GreenDIGIT: [Green]er Future [DIGIT]al Research Infrastructures

Artificial Intelligence and Environmental Sustainability: The GreenDIGIT Perspective

GreenDIGIT

Yuri Demchenko, UvA

GreenDIGIT Project

e-IRG2025 Meeting, 24-25 June 2025, Poznan

Outline

- GreenDIGIT project scope and goals for RI/ESFRI community
  - Founding EU ESFRI Research Infrastructures: SoBigData, EBRAINS, SLICES, EGI
- Shared Responsibility Model for Sustainability
  - Sustainability aspects in Digital Infrastructure: From researcher/experimenter to developer and operator
- Sustainability by Design: Architecture Framework and Energy/Impact Monitoring Infrastructure
- Monitoring and metrics for Energy Efficiency optimisation and Environmental Impact
- Research Infrastructure Lifecycle Model (RILM) and EU Regulations Compliance
- Standardisation on Environmental Sustainability and technical requirements

Funded by the European Union.
Grant ID: 101131207

Sustainability Aspects: Energy Efficiency – Decarbonisation – Environmental Impact

- **Energy Efficiency of Digital Infrastructures**:
  - **Definition**: This refers to optimising digital infrastructures to consume as little energy as possible for a given workload or service. It's about achieving more computational or storage results with less energy input.
- **Decarbonization of Digital Infrastructures**:
  - **Definition**: This specifically targets the reduction of carbon emissions associated with the operation and maintenance of digital infrastructures.
- **Reducing Environmental Impact of Digital Infrastructures**:
  - **Definition**: This is a more comprehensive consideration of the various ways digital infrastructures might affect the environment, going beyond just energy consumption and carbon emissions.

