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

Dr. Yuri Demchenko
University of Amsterdam

BDAA2020 Symposium, HPCS2020
February 2021 (Virtual)

Outline

- Background for this research
  - European Research Infrastructure
  - European Open Science Cloud (EOSC)
- Research Infrastructure development and enabling technologies
  - Digitalisation, AI and 5G technologies
  - Leveraging Platform concept for RI platform model
  - TMForum Digital Platform Reference Architecture (DPRA)
- Proposed Platform RI as a Service Architecture (PRIaaS)
- SLICES-DS Project and PRIaaS implementation
- Discussion

Background for this research

- Project
  - EGEE – Enabling Grids for E-science
  - EGI – European Grid Initiative
  - ENVRI - Environmental Research Infrastructure
  - EOSC FAIRsFAIR
  - SLICES-DS
- Standardisation and Best Practices
  - NIST Big Data Architecture
  - RDA – Research Data Alliance, support Research Data Management best practices

European Research Area and Initiatives

- European Research Area (ERA) is an important area of the European policy development and funding
- Research Infrastructure (RI) is one of pillar in support of European science
  - Coordinated by ESFRI (European Strategy Forum on Research Infrastructures)
    - ESFRI Roadmap is published bi-annually to start new call for RI process, since 2006
  - More than 54 European RIs are listed in the EU HLEG study “Supporting the Transformative Impact of Research Infrastructures on European Research”, 2020
    - 34 distributed and 9 single sited serving EU research community
  - More than 1800 RIs operating in Europe
- RIs are supported by e-Infrastructure programme funded by EU Horizon 2020 and next Horizon Europe
  - Providing common integration platform for individual RIs and other research and industry domains
- ERA and ESFRI supports international cooperation and research
- European Open Science Cloud (EOSC) initiatives to create a common platform for European RI integration

ESFRI Research Infrastructure Domains

- Social and cultural Innovation
- Environment
- Health and Food
- Energy
- Physical Science and Engineering
- New domain included: DIGIT – Digital and Information technologies
  - To integrate modern digital and data technologies
- Former e-Infrastructure projects: GEANT, EGEE, EGI

- ESFRI Roadmap 2018 - http://roadmap2018.esfri.eu/media/1066/esfri-roadmap-2018.pdf

European Open Science Cloud (EOSC)

- EOSC is an overarching concept and framework to integrate existing RIs and facilitate information and data exchange between RI, organisations and researchers
  - First phase 2016-2020 with funded projects 2016-2022
    - 53 projects in total
- EOSC main projects and co-creation activities
  - EOSCpilot – Initial EOSC architecture and requirements
  - EOSChub – Technical integration platform, RI marketplace and API/services directory
  - EOSCsecretariat https://www.eoscsecretariat.eu/
    - Establishment of the Governance structure and EU EOSC association to be co-funded by EC and Member States (MS) Co-creation model and European Open Science Commons
- Built on experience of the past successful initiatives and project
  - EGEE and WLCG, EGI, RDA (co-founded by EC and NSF), GEANT/TERENA
- Provides a model experience for future EU initiatives, such as GAIA-X European Federated Data Cloud

