Management of Research Data in Computational Fluid Dynamics and Thermodynamics

Björn Selent[1], Hamzeh Kraus[2], Niels Hansen[2], Björn Schembera[3], Anett Seeland[4,5] and Dorothea Iglezakis[5]

[1]Institute of Aerodynamics and Gasdynamics, University of Stuttgart, Germany;
[2]Institute of Thermodynamics and Thermal Process Engineering, University of Stuttgart;
[3]High-Performance Computing Center Stuttgart;
[4]Communication and Information Center, University of Stuttgart;
[5]University Library, University of Stuttgart

The performance increase of the available resources in the HPC area offers the possibility to investigate fundamental questions of fluid mechanics with numerical tools in high temporal and spatial resolution. In particular, turbulent flows, which account for the largest part of flow processes in nature and technology and do not exhibit a closed analytical solution, can be investigated with increasing precision. Nowadays, one trillion data points are stored and processed per simulation. It is not clear from the outset which data is relevant for understanding the physical processes. Therefore, hundreds to thousands of simulations are carried out in the course of a research project in order to investigate the influence of individual parameters. Similarly molecular simulations underwent a huge increase in algorithmic and technical performance making it now simple to generate large amounts of data. The pure amount of data can be reduced by suitable data compression algorithms. However, it remains an important and challenging task to manage these simulations in a structured way to ensure reproducibility, retrievability and clarity. Equally important is the subsequent question of how the methodology, process and results of the research project can be secured in the long term and shared if necessary. So far, publication of the data is not anchored in the professional culture and is not easy to achieve due to the technical circumstances. Archiving data for more than 3-4 years seems neither possible nor sensible. Based on this initial situation, members of the IAG and ITT, in cooperation with the infrastructure facilities (UB, TIK, HLRS) of the University of Stuttgart, develop and test a working process in which, immediately after data generation, metadata is automatically extracted from log and input files, supplemented by rarely changing information on authors and projects, and stored in the DaRUS data repository. The repository, based on the open-source software *Dataverse*, is used for the local administration of the data. The easy searchability of the descriptive data helps to reuse the existing data and to document the research process. Last but not least, the data already described can be published or meaningfully archived without much additional effort. For the future, we are also striving for clear criteria for the selection, quality control and retention period of the data and mechanisms for the automated linking of datasets.

Björn Selent et al.: Management of Research Data in Computational Fluid Dynamics and Thermodynamics.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 128–139. DOI: https://doi.org/10.11588/heibooks.598.c8422

1. Initial Situation and Requirements

1.1. Fluid Mechanics

Fluid mechanics is a branch of mechanics that considers the motion of liquids and gases and the forces associated with them. Besides analytical and experimental methods, numerical fluid mechanics (CFD) is a means of capturing and analyzing fluid mechanical processes. CFD facilitates the formulation and calculation of the underlying equations of conservation for mass, momentum and energy even if there is no closed solution available. In comparison to experiments, complex geometries and environmental conditions can also be varied more easily and economically. Numerical methods have been used in fluid mechanics for about 90 years (e.g. [3]), albeit they saw their breakthrough with the arrival of digital computers. Since 1990 the fastest supercomputer has seen a performance boost of six orders of magnitude as can be seen in Figure 1(a). This massive increase in computational performance has led to the ability to use numerical methods based on more evolved equations which model the physical processes in much more detail (cf. Fig.1(b)). The gain in additional accuracy though, led to an increase in both the complexity of the numerical set-up and the associated computational effort. This is exemplified in Figure 1(a) by plotting the storage demand of recent CFD simulations. The amount of data produced by recent computations lies in the range of several trillion data points ($\mathcal{O}(10)$ TB). The extended parameter space furthermore demands more simulations to merely test the computational set-up. And finally when the production set-up has been established the number of parameters to be examined is also a lot larger than in previous decades. This makes a proper management of the computational campaign and the associated data an ever increasing and urgent demand. Data management in this context is not limited to the eventually published scientific article but to all data necessary to produce these published results along the complete research cycle.

