A FAIR Problem-Solving Lifecycle Architecture

Nikolay A. Skvortsov[0000-0003-3207-4955]

Federal Research Center "Computer Science and Control"
of the Russian Academy of Sciences, Moscow, Russia
nskv@mail.ru

Abstract. Research infrastructures are intended to provide access to scientific data and resources needed for problem-solving. Approaches to support interoperability and reuse of heterogeneous resources in such infrastructures need investigation. Researchers tend to integrate and reuse existing resources but do not spend much effort to publish the resources created during solving scientific problems to make them reusable in communities. As the result, further users of these results have to spend their time and effort on integrating resources to reuse them. We propose an architecture of research infrastructures that are initially based on the lifecycle of problem-solving in research communities providing interoperability and reuse of the sources and previous research results. In this architecture, most of the data maintenance tasks are moved from the data integration stage to the data publishing stage, and data are manipulated following the domain specifications accepted by communities. This makes providing the interoperability and reuse of resources in the infrastructure more competent and easy and avoids repeated integration.

Keywords: research infrastructures, data reuse, problem-solving lifecycle, software architecture.

1 Introduction

Contemporary research is mainly based on data about the research object, often collected from different sources, initially focused on various aspects and purposes of object describing and obtained by various methods and instruments. At the same time, there are continuously growing needs for covering a wide variety and large volumes of data and in possible directions of their analysis. That applies to almost any scientific, commercial, or state field of research. There is an urgent need to create research infrastructures that provide both the data themselves and tools for comprehensive support of research on them.

The creation of data infrastructures was due to big volumes of research data needed to be stored, replicated in data centers in different regions, and shared to different research groups for immediate and probably long-term processing for obtaining new knowledge from them. Thus, data infrastructures were aimed at data collecting, long-term preserving, suitable organization of data for processing and analysis, and sharing

between research groups with different interests. For instance, the CERN data processing problems were similar. However, later, with the growth of data centers preserving and providing data of different nature and in different research domains. So the problems of data preserving had shifted to the problems of interoperability and reuse of heterogeneous research data for problem-solving.

Such infrastructures, in addition to long-term data preservation, need search and access tools, description of data semantics, tools for data integration, automation of research on data. Ehe ultimate goal of these tools is to support unimpeded, useful, and productive reuse of data and related resources for solving research tasks. The guiding principles for providing reusable data, known by the abbreviation FAIR (findable, accessible, interoperable, reusable data) [1] have become one of the main directions for the actively discussed foundations for creating global interdisciplinary research infrastructures.

Despite the universal recognition of the FAIR data guiding principles as a measure of the quality of technologies to select, positions on organizing, automating, and simplifying the process of solving problems in research infrastructures have not yet been established and are actively being discussed. Moreover, there are obvious problems, the overcoming of which is seen not so much in technical solutions, but in changing the established research paradigm. Even if research infrastructures provide a wide range of services, there are not enough incentives for researchers to use them in a way to meet the FAIR data principles, to make the research process easier, and the results more accessible and convenient for reuse by other members of the research community.

So there is a need for infrastructures with an architecture itself that would direct researchers to a new paradigm and culture of research, in which providing the ability to reuse any results of data management and data analysis and the continuity of research in domain communities are put at the forefront.

In [2], a lifecycle for problem-solving in research infrastructures was proposed, which was the result of thinking on the guiding principles of FAIR data. It uses semantic approaches, domain specifications developed by research communities, and timely publishing data at all stages of solving research problems, from the selection and integration of source data to long-term preservation of research results. Each stage of such a life cycle is designed in such a way as to preserve the results of passing it and provide their reuse in subsequent studies.

The purpose of this work is to propose an architecture of research infrastructures that would provide the implementation of the problem-solving lifecycle mentioned above, and therefore, fulfill the principles of FAIR data in them. The next two sections discusses the accustomed approaches used in research infrastructures. Section 4 briefly introduces the theses of the proposed problem-solving lifecycle to overcome these deficiencies. Section 5 describes the architecture of research infrastructures that implement the discussed approach. Then, in section 6, it is explained how the proposed architecture provides the principles for managing the problem-solving declared as part of FAIR.

2 Issues of the Research Process

Data infrastructures were originally designed for the consolidation of research data, access to them, and the ability to share them. In some specialized infrastructures, methods and workflows were provided to be used together with data. Later research infrastructures appeared that include wider functions and provide tools for processing and analyzing data. In other words, these were environments for supporting research on data.