EOSC Status: Minimum Viable EOSC and EOSC Core

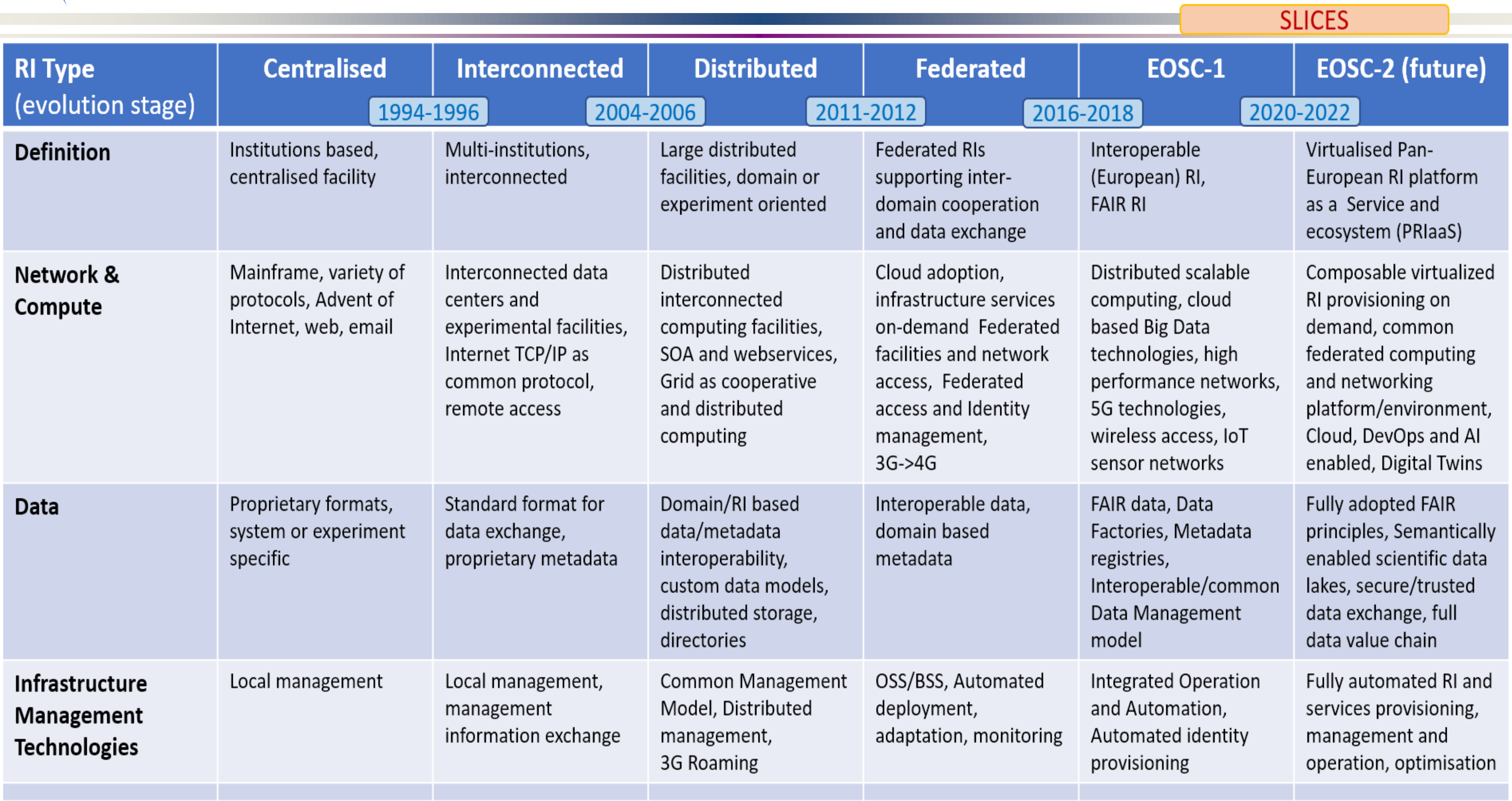
- Architecture defining (infrastructure) components
  - Data Access framework
  - PID framework and service
  - FAIR data enabling services
  - Service Management and Access framework
  - Authentication, Authorisation Interoperability Framework

- Policy and Governance
  - Shared Open Science policy framework
  - A minimum legal metadata framework as part of the FAIR compliance framework
  - An open metrics framework

Demand for modern RI platform using recent development by industry (for future technology exchange)
EOSC challenges: to incorporate recent technology development into ERA

- Portal providing web access to the EOSC services
- EOSC is an important stage in the European RI integration

Timeline European RI evolution



SLICES

RI Type (evolution stage)	Centralised	Interconnected	Distributed	Federated	EOSC-1	EOSC-2 (future)
	1994-1996	2004-2006	2011-2012	2016-2018	2020-2022	
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