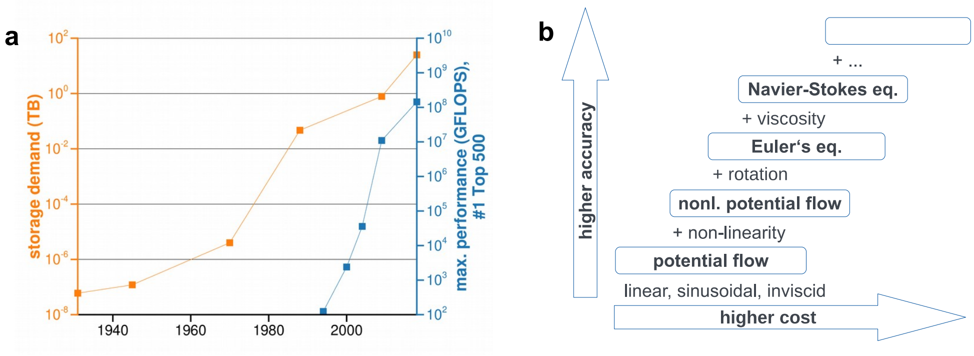

Figure 1.: (a) Storage demand and performance increase of fastest supercomputers.
(b) Complexity over cost of equations governing fluid flows.

The remainder of the present text describes the development of tools necessary to support researchers in the handling of research data and processes throughout the full lifespan of their respective project and beyond.

In order to develop a suite of tools for research data management (RDM) five fundamental stages of a research cycle stage have been defined. These stages albeit deduced from a particular project in the field of CFD may, due to their general character, more or less be found in any research area. Each of the stages has specific demands on the RDM which are partially induced by mutual dependencies as shown in Figure 2.

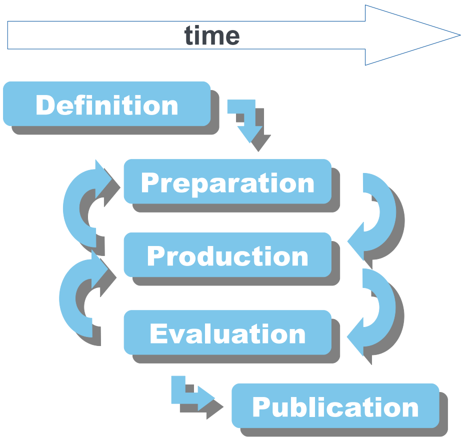

Figure 2.: Stages of research cycle.

The *Definition* stage consists of determining the project's goals, matching them with the state-of-the-art knowledge in its scientific field, collecting information of existing tools and methods to accomplish these goals and making a timetable of the project's progress. These tasks are usually data-poor, i.e. the recordings and collections consist of a manually manageable amount of documents. The data is typically generated on a personal workstation and transferred to any type of repository system. The RDM has to take care of the data being findable and saved from loss, at least for the project's lifespan.

The *Preparation* stage is devoted to preparing the numerical set-up. This includes so-called mesh studies to identify the numerical parameters associated with the computation and a thorough testing of any additional numerical methods involved with the models. Assuming a 3-dimensional computational domain, a set of four different boundary conditions and three additional parameters, a total number of $N = 3^3 \cdot 4^3 \cdot 3 = 5184$ simulations can easily occur. The simulations are usually run on some kind of High Performance Computing (HPC) system. At this stage it is vitally important that the RDM processes are automated and transparent for the researcher. The important metrics of the simulations have to be captured and transferred into a searchable format on a long lasting repository. The main focus of RDM thus lies in the preservation and documentation of the processes leading to a functional environment for succeeding numerical simulations and in making this information retrievable at later times. Albeit simulations at later stages are done for settings determined during the preparation, what might occur is that some limited additional testing is necessary. The RDM tools should therefore allow for dynamic changes at any time.

