Data Spaces for Interoperable Research Data Management and Governance

Yuri Demchenko

Session on Data Spaces for RDM

SLICES Workshop @ IoT Week

23 June 2022

Outline

- Introduction and goals
  - SLICES-RI and Research Data Management
- European Data Spaces and related documents
  - Data Act - Data Governance Act – GDPR - Data Altruism
- Industrial Data Spaces
  - IDSA Reference Architecture Model and Data Sovereignty
  - Open DEI Data Spaces Design principles
- EOSC Interoperability Framework and FAIR data principles
- Discussion: Benefits of adopting Data Spaces concept for RDM

Motivation and Goal: SLICES-RI and Research Data Management

- Can Data Spaces concept facilitate interoperability and sustainability in research data sharing?
- Best practices and standardisation: SLICES-RI, EOSC, RDA, and IDSA/Industry
- Infrastructure services for data sharing
- Interoperability in data sharing including between Research and Industry

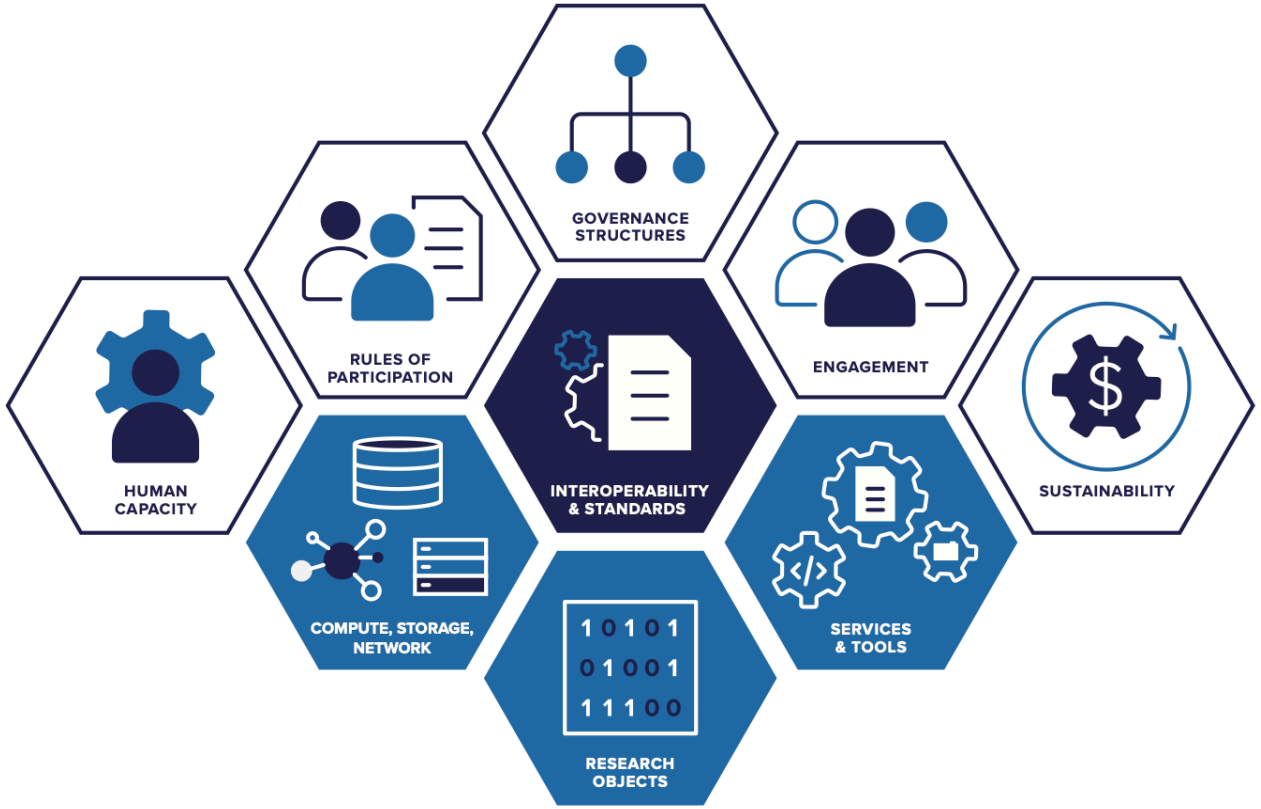
SLICES-RI and Research Data Management & Governance

- ESFRI SLICES-RI is a future RI for digital technologies experimentation
  - Require infrastructure for experiments access and management
  - Require supporting infrastructure for experimental and research data sharing (both computer and semantic)
- Experimentation data are big, distributed, domain specific, serving specific community
  - Require effective models and infrastructure services for RDM and secure data sharing

