








# NFDI4Earth

Deliverable D4.3.2

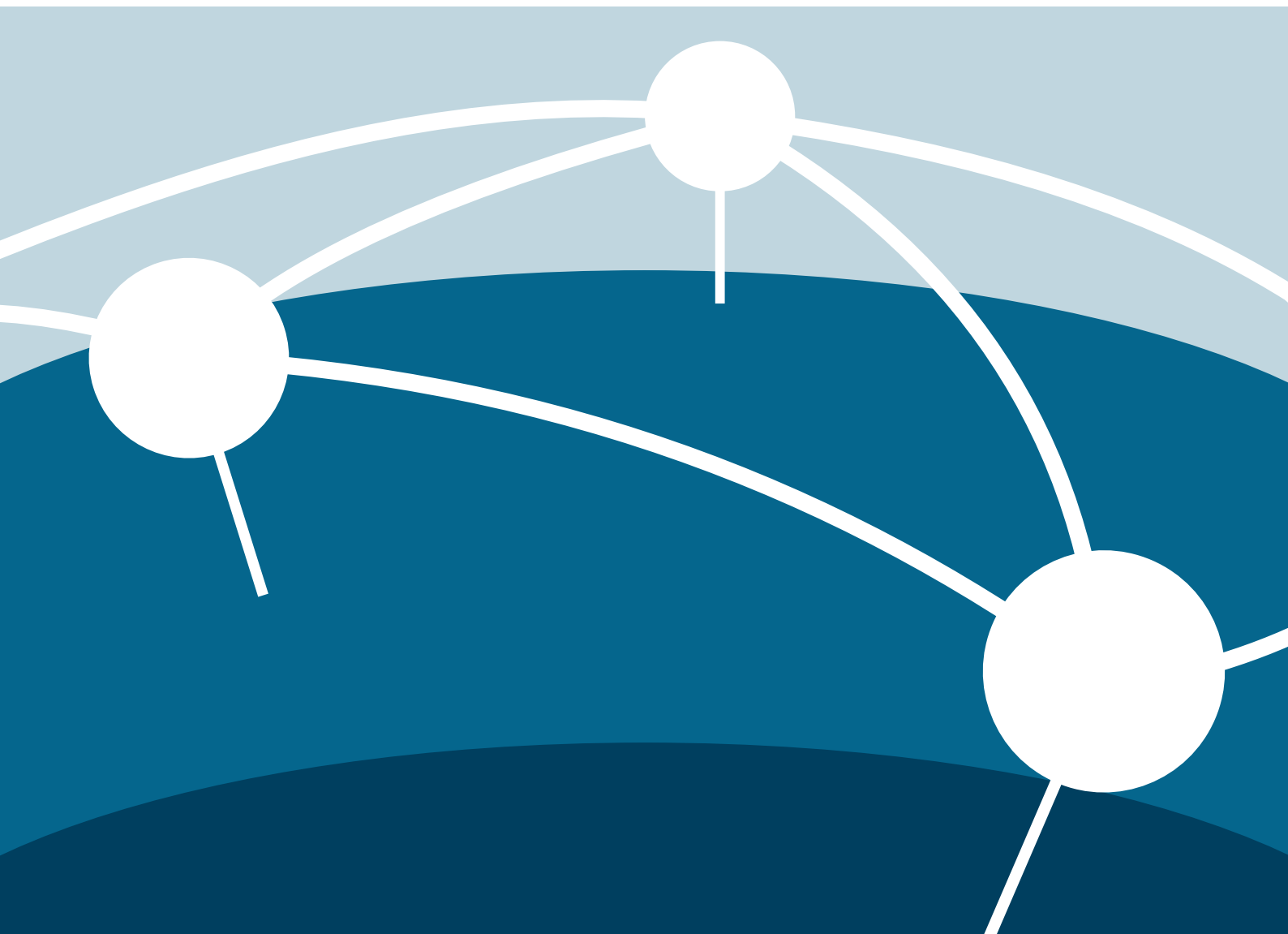
## NFDI4Earth Knowledge Hub - Concept

Auriol Degbello  ([auriol.degbello@tu-dresden.de](mailto:auriol.degbello@tu-dresden.de)), Christin Henzen ,  
Jonas Grieb , Ralf Klammer , Claus Weiland , Lars Bernard ,  
Ralph Müller-Pfefferkorn 

2023-01

DOI: [10.5281/zenodo.7950860](https://doi.org/10.5281/zenodo.7950860)

[nfdi4earth.de](https://nfdi4earth.de)



## Citation

Auriol Degbelo, Christin Henzen, Jonas Grieb, Ralf Klammer, Claus Weiland, Lars Bernard, Ralph Müller-Pfefferkorn. 2023. *NFDI4Earth Knowledge Hub - Concept (NFDI4Earth Deliverable D4.3.2)*. Zenodo. <https://doi.org/10.5281/zenodo.7950860>

## License

This work is licensed under a [Creative Commons "Attribution 4.0 International"](#) license.



## Acknowledgement

This work has been funded by the German Research Foundation (DFG) through the project NFDI4Earth (DFG project no. 460036893, <https://www.nfdi4earth.de/>) within the German National Research Data Infrastructure (NFDI, <https://www.nfdi.de/>).

## Executive summary

Digital Research products from the Earth System Sciences (ESS) are increasingly difficult to find. There is a need for tools that automate their discovery. The Knowledge Hub is one component of the NFDI4Earth architecture that addresses that gap. It will collect information on various ESS research products and serve as the central information source for other NFDI4Earth components (OneStop4All, Living Handbook, EduTrain).

This deliverable introduces the vision for the Knowledge Hub within NFDI4Earth and presents our concepts regarding its implementation and continuous evaluation. The deliverable primarily addresses the technical base of the Knowledge Hub. Strategies on elaborating the content of the Knowledge Hub will be provided by the cooperating NFDI4Earth measures (M1.3, M2.2, M3.1, and M3.2).