During the *Production* stage the number of simulations is usually one order smaller than during preparation. But now the results, i.e. actual output data, are valuable as well, at least in parts. The processes during the production stage involve further tools to extract and summarize the results from large datasets. This includes programs to compute derived data and procedures to visualize the results in graphs and pictures. Thus any suitable RDM should allow extending the simulations' metrics with post-processing steps and the generated graphics and data-tables. The RDM has to ensure a strong link between the set-up and the result metrics in order to enable any repetition at later times.

The *Evaluation* stage is strongly interwoven with the *Production* stage, the main difference being that the results are not only observed but also checked for plausibility and that results from several runs are compared with each other and with existing literature. This almost automatically leads to a further requirement for the RDM at this point, namely compound results depicted in a graph should be assigned to the correct simulation's metrics. Finally it is worth mentioning that the integrity of the data has to be ensured as well in order to preserve good scientific practice. *Preparation*, *Production* and *Evaluation* are data-rich stages which lead to the strong demand that any RDM measures should alleviate the work flow and must not add any additional load. Otherwise the acceptance of any RDM is strongly undermined.

The final stage of the cycle consists of the *Publication* of the conclusions and discoveries made. Here the typical requirements for published data are to be applied, i.e. it has to be findable, accessible, inter-operable and re-usable (FAIR) [8] for the public, verifiable and long-lasting. The repository system therefore has to allow access from the outside without endangering any non-published data on the same resources form being compromised. Nevertheless the results shown in any publication should readily be available in both graphical and tabulated form in order to allow third parties to compare them with their own results. In ideal circumstances, the whole production tool-chain should be made available to third parties, including the programs to compute the data along with all parameters and post-processing routines. To achieve this, it is advisable to resort to open and established standard formats and open source programs whenever possible. There are first attempts to share and disseminate large datasets in the community [2, 6], but there is a lack of standards to describe and exchange these datasets in a structured way.

1.2. Thermodynamics

Thermodynamics provides the scientific fundamentals for energy and material conversion processes which occur in almost all areas of modern societies. This makes thermodynamics a key discipline for pressing technological issues in our society such as the rising demand on energy supply and storage, the development of new materials and the optimization of chemical and biotechnological processes. The ITT conducts research in the fields of molecular thermodynamics, molecular simulation as well as simultaneous process and solvent design. In recent years molecular simulation in particular has matured into a powerful tool for predicting microscopic processes and material properties with a high degree of versatility. However this versatility comes at a price in that it is increasingly difficult for researchers to find their way through the maze of available computational techniques and models.

In principle, molecular models and algorithms can be separated into disjunct groups, but in practice a more or less strong coupling exists, making the connection between the molecular model and the predicted substance property sometimes ambiguous. This may also depend on algorithmic details that are not, or incompletely outlined [1, 4, 7]. As a consequence the reproducibility of literature data, only based on the textual information, is often hardly possible.

In the field of thermodynamic property measurement, the Thermodynamics Research Center (TRC) at the National Institute of Standards and Technology (NIST) established a cooperation with various journals more than 10 years ago in order to develop standards for the storage and exchange of thermophysical property data (`https://trc.nist.gov/ThermoM-L.html`). Numerous journals oblige authors to adhere to these standards. In contrast, molecular simulation standards for the exchange and communication of simulation conditions and results have yet to be developed.

Similar to the fluid mechanics workflow towards a publication, different stages can be defined for a molecular simulation study. The definition stage is comprised of determining the systems that need to be examined in order to explore a specific property, and how many variations of e.g. solvent composition, temperature or pressure are required. The system will then be prepared in the next stage, by generating an initial structure containing all atom positions, and files containing all the systems parameters. These two stages are critical and play a major role, since the actual production takes a long simulation time to create physically and statistically adequate data. Thus deep understanding and initial test cases are required before moving to the production stage.

