Research Data Management in Computer Science - the NFDIxCS Approach

Michael Goedicke[1], Ulrike Lucke[2]

Abstract: This contribution discusses the challenges and architectural considerations for research Data management in computer science and related infrastructure for implementing the so-called FAIR principles. The main challenge is, to cover the research data management requirements of the various sub disciplines of computer science. This diversity must be managed in a uniform way which entails a common structure for this task. We outline these requirements briefly and discuss then the concept of the so-called research data management container (RDMC) which encapsulates a given research data set in conjunction with all accompanying information and support (software, execution environment etc) in order to provide a portable unit for management, distribution and access control.

Keywords: Research Data Management, FAIR Principles, Nationale Forschungsdaten Infrastruktur

1 Introduction

Recently, the discussion of science and scientific processes in general have gained important public attention due to the lack of transparency in certain cases which have been made public. One general factor has been identified in these cases that for a better and transparent scientific process the used data be it measured or generated by e.g. extensive simulation was in many cases not available for scrutiny or even turned out be manipulated. In addition, the desire was apparent to reuse this – in many cases – expensively obtained research data. The actual needs in these areas created baroque effects especially in attributing research data in publications – like including the data owners as authors of papers who had no other part in the actual paper at hand.

These are general observations. In computer science (CS) specific observations regarding research data can be made. If a piece of research in CS generates, observes or uses a data set to support the published results it is was in many cases quite easy to "publish" them: create a specific website for the purpose and put it there. In addition, further artifacts (e.g. software and scripts) and context information (e.g. execution environment) can be included there as well. However, this ad-hoc measures suffer from a range of problematic aspects: the site is often not maintained thus the data and artifacts are not useable after a while and the entire piece of information is not findable in a systematic way due to the

[1] paluno/University of Duisburg-Essen, Faculty of Economics and Business Administration, Gerlingstraße 16, Essen, 45117, michae.goedicke@paluno.uni-due.de,
[2] Complex Multimedia Architectures/University of Potsdam, Department of Computer Science, An der Bahn 2, Potsdam, 14476, ulrike.lucke@uni-potsdam.de

doi:10.18420/inf2022_112

lack of standardized meta-data.

The approach to realize more transparency at a general scientific level created the so-called FAIR[3] (Findable, Accessible, Interoperable, Reusable) principles for research data management. These must be translated into actual activities for each specific scientific discipline. In CS a consortium has been created to define such activities and measure and funding is requested in a proposal (NFDIxCS) as part of the NFDI[4] – initiative. This national initiative has been started to create a research data management (RDM) infrastructure to support scientific communities in a discipline specific way. Thus, we are presenting here the CS-way to build such an infrastructure as has been put forward in a related funding proposal (see also the related website[5] where more detail regarding the consortium can be found).

The overall approach is based on the CS scientific community which is included by reaching out using the existing infrastructure of scientific societies. Here the GI[6] plays a key role. While this is important to build a common understanding of the types of research data and how the much-needed transparency of scientific processes is improved for CS we will concentrate on the infrastructure in this contribution. The accompanying measures to reach out to the CS-community will be also supported by the infrastructure, of course.

Thus, this contribution is an account of the key elements of the proposed infrastructure. We briefly describe how this relates to the FAIR-principles and how basic requirements of our scientific discipline are mapped onto related RDM concepts. Finally, the architecture of the infrastructure is sketched where the so-called Research Data Management Container plays a major role.

2 Transparency in Science: The FAIR Principles for RDM in CS

Based on an intensive discussion in the structures of the GI an understanding the FAIR principles and their application to CS was created. We discuss this implementation of these principles in NFDIxCS briefly below. This will include also measures to quality assurance for research data and services to involve the community continuously which we will not address here. This altogether will enable the cultural change towards FAIR research in CS.