Among the basic set of services in existing research infrastructures are registration, long-term preservation, and search for datasets, and support of data analysis.

The architecture of the EUDAT CDI [3] research infrastructure is initially designed to support the collaboration of large communities as well as individual users in work with data. EUDAT CDI can be considered as a research infrastructure with a classic set of services. Long-term data preservation with support of permanent identifiers and provenance is provided by B2SAFE and B2HANDLE services. Digital objects [4] are supported that contain several datasets under one identifier, and all of them have their identifiers. The B2SHARE and B2ACCESS provide data storage and access services. The B2STAGE service provides replication of datasets for problem-solving in intermediate storage. A data type registry (DTR) is created to define descriptive metadata. The B2NOTE service provides semantic annotation of data, and the B2FIND metadata catalog allows searching for data by annotations.

The architecture of the open science cloud EOSC [5, 6] is designed to combine a set of research infrastructures. It is defined as a system of systems based on registering services by providers. EOSC functionalities are provided as services on nodes distributed across some organizations and regions. It is declared that EOSC services should promote and support the principles of FAIR.

The architecture defines several dozen classes of services that are considered as a minimum viable product. Such services include those specifically designed to serve researchers, research administrators, third-party service providers, as well as EOSC managers. Like in EUDAT, the metadata structure is an enhanced set of fields, which were started in Dublin Core. Interestingly, the user interface and the search engine services are related to EUDAT services.

GO FAIR [7] is an initiative aimed at the interaction of experts and the wide community to implement the FAIR data principles. Discussions, events, and initiatives in it are held for popularization, making incentives, education, and direction for implementing the principles. Technologies and components of architectures are proposed, investigation of certain aspects of the principles are initiated, the best practices are chosen, and the response of the community is evaluated. One of the priorities of GO FAIR is the coordination of different studies necessary to create the EOSC infrastructure.

An important initiative for the development of the Internet of FAIR Data and Services (IFDS) [8] is also being discussed in the GO FAIR community in connection with the EOSC infrastructure. It is aimed at the support of type-driven automatic data processing. It links data with tools and calculations semantically relevant to them. Thus, data processing research infrastructures can be performed automatically based on the

data semantics using relevant tools. Such studies comply with the guiding principles of FAIR data as a kind of machine-actionable approach.

The FAIRsFAIR [9] project is aimed at promoting selected practices and support representative projects such as data centers or repositories that support the FAIR data principles and use certain sets of technologies for it. For this purpose, criteria for evaluating projects for FAIRness have been developed. Explanations of the meaning of various aspects of the FAIR data principles and possible ways to follow these concepts have been developed. Specifying ontologies, data schemas, interfaces, and protocols, and mapping them to one another are recommended [10]. This simplifies the understanding of the principles themselves and in the research community, and technologies implementing them can be selected from the project case studies. These projects show the advantages of following the FAIR data principles for solving problems of interoperability and reuse of data in research.

The FREYA Project [11] was devoted to the study of global persistent identifiers. As a result of its implementation, tools have been created to support identifiers of various entities types (such as research publications, data, programs, people, and organizations), related standards, and their integration. In the extensible registry of the Knowledge Hub, there are services to support existing standards of global persistent identifiers, domain-specific identifier types, linking data with their metadata and provenance information based on the graph of identifiers of different types [12], cross-resolution of identifiers in different standards, entity annotation, and others. This work was carried out in connection with the needs of the EOSC research infrastructure and the results are proposed as a part of the EOSC-hub. The PID Graph is a conceptual part of EOSC. Another result of the project is a community that continues to promote the results of the project [13].

All the described projects and initiatives indicate that research communities are aware of the need to conceptualize their domains for greater interoperability and data processing and research automation. However, in those projects, it is recognized that the mentality of researchers is changing slowly, and willing and organizational efforts are needed to use semantic descriptions of domains as community requirements, implement FAIR principles and change the situation with machine-actionable data processing.

When solving problems in research infrastructures, users usually have to work with heterogeneous data. In general, the mapping of data models and the integration of heterogeneous structures are considered as a non-trivial task, so solutions for it are avoided. To overcome them, it is proposed to use widely used universal formats and trivial data representation. However, this does not keep users from the semantic heterogeneity of the data. Data and other resources are available in their original formats, and the use of them for problem-solving, in any case, requires understanding and reconciliation. To reconcile heterogeneous data, research infrastructures can provide services that help to integrate resources. These tools solve the problems of data interoperability but do not yet ensure the effectiveness of applying them.