Many activities related to Data Spaces – Up-to-date

- IDSA: Continuous activity of multiple groups in International Data Spaces Association (IDSA)
- RDA: BoF on Data Spaces for Research Data Taxonomy scheduled for RDA19 Plenary on 21 June 2022 (morning due to time difference)
  - Reviewing existing Data Spaces definitions and suggesting harmonization and apporach
- RDA: BoF on Data Commons (actually similar to Data Spaces concept in US and Australia) – 22 June 2022 (night)
  - Data Commons Typology diagram: part od the concept by RDA activity GORC (Global Open Research Commons)
- DSBA (BDVA + IDSA) organised DataWeek2022 event 24 May - 9 June 2022, online on on-site meeting on 8-9 June in Naples.
- SLICES workshop and session "Data Spaces for interoperable Research Data Management and Governance (DS4RDM)" during IoT Week on 23 June 2022 (this event)
  - The BoF focus is DS for RDM and relevance to RI via SLICES.
- During IotWeek2022, few meetings about DS are organised by DSBA, IDSA, project Team Data Spaces (https://dataspaces4.eu/) and many projects focusing on Data Markets

RDA GORC Typology Diagram



- Interoperability and standards
- Governance Structures
- Builds on participation
- Engagement
- Human capacity
- Sustainability
- Computer, Storage, Network
- Services and tools
- Research objects

Data Spaces Taxonomy BoF - 21 June 2022

Two-three line summary of session topic / focus and discussion:

Term 'Data Spaces' is appearing recently in many contexts related to data governance and management, data exchange and sharing. The session focused on the nature of the data spaces and taxonomies. Also differences and similarities with data commons were discussed.

Key Outcomes/Actions/Takeaways:

1. *There are multiple use cases for data spaces concept for example in the area of the health data*
2. *Differences and similarities with "data commons" and other related concepts needs further discussions*
3. *Taxonomy of data spaces (and data commons) is needed to produce or enhance interoperability between data sets and infrastructures*

Slide courtesy Ville Tenhunen, EGI

European Data Space(s)

- European Data Governance Act
  - Support for set-up and development of common European data spaces in strategic domains, involving both private and public players
- European rules and values to ensure free flow of data between EU countries and outside
  - Sharing and re-using data, data monetisation
  - Enabling regulated environment for data processing
- Industrial Data Space: Sharing data while ensuring data Sovereignty
  - Data Sovereignty achieved by attaching policy and usage conditions to data
  - Enabling regulated environment for data processing
- International Data Spaces Association and GAIA-X

European Data Space – Policy Aspects (since 2018)

European strategy for data COM(2020) 66 final, 19.02.2020
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066
The European data space will give businesses in the EU the possibility to build on the scale of the Single market. Common European rules and efficient enforcement mechanisms should ensure that:

- Data can flow within the EU and across sectors;
- European rules and values, in particular personal data protection, consumer protection legislation and competition law, are fully respected;
- Rules for access to and use of data are fair, practical and clear, and there are clear and trustworthy data governance mechanisms in place; there is an open, but assertive approach to international data flows, based on European values.