The CS discipline is nowadays very broad since CS also reaches out to quite big range of applications. In this work we address not the applications of CS but the findings and research which relates back to CS from applications e.g. search in biological structures like DNA provides new insights into general searching algorithms. Thus, the entire types of research data become quite diverse and in short covers all kinds of data structures from e.g. unstructured point clouds, semi structured texts like software artifacts or automated

[3] https://force11.org/info/the-fair-data-principles/
[4] https://www.dfg.de/nfdi
[5] https://www.nfdixcs.org
[6] Gesellschaft für Informatik https://www.gi.de

theorem proofs to highly organized relational data in tabular form. We used a categorization of the DFG and the subdivision of CS in the substructures of the GI to create related profiles of research data types in CS. This would fill another paper thus whenever the reader considers her/his own subdiscipline just think of it being covered by the NFDIxCS approach. Below we address this as the *profiles of [CS-] research data*.

2.1 Measures to Implement the FAIR Principles across CS

The FAIR principles are the guide and compass for the work within NFDIxCS. The design of the infrastructure, the services and the processes will always carry the FAIR principles as general requirements. It is clear that these requirements have to be complemented with possible conflicting requirements e.g. in terms of legal, intellectual property rights or privacy requirements, and not all FAIR principles can be realized to the utmost extent. We foresee creating separate views on the metadata and data that satisfy the varying requirements. The use of a specific view is then controlled by the role a user has on the data (e.g. author, reviewer, reader). Also, several rule sets, process models etc. have been set forth to guide RDM in general (see e.g. [RISE] and [DIAMANT]) and support ways to implement the FAIR principles within scientific practice. Here, we address the measures at an abstract level to support and guide the work within NIFDxCS by using these knowledge sources. We walk through the principles to point to the planned measures.

- To be **Findable**: This principle addresses the role and properties of identifiers and metadata which are part of a search infrastructure identifying the data and related context. The general related concepts for CS research data as characterised in the profiles as mentioned above. In order to be able to connect to and interoperate with other national and international partners, specifications need to be standardized between stakeholders beyond NFDIxCS.
- To be **Accessible**: This principle addresses the way research data is actually stored and can be accessed through the metadata and identifiers via a search infrastructure. Related open protocols using open authorization and authentication will be key for this topic. NFDIxCS related work packages will address its realization.
- To be **Interoperable**: Interoperability must be realized at all levels of our research data management architecture (see below). The metadata and related protocols need to be open and shared among the community at large and need to be integrated into the general infrastructure of NFDI and possibly beyond. Thus, the overall architecture, meta data, protocols and common standards address this challenging part of the work. Part of the community building will address the related dissemination.
- To be **Re-Usable**: This principle addresses the way the research data can be reused easily and openly. Plurality of attributes, access license(s), provenance and community standards are the aspects addressed by respective work packages. A special aspect supporting reusability will be addressed in work packages addressing reusable execution environments for the long term and architectures / interfaces. While the

former will provide concepts and implementations of the RDMC concept (see below) the latter work addresses the whole structure where instances of RDMCs can be maintained and provided for storing and accessing the data via the appropriate role-based interfaces.

Also, the composition of the ruling bodies of NFIDxCS as well as the operating model are formed in such a way that the continuous observation of the FAIR principles guides the monitoring, discussion and decision making within NFDIxCS as well. We will design a specific code of conduct for NFDIxCS to explicate our common understanding of the FAIR principles against our disciplinary background. This approach will be based on the existing GI statutes and ethical guidelines, leading to the NFDIxCS bylaws.

2.2 Research Data in Computer Science

The CS community is methodically well-prepared to specify abstract-level descriptions of an entity, here research data and its accompanying information. There are plenty of formal description languages to model, for instance, domain-specific data structures (syntax and semantics), system components (structure and behaviour), environments (input from and output to human and/or technical context) etc.. These vocabularies will be a valuable basis to derive metadata standards by a) categorization of possible dimensions and features of different CS research data and b) relating them to each other within a taxonomy. For these semantic aspects we rely on previous work on the Open Research Knowledge Graph [ORKG]. These quite abstract concepts are more at the level of publications and experiments. However, it can be used to e.g. link papers with the datasets used in the paper. Structures in the ORKG are extensible, which we intend to use for CS-specific approaches to characterize our dedicated types of research data according to the needs of the identified sub-disciplines.

