HELIX

Hellenic
Data
Service

HELIX

Hellenic
Data
Service


ATHENA
Research & Innovation
Information Technologies

Athena RC

Why us?

- Our **scientific focus** is on cross-disciplinary, transformative **data-intensive** research (Big, Open, Linked data)
- We champion **Data Economy**, **Big Data** and **Data Science** for national economic growth
- We lead **EU/national policies** and **technical interventions** on **Open Access**/**Data** (scientific infrastructures, data catalogues)
- Our collective insights and knowledge shaped the vision, implementation, an governance of HELIX grounding it on **real-world challenges and considerations**

Motivation

Motivation

Converging Policy landscape

- **Data Economy** – a strategic priority for EU's sustainable future growth integrating policy, technology, and innovation actions
- **Public Sector Information** – open up and create value from public-sector and publicly-funded Data (open data, INSPIRE, OGP, …)
- **Industrial Data Platforms** – emerging organization & technical instrument to facilitate data sharing and valorization within EU industrial value chains
- **Research -**
  - **Open Access** – de jure policy for sharing EU-funded scientific output
  - **Data Management Plans** – formalize data handling on project/organization-level
  - **FAIR data** – de facto international policy for scientific data

Motivation

The Big Picture

Economic growth, scientific progress, and societal prosperity are about **searching, sharing, using, experimenting, building, and valorizing**

(*frictionless)

Data

* = simple, fast, inclusive,

Motivation

Archiving-focused Data Platforms

- **Flexible, low-cost, open,** collaborative services for simplifying **sharing**, **discovery**, **use**, **analysis**, and **visualization** of scientific data
  - Lower the **entry barrier**, embrace **all** types of data
  - Make data **useful** to **most** scientists, **most** of the time

I am not a librarian

Motivation

Challenging the status quo

- **Data useful for research** are well **beyond** and **above** those linked with a **publication**
- We live in the **Data Economy** and **Big Data** age; everything is becoming **data-focused** and **data-intensive**
- **Let's change**: the explicit **assumption** is that they serve **scientists**
  - Most useful data are not linked with publications
  - Make it easier to publish data, why the strict rules?
  - Help me use and experiment with data

I am a scientist

Motivation

Why is this needed?

- Key lessons learned from **Open Data** are highly relevant for Research Data Platforms
  - Lower the **entry barrier**, making it **easy, simple, and fast** to publish and find data
  - No **walled gardens**; **all data**, from **any field** are welcomed, at **any point** of their lifecycle
  - Make data **useful** to **more people**, **most of the time** (80/20) through visualization and services
  - Ensure **sustainability**

Motivation

Sustainability

- A Research Data Platform must be **diachronic**, ensuring data are always accessible, and **evolving**, addressing the ever-growing data-intensive needs of scientists
  - Relatively low CAPEX (setup), **higher and fluctuating OPEX** (operation, growth)
  - Public funding may not **suffice** or be timely **available**
  - **Devaluation** is (only) a few steps away (stale/missing data, no QA/SLA)
  - Need to introduce additional **revenue streams**, but from where?
- Again, lesson from Open Data
  - **Industry** amongst the first and leading users of Open Data, generating **value** from new/improved services
  - Sharing and using **industrial data** in commercial **value chains** remains a challenge

Motivation

Industrial Data Platforms

- Data Platforms for securely **sharing, discovering, licensing, using**, and ensuring fair **reimbursement** of industrial data
  - Concept follows the **paradigm** of open data (simplicity, fit for purpose, benefits, fast, low cost)
  - Same technical **foundations** with key differences (confidentiality, contract management, IPR protection)
- **We can inherently serve these needs, provide a parallel industrial data platform by-design, and tap into the additional revenue streams**
  - **USPs**: scalable production-grade **data processing/analysis services**, **unified** proprietary & open data, **data science** as a service

HELIX

Hellenic
Data
Service

HELIX

Hellenic Data Service

- **Scientific Infrastructure for data-intensive research**
  - Supports the full lifecycle of scientific **data management, processing, sharing, and reuse**
  - Inherently **scalable**, **cloud-based**
  - Nation-wide, **horizontal**, **cross-domain**
  - **Low-cost**, economies of scale, network effects, maximize ROI
  - **Multiple roles**: Open Access, FAIR Data, Public Data, Industrial Data Platform