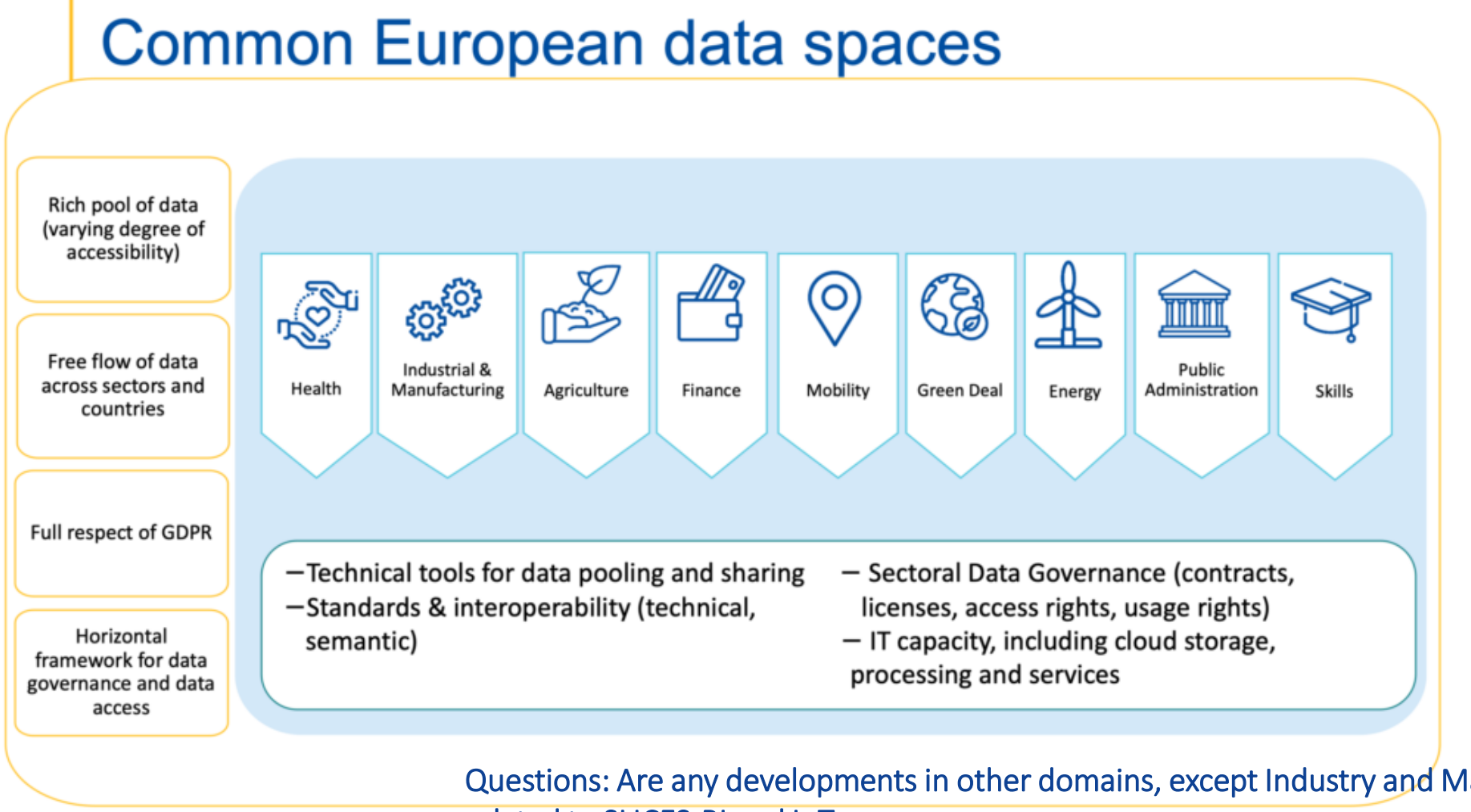
Key documents

- European Data Governance Act, Nov 2020
- Digital Services Act package (Q4 2020)
- Data Act, February 2022

European Data Governance – Covered Aspects

- General aspects
  - (a) conditions for the re-use, within the Union, of certain categories of data held by public sector bodies
  - (b) a notification and supervisory framework for the provision of data sharing services
  - (c) a framework for voluntary registration of entities which collect and process data made available for altruistic purposes.
- Re-use of certain categories of ***protected data*** held by public sector bodies which are protected on grounds of:
  - (a) commercial confidentiality
  - (b) statistical confidentiality
  - (c) protection of intellectual property rights of third parties
  - (d) protection of personal data
- Requirements to data sharing services and providers
- Data Altruism and organisations

Data Governance Act: Common European Data Spaces



- Health
- Industry and Manufacturing
- Agriculture
- Finance
- Mobility
- Green Deal
- Energy
- Public Administration
- Skills

