Challenges of Research Data Management for High Performance Computing

Björn Schembera(✉) and Thomas Bönisch

High Performance Computing Center Stuttgart (HLRS),
University of Stuttgart, Nobelstr. 19, 70569 Stuttgart, Germany
{schembera,boenisch}@hlrs.de

Abstract. This paper targets the challenges of research data management with a focus on High Performance Computing (HPC) and simulation data. Main challenges are discussed: The Big Data qualities of HPC research data, technical data management, organizational and administrative challenges. Emerging from these challenges, requirements for a feasible HPC research data management are derived and an alternative data life cycle is proposed. The requirement analysis includes recommendations which are based on a modified OAIS architecture: To meet the HPC requirements of a scalable system, metadata and data must not be stored together. Metadata keys are defined and organizational actions are recommended. Moreover, this paper contributes by introducing the role of a Scientific Data Manager, who is responsible for the institution's data management and taking stewardship of the data.

Keywords: Research data management · HPC · Simulation · Big data · Archive · OAIS · Metadata · Data life cycle

1 Introduction

Today's science can be considered as data-driven. Research data is all scientific data generated or recorded from experiments, studies or simulations. In contrast to theory and classical experiments, simulations produce huge amounts of big research data [19], usually in size of Petabytes (PB). High performance computing (HPC) is one of the driving forces behind big research data enabling large-scale simulations in climate research, engineering or particle physics just to name a few. For researchers it is crucial to keep their data for review or later resumption of work. However, the ability to store and especially manage research data is lagging behind the ability to generate data [15] in HPC: For example, the sheer volume of the data is a specific problem, but not the only critical one.

The following work presents the challenges of research data management of simulation data in the scope of HPC. The first challenge is the problem of Big Data: volume and variety. As a second challenge, insufficient data management concepts will be discussed. Moreover, research data management is not only a technical but also a organizational problem in HPC: A lack of data management plans, regulations and incentives.

J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, pp. 140–151, 2017.
DOI: 10.1007/978-3-319-67008-9_12

Derived from these challenges, requirements for a feasible research data management in HPC are specified in Sect. 3. This requirement analysis is the main contribution of this paper and includes data management requirements such as metadata, persistent identifiers and data security. Since Open Access will become a key requirement in the future, it will be discussed in a separate subsection. For HPC, scalability requirements are important: Research data management has to cope with the volume and has to provide efficient indexing mechanisms for feasible search of millions of data objects. Since research data management is not only a technical problem, one contribution is the introduction of a new role: The Scientific Data Officer (SDO) that is in charge of the research data management efforts of an institution. All these efforts lead to an improved data life cycle being able to reduce "dark" data. Related work is presented in Sect. 3.7.

2 Challenges

2.1 Big Research Data: Volume and Variety

The data volume produced on an HPC system strongly depends on the amount of main memory of a supercomputer. A DoE study [12] estimates the data volume factor to be 1:35 in worst case: For each Byte of main memory of the compute system, 35 Byte of data to archive is created per year, so the amount of data scales linear with the amount of main memory of the HPC system. This means every time a new HPC system is deployed, an increase in data production (due to a more fine-grained resolution or due to larger scales) has to be expected. The growth over time of the research or during the overall system lifetime data volume must be expected to be exponential, the study concludes. A follow-up study reminds that a storage technology gap exists for Exascale [13]. Figure 1 shows a sample trend (2010–2017) of the data stored in a tape archive at the HLRS related to the main memory of the corresponding HPC flagship systems.

The number of files does not depend on the amount of main memory but on the number of cores and processors, however no estimate can be given. A reason for this is that it strongly depends on the behavior of the researchers how many files are written or if file aggregation is used. Nevertheless, the study states that the number of files to archive per system will grow exponentially from millions to billions during the next 5 years.

Regarding variety, research data is strongly diverse. Most data is formatted according to the researchers bias. For example, as comma-separated values, tables or stored in files of different formats [22]. How this data is organized and managed is another challenge that will be discussed in the following subsection.