Anyway, an accustomed approach to starting new research is to search for data relevant to the problem being solved, find out their structure, semantics, and interfaces, and reconcile them for the possibility to use them in a problem-solving application.

This usually takes a lot of time, effort, and manual work, moreover, it is done multiple times for the same data in each independent research. The results of this big work often remain hardly applicable for reuse.

The lack of support for specialized methods related to data in the domain community becomes a problem due to the need to develop the same methods many times in each research. Creating libraries of methods is not enough, since extending them with new research results is difficult, and they remain unprepared for search and poorly integrated.

The publishing results of problem-solving in research infrastructures is usually supported or even required to make the results available to reuse or reproduce in further research. However, simple procedures of data and service publishing in research infrastructures allow providers of newly obtained data and developed services to publish them as-is with simple descriptions for potential users. This immediately becomes a problem for their reuse, since users have to do a lot of work to reconcile them before using them.

Developers of research infrastructures, including pilot projects of new infrastructures, discuss the weak development of researchers' skills in interaction with infrastructures. This probably happens because publishing is postponed until the end of the project and so it is not done properly. Thus, the research paradigm of users is changing slowly, the problem-solving in them remains involving a lot of manual work. This, in turn, is a weak incentive for the use of infrastructures by research communities.

So leaving the final research results unintegrated, the inevitability of repeated manual work with small variations for data integration and method implementation, and the inability to reuse the results obtained at these intermediate stages of problem-solving mean contradicts the principles of FAIR data, since data are not reusable in domain communities.

3 Formal Domain Specification in Different Disciplines

The use of ontologies in some research domains is very natural and convenient. Ontologies are especially useful for classifying a large number of different types of entities defined by their certain properties and the values of attributes inherent to these entities.

For example, the domains where research communities have developed and widely used ontologies for a long time are bioinformatics and biomedicine. The gene structures or proteins define physiological characteristics, organ functions, pathologies, and other aspects of the living organism description. It allows to classify the properties and identify the types of structures that affect the characteristics of the organism. There are a lot of subdomains in biomedicine. Each of them can be used by specific communities, described by specific concepts, and participate in the classification of entities based on correlation with other subdomain concepts. These subdomains are supplemented with domains for the description of the observation tools and methods. Those tools and methods define sets of the observed and evaluated characteristics of domain entities. BioPortal [15] collects ontologies describing different aspects of the biomedicine domain. The need to harmonize these ontologies was realized and some of them were emerged and

reconciled with each other. The conceptual framework of the research process contains concepts for laboratory research on physical entities, creating information entities, and then analyzing information objects to generate secondary data and new knowledge. Semantic annotation of data and resources in terms of these ontologies is widely used in biomedicine and it is a good basis for achieving data interoperability in research domain communities and infrastructures. Many of the rest ontologies in BioPortal remain overlapping, contradicting, and unformal.

Similarly, in materials science, the description of the microstructure of materials defines their chemical and physical properties. Thus ontologies allow to define and classify all the variety of such characteristics. Domains with parameters dependent on the microstructure may include chemical, magnetic, optical characteristics, crystal structures, and others. The OntoCommons [14] project aims to standardize a set of ontologies and data representations in materials science and to offer practical tools for it as well. A system of ontologies related to each other is being developed and form the hierarchy. At the most abstract level, there are widely used higher-level ontologies, then middle-level ontologies related to the domain, in particular, those included in EMMO [16] and their extension, and finally, ontologies of various subdomains based on them. A set of use cases demonstrate the effectiveness of the approach.

In astronomical research, the correlation of the observed parameters of astronomical objects with their astrophysical parameters is important. Experiments [17, 18] with formal semantic specifications of various domains in this discipline show that during solving representative domain research problems the vocabularies of domain concepts have been quickly saturated and become mostly sufficient for solving other problems in the same domains. On the other hand, Observation Core Data Model (ObsCore) [19] can be considered as one of the best standards in astronomy for common domain specifications and FAIR data management. It includes features of general domains of astronomical observation such as descriptions of the spatial axis and time, observational properties and spectral characteristics of astronomical objects, annotations of data with semantic definitions (descriptors in the UCD standard are unfortunately are sufficiently ambiguous and not formally defined), metadata, and provenance model and allows querying it all simultaneously.

The ontologies, data schemas, interfaces, and protocols in specific research domains confirm that research communities move towards a formal conceptual definition of their research domains. Some of the most common domains are ready for formal semantic approaches since domain specifications are enhanced and used intensively used in them.

