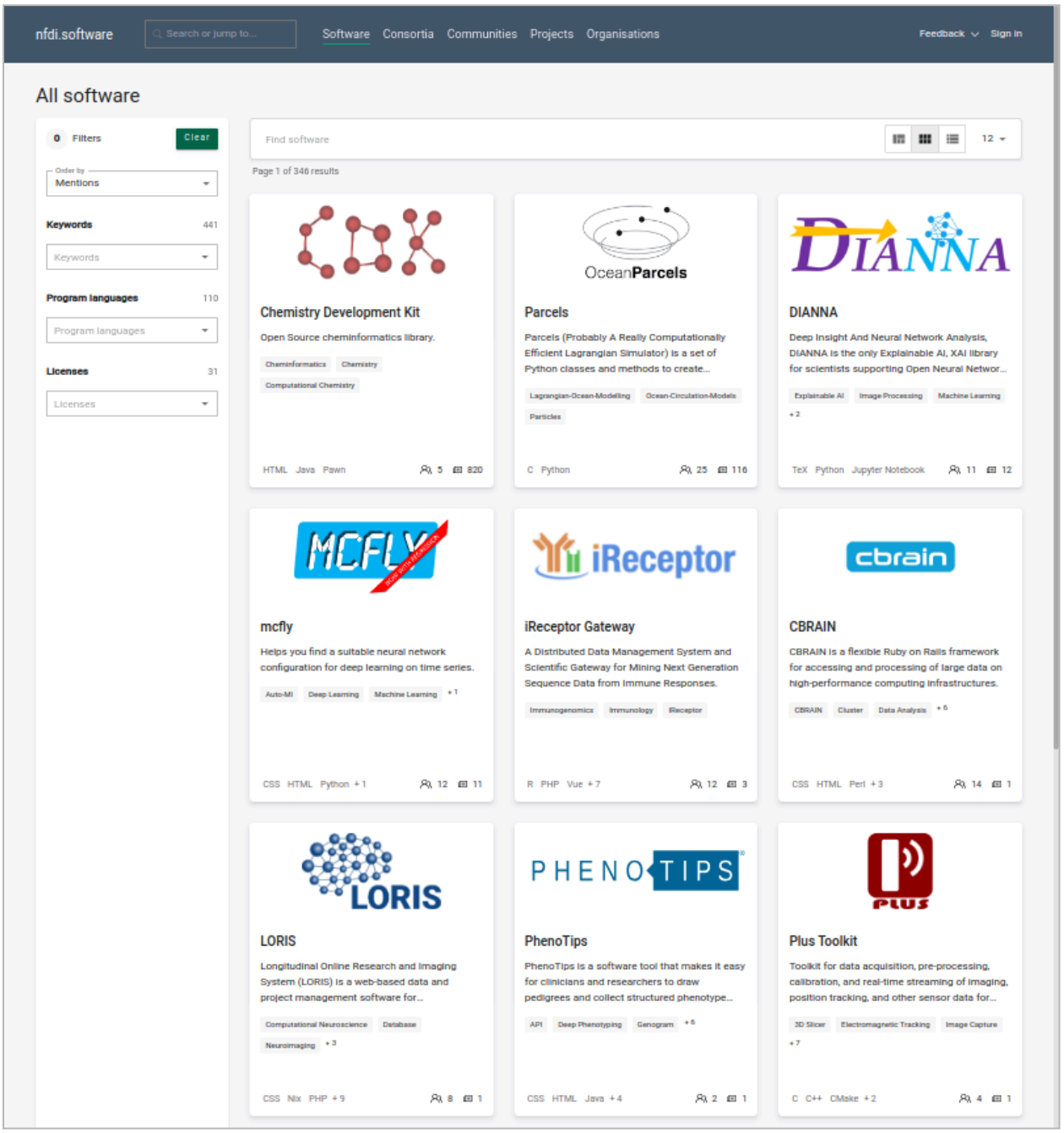
Proposal - Initialisation Phase

nfdi.software

Unlock software, activate research data.



