energiedatenmanagementsystems unter beachtung von datenschutz und privatsphäre *VDE-Kongress 2014* ed VDE VDE (VDE VERLAG GmbH)

[6] Müller P, Weiland Y, Kaufmann R, Gunkel M, Hillebrandt S, Cremer C and Hausmann M 2012 *Analysis of fluorescent nanostructures in biological systems by means of Spectral Position Determination Microscopy (SPDM)* vol 1 pp 3–12

[7] Cremer C *et al.* 2011 *Biotechnology Journal* **6** 1037–1051

[8] Prabhune A, Stotzka R, Jejkal T, Hartmann V, Bach M, Schmitt E, Hausmann M and Hesser J 2015 An optimized generic client service api for managing large datasets within a data repository *Big Data Computing Service and Applications (BigDataService), 2015 IEEE First International Conference* (IEEE) pp 44–51 URL `http://ieeexplore.ieee.org/abstract/document/7184863/`

[9] Jejkal T, Vondrous A, Kopmann A, Stotzka R and Hartmann V 2014 *Large-Scale Data Management and Analysis-Big Data in Science-*

[10] Prabhune A, Zweig A, Stotzka R, Gertz M and Hesser J 2016 Prov2one: An algorithm for automatically constructing provone provenance graphs *International Provenance and Annotation Workshop* (Springer) pp 204–208 URL `https://link.springer.com/chapter/10.1007/978-3-319-40593-3_22`

[11] et al M G 2017 *Journal of Physics: Conference Series* To be published

[12] Strutz M, Gasthuber M, Aplin S, Dietrich S, Kuhn M, Ensslin U, Smirnov G, Lewendel B and Guelzow V 2015 *Journal of Physics: Conference Series* **664** 042053 URL `http://stacks.iop.org/1742-6596/664/i=4/a=042053`

[13] Schwarz K 2011 *Journal of Physics: Conference Series* **331** 052018 URL `http://stacks.iop.org/1742-6596/331/i=5/a=052018`

[14] Bagnasco S, Betev L, Buncic P, Carminati F, Cirstoiu C, Grigoras C, Hayrapetyan A, Harutyunyan A, Peters A J and Saiz P 2008 *Journal of Physics: Conference Series* **119** 062012 URL `http://stacks.iop.org/1742-6596/119/i=6/a=062012`

[15] Axer M, Amunts K, Grssel D, Palm C, Dammers J, Axer H, Pietrzyk U and Zilles K 2011 *NeuroImage* **54** 1091 – 1101 ISSN 1053-8119 URL `http://www.sciencedirect.com/science/article/pii/S105381191001178X`

[16] Amunts K, Bücker O and Axer M 2014 *Towards a Multiscale, High-Resolution Model of the Human Brain* (Cham: Springer International Publishing) pp 3–14 ISBN 978-3-319-12084-3 URL `http://dx.doi.org/10.1007/978-3-319-12084-3_1`

[17] Streit A, Bala P, Beck-Ratzka A, Benedyczak K, Bergmann S, Breu R, Daivandy J M, Demuth B, Eifer A, Giesler A, Hagemeier B, Holl S, Huber V, Lamla N, Mallmann D, Memon A S, Memon M S, Rambadt M, Riedel M, Romberg M, Schuller B, Schlauch T, Schreiber A, Soddemann T and Ziegler W 2010 *annals of telecommunications - annales des télécommunications* **65** 757–762 ISSN 1958-9395 URL `http://dx.doi.org/10.1007/s12243-010-0195-x`

[18] Biancini A, Florio L, Haase M, Hardt M, Jankowski M, Jensen J, Kanellopoulos C, Liampotis N, Licehammer S, Memon S, van Dijk N, Paetow S, Prochazka M, Sall M, Solagna P, Stevanovic U and Vaghetti D 2016 *Arxiv* URL `https://arxiv.org/abs/1611.07832`