4 The Research Problem-Solving Lifecycle

In [2], a problem-solving lifecycle was proposed, which was the result of reasoning about the guiding principles of FAIR data. The core of such a lifecycle is the support of formal domain specifications by communities since data semantics plays an important role in the integration of data and the capabilities of machine-controlled data management processes. Based on the management of domain descriptions, the search

for relevant data, methods, and other resources related to data for their reuse is provided. The problem-solving and the presentation of the research results are presented in the domain terms. Thus, the research results remain available for reuse in subsequent research within the community.

Communities of researchers in such infrastructures become interested in describing the domains of their interests and in semantic integration of data related to their domains for repeated reuse without additional integration.

The lifecycle of research problem-solving includes a framework of basic activities (Fig. 1, highlighted in blue):

- description of the domain, formulation, and analysis of the problem in it;
- selection of data resources relevant for problem-solving;
- selection of methods for solving the problem or implementing them;
- getting the results of problem-solving and publishing them.

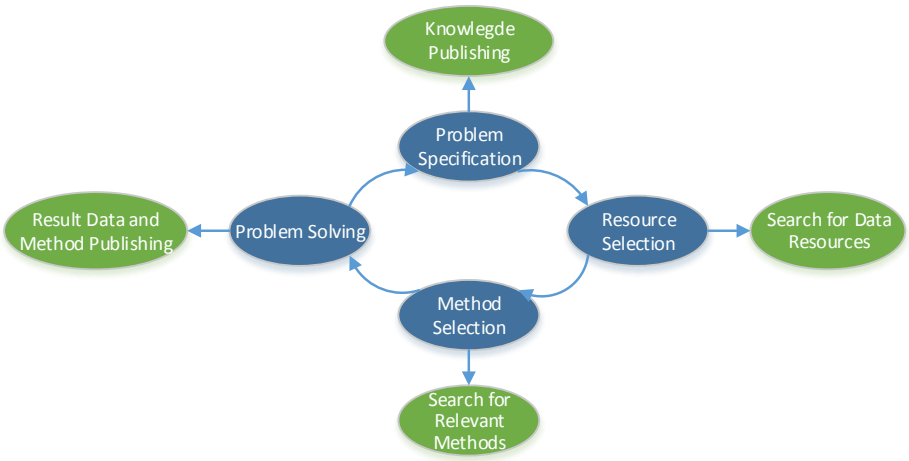

Fig. 1. The framework of the lifecycle of problem-solving on data

However, other important intermediate stages of problem-solving are put on this basic framework. An infrastructure should provide the possibility of a formal search for semantically relevant resources based on domain ontologies (Fig. 1, highlighted in green):

- search for domain concepts, related requirement models of previously solved (sub-) problems, and publishing the knowledge obtained during the problem analysis;
- search for relevant data resources that were previously registered in the community;
- search for suitable methods and workflows for solving the problem, previously implemented for use in the community (the principles of the formal ontological description of methods are presented in [20]);

- publishing the results of problem-solving and their metadata in terms of domain ontologies to provide the search for them during solving subsequent research problems in the community.

If relevant resources have been integrated with the domain specifications, and the results of integration have been published, they can be reused with the domain specifications without additional integration.

If some resources have not been previously used by the community at all or found resources have been registered but not integrated with the domain descriptions, the lifecycle offers the tools for integrating them with the domain specifications:

- finding correspondences between the elements of the integrated data model and the elements of the canonical model or already known extensions of the canonical data, making data model mapped to the canonical one and work with a unified model [21] to solve problems in it;
- finding correspondences of the elements of the integrated schema with the conceptual schemas and formats used in the domain of the community, and mapping them into each other, as well as integrating data at the object level between the reconciled schemas;
- search for matches of elements of the method and workflow specifications for integration and implementation.