Proposal for the Initialisation Phase of Base4NFDI

Submitted: scheduled for 1st of May 2024

On behalf of: WG Research Software Engineering, Section Common Infrastructures

1. General Information

- Name: NFDI Research Software Marketplace
- Acronym: nfdi.software
- Subtitle: “Unlock software, activate research data.”
- Corresponding section: Common Infrastructures
- Lead institution: Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences, Telegrafenberg, 14473 Potsdam
- Name of lead institution principle investigator: Martin Hammitzsch
- Participation institutions (Table 1):

Principal Investigator	Institution, location	Contact Email	Member in [consortium][1]	Funding requested [yes | no]
Martin Hammitzsch, Beate Hetényi	Helmholtz-Zentrum Potsdam Deutsches GeoForschungsZentrum GFZ (GFZ)	hammitzsch@gfz-potsdam.de hetenyi @gfz-potsdam.de	NFDI4Earth	yes
Björn Grüning	Albert-Ludwigs-Universität Freiburg (ALUF)	gruening@informatik.uni-freiburg.de	DataPLANT, NFDI4BioImaging*	yes
Andreas Rausch	Technische Universität Clausthal (TUC)	stefan.wittek@tu-clausthal.de	NFDI4Ing	yes
Michael Zacharias	Landessternwarte, Universität Heidelberg (UHD)	m.zacharias@lsw.uni-heidelberg.de	PUNCH4NFDI	yes
Aleksander Marčić	Nordrhein-Westfälische Akademie der Wissenschaften und der Künste (AWK NRW)	amarcic2@uni-koeln.de	Text+, NFDI4Culture*	yes
Matthias Löbe	Universität Leipzig (UL)	matthias.loebe@imise.uni-leipzig.de	NFDI4Health	yes
Astrid Nieße	OFFIS e.V. (OFFIS)	astrid.niesse@uni-oldenburg.de, oliver.werth@offis.de	NFDI4Energy	yes
Leyla Jael Castro	ZB MED – Information Centre for Life Sciences (ZB MED)	ljgarcia@zbmed.de	NFDI4DataScience	yes
Felix Engel	TIB – Leibniz-Informationszentrum Technik und Naturwissenschaften	felix.engel@tib.eu	NFDI4Ing	no
Sonja Schimmler	Fraunhofer FOKUS	sonja.schimmler@fokus.fraunhofer.de	NFDI4DataScience, NFDI4Cat*	no

Table 1: List of participating institutions

[1] Name one DFG consortium the institution is or has a route to become a member of and through which funds should be appropriated if this proposal is approved. (* - involved, but without routing of funds)

- Planned duration of the project: 12 months
- Significance: Software is essential for the generation and analysis of research data. The scientific disciplines each have their own requirements, which lead to specific software solutions developed, maintained and operated along the use and analysis of data in the research process. Despite its accuracy of fit, research software has extensive utilisation options beyond the initially targeted data. nfdi.software will serve a single entry point in Germany that brings together information about research software from the different scientific communities. In this sense, it will make it easier to find, understand, evaluate and reuse software for specific research questions and for one's own research.

Summary of the proposal in English:

nfdi.software aims to design, implement and test a central marketplace for research software. The aim is to improve the accessibility of NFDI research data and its access with specific software packages and enable more complex (re)use of metadata and software. The need for a central access portal for research software of the NFDI consortia arises from the growing need of the scientific disciplines of the cultural sciences, humanities and social sciences, engineering and the natural and life sciences to ensure the sustainable use and further development of research software. nfdi.software is intended to link and coordinate independent individual developments from these areas in a federated data infrastructure. The networking and contextualisation of research software opens up the potential to use research data sustainably and to expand the spectrum of analysis and processing of research data. In order to permanently update this, metadata of research software is extracted, standardised and enriched and relevant AI solutions are implemented. The marketplace is therefore a contemporary platform that also reflects the relevance, acceptance and possibilities of the respective software within the research domains through continuous monitoring. nfdi.software provides the foundation for establishing standards and for networking across the NFDI. During the initialisation phase, an approach is developed based on the experience of previous initiatives, leading to a 'Prototype for Integration'.

Summary of the proposal in German:

Mit nfdi.software wird ein zentraler Marktplatz für Forschungssoftware konzipiert, implementiert und getestet mit dem Ziel, verteilte Softwaremetadaten abzubilden und zu verknüpfen. Dadurch sollen die in der NFDI verfügbaren Forschungsdaten und ihre Erschließung durch spezifische Programme unmittelbarer zugänglich gemacht werden, sowie eine komplexere (Nach-)Nutzung der Metadaten und Software ermöglicht werden.

