Future Scientific Data Infrastructure: Towards Platform Research Infrastructure as a Service (PRIaaS)

Yuri Demchenko, Cees de Laat, Wouter Los
Complex Cyber Infrastructure Group
University of Amsterdam
Amsterdam, The Netherlands
e-mail: {y.demchenko, C.T.A.M.deLaat, W.Los}@uva.nl

Abstract— **Modern Science is becoming increasingly data driven and works with a large amount of data, which are heterogeneous, distributed and require special infrastructure for data collection, storage, processing, and visualisation. Science digitalization, likewise industry digitalization, is facilitated by the explosive development of digital technologies and cloud based infrastructure technologies and services. This paper presents two lines of analysis: one is retrospective analysis related to the European Research Infrastructure (RI) development stages and timeline from centralized to distributed and current Federated Interoperable; another line provided analysis of digital technologies trends and identified what technologies will impact the future Scientific Data Infrastructure (SDI). Based on this analysis, the paper proposes a vision for the future RI Platform as a Service (PRIaaS) that incorporates recent digital technologies and enables platform and ecosystem model for future science. Notably the proposed PRIaaS adopts TMForum Digital Platform Reference Architecture (DPRA) that will simplify building and federating domain specific RIs while focusing on the domain specific data value chain with data protection and policy based management by design.**

Keywords- Scientific Data Infrastructure, Science Digitalisation, Platform Research Infrastructure as a Service (PRIaaS), TMForum Digital Platform Reference Architecture (DPRA), Big Data Infrastructure, Research Data Management, EOSC, FAIR

I. INTRODUCTION

Modern Science is becoming more and more data driven and works with a large amount of data, which are heterogeneous, distributed and require special infrastructure for data collection, storage, processing, and visualisation. Science digitalization, likewise industry digitalization, is facilitated by the explosive development of digital technologies as well as infrastructure technologies and services.

Targeting new large scale problems such as the genome, climate, global warming, which became possible with Big Data and Cloud Computing technologies, requires further research platforms advancement by new technologies implementation and consolidation.

Future digital science opens new possibilities of cross-domain/cross-sector integration and consolidation of resources and capacities. It will require new type of infrastructure that would provide extended functionality to collect, store, distribute, process, exchange and preserve research data to support common knowledge grow and exchange [1, 2, 3]: We will refer to this new infrastructure as Future Scientific Data Infrastructure (FutureSDI or FutureRI).

Recent European initiatives and projects such as the European Open Science Cloud (EOSC) [4], Research Data Alliance (RDA) [5] facilitated implementation of the FAIR (Findable, Accessible, Interoperable, Reusable) data principles [6] that allow for effective data exchange and integration across scientific domains, making scientific data a valuable resource and a growth factor for the whole digital economy and society. To uncover the potential of the future digital and data driven science, the FutureSDI must provide a platform for effective use of scientific data by allowing creating specialized consistent ecosystems supporting full cycle of the value creation from data collection to model creation and knowledge acquisition and exchange. Shift of the focus from infrastructure operation to value creation will require new FutureSDI design approach, operation and evolution to respond to changing requirements and evolving technologies. Growing infrastructure complexity will require automation of the infrastructure provisioning and operation, allowing researchers to focus on problem solving.

This paper attempts to analyse current technology that can advance SDI development and support future digital science. Based on this analysis, the paper proposes a vision for the future RI Platform as a Service (PRIaaS) that incorporate recent digital technologies and enables platform and ecosystem model for future science.

The proposed analysis and PRIaaS architecture are based on the authors' long time involvement in numerous EU and national projects on RI development, studies, and initiatives, including current on-going projects GEANT4 [7], FAIRsFAIR [8], SLICES-DS [9] dealing with different of the modern Research Infrastructure and e-Infrastructure developments. The paper refers to the previous authors' works on defining the Big Data Architecture Framework (BDAF) [10], Scientific Data Infrastructure requirements [11] and developing practical aspects of the cloud services network infrastructure [12] that provide a strong foundation for current research.

The paper is organised as follows. Section II provides a short reference to recent regulations, initiatives, and projects in the European Research Area that drive future SDI and RI development. Section III provides an overview of the key technologies development that may facilitate FutureSDI development. Section IV describes the main features of the future digital science, analyses the timeline of the European

RI development and proposes a vision for the key technologies that can shape the FutureSDI marked as EOSC-2. Sections V summarises the general requirements to FutureSDI and describes the proposed PRIaaS architecture and its operation. Section VI discusses important aspects of the research data management to support FutureSDI requirements and PRIaaS functionality. Section VII presents a conclusion and refers to ongoing and future developments.

II. Transforming European Research Area

A. European Research Infrastructures and ESFRI Roadmap

European Research Area (ERA) is an important area of the European policy development and funding to support European science and ensure its competitiveness while facilitating European cooperation and integration. The Research Infrastructures (RI) is one of the pillars of ERA designated to connect research, higher education and innovation [13].

A European Research Infrastructure (RI) is a facility or (virtual) platform that provides the scientific community with resources and services to conduct top-level research in their respective fields. The research infrastructures can be single-sited or distributed or an e-infrastructure, and can be part of a national or international network of facilities, or of interconnected scientific instrument networks.