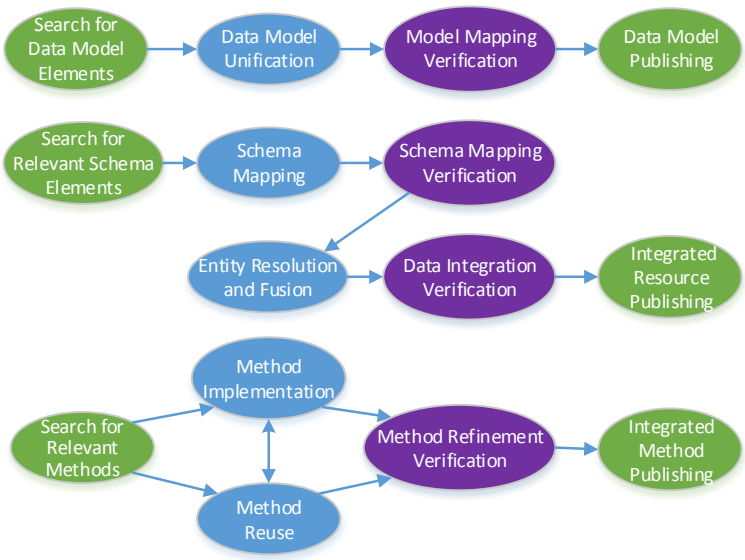

Fig. 2. Integration of heterogeneous resources for problem-solving

Each of these stages starts with a formal search for relevant resource elements from the ontological point of view and finishes on publishing the integrated resources (data models, data resources, methods, and workflows). At the same time, the integration

results themselves are published too. The established correspondences between the integrated resources and the specifications of the domains to which they are mapped, so that the integration process is not repeated in the future, but reused (see Fig. 2, highlighted in green). Activities for formal verification (highlighted in purple) include services for verifying the results of integration stages or reuse based on proof of the refinement relationship between specifications [22]. The integration support processes themselves are highlighted in light blue in Fig. 2.

Semantic approaches to managing heterogeneous data and formally proving the correctness of their integration or reuse are combined with tools for quality publishing of research and data management results at different stages.

To implement the principles of FAIR data, the principles for creating research infrastructures based on the presented problem-solving lifecycle on heterogeneous data could be the following.

- The infrastructure is aimed at supporting communities working in certain research domains. The basis of the infrastructure is the research problem-solving lifecycle providing data and resource reuse in communities.
- Domain knowledge is the basis for uniting research communities. Formal specifications of ontologies with the reasoning feature are used for the description of domains and related resources.
- Metadata in terms of domain specifications annotate and describe available resources. They are used for the semantic search for resources in communities. Different kinds of resources can be semantically annotated including specifications of datasets, conceptual schemas, methods signatures, data model elements, and others. Any manipulations with resource metadata are performed over metadata registries.
- Providers should publish resources to make resources reusable in communities. The publishing of problem-solving results includes describing them in terms of domain ontologies and storing the metadata in registries to make available the search for them in communities. Described data should be preserved and accessible by identifiers. All useful results should be published as soon as they appear following the problem-solving lifecycle or a special data management plan. The infrastructure ensures the reuse of different kinds of problem-solving results, including data, methods, integration results, so that any repeated work in communities is minimized.
- Automated reasoning on semantic specifications and metadata is used as much as possible at all stages of problem-solving (search for relevant specifications, resource integration and reuse) so that machine-controlled data management and problem-solving at a semantically significant level are possible.

5 The Proposed Architecture of Research Infrastructures

Following the described lifecycle of research problem-solving and the principles of creating research infrastructures based on such a lifecycle, the following software architecture is proposed for their implementation (Fig. 3).

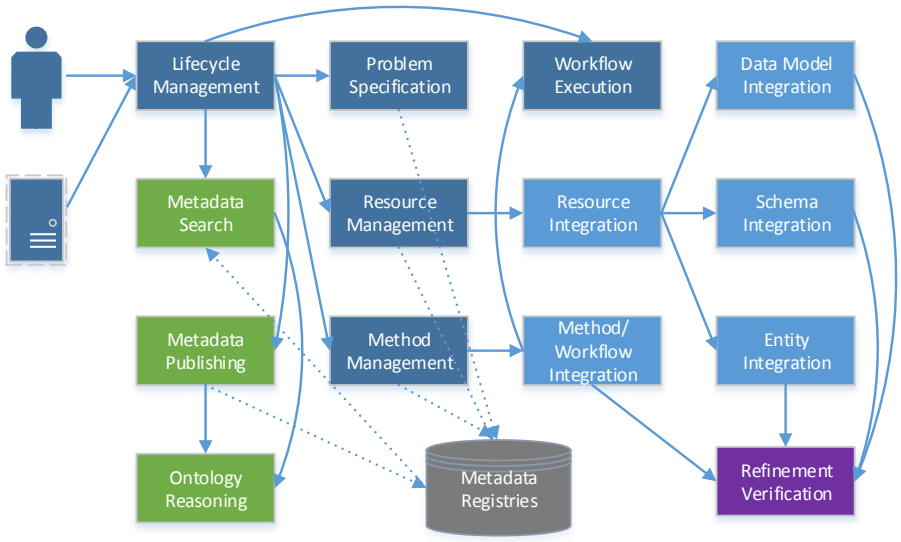

Fig. 3. The architecture supporting the problem-solving lifecycle

A person or a machine (possibly controlled by an expert) as an agent interacts with the problem-solving lifecycle management workflow to use the components of the architecture as its activities.