Die Notwendigkeit eines zentralen Einstiegsportals für Forschungssoftware der NFDI Konsortien ergibt sich aus dem wachsenden Bedarf der Wissenschaftsdisziplinen der Kultur-, Geistes- und Sozialwissenschaften, den Ingenieurwissenschaften und den Natur- und den Lebenswissenschaften, eine nachhaltige Nutzung und Weiterentwicklung von Forschungssoftware sicherzustellen. nfdi.software soll aus diesen Bereichen unabhängige Einzelentwicklungen miteinander in einer föderierten Dateninfrastruktur verknüpfen und aufeinander abstimmen. Die Vernetzung und Kontextualisierung von Forschungssoftware eröffnet das Potential, Forschungsdaten nachhaltig zu nutzen und das Spektrum der Analyse und der Bearbeitung von Forschungsdaten zu erweitern. Um dies permanent aktuell zu halten, werden Metadaten zu Forschungssoftware extrahiert, standardisiert und angereichert, sowie relevante KI-Lösungen implementiert. Der Marktplatz repräsentiert dadurch eine aktuelle Plattform, welche zusätzlich durch fortlaufendes Monitoring die Relevanz, Akzeptanz und Möglichkeiten der jeweiligen Software innerhalb der Forschungsdomänen widerspiegelt. Mit nfdi.software wird innerhalb der NFDI die Grundlage geschaffen, Standards zu etablieren und sich zu vernetzen. In der Phase der Initialisierung wird ein gemeinsames Herangehen erarbeitet, welches auf Erfahrungswerten vorangegangener Initiativen aufsetzt und in einen „Prototype for Integration“ mündet.

2. State-of-the-Art of Proposed Basic Service

2.1 Background and Motivation

Given its importance, research software should be findable, accessible, interoperable and reusable (FAIR) [1]. Currently, findability of research software as well as its connection to other research artefacts remains a challenge [2]. NFDI4Ing, NFDI-MatWerk, NFDI4Energy, NFDI4Biodiversity, and NFDI4DataScience aim to address this issue in their respective communities. NFDI4Ing introduces the Archetype Betty, who is mainly concerned with research software, where to find it, how to reuse it and what to benchmark it with. NFDI-MatWerk contains a Task Area named ‘Workflows and Software Development’ (TA-WSD) with a focus on the creation, standardisation, and implementation of scientific workflows and automated protocols for experiments as well as the development of a software framework. They collect and list software related to workflows and software development as well as software services related to material data. NFDI4Energy is establishing research community services to allow reproducibility and transparency of research artefacts, such as a registry to find suitable data sets and software modules as well as existing simulations. TA3 ‘Sustainable Data, Tools and Services’ (NFDI4Biodiversity) captures and documents the current range of software solutions, repositories, and services in biodiversity research. The TA 'Infrastructure & Services' with the

measure 'Unified access portal and interface' is focused on developing the NFDI4DS Gateway and Portal, which references research objects, also including research software from the data science and artificial intelligence community (e.g. from GitLab and GitHub instances).

Although much work regarding research software is planned and conducted within at least 20 consortia, there is currently a lack of a cross-consortia initiative focused on common objectives, bringing together reliable contributions from these consortia with the necessary impetus and producing sustainable results that can be used by the majority of the NFDI consortia. The initialisation phase of nfdi.software establishes the required communication for a concerted effort, brings together relevant players for jointly designing the interplay of the related consortia services, and offers an integrative prototype for first-hand experiences supporting the design phase and for the practical identification of problems and gaps. **As a central service, nfdi.software will harvest, aggregate and harmonise software metadata from NFDI consortia marketplaces and registries, and will provide back enriched and curated metadata.**

2.2 State-of-the-art / Status of work results in preparation for the Basic Service

In its initial phase, nfdi4.software will build on top of existing solutions, thus minimising duplication of work while maximising compatibility to other similar and/or complementary efforts. **Bio.tools** and the Research Software Ecosystem[2] [3] is a portal supported by the European Infrastructure for Biological Information (ELIXIR)[3], where researchers can find and compare bioinformatics tools thanks to curated metadata. The main objective of bio.tools is to build and maintain a comprehensive registry of high-quality software metadata. The **Research Software Directory**[4] aims at finding and reusing cross-domain research software, stimulating reuse, encouraging proper citation, and making the impact of research software more visible. **Betty's Research Engine (BRE)**[5] is a search engine [4] that finds software repositories from GitHub and GitLab and then links them to corresponding scientific publications or entries in third-party databases. Results are enriched with harmonised metadata extracted from the code repository, the publication and the related databases. **Physics.tools**[6] searches in publications for referenced software and automatically retrieves accessible metadata information from the code repository (e.g. GitHub). This also allows for direct access to the number of papers that reference a given product, which is a measure of the product's success. The criteria catalogue

[2] https://github.com/research-software-ecosystem/content
[3] https://elixir-europe.org/
[4] https://research-software-directory.org/
[5] http://nfdi4ing.rz-housing.tu-clausthal.de/
[6] https://physics.tools/

of the TMF ToolPool Health Research for ensuring the high quality of registry entries will be incorporated into the project as conceptual preparatory work [5].