In short, a molecular dynamics simulation numerically integrates Newton's equation of motion for all atoms. Due to physical and numerical constraints, an integration time step is in the range of femto seconds (10^{-15} s). Depending on the scientific question, simulation times up to micro seconds (10^{-6} s) might be needed, creating a computational time up to a month or longer in some cases. Once a physically converged system is achieved, dynamic and static system properties, like diffusion coefficient and density distribution, can be calculated in the evaluation stage. Depending on the results, further simulations with different parameters might be needed. In the final stage, the computed properties will be published. Subsequent to acceptance however, the produced data will be archived or discarded most of the time and also rarely made publicly available due to a data size of hundreds of gigabytes. Therefore a database for the produced data will be needed to incite more clarity into the simulation model and workflow. The approach of a data repository for input and output files of a simulation might not be a replacement for a structured simulation workflow, but it surely is an approach leading in the right direction.

2. Realization

The goal of the implementation was to support the two central requirements—to provide an overview over a large amount of data with an expanded parameter space and to prepare and link data from the different stages (preparation, production, evaluation) for a FAIR way [8] to archive and publish.

The basis for both requirements is a structured description, capturing not only all relevant search criteria to make the data findable but also a representation of the research process and the observed system.

To create and handle this metadata description together with the actual data, a toolchain that supports the researcher with automation, a safe location for the data and an intuitive interface for interaction with the data is necessary.

2.1 Description of the Data - EngMeta as a Metadata Scheme for Engineering Data

The first step of building a metadata description model was the identification of relevant categories: For search and overview, the most important criteria besides general descriptive information like author, year, and project turned out to be quite discipline-specific information about the observed system and their components (like force fields in thermodynamics, the observed and controlled variables), parameters of the used methods (like filtering in aerodynamics or the simulation metrics) and parameters of the observation/simulation itself (like the grid). To make the data understandable and reproducible information about the computing environment (for computational engineering) and used instruments (for experimental engineering), the used software and tools are also necessary.

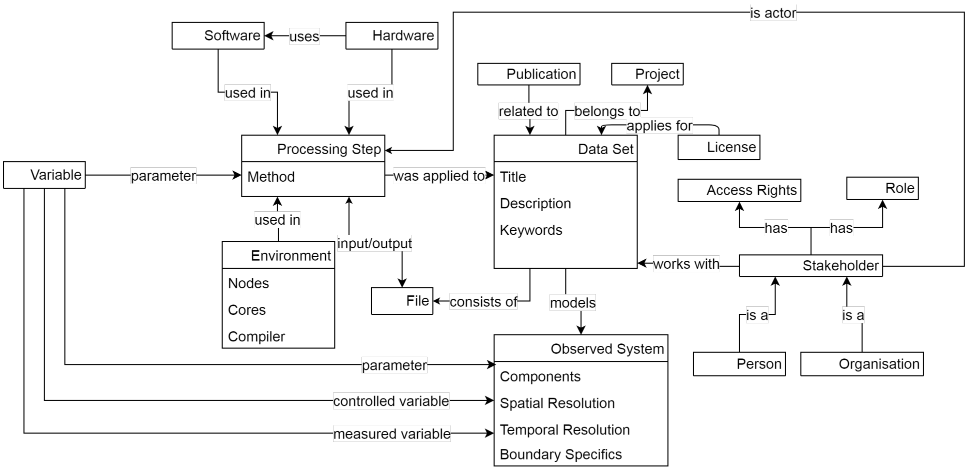

Figure 3.: Object model to describe engineering data.

The object model in Figure 3 visualizes the objects of the research process whose properties are important for the description of the data. Central are not only the characteristics of the data itself but also the steps taken to generate this data. Important for the local management of hot data was also the possibility to mark and comment on negative results. Starting from this object model we looked for existing metadata schemata covering these properties. As we did not find a scheme that addresses all criteria, we created an application profile building on existing standards like DataCite (for general descriptive information), PREMIS (for technical information), CodeMeta (for describing software), ExptML (for information about experimental instruments) and PROV (for a description of the process) and added additional description fields for the discipline-specific information. Figure 4 shows the main metadata categories and fields that later became EngMeta, a metadata model for computational engineering applications[5]. EngMeta is serialized in an XSD-scheme and publicly available.[1] For the data from the *Definition* stage, the general descriptive metadata fields of EngMeta are sufficient with the possibility to link the data to the project (context → project). For the *Preparation* stage, the capture of the simulation parameters (provenance → method/parameters, observed system → bounding