The challenge of metadata management in NFDIxCS is the availability of vocabularies optimally reflecting necessary technical specifications, formats, languages, notations etc. for all categories of data in CS. We compiled a sample of relevant approaches for these categories of data and thus exemplifies the range of items to be covered by metadata vocabularies to be developed for CS.

Table 1 Dimensions for specification of selected CS data (sub)categories

(only samples given in no way a claim for completeness)

Data Category	Data Subcategory	Examples for technical specifications, protocols, formats, languages etc.
Data sets	strongly structured	SQL, RDF, SPARQL, GraphQL, JSON, XML
	semi-structured	CSV, TSV, NetCDF, HDF5, Apache

		Parquet
	unstructured	txt, mp3, mp4, jpg
Software	formal models	UML, BPMN, DMN, ArchiMate, ERM
	scripts	Python, R, MatLab, Visual Basic
	web/microservices	HTTP, REST, JSON, AMQP, MQTT
	complex systems	YAML
Context	operating systems	MS Windows, macOS, iOS, Android, various Unix versions and derivatives
	Programming models & runtimes	C, C++, Fortran, Java, OpenMP, MPI, CUDA
	supporting services	CI, CD, container services, server virtualization
	hardware	Configuration files, specifications, emulators if available/necessary
	Physical location	GPS, Address, location in rack, room no.

It is important to strengthen process-oriented metadata management along the data lifecycle and the re-usability of metadata to reduce the burden of documenting data as well as to assure quality of that metadata. Researchers must be encouraged in their roles as data users and data producers by supporting the documentation of data as early as possible during the preparation of the data collection using custom metadata-tools. An improvement processes will take into account international standards in the field, with a focus on harmonising, internationalising and standardising current developments e.g. in the context of RDM and other related policy-making bodies.

2.3 Summary: The Requirements

Based on the aforementioned profile descriptions of our sub-disciplines, the CS community is dealing with a large variety of data types that is or has to be considered for sharing, archiving and publishing. Besides (1) traditional datasets (usually originating from a specific field of application), our work focuses on (2) software as research data, and the preservation of its (3) environmental conditions (context) as a third category of CS data. Data can be related to each other, e.g. several datasets are linked to the same software for production or processing, or a software requires a certain execution environment. Moreover, data can be associated not only with (4) metadata (as familiar from other disciplines) but also with different forms of (5) documentation (e.g. admin, developer or user manuals

as well as abstract level design artifacts).

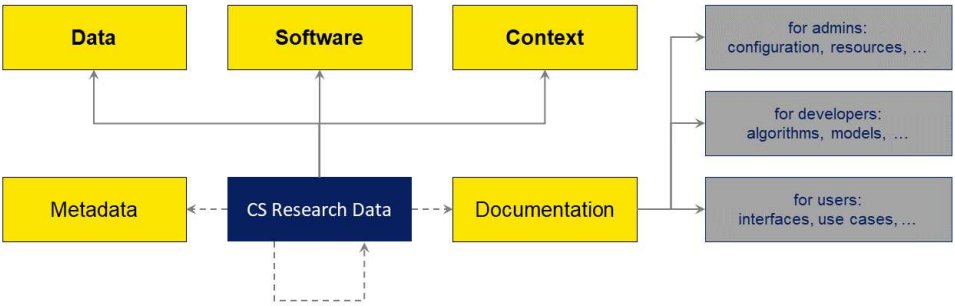

Figure 1 The conceptualization of CS research data in NFDIxCS addresses the characteristics of the discipline in the variety of data types involved.