In the **Section (Meta)data, Terminologies and Provenance** we will exchange with WG Research Software Metadata and WG Search and Harvesting (Harvesting and Findability Enhancing Services). BRE is actively collaborating with the **Open Research Knowledge Graph** [6] (ORKG[7]), by utilising the ORKG's REST API to provide the user with a direct interface which can be used to semi-automatically create data entries. With **Section Common Infrastructures** nfdi.software will cooperate with WG Identity and Access Management (IAM), WG Persistent Identifiers (PID), WG Long Term Archiving (LTA), and WG Data Science and AI (DSAI). **Section ELSA** could provide detailed information to add forms collecting licence information, data protections and dual use information, which would integrate recommendations of ELSA practically in the software marketplace. Exchange with **Section Education and Training**, WG Training Infrastructure is also planned.

2.3 Current Technical Readiness Level (TRL) of the proposed Basic Service

For the overall approach of nfdi.software the technology concept has been discussed and formulated (TRL 2) for the initialisation phase and in parts also further options for successive phases. The required components do exist and range from a system prototype (TRL 7) to systems proven in an operational environment (TRL 9).

Description	Software	TRL
Overall service	nfdi.software	2
Component: metadata storage and scraper	bio.tools	9
Component: metadata harvester and LLM engine	physics.tools	7
Component: metadata migration service	Betty's Research Engine	9
Component: user frontend	RSD	9

2.4 Lessons learned and other Use Cases

nfdi.software envisions a central marketplace to map and link distributed software metadata. With this vision, we find ourselves amongst similar, pre existing solutions that we want to learn from and reflect on how we differ, to highlight the additional value that we are aiming for. Papers with Code[8] is a community-led effort to enhance the discoverability and accessibility of research software. It provides references to scholarly articles in areas such as mathematics, physics, machine learning, astronomy, and computer science, along with links to associated data and software repositories. Content creation and quality assurance are carried out by the user

[7] https://orkg.org/
[8] https://paperswithcode.com/

community. Zenodo[9] provides a service where users can share research-related digital items, including software, and assign them a DOI (Digital Object Identifier), thus making them referenceable. For digital works also housed on GitHub, Zenodo enhances visibility with a Zenodo-Badge, a compact image that can be embedded in the README file to prominently display the Zenodo DOI. The citation file format (CFF[10]) is a file format that is human and machine readable. If a citation file is placed in a GitHub Repository it unlocks the "cite this repository" function on GitHub, which further increases visibility and the probability of getting cited. The European Open Science Cloud (EOSC) is envisioned as a 'Web of FAIR Data and Services' for science in Europe, making it easier for researchers to publish, find, and reuse data, tools, and services. It also provides a platform for accessing learning and computing resources, and finding funding opportunities.

Despite these existing solutions that already provide researchers with the technical infrastructure to produce FAIR research software, the needs and use cases of our national, German research landscape are not sufficiently addressed. This becomes apparent when reflecting all the initiatives described in 2.1 and 2.2. This extra need of a *national research data infrastructure* is (as we see it) a chance to advance the state of the art in research data management. However, nfdi.software will not be an island. We will utilise existing solutions, learn from them and integrate. [6] for example, uses Zenodo and CFF to link software repositories to their respective references and the EOSC is our target destination when it comes to aligning ourselves with European efforts. In Papers With Code we see a proof of concept that community driven marketplaces for research software work. We aim to push this concept by focusing on the needs in our national context towards One NFDI.

4. Working Concept for the Development of the Basic Service

The nfdi.software initialisation phase will offer an early proof-of-concept service implementation that is further developed and regularly updated towards a prototype service until the end of the funding. This service will build upon BRE, RSD, physics.tools and bio.tools. Further the integration of the NFDI IAM is also planned. In this way, the service aims to provide a tangible starting point while also enabling user feedback to identify strengths and weaknesses that will guide the future development in a follow-up integration phase. In addition to the early service, nfdi.software will establish bridges to the domain-specific NFDI infrastructures and services, and the corresponding communities with a focus on research software. This community building across the NFDI will safeguard a concerted effort in the NFDI, aims at the joint discussion of

[9] https://zenodo.org/
[10] https://citation-file-format.github.io/

commonly shared requirements and community specific needs, and feeds into the concept and design of a follow-up integration phase with a corresponding proposal.