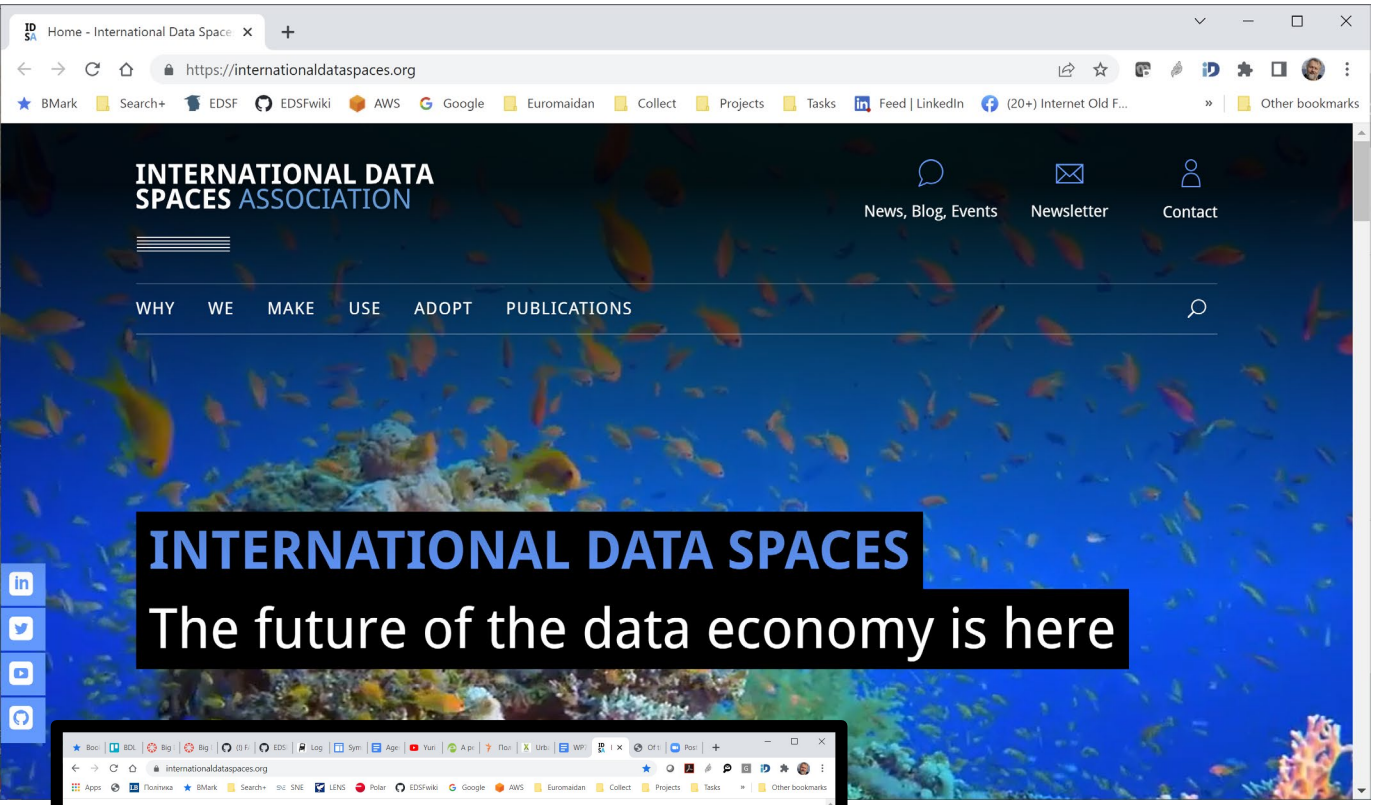
Questions: Are any developments in other domains, except Industry and Manufacturing? What is domain related to SLICES-RI and IoT

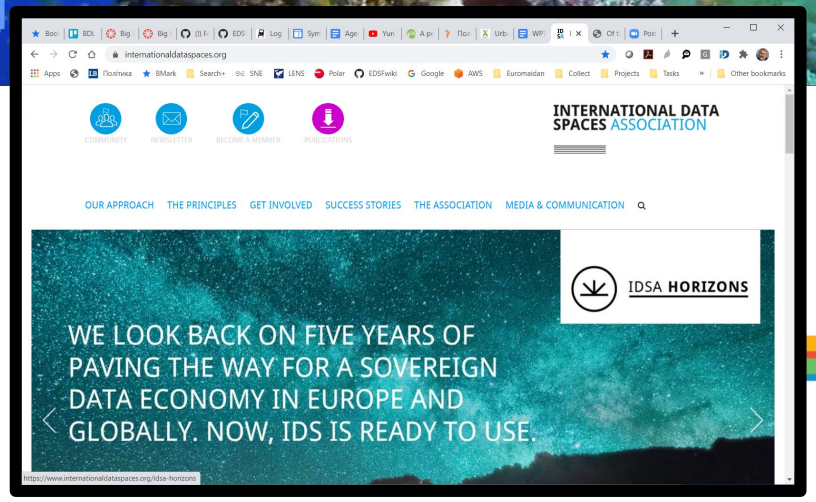
[ref] Common European Data Spaces http://dataspaces.info/common-european-data-spaces/

Industrial Data Spaces: Sharing Data

- Data sharing along the whole data handling flow in industrial or business processes – Allow for 3rd party services
  - Sensors - IoT – Ingest/Load – Ingest/Transform – Process – Act/Operate
- Sharing and Sovereignty are key to make value of data
  - Network of trusted data
- Data Sovereignty achieved by attaching policy and usage conditions to data
  - In non-fungible/blockchain way or protected environment
- Enabling regulated environment for data processing
  - Enclave computing in modern cloud technology