- Based on the authors first hand experience
- SLICES-RI project positioning and intended contribution

From EOSC-1 to EOSC-2: Four Technology Aspects

2016-2018 2020-2022

RI Type	EOSC-1	EOSC-2 (future)	Beyond 2025
Definition	Interoperable Federated (European) RI, FAIR RI	Virtualised Pan-European **RI platform as a Service and Ecosystem (PRIaaS)**	
Network & Compute	• Distributed scalable computing • Cloud based Big Data technologies • High performance networks • 5G technologies, wireless access • IoT sensor networks • **Portal and Services Catalog** • 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 • DevOps and AI enabled services • Digital Twins • Interoperability and Integration with Industry infrastructure (e.g. IDSA+, Industrial Internet)	
Data Infra	• FAIR data • Data Factories and PID • Metadata registries • 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	• Zero trust security, Trust Bootstrapping • Homomorphic encryption and data processing • Quantum ready encryption, Quantum enabled key management • Federated Identity Management, Federated Access Control • Automated identity provisioning	
Infra Managnt Technolog	• Integrated Operation and Automation	• Fully automated RI and services provisioning, management and operation • Optimisation of infrastructure and operation • DevOps and AI enabled (re-usable design patterns)	

Technology Trends for Science Transformation

- Science and industry digitalisation make **easy exchange of technologies**, solutions, application
- Hyperconverged Infrastructure model
  - Cloud based infrastructure integrations
  - Benefitting from global cloud infrastructure and cloud native technologies
- Transformation effect of Artificial Intelligence and Machine Learning technologies
- Data Management and Governance as an important asset
- 5G and Telecom cloud
  - Infrastructure deployment and operation using cloud native technologies

Transformational Role of Artificial Intelligence

- Extending possibilities of research when working with Big Data
- Automating data preparation, processing, and analysis
  - AI enabled data management
- 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
- Role of data will change: the learned model will replace data; theory becomes data for next generation AI
  - – AI require reliable data infrastructure, with new metadata model

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

Data Management and FAIR Data Principles

- Importance of proper data management and data quality
  - Re-usable research data
  - Reference data sets for ML and AI algorithms
    - Explainable AI
  - Digital Twins in industry and science
    - Depend on data quality and proper data management flow
- FAIR data principles
  - Findable – Accessible – Interoperable - Reusable
  - One of major coordination area and co-creation in EOSC
  - GO FAIR Initiative - https://www.go-fair.org/
    - GO FAIR Implementation network
    - Internet of FAIR Data and Services

Promises of 5G and Value for RI&Science

- 5G main use cases (or usage scenarios) that can be adopted by the FutureSDI
  - 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
- 5G architecture solutions
  - e2e network slicing technology providing isolated virtual overlay networks using Network Functions Virtualisation (NFV) and cloud native services deployment model and mechanisms
    - Extended to 5G Radio Access Network (RAN) for sensor networks
  - Consistent security model that enables Trusted Execution Environment (TEE)

Leveraging Platform Concept for RI

- There is a need for defining and building new type of infrastructure for EOSC projects
  - Current EOSC pilot projects successfully demonstrated inter-/multi-domain data integration
  - However, each pilot project built own underlying infrastructure
- Future EOSC infrastructure should provide functionality
  - (1) automate deploying specialized RIs with focus on scientific data integration
  - (2) create a repository of infrastructure/services design patterns and common templates
  - (3) facilitate cooperative/business relations between partners
  - (4) apply governance and compliance policies by-design
- Learn from and leverage best industry practice and infrastructure development trends
  - Hyperconverged Infrastructure and 5G e2e network slicing
  - Industry developed platform and cloud native models
  - DevOps and DevSecOps