4.1 Service initialisation concept

The nfdi.software proof-of-concept service consists of a website and an API. The website offers an interface to search for and discover research software available via the NFDI. In the background it contains functionalities for harvesting and looking-up information available in the federated infrastructure of the NFDI. To explore detailed information about software, also associated information will be provided such as related publications and data, involved persons, communities, organisations and projects. Parts of this are already available within the range of functions provided by the RSD. Other existing parts will be integrated such as a bio.tools harvester along with the necessary performance improvements for data storage and data access. Based on BRE and physics.tools, an extended dynamic on-demand search and recommender will be integrated too.

In preparation for an integration phase following the initialisation phase, different metadata vocabularies for research software will be investigated and discussed with the help of the Research Software Metadata WG. This will build on the results of previous projects and activities (like workshops, etc.) and develop into a concept for a common cross-sector metadata vocabulary. Furthermore metadata efforts and services such as Knowledge Graphs and Jupyter4NFDI will be evaluated. Their concrete application and further use will be discussed with the relevant stakeholders in the NFDI. The results will be documented in an initial concept and design report containing an architecture blueprint for potential integration. The concept will be used in a workshop and discussions with broad participation of the consortia and (potential) future basic services. In this way, feedback is received to subsequently improve the architecture blueprint ready for the follow-up integration phase.

4.2 Development and integration outlook

Based on the concept and design developed in the initialisation phase, the integration phase will be proposed in cooperation with the consortia and basic services that are ready and available at the time. Connectivity to the solutions of the consortia as well as to further basic services will be implemented, e.g. with REST-based and metadata-compliant APIs enabling and supporting harvesting and on-demand searches of consortium-specific marketplaces, or search-and-discovery portals for software-related information. This will create an offer to other NFDI marketplaces and portals of the consortia to easily reuse and integrate pieces of the research software metadata conveyed by nfdi.software.

Furthermore, NFDI consortia can reuse technologies and components underlying nfdi.software so that community-specific instances either in parts or as a whole of nfdi.software can be set up and provided with further necessary customisation as appropriate. These instances are consequently linked with nfdi.software to ensure connectivity and exchange of software-related information across the NFDI.

Further information and research artefacts could be associated with research software by linking APIs of marketplaces across the consortia with a knowledge graph infrastructure. Along with this, the integration of further basic services for terminologies, ontologies, their mapping, and their harmonised provision is also conceivable. The construction of knowledge graphs could be further promoted to facilitate mapping, transformation, and representation of the collected information about research software and its linkage with research artefacts in models, knowledge graphs, and standardised metadata schemes.

The outlined scenarios will be clarified and discussed with stakeholders in the initialisation phase to enable testing and implementation in the integration phase. This increases the requirements for the metadata vocabulary, which needs to support semantic technologies. Nevertheless, there are also services whose integration already seems directly possible today, such as the embedding of directly executable and reusable research software in JupyterHubs, which could be found directly in nfdi.software and used from there via the respective JupyterHub.

4.3 Ramping up for Operation

The basis for the ramping-up-for-operation phase is an already well-functioning service that is well integrated into the NFDI. At the moment, it is difficult to fully understand the dependencies and implications created by the interaction with other services that are emerging. Initial thoughts will be shared and discussed in the first year of nfdi.software and will be fleshed out in the second and third year in preparation for an operation phase.

4.3 Risks and Challenges

The integration of the already developed software solutions BRE, bio.tools, physics tools and RSD greatly limits potential risk factors from a technological point of view. This provides a solid basis for the actual task of designing a marketplace for research software. Further preliminary work that has already been carried out through previous activities and events is crucial. Through regular community activities over the last two years, there is also a broad network of stakeholders and interested parties on which to build. On this basis, potential risks such as the degree of dissemination and awareness can be adequately countered. Furthermore, the

evaluation of existing technological developments as well as the assessment of user behaviour and user needs ensures a targeted prioritisation of functionality and user acceptance.

5. Work Programme

In the initialisation phase four work packages (WPs) bundle the activities to realise nfdi.software, design the foundations for the successive integration phase, and help with processes needed to steer the planned activities community-wise and projectwise.