Architecture, Design,
Recommendations

Operation,
Monitoring, KPI

Lifecycle, Policy,
Training

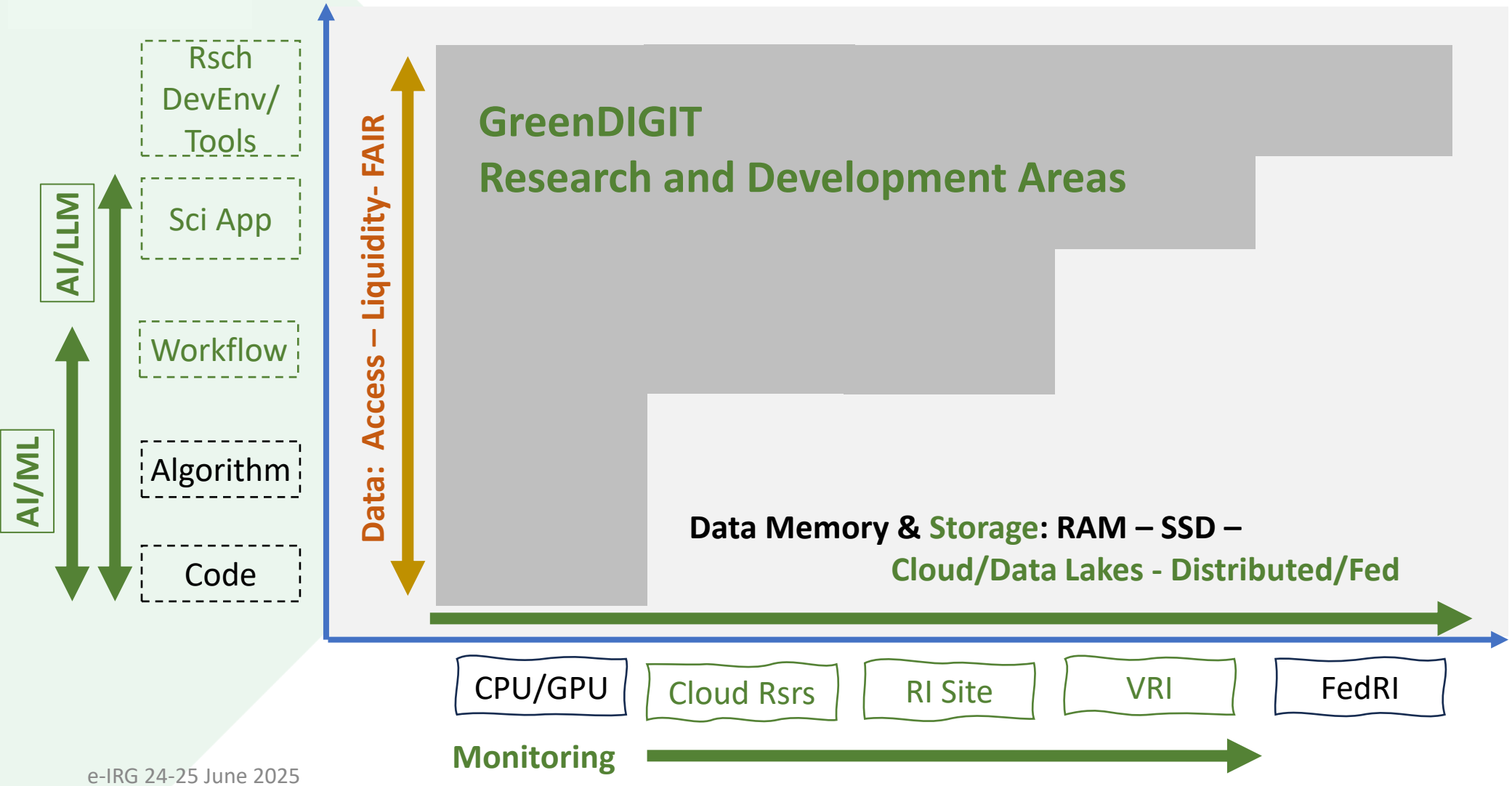
AI/ML/LLM and Research Practices/Use

Energy Efficiency and Environmental Impact

- Algorithms and workflow optimisation
- Effective data management through the ML lifecycle
  - Extended data model enriched with semantics, metadata, vocabularies
- **Infrastructure monitoring and metrics optimisation**
  - **Smart energy sources management**
- User side: LLM input and output caching and embedding

- **Research on ML/Data Analytics**
  - Common algorithms and models for different scientific domains and data – require algorithms profiling and efficient management (data liquidity)
  - Access to developer controlled computation platforms
- **AI/ML enabled research environment** to assist in domain-specific research and hypothesis testing
  - Expected growing role of LLM – LLM infrastructure to be optimised for operation/inference and the creation of domain specific knowledge base(s)
    - Instrument for *conversation research* methodology
- AI4EOSC Flagship project: Focusing on AI/ML tools and EOSC targeted services
  - Software stack, composite serverless AI platform (cross-discipline)
  - Data and models repository and management – extending FAIR principles for actionable data management (beyond data storage and sharing)
    - Rich data models (beyond textual and relational) including semantics, ontology, metadata, vocabulary

GreenDIGIT Scope Areas: RI & Sustainability AI/ML Aspects



Multiple RI dimensions

- Code – Algorithms – Workflow
- CPU/GPU – Cloud – Site – VRI/VRE

Data dimensions

- Memory – Storage – Distributed
- Access – Liquidity (Data & Model) - FAIR

GreenDIGIT Approach: Architecture Consolidated

- **Architecture** is a way to coordinate and synchronise/unite
  - Developers of Infrastructure and Applications
  - Operators
  - Users/Researchers
  - Policy and decision makers
  - Refer to TOGAF architecture principles (as an approach accepted by the majority of businesses)
- A basis for linking standards and regulations to architecture functional components
  - Ensure compliance of the designed/developed RI and services (including for audit)