2.2 Research Data Management

Research data management in HPC is often handled by the directory structures and an appropriate naming of files and directories, such as

```
/group/project/user/simulation/run/description.format
```

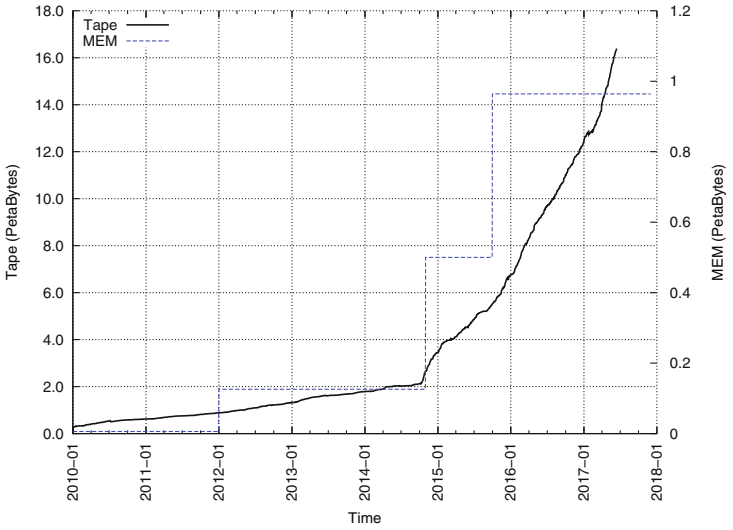

Fig. 1. Data stored (left y-axis/continuous line) in the HLRS HPC center in Stuttgart, Germany in relation to the main memory of a system (right y-axis/dashed line). A dependency is easy to see. The ratio is higher than 1:10 in this center. As of June 2017, 16 PB of data in approximately 9.2 million files of almost 200 users are held on tape.

In doing so, searching, finding and retrieving research data becomes a burden [1,15]: A third person is unable to find data if the person does not have information on who ran the simulation, which group was leading which project or which project acronym is used. Additionally, the directory structures are highly dynamic, for example when a parameter has to be varied unexpectedly.

Moreover in most cases, no explicit metadata is attached to these files. This means that the description what the files contain and how they can be interpreted is nowhere (formally) written down. If metadata is attached to the files, besides file names, it may reside in text files, spreadsheets or encoded in the output files [15]. There are only little common formats for storing and especially describing research data [22], such as NetCDF. It strongly depends on the community whether these possibilities are used, for example in earth sciences with NetCDF [18] or with HDF5 at NASA [19].

2.3 Organizational Workflows

The challenges discussed above have been rather technical issues of scalability. However, research data management is not only a technical management task. What is lacking nowadays for example are incentives for researchers to perform the additional work of tagging their data with metadata and storing the data in a research data management system.

Moreover, plans how to organize data are often missing; specifications that define roles, responsibilities, timelines and descriptions of the data are lacking.

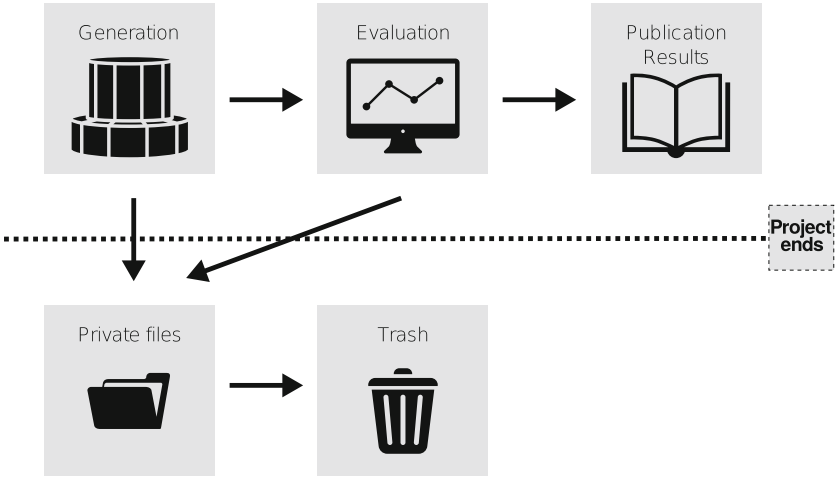

Fig. 2. Today's data life cycle: After generation, evaluation of data and latter publication of the results, the data is put to the private files. There it is deleted or forgotten.

It is often unclear who is taking the stewardship of big research data [19]. Typically, one person of a research group is pushed into a vague role of handling the institution's data. Researchers are acting on their own because there is no assistance offered. This is due to the fact that either the expertise is not available or distributed through the research institution. This is also a question of missing training: How will a researcher get in touch with the tools, practices and regulations of data management? According to a study conducted in the UK [3], lacking skills are one main challenge of data management in general.