International Data Spaces Association (IDSA)
https://www.internationaldataspaces.org/

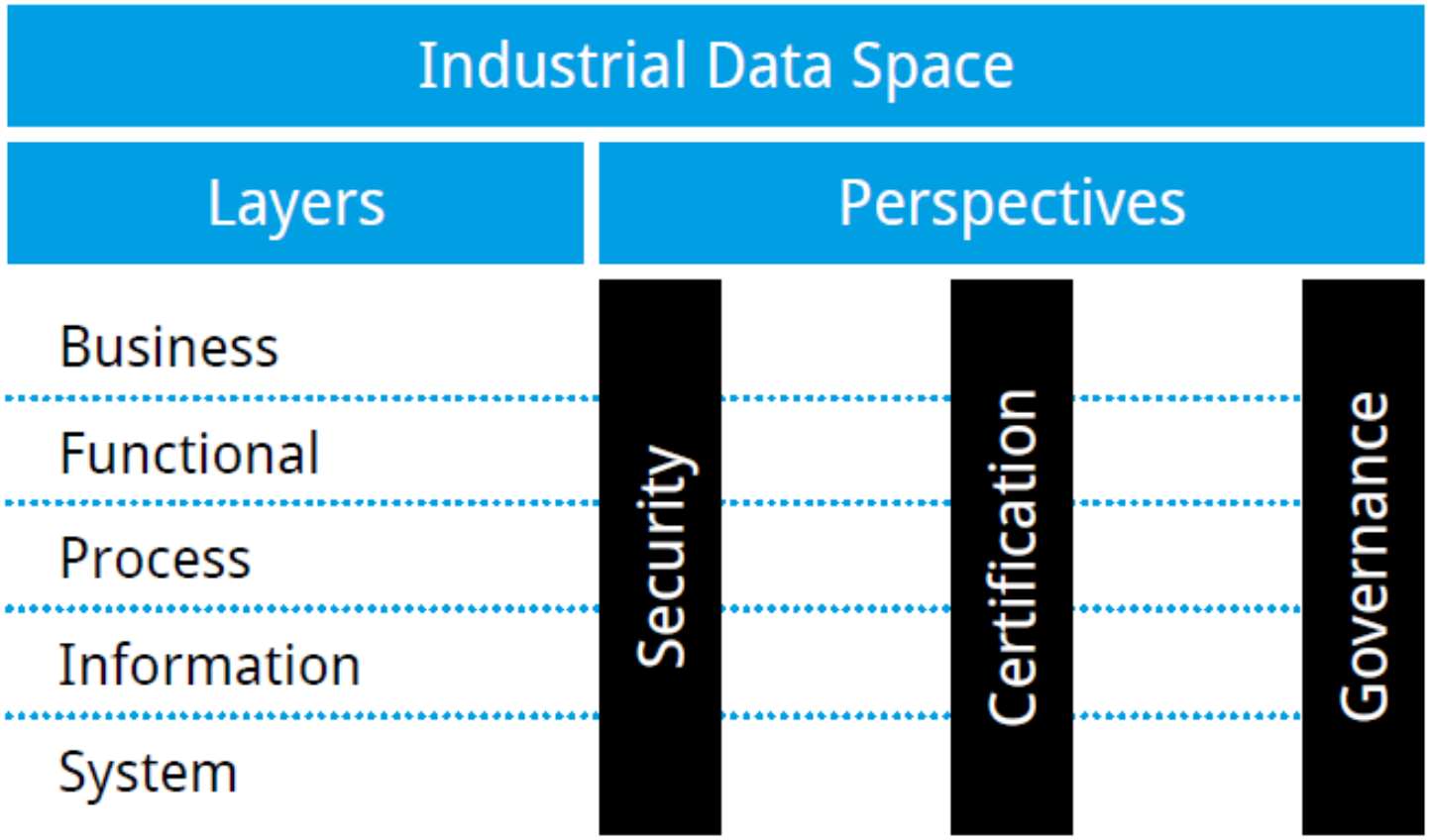




- More than 120 members
- Started 2016 as Industrial Data Space initiative (supported by German project)
- Re-defined as International Data Space Association (IDSA)
- Published IDSA Reference Architecture Model Version 3.0 (RAM3.0, 2018)
  - Whitepaper and use cases
- Focused on Industrial data sharing
- Key Issue is Data **Sovereignty**
- Number od associated H2020 and HE projects