GreenDIGIT implementation and current outcomes

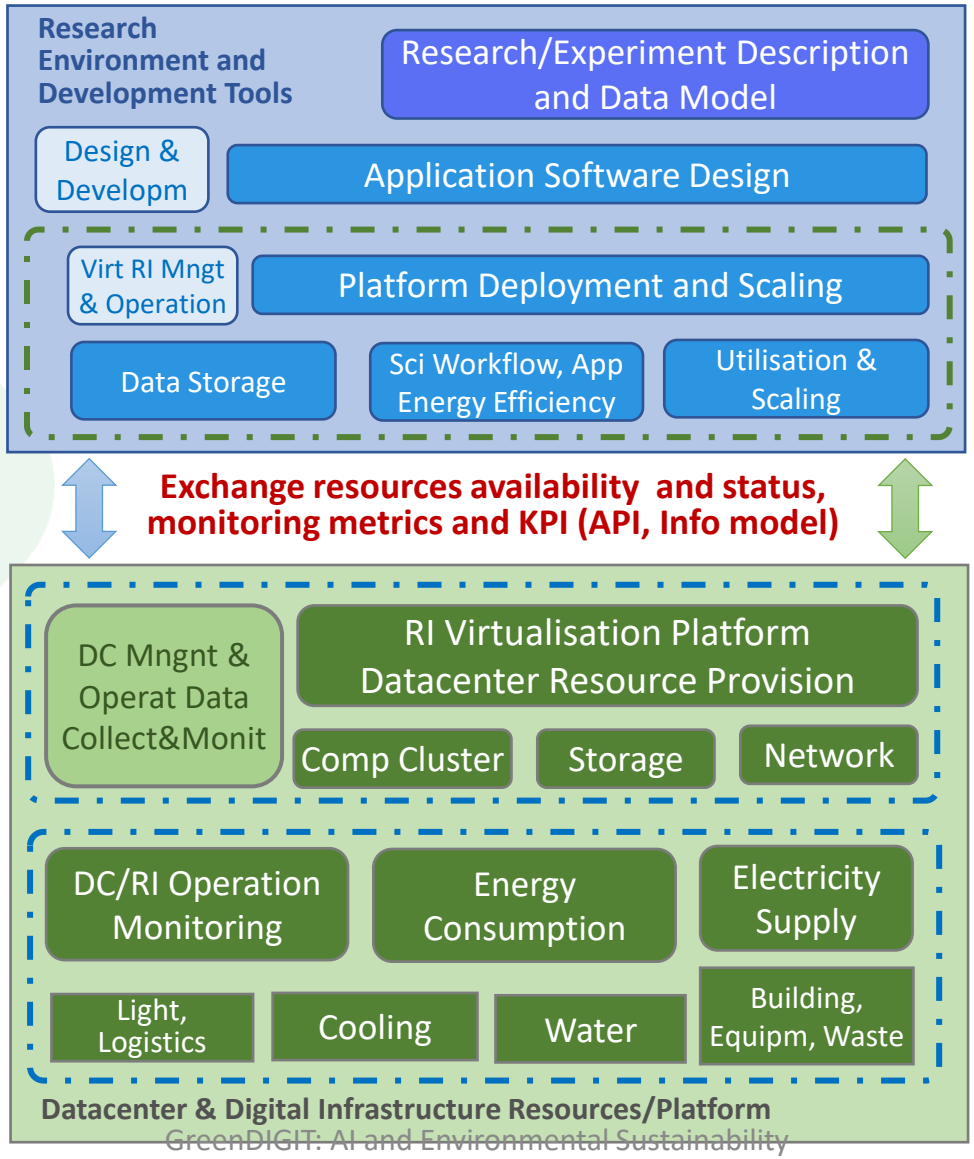
- Landscape analysis (2024)
- Architecture Framework for Sustainability by Design
- **Shared Responsibility Model for Sustainability**
- Metrics and Monitoring Architecture
  - Researcher Tools for Energy and Environmental Impact Awareness
- RI Lifecycle Model and Impact Assessment Framework
- Green Competences Framework and Thematic Tutorial Groups

Shared Responsibility in Sustainability – Reflecting Operational and Management Aspects and Roles

GreenDIGIT

Users responsible for sustainability ***on*** the RI

Providers responsible for the sustainability ***of*** the RI



Project/Researcher Responsibility:
- Applications Development,
- Deployment, Jobs submission & Energy Efficiency optimisation
- Operation, Workflow execution
- Energy usage and KPI monitoring

Experimenter/Researcher Demand:
- Access to datacenter (internal) monitoring data and custom configuration
- Energy usage and KPI & metrics monitoring

Provider/Operator Responsibility:
- Operation of Research Infrastructure or Datacenter,
- Monitoring Energy: environmental impact metrics and KPI
- Waste
- Lifecycle and evolution

RI Sustainability by Design Components/Aspects: Motivated by the Shared Responsibility Model

Users responsible for sustainability ***on*** the RI

Providers responsible for the sustainability ***of*** the RI

Researcher/ Project Responsibility:
Applications Development, Energy usage and KPI monitoring

Sustainability by Design Pillars/Challenges

RI/Datacenter Provider/Operator Responsibility:
Monitoring Energy and environmental impact, metrics and KPI

Pillar 1 – Layered and Modular Architecture for Sustainable Digital Infrastructur
- Functional components, Layers, API, Requirements