Open Access is the paradigm to make research data publicly available and is another challenge in HPC context. Open Access is still not commonly accepted - this is also a problem of regulations and conventions by institutions as well as a legal challenge. However, Open Access will become a top priority in the future [3]. A European Union report argues that in the future, all generated research data has to be made available to the public [7]. In addition, the German Research Foundation (DFG) sets as a vision that publicly funded research data should be publicly available for a long-term period [5]. This prevents duplicate work and saves resources.

2.4 Research Data Life Cycle

Today's HPC research data life cycle is shown in Fig. 2 and can be described as follows: While computing the simulation jobs, data is continuously written to a parallel filesystem, such as Lustre. After the generation of the data, evaluation follows. After the end of the project, the knowledge gained through the results is published. Researchers see the end of data management when the publication is written: Scientific results are communicated and published in papers,

where condensed information and statistics are presented. There is no room for referencing research data, so after publication the data is either forgotten in the private files, on tape storage systems or even deleted. In this way, the data becomes dark data, since reusing or reviewing the data is hardly possible [10]. This is not only a problem in HPC but in all big data sciences [11] and runs contrary to good scientific practice, as for example the DFG recommends [5].

3 Requirements for HPC Research Data Management

3.1 Scalability Requirements

To cope with the above challenges, a three-layered architecture is proposed as a technical foundation. It is based on the OAIS model and Askhojs approach, as described in Sect. 3.7 and consists of a storage layer, an object layer and a user layer. The two bottom layers are critical with respect to the scalability challenges of volume and variety.

The storage layer is logically the lowest component and has to handle exponential data growth. It must be possible to easily extend the storage space. This requirement is not specific to HPC but gets critical here [13]. Tape media is the only media that meets the HPC requirements implying low total cost of ownership, large data volume and low error rate [8]. The storage layer of a research data management system needs the possibility to be distributed if the amount of data is too large to store in one single data center[1]. Classic DBMS like DB2 are not suitable for storing the actual data since they are unable to store data in the size of TeraBytes. The storage layer has to be a cost-effective mass storage system on which the object management can be built.

The management of the data objects is located on the object layer which is logically on top of the storage layer. The object layer makes the data stored in the storage layer search- and findable. In HPC this is an important requirement since millions of files are not unusual. According to a report, metadata performance is already critical but will become crucial in the Exascale [13]. Searching must be performed in a feasible time that can only be achieved if metadata and data are not stored together albeit violating the OAIS paradigm of containers holding both data and metadata. Storing only the reference to data residing on another layer of the system, queries can be performed as a first step for retrieving the actual data. Askhojs layered architecture is preferable for HPC due to reasons of flexibility and the distributed character of the data management system. However, Askhojs design to bring OAIS to the cloud is not suitable for the HPC use case since it would not scale for big data volumes and growth. The connection to a cloud storage service would be too poor to transfer huge amounts of data in a feasible time (i.e., retrieving 50TB via a 1GB/s cloud link would take approx. 5 days). Moreover, due to a lack of integration in the HPC workflow and security considerations, cloud services as a backend are not preferable.

[1] For example, the data produced by the CERN/LHC experiments is distributed to data centers all over Europe. See: https://home.cern/about/computing/worldwide-lhc-computing-grid, last accessed Nov 28th 2016.

3.2 Data Management Requirements

Metadata is the main concept to handle data [9,23]. Metadata is "data about data" and describes data from a logical point of view as well es from its attributes in a structured form. Enriched with metadata, data becomes a valuable object. For a feasible HPC research data management, the following parts have to be implemented: First, a reasonable metadata scheme that identifies generic as well as domain-specific characteristics of HPC simulation data has to be defined (for example in XML). Following the OAIS metadata scheme (details in Sect. 3.7), structural, administrative and descriptive metadata has to be included. Second, a suitable storage for metadata must be located in the object layer and has to be reliable, safe and performant, for example on mirrored SSDs. Third, efficient index mechanism to search and explore billions of files is needed, which is also a matter of scalability. Following metadata keys for HPC are mandatory:

Descriptive Metadata describes the data content. Important descriptive keys for searching in an HPC context are *Authors*, *Name*, *Filename*, *Creation Date*, *Access Date*, *Change Date*, *Keywords*, *System*, *Compiler*, *Compiler Flags*, *Batch System*, *Size*, *Algorithm*, *Context*, *Publications*. There must be an additional field for *domain-specific metadata*. For example in CFD, this could include the Reynolds number, the cases and the exact turbulence model used.