Platform Concept and Platform Economy

- "The platform business model enables interactions between producers and consumers of value. It achieves this goal through two mechanisms**. First, a platform provides a plug and play infrastructure** which encourages open participation by an external ecosystem of producers and consumers. **Second, it lays out the rules of governance for the interactions** that ensue."
  - Source: Sangeet Paul Choudary , http://platformthinkinglabs.com/start here/
- "A platform is a business based on **enabling value creating interactions between external producers and consumers**. The platform provides an open, participative infrastructure for these interactions and sets governance conditions for them. The platform's overarching purpose: to consummate matches among users and facilitate the exchange of goods, services, or social currency, thereby enabling value creation for all participants
  - Source: Choudary , Sangeet Paul; Van Alstyne , Marshall W.; Parker, Geoffrey G.; Platform Revolution: How Networked Markets Are Transforming the Economy and How to Make Them Work for You
- Example platform based businesses: Airbnb, Alibaba, Amazon, Azure (Microsoft), eBay, Facebook, Instagram, KAYAK, Pinterest, YouTube, Twitter, Wikipedia, Uber, Upwork
- This is in contrast to pipeline based businesses with linear value chains

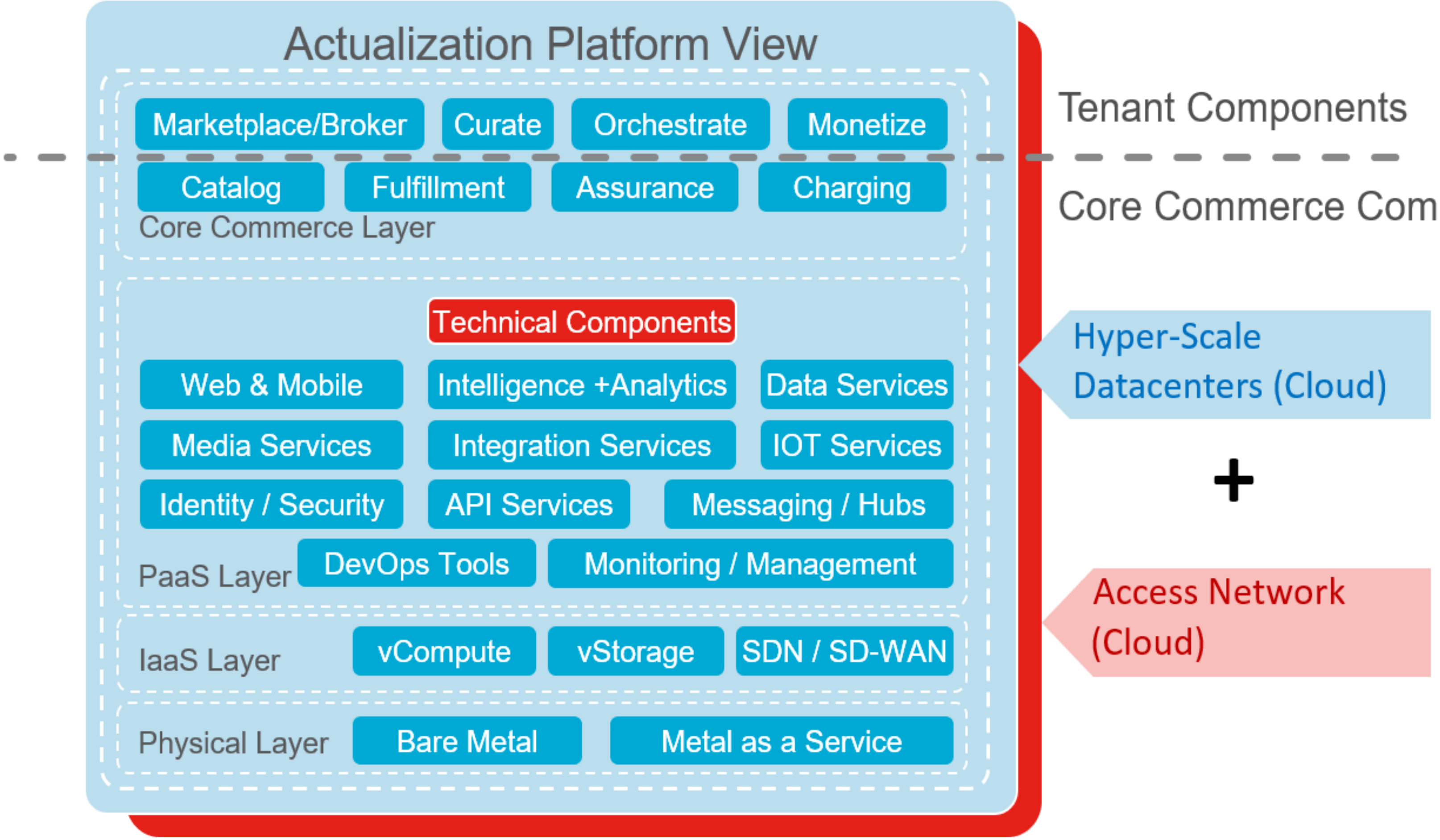
Platform vs Pipeline Model

- Source: Choudary , Sangeet Paul; Van Alstyne , Marshall W.; Parker, Geoffrey G.; Platform Revolution: How Networked Markets Are Transforming the Economy and How to Make Them Work for You
  - “Platforms beat pipelines because platforms scale more efficiently by eliminating gatekeepers.” Example: Coursera vs. a college / university
  - “Platforms beat pipelines because platforms unlock new sources of value creation and supply.” Example: Airbnb vs. an hotel chain
  - “Platforms beat pipelines by using data based tools to create community feedback loops.”
    - For virtualized functions, objective metrics could also be used, in addition to subjective consumer feedback.
  - “Platforms invert the firm”
- Pipeline Model uses linear value chains for a vertical business stack

Adopting TMForum DPRA

- TMForum DPRA (Digital Platform Reference Architecture) defines a telecom services provider platform that allows delivering a fully functional service platform/infrastructure for customers
  - 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/
  - Actualisation Platform is defined as the main DPRA component that enables creating customer/tenant service ecosystem
  - Implements platform economy concept
- Part of the TMForum Open Digital Architecture (ODA)
  - 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/

TMForum: Actualisation Platform View



- TM Forum Actualization Platform View The underlying infrastructure of edge to hyper scale datacenters and networks that host the software components that make up the Business Platform, enabled by reusable technical capabilities that are required to operate in an agile and efficient manner.