Metadata registries support long-term preservation and access to domain specifications, metadata of the integrated data resources, available implementations of methods applicable to domain objects, descriptions of data models mapped to the canonical model. In addition to metadata registries, repositories for replicated data and digital objects are used.

Metadata search services are based on tools that support logical reasoning in terms of formal ontologies (highlighted in green) over metadata in registries. They support the search for data resources, methods, workflows relevant to the problem, and similar specification elements during the process of heterogeneous resource integration. Metadata publishing services relate resource identifiers with metadata and preserve them in the registries. They support the publishing of data models, schemas, data resources, methods, and workflows at different stages of the problem-solving lifecycle.

Resource integration services (highlighted in light blue) can provide support for the integration of data resources at all levels, including the integration of their data models, schema mapping, instance integration. Integration of methods and workflows follow the resource integration.

Formal semantics refinement verification services (highlighted in purple) can provide formal proof of the correctness of mapping data models, schemas, objects, methods, or workflows to one another.

The issues of authorization, authentication, licensing, and access restriction are tied to domains, communities and participating in them. Domains are organized as partially

ordered set where subdomains use knowledge of some more general domains. Communities of domains have access to resources of domains and more common domains. Researches, research groups, projects, and machine agents must join the community to get access to its resources. Access to community implies commitment to its domain specifications and standards. Joining communities is regulated by license politics and may be open or restricted. Digital objects are considered as resource access units.

The prototype of the research infrastructure architecture is being implemented using Hadoop technologies. To materialize data from integrated sources in the form of files, the distributed storage is applied. The Spark framework was chosen as a distributed computing platform for data analysis. However, for structured data, it is possible to use distributed databases and frameworks that work in Hadoop too, for example, HBase and Hive. The implementation is also considered to be moved to a cloud or other distributed technologies. The variety of data presented here can be very large: copies of the source data resources published intermediate data (collected, selected, transformed, processed, generated data), data on resource integration, data obtained as a result of solving problems (new data resources, models, programs, libraries, and others).

To implement metadata registries based on formal ontologies, an RDF repository or a framework for RDF can be chosen. It must support the OWL language, processing in RAM, and store or interact with a distributed DBMS for storing domain specifications and metadata. The framework needs to integrate a tool for reasoning in description logics for resource classification (for example, Pellet), and the SPARQL endpoint for metadata queries.

A workflow execution framework is used to activate the problem-solving lifecycle as well as problem-specific workflows.

Programming languages that support Spark, for example, Java, Python, Scala can be used for the implementation of some tools in this architecture or problem-specific methods. Data can be stored in a distributed HDFS file system with the ability to extract them by global identifiers. But it can be accessed without materialization in the research infrastructure repositories by the same identifiers from the original long-term preservation locations.

6 Following the Principles of FAIR Data by the Research Infrastructure Architecture

The presented architecture of research infrastructures for problem-solving on data aims to cover the guiding principles of FAIR data as much as possible since any resources related to FAIR data should be FAIR as well. The arguments given below allow evaluating the decisions from the view of every FAIR data principle.

6.1 Findable Data

According to the FAIR data principles, finding datasets and services requires human- and machine-readable metadata, global data identifiers that explicitly link metadata to data, and tools of indexing or registering data based on such metadata.

Data and related resources are provided in the proposed architecture with metadata representing their semantic annotations. They widely, comprehensively, and formally describe the meaning and properties of data from different perspectives defined by a set of domain ontologies. Any concepts existing in the domain should be expressible in the annotations. Semantic annotations refer to the data using permanent unique global identifiers.

The publishing data consists of defining metadata and registering them in registries. This makes it possible to search for data and resources by specifying the necessary concepts as queries as well.

6.2 Accessible Data

To make data accessible, compatibility at the level of access protocols, the ability to extract data by global identifiers are important. Access restriction rules are defined. And the access to metadata is maintained even if the data itself is no longer available.