Important instruments in defining European in RI development and evolution is the ESFRI (European Strategy Forum on Research Infrastructures) Roadmap [14]. The new ESFRI Roadmap 2021 defines the important priorities set for the ESFRI Roadmap 2021 include consolidating the landscape of European RIs, opening, interconnecting and integrating RIs to develop the full potential of data generated by RIs and increase the innovation potential of ERA/European science in its cooperation with industry [15]. Research Infrastructures constitute a powerful resource for industry, a prerequisite for collaboration between industry and academia.

To facilitate RI and science digitalization, the new ESFRI Roadmap includes a new DIGIT area whose focus is to support research on digital technologies.

e-Infrastructure is another area of the European policy and funding that is designated to support ESFRI and constitute the essential building block for ERA. e-Infrastructures address the needs of European researchers for digital services in terms of networking, computing and data management. e-Infrastructures provide digital-based services and tools for data- and computing-intensive research in virtual and collaborative environments.

e-Infrastructures are key in the future development of research infrastructures, as activities go increasingly online and produce vast amounts of data. Current European e-Infrastructure capacity includes such Trans-European operational infrastructures as GÉANT – the high-capacity and high-performance communication network [16], and PRACE – European HPC services for European research [17].

B. European Open Science Cloud (EOSC)

The European Open Science Cloud (EOSC), started in 2016, is the part of the "European Cloud Initiative - Building a competitive data and knowledge economy in Europe" [18, 19] that is targeted to capitalise on the data revolution. Under this initiative, EOSC federates existing and emerging e-Infrastructures to provide European science, industry and public authorities with world-class data infrastructure connected to high performance computers (HPC).

The EOSC goal is to enable the Open Science Commons [20]. At the present time, the EOSC projects created the foundation for research data interoperability and integration for European IRs. The Minimum Viable EOSC (MVE) achieved by the end 2020, will create a starting point for future EOSC development [21].

MVE defines EOSC Core that is designed to provide a federated data exchange environment for research projects and communities where data comply FAIR principles. EOSC Core includes the following components/functionalities:

- Shared Open Science policy framework
- Authentication, Authorisation Interoperability Framework
- Data Access framework
- Service Management and Access framework
- A minimum legal metadata framework
- An open metrics framework
- PID framework and service
- Portal providing web access to the EOSC services.

The further EOSC development based on MVE (which we can refer to as EOSC-1) will require designing a new type of infrastructure that can benefit from existing and emerging digital and infrastructure technologies.

III. Technology driven Science Transformation

A. Science Digitalisation and Industry 4.0

Science digitalization is a demand of time and advised by the OECD report [1]. Science and industry digitalisation make easier exchange of technologies, solutions, application and also adopting recent industry trends such as Industry 4.0 [22] and platform based ecosystems.

Industry 4.0 will bring tremendous changes to both business models and the way future factories will operate. The key Industry 4.0 elements that both empower new data-economy and will be facilitated by the new business and consumer models include: Cyber-physical systems; Internet of Things; Internet of services; Smart factories; Mobile technologies and highspeed access networks; Cloud Computing and distributed data processing; Big Data; Artificial Intelligence and Machine Learning; Automation, Robotics and Digital Twins.

The digital nature of ongoing economy transformation opens opportunity for faster technologies and solutions exchange with science and research. Science can benefit from massive investments into industrial digital and data driven technologies that can be directly used in digital

science, in particular, experimental research automation and following data processing and management. The scientific community should follow the development and be open to wider use of technologies that are advanced by industry, actually all technologies powering Industry 4.0 can be effectively used both in the Future SDI and domain specific scientific applications.

B. *Trasformational role of Artificial Intelligence*

Similar to Industry 4.0, Artificial Intelligence will have a strong transformative effect on future science [23]. Benefits that AI can bring to scientific research and SDI include but not limited to:

- Extending possibilities of research when working with big data
- Automating data preparation, processing, and analysis
- Smart infrastructure and tools operation and management
- AI driven and Machine Learning powered scientific discovery and decision support, digital models creation (Digital Twins)
- AI powered self-learning assistant to a researcher/scientist capable of creating domain related intelligence; many research questions will be pursued semi-automatically [24]
- Role of data will change: the learned model will replace data; theory becomes data for next generation AI [24]

It is recognised that an effective work of AI and ML technologies is critically dependent on the quality of data and their availability at all stages of the AI lifecycle [25]. This will impose the specific requirement to the FutureSDI, including general compute and storage, distributed federated ML algorithms, edge computing and highspeed access network.

Consistent data management including FAIR compliance, quality assurance, data lineage and privacy protection are general preconditions for successful AI implementation [26].

C. *Promises of 5G technologies*

5G technologies promise to solve not only high-speed mobile communication for smart(phone) applications but also e2e land/terrestrial network communication. 5G architecture defines three main future use cases (or usage scenarios) that can be adopted by the FutureSDI [27]

- Enhanced Mobile Broadband (eMBB): this also covers IoT, robotics, sensor network
- Massive Machine Type Communications (mMTC) to support HPC and large scale distributed data processing
- Ultra Reliable and Low Latency Communications (URLLC): industry automation, process control, real time applications

To address these use cases and corresponding requirements, 5G architecture offers e2e network slicing technology that allows proving isolated virtual overlay networks using Network Functions Virtualisation (NFV) and cloud native services deployment model and mechanisms. In addition to slices isolation, the 5G architecture is also offering a consistent security model that enables Trusted Execution Environment (TEE) [28] for running secure and trusted services by using the hardware Root of Trust (which idea is originated from the Trusted Computing Platform architecture [29]).