Independently of the respective sub-discipline, this conceptualization of CS research data (see Fig. 1) can be considered a general model for data handling across CS.

In summary, this leads to the overall architecture of the NFDIxCS infrastructure as presented below derived from the profile descriptions of the sub-disciplines and their declared RDM needs and solutions – for shaping and re-organizing RDM in CS.

3 The Architecture and the RDMC

The major goal of NFDIxCS is to increase sustainability and reusability of research results in CS. Thus, FAIR research data is intended to become an integral element of CS research in general and its publication culture in particular. The core of the efforts is thus to build an infrastructure which addresses the goals of the community as discussed above.

Below we sketch the architecture of the NFIDxCS services as well as the important constituent component: the RDMC for storing and accessing the research data plus accompanying information.

3.1 The Overall Architecture of the NFDIxCS Infrastructure

The infrastructure of NFDIxCS is a stratified set of services. These services are hosted by service providers within the consortium and centrally managed by the NFDIxCS executive management group.

The Access Services (see Fig. 2) provide the direct ability to search, use the search results in form of access to the research data and, in addition, quite a range of (semi-) automated community and management services to NFDIxCS users including central first contact

points for help and support. There is also a machine level interface facilitating the outreach to other organisations within NFDI and beyond, nationally and internationally.

In the middle of the architecture, a set of Core Services provide dedicated functionality to establish the FAIR principles, to support community processes and necessary management operations like ID management and monitoring.

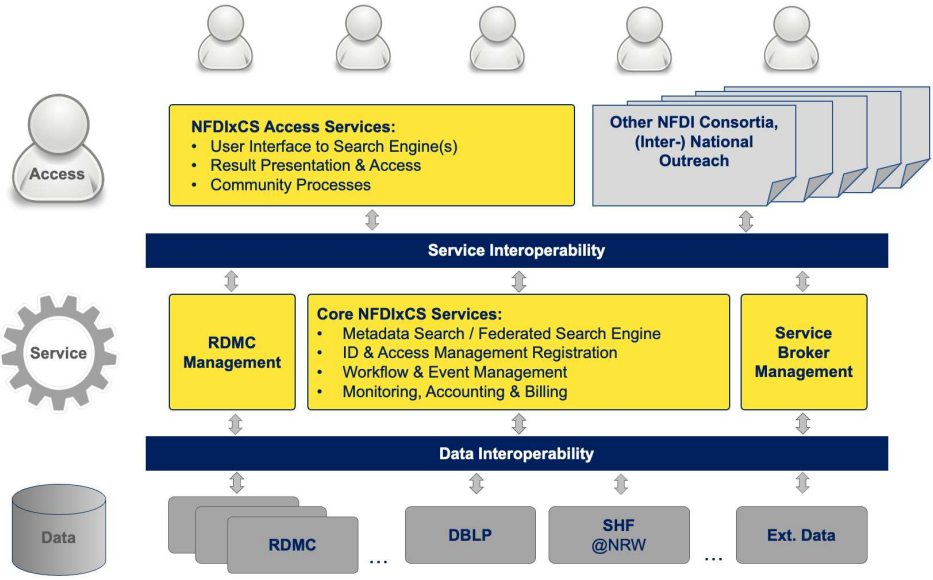

Figure 2 Overview of NFIDxCS Services

At the bottom of the architecture, a set of discipline-specific Data Services are provided to store and access research data and other related databases and repositories. Here, we will include the services of dblp [dblp], facilitating the improved CS related publication processes, and the services of the Software Heritage Foundation [SHF] which are provided as a mirror by the University of Duisburg-Essen to give access to all needed software including all past versions for long term storage (SHF@NRW). Other external repositories and services will be integrated in the NFDIxCS architecture to enable cooperation with other NFDI consortia (using the agreed container interfaces) or basic services like DFN-AAI for identity and access management, for example.