In the proposed architecture, in this regard, data can be extracted from the original places of their long-term preservation using standard Internet protocols and global identifiers. Or copies of them are available in distributed repositories using the same identifiers. Digital objects also provide identification and access through other integrated protocols. Problems of identifying data fragments and reconciliation of identifiers of partially mismatched data entities can be solved by using the identifier graph [12], although the primary idea of creating it had other goals.

After publishing, the metadata are permanently preserved in the registries. If changes are necessary, the metadata is supplemented with the expiration as provenance information, and new versions of the metadata are created.

Any types of resources, such as ontologies, data resources, services, workflows, the results of resource integration are also published, can be found and extracted by identifiers. Thus, they meet the principles of FAIR data in their turn.

6.3 Interoperable Data

The principles of achieving data interoperability include the use of a formal, accessible, and widely used knowledge representation language at the core of data management. Various types of resources, in particular, dictionaries for describing related to data and necessary for their description, should themselves comply with the principles of FAIR data. It is also important to provide a qualitative description of linking the data related to each other or used together. These principles are designed to solve the issues of compatibility of applications and workflows with data during their processing, analysis, and preservation. It is noteworthy that the main principle of data interoperability is the use of a formal language of knowledge representation. This gives hope for machine-controlled data management but would not recognize machine learning methods perhaps because of their probabilistic character and poorly interpretability.

In the proposed architecture, formal models, in particular, description logics, are used for describing the domain knowledge and semantic annotation of data and resources. During publishing in the registries, automatic formal logical reasoning is used

for the classification of data and resources from the point of view of ontologies. Automatic reasoning provides the ability to interpret the meaning of data and resources when solving problems by both a human and a machine.

Domain ontologies as dictionaries are themselves formal and available for reuse after their publishing in registries, thus they comply with the principles of FAIR data.

The proposed method of semantic annotation allows not only to use the concepts of ontologies directly to define the data semantics but to describe data with expressions that define subconcepts in terms of existing ontologies using the entire expressive power of the knowledge representation language. Thus, it becomes possible to express necessary constraints and relationships of concepts within one ontology, or between several ontologies, and then use logical reasoning to classify data by annotations.

Data, method implementations, the results of resource integration, workflows, metadata, and other resources refer to each other within digital objects. They are described in the registries, and they are accessible using an ontology-based search and by storing identifiers of each other.

6.4 Reusable Data

The reusability of data in solving various problems as the ultimate aim of applying the principles of FAIR data is provided as follows. Datasets should be widely described by a set of accurate and relevant attributes to assess their applicability in the problems being solved. They should be supplied with a license for their use, accompanied by detailed provenance information, and should comply with the community domain standards. Thus, it is possible to deal with data copies and combinations for various purposes.

Following these principles, in addition to the research domain specifications, knowledge specifications include special ontologies for describing non-functional properties (attributes) of data and evaluating their applicability. These are such domains as data quality, measurement quality, and data provenance [23]. Non-functional data requirements can be expressed in terms of such special ontologies, and become part of queries when searching for relevant data for problem-solving. Metadata in terms of these ontologies describe attributes of data.

Special standards supported by research communities are registered like data models or conceptual schemas used in their domains. Within the framework of digital objects, they are supplied with information about their integration. The data that meet these standards refer to those data models and schemas. Licenses related to domain communities allow organizing access to digital objects and ensure commitment to formal specifications used for access, publishing and reuse of their resources.

7 Conclusion

An analysis of the principles for building research infrastructures that provide management of the problem-solving lifecycle ensuring the reuse of various types of resources is presented in the paper. An architecture designed to implement such a lifecycle has

been developed. It is shown how the presented architecture of research infrastructures complies with the principles of FAIR data.

The investigations related to the FAIR data principles in big research infrastructures are wide. The real focus of them remains to provide services the research community tends to use. And the mentality is being changed slowly. Therefore, formal semantic technologies and commitment to specified domain knowledge are not in priority. The proposed research infrastructure architecture is based on prospective semantics-based solutions for changing the paradigm of the lifecycle of research over data. It does not avoid them in favor of the short-term broad needs of communities in data infrastructures. It gives the possibility to reuse heterogeneous source data, results of their integration, methods developed for problem-solving, and research results without multiple integrations inside the research community.

Acknowledgments. The work was carried out using the infrastructure of shared research facilities CKP "Informatics" of FRC CSC RAS [24], supported by the Russian Foundation for Basic Research, grants 19-07-01198, 18-29-22096.