TMF DPRA Actualisation Platform Elements

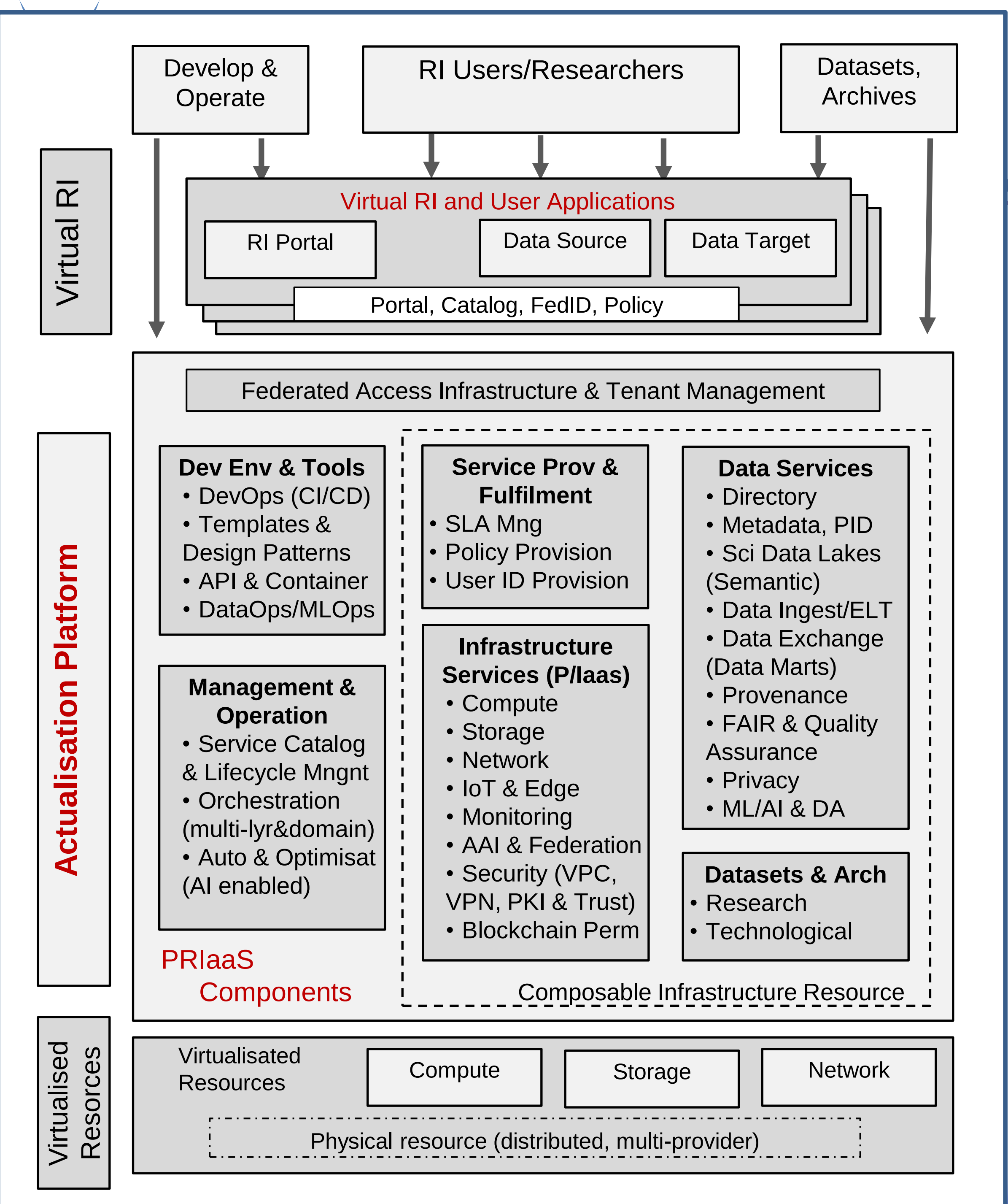
The Actualisation Platform includes the following essential (group of) components:

- 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

Paradigm Change in Modern Data Driven Science

- **Availability** of Pan-European Research Infrastructure Platform as a Service
- **Automation of scientific experiments** and all data handling processes
  - Adopting and leveraging **DevOps and DataOps/MLOps** technologies and using cloud automation tools
- **Digitising existing artifacts** and creating their digital twins, AI assisted documenting and cataloging, building subject/domain knowledge base using self-learning algorithms.
- Adoption of **FAIR data principles**, both prospective and retrospective
- Support **data value creation** model and flow (e.g. STREAM data properties)
- Availability of new algorithms for **distributed secure data processing** (e.g. federated machine learning, blockchain enabled policy enforcement)
  - **Enclave computing** as a new service by cloud providers
- **Global data availability and access** for cooperative group of researchers, however subject for the data sharing and access policies, in particular GDPR.
- Advanced security, access control and identity management technologies

PRIaaS Architecture Model (in progress)



Develop & Operate

RI Users/Researchers

Datasets, Archives

Virtual RI

Virtual RI and User Applications

RI Portal

Data Source

Data Target

Portal, Catalog, FedID, Policy

Actualisation Platform

Federated Access Infrastructure & Tenant Management

Dev Env & Tools
- DevOps (CI/CD)
- Templates & Design Patterns
- API & Container
- DataOps/MLOps

Management & Operation
- Service Catalog & Lifecycle Mngnt
- Orchestration (multi-lyr&domain)
- Auto & Optimisat (AI enabled)