5.1 Overview of work packages

In WP1 the initial provision of nfdi.software and iterative updates are ensured. It covers the implementation and initial integration of complementary components and services so that nfdi.software is available for use with a proof-of-concept first and later on as a continuously updated prototype service for first-hand experiences. The prototype shall pave the way for the successive integration phase. WP2 ensures community-wise interactions and NFDI stakeholder involvement. Thus it aims at nurturing an active exchange towards a concerted effort within the NFDI. WP3 covers the design and concept for implementation in a follow-up project and the writing of the proposal for the integration phase. This WP therefore feeds from the prototype experiences and feedback as well as the stakeholder exchange and expressed community needs. While WP1-3 aim at specific results and deliverables, WP4 takes care of coordination and management tasks as well as the required quality assurance. The following table lists the work packages, their deliverables and overall checkpoints C1-6.

Project month	1	2	3	4	5	6	7	8	9	10	11	12
WP1 Prototype Service		1.1	1.2		1.3	1.4				1.7		1.8
					1.5	1.6						
WP2 Community	2.1		2.2			2.3	2.4			2.5		2.6
WP3 Concept & Design			3.1			3.2				3.4		3.5
WP4 QA & Coordination	C1		C2			C3	C4			C5		C6

Figure 1: Gantt chart with work packages and deliverables.

5.2 Detailed work programme

The work programme planned is depicted in Fig. 1 with a Gantt chart showing the activities in each of the interlinked work packages and their alignment using six checkpoints. The initialisation phase aims at starting in November 2024 and will last for twelve months. The activities within the work packages are shortly summarised with their deliverables. The overall approach covered by the work packages already is described in chapter 4, specifically 4.1.

5.2.1 WP1 Prototype Service (Lead: TUC; WP members: GFZ, TUC, ALUF, UHD)

Based on RSD, WP1 starts with the initial provision of the foundation for nfdi.software and the integration with the NFDI AAI. Further iterative updates ensure the integration with BRE, bio.tools, and physics.tools towards a proof-of-concept release. An agile software development approach provides updates of nfdi.software on a regular basis, e.g. bi-weekly or monthly. Each update or release of the service populates further prototypic integration of the components and identified features. So this WP is primarily concerned with connecting bio.tools, RSD, BRE and physics.tools in a prototype so that the information on research software stored in these systems is available in nfdi.software and can thus be found and discovered. Whereas bio.tools offers a mature approach and proven solution from a broad community, and serves as an implementation reference in nfdi.software for linking further solutions from other communities in a successive integration phase, the RSD, BRE and physics.tools offer a user interface, user interface components and alternative search approaches for providing information about research software. Currently, findability of research software remains a challenge. nfdi.software aims to address that challenge with (amongst others) a search and recommender system. BRE's search function will be integrated as a dynamic front end plugin. With this feature, nfdi.software is capable of considering software repositories that are not yet stored or linked in the main metadata repositories. BRE and physics.tools provide complementary approaches to the same cascading search process. While BRE searches for identifiers within the description of the software repository and tries to link them to corresponding publications, physics.tools obtains software links from publications and then links it to the corresponding page or repository. Both approaches provide valuable information on citations, datasets, and which can be found on third party service providers (Zenodo, OpenCitations, ORKG). This additional feature will help researchers to assess the relevance of a given software repository, without having to directly look at it. Aligning both approaches provides a powerful database for the implementation of the prototype. The LLM-based search engine, which is currently being developed for physics.tools, will also be implemented in nfdi.software. It is trained on paper content filling a vector database. It allows the user to find publications and related software products without needing to know the name of a given software. In this way, it also allows the user to find multiple, similar products.

- D1.1 Set-up and initialisation of the base for the proof-of-concept includes the setup of a fork of the RSD for nfdi.software, including the configuration of a CI/CD pipeline for the deployment, as well as the provisioning of server infrastructure (Lead: GFZ).
- D1.2 Migration to B4N-IAM: extension of the current support within RSD for identity providers in order to support user login via IAM4NFDI (Lead: GFZ).

- D1.3 Access of bio.tools and integration of physics.tools in RSD enabling connectivity and harvesting of registry with software tools in the life sciences and enabling mapping, transformation, and representation of harvested software metadata to common metadata schemes to provide access and utilisation by existing search (Lead: ALUF)
- D1.4 Implementation of performance improvements required due to increased amount and availability of software entries and related metadata (Lead: ALUF)
- D1.5 Integration of a search recommender for search results listing existing software entries dynamically enriched by BRE with new, not yet stored nor linked information (Lead: TUC)
- D1.6 Integration of software recommender for software pages displaying specific software found by the search recommender (Lead: TUC)
- D1.7 Integration of AI-based LLM search engine for full-text topical search to display topic related publications and their related software products (Lead: UHD)
- D1.8 Provision of prototype service with the features previously added, further iterative improvements, fixes and polishing (Lead: GFZ)