D. *Adopting platform and ecosystems business model for future SDI*

The platform economy [30, 31] and digital ecosystems [32] are the two trends shaping ongoing transformation of the modern economy facilitated by digitalisation. The wide adoption of the platform business and operational model (as an alternative to the pipeline model) facilitates the creation of the value chain between producers and consumers when using (composable) platform services powered by extended data collection and availability from the platform providers. This allows creating consistent business oriented digital ecosystems as loose associations of stakeholders and capabilities instantiated on the platform provider facilities. An ecosystem has members that interact in the context of a defined set of services and offerings.

TeleManagement Forum (TMF) defines the Open Digital Architecture (ODA) [33] and the Digital Platform Reference Architecture (DPRA) [34], where the infrastructure provisioning component is defined as the Actualisation Platform which architecture is illustrated in Figure 1.

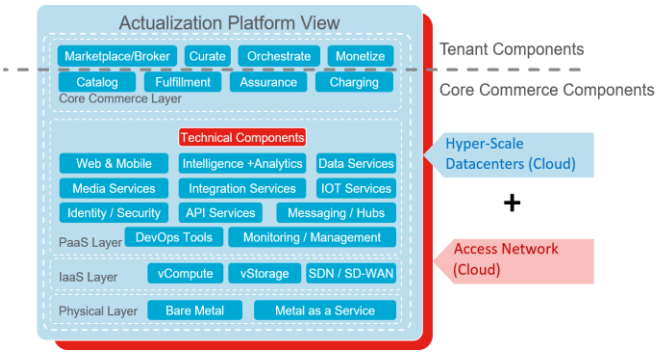

Figure 1. TMForum Actualisation Platform as a core part of DPRA [348].

The Actualisation Platform includes the following essential (group of) components:

- Common infrastructure and platform services
- Data and digital content services
- Catalog Lifecycle Management & Federation Platform
- Integration, orchestration, and DevOps
- Security and Identity Management
- Core commerce services including Fulfillment Platform Component and customer facing services

The Fulfillment Platform defined in DPRA "allows for user/service configuration and activation data to be sent for each individual component service, and also for fully composed product offers (of the customisable templates or design patterns). It allows a product creator to configure (fulfill) a service that is being composed into an e2e offer –

this could involve adding an end-user (authorization credentials, establishing an account), or any other actions required for configuration management" [34].

ODA and DPRA are adopted by many telecom providers, and we can benefit from adopting it for FutureSDI that could serve to create instant virtualized RI and ecosystems for specific user communities.

E. Other infrastructure technologies and trends

The following are recent technologies that can be adopted to build the Future SDI:

- Cloud based federated hyperconverged infrastructure allowing for provisioning on demand secure private infrastructure [35]
- IDSA architecture and IDS Trusted Connector enabled data exchange infrastructure [36, 37]
- Infrastructure automation technologies and tools (virtualization, microservices, composability, containerization, code libraries, API)
- DevOps and CI/CD that trends to become integrated into the change management process to ensure the continuous evolution of the target system [38]
- Data centric models DataOps/MLOps (which examples are services offered by Azure cloud platform) [39, 40]
- Semantic Data Lakes as integrated data storage and data analytics platform (which example is Azure Data Lake gen2 that offers storage for heterogeneous data and provide integrated data analytics) [41, 42]
- Infrastructure related security technologies that propose solutions for trust bootstrapping, data lineage, creating secure trusted virtual execution environment for data

IV. DEFINING FUTURE SCIENTIFIC DATA INFRASTRUCTURE

A. Paradigm change in modern Data Driven/Digital Science

Ongoing Science digitalization is powered by the rapid development of Cloud Computing, Big Data, Artificial Intelligence, and DevOps based infrastructure automation technologies.

The FutureSDI should consolidate existing and future RIs focusing on specific scientific domains, minimise costs and efforts of creating specialized RI for different scientific communities. Achieving MVE/EOSC-1 will create a platform for FAIR data interoperability and sharing, a key step in the future digital transformation of science.

Here we summarise the main characteristics of the (future) digital science powered by recent advancement in data driven technologies and AI (also refer to our previous analysis [10]):