[1] `https://www.ub.uni-stuttgart.de/engmeta`, Last accessed 26/04/2019

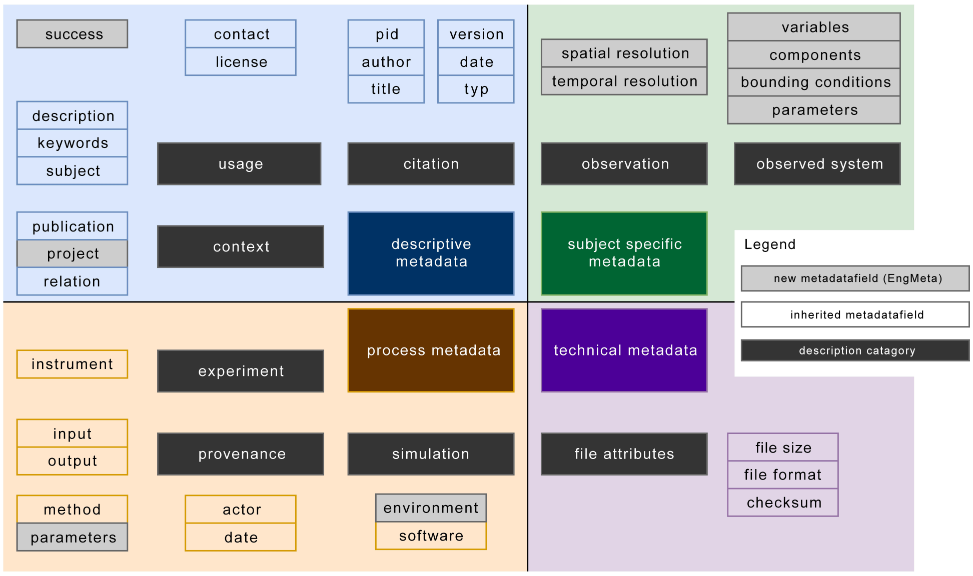

Figure 4.: Metadata categories of EngMeta.

conditions/components) and the documentation of successful and failed configurations (usage → success) is more crucial than storing the actual data. To enable reproduction and understandability, the data of the *Production* stage also has to be connected with a detailed description of the tools and environment (simulation → software/environment), methods (provenance → method/parameters) and the definition of input and output data (provenance → input/output). For the *Evaluation* stage, EngMeta provides different relation types (context → relation) to assign data from different runs and stages to each other.

2.2. Extracting Structured Metadata Using Automated Metadata Extraction

A pure metadata model leaves the annotation of metadata to the researchers. Since the effort for tagging the data in such a expanded parameter space is extraordinary high, we developed an automated metadata extraction as a second step. Lots of meta information already available through the simulation outputs can be extracted without manual intervention. This especially holds true for process metadata, such as information on the computational environment, as well as for domain-specific metadata, such as information about the controlled variables. This information is mostly available through job-, log- and output-files of the simulation run. Technical metadata is also easy to collect, since it is available through file system attributes, whereas descriptive metadata is all but impossible to collect, since it describes the data from a higher point of view. With this background in mind, we developed an automated approach as a light-weight Java tool. The metadata extraction is generic in two ways: First, the metadata information to extract from the simulation codes is configurable via a specific file. This configuration file defines the following for each metadata key: In which file can it be found? What is the search key required to find it? What is the delimiter to separate the key from the value? We chose this approach of configuring the extraction outside the code because only in this way we could overcome one's inhibitions. This configuration file only has to be written once for a simulation code or at least only once for many runs to be accomplished. User-defined descriptive metadata can be put into extra files which can be parsed along with all the output files from the simulation runs. Second, the format of the output files, which includes all the extracted metadata values, can be changed. This was originally implemented for EngMeta, returning an XML document according to the EngMeta specification. Changing the format can only be accomplished by writing output classes in Java. However, the extractor tool additionally outputs the plain information as a (key, value) listing, so one can also convert from this plain text information to any user-defined format.