Research data to be included into NFDIxCS will be available in a packaged format: the so-called Research Data Management Container (RDMC), which is a portable object that manages itself in terms of access, specific workflow and all the data, software and further information to describe the data and all the means to access / bring the data back to life. These RDMCs will be hosted by a few service providers initially within the consortium and later, especially after the funding periods, by trusted (probably also paid) service providers which will possibly be available by EOSC / GAIA-X or similar provisioning

schemes. The components encapsulated by an RDMC are sketched in Fig. 3.

We are aware that the idea of encapsulating research data in a container gains much attention in the community (e.g. code ocean, GIDA , Fair Data Objects etc.). We observe a very dynamic situation and currently each of them has a certain profile and is in a different state of maturity. Of course, the RDMC is also the unit for evolution should a specific technology be no longer supported and the entire set of research data and accompanying information needs to be moved to a new set of technologies.

The CS sub-disciplines will provide – managed by the overall compliance and consultancy processes of NFDIxCS – templates for the different research data types in CS. The ability to store the raw and processed data together with (a reference to) the needed software and execution context has been mentioned already ([Str21] and many others). In addition to this, the special NFDIxCS feature are the three components (Access Control, Workflows and Filters & Transformation) on top, which rule the access and the work state of the data, including filters and role specific transformation within the container.

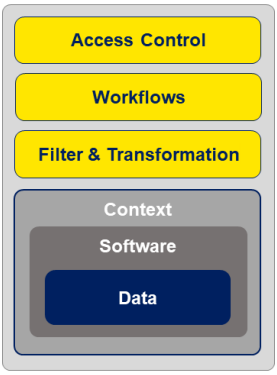

Figure 3 Overall Structure of a Research Data Management Container (RDMC)

Given the envisaged service structure and the RDMC concept, we now go into detail, which aspects of CS RDM we will address in NFDIxCS. Based on such an infrastructure, we support the current research methods and the types of data that typically are generated

3.2 The RDMC

A centrepiece of the storage and management of the actual research data management is the concept and realization of the RDMC. This provides an encapsulation facility to support and execute the rules which will be designed by NFDIxCS to achieve the goal of the project for each specific research data set being eligible for storage and access by the consortium's resources.

It contains everything to encapsulate the research data in question, the software

components (or reference the specific version or variant in a separate repository as, e.g., the Software Heritage repository) and information on the run-time execution environment and further context necessary to support common tasks like the replication of the creation, purging, analysis and visualization of the data.

Essential features of a RDMC include role-based access control to the encapsulated resources. Additionally, there are plug-ins included to provide proper protection of IPR, privacy and security and potential transformation, depending on context, role and purpose of the access in question. If needed, different specific views must be provided to realize possibly divergent requirements. The general structure of a RDMC is sketched in Fig. 4.