Preservation Metadata or administration metadata is all the metadata ensuring the long-term preservation. Reasonable keys can be derived from the OAIS metadata model and include a *Persistent identifier* (PID) key for the location of the data, such as Handle[2] or ePIC[3]. A PID allows data citation, since data can be uniquely identified. *Provenance information* incorporates information about the origin and the changes in form of list. The *Context information* key holds information that links the data object to others and is a list of persistent identifiers of other data objects. A *fixity* metadata key contains information that ensures the integrity of a data object. This is a checksum of the data object, such as a MD5 hash. An *access rights* key carries information on the access right to the data object. This also includes time limits and embargoes. Additionally to the management of data, preservation metadata plays a crucial role for fulfilling the requirements of data security, Open Access due to the usage of PIDs as well as possibilities of distributing the data over several data centers.

Content Metadata is mostly held by the data itself. However, it may be useful to store it additionally in the metadata store. For example, a key as *Format* should be stored here.

User Interface and Workflow Integration. A user interface that allows browsing, searching, injecting, manipulating of data objects in the research data management system builds the user layer. Possibilities to enter metadata and link it to data is mandatory. The user interface must be able to run in a distributed environment: Since the location of the data stored may differ from the location

[2] http://handle.net/, last accessed Nov 26th 2016.
[3] http://www.pidconsortium.eu/, last accessed Nov 26th 2016.

from where data is accessed or managed, location transparency is required that allows a single system view. The system must be accessible via a low-level client on the HPC frontends to perform metadata tagging of created files or retrieving data back for further analysis. All the above points have to be integrated in the HPC workflow seamlessly. Tagging of metadata has to be possible both at creation time of the files (like in the iCurate system discussed in Sect. 3.7) and at the time the files are injected into the system.

3.3 Security Requirements

Data security must be guaranteed by the data management system [23]. First, this means bitstream preservation which is the preservation of bits on a physical layer and also includes the ability to retain the bits if technology changes. Two copies of the data on tape are recommended, preferably in physically distinct locations. Second, this also has to include the fixity of the data objects, that is its integrity. Checksums can guarantee that the data has not been altered. Fixity checks have to be included. Third, this also includes end-to-end data integrity [13]. Fourth, encryption must be possible. These requirements are general requirements and not bound to HPC, but have to take into account the characteristics of big research data, such as how to feasibly perform integrity checks on Petabytes of data. This topic overlaps with the scalability requirement discussed in Subsect. 3.1 and affects components in both the storage and the object layer.

3.4 Open Access Requirements

Open Access is crucial for HPC as a data-driven science and will be raised as a key requirement in the future, as discussed in Sect. 2.3. This is also a technical challenge: On the user layer, an interface has to be provided that is accessible for the world to retrieve data publicly for the concrete use case. Transfer technologies have to be found that are able to move big research data. While publishing or moving the data, metadata annotations have to remain. This can be achieved by including a PID, as described in Sect. 3.2. Moreover, incentives have to be pushed and regulations have to be implemented that make researchers publishing their research data as Open Access.

3.5 Organizational Requirements

Incentives. There are extrinsic and intrinsic reasons why scientists and institutions should participate in research data management efforts. Extrinsic reasons are external influences, regulations or institution-wide standards. For example in Germany, the DFG advises to store scientific research data for at least 10 years [5]. Intrinsic reasons aim for the stakeholders themselves: Reasons for the researcher like the higher reputation when publishing the research data or an easier way to reuse the data after years. These incentives have to be analyzed case-by-case, pushed and incorporated by all future data management approaches.

Data Management Plans (DMP) are formal documents that specify plans and numbers on data management and fulfill the requirement of organizational security [14]. This documentation has to name how research data is handled during and after a research project. Research proposals already require management plans, for example those of the European Union [6] or of the National Science Foundation [20]. A DMP has therefore to be mandatory for all HPC research data management efforts and must be raised as a general requirement when handling research data. All persons involved have to negotiate on a DMP[4].

The DMP has to include a part *Data Description and Metadata*, where the data (and their provenance) should be described as well as the metadata keys used. It has to be specified how, when and where the data is produced. This is of importance for HPC: For example, the data has to be treated differently if it resides on a parallel scratch filesystem than if it resides on tape.