[19] Sakimura N, Bradley J, Jones M B, de Medeiros B and Mortimore C 2014 *The OpenID Foundation* URL `http://openid.net/specs/openid-connect-core-1_0.html`

[20] Cantor S, Kemp J, Philpott R and Maler E 2006 *OASIS Standard* URL `http://docs.oasis-open.org/security/saml/v2.0/saml-core-2.0-os.pdf`

[21] Ceccanti, Hardt, Wegh, Millar, Licehammer, Caberletti and Vianello 10-14 October 2016 *22nd International Conference on Computing in High Energy and Nuclear Physics (CHEP2016)* Poster

[22] Amunts K, Ebell C, Muller J, Telefont M, Knoll A and Lippert T 2017 *Neuron* 574 – 581

[23] Benedyczak K, Schuller B, Petrova M, Rybicki J and Grunzke R 2016 Unicore 7 - middleware services for distributed and federated computing *International Conference on High Performance Computing Simulation (HPCS)* URL `http://dx.doi.org/10.1109/HPCSim.2016.7568392`

[24] Grunzke R 2016 *Generic Metadata Handling in Scientific Data Life Cycles* Ph.D. thesis Doctoral Thesis, Technische Universität Dresden URL `http://nbn-resolving.de/urn:nbn:de:bsz:14-qucosa-202070`

[25] Grunzke R, Birkenheuer G, Blunk D, Breuers S, Brinkmann A, Ge sing S, Herres-Pawlis S, Kohlbacher O, Krüger J, Kruse M, Müller-Pfefferkorn R, Schäfer P, Schuller B, Steinke T and Zink A 2012 A data driven science gateway for computational workflows *UNICORE Summit 2012 Proceedings* (*IAS Series* vol 15) pp 35–49 ISBN 978-3-89336-829-7 URL `http://hdl.handle.net/2128/4705`

[26] Krüger J, Grunzke R, Gesing S, Breuers S, Brinkmann A, de la Garza L, Kohlbacher O, Kruse M, Nagel W E, Packschies L, Müller-Pfefferkorn R, Schäfer P, Schärfe C, Steinke T, Schlemmer T, Warzecha K D, Zink A and Herres-Pawlis S 2014 *Journal of Chemical Theory and Computation* **10(6)** 2232–2245

[27] Grunzke R, Hartmann V, Jejkal T, Prabhune A, Herres-Pawlis S, Hoffmann A x, Deicke A, Schrade T, Herold H, Meinel G, Stotzka R and Nagel W E 2016, accepted Towards a Metadata-driven Multi-community Research Data Management Service *2016 8th International Workshop on Science Gateways (IWSG)*

[28] Chasapis K, Dolz M, Kuhn M and Ludwig T 2014 Evaluating Power-Performace Benefits of Data Compression in HPC Storage Servers *IARIA Conference* ed Fries S and Dini P (IARIA XPS Press) pp 29–34 ISBN 978-1-61208-332-2 ISSN 2308-412X

[29] Kuhn M, Chasapis K, Dolz M and Ludwig T 2014 Compression By Default - Reducing Total Cost of Ownership of Storage Systems *Supercomputing* (*Lecture Notes in Computer Science* no 8488) ed Kunkel J M, Ludwig T and Meuer H W (Berlin, Heidelberg: Springer International Publishing) ISBN 978-3-319-07517-4 ISSN 0302-9743

[30] Kunkel J, Kuhn M and Ludwig T 2014 *Supercomputing Frontiers and Innovations* 116–134 URL `http://superfri.org/superfri/article/view/20`

[31] Grunzke R, Jug F, Schuller B, Jäkel R, Myers G and Nagel W E 2017 Seamless hpc integration of data-intensive knime workflows via unicore *Euro-Par 2016: Parallel Processing Workshops: Euro-Par 2016 International Workshops, Grenoble, France, August 24-26, 2016 , Revised Selected Papers* (Springer International Publishing) pp 480–491 ISBN 978-3-319-58943-5 URL `http://dx.doi.org/10.1007/978-3-319-58943-5_39`