- Community activities
  - Weekly architecture design meetings
  - Governance task force
  - Winter/Summer schools

General Structure of IDS Architecture



IDS Governance Perspective

- Governance aspects on different architecture layers
- Data as an economic goods
- Data Ownership
- Data Sovereignty
- Data Quality
- Data Provenance

- Specification defines functionalities by layers
- Details are sufficient to define processes, functional components and API

Data Spaces definition: Ongoing work by the Data Spaces Taxonomy BoF

- European documents
  - Data Spaces are domain specific ecosystems for data sharing governed by common policies (complying EU Data Governance policies)
- IDSA, OPEN DEI and GAIA-X
  - Data space can be defined as ***a federated data ecosystem*** within a certain application domain and based on shared policies and rules. A data space is made up of **multiple actors, which together form a data ecosystem**.
  - **Decentralized infrastructure for trustworthy data sharing and exchange in data ecosystem based on commonly agreed principles**

Data Spaces Interoperability Framework – Infrastructure

- Technical – Infrastructure
- Semantic – Metadata + Infrastructure Services
- Organisational
- Legal

EOSC Interoperability Framework (EOSC-IF) is about Data – To support FAIR data sharing

- FAIR Digital Object (FDO) is a key concept
- Technical Interoperability:
  - Artefact Common Protocols and Data Formats
- Semantic Interoperability:
  - Contextual Semantics related to Common Semantic resources
- Organisational Interoperability:
  - Contextual Semantics related to Common process resources
- Legal Interoperability:
  - Contextual licenses related to Common Licenses resources

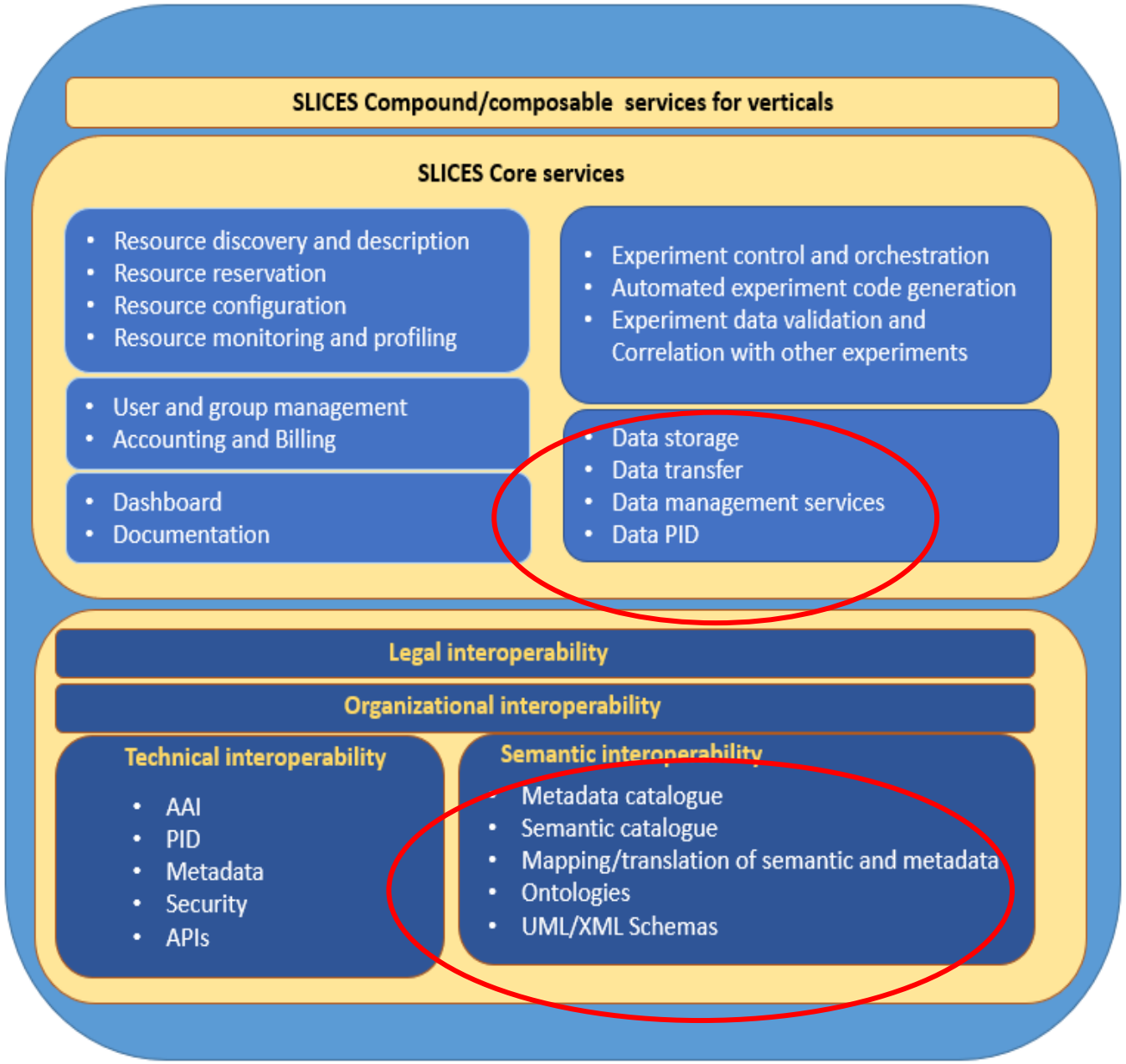
- FDO is actively promoted by GO FAIR Technical Center and Peter Wittenburg
- **EOSC Interoperability profiles to facilitate interoperability**