A *Timeline* has to be sketched out. It has to be specified how data is managed during all phases of the HPC workflow. The DMP has to define what tools are used in each step to transfer data back and forth and keep track of the data.

Moreover, *Organizational Topics* have to be covered by the DMP. In the document, the responsibilities have to be defined, that means a SDO has to be named. Legal issues have to be addressed. Access rights have to be defined: Who will be able to access the data at which time? Will the data be made publicly available as Open Access? This part also has to include how data management costs will be covered.

Qualification. Another organizational requirement is qualification [19]. Courses, trainings and integration into curricula have to be provided by and in institutions that apply data management. Persons that later can act as multipliers and end-users have to get data management skills. Only by increasing training activities, scientists will benefit from research data management efforts.

Scientific Data Officer. A person within a project, a department or an institution has to be defined taking the responsibility for data management and the stewardship of the data. This person assumes the role of the Scientific Data Officer (SDO). The SDO has the same position as the security officer with respect to data: Being aware of the trends in research data management, HPC-related tools and infrastructures and knowledge of in-house data storage facilities. The person has to be trained in respective courses. Within the research group or institute, this person has to act as a multiplier and transfer knowledge to the group in talks or on request. Moreover, the data officer has the stewardship of orphaned data that is preliminary dark data still existing when a person leaves the group. If the institution runs own data management systems, technical administration and support for these systems may also be in charge of the SDO.

[4] There are online tools available for specifying a DMP, such as: https://dmponline.dcc.ac.uk/, last accessed Nov 25th, 2016.

3.6 An Alternative Data Life Cycle

All the requirements for a feasible research data management for HPC should in the end lead to an improved, ideal data life cycle, as depicted in Fig. 3. Research data management are combined measures of both systems and organization and they have to take effect in all the steps if research data is involved. When data is generated, it has to be enriched with metadata and put to the data management system along with the metadata and a PID. Data with according metadata gained for the evaluation has to be archived as well. Only with data and metadata, the results can be re-evaluated, or checked after the project end. A well-integrated user-interface must enable Open Access to the data. A DMP has to define all the processes, data movements and timelines that occur within the data management process and the SDO has to take responsibilities and coordinate all actions. The SDO also has to take stewardship of all orphaned data. The data life-cycle does not end when the project is finished: The data and metadata remains active in the research data management system for reuse or Open Access.

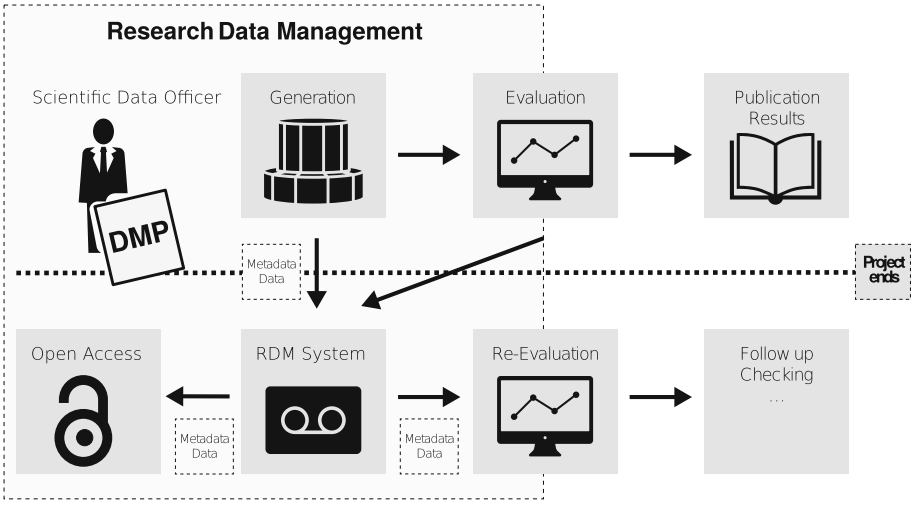

Fig. 3. Ideal data life cycle: Data management affects the generation, evaluation and archiving steps in an HPC ecosystem. Data management is not only a technical, but also a organizational task and is led by the Scientific Data Officer (SDO). A Data Management Plan (DMP) defines all processes and timelines.