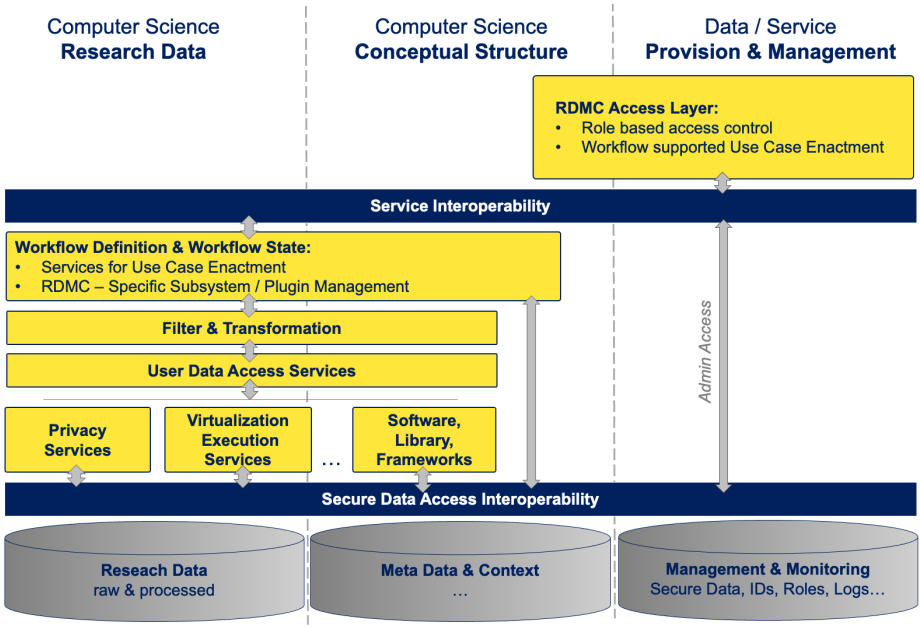

Figure 4 General Structure of Research Data Management Containers (RDMCs) and their management across the levels of the NFDIxCS architecture

For each type of research data identified and described by a CS sub-discipline a type of RDMC is defined and created as a template to collect, store and index a certain dataset with all its belongings as sketched above. Depending on this type, the components as depicted in Fig. 4 might slightly adapt their behaviour to the characteristics of the respective data:

- The RDMC Access Layer restricts the access to the container's resources according to the user's permissions. It uses the same service for access control as the infrastructure.

- The container's workflow defines its possible use cases, e.g. review or reuse processes. The related workflow execution uses the workflow engine from the infrastructure, but the state of the workflow is encapsulated in the container.
- Filter and transformation are plugins which get configured and used based on the use case realized in the container, for instance to implement pseudonymization, anonymization and/or differential privacy. Other transformations, e.g. aggregation of data in case of IPR protection or measures to implement simple privacy protection by eliminating personal data in log files, can be implemented as a plugin here.
- User Data Access Services provides consistent access to the various forms of research data to comply with the given privacy rules defined for the general type of research data and for the specific container in question in addition. Of course, in order to support review in exceptional cases (e.g. a legal inquiry) or finishing a RDMC off at the end of its defined life cycle, admin access is provided as well - secured by additional means like four eyes principle or adequate alternatives.
- Further components help to run the software to access the data using the components privacy services, virtual execution services and additional libraries. This can also include hardware emulation services like Qemu.

Additional considerations are necessary for management of such RDMCs. For instance, it is compulsory for each specific RDMC that the metadata will be made available through the search engines sketched in the overall architecture above. The RDMC Management implements the services to realize the RDMC lifecycle from its inception, registration in the various repositories of the NFDIxCS infrastructure to find it, make it accessible and interoperable and reuse it.

3.3 Sample User Story of the RDMCs

To illustrate the technical considerations above, we provide a sample user story of how the RDMC concept can be used in the realm of the NFDIxCS infrastructure.

In studies on HMI research data is collected in the form of measurements over time. Typically, human participants have to perform a given task for testing the given hypotheses, e.g. on the performance of different interaction techniques. This may be done by capturing data in the system (events like changing a viewpoint, selecting an object or manipulating its characteristics, along with the related timecode) as well as data on the participants (spatial data like movements of the person and his/her hands or head, sensor data like direction of gaze, skin conductance or heart rate). The nature of this data and such studies brings up several aspects to be considered when publishing the data. To name just three:

- Gathered data is closely related to the environment that was used in the study. This includes software aspects, e.g. the used 3D engine or 3D model, the integrated measurement component or the applied analysis tool. Moreover, this in turn includes execution environments, e.g. the used host computer (hardware, operating system,

graphics frameworks and libraries) and virtual reality equipment. For replication of the study to either verify the results or modify the parameters, availability of this software and hardware (at least in an emulated way, as far as this is possible) is a necessary condition. Thus, our RDMC approach encapsulates the data along with these components in order to publish them.