5.2.2 WP2 Community Building & Networking (Lead: UL; WP members: all)

Community acceptance is of utmost importance for a central service. For becoming an integral part of the NFDI nfdi.software not only requires relevant and usable features and convenience functionalities facilitating the operation of consortia services but also requires networking and community building activities around the software marketplace to ensure close interlocking with the consortia and further base service offers. Also, a constant connection and exchange on the progress of nfdi.software to its intended community is crucial to fulfil the user needs and collect requirements appropriately. The nfdi.software partners already have several connections to other NFDI consortia and are part of several WGs. Additionally, they will also monitor international developments in RDA and EOSC. As a result, there is already survey data and overviews available that will be reused and further developed in this WP. The process of this work package can be subsumed in the following Deliverables

- D2.1 Internal Kick-off Workshop (Lead: GFZ)
- D2.2 Descriptive review of already existing solutions, plans and policies regarding software marketplaces in NFDI consortias (Lead: UL)
- D2.3 Survey for needs and current implementations in NFDI (Lead: OFFIS)
- D2.4 Public Interim presentation and workshop with consortia and base services providers for presenting prototype, survey results, collecting further needs and requirements, and drafting options for integration (Lead: UL)

- D2.5 Develop proposals for curation and quality criteria for entries/records that were included in nfdi.software (Lead: ZB Med)
- D2.6 Public Final Presentation (Lead: UL)

5.2.3 WP3 Concept & Design (Lead: OFFIS; WP members: all)

WP3 will explore and analyse connectivity at three levels: common metadata schema, use of and provision for other Base4NFDI infrastructures; metadata collection and enrichment with regards to NFDI; and third-party infrastructures. A common core metadata layer will ensure basic interoperability between nfdi.software and NFDI consortia marketplaces (e.g., NFDI4Earth OneStop4All, NFDI4DataScience Gateway and Portal, Text+ Curated Software Plattform, NFDI4Culture Registry, SSH Open Marketplace, PUNCH4NFDI marketplace and physics.tools) while also enabling further integration of domain-specific vocabularies (e.g., to describe tasks or operations supported by software particular to a domain – for instance protein function prediction). The nfdi.software common core metadata layer will consider ongoing efforts including CodeMeta, Bioschemas and the FAIR Impact Recommendations for Research Software Metadata [7]. It will also serve as a starting point for existing NFDI consortia marketplaces to align with the common metadata.

nfdi4.software will rely on metadata to connect to other base services and platforms. For instance, nfdi.software could expose the metadata schema in the Terminology Service, consume related software metadata from NFDI knowledge graphs, and provide metadata back to base services. To do so, this WP will analyse current and planned NFDI services. In addition, it will determine how metadata can be programmatically collected from external sources (e.g., GitHub), curated and enriched, and how it can be consumed by other services (e.g., integrating it back to the original GitHub repository or to third-party software marketplaces).

- D3.1 Identification of relevant metadata recommendations and vocabularies for research software and crosswalks to NFDI consortia marketplaces and registries, in cooperation with NFDI-Meta RSMeta WG (Lead: AWK NRW)
- D3.2 Initial concept for metadata exchange with other Base4NFDI services (Lead: ZB MED)
- D3.3 Initial concept to integrate enriched and curated metadata to a selection of target third-party resources (Lead: AWK NRW)
- D3.4 Define the interplay and linking with the EOSC marketplace so that nfdi.software may become a national hub for Germany to provide harmonised information about research software (Lead: OFFIS)

- D3.5 Development of a concept and the design for nfdi.marketplace for implementation in the successive integration phase (Lead: OFFIS)

5.2.4 WP4 Quality Assurance and Coordination (Lead: GFZ, WP members: all)

WP4 covers standard project management activities with regular meetings keeping the implementing partners overall on track and in touch for mutual coordination for the joint tasks. Since it is an overarching WP to keep everyone in sync, to ensure coordinated efforts, to resolve emerging problems, and to finish the WP1-3 deliverables in time with the required quality, WP4 carries out activities continuously without any specific deliverables. We plan quality assurance as an ongoing effort, that we aim to uphold throughout the entire nfdi.software development. The aim is to:

- Ensure metadata standards that harmonise with NFDI efforts and beyond to ensure sustainability and long term accessibility.
- Keep up to date with existing and emerging technologies, best practices and community needs to ensure nfdi.software's reliability and competitiveness
- Document our development efforts, discussions, impediments and results
- Explore automatic and user based feedback mechanisms to keep our data up to date and track changes
- Keep track of community feedback and allocate efforts for continuous improvement
- Ensure coordinated and smooth work on deliverables as scheduled for specific checkpoints

5.2.9 Checkpoints

The initialisation phase with a duration of 12 months contains six checkpoints C1-C6 with a specific set of deliverables associated to each checkpoint and expected to be ready for the next phase of working on deliverables until the next checkpoint.