3.7 Related Work

Archiving. The Open Archive Information System [21] offers a reference model for archive systems which defines the interaction between humans and machines and moreover proposes a metadata model. As a specification, OAIS is focused on conceptual work and not implementation. The model comprehends three roles:

The producer of data, the consumer and the management. These three parties have to negotiate on preservation planning, administration and data management. Procedures for data ingest and access have to be defined and archival storage must be planned as a reliable, long-term storage. As a framework, OAIS proposes concepts and it is always up to the specific use case in which the archive is used. Emerging from the field of record preservation in the cloud, Askhoj et al. [2] propose to map OAIS to a layered architecture and bring the archive to the cloud in order to combine a well-established concept for archiving with the flexibility and scalability of cloud systems. A PaaS-layer, handling objects as binary strings and ensuring bitstream preservation. The SaaS layer handles digital objects that emerge as objects when they are packaged at the next layer, the packaging layer. A fourth layer called Archives and Records Management layer incorporates all management capabilities and a user interface. In contrast to the OAIS specification, Askhoj et al. argue that it is more beneficial to split data and metadata. They introduce persistent identifiers to reference the data. iCurate is a data management system that is not based on OAIS but has a strong focus on the management layer and metadata [17]: Annotation, retrieval, and validation of metadata should be possible. The system is adapted to the HPC workflow, that means users can specify already in the Portable Batch System (PBS) file some metadata which is added to the output files when the job is complete. The iCurate system can then also harvest the output files for technical metadata to be automatically added. The main contribution of iCurate is the automating of workflows for the annotation of metadata.

Metadata. It is specific to OAIS, that data and metadata is bonded together to a data object. According to the OAIS reference model, there are three major metadata categories. *Content Information* or structural metadata refers to the data object itself with associated, necessary representation information, such as information on the format of the data. Without this representation information, data would not be machine-readable any more and hence become worthless. *Preservation Description Information* or administrative metadata consists of five subcategories. First, Reference Information in general is a unique identifier to locate the data. Second, Provenance Information holds information about the origin of data as well as the history of changes. Third, Context Information describes the relations between digital objects. Fourth, Fixity Information protects the content from unauthorized alteration and may be realized by a checksum. Lastly, Access Rights are access policies. *Descriptive Information* can consist of a whole set of attributes describing properties of the object that emerge to a higher level description of the data. Other existing schemes like DataCite [4] introduce general elements such as *identifier* or *format* to tag and identify data and can be used as a framework to add more specific metadata attributes. The Climate and Environmental Retrieval and Archive system (CERA) data model is a domain-specific metadata model developed for earth sciences at the DKRZ [16].

4 Conclusion

The paper presented the challenges of research data management for HPC. In HPC, simulation data is big in volume and variety. Data reproducibility to diminish the problem of volume is not given for HPC since machines and compilers are renewed every 3 to 5 years. Data management becomes a burden since current solutions disrespect HPC characteristics such as having millions to billions of huge files. Moreover, organizational challenges such as Open Access policies get critical in HPC: Besides legal issues for example, publishing Petabytes of data is not a trivial task. Nowadays, the data life cycle produces a lot of dark data becoming worthless.

Derived from these challenges, requirements for a feasible research data management have been outlined: On the side of technical management, those were metadata, persistent identifiers, data security and workflow integration. To cope with Open Access, research data management has to incorporate the idea and include an appropriate user interface. Scalability requirements aim for providing technologies that can deal with the huge data volume and variety of files. This can be accomplished by a three-layered architecture, separating data storage and metadata storage due to performance considerations. Since research data management is not only a technical task, organizational requirements have been defined, such as qualification and planning documents. Moreover, the role of the SDO has been introduced as one contribution of this paper: Only with a skilled person taking stewardship of research data activities in an institution, research data management can be successful in the interplay of human and machine. In the end, the requirements lead to a data life cycle for HPC where research data management affects all the stages where data is involved and does not end when the project is over.

Acknowledgments. We would like to thank *Wanda Spahn* for proofreading.