Data first

ATHENA
Research & Innovation
Information Technologies

HELIX

The 3 pillars of HELIX

- **Publications**
  - Nation-wide, cross-domain discovery of **publications**
  - Adapt and localize **OA OpenAIRE CRIS** services
- **Data**
  - Data **catalogue** and **repository** for **FAIR scientific** and **industrial data**
  - Discover, collect, evaluate, download, and use
- **Labs**
  - **Generic-purpose** and **domain-specific** services and APIs for **data analysis**, **processing**, and **experimentation**

Data alone is not enough

Target groups

- **Scientists**: data sharing, OA publishing, data experimentation
  - All scientific fields, including **citizen scientists**
- **Organizations**: institution-wide services augmenting, exposing, or replacing existing publication & data catalogues/repositories
  - Academia, Research, Public Administrations (PSI), special-interest groups
- **Scientific Infrastructures**: building block; scalable data processing services for very large, heterogeneous scientific data
  - Upcoming: ELIXIR (bio), APOLLONIS (linguistic)
- **Industry & innovators**: value-added services; ad hoc analysis services
  - **Industrial Data Platform**: low-cost data processing infrastructures; Data Science as a Service, training data for ML

HELIX

Core Concepts 1/2

- **Data-first:** make it simple, easy, and fast to share data (<10 secs); this is what is truly missing; build critical mass (data & users)
- **Scientists first:** serve the scientists, not librarians or standardization bodies; all too often this is lost, raising the entry barrier and thus failing (see open data)
- **Just another tool:** ensure inclusiveness and downplay our potential impact on the scientific process – be useful and in the background (just another hammer)
- **Love ALL data:** any data used during research (not only in pubs); we do not know what/how/where data will be useful; no data is too little, no data is too small

HELIX

Core Concepts 2/2

- **Cross-disciplinary:** actively avoid walled-gardens and domain silos; facilitate data-driven cross-disciplinary research (introduce data & problems, facilitate networking); ‘scientist’ role is fluid
- **Bundle data with services:** software, tools, and knowhow on how to use data is the 2nd greatest bottleneck behind data availability; think equally big (e.g. spark) and small (e.g. fast visualization);
- **Openness as a principle:** open software, open standards, open services (learn from others, give back)
- **Agility:** flexibility and reusability across all provided services and sub-systems; also during design and development of the system itself
- **All Scientists are Data Scientists:** data management, processing and analysis skills are integral in modern scientific practice

Development Roadmap

- **Phase 0 (incubation)**: 2012-2017
  - Original concept & funding proposal; core technology developed in other R&D projects; National Research Infrastructures Roadmap
- **Phase 1 (MVP):** 2018-2019
  - MVP for technical/policy foundations; core services & lighthouse apps/communities; prepare follow-up
- **Phase 2 (Beta):** 2020-2024
  - Scale services and expand reach to more scientific communities; integration in 3rd infrastructures; first industrial clients; governance structure; industrial data platform
- **Phase 3 (Production):** 2025-
  - Sustainable diachronic operation

HELIX Architecture

HELIX Architecture

Three pillars

hellenic**data**service**.gr** || **helix.gov.gr**

HELIX
Hellenic
Data
Service

HELIX
PUBS
Hellenic
Data
Service

pubs.hellenicdataservice.gr

HELIX
DATA
Hellenic
Data
Service

data.hellenicdataservice.gr

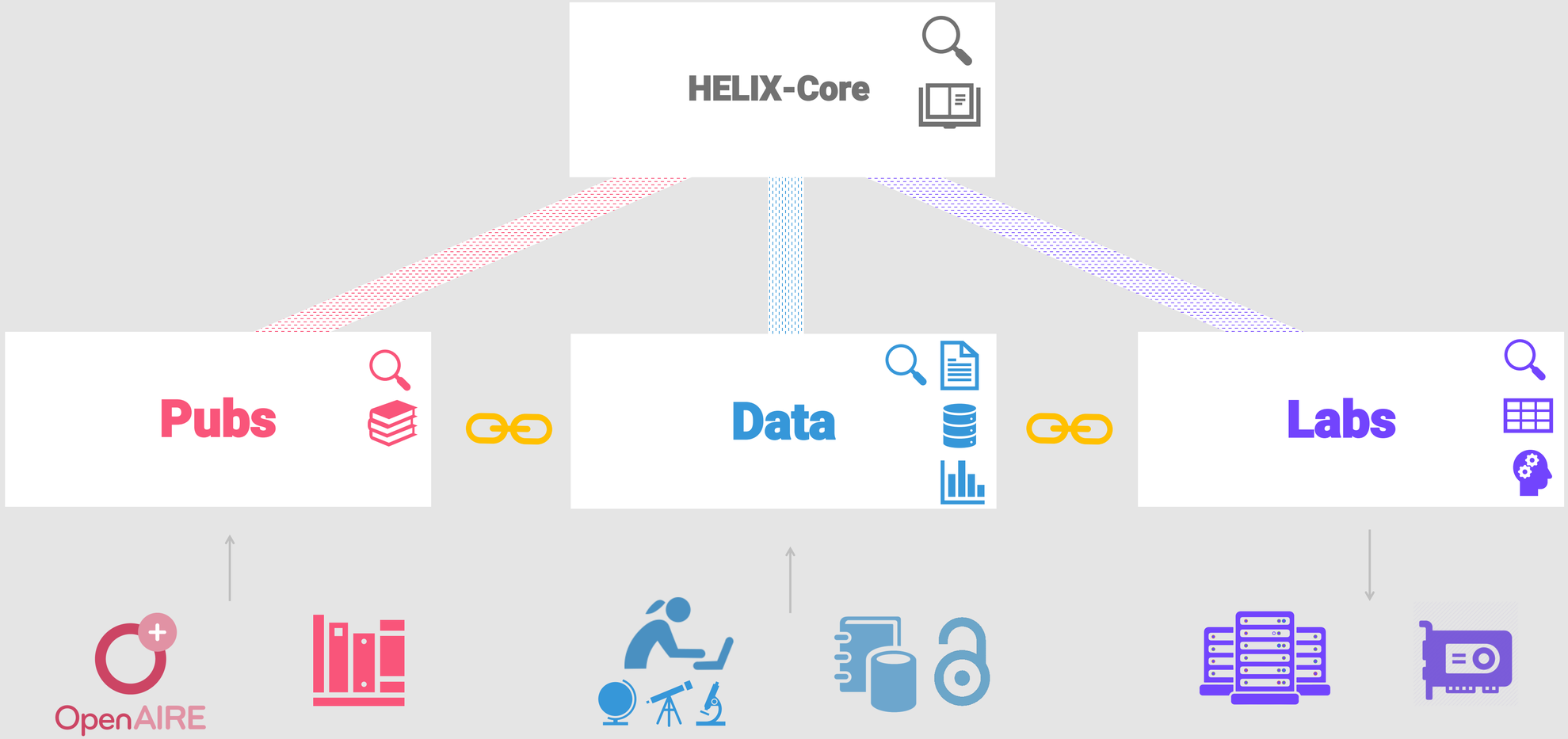
HELIX
LAB
Hellenic
Data
Service

lab.hellenicdataservice.gr

ATHENA
Research & Innovation
Information Technologies

HELIX Architecture

Birds-eye view



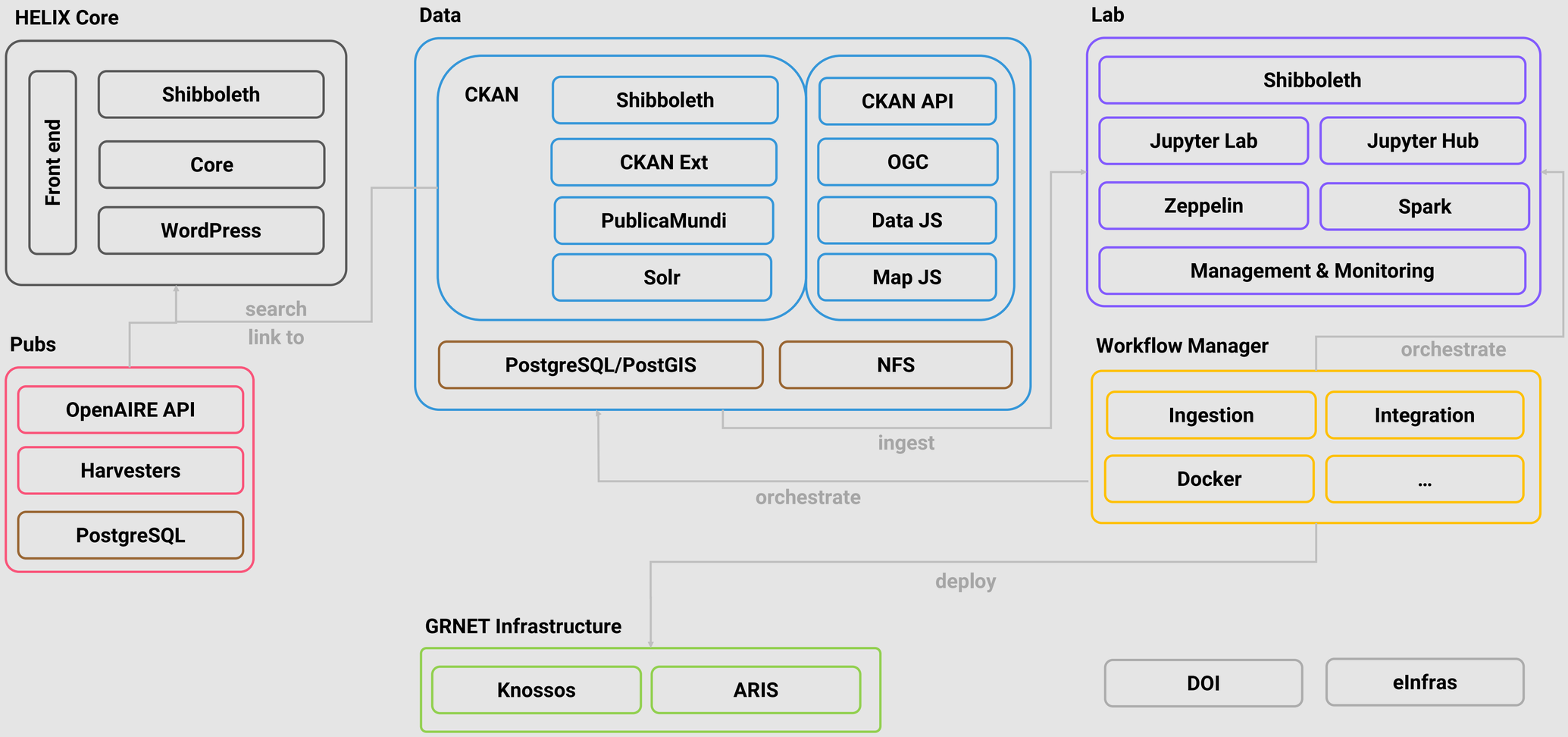
HELIX Architecture

Core Principles

- Not a single monolithic system, but an assembly of loosely coupled, highly-scalable independent components
  - Repurpose/extend software/APIs, flexible prototyping & experimentation, asynchronous parallel development tracks
  - Independently scale as/when/where needed, no single-point of failure
  - Workflow orchestration, management & monitoring via in-house Spring Boot system
- Cloud-based
  - Leverage and valorize GRNET's IaaS cloud (Knossos-okeanos) & HPC (Aris)
  - Docker-based, ported to Kubernetes
- Open Source/Open Standards
  - Exclusively open: build on existing great software, give back to the community, help others
  - Majority of software originally developed in the context of EU/national R&D projects, powering world-scale systems
- Shibboleth-based federated authentication for members of the Greek scientific community
  - Authorization handled individually by each applications by custom roles (SSO not advisable)

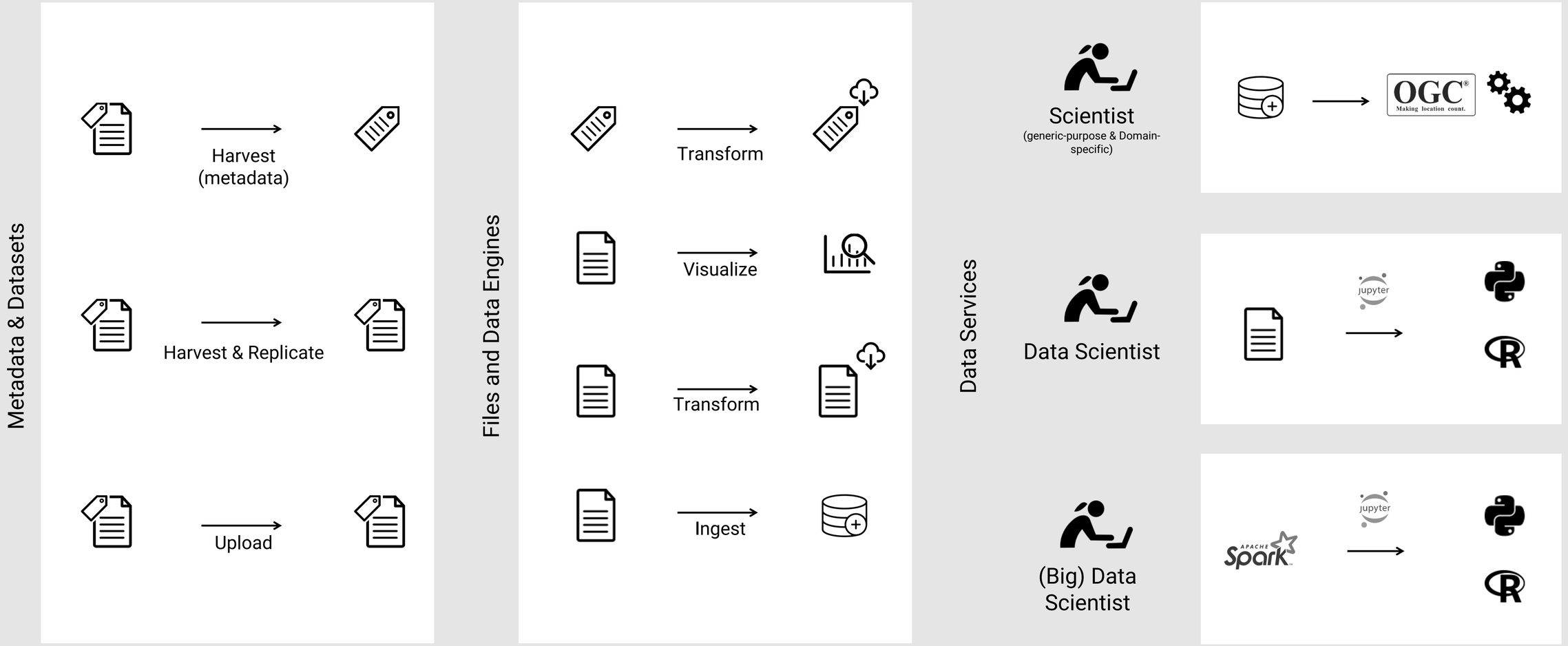
HELIX Architecture

Logical Architecture



HELIX Architecture

The data lifecycle



HELIX Architecture

HELIX-Core

- Entry point for discovering all HELIX services, resources, and guides
  - Provides the illusion of a single application (common theme)
  - Loose, API-based integration of search results for all other services (Pubs, Data, Lab)
  - Direct entry points also available (e.g., data.helix.gr)
- Custom Spring app
  - Workflow management (data ingestion, housekeeping)
  - WordPress (content management)
  - Services/code reused in other services for AAI, multilinguality support, monitoring/logging

HELIX Architecture

Publications

- **Search for Publications**
  - Harvested from EU-wide institutional, thematic, or ad-hoc repositories
  - Provide publications published from Greek S&T organizations
  - OAI-PMH v2.0, OAI-DC
- **Value added services** (under development/testing)
  - Infer data from publications (link data with pubs)
  - Analytics & KPIs
- **OA Training & support**

HELIX
PUBS
Hellenic
Data
Service

HELIX Architecture

Data

- **CKAN-based Data Catalogue & Repository** extended via multiple plugins
  - Core **CKAN v2.8** (started from v2.2, soon will port to v3.0)
  - Plugins: **CKAN + PublicaMundi** (metadata, geo) + **HELIX** (metadata/workflow)
  - Custom roles/profiles/organization structure
- **Core CKAN services & HELIX-specific services**
  - **Search, view, visualize, download**
- **Data management**
  - Dataset upload (files) **open** to all publishers (size-limited, admin QA & sanitization)
  - Multiple **replication** policies for harvested datasets
  - Automated **independent** and **asynchronous** data **ingestion** policies (files to data)

HELIX DATA
Hellenic Data Service

ATHENA Research & Innovation Information Technologies

Data Services

- Core Metadata and Standard Schemas
  - **DataCite-based** schema (default, common, simple)
  - Support for **domain-specific** metadata schemas (e.g., ISO 19131)
    - **Upload/harvest** (e.g., INSPIRE or Public Data catalogues)
    - Extensible programmatic homogenization/**mapping** (to Core), **UI generation** (editor) and **on-the-fly transformations** (all metadata files available)
- **Personal data collections** (check later, send to others, use in Lab)
- Datasets **linked** with **Data Services** (how/where to use) & **Pubs** (manual & automated via OpenAIRE)
- User **hierarchies/rights** (organization, curators, authors)
- Flexible **DMP** support (confidential, embargo)

Data as a Service

- Data catalogue & repository provided as **a Service** to **Research** Organizations, Scientific **Infrastructures**, Domain-specific **communities**, **Government/NGOs**
  - **Follow the data and the users** (e.g., high-value data, large user groups) and bring the services closer to their **needs** (e.g., domain-specific schemas and services)
  - Low-cost, low-effort, inclusive **institutional** data catalogues/repos with integrated OA support & DMP facilities
- Sub-domain in HELIX (group)
- White-labelling

HELIX Architecture

Lab

- **Open-ended collection of independent services and applications for experimenting and using data**
  - No interdependencies or single point of failure
  - Fast and simple to replace/extend services in operation
  - Service portfolio constantly expanding, with varying TRL/access levels
  - Replicate/expand the industry emerging paradigms (e.g., Azure, Google)
- **All have automated & configurable access to the repository's data**
  - Data available as files or databases/data processing frameworks
  - Flexible data availability policies per type/data set (e.g., depending on size, popularity, importance, domain, resource-utilization)

Data Science Notebooks

- Jupyter Lab/Hub (open beta)
  - **Tiered** kernel/resource access per user type (from R, to HPC)
  - **Repository** data available in user's notebooks (my data collection; minimize time/effort to discover & use data)
  - Support for under/post-grad **courses** (share data/exercises) and **research teams** (collaborative editing)
  - Constantly expanded with additional facilities & services to support **Data Science** and targeted domain needs
- Apache Zeppelin (invitational beta)
  - Notebook-like facility for **Apache Spark** clusters (Java/Scala)
  - Dedicated clusters for **Big Data** experimentation & benchmarking

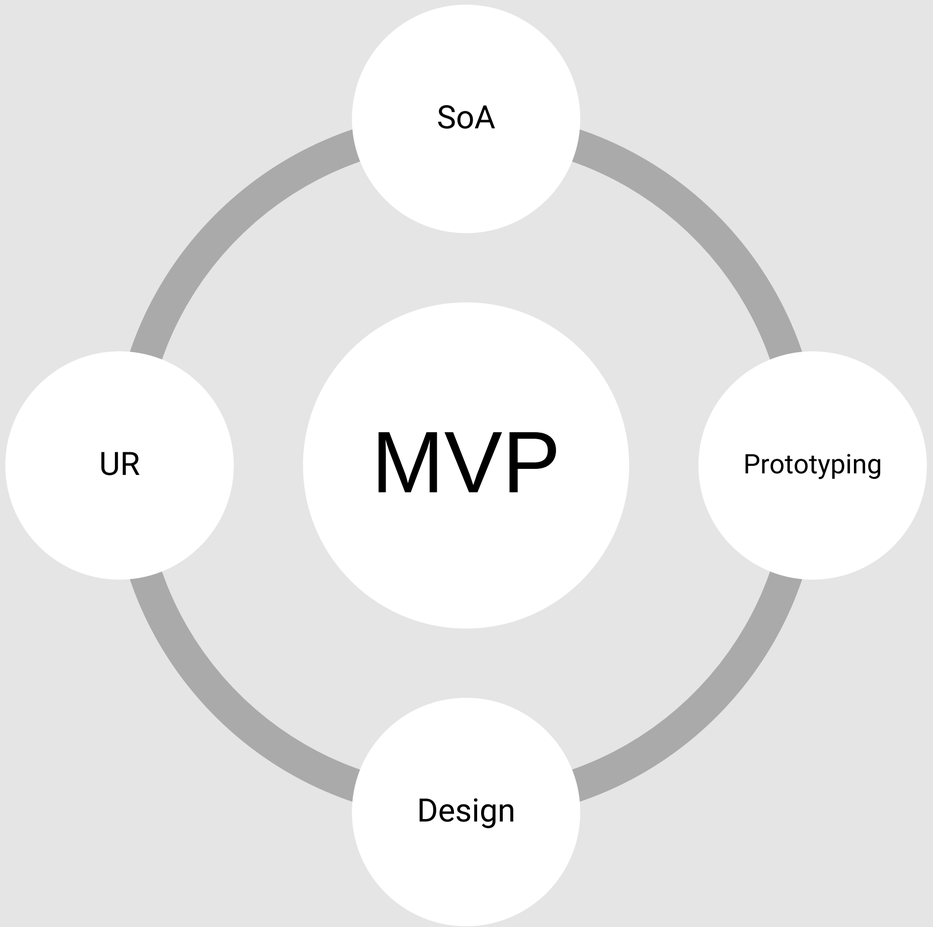
Other Services

- Interactive Data Services/widgets (evaluate & use)
  - Presentational (tables, charts, maps) for tabular data
  - File transformations (schemas/formats, CRS)
- End-points & APIs (for third system/apps)
  - OGC Services for geospatial (Catalogue, WMS, WFS, WPS-experimental)
  - Linked Open Data (SPARQL, GeoSPARQL end-points)
  - JavaScript Data API (simple filter/SQL-type queries over tabular data)
  - JavaScript Mapping API (custom standalone/embeddable maps)

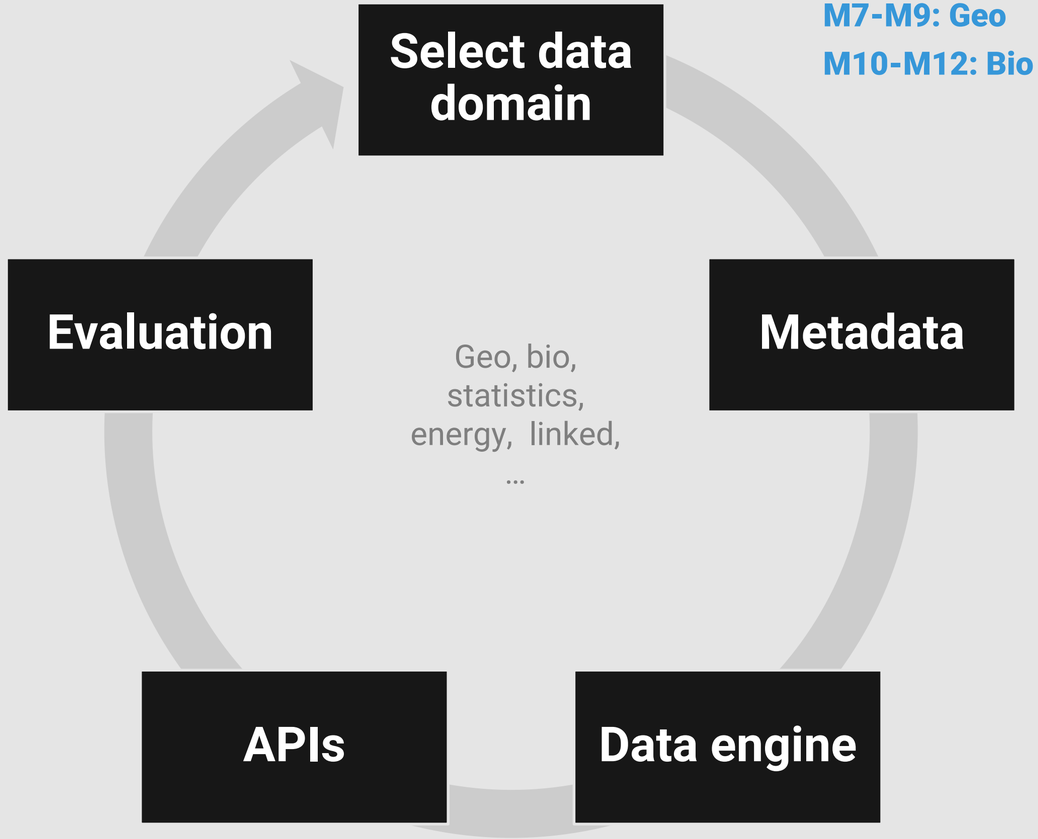
HELIX Architecture

Data-drive & Agile development

M1 – M6



M7 – M24



M7-M9: Geo
M10-M12: Bio

ATHENA
Research & Innovation Information Technologies

DEMO

HELIX

DATA

Hellenic
Data
Service

Find, view, and use open scientific data

HELIX

PUBS

Hellenic
Data
Service

Discover and share open scientific publications

HELIX

LAB

Hellenic
Data
Service

Learn, experiment, and build with data

Thank you