- Availability of Pan-European Research Infrastructure Platform as a Service (later defined as PRIaaS) that uses cloud-native technologies (S/P/IaaS) for on-demand provisioning of the fully operational infrastructure for end-to-end scientific research (both experiments and data processing) by using composable infrastructure and applications design templates, supported by DevOps tools.
- Automation of scientific experiments and all data handling processes, including data collection, storing, classification, pre-processing and curation, provenance.
- Adopting and leveraging DevOps and DataOps/MLOps technologies found rapid adoption in the industry and supported with a variety of tools available with cloud-based infrastructure platforms such as AWS, Azure, Google Cloud Platform, and from multiple vendors.
- Digitising existing artifacts and creating their digital twins, AI assisted documenting and cataloging, building subject/domain knowledge base using self-learning algorithms.
- The full adoption of the FAIR data principles, both prospective and retrospective, to ensure re-usability of available data/datasets in the cross-domain and secondary research.
- Adopting STREAM data properties and corresponding infrastructure to enable trusted multipurpose data sharing and exchange, including data trading as economic goods and enabling different economic models for data sharing.
- Availability of new algorithms for distributed secure data processing such as federated machine learning, or blockchain enabled policy aware distributed data processing.
- Global data availability and access over the network for cooperative group of researchers, including wide public access to scientific data, however subject for the data sharing and access policies, in particular GDPR.
- Advanced security, access control and identity management technologies that ensure the secure operation of the complex research infrastructures and scientific instruments and allow creating a trusted secure environment for cooperating groups and individual researchers

The future SDI should support the whole data lifecycle and explore the benefit of data aggregation and provenance at a large scale and during a long/unlimited period of time.

Data security is not limited by a secure and trusted storage but also requires a secure and trusted data processing environment that would allow data processing using proprietary algorithms. Demand for RI trustworthiness and security is increasing to address both personal data protection and the trustworthiness of the research process itself. Data infrastructure must ensure data security (integrity, confidentiality, availability, and accountability), trustworthiness and, at the same time, data sovereignty that include both data ownership protection and control of data sharing and processing by data owners. There should be a possibility to enforce data/dataset policy (sharing, processing, derivative/secondary data) in the distributed data storage, sharing and processing environment.

SLICES

RI Type (evolution stage)	Centralised (1994-1996)	Interconnected (2004-2006)	Distributed (2011-2012)	Federated (2016-2018)	EOSC-1 (2020-2022)	EOSC-2 (future)
Definition	Institutions based, centralised facility	Multi-institutions, interconnected	Large distributed facilities, domain or experiment oriented	Federated RIs supporting inter-domain cooperation and data exchange	Interoperable (European) RI, FAIR RI	Virtualised Pan-European RI platform as a Service and ecosystem (PRIaaS)
Network & Compute	Mainframe, variety of protocols, Advent of Internet, web, email	Interconnected data centers and experimental facilities, Internet TCP/IP as common protocol, remote access	Distributed interconnected computing facilities, SOA and webservices, Grid as cooperative and distributed computing	Cloud adoption, infrastructure services on-demand Federated facilities and network access, Federated access and Identity management, 3G->4G	Distributed scalable computing, cloud based Big Data technologies, high performance networks, 5G technologies, wireless access, IoT sensor networks	Composable virtualized RI provisioning on demand, common federated computing and networking platform/environment, Cloud, DevOps and AI enabled, Digital Twins
Data	Proprietary formats, system or experiment specific	Standard format for data exchange, proprietary metadata	Domain/RI based data/metadata interoperability, custom data models, distributed storage, directories	Interoperable data, domain based metadata	FAIR data, Data Factories, Metadata registries, Interoperable/common Data Management model	Fully adopted FAIR principles, Semantically enabled scientific data lakes, secure/trusted data exchange, full data value chain
Infrastructure Management Technologies	Local management	Local management, management information exchange	Common Management Model, Distributed management, 3G Roaming	OSS/BSS, Automated deployment, adaptation, monitoring	Integrated Operation and Automation, Automated identity provisioning	Fully automated RI and services provisioning, management and operation, optimisation

Figure 2. Timeline RI evolution and SLICES positioning.

Table 1. Details of the technologies used in current EOSC-1 and future EOSC-2

RI Type	**EOSC-1 (2016-2021)**	**EOSC-2 (future 2022-2025+)**
Definition	**Interoperable Federated RI, FAIR RI**	**Virtualised Pan-European RI platform as a Service (PRIaaS)**
Network & Compute	• Distributed scalable computing • Cloud based Big Data technologies • High performance networks • 5G technology readiness • IoT sensor networks • Portal • Services Catalog and API repository) • Industry standards and IDSA adoption	• Composable virtualized RI provisioning on demand (including for services integration) • Common federated computing and networking platform/environment, enabling virtual RIs • Cloud based and cloud enabled • 5G adoption: wireless access network, e2e network slicing • DevOps and AI enabled services • Digital Twins • Interoperability and Integration with Industry infrastructure (e.g. IDSA+, Industrial Internet)
Data Infrastructure	• FAIR data management and exchange • Metadata registries • PID and Data Factories • Interoperable/common Data Management model	• Fully adopted FAIR principles, extended to ontologies • Semantically enabled scientific data lakes, common vocabularies Secure/trusted data exchange (data markets) • Full data value chain supported (cross-domain)
Security	• Federated Identity Management, Federated Access Control • Automated identity provisioning	• Federated Identity Management, Federated Access Control • Automated identity provisioning • Zero trust security, Trust Bootstrapping • Homomorphic encryption and data processing • Quantum ready encryption, Quantum enabled key management
Infrastructure Management Technology	• Integrated Operation and Automation • Automated identity provisioning	• Fully automated RI and services provisioning, management and operation • AI enabled Optimisation of infrastructure and operation • DevOps and re-usable design patterns

Legend: Evolution from EOSC-1 to EOSC-2 means that new advanced EOSC-2 services are built on the available and operational EOSC-1 services.