References

1. Wilkinson M., et al. The FAIR Guiding Principles for scientific data management and stewardship. In: Scientific data, Vol. 3 (2016).
2. Skvortsov N. A., Stupnikov S. A. Managing Data-Intensive Research Problem-Solving Lifecycle. Data Analytics and Management in Data Intensive Domains (DAMDID 2020), CCIS, vol. 1427. Springer. 2021.
3. Technical report, EUDAT. Available at http://hdl. handle. net/11304/2433d23a-6079-49a6-9010-ca534f6e348d, 2015.
4. Wittenburg, P. From Persistent Identifiers to Digital Objects to Make Data Science More Efficient. In: Data Intelligence, Vol. 1, Iss. 1, P. 6-21 (2019). DOI: 10.1162/dint_a_00004.
5. Candela, L.; Castelli, D.; La Rocca, G.; Lukkarinen, A.; Manghi, P.; Pagano, P.; Papadopoulou E. (2018) Initial EOSC Service Architecture. EOSCpilot Deliverable D5.1 https://eoscpilot.eu/content/d51-initial-eosc-service-architecture
6. Candela L., Castelli D., Zoppi F. Final EOSC Service Architecture. EOSCpilot Deliverable D5.4. https://eoscpilot.eu/content/d54-final-eosc-service-architecture
7. GO FAIR Initiative. https://www.go-fair.org/go-fair-initiative/
8. The Internet of FAIR Data & Services. https://www.go-fair.org/resources/internet-fair-data-services/
9. FAIRsFAIR. FAIR Semantics, Interoperability, and Services. https://www.fairsfair.eu/fair-semantics-interoperability-and-services-0
10. Franc, Y. L. et al. D2.5 FAIR Semantics Recommendations. Second Iteration. FAIRsFAIR 2020. DOI: 10.5281/zenodo.4314321
11. FREYA. Project outputs. https://www.project-freya.eu/en/resources/project-output
12. Fenner, M., Aryani, A. Introducing the PID Graph. DataCite (2019). DOI: 10.5438/jwvf-8a66
13. The PID Forum. https://www.pidforum.org/
14. OntoCommons. Ontology-driven data documentation for Industry Commons. https://ontocommons.eu/

15. Noy, N. et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research, 37(suppl_2), pp. W170–W173 (2009). DOI: 10.1093/nar/gkp440
16. The European Materials & Modelling Ontology (EMMO). https://github.com/emmo-repo/EMMO/
17. Skvortsov N. A. et al. Conceptual Approach to Astronomical Problems. Astrophysical Bulletin, 71(1), pp. 114-124. Pleiades Publishing (2016). DOI: 10.1134/S1990341316010120
18. Skvortsov, N. A. Conceptual Model Reuse for Problem Solving in Subject Domains. Modelling to Program, p.191-211. Springer CCIS (2021). DOI: 10.1007/978-3-030-72696-6_10.
19. Louys, M., et al. Observation Data Model Core Components and its Implementation in the Table Access Protocol. Version 1.1. IVOA Recommendation. IVOA (2017). http://www.ivoa.net/documents/ObsCore/
20. Nikolay A. Skvortsov, Leonid A. Kalinichenko, Dmitry Yu Kovalev. Conceptualization of Methods and Experiments in Data Intensive Research Domains // Data Analytics and Management in Data Intensive Domains, XVIII International Conference, DAMDID/RCDL 2016, Ershovo, Moscow, Russia, October 11-14, 2016, Revised Selected Papers, Springer International Publishing AG 2017 Springer. Communications in Computer and Information Science, Vol. 706, P. 3-17, 2017.
21. Stupnikov S., Kalinichenko L. (2019) Extensible Unifying Data Model Design for Data Integration in FAIR Data Infrastructures. In: Manolopoulos Y., Stupnikov S. (eds) Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2018. Communications in Computer and Information Science, vol 1003, P. 17-36. Springer, Cham. doi.org/10.1007/978-3-030-23584-0_2
22. Abrial, J.-R.: The B-Book: Assigning Programs to Meanings. Cambridge: Cambridge University Press, 1996
23. Belhajjame K., Cheney J., Corsar D., Garijo D., Soiland-Reyes S., Zednik S., Zhao J. PROV-O: The PROV Ontology. W3C Recommendation, World Wide Web Consortium (W3C), 2013. https://www.w3.org/TR/prov-o.
24. Regulations of CKP “Informatics”. http://www.frccsc.ru/ckp