- Involvement of human participants and capture of motion data or physiological data require additional efforts for conservation of privacy. As an example, biometric information could be obtained from typical movement or gaze patterns, making it possible to identify this person. Thus, the data must be anonymized before publication. The required effort needed – pseudonymization, anonymization, additional differential privacy – depends on the actual study and experimental design as well as the scenario of use for this data. Our RDMC approach is equipped with specific plugins and workflow models to implement transformations and filters on the data. Thus, e.g. for quality assurance in the publication process, reviewers might get full access to the raw data. Upon acceptance for publication, data is only accessible in an anonymized way.
- The developed technology or interaction technique might be associated with certain intellectual property rights. Thus, authors might refrain from publishing the whole dataset to the broader public. However, selected parts of the data or an aggregated version might still be worth publishing beyond academic reuse. This depends on the license to be issued for the data and is associated with, again, dedicated transformations of the dataset for publication and reuse.

The figure below depicts these possibilities using the example from HMI sketched above in accordance with the layers of the RDMC as familiar from Fig. 3.

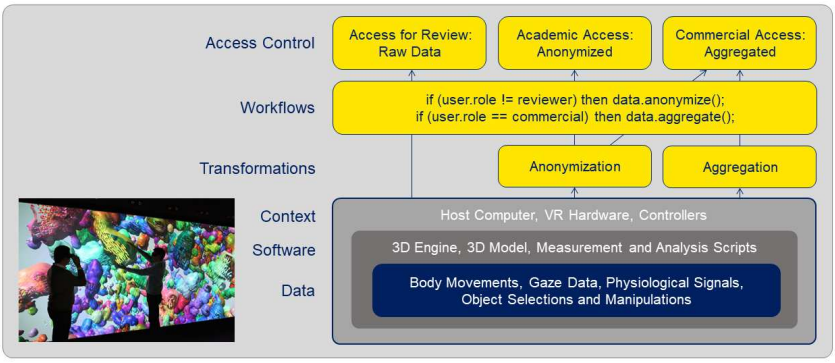

Figure 5 Example of Interaction Data and PlugIns in an RDMC

This way, a trust relation is established between the various stakeholders (authors, reviewers and users of the data) in the research process to ensure that the scientific contribution fulfils the community standards. Authors get their results published in high quality form without harming the privacy of their study participants or their own exploitation interests.

and reviewers can perform a quality check. Users can check and replicate the results to a greater or lesser depth, depending on their role in the academic system – without getting access to the raw data.

Thus, the RDMC is a versatile envelope which provides a great range of configuration possibilities for fulfilling the identified RDM requirements for CS.

4 Summary and the Way ahead

It is obviously quite a work programme for the consortium and the CS community as well. First little experiments have been done to establish proofs of concepts and outreach to national and international societies are developing.

This text is in parts adapted from the proposal for NFDIxCS. Of course, this is a multi-person effort but the responsibility for this text is by the authors.

We are grateful to all who participated and supported us. These are the especially the coapplicants and the participants of NFDIxCS.

We are looking forward to a rewarding journey and to meet you and your RDM requirements.

Bibliography

[DIAMANT] The DIAMANT-Model 2.0 Reference Process https://ubt.opus.hbz-nrw.de/frontdoor/index/index/docId/1432

[dlbp] Computer Science Bibliography https://dblp.org/

[ORKG] Open Research Knowledge Graph https://www.orkg.org/orkg/

[RISE] The RISE-DE Reference Model https://zenodo.org/record/3585556#.X3DX5C-21VQ

[SHF] Software Heritage Foundation https://www.softwareheritage.org/

[Str21] Strickroth, S.; Bußler, D.; Lucke, U., Container-based Dynamic Infrastructure for Education On-Demand. In: Kienle, A., Harrer, A., Haake, J. M. & Lingnau, A. (Hrsg.), DELFI 2021. Bonn: Gesellschaft für Informatik e.V. (p 205-216). https://doi.org/10.1365/s40702-021-00771-7