The tool itself was implemented with the Java Scanner API[2], which is feasible for log files in the size of < 20 MB (the usual log files size of simulation output is only some MegaBytes). However, we developed a parallel solution using the Apache Spark Data Analytics Framework[3], being capable of analyzing multiple 100MB of textual log-files in a very short period of time. The tool can be run directly after the actual simulation, writing a sub-directory *.metadata*. This sub-directory now contains two files: *metadata.txt* contains the plain listing whereas *engMeta.xml* contains the same information according to the EngMeta model specification as an XML document. Since the tool is written in Java, it can be run on several operating systems and architectures, ranging from workstations (including Linux and Windows machines) to clusters (tested on bwUniCluster[4]) and high-performance computing systems (tested on Cray XC40 „Hazel Hen“[5]).

2.3. Overview over the Data - Using Dataverse as a Data Repository for Metadata Management

Without a repository to index, store and manage the metadata as well as a link to the data itself, both the scheme and the extracted metadata are useless for the researchers. So, the third step was to identify a system, that is able to handle EngMeta as a metadata scheme, has a detailed role and rights management and offers an intuitive interface for data management and search. After an overview of existing software systems for (data) repositories, three systems were shortlisted for a detailed comparision: Dataverse[6], DSpace[7] and Invenio[8]. While none of the systems met all the requirements, we chose Dataverse as our data platform primarily for the following reasons: The ability to define custom metadata per collection, programmable APIs to enable automation, an intuitive web interface, an active developer community, and optimization for data instead of text publications.

[2] `https://docs.oracle.com/javase/8/docs/api/java/util/Scanner.html`, Last accessed 26/04/2019

[3] `https://spark.apache.org/`, Last accessed 26/04/2019

[4] `https://www.scc.kit.edu/dienste/bwUniCluster.php`, Last accessed 26/04/2019

[5] `https://www.hlrs.de/systems/cray-xc40-hazel-hen/`, Last accessed 26/04/2019

[6] `www.dataverse.org`, Last accessed 03/05/2019

[7] `https://duraspace.org/dspace/`, Last accessed 03/05/2019

[8] `https://invenio-software.org/`, Last accessed 03/05/2019

Dataverse is an Open Source software platform for data repositories, developed at Harvard IQSS, supported by an active international community of contributors. It is the software base of the data repository of Harvard University and another 41 installations existing around the world. Dataverse offers a user interface to upload, search, manage, share and publish datasets and different APIs for programmable access to the functionalities of the platform. A differentiated rights and role management gives the possibility to share datasets with the desired community—be it within the own working group, with international project partners or the whole research community. Each dataset is described with structured metadata and is managed within a so called dataverse, a collection of datasets with its own set of metadata categories and rights. So, this system allows to create discipline-specific data spaces inside an institutional repository. Dataverse comes out of the box with a set of metadata categories: general descriptive metadata and some discipline-specific additions for the social and geospatial sciences and astronomy. Basing on this software, DaRUS – the data repository of the University of Stuttgart – offers all its members and partners the possibility to manage, share and publish their datasets. We added the procedural and discipline-specific parts of EngMeta to allow the differentiated description of datasets from our two application areas. Normally, Dataverse is meant for the publication of datasets. With DaRUS, we are using the system mainly for the management of the research data within a research group or project, to give an overview over the produced datasets and to help keep the datasets findable and understandable for third parties.

For the data itself, object storage (Netapp StorageGRID) was introduced at our university. Unlike other storage solutions, such as file systems, data on an object storage is organized as objects with an unique identifier, some metadata information and the data itself. Object storage has its strengths in write-once-read-many scenarios, typical for research data, and can easily be scaled for huge amounts of objects. Further, it can simultaneously handle different physical storage, like discs and tapes, enabling cost-efficiency for hot and cold data. Eventually, the identifier URL does nor change, even if the object is moved to a different physical location, which makes it easy to assign DOIs to objects. To avoid data loss, the objects are replicated between two independent data centers. Dataverse supports objects storage via the S3-interface.