Example of EOSC Provider profile (EPP): *Provider data model*

Category	Components
Basic information	ID, Name, abbreviation, Website, legal entity(Y/N), status
Marketing information	Description, logo, multimedia
Classification	Scientific domain and sub-domain(s), tags(relevant keywords)
Location information	Street, postcode, city, region, country
Contact information	Main contact, public contact, Email, Phone number, position
Maturity information	Lifecycle status, certifications
Other information	Name of the organization, participating countries, affiliations, structure(single sited, distributed, etc.), ESFRI domain (e.g., energy, environment…), societal grand challenge defined by EC (e.g., health, food ...), part of the National roadmap for infra, etc.

Example of EOSC Resource profile (ERP): *Resource data model*

Category	Components
Basic information	ID, Name, resource organization, resource providers, Website
Marketing information	Description, tagline (catch-phrase), logo, multimedia, usecases supported
Classification	Scientific domain, scientific sub-domain(s), scientific category, scientific subcategory, target users, access type, access mode, tags (relevant keywords)
Geo and language information	Resource Locations, languages of the UI, resource availability location
Main Contact/Resource owner	Name, position, organization Email, Phone number, public contact, Helpdesk information (email. phone), security contact
Maturity information	TRL, lifecycle, certifications, standards, open source tech supported by resource, version, last update, change log
Dependency information	List of other resources required, commonly used with and platforms, etc.
Funding	Funding body, program and grant/project, etc.
Management information	Helpdesk page, manual, terms of use, privacy policy, access policy, service level, training info, status monitoring, webpage with planned maintenance info
Access and order info	Ordering procedure
Financial information	Payment model, pricing, etc.



Conceptual view of SLICES Interoperability Architecture

- Provides vision and roadmap to achieving interoperability with EOSC
- Some services can be used from EOSC, some services will require API with EOSC services of metadata mapping
- Data Interoperability and sharing is an important component of SLICES-RI
  - Compliance with the Open Science and FAIR data principles
  - Semantic interoperability
  - Supported by robust data infrastructure
  - Data Management and Governance

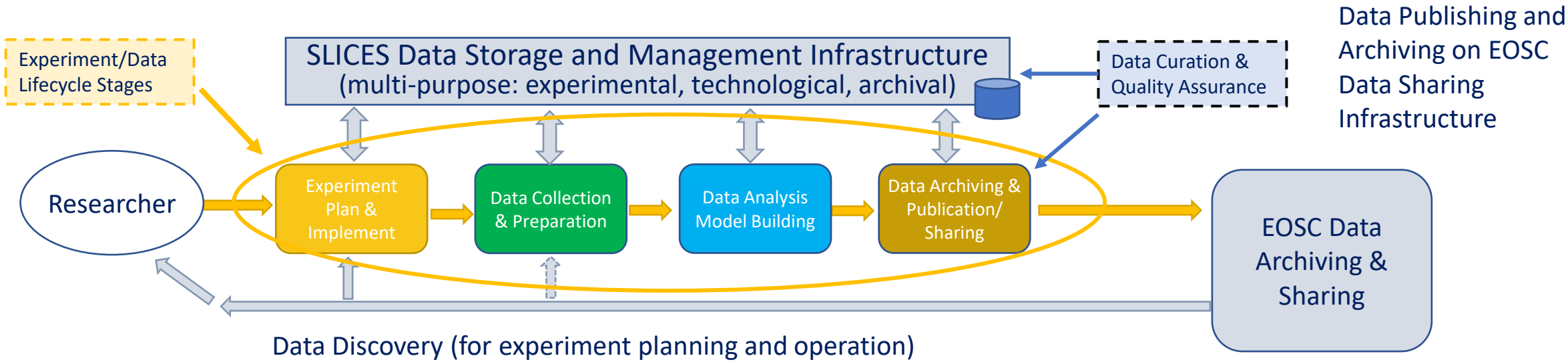
EOSC Provider and Resource Metadata Profiles