PRIaaS Components

Service Prov & Fulfilment
- SLA Mng
- Policy Provision
- User ID Provision

Infrastructure Services (P/Iaas)
- Compute
- Storage
- Network
- IoT & Edge
- Monitoring
- AAI & Federation
- Security (VPC, VPN, PKI & Trust)
- Blockchain Perm

Data Services
- Directory
- Metadata, PID
- Sci Data Lakes (Semantic)
- Data Ingest/ELT
- Data Exchange (Data Marts)
- Provenance
- FAIR & Quality Assurance
- Privacy
- ML/AI & DA

Datasets & Arch
- Research
- Technological

Composable Infrastructure Resource

Virtualised Resorces

Virtualisated Resources

Compute

Storage

Network

Physical resource (distributed, multi-provider)

Actualisation Platform Components

- Core Infrastructure Services (IaaS & PaaS)
- Data Services
- Management and Operation
- Development Environment and Tools
  - DevOps
  - Templates and Patterns
- Service Provisioning and Fulfilment
- Datasets and Archives
- Federated Access Infrastructure + IoT Edge and Tenants Management
- Virtual RIs and Portal

From EOSC-1 to EOSC-2: Four Technology Aspects

2016-2018

2020-2022

RI Type	EOSC-1	EOSC-2 (future)	Beyond 2025
Definition	Interoperable Federated (European) RI, FAIR RI	Virtualised Pan-European **RI platform as a Service and Ecosystem (PRIaaS)**	
Network & Compute	• Distributed scalable computing • Cloud based Big Data technologies • High performance networks • 5G technologies, wireless access • IoT sensor networks • **Portal and Services Catalog** • 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 • DevOps and AI enabled services • Digital Twins • Interoperability and Integration with Industry infrastructure (e.g. IDSA+, Industrial Internet)	
Data Infra	• FAIR data • Data Factories and PID • Metadata registries • 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	• Zero trust security, Trust Bootstrapping • Homomorphic encryption and data processing • Quantum ready encryption, Quantum enabled key management • Federated Identity Management, Federated Access Control • Automated identity provisioning	
Infra Managnt Technolog	• Integrated Operation and Automation	• Fully automated RI and services provisioning, management and operation • Optimisation of infrastructure and operation • DevOps and AI enabled (re-usable design patterns)	

PRIaaS Technology Development in Ongoing and Past Projects

- GEANT GN4-3 Research WPs and Tasks
  - Operation, Automation, Virtualisation Architecture (OAV) and Service Provider Architecture (SPA)
    - Adopts/recommends TMF ODA for GEANT OAV
- EOSC - Minimum Viable EOSC (MVE)
  - EOSC Architecture and FAIR data principles
  - Apparently, vision for EOSC-2
- SLICES-DS – Future RI technology research
- Past projects
  - GN4-1 Open Cloud eXchange (OCX) and Big Data Architecture Framework (BDAF)
  - GN4-2 ZeroTouch Provisioning, Operation and Management (ZTPOM)
  - GN3 Composable Services Architecture (CSA)
  - GEYSERS Intercloud Architecture Framework (ICAF)

SLICES-RI and SLICES-DS Projects

- European Project involving 23 partners
  - SLICES-DS Design stage – 10 partners
- SLICES – RI is under evaluation for a new ESFRI DIGIT domain
- SLICES-RI Research Areas
  - Advanced wireless networking
    - 5G and radio access networks
    - Integrated sensing and communication
  - Smart/intelligent infrastructure operation and management
    - Distribution of intelligence into the Edge and beyond the Edge of the network
  - Design and validation of new Edge/Fog and hyper-converged infrastructures
    - Distributed resource management & microservices
    - Federated deep-learning
  - Advanced functionalities
    - Composable infrastructure services on-demand (RI as a Service)
    - Security and privacy