B. Timeline of the European RI development/evolution

In our research on the technologies for FutureSDI we analysed the development and evolution of the European RIs. Figure 2 below illustrates the timeline of the European RI evolution (based on the authors expertise and wide community discussions aligned with technology evolution and trends) that covers past stages: Centralised, Interconnected, Distributed, Federated, where the current stage is labeled as EOSC-I (actually implementing EOSC Core) and foreseeing future stage labeled as EOSC-II. Table 1 provides extended details about technologies that are suggested to drive the transition from EOSC-1 to EOSC-2.

Past stages (before EOSC) delivered Federated Research Infrastructures supporting inter-organisational and interdomain cooperation and data sharing using well defined metadata ensuring data interoperability, however in many cases limited to a science domain. Examples of such RIs are EGI, EUDAT, GEANT, PRACE and other landmark RIs as reviewed in the ESFRI Roadmap 2018 [43]. The European Open Science Cloud (EOSC) provides a basis for European RI integration and interoperability based on adoption of the FAIR principles both for data and for RIs themselves. H2020 EOSC-hub project established and operates EOSC Portal offering services Catalog and Marketplace that enables services and data findability, interoperability and re-usability based on published APIs [44].

Future progression and adoption of modern technologies such as Cloud and Edge Computing , Big Data, AI, IoT, and Digital Twins will enable fully virtualised Pan-European RI platform as a Services (PRIaaS) that will allow virtualized RI provisioning on demand for specific scientific domain and community; advanced data management and processing technologies will allow full FAIR principles implementation and trusted data exchange, supporting whole data lifecycle and value chain with the necessary infrastructure services. Adoption of the 5G technologies is expected to start a preparatory stage at the EOSC-I stage in some individual projects and testbeds and will become the main enabling technology for virtualizing/slicing network and RI in the future, combining with the Virtual Private Cloud (VPC) [45] technologies supported by modern cloud platforms.

The envisioned PRIaaS definition leverages the TMForum DPRA concepts and principles that define the provider actualization platform as a way to enable provisioning customer tailored services platform/ecosystem on demand.

Recently started the SLICES-DS project [9] intends to bridge the current EOSC-I stage and future EOSC-II stage by advancing infrastructure technologies to fully virtualized customised domain specific RI provisioning on-demand. Many modern advanced and emerging technologies need to be tested, adopted and prototyped to make them easily usable by different RIs and embedded into the PRIaaS platform (see section V for PRIaaS architecture).

C. General Requirements to Future Data driven Research Infrastructures

From the overview, we just gave we can extract the following general infrastructure requirements to SDI for emerging Big Data Science:

- Cloud based provisioned (on-demand) instant RI, fully functional including virtual user organisation – multi-cloud and hybrid
- On-demand infrastructure provisioning to support data sets and scientific workflows, mobility of data-centric scientific applications, with enabled automation and operation
- Support long running experiments and large data volumes generated at high speed
- Multi-tier inter-linked data distribution and replication
- Secure trusted data infrastructure, ensuring data sovereignty and trustworthiness, FAIR compliant and supporting STREAM properties for effective data exchange
- Support of virtual scientist communities, addressing dynamic user groups creation and management, federated identity management – to enable cooperation and data sharing
- Trusted environment for data storage and processing
- Support for data integrity, confidentiality, accountability, provenance, sovereignty
- Mechanisms for policy binding to data to protect privacy, confidentiality and IPR that ensure the policy is attached to data during the whole data lifecycle; mechanisms for policy provisioning and roaming as part of the provisioned infrastructure to ensure policy enforcement by design in a diverse heterogeneous environment.

V. PROPOSED PRIAAS ARCHITECTURE MODEL

We propose the PRIaaS Architecture for FutureSDI as illustrated in Figure 3. This model contains the three generalized layers:

Virtualised Resources (VR): Virtualised general compute, storage and network resources that are composed to create infrastructure components and are used by other services and applications.

Actualisation Platform: This is the main component and layer that enables provisioning, monitoring and operating fully functional instant Virtual RIs for specific scientific domains, projects, or communities.

Virtual (Private) RI (VirtRI): Virtual RI provisioned on demand that contains a full set of services, resources and policies needed to serve the target scientific community and create full value change of data handling. VirtRI is operated by the specific community and uses services provided by the Actualisation platform, including the possibility of cross-platform data sharing.

Users and external resources include researchers, developers and operators, and external datasets.

Federation Access Infrastructure and Tenants Management FAI&TM) layer serves as interface layer enabling communication between distributed Actualisation Platform resources and services and generally distributed and

multiorganizational VirtRI. FAI&TM is also the place where VirtRI and Actualisation Platform policy are enforced and managed.