References

1. Arora, R.: Data management: state-of-the-practice at open-science data centers. In: Khan, S.U., Zomaya, A.Y. (eds.) Handbook on Data Centers, pp. 1095–1108. Springer, New York (2015). doi:10.1007/978-1-4939-2092-1_37
2. Askhoj, J., Sugimoto, S., Nagamori, M.: Preserving records in the cloud. Rec. Manage. J. **21**(3), 175–187 (2011). https://doi.org/10.1108/09565691111186858
3. Cox, A.M., Pinfield, S.: Research data management and libraries: current activities and future priorities. J. Librarian. Inf. Sci. **46**(4), 299–316 (2014). http://dx.doi.org/10.1177/0961000613492542
4. DataCite: (2016). http://schema.datacite.org/. Accessed 6 Dec 2016
5. DFG: Safeguarding good scientific practice (2013). http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf. Accessed 6 Dec 2016
6. EU: H2020 programme guidelines on FAIR data management in Horizon 2020 (2016). http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf. Accessed 6 Dec 2016

7. EU: European Cloud Initiative - Building a competitive data and knowledge economy in Europe (2016). http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=15266. Accessed 6 Dec 2016
8. Faulhaber, P.: Investing in the future of tape technology. Presentation, HPSS User Forum, New York City (2015)
9. Gray, J., Liu, D.T., Nieto-Santisteban, M., Szalay, A., DeWitt, D.J., Heber, G.: Scientific data management in the coming decade. SIGMOD Rec. **34**(4), 34–41 (2005). http://doi.acm.org/10.1145/1107499.1107503
10. Heidorn, P.B.: Shedding light on the dark data in the long tail of science. Libr. Trends **57**(2), 280–299 (2008). http://doi.org/10.1353/lib.0.0036
11. Helly, J., Staudigel, H., Koppers, A.: Scalable models of data sharing in earth sciences. Geochem. Geophy. Geosyst. **4**(1) (2003). http://dx.doi.org/10.1029/2002GC000318
12. Hick, J.: HPSS in the Extreme Scale Era: Report to DOE Office of Science on HPSS in 2018–2022. Lawrence Berkeley National Laboratory (2010)
13. Hick, J.: The Fifth Workshop on HPC best practices: File systems and archives. Lawrence Berkeley National Laboratory. LBNL Paper LBNL-5262E (2013)
14. Jensen, U.: Datenmanagementpläne. In: Büttner, S., Hobohm, H.-C., Müller, L. (eds.) Handbuch Forschungsdatenmanagement. Bad Honnef: Bock u. Herchen (2011)
15. Jones, S.N., Strong, C.R., Parker-Wood, A., Holloway, A., Long, D.D.E.: Easing the burdens of HPC file management. In: Proceedings of the Sixth Workshop on Parallel Data Storage, PDSW 2011, NY, USA, pp. 25–30 (2011). http://doi.acm.org/10.1145/2159352.2159359
16. Lautenschlager, M., Toussaint, F., Thiemann, H., Reinke, M.: The CERA-2 data model (1998). https://www.pik-potsdam.de/cera/Descriptions/Publications/Papers/9807_DKRZ_TechRep.15/cera2.pdf
17. Liang, S., Holmes, V., Antoniou, G., Higgins, J.: iCurate: a research data management system. In: Bikakis, A., Zheng, X. (eds.) MIWAI 2015. LNCS, vol. 9426, pp. 39–47. Springer, Cham (2015). doi:10.1007/978-3-319-26181-2_4
18. Malik, T.: Geobase: indexing NetCDF files for large-scale data analysis. In: Big Data Management, Technologies, and Applications, pp. 295–313. IGI Global (2014). http://doi.org/10.4018/978-1-4666-4699-5.ch012
19. Mattmann, C.A.: Computing: a vision for data science. Nature **493**(7433), 473–475 (2013). http://dx.doi.org/10.1038/493473a
20. NSF: Grant proposal guide chapter ii.c.2.j (2014). https://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/gpg_2.jsp#dmp. Accessed 6 Dec 2016
21. OAIS: Reference model for an Open Archival Information System. Technical report, CCSDS 650.0-M-2 (Magenta Book) Issue 2 (2012)
22. Parker-Wood, A., Long, D.D.E., Madden, B.A., Adams, I.F., McThrow, M., Wildani, A.: Examining extended and scientific metadata for scalable index designs. In: Proceedings of the 6th International Systems and Storage Conference, SYSTOR 2013, NY, USA, pp. 4:1–4:6 (2013). http://doi.acm.org/10.1145/2485732.2485754
23. Potthoff, J., van Wezel, J., Razum, M., Walk, M.: Anforderungen eines nachhaltigen, disziplinübergreifenden Forschungsdaten-Repositoriums. In: DFN-Forum Kommunikationstechnologien, pp. 11–20 (2014)