## Acronyms

API - Application Programming Interface

ESS - Earth System Science

KG - Knowledge Graph

KH - Knowledge Hub

LH - Living Handbook

ORKG - Open Research Knowledge Graph

RDF - Resource Description Framework

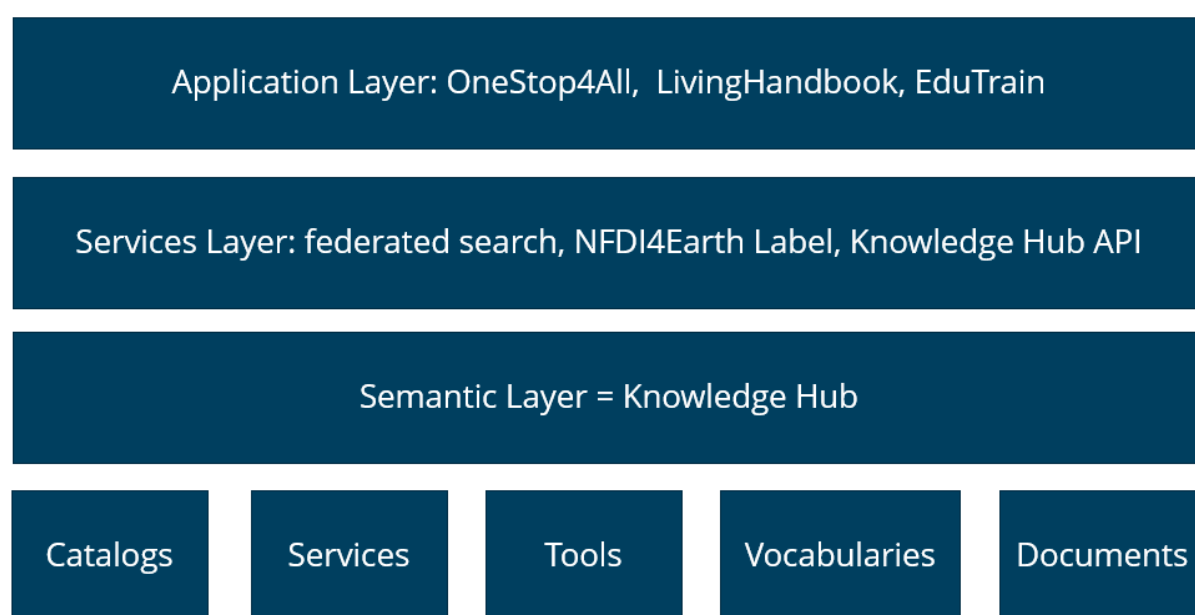
## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Research knowledge graphs</b>	<b>2</b>
<b>3</b>	<b>Vision of the NFDI4Earth Knowledge Hub</b>	<b>3</b>
<b>4</b>	<b>Implementation Concept</b>	<b>5</b>
4.1	Step 1 - Benchmark questions . . . . .	5
4.2	Step 2 - Data collection . . . . .	6
4.3	Step 3 - Data transformation . . . . .	7
4.4	Step 4 - Linking to existing knowledge bases . . . . .	7
4.5	Step 5 - Licensing . . . . .	7
<b>5</b>	<b>Knowledge Hub Maintenance</b>	<b>7</b>
5.1	Provenance tracking of statements . . . . .	8
5.2	Temporal scoping of statements . . . . .	8
5.3	Maintenance of harvesting scripts . . . . .	8
<b>6</b>	<b>Technical setup</b>	<b>9</b>
<b>7</b>	<b>Evaluation of the Knowledge Hub</b>	<b>10</b>
7.1	Accessibility . . . . .	10
7.2	Domain coverage . . . . .	10
7.3	Richness and formality of the represented knowledge . . . . .	11
7.4	Usage . . . . .	11
<b>8</b>	<b>Conclusion</b>	<b>12</b>
	<b>Acknowledgements</b>	<b>12</b>
	<b>References</b>	<b>13</b>
	<b>Appendix: Possible Services on Top of the Knowledge Hub</b>	<b>17</b>

# 1 Introduction

The mission of NFDI4Earth is to address the digital needs of the Earth System Sciences (ESS). The scope of the term 'ESS' is arguably challenging to pinpoint, but for the purposes of this report, it will suffice to say that it encompasses a broad range of fields including (in alphabetical order): Astrophysics/Astronomy, Atmospheric Sciences (e.g., Climatology, Meteorology), Geochemistry, Geodesy/ Cartography, Geography, Geoinformatics, Geology/Paleontology, Geophysics, Hydrology, Ecology, Mineralogy/Crystallography, Oceanography, Planetology, Remote Sensing/ Photogrammetry and Soil Sciences.

Researchers from the ESS produce an increasing number of artefacts during the process of knowledge creation and exchange. In this report, these artefacts are referred to as 'research information products' (or 'digital research products' or 'research products' for short). Examples include: 1) datasets, 2) services, 3) software tools and processing workflows, 4) vocabularies, 5) technical reports, 6) scientific papers, and 7) peer reviews. The aim of NFDI4Earth is to make these digital research products FAIR: findable, accessible, interoperable, and reusable. This report presents the concept of the NFDI4Earth Knowledge Hub, which is one key component of the NFDI4Earth portfolio next to the OneStop4All, the EduTrain and the Living Handbook (see Figure 1).



**Figure 1:** Simplified version of the key components of the NFDI4Earth architecture

The Knowledge Hub is a major backend service of NFDI4Earth and its central information source. It integrates metadata about all NFDI4Earth resources using the Resource Description Framework (RDF) as an abstract data model and is accessed via an Application Programming

Interface (API). Hence, the Knowledge Hub is a knowledge graph in the sense of (Hogan *et al.*, 2022). This report will:

- Briefly present the key idea behind knowledge graphs and provide examples of past/ongoing projects addressing research knowledge graphs;
- Present the vision and scope of the NFDI4Earth Knowledge Hub;
- Articulate the implementation approach of the Knowledge Hub within NFDI4Earth.

## 2 Research knowledge graphs

There are several definitions of knowledge graphs (KGs) in the literature. In this report, the term is defined after (Hogan *et al.*, 2022) as “a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities”. In line with (Brack *et al.*, 2022), a knowledge graph consists of (i) an ontology describing a conceptual model (i.e. classes, relation types, and axioms), and (ii) instance data (e.g. objects, literals, and <subject, predicate, object>-triplets) that follow the constraints set by the ontology. Hence, the construction of KGs involves the design of ontologies and their population with instances.