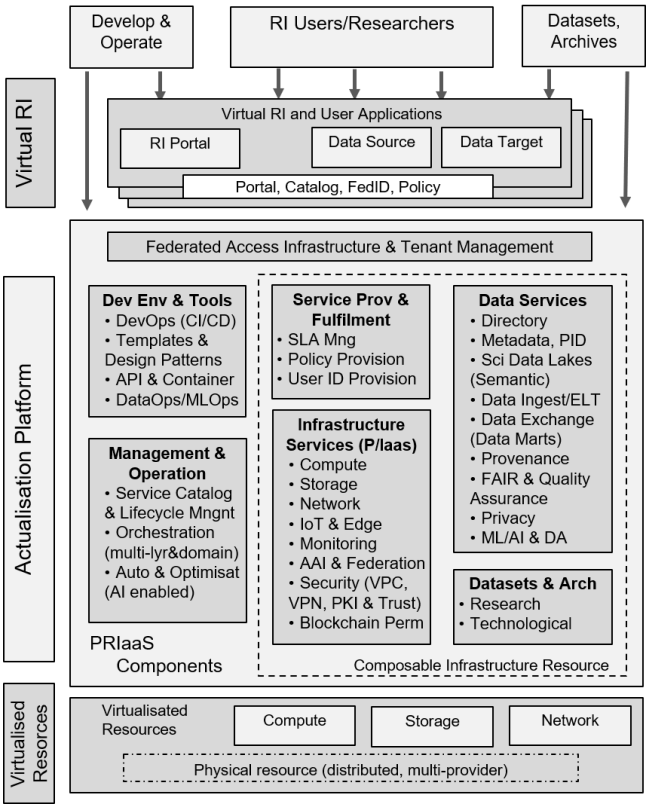

Figure 3. The proposed PRIaaS architecture

A. *Actualisation Platform Components*

Actualisation Platform includes the following groups of services required to develop, deploy, manage and operate the Virtual RI during its whole lifecycle, including resources and users that can be grouped into Virtual Organisations.

- Core Infrastructure Services (IaaS & PaaS) including compute, storage, network, IoT&Edge, blockchain, Access Control and Federated Identity management, infrastructure security
- Data Services including directory, metadata/PID, lineage/ provenance, FAIR &QA, semantic data lakes, data analytics and AI tools
- Management and Operation including Service Catalog and Lifecycle Management, orchestration and management
- Service provisioning and Fulfilment including user provisioning, SLA management and policy provisioning.
- Development Environment and Tools that support DevOps process related to platform and VirtRI development, provisioning and operation; this group also maintains the repository of API, containers and design templates that can facilitate VirtRI design and provisioning.

VirtRI provisioning process is based on well-known and commonly used DevOps tools and is supported by the Management and Operation functions. As the PRIaaS platform will progress, the repository of the design patterns, templates and containerized applications and functions will grow. A starting point for such a repository can be the EOSC Catalog [44] that already contains information about API for applications and services offered by existing RIs and service providers.

The policy provisioning, management and enforcement are important functions of the Actualisation Platform that can be attributed to the Fulfilment function. The policy that is defined by the target community is provisioned as a part of VirtRI provisioning. Policy management and enforcement infrastructure should support policy roaming and combination for the multi-domain distributed resources and tenants.

VI. Research Data Management in Future SDI

A. *European wide and international initiative and projects*

The importance of data and research information sharing has been central in a number of European wide initiatives and projects, such as Open Access Open Data, Open Science, Open Commons. The Research Data Alliance (RDA) that was created in 2012 jointly by the National Science Foundation of USA (NSF) and European Commission, became a key community coordination body to exchange and develop best practices in research data management. One of the important RDA developments became Persistent Identifiers (PID for data objects to enable data interoperability and findability. [46].

To facilitate research data sharing and implementation of the FAIR principles, European Commission started Open Research Data (ORD) Pilot [47] and currently all EU funded projects are required to develop and implement the Data Management Plan (DMP) at the initial stage of the project. Data produced in the project must be stored in the open available but secure repositories (operated by the project or using national or European data archive services. Metadata must be published and quality of data ensured, in particular, compliance with the FAIR principles

B. *From FAIR data principles to STREAM data properties*

FAIR data principles are important for creating trusted research-friendly environment for data sharing. FAIR data is a key element/layer of the EOSC core. However, the data exchange infrastructure requires additional data properties that would allow trusted and economical data exchange, also supporting data value chain creation.

Data exchange and data trading/market have been long time interest area by/from the industry where data represent also companies' intellectual property and companies want to remain in control of their data what is defined as data sovereignty.

Data Sovereignty is a key principle of the industrial data exchange as defined by the International Data Spaces Association (IDSA) Reference Architecture Model (RAM) [36].

Data involved in industrial processes and business relations are becoming a part of the economic relations and added value creation process. However, data as economic

goods are in many aspects are different from the traditional economic goods and commodities. We refer to our research on data properties as economic goods [48] as part of the RDA Interest Group on Data Economics (IG-DE) [49].

Emerging data driven economy and modern Big Data technologies facilitate interest in making data a new economic value (data commoditisation) and consequently the identification of new properties of data as economic goods. The STREAM data properties for industrial and business data have been proposed by the authors in [48]. To become an economic goods and bring business value to data producers and data consumers data must be: [**S**] Sovereign, [**T**] Trusted, [**R**] Reusable, [**E**] Exchangeable, [**A**] Actionable, [**M**] Measurable

Other data properties important to enabling data commoditisation and allow data trading and exchange for goods include: Quality, Value, Auditability/Trackability, Branding, Authenticity, as well as original FAI(R) properties Findability, Accessibility, Interoperability, Reusability. Special features that must be managed in all data transfers and transformations are data ownership, IPR and privacy. The data property originated from its digital form of existence defined as not-Rivalry, on one hand, makes data exchange (copying, distribution) easy, but on the other hand, it creates a problem when protecting proprietary, private or sensitive data or IPR.