2.4. Automation of Data Ingest

After extracting all available metadata from the simulation files, we had to map EngMeta as a deeply nested XML schema to the flat key-value structure of Dataverse's metadata in order to automate the creation of datasets in DaRUS. For this purpose a Python API was developed. It accesses DaRUS through a provided REST-protocol. The API takes as input the generated metadata file from the extraction and other user-specific entries, like the desired database and repository name, and automatically uploads the specified files into DaRUS.

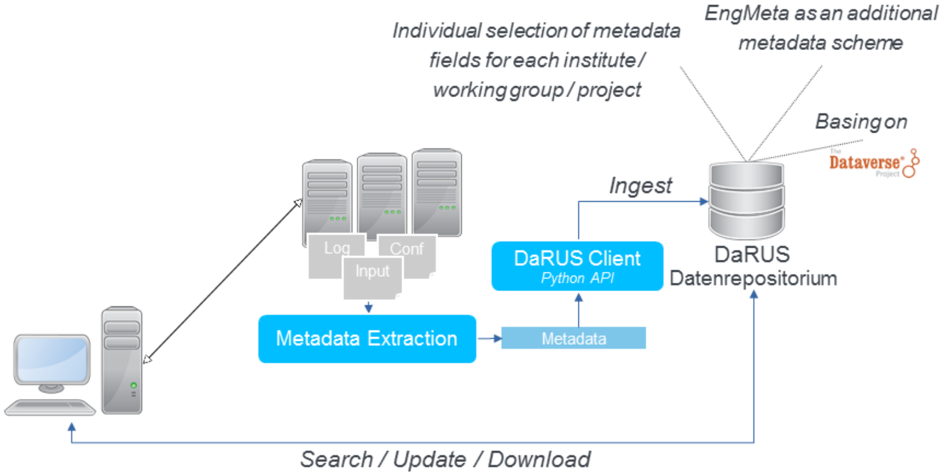
Individual selection of metadata fields for each institute / working group / project
EngMeta as an additional metadata scheme
Basing on
The Dataverse Project
Ingest
Log
Conf
Input
DaRUS Client
Python API
DaRUS
Datenrepositorium
Metadata Extraction
Metadata
Search / Update / Download

Figure 5.: Interaction of Repository, Metadata Extraction and Automation

3. Experiences

The whole process illustrated in Figure 5 is evaluated since the end of 2018 while part of the components were still in development. The major findings will be discussed briefly.

First and foremost we saw that the cooperation between researchers and infrastructural services has been very productive. Especially the metadata model could not have been developed without this combined effort. The researchers contributed deep insight to their research process and the domain-specific view while the infrastructural services contributed their knowledge on metadata standards, repositories and system integration. With the combination of the individual components we have built a fully functional toolchain for extracting, storing and managing data and processes. Even though there are still open questions we think of this as Proof of Concept for a method to assist scientists at all stages of their work in a transparent fashion without increasing their workload.

One of the major experiences was the impact of automation. We learned that automation is crucial if research data management efforts should be accepted. Without automation researchers add only the basic metadata and describe the data in a non-structured way, i.e. by specifying parameters in a general free-text description field. This approach is not sustainable in terms of FAIR data management, so automation was one of our primary goals.

This led to the development of the automated metadata extraction. One of the findings was that initially, the development of the configuration files for the automated metadata extraction process takes some additional time. One has to choose reasonable search terms and know the output files of the simulation codes well. After this initial extra effort, one can profit from the automated approach. We were able to attach the automated extraction directly to the job scripts of the simulations, so after each simulation's run, the extraction could be performed automatically.

Management of data requires a different view of the data than publishing: Dataset listings in Dataverse focus on the citation of data, but for an overview of ones own data there has to be a sortable and filterable view focused on the parameters of the data.