SLICES-RI Research Topics: UvA and NL partners

- Cloud and Network Infrastructure research
  - Architecture and design patterns of the future RI Platform as a Service (PRIaaS)
  - Federated multi-cloud and inter-cloud infrastructure integration and management
  - Decentralised network/compute optimisation in edge/fog environments
  - Sustainable cloud services with energy consumption monitoring and optimization
- Data Infrastructure
  - Big Data Infrastructure and Technologies (cloud enabled)
  - Trusted data exchange and processing with policy/rules enforcement, preserving data sovereignty and protecting data privacy
  - Data management and quality assurance aspects in Industry 4.0 experimentation and Digital Twins applications
- New security and compliance models for Complex Cyber Infrastructure
  - Distributed Cyber Security techniques and architectures
- Federated Data Analytics and Deep Learning, in particular for predictive maintenance, logistics and smart cities
- Support of education on key technologies of the future data centric and cloud enabled infrastructures by provisioning educational platforms and resources for universities on demand

Discussion and Questions

- Presented PRIaaS Architecture and technology overview as a Request for Comments and call for professional community contribution
- Consistent framework for planning and aligning future research areas
- PRIaaS platform provider model and federated model for multidomain research

Research Data in Europe: FAIR, EOSC, ORD Pilot and DMP

- EU policy on Open Research Data
- Horizon 2020 Open Research Data (ORD) Pilot
- Horizon 2020 Data Management and Data Management Plan

Requirements to Future SDI

- R01: Cloud based provisioned (on-demand) instant RI, fully functional including virtual user organisation – multi-cloud and hybrid
- RI02: On-demand infrastructure provisioning to support data sets and scientific workflows, mobility of data-centric scientific applications
- RI03: Trusted environment for data storage and processing
- RI04: Mechanisms for policy binding to data to protect privacy, confidentiality and IPR
- RD01: Multi-tier inter-linked data distribution and replication
- RD02: Secure trusted data infrastructure, ensuring data sovereignty and trustworthiness
  - FAIR compliant and supporting STREAM properties for effective data exchange
- RD03: Support for data integrity, confidentiality, accountability, provenance, sovereignty
- RO01: Support long running experiments and large data volumes generated at high speed
- RO02: Support of virtual scientist communities, addressing dynamic user groups creation and management, federated identity management

EU policy on Open Research Data

- Research data can be defined as whatever is either produced in the research process or evidences research outputs such as articles
- The European Commission’s Research Data definition
  - **“Information, in particular facts or numbers, collected to be examined and considered and as a basis for reasoning, discussion, or calculation”**
  - https://ec.europa.eu/research/openscience/index.cfm?pg=openaccess
  - https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm
  - https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm
- Examples include: statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings, images
- **Open data** are deposited in institutional or specialist repositories and licensed appropriately so that prospective users know clearly any limitations on re-use.

Horizon 2020 Open Research Data (ORD) Pilot

- ORD pilot aims to improve and maximise access to and re-use of research data generated by Horizon 2020 projects, taking into account
  - the need to balance openness and protection of scientific information
  - commercialisation and IPR
  - privacy concerns
  - security
  - data management and preservation questions
- Applying principle '**as open as possible, as closed as necessary**’
- Complying with **FAIR Data principles**
- ORD applies primarily to the data needed to validate the results presented in scientific publications.
  - Other data can also be provided by the beneficiaries on a voluntary basis.

Horizon 2020 Data Management and Data Management Plan

- Data Management Plans (DMPs) are a key element of good data management.
  - A DMP describes the data management life cycle for the data to be collected, processed and/or generated by a Horizon 2020 project.
  - Help making research data Findable, Accessible, Interoperable and Reusable (FAIR)
- DMP should include information on:
  - handling of research data during & after the end of the project
  - what data will be collected, processed and/or generated
  - which methodology & standards will be applied
  - whether data will be shared/made open access and
  - how data will be curated & preserved (including after the end of the project).
- The project **must submit a first version of DMP** (as a deliverable) within the **first 6 months** of the project.
  - DMP is updated if data are changed
  - DMP is mandatory for projects participating in ORD Pilot

[ref] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm