EOSC-Pillar

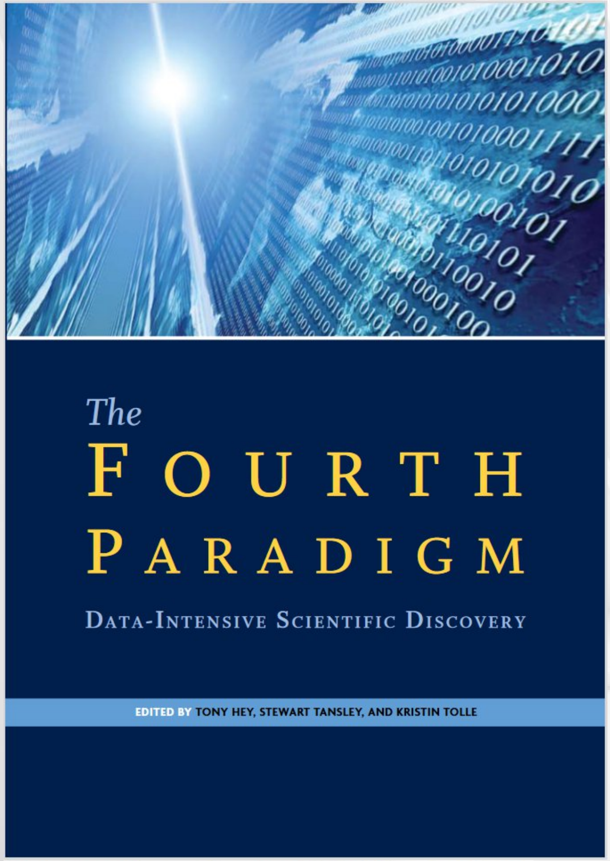
Coordination and Harmonisation of National & Thematic Initiatives to support EOSC

Solving the data aggregation and data integration problem

Philipp von Hartrott

EOSC-Pillar has received funding from the European Union's Horizon 2020 research and innovation Programme under Grant Agreement No. 857650.

Paradigm changes



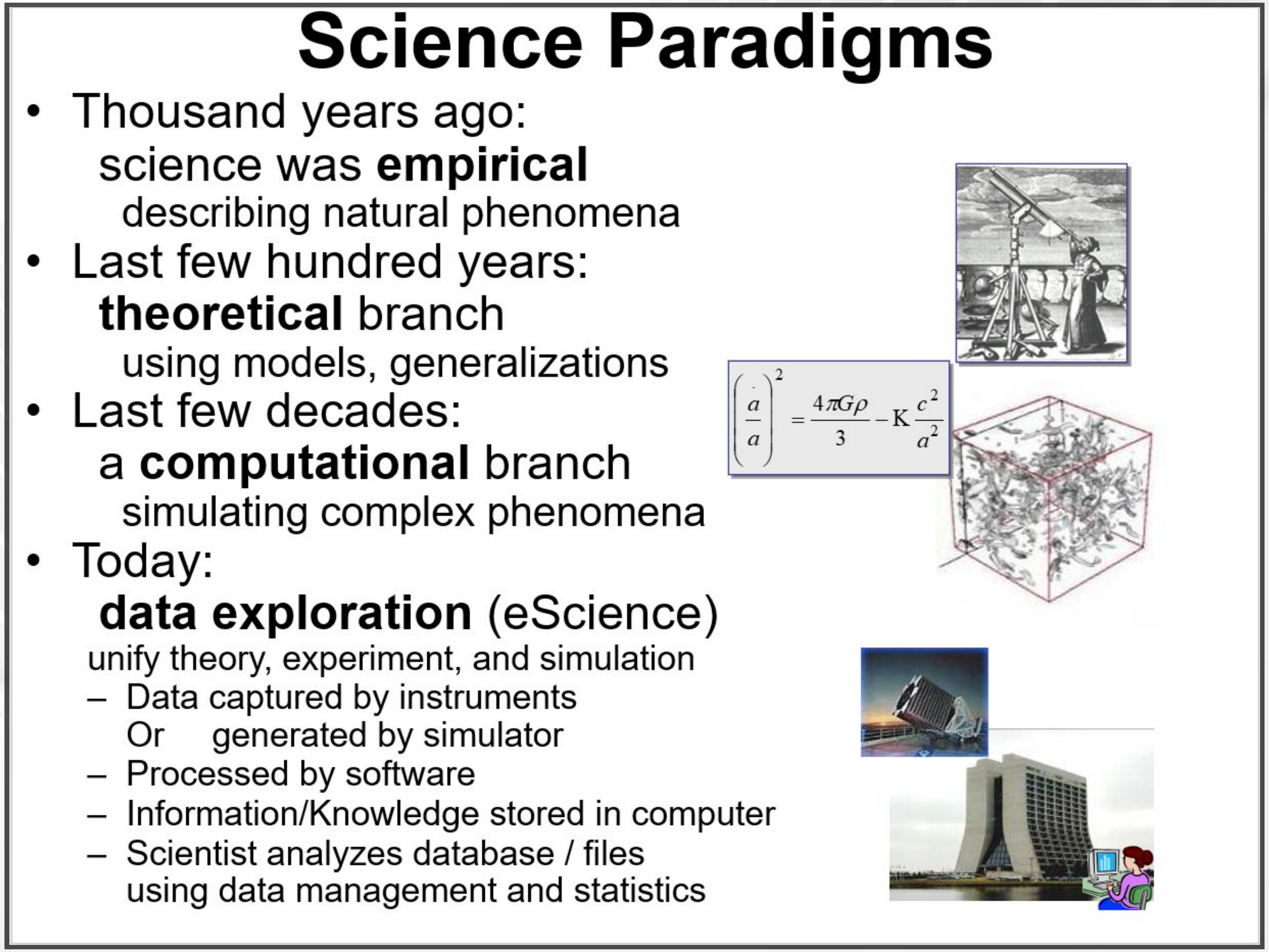
The
FOURTH PARADIGM

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

Tony Hey, Stewart Tansley, Kristin Tolle, Jim Gray
Published by Microsoft Research | October 2009
ISBN: 978-0-9825442-0-4

https://www.microsoft.com/en-us/research/wp-content/uploads/2009/10/Fourth_Paradigm.pdf



Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today:
  data exploration (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments
    Or generated by simulator
  - Processed by software
  - Information/Knowledge stored in computer
  - Scientist analyzes database / files
    using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$

Jim Gray, Alex Szalay, 2007 http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt

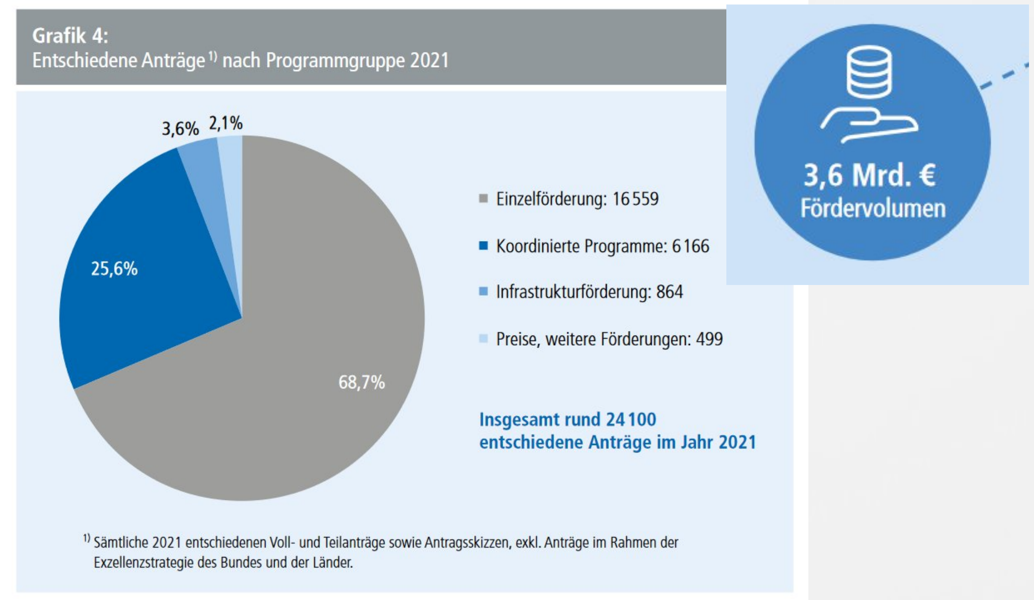
Why are **we** doing it?

* Gigaprojects can afford creating their own IT infrastructure.

* Megaprojects can contribute towards a collective scientific IT infrastructure.

* The vast majority of research projects depend on collective scientific IT infrastructure.

German Science Foundation DFG Yearbook 2021

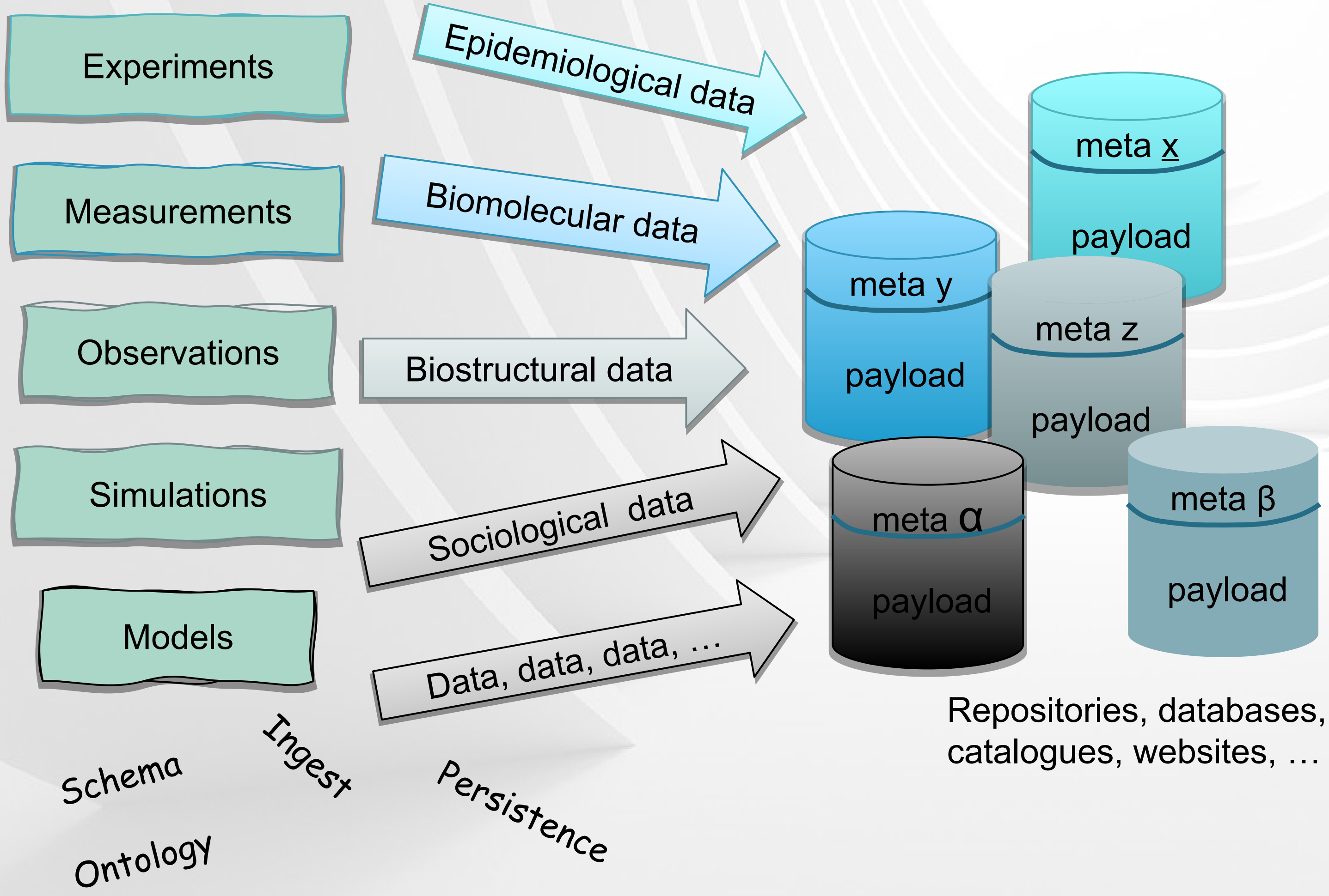


Grafik 4:
Entschiedene Anträge[1] nach Programmgruppe 2021

[1] Sämtliche 2021 entschiedenen Voll- und Teilanträge sowie Antragsskizzen, exkl. Anträge im Rahmen der Exzellenzstrategie des Bundes und der Länder.

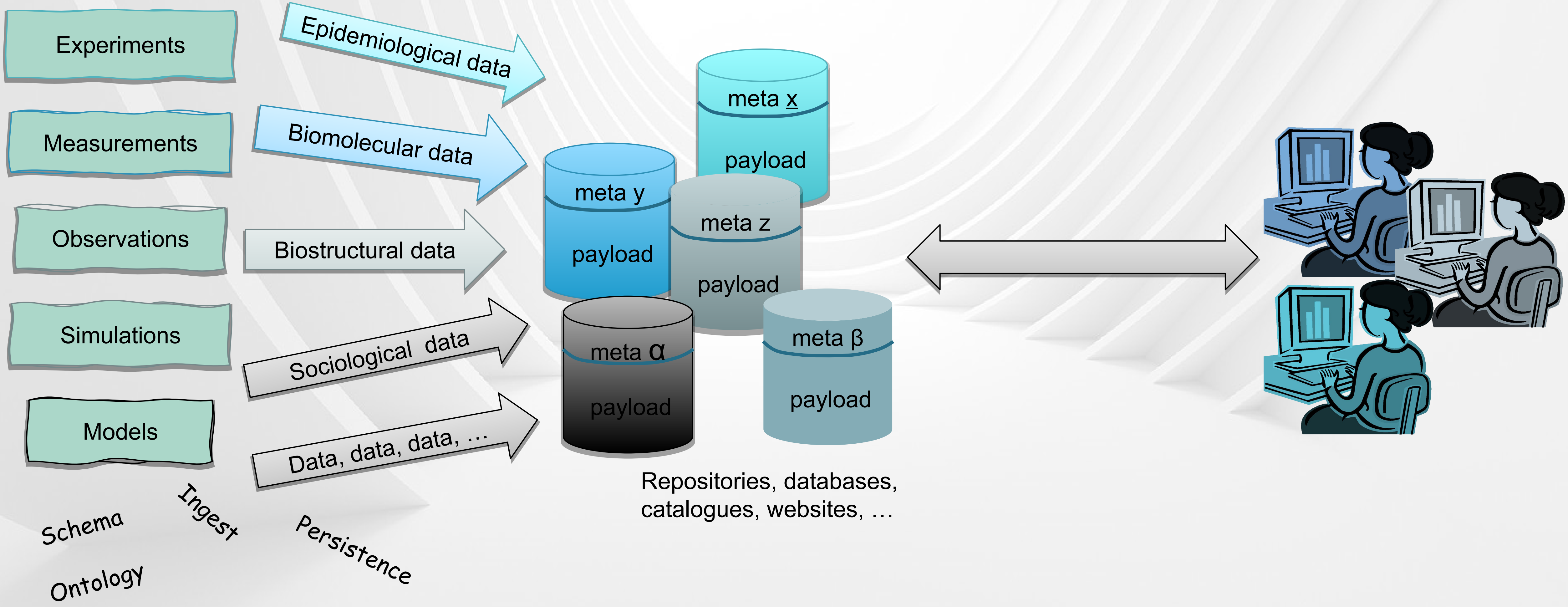
https://www.dfg.de/download/pdf/dfg_im_profil/geschaeftsstelle/publikationen/dfg_jb2021.pdf

Are we ready?



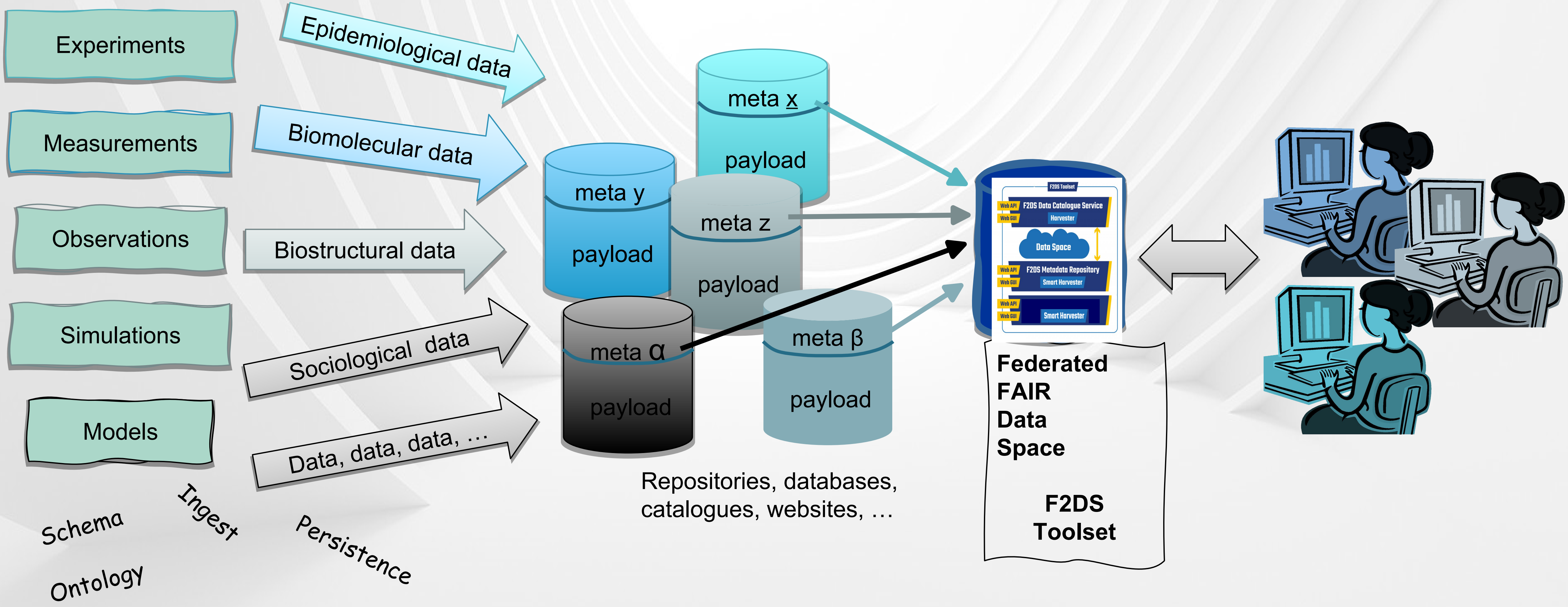
Repositories, databases, catalogues, websites, …

Are we ready?



Repositories, databases, catalogues, websites, …

Are we ready?



Experiments

Measurements

Observations

Simulations

Models

Epidemiological data

Biomolecular data

Biostructural data

Sociological data

Data, data, data, …

Schema

Ingest

Ontology

Persistence

meta x

payload

meta y

payload

meta z

payload

meta α

payload

meta β

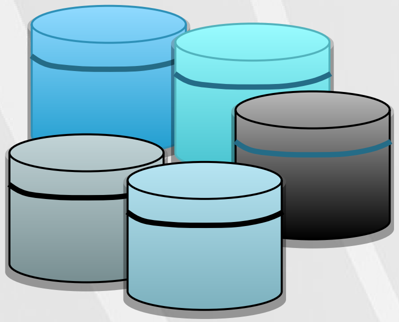
payload

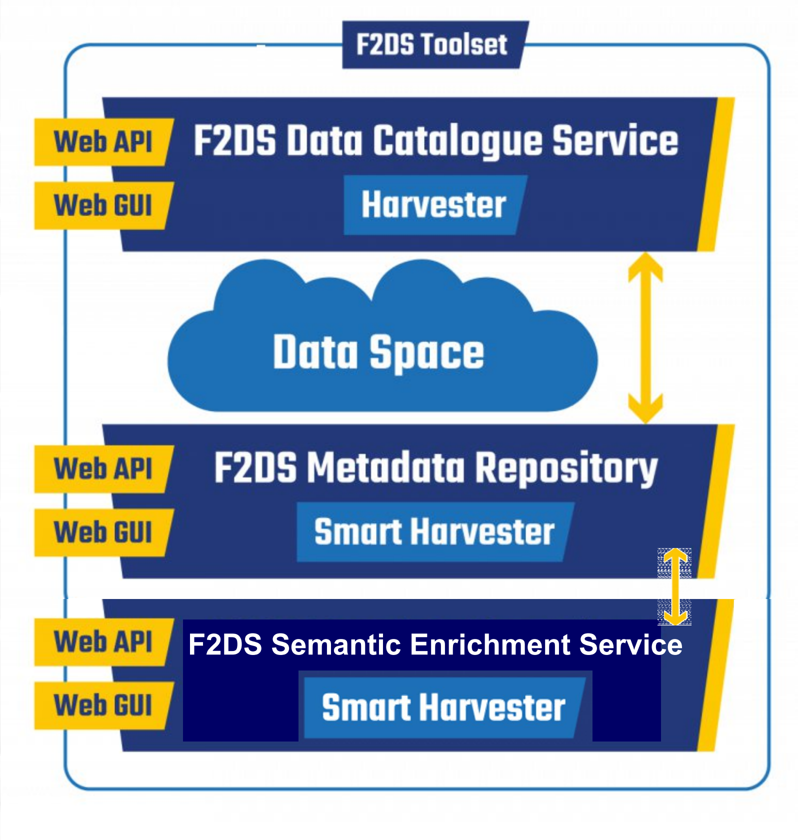
Repositories, databases,
catalogues, websites, …

Federated
FAIR
Data
Space

F2DS
Toolset

F2DS Toolset





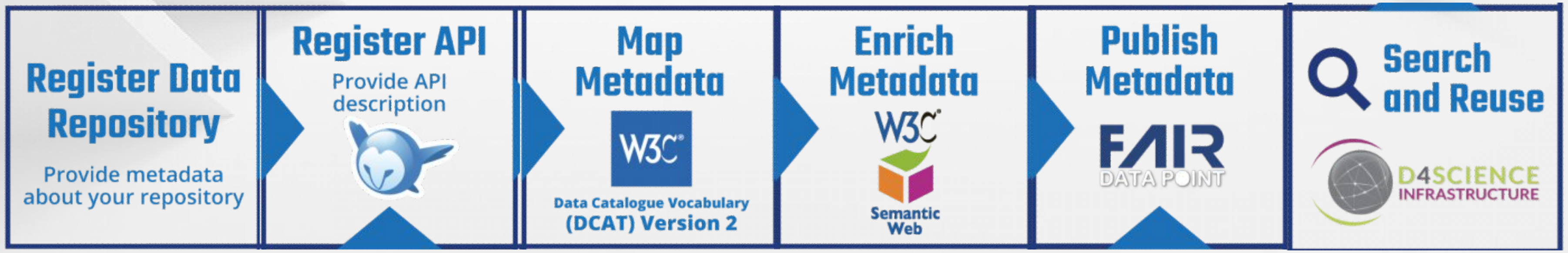
F2DS Toolset

Web API
Web GUI
F2DS Data Catalogue Service
Harvester

Data Space

Web API
Web GUI
F2DS Metadata Repository
Smart Harvester

Web API
Web GUI
F2DS Semantic Enrichment Service
Smart Harvester



Register Data Repository
Provide metadata about your repository

Register API
Provide API description

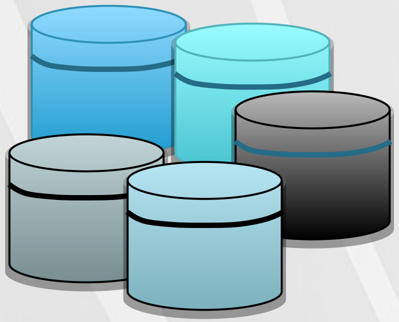
Map Metadata
Data Catalogue Vocabulary (DCAT) Version 2

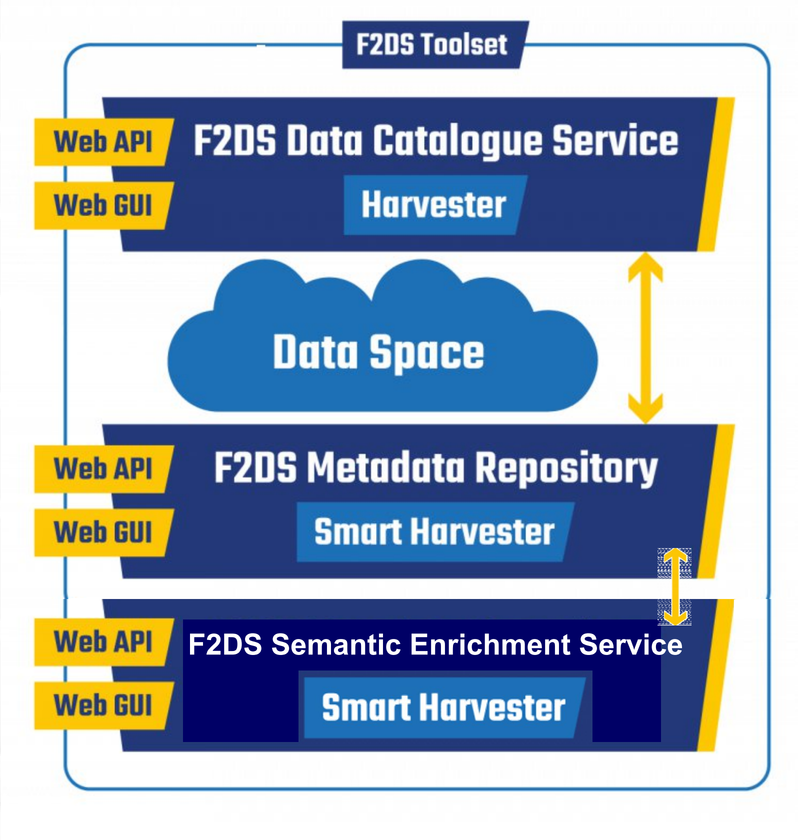
Enrich Metadata
Semantic Web

Publish Metadata
FAIR DATA POINT

Search and Reuse
D4SCIENCE INFRASTRUCTURE

F2DS Toolset





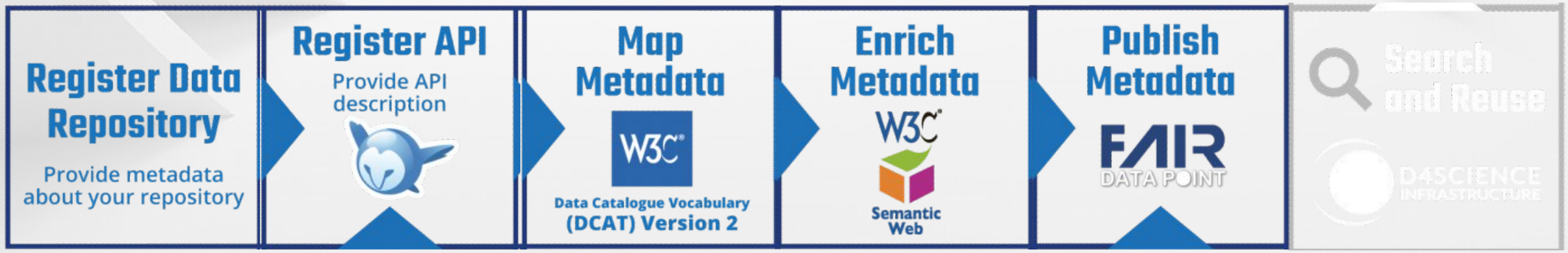
F2DS Toolset

Web API
Web GUI
F2DS Data Catalogue Service
Harvester

Data Space

Web API
Web GUI
F2DS Metadata Repository
Smart Harvester

Web API
Web GUI
F2DS Semantic Enrichment Service
Smart Harvester



Register Data Repository
Provide metadata about your repository

Register API
Provide API description

Map Metadata

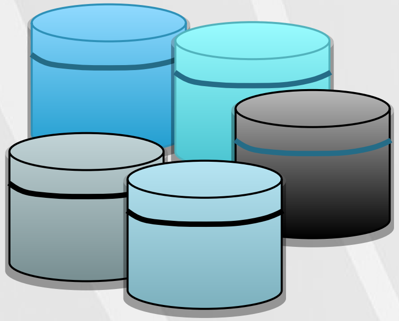
Data Catalogue Vocabulary (DCAT) Version 2

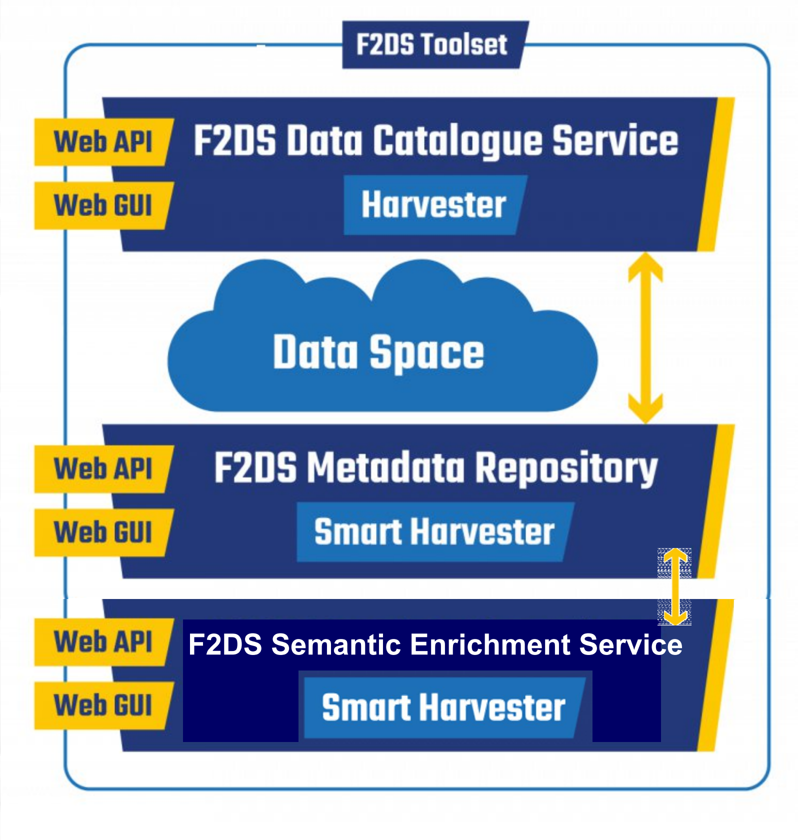
Enrich Metadata

Publish Metadata

Search and Reuse

F2DS Toolset





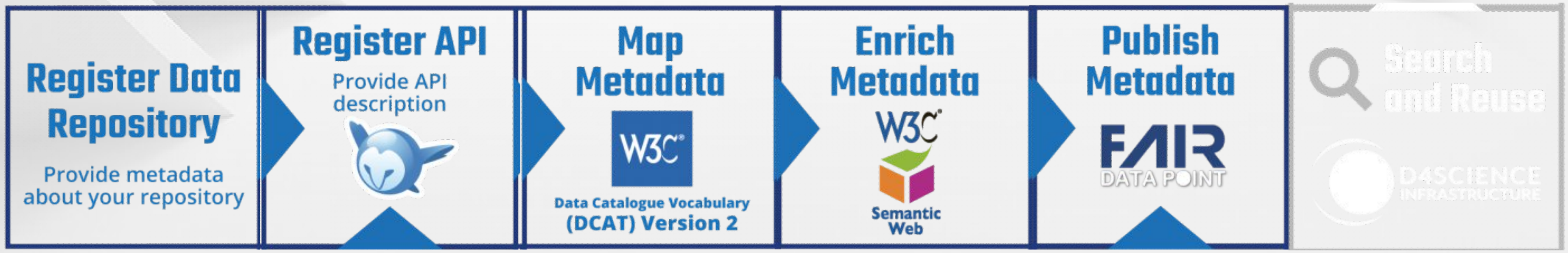
F2DS Toolset

Web API
Web GUI
F2DS Data Catalogue Service
Harvester

Data Space

Web API
Web GUI
F2DS Metadata Repository
Smart Harvester

Web API
Web GUI
F2DS Semantic Enrichment Service
Smart Harvester



Register Data Repository
Provide metadata about your repository

Register API
Provide API description

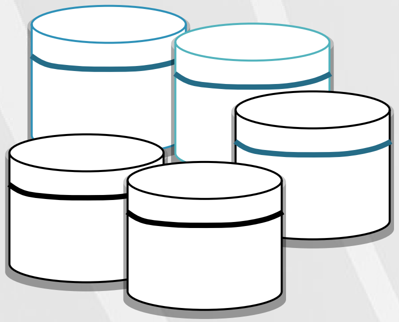
Map Metadata
Data Catalogue Vocabulary (DCAT) Version 2

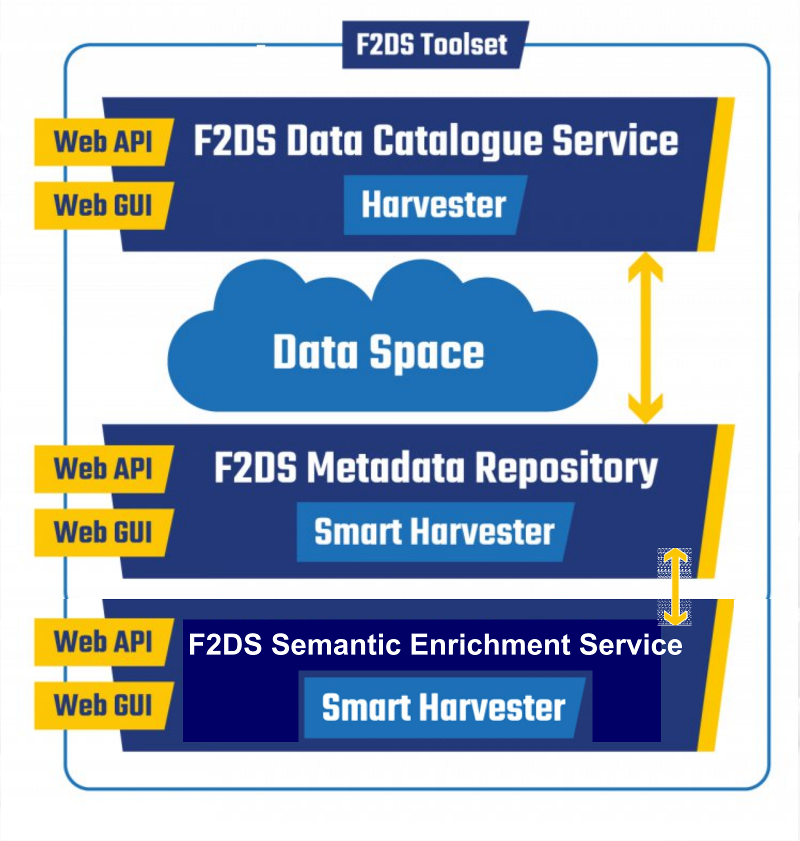
Enrich Metadata
Semantic Web

Publish Metadata
FAIR DATA POINT

Search and Reuse
D4SCIENCE INFRASTRUCTURE

F2DS Toolset





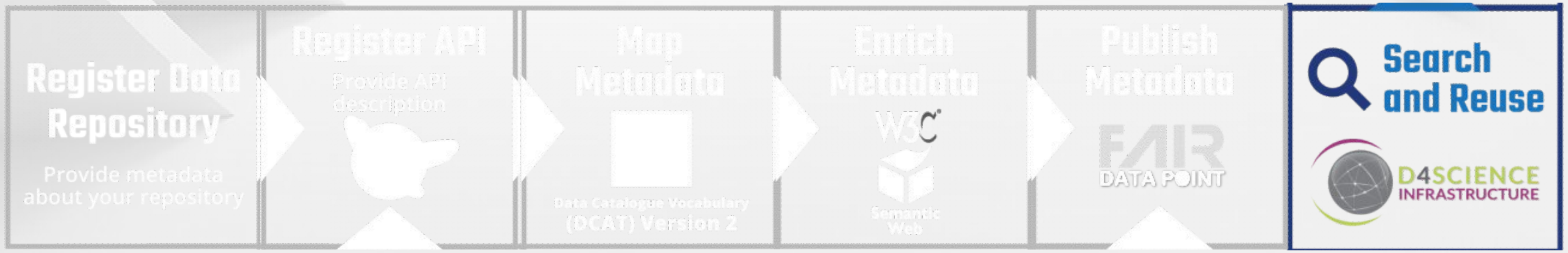
F2DS Toolset

Web API
Web GUI
F2DS Data Catalogue Service
Harvester

Data Space

Web API
Web GUI
F2DS Metadata Repository
Smart Harvester

Web API
Web GUI
F2DS Semantic Enrichment Service
Smart Harvester



Register Data Repository
Provide metadata about your repository

Register API
Provide API description

Map Metadata
Data Catalogue Vocabulary (DCAT) Version 2

Enrich Metadata
W3C
Semantic Web

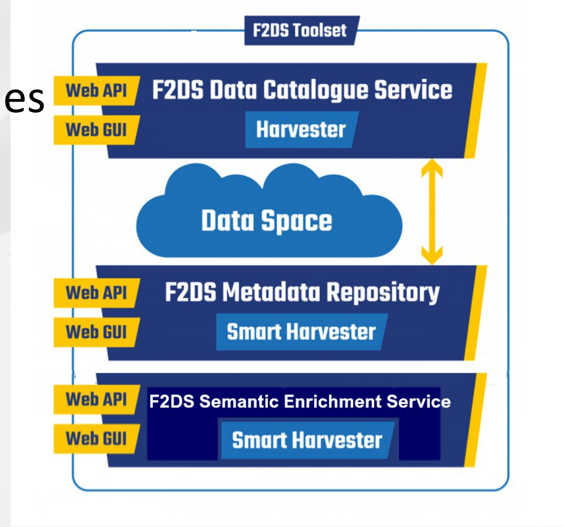
Publish Metadata
FAIR DATA POINT

Search and Reuse
D4SCIENCE INFRASTRUCTURE

Outlook

* Improve F2DS with feedback from community and testers.
* Provide finalized version with registered repositories from different communities.
* Will have integrated a previously unconnected service including proof of concept for search engine for semantic concepts.

* Domain specific metadata should receive more attention
* Semantic enrichment foundation is available.
  —> Build tools that exploits this enrichment, based on knowledge encoded in ontologies
* Payload data if often not readily actionable.
  —> Promote FAIR principles and generic tools
* Interested in more technological details and a demo
  about the F2DS ? **Visit session S7 today 14:00-15:30!**



Data aggregation: informal definition

Data aggregation is the compilation of data from different sources into a common schema with intent to process and amplify the information content.

Data integration: formal definition

Data integration system tuple: <G, **S**, **M**>

G - Global schema (here DCAT)

S - Source schemas (multiple, e.g.….)

M - Mappings (here part of FDP)

Jim Grays action items

* Foster Tools and Foster Tool Support
* Invest in Tools at all levels
* Foster Generic LIMS (Lab info management systems)
* Foster Data Management-, Data Analysis-, Data Visualization- Algorithms &Tools
* Do for other sciences what NLM has done for BIO Genbank-PubMedCentral…
* Foster new document authoring and publication models and tools
* Foster Digital Data Libraries (not metadata, real data) and integration with literature

Typical dataset catalogs

data.gov (US), CKAN REST, 330 000 datasets

worldbank.org (int), proprietary REST, 5 500 datasets

eea.europa.eu (EU), proprietary SPARQL, 1 700 datasets

fao.org (int), CKAN REST, 2 100 datasets

dkrz.de (DE), proprietary REST, 840 000 datasets

inrae.fr (FR), dataverse REST, 1 900 datasets

nakala.fr (FR), proprietary REST, 450 000 datasets

Different disciplines require different metadata properties to their data. This motivates the creation of proprietary search APIs. (Example: Geographic, Keywords, …)

EOSC-Pillar

Coordination and Harmonisation of National & Thematic Initiatives to support EOSC

Thank you!

Get in touch with us!

🔗 www.eosc-pillar.eu

@EoscPillar

/company/eosc-pillar

EOSC-Pillar has received funding from the European Union’s Horizon 2020 research and innovation Programme under Grant Agreement No. 857650.