VII. Future Research and Development

In this paper, we presented analysis of current trends in digital technologies that can use to build Future Scientific Data Infrastructure, and in particular can be used to progress the current EOSC infrastructure, also proposing a common platform for future European RI integration. Further research will require a closer analysis of the typical use cases in ESFRI and EOSC projects. The presented research and proposed PRIaaS are based on the authors long time experience in infrastructure research and developing/implementing practical solutions in a number of national EU funded projects such as EGEE, GEANT, GEYSERS, as well as standardisation activity in such bodies as IETF, OGF, NIST, CEN.

The proposed PRIaaS architecture and DPRA inspired operational model require a variety of technologies to work together realizing data centric data exchange and transformation to enable data based applications and services and added value data service creation. New functionality and technology combinations will require re-thinking existing concepts and models, extending usage scenarios.

Further development of the proposed PRIaaS and its components will be done in the ongoing project SLICES-DS. This work also intends to contribute to the EOSC Architecture Working Group.

Acknowledgment

This work is supported by EU funded projects SLICES-DS (Grant Agreement No. 951850), FAIRsFAIR (Grant Agreement No. 831558), GN4-3 (Grant Agreement No. 856726). The authors value wide discussions in the RDA and EOSC forums on the different aspects of the existing research infrastructures, data management and ongoing research and developments.

References

[1] The Digitalisation of Science, Technology and Innovation: Key Developments and Policies, OECD Publishing, Paris, 2020 [online] https://doi.org/10.1787/b9e4a2c0-en.

[2] Measuring the Digital Transformation: A Roadmap for the Future, OECD Publishing, Paris. [online] https://doi.org/10.1787/9789264311992-en

[3] A European strategy for data, EC COM(2020) 66 final, Brussels, 19.2.2020 [online] https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf

[4] European Open Science Cloud (EOSC) [online] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

[5] Research Data Alliance [online] https://rd-alliance.org/

[6] Turning FAIR into reality. Final Report and Action Plan from the European Commission Expert Group on FAIR Data, 2018 [online] https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf

[7] GÉANT Project GN4-3, Accelerating research, driving innovation and enriching education. [online] https://www.geant.org/Projects/GEANT_Project_GN4-3

[8] FAIRsFAIR Project: Fostering FAIR data practices in Europe [online] https://www.fairsfair.eu/

[9] SLICES-DS Project [online] http://slices-ri.eu/

[10] Defining architecture components of the Big Data Ecosystem Y Demchenko, C De Laat, P Membrey 2014 International Conference on Collaboration Technologies and Systems (CTS 2014), May 19-23, 2014, Minneapolis, USA

[11] Addressing big data issues in scientific data infrastructure Y Demchenko, P Grosso, C De Laat, P Membrey 2013 International Conference on Collaboration Technologies and Systems (CTS 2013), May 20-24, 2013, San Diego, California, USA. ISBN: 978-1-4673-6402-7

[12] Open cloud exchange (OCX): Architecture and functional components
Y Demchenko, J Van Der Ham, C Ngo, T Matselyukh, S Filiposka, ...
2013 IEEE 5th International Conference on Cloud Computing Technology and Science (CloudCom2013), 2-5 Dec 2013, Bristol, UK

[13] Euroepan Research Area (ERA) [online] https://ec.europa.eu/info/research-and-innovation/strategy/era_en

[14] Landscape of Research Infrastructures and evolution of the ESFRI Roadmap. ESFRI Roadmap 2021 InfoDay 25 Sept 2019 [online] https://www.esfri.eu/esfri-events/roadmap-2021-infoday?qt-event=5#qt-event

[15] Making Science Happen: A new ambition for Research Infrastructures in the European Research Area. ESFRI Whitepaper, March 2020 [oline] https://www.esfri.eu/esfri-white-paper

[16] GEANT - https://www.geant.org/

[17] PRACE – Partnership for Advanced Comp in Europe - https://prace-ri.eu/

[18] "European Cloud Initiative - Building a competitive data and knowledge economy in Europe" [online] https://www.eesc.europa.eu/en/our-work/opinions-information-reports/opinions/european-cloud-initiative-building-competitive-data-and-knowledge-economy-europe

[19] Building the European Open Science Cloud, By Drago, Federico; Ferguson, Nicholas, 19 March 2020 https://zenodo.org/record/3716192#.X1aYzXkzZPY].