- C1 ensures the proper start of activities, the set-up of agile processes and regular meetings. It also contains the kick-off meeting in month 1.
- C2 covers the provision of the base for the proof-of-concept with a first tangible, clickable service version available. It includes the setup necessary for professional software engineering, delivers the base version for successive updates with further features over the next months, and enables first-hand experiences for feedback. Also a review of solutions, plans and policies regarding (software) marketplaces in NFDI consortias as well as the identification of relevant metadata recommendations should be reached at this checkpoint.

- C3 extends the proof-of-concept version with core functionality and provides a first release linking with bio.tools and offering a recommender system based on BRE. C3 also identifies relevant aspects discovered by a survey together with an initial concept for metadata exchange, both, enabling crosswalks to NFDI consortia marketplaces and registries. After completion, a deeper understanding of the challenges and opportunities for the development will be shared in the upcoming workshop at C4.
- C4 marks the mid of the initialisation phase. The interim results, namely the proof-of-concept, the survey analysis outcome and the initial concept for metadata exchange, form the base for the interim presentation and the workshop with the consortia and an interested, broader audience. The workshop allows further insights in NFDI wide needs, requirements and emerging services and the interim presentation provides an opportunity of advertising nfdi.software for further potential collaborations.
- C5, together with C6, builds upon the discussions and results of the workshop so that the design for the successive integration phase will be ready at C6. C5 already defines the potential future linking with the EOSC marketplace and comprises a concept for the envisaged curation support with a curation assistant and curator processes, both concerned with keeping already existing data on various platforms up to date with the help of nfdi.software. Also the service is extended with an AI-based LLM search engine.
- C6 marks the end of the proposed initialisation phase with the completion of design for the integration phase and the corresponding proposal. The design forms the foundation for a final public presentation of results and next steps. It is accompanied with a demo of the prototype service discussing its pros and cons.

III Appendix

Bibliography and list of references

[1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., “The fair guiding principles for scientific data management and stewardship”, Scientific data, vol. 3, no. 1, pp. 1–9, 2016.

[2] H. Anzt, F. Bach, S. Druskat, F. Löffler, A. Loewe, B. Y. Renard, G. Seemann, A. Struck, E. Achhammer, P. Aggarwal, et al., “An environment for sustainable research software in germany and beyond: current state, open challenges, and call for action”, F1000Research, vol. 9, 2020.

[3] J. Ison, H. Ienasescu, P. Chmura, E. Rydza, H. Ménager, M. Kalaš, V. Schwämmle, B. Grüning, N. Beard, R. Lopez, et al., “The bio. tools registry of software tools and data resources for the life sciences”, Genome biology, vol. 20, no. 1, pp. 1–4, 2019.

[4] V. Seibert, A. Rausch, and S. Wittek, “Betty’s (re) search engine: A client-based search engine for research software stored in repositories.”, 2023.

[5] M. Löbe, M. Bialke, J. Bienzeisler, J. Drepper, T. Ganslandt, S. Haderer, D. Kraska, M. Lablans, U. Sax, R. Speer, et al., “Toolpool gesundheitsforschung-a repository for software and services focused on supporting clinical and epidemiological research.”, in dHealth, pp. 19–27, 2022.

[6] M. Y. Jaradeh, A. Oelen, K. E. Farfar, M. Prinz, J. D’Souza, G. Kismihók, M. Stocker, and S. Auer, “Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge”, in Proceedings of the 10th International Conference on Knowledge Capture, pp. 243– 246, 2019.

[7] M. Gruenpeter, S. Granger, A. Monteil, N. Chue Hong, E. Breitmoser, M. Antonioletti, D. Garijo, E. González Guardia, A. Gonzalez Beltran, C. Goble, S. Soiland-Reyes, N. Juty, and G. Mejias, “D4.4 - guidelines for recommended metadata standard for research software within eosc”, 2023.

The authors an consortium of the proposal:

Martin Hammitzsch, Beate Hetényi, Björn Grüning, Aleksander Marcic, Matthias Löbe, Leyla Jael Castro, Felix Engel, Konrad Höfner, Michael Zacharias, Oliver Werth, Vasiliy Seibert, Andreas Rausch