Dataverse offers a full text search and search facets that are good to filter on discrete metadata categories. For continuous parameters, we need different tools to filter datasets like range queries or a sortable tabular overview over the datasets.

3.1. Future Work

The amounts of data usually generated in fluid mechanics and thermodynamics are expensive to store and time-consuming to move. Dataverse has its root in the social sciences, and is therefore not designed for file ingest in the gigabyte and terabyte range. Files of this magnitude produce time-outs during the ingest and download. To enable uploads of data larger than 2 GB, the time outs for API calls are prolonged. But still, this procedure does not scale sufficiently for datasets larger than 100 GB. Therefore, we work on different ways to handle, save and manage these files. One starting point is to leave the data where it is and to connect Dataverse to several different data backends, be it through links in the metadata or through connecting Dataverse with different storage technologies in distinct locations. Another approach is the development of clear criteria to decide which part of the data shall be stored for what period of time. The data from the *Preparation* stage can be deleted quite soon, while published data has to be securely saved for a long period. In order to use the available storage space efficiently, we will need measures for the usefulness of data and automation not only for the ingest, but also for the deletion of data (records) no longer used.

Even if the data is read in automatically, the challenge remains to link data from different stages in the research process in a meaningful and automated way. Data from different simulation runs are linked, compared and integrated during subsequent research steps as described in Section I. A further goal for the future is to automate the mapping of these links into the metadata of the data records concerned and to represent and visualize the research process in a comprehensible and reproducible way.

4. Conclusions

To manage and handle large amounts of data from computational fluid dynamics and thermodynamics we used a data repository based on the Dataverse platform together with an automated description and upload of the data. We use Dataverse not only in its actual function – publishing – but in addition for managing and sharing data within a defined public. There are still steps to take in making this system a feasible management of this data, especially in the linking of datasets and the handling of large data files.

Due to their configurability, the approaches developed can be easily transferred to other fields. This applies in particular to disciplines that also use (simulation) codes to obtain their results. DaRUS is already used as an institutional repository. The content is adapted to the different subject areas through the individual configuration of the metadata categories. The automated metadata extraction can principally work on all text-based log, input or user-generated files that contain metadata information in a semi-structured form. It is currently being tested by working groups from various disciplines.

Thanks

The DIPL-ING project is funded by the Federal Ministry of Education and Research under grant number 16FDM008.

Bibliography

[1] Loeffler, Hannes H., et al. "Reproducibility of Free Energy Calculations across Different Molecular Simulation Software Packages." Journal of Chemical Theory and Computation 14.11 (2018): 5567-5582.

[2] Meneveau, Charles and Marusic, Ivan "Turbulence in the Era of Big Data: Recent Experiences with Sharing Large Datasets." In A. Pollard, L. Castillo, L. Danaila and M. Glauser (eds.), Whither Turbulence and Big Data in the 21st Century?. Springer, (2017): 497-507

[3] Rosenhead, L. "The Formation of Vortices from a Surface of Discontinuity." Proc. Roy. Soc. London A, 134.823 (1931): 170-192.

[4] Schappals, Michael, et al. "Round Robin Study: Molecular Simulation of Thermodynamic Properties from Models with Internal Degrees of Freedom." Journal of Chemical Theory and Computation 13.9 (2017): 4270-4280.

[5] Schembera, Björn, and Iglezakis, Dorothea. "The Genesis of EngMeta-A Metadata Model for Research Data in Computational Engineering." Research Conference on Metadata and Semantics Research. Springer, Cham, (2018).

[6] Sillero, Juan A. and Jiminéz, Javier "Public Dissemination of Raw Turbulence Data." In A. Pollard, L. Castillo, L. Danaila and M. Glauser (eds.), Whither Turbulence and Big Data in the 21st Century?. Springer, (2017): 509-515.

[7] van Gunsteren, Wilfred F., et al. "Validation of Molecular Simulation: An Overview of Issues." Angewandte Chemie International Edition 57.4 (2018): 884-902.

[8] Wilkinson, Mark D., et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." Scientific data 3 (2016).