With (digital) scientific papers and increasingly other digital artefacts generated during the research lifecycle (e.g., datasets, software) becoming available at a fast pace, a number of initiatives and projects have been brought forth to transition from document-centric information communication to knowledge-based information workflows. The concept of ‘science graph’ (a.k.a. ‘research knowledge graph’) is a key enabler of this transition. As Auer *et al.* (2018) put it: “The science graph is a knowledge graph for scholarly communication. [...] The science graph represents scientific information. It does not merely link (metadata about) people, documents, datasets, institutions, grants, etc. but rather represents research contributions semantically, i.e., explicitly and formally”. So far, several research knowledge graphs (RKGs) have been proposed. Below is a non-exhaustive list of examples, in alphabetical order:

- Academia/Industry DynAmics (AIDA) Knowledge Graph (Angioni *et al.*, 2021): represents and collects information about publications and patents in the field of Computer Science.
- CovidPubGraph (Pestryakova *et al.*, 2022): represents and collects information about scientific publications (related to the topic of COVID-19), sections of these scientific publications, authors, bibliographic entries, and named entities.
- Dataset Knowledge Graph (Färber and Lamprecht, 2021): represents and collects information about data sets that were mentioned in at least one publication of the Microsoft Academic Knowledge Graph.

- Geolink Knowledge Graph (Cheatham *et al.*, 2018): integrates data from several geoscience metadata repositories in the United States and offers them as RDF.
- KnowWhere Graph (Janowicz *et al.*, 2022): represents and collects information about people, places and natural hazards (e.g., hurricanes, wildfires, smoke plumes).
- Microsoft Academic Knowledge Graph (Färber, 2019): represents and collects information about scientific publications, authors, institutions, journals, and fields of study.
- OceanGraph (Zárate *et al.*, 2019): represents and collects information about marine species, oceanographic campaigns, scientific publications, environmental variables (e.g., temperature, humidity, oxygen) and geographic locations.
- Open Research Knowledge Graph (Auer *et al.*, 2020): represents and collects information about research papers from the Arts and Humanities, Engineering, the Life Sciences, the Physical Sciences & Mathematics, and the Social and Behavioral Sciences, in a structured manner.
- Springer Nature SciGraph<sup>1</sup>: represents and collects information about nine types of entities: articles, books, chapters, clinical trials, grants, journals, organizations, patents and persons.

While these research knowledge graphs are useful, knowledge graphs that model concepts and workflows of the Earth System Sciences are still needed. This gap is addressed in NFDI4Earth through the Knowledge Hub. For additional work on knowledge graphs within NFDI, see for example (Stocker *et al.*, 2023).

### 3 Vision of the NFDI4Earth Knowledge Hub

Through the integration and interlinking of metadata about ESS research products, the Knowledge Hub intends to facilitate knowledge discovery within the Earth System Sciences.

**The Knowledge Hub will support question answering about ESS research products.**

We are particularly interested in questions about entities with a spatial reference as well as spatiotemporal relationships among these entities. These entities may be the ESS research products themselves (e.g., a dataset) and the things talked about in a ESS research product (e.g., deforestation, hazards, mountains, etc).

*Examples of research products of interest include:* 1) scientific articles (along with their contributions, e.g., hypotheses, methods, study areas, study periods and findings), 2) datasets, 3) data repositories (a.k.a. geoportals), 4) vocabularies (a.k.a. geographic ontologies), 5) software code, 6) software tools, 7) web services (e.g., WMS, WPS, WFS from the Open

---

<sup>1</sup> [https://scigraph.springernature.com/explorer/datasets/data\\_at\\_a\\_glance/](https://scigraph.springernature.com/explorer/datasets/data_at_a_glance/) (accessed: December 28, 2022).



Geospatial Consortium), 8) virtual research environments, 9) researchers, 10) research organizations, 11) peer-reviews and 12) standards/specifications.

As for the types of questions supported, one must distinguish between two types of questions in this context: Type I) questions about ESS products and stakeholders (e.g., questions about scientific publications and their authors/institutions as well as data or software related to a scientific paper), and Type II) questions about the spatial entities talked about in the ESS research products (e.g., questions about a glacier investigated in a scientific publication). The two types of questions are relevant to Earth System Sciences researchers and NFDI4Earth intends to collect meta/data to help answer both.

- *Examples of questions of Type I include:* “Which (OGC) services are available for a given spatial area?”, “Which agent is the contact point for a software S?”, “Which researchers in country C are doing research on topic T?”, “At which repository can I archive my [geophysical] data of [2] GB?”, “Where has a given vocabulary describing geographic entities already been used?”
- *Examples of questions of Type II include:* “What entities are located here?”, “What events occurred here in the past?” and “How does region X compare with region Y?” (Baru *et al.*, 2022). More generally, the questions of type II touch upon known facts, extracted from the scientific literature, about spatial entities investigated by researchers. These facts may touch upon properties of individuals, properties of geographical categories<sup>2</sup>, and relations between geographic entities, for instance: “What are all known facts about the Thwaites Glacier [individual]?”, “What are all known facts about Glaciers [category]?”, and “What are known facts about the relationship (correlation, causation) between the melting of the Thwaites Glacier [property of an individual] and sea-level rise over a period of time [individual]?”. Additional examples of questions related to spatial/geographic aspects of information can be found in (Kuhn, 2012) and (Mai *et al.*, 2021).

Given that the Knowledge Hub (KH) is the backend service of NFDI4Earth, it interacts with other components of NFDI4Earth. On the one hand, the KH stores metadata harvested about these components in a structured format (e.g., metadata about Living Handbook articles and EduTrain material). That is, we will setup metadata harvesting (and conversion) pipelines for Living Handbook articles and EduTrain material, so that structured metadata in RDF about these articles and material can be retrieved from the KH via a SPARQL endpoint. On the other hand, the OneStop4All will use the Knowledge Hub’s SPARQL endpoint to retrieve its content. In particular, metadata information in the KH will be used as input for the search/browsing functionalities offered in the OneStop4All. We also envision that the Living Handbook (LH) articles will embed facts directly extracted from the Knowledge Hub (e.g., facts about the

---

<sup>2</sup> Examples of geographical categories: Mountain, Hill, Valley, River, Rock, Lake, Canyon, Cliff, Ocean, Cave, see (Smith and Mark, 2001) for a thorough discussion.

number of repositories operated by NFDI4Earth could be automatically retrieved from the KH through a SPARQL query and inserted in a paragraph of an introductory LH article about NFDI4Earth).