[20] Tiziana Ferrari, Diego Scardaci, Sergio Andreozzi, The Open Science Commons for the European Research Area, Part of the ISSI Scientific Report Series book series (ISSI, volume 15) [online] https://link.springer.com/chapter/10.1007/978-3-319-65633-5_3

[21] Solutions for a Sustainable EOSC. A tinman report from the EOSC Sustainability Working Group, Draft 2 December 2019, https://www.eosc-nordic.eu/content/uploads/2020/03/Tinman_draft_19_compressed.pdf

[22] Industry 4.0: the fourth industrial revolution – guide to Industrie 4.0 [online] https://www.i-scoop.eu/industry-4-0/

[23] AI for Science: Report on the Department of Energy (DOE) Town Halls on Artificial Intelligence (AI) for Science, July-September 2019 [online] https://www.anl.gov/ai-for-science-report

[24] AI for Science, by Barbara Helland, AI for Science Town Hall, Oct 2019 [online] https://science.osti.gov/-/media/ber/berac/pdf/201910/Helland_BERAC_Oct2019.pdf

[25] AI Development life cycle: explained [online] https://www.devteam.space/blog/ai-development-life-cycle-explained/

[26] Data quality in the era of Artificial Intelligence, blog by George Krasadakis [online] https://medium.com/innovation-machine/data-quality-in-the-era-of-a-i-d8e398a91bef

[27] 5G and The Cloud, 5G Americas White Paper, December 2019 [online] https://www.5gamericas.org/5g-and-the-cloud/

[28] The Evolution of Security in 5G- 5G Americas White Paper, 5G Americas, July 2019 [online] https://www.5gamericas.org/wp-content/ uploads/2019/08/5G-Security-White-Paper-07-26-19-FINAL.pdf

[29] INSPIRE-5Gplus Deliverable D2.1: 5G Security: Current Status and Future Trends, 2020 INSPIRE-5Gplus Consortium Parties [online] https://www.inspire-5gplus.eu/wp-content/uploads/2020/05/i5-d2.1_5g-security-current-status-and-future-trends_v1.0.pdf?x51934

[30] Parker, Geoffrey G.. Platform Revolution: How Networked Markets Are Transforming the Economy and How to Make Them Work for You. W. W. Norton & Company, March 2016.

[31] Key Enablers for a Hybrid Infrastructure Platform, TMForum, 15-18 May 2017, Nice, France [online] https://dtw.tmforum.org/wp-content/uploads/2017/05/12.-Stephen-Fratini-Milind-Bhagwat-Takayuki-Nakamura.pdf

[32] In the Ecosystem Economy, What's Your Strategy? By Michael G. Jacobides, Harvard Business Report, Issue Sept–Oct 2019 [onlie] https://hbr.org/2019/09/in-the-ecosystem-economy-whats-your-strategy

[33] IG1167 TM Forum Exploratory Report ODA Functional Architecture, 31 Jan 2020 [online] https://www.tmforum.org/resources/exploratory-report/ig1167-oda-functional-architecture-v5-0/

[34] IG1157 Digital Platform Reference Architecture Concepts and Principles v5.0.1, 21 July 2020 [online] https://www.tmforum.org/resources/reference/ig1157-digital-platform-reference-architecture-concepts-and-principles-v5-0-0/

[35] 11 main benefits of hyper-converged infrastructure, August 2020 [online] https://searchconvergedinfrastructure.techtarget.com/tip/11-main-benefits-of-hyper-converged-infrastructure

[36] IDSA Reference Architecture Model (RAM3.0), Version 3.0, April 2019 [online] https://www.internationaldataspaces.org/wp-content/uploads/2019/03/IDS-Reference-Architecture-Model-3.0.pdf

[37] Trusted Connector, Industrial Data Space, Draft DIN Spec 27070 [online] https://industrial-data-space.github.io/trusted-connector-documentation/

[38] DevOps & Change Management In The Enterprise World [online] https://clearbridgemobile.com/devops-change-management-in-the-enterprise-world/

[39] MLOps: Methods and Tools of DevOps for Machine Learning, 23 Jul, 2020 [online] https://www.altexsoft.com/blog/mlops-methods-tools/

[40] Design a machine learning operations (MLOps) framework to upscale an Azure Machine Learning lifecycle, Microsoft Azure [online] https://docs.microsoft.com/bs-latn-ba/azure/architecture/example-scenario/mlops/mlops-technical-paper

[41] Ten Reasons to Dive into the Smart (Semantic) Data LakeCambridge Semantics, 2017 [online] https://blog.cambridgesemantics.com/ten-reasons-to-dive-into-the-smart-semantic-data-lake

[42] Building your Data Lake on Azure Data Lake Storage gen2, Blog by Nicholas Hurt, March 2020 [onine] https://medium.com/@Nicholas_Hurt/building-your-data-lake-on-adls-gen2-3f196fc6b430

[43] ESFRI Roadmap 2018 [online] http://roadmap2018.esfri.eu/media/1066/esfri-roadmap-2018.pdf

[44] EOSC Catalog [online] https://catalogue.eosc-portal.eu/

[45] Virtual private Cloud [online] https://en.wikipedia.org/wiki/Virtual_private_cloud

[46] Persistent Identifiers, Groups of European Data Experts, 27 Nov 2017, RDA [online] https://www.rd-alliance.org/system/files/PID-report_v6.1_2017-12-13_final.pdf

[47] Data management, Extension of the Open Research Data Pilot in Horizon 2020, Horizon2020 Manual [online] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

[48] Demchenko, Yuri, Wouter Los, Cees de Laat, Data as Economic Goods: Definitions, Properties, Challenges, Enabling Technologies for Future Data Markets, ITU Journal: ICT Discoveries, Special Issue "Data for Goods", December 2018

[49] RDA Interest Group on Data Economics (IG-DE) [online] https://www.rd-alliance.org/groups/data-economics-ig