- Q1. How Data Spaces concept (as domain specific data ecosystem) can facilitate data sharing in your domain?
- Q2. Data Spaces and FAIR data principles: What place in the Data Spaces architecture?
- Q3. Is there a specifics for defining metadata profiles for experimental data sharing? What level of standardisation is required?
- Q4. What infrastructure services are needed to enable expected Data Spaces implementation and operation?
- Q5. What aspect of inter-domain, inter-sector or international cooperation do you see?

Additional/Context information

- SLICES Data Lifecycle Model and Dataflow
- SLICES Data Management and Governance: Infrastructure, Services, Policy
- FAIR Data Principles: Metadata Management

SLICES Data Lifecycle Model and Dataflow



- **Each Data Lifecycle stage** – experiment, data collection, data analysis, and finally data archiving, works with own **data set, which are however connected**. All these data sets need to be stored and possibly re-used in later processes.
- Many experiments and research require already existing datasets that will be available in the SLICES data repositories or can be obtained/discovered in EOSC data repositories

Robust Data Infrastructure for Data Driven Research

- Distributed data storage and experimental data(set) repositories
  - Supporting recognized data interoperability standards (data formats and metadata)
  - Eventually certified: RDA endorsed Maturity and certification practice
- Support the whole data lifecycle
  - Connected to research/experiment lifecycle or workflow
- Linked data and data discovery
  - PID (Persistent IDentifier) and FDO (FAIR Digital Object) infrastructure
    - Supported by EOSC and developed by RDA and number of H2020 and HE projects
- (Trusted) Data exchange and transfer protocols
  - Supported by rules and policies
  - Interoperable metadata
- Data management and data curation and quality assurance

SLICES Data Management and Governance: Infrastructure, Services, Policy

- Data Infrastructure
  - Distributed data storage and repositories (for experiments support and data publishing)
  - PID/FDO infrastructure for data discovery and linked data
  - Data Catalogs and Metadata Catalogs
- Data Management
  - FAIR data principles and metadata registries (schemas)
  - Data Management Plan and Data Stewards (to maintain DMP and data quality)
  - Define metadata for sharing and publication of data produced by SLICES
  - Define metadata for the resource and experimental facilities description
- Data Governance:
  - Organisational and legal aspects in data handling
  - Policies and IPR
  - personal data protection

FAIR Data Principles: Metadata Management (GO FAIR recommendations)

Findable:

- F1 (meta)data are assigned a globally unique and persistent identifier;
- F2 data are described with rich metadata;
- F3 metadata clearly and explicitly include the identifier of the data it describes;
- F4 (meta)data are registered or indexed in a searchable resource;

Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles;
- I3. (meta)data include qualified references to other (meta)data;

Accessible:

- A1 (meta)data are retrievable by their identifier using a standardized communications protocol;
  - A1.1 the protocol is open, free, and universally implementable;
  - A1.2 the protocol allows for an authentication and authorization procedure, where necessary;
- A2 metadata are accessible, even when the data are no longer available;

Reusable:

- R1 meta(data) are richly described with a plurality of accurate and relevant attributes;
- R1.1 (meta)data are released with a clear and accessible data usage license;
- R1.2 (meta)data are associated with detailed provenance;
- R1.3 (meta)data meet domain-relevant community standards;

slicesRI

FAIR from the technical point of view

- Findable
  - Metadata and PDI – infrastructure and tools
  - Registries and handles resolution, API
  - Policies and SLA
- Accessible
  - Repositories and data storage: infrastructure and management
  - Policy and access control: infrastructure and API management
  - Data access protocols
  - Usage Policy and Sovereignty
  - Data protection, compliance, privacy and GDPR
- Interoperable
  - Standard data formats
  - Metadata and API
  - FAIR maturity level and certification
- Reusable
  - Data provenance and lineage
  - Preservation
  - Metadata, PID and API – linked or embedded into datasets

Require comprehensive **data infrastructure** to support

- Data Storage
- Data publication
- Data discovery
- Linked data and data lineage (provenance)
- Multiple datasets access for analysis