## 4 Implementation Concept

The implementation of the Knowledge Hub is iterative, with every iteration intended to cover the modelling of, and question answering about a specific type of entity (e.g., dataset, web service), as well as the linking of KH items to other KGs (e.g., Wikidata, DBpedia, Geonames, YAGO or GEMET<sup>3</sup>). Every iteration has at least the following five steps, and ends with a new release of the Knowledge Hub. A release involves the population of the Knowledge Hub with new content, the publication of all scripts needed to create this content (e.g., harvesting scripts) on GitHub and/or GitLab, and a documentation of the types of questions supported by the Knowledge Hub's SPARQL endpoint. The software running the Knowledge Hub is presented in the section "Technical setup". This section focuses on the process of *creating content* for the Knowledge Hub. The following steps are envisioned:

- Step 1 - Formulation of the benchmark questions
- Step 2 - Data collection
- Step 3 - Data conversion to RDF
- Step 4 - Linking to existing knowledge bases
- Step 5 - Licensing

### 4.1 Step 1 - Benchmark questions

In the spirit of competency questions used in the field of ontology engineering (see e.g., (Grüninger and Fox, 1995; Degbelo, 2017)), we plan to explicitly list all questions the NFDI4Earth Knowledge Hub will answer. The listing of these questions helps (i) to delineate the scope of the KH, and (ii) compare it to other KHs. The questions will be listed first in natural language and then translated into SPARQL. As a result, they may evolve into benchmark questions for research knowledge graphs for the Earth System Sciences (much in the same way as SPARQL-based queries were used as benchmarks to test the compliance of triple stores to the GeoSPARQL standard in (Jovanovik, Homburg and Spasić, 2021)). Each new release of the Knowledge Hub will come with a documentation of the answers supported.

---

<sup>3</sup> <https://www.eionet.europa.eu/gemet/en/about/> (accessed: February 6, 2023).

## 4.2 Step 2 - Data collection

Once the questions are listed, datasets about the entities of interest will be collected. The entities of interest (e.g., scientific articles, web services and so on) were mentioned in the Section “Vision of the NFDI4Earth Knowledge Hub”. There are several sources to start with, e.g., re3data<sup>4</sup> (dataset about research repositories), GitHub<sup>5</sup> (dataset about software code), and ROR<sup>6</sup> (dataset about organizations), and these will be used in the process of data collection about entities where appropriate.

A strategy to collect datasets to populate a knowledge graph is through *knowledge harvesting* from structured and unstructured data sources. This strategy was at the heart of projects such as DBpedia or YAGO. Another strategy, followed by the Open Research Knowledge Graph project is to rely on the use of *crowdsourcing*. The project funds curation grants, where grantees are tasked with adding key research questions and corresponding research contributions in their research field to the ORKG<sup>7</sup>. The key idea is that human annotators add semantic representations to artefacts a posteriori. A third strategy proposed in (Kuhn and Dumontier, 2017) is *genuine semantic publishing* (that strategy may be called Metadata/FAIRness by Design). Knowledge harvesting and crowdsourcing rely on semantic annotation to add semantic representations to existing artefacts. By contrast, genuine semantic publishing is about including semantic representations that originate from the researchers themselves, from the start (see (Kuhn and Dumontier, 2017)). These semantic representations are a primary component of the artefact, available at the time of publication and are fine-grained in the sense that they can refer to the whole artefact (e.g., a research article) or components of that artefact (e.g., a section or paragraph of that research article). Answering questions of type II mentioned in the Section “Vision of the NFDI4Earth Knowledge Hub” in particular will require either crowdsourcing or genuine semantic publishing. Hence, NFDI4Earth intends to generate its knowledge graph through a combination of the three strategies. In particular, we intend to use the creation of articles in the Living Handbook and of educational material in the EduTrain to experiment with (i.e., design and test) approaches that support easy and scalable genuine semantic publishing by contributors.

---

<sup>4</sup> <https://www.re3data.org/> (accessed: December 28, 2022).

<sup>5</sup> <https://github.com/> (accessed: December 28, 2022).

<sup>6</sup> <https://ror.org/> (accessed: December 28, 2022).

<sup>7</sup> [https://orkg.org/about/28/Curation\\_Grants](https://orkg.org/about/28/Curation_Grants) (accessed: December 28, 2022).

### 4.3 Step 3 - Data transformation

The key goal of this step is the conversion of the datasets from Step 2 to RDF. At this stage, a choice of ontologies/vocabularies is necessary to annotate these datasets semantically. The LOV platform (Vandenbussche *et al.*, 2016) is a good source for the search of existing ontologies/vocabularies. Ontologies specifically created for the annotation of scholarly publications and scientific documents were reviewed in (Ruiz Iniesta and Corcho, 2014). These existing ontologies and vocabularies will also be considered for re-use. New ontologies/vocabularies will be created on an as-needed basis. The description of spatial/geographic aspects of entities or phenomena will follow the best practices for spatial data publishing on the Web<sup>8</sup>. The Open Geospatial Consortium has produced several standards for the description of datasets, sensors and services within the geospatial domain<sup>9</sup>. Specifying mappings to enable the conversion from these standards to RDF format is also a core task at this stage.

### 4.4 Step 4 - Linking to existing knowledge bases

Linking a dataset to other datasets is useful to provide context<sup>10</sup> and this step intend to link the Knowledge Hub to other relevant knowledge graphs. Examples already mentioned above include: Wikidata, DBpedia, Geonames, YAGO and GEMET. Depending on the topic at hand, some of the research knowledge graphs mentioned in the Section “Research knowledge graphs” could be linked to as well.

### 4.5 Step 5 - Licensing

The choice of an open licence is key to facilitate the reuse of the datasets converted to RDF through the knowledge graph construction process. The open licence for the Knowledge Hub will be clarified at the release stage.

## 5 Knowledge Hub Maintenance

Knowledge bases are not static, but dynamic. That is, facts, relationships and the rules represented in the Knowledge Hub may change (e.g., students graduate and ‘lose’ their student status, the affiliation of researchers may change, the postal address of an organization

---

<sup>8</sup> <https://www.w3.org/TR/sdw-bp/> (accessed: December 28, 2022).

<sup>9</sup> <https://www.ogc.org/docs/is> (accessed: December 28, 2022).

<sup>10</sup> <https://www.w3.org/DesignIssues/LinkedData.html> (accessed: December 28, 2022).

may change, the maintenance of data repositories is discontinued, values for environmental variables [e.g., temperature, humidity] fluctuate). There is a need of a strategy to properly represent these changes. This research area is still in need of mature approaches (see (Xi, 2020; Weikum *et al.*, 2021)), but the recommendations made in (Weikum *et al.*, 2021) are sensible. The development of the metadata maintenance strategies is ongoing at the moment of this writing and the documentation of the maintenance concept will be presented in subsequent deliverables about the Knowledge Hub. For now, we are committed to the following ideas regarding provenance tracking and the temporal scoping of statements.

### 5.1 Provenance tracking of statements

Weikum *et al.* (2021) suggested that each statement in a knowledge base should be annotated with provenance metadata about:

- the source (e.g., web page) from where the statement was obtained, or the different sources when multiple inputs are combined;
- the timestamp(s) of when the statement was acquired; and
- the extraction method(s) by which it was acquired, for example, the rule(s), pattern(s) or classifier(s) used in case of automatic knowledge extraction.

This recommendation is particularly relevant in this context, since we are envisioning data collection using different strategies: knowledge harvesting, crowdsourcing and genuine semantic publishing.

### 5.2 Temporal scoping of statements

Temporal scoping denotes the annotation of the statements in the Knowledge Hub with the time at which they are valid (e.g., timepoints or time intervals, see (Weikum *et al.*, 2021)). In GIScience, temporal scoping has been discussed for researchers' life lines (Trame, Keßler and Kuhn, 2013), but temporal scoping needs to be extended to many more statements (e.g., model the time at which a statement about the value of an environmental variable is valid). We will explore approaches to realize (scalable) temporal scoping for the statements of interest in the Earth System Sciences and document them in upcoming deliverables about the Knowledge Hub.

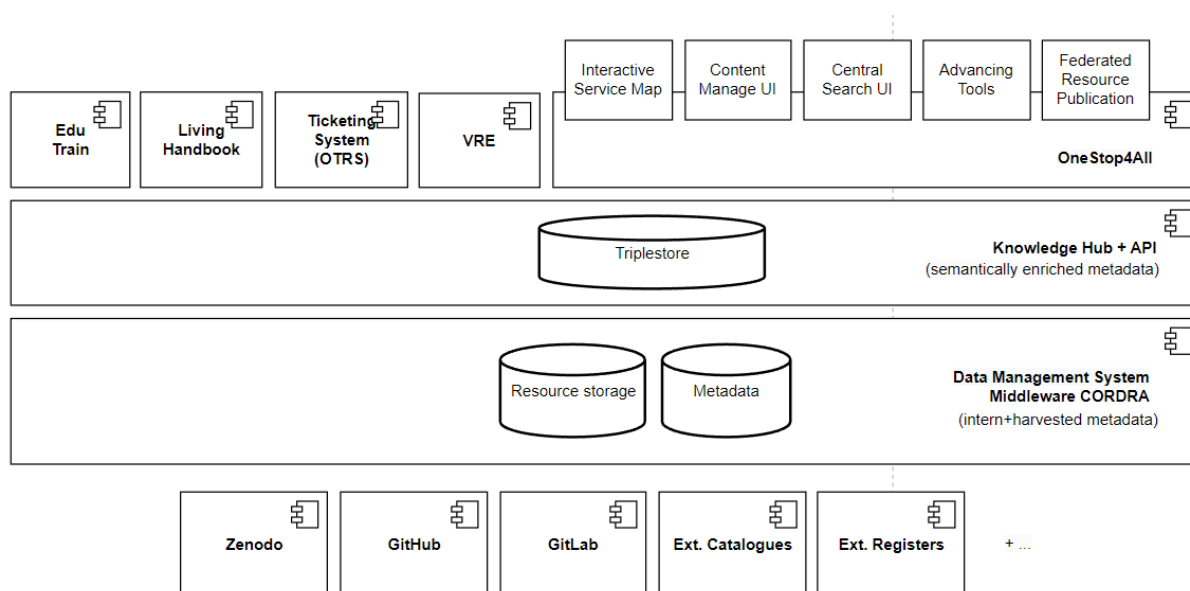
### 5.3 Maintenance of harvesting scripts

In addition to provenance tracking and temporal scoping discussed above that touch upon how changes are represented in the Knowledge Hub, there is a need to maintain the harvesting

pipelines through which information is harvested and updated to the KH. To that end, automated tests for the different harvesting scripts will be developed. These tests are intended to inform the developer teams of NFDI4Earth about the scripts where adjustments are needed (e.g., the harvesting fails because the original API to retrieve content has changed).

## 6 Technical setup

Figure 2 presents the current version of the NFDI4Earth architecture and their components. Particularly relevant to this deliverable about the Knowledge Hub are the triple store (to store all metadata encoded in RDF) and the middleware layer to manage that metadata. Popular options to store RDF-based metadata in the context of research knowledge graphs are GraphDB (used for example by (Zárate *et al.*, 2019; Liu *et al.*, 2022)) and Virtuoso (used for instance by (Färber, 2019; Pestryakova *et al.*, 2022)). Change is a key requirement in this context (see Section “Knowledge Hub Maintenance”). Hence, we will use the open-source software Cordra to manage metadata (e.g., edit, delete statements). That is, the metadata harvested (from external catalogues, from code repositories, or document repositories) will be stored first as a JSON-LD-serialized digital objects in Cordra<sup>11</sup>, and then exported as RDF to the triple store. Every digital object in Cordra is corresponding to a named graph in the triple store and contains all triples about one entity of interest in the real world. The central search (i.e., algorithms for question answering) as input for the OneStop4All will be performed on the triple store.



**Figure 2:** NFDI4Earth architecture (as of January 2023)

<sup>11</sup> <https://www.cordra.org/> (accessed: December 28, 2022).

The additional abstraction and management layer represented by Cordra allows us to track changes and provenance in the knowledge graph as described in the previous chapter. Every update of the knowledge graph consists in a new JSON-LD document for a digital object. During update operations, the change between the old and the new version can be assessed and recorded as a JSON Patch. Additional metadata of the change operation, such as who is the acting agent and the timestamp, is also recorded and stored in a provenance graph for the digital object in accordance with the PROV Data Model<sup>12</sup>.

Since the technical setup of the Knowledge Hub is largely based on the configuration and extension of existing software, the main focus of the technical implementation is on the development of harvesting pipelines. This will follow the iterative approach presented in the chapter “Implementation Concept”. Each harvesting pipeline is a programmable job that is responsible for harvesting knowledge from a specific source, mapping the information to RDF triples according to the data model of the Knowledge Hub and pushing the triples to the knowledge graph. The jobs must be able to be scheduled to run regularly and also upon request.

## 7 Evaluation of the Knowledge Hub

Several criteria have been proposed in the literature to evaluate knowledge graphs. A few are particularly of interest.

### 7.1 Accessibility

As discussed in (Färber *et al.*, 2017), this refers to aspects on how the data within the Knowledge Hub can be accessed. Two criteria are worth mentioning:

- Availability: the probability that a feasible query is correctly answered in a given time range
- Response time: the delay between the point in time when the query was submitted and the point in time when the query response is received

### 7.2 Domain coverage

The benchmark questions mentioned in the Section “Implementation Concept” are a means of documenting the topical coverage of the Knowledge Hub. Furthermore, Heist *et al.* (2020) listed the following instance-based metrics:

---

<sup>12</sup> <https://www.w3.org/TR/2013/REC-prov-dm-20130430/> (accessed: January 31, 2023).

- The number of instances in a graph
- The number of assertions (or edges between entities)
- The average and median linkage degree (i.e., how many assertions per entity does the graph contain?)

Furthermore, Weikum *et al.* (2021) mentioned precision and recall, in combination with a sampling strategy to collect ground truth data from human annotators as a means of assessing both the domain coverage and the correctness of the statements in the KGs. These two criteria will be used to assess the domain coverage and correctness of the NFDI4Earth Knowledge Hub as well.

### 7.3 Richness and formality of the represented knowledge

Heist *et al.* (2020) listed the following schema-level metrics:

- The number of classes defined in the schema
- The number of relations defined in the schema
- The average depth and width (branching factor) of the class hierarchy
- The complexity of the schema

### 7.4 Usage

This is an indicator of the adoption of the Knowledge Hub by the community and its utility. Three criteria are of relevance:

- The number of API calls
- Their spatial and temporal distribution
- The number of services build on top of the Knowledge Hub. This is an adaption of the idea of “data stories” used by the European Data Portal to document use cases that result from the use of open data<sup>13</sup>. Examples of services anticipated are presented in the Section “Appendix: Possible Services on Top of the Knowledge Hub”. NFDI4Earth intends to set up a feedback workflow to collect these services systematically.

---

<sup>13</sup> <https://data.europa.eu/en/publications/datastories> (accessed: December 28, 2022).



## 8 Conclusion

This report has presented our concepts regarding the implementation and evaluation of the NFDI4Earth Knowledge Hub. Below is a recap of the key ideas:

- The Knowledge Hub is a knowledge graph that will serve as the central information source for other NFDI4Earth components.
- The Knowledge Hub integrates metadata about all NFDI4Earth resources using the Resource Description Framework (RDF) as an abstract data model and is accessed via an Application Programming Interface (API).
- The data model for the information represented in the Knowledge Hub is built incrementally and iteratively. It will reuse terms of existing ontologies and vocabularies as much as possible. The requirements for this data model are provided by the dataset(s) to integrate in the Knowledge Hub and the envisioned (spatio-temporal) questions to answers.
- Triple stores often lack capabilities for metadata management. Hence, the open-source software Cordra will be used for metadata management (e.g., edit, delete statements).
- Maintenance aspects addressed in the project include: provenance tracking of statement, temporal scoping of statement, as well as the automated testing of harvesting scripts.
- The Knowledge Hub will be evaluated using criteria such as availability, response time, domain coverage, richness and formality of the represented knowledge, and usage.

## Acknowledgements

We thank members of the NFDI4Earth Data Model Expert Group and the NFDI4Earth Knowledge Hub Working Group for their contributions in shaping these ideas (in alphabetical order): Ivonne Anders, Stefan Hachinger, Sibylle Haßler, Claudia Müller, Valentina Protopopova, Thomas Rose, Farzaneh Sadeghi, Markus Stöcker, Freya Thießen, Peter Valena, Alexander Wellmann.

## References

- Angioni, S., Salatino, A., Osborne, F., Recupero, D.R. and Motta, E. (2021) 'AIDA: A knowledge graph about research dynamics in academia and industry', *Quantitative Science Studies*, 2(4), pp. 1356–1398. Available at: [https://doi.org/10.1162/qss\\_a\\_00162](https://doi.org/10.1162/qss_a_00162).
- Auer, S., Kovtun, V., Prinz, M., Kasprzik, A., Stocker, M. and Vidal, M.-E. (2018) 'Towards a knowledge graph for science', in R. Akerkar, M. Ivanovic, S.-W. Kim, Y. Manolopoulos, R. Rosati, M. Savic, C. Badica, and M. Radovanovic (eds) *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics (WIMS 2018)*. Novi Sad, Serbia: ACM, p. 1:1--1:6. Available at: <https://doi.org/10.1145/3227609.3227689>.
- Auer, S., Oelen, A., Haris, M., Stocker, M., D'Souza, J., Farfar, K.E., Vogt, L., Prinz, M., Wiens, V. and Jaradeh, M.Y. (2020) 'Improving access to scientific literature with knowledge graphs', *Bibliothek Forschung und Praxis*, 44(3), pp. 516–529. Available at: <https://doi.org/10.1515/bfp-2020-2042>.
- Bartoschek, T., Pape, G., Kray, C., Jones, J. and Kauppinen, T. (2014) 'Gestural interaction with spatiotemporal linked open data', *OSGeo Journal*, 13(1), pp. 60–67.
- Baru, C., Campbell, L., Dade, A., Fulay, P., Loewi, A., Maughan, D., Mohedas, I., Molnar, L., Pozmantier, M., Reksulak, M., Smith, S. and Tehrani, N. (2022) 'The NSF Convergence Accelerator program', *AI Magazine*, 43(1), pp. 6–16. Available at: <https://doi.org/10.1002/aaai.12032>.
- Brack, A., Hoppe, A., Stocker, M., Auer, S. and Ewerth, R. (2022) 'Analysing the requirements for an Open Research Knowledge Graph: use cases, quality requirements, and construction strategies', *International Journal on Digital Libraries*, 23(1), pp. 33–55. Available at: <https://doi.org/10.1007/s00799-021-00306-x>.
- Cheatham, M., Krisnadhi, A., Amini, R., Hitzler, P., Janowicz, K., Shepherd, A., Narock, T., Jones, M. and Ji, P. (2018) 'The GeoLink knowledge graph', *Big Earth Data*, 2(2), pp. 131–143. Available at: <https://doi.org/10.1080/20964471.2018.1469291>.
- Degbelo, A. (2017) 'A snapshot of ontology evaluation criteria and strategies', in R. Hoestra, C. Faron-Zucker, T. Pellegrini, and V. de Boer (eds) *Proceedings of the 13th International Conference on Semantic Systems - SEMANTICS 2017*. Amsterdam, The Netherlands: ACM Press, pp. 1–8. Available at: <https://doi.org/10.1145/3132218.3132219>.
- Degbelo, A. (2021) 'An ontology design pattern for geovisualization content description', in E. Blomqvist, T. Hahmann, K. Hammar, P. Hitzler, R. Hoekstra, R. Mutharaju, M. Povedaf, C. Shimizuc, M. Skjaeveland, M. Solanki, V. Svátek, and L. Zhou (eds) *Advances in Pattern-based Ontology Engineering*. IOS Press, pp. 279–291. Available at: <https://doi.org/10.3233/SSW210019>.
- Färber, M. (2019) 'The Microsoft Academic Knowledge Graph: A linked data source with 8 billion triples of scholarly data', in C. Ghidini, O. Hartig, M. Maleshkova, V. Svátek, I.F. Cruz, A. Hogan, J. Song, M. Lefrançois, and F. Gandon (eds) *The Semantic Web - ISWC 2019 - 18th International*

*Semantic Web Conference*. Auckland, New Zealand: Springer (Lecture notes in computer science), pp. 113–129. Available at: [https://doi.org/10.1007/978-3-030-30796-7/\\_8](https://doi.org/10.1007/978-3-030-30796-7/_8).

Färber, M., Bartscherer, F., Menne, C. and Rettinger, A. (2017) 'Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO', *Semantic Web*. Edited by A. Zaveri, D. Kontokostas, S. Hellmann, J. Umbrich, A. Zaveri, D. Kontokostas, S. Hellmann, and J. Umbrich, 9(1), pp. 77–129. Available at: <https://doi.org/10.3233/SW-170275>.

Färber, M. and Lamprecht, D. (2021) 'The data set knowledge graph: Creating a linked open data source for data sets', *Quantitative Science Studies*, 2(4), pp. 1324–1355. Available at: [https://doi.org/10.1162/qss\\_a\\_00161](https://doi.org/10.1162/qss_a_00161).

Gil, Y., Pierce, S.A., Babaie, H., Banerjee, A., Borne, K., Bust, G., Cheatham, M., Ebert-Uphoff, I., Gomes, C., Hill, M., Horel, J., Hsu, L., Kinter, J., Knoblock, C., Krum, D., Kumar, V., Lermusiaux, P., Liu, Y., North, C., Pankrati, V., Peters, S., Plale, B., Pope, A., Ravela, S., Restrepo, J., Ridley, A., Samet, H. and Shekhar, S. (2019) 'Intelligent systems for geosciences: an essential research agenda', *Communications of the ACM*, 62(1), pp. 76–84. Available at: <https://doi.org/10.1145/3192335>.

Goodchild, M.F., Guo, H., Annoni, A., Bian, L., de Bie, K., Campbell, F., Craglia, M., Ehlers, M., van Genderen, J., Jackson, D., Lewis, A.J., Pesaresi, M., Remete-Fulopp, G., Simpson, R., Skidmore, A., Wang, C. and Woodgate, P. (2012) 'Next-generation digital earth', *Proceedings of the National Academy of Sciences*, 109(28), pp. 11088–11094. Available at: <https://doi.org/10.1073/pnas.1202383109>.

Grüniger, M. and Fox, M.S. (1995) 'Methodology for the design and evaluation of ontologies', in *Proceedings of the IJCAI Workshop on Basic Ontological Issues in Knowledge Sharing*. Montreal, Quebec, Canada. Heist, N., Hertling, S., Ringler, D. and Paulheim, H. (2020) 'Knowledge graphs on the web - an overview', in I. Tiddi, F. Lécué, and P. Hitzler (eds) *Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges*. IOS Press (Studies on the semantic web), pp. 3–22. Available at: <https://doi.org/10.3233/SSW200009>.

Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G.D., Gutierrez, C., Kirrane, S., Gayo, J.E.L., Navigli, R., Neumaier, S., Ngomo, A.-C.N., Polleres, A., Rashid, S.M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S. and Zimmermann, A. (2022) 'Knowledge graphs', *ACM Computing Surveys*, 54(4), pp. 1–37. Available at: <https://doi.org/10.1145/3447772>.

Janowicz, K., Hitzler, P., Li, W., Rehberger, D., Schildhauer, M., Zhu, R., Shimizu, C., Fisher, C.K., Cai, L., Mai, G., Zalewski, J., Zhou, L., Stephen, S., Gonzalez, S., Mecum, B., Lopez-Carr, A., Schroeder, A., Smith, D., Wright, D., Wang, S., Tian, Y., Liu, Z., Shi, M., D'Onofrio, A., Gu, Z. and Currier, K. (2022) 'Know, Know Where, KnowWhereGraph: A densely connected, cross-domain knowledge graph and geo-enrichment service stack for applications in environmental intelligence', *AI Magazine*, 43(1), pp. 30–39. Available at: <https://doi.org/10.1002/aaai.12043>.

- Jovanovik, M., Homburg, T. and Spasić, M. (2021) 'A GeoSPARQL compliance benchmark', *ISPRS International Journal of Geo-Information*, 10(7), p. 487. Available at: <https://doi.org/10.3390/ijgi10070487>.
- Kuhn, T. and Dumontier, M. (2017) 'Genuine semantic publishing', *Data Science*. Edited by S. Peroni, 1(1–2), pp. 139–154. Available at: <https://doi.org/10.3233/DS-170010>.
- Kuhn, W. (2012) 'Core concepts of spatial information for transdisciplinary research', *International Journal of Geographical Information Science*, 26(12), pp. 2267–2276. Available at: <https://doi.org/10.1080/13658816.2012.722637>.
- Liu, Z., Gu, Z., Thelen, T., Estrecha, S.G., Zhu, R., Fisher, C.K., D'Onofrio, A., Shimizu, C., Janowicz, K., Schildhauer, M., Stephen, S., Rehberger, D., Li, W. and Hitzler, P. (2022) 'Knowledge explorer: exploring the 12-billion-statement KnowWhereGraph using faceted search (demo paper)', in M. Renz and M. Sarwat (eds) *Proceedings of the 30th International Conference on Advances in Geographic Information Systems (SIGSPATIAL 2022)*. Seattle, Washington, USA: ACM, p. 73:1–73:4. Available at: <https://doi.org/10.1145/3557915.3561009>.
- Mai, G., Janowicz, K., Zhu, R., Cai, L. and Lao, N. (2021) 'Geographic question answering: challenges, uniqueness, classification, and future directions', in P. Partsinevelos, P. Kyriakidis, and M. Kavouras (eds) *Proceedings of the 24th AGILE Conference on Geographic Information Science (AGILE 2021)*, pp. 1–21. Available at: <https://doi.org/10.5194/agile-giss-2-8-2021>.
- Pestryakova, S., Vollmers, D., Sherif, M.A., Heindorf, S., Saleem, M., Moussallem, D. and Ngomo, A.-C.N. (2022) 'CovidPubGraph: A FAIR Knowledge Graph of COVID-19 Publications', *Scientific Data*, 9(1), p. 389. Available at: <https://doi.org/10.1038/s41597-022-01298-2>.
- Ruiz Iniesta, A. and Corcho, O. (2014) 'A review of ontologies for describing scholarly and scientific documents', in A. García Castro, C. Lange, P. Lord, and R. Stevens (eds) *4th Workshop on Semantic Publishing (SePublica)*. Anissaras, Greece (CEUR workshop proceedings). Available at: <http://ceur-ws.org/Vol-1155>.
- Smith, B. and Mark, D.M. (2001) 'Geographical categories: an ontological investigation', *International Journal of Geographical Information Science*, 15(7), pp. 591–612. Available at: <https://doi.org/10.1080/13658810110061199>.
- Stocker, M., Rossenova, L., Shigapov, R., Betancort, N., Dietze, S., Murphy, B., Bölling, C., Schubotz, M. and Koepler, O. (2023) *Knowledge graphs - working group charter (NFDI section-metadata)*. Zenodo. Available at: <https://doi.org/10.5281/zenodo.7515324>.
- Trame, J., Keßler, C. and Kuhn, W. (2013) 'Linked Data and time - modeling researcher life lines by events', in T. Tenbrink, J. Stell, A. Galton, and Z. Wood (eds) *Spatial Information Theory - 11th International Conference*. Scarborough, UK: Springer International Publishing, pp. 205–223. Available at: [https://doi.org/10.1007/978-3-319-01790-7\\_12](https://doi.org/10.1007/978-3-319-01790-7_12).

Vandenbussche, P.-Y., Ateazing, G.A., Poveda-Villalón, M. and Vatan, B. (2016) 'Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web', *Semantic Web*. Edited by M. Dumontier, 8(3), pp. 437–452. Available at: <https://doi.org/10.3233/SW-160213>.

Weikum, G., Dong, X.L., Razniewski, S. and Suchanek, F. (2021) 'Machine knowledge: creation and curation of comprehensive knowledge bases', *Foundations and Trends® in Databases*, 10(2–4), pp. 108–490. Available at: <https://doi.org/10.1561/19000000064>.

Xi, J. (2020) *A brief survey on knowledge graph update*.

Zárate, M.D., Rosales, P., Braun, G.A., Lewis, M., Fillottrani, P.R. and Delrieux, C. (2019) 'OceanGraph: Some initial steps toward a oceanographic knowledge graph', in B. Villazón-Terrazas and Y. Hidalgo-Delgado (eds) *Knowledge Graphs and Semantic Web - First Iberoamerican Conference (KGSWC 2019)*. Villa Clara, Cuba: Springer (Communications in computer and information science), pp. 33–40. Available at: [https://doi.org/10.1007/978-3-030-21395-4\\_3](https://doi.org/10.1007/978-3-030-21395-4_3).

## Appendix: Possible Services on Top of the Knowledge Hub

Beyond NFDI4Earth components to explore the Knowledge Hub's content (OneStop4All, LivingHandbook, EduTrain), we anticipate that the Knowledge Hub will serve as a backbone for intelligent systems in the earth system sciences, much in line with the vision of (Gil *et al.*, 2019). The possibilities may be numerous, but here are some promising first ideas:

- Evidence checker: help users find evidence for and against a statement, based on the knowledge available
- Hypotheses generator: find (spatial) hypotheses for occurrences of phenomena based on the knowledge available
- Intelligent question answering: a chatbot gives meaningful answers to questions asked by users about research data management
- Interoperable software agents: build software agents that communicate with each other to exchange datasets
- User interfaces of diverse kinds, for instance:
  - An interactive 2D environment that uses the map as an entry point to visualize all that is known about a location, in line with the idea of 'vertical context elicitation' (Goodchild *et al.*, 2012)
  - An interactive 3D environment to navigate through resources: nodes as places, edges as paths between these places
  - An interactive user environment that enables researchers to record their insights while interacting with the data, based on the design pattern from (Degbelo, 2021)
  - Exploration of the data through gestural interaction on large displays, as done for example by (Bartoschek *et al.*, 2014)