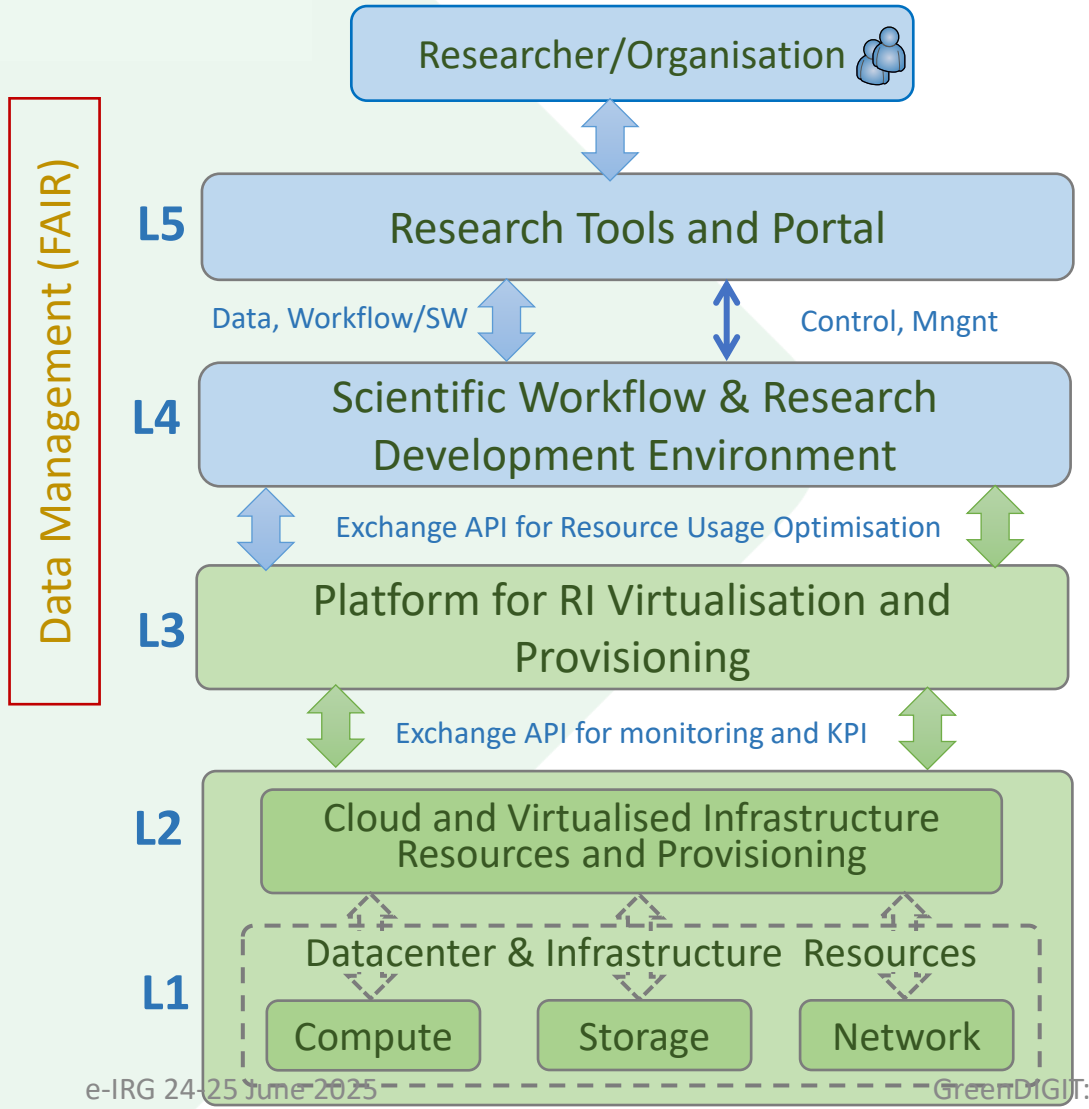
Pillar 2 - Software and Scientific Applications Design Patterns for Energy Efficiency and Environmental Impact Reduction
- Green aware API including necessary energy, performance, environment information

Pillar 3 - Lifecycle-Oriented Sustainability Management of Research Infrastructures and Scientific Applications
- RI lifecycle stages (concept, design, development, deployment, operation, decommissioning) and scientific workflow and research data

(!) Pillar 4 – Interoperable/common Information Models for **Sustainability Metrics and Monitoring**
- Including **Requirements, KPI, Metrics** + FAIR for Sustainability

Pillar 5 - Sustainable Data Management and Energy-Efficient Storage Systems

RI Sustainability by Design Components/Aspects:
5 Pillars Architecture Framework for Sustainability by Design



Dev Tools, IDE/SDK, Advice/ Assess

Pillar 1 – Layered and Modular Architecture for Sustainable Digital Infrastructur

- Functional components, Layers, API, Requirements

Pillar 2 - Software and Scientific Applications Design Patterns for Energy Efficiency and Environmental Impact Reduction

- Green aware API including necessary energy, performance, environment information

Pillar 3 - Lifecycle-Oriented Sustainability Management of Research Infrastructures and Scientific Applications

- RI lifecycle stages (concept, design, development, deployment, operation, decommissioning) and scientific workflow and research data

(!) Pillar 4 – Interoperable/common Information Models for **Sustainability Metrics and Monitoring**

- Including **Requirements, KPI, Metrics** + FAIR for Sustainability

Pillar 5 - Sustainable Data Management and Energy-Efficient Storage Systems

Architecture Framework Pillars – Technical – Pillars 1-5

Pillar 1 – Layered and Modular Architecture for Sustainable Digital Infrastructure

Pillar 2 - Software and Scientific Applications Design Patterns for Energy Efficiency and Environmental Impact Reduction

Pillar 3 - Lifecycle-Oriented Sustainability Management of Research Infrastructures and Scientific Applications

Pillar 4 – Interoperable/common Information Models for Sustainability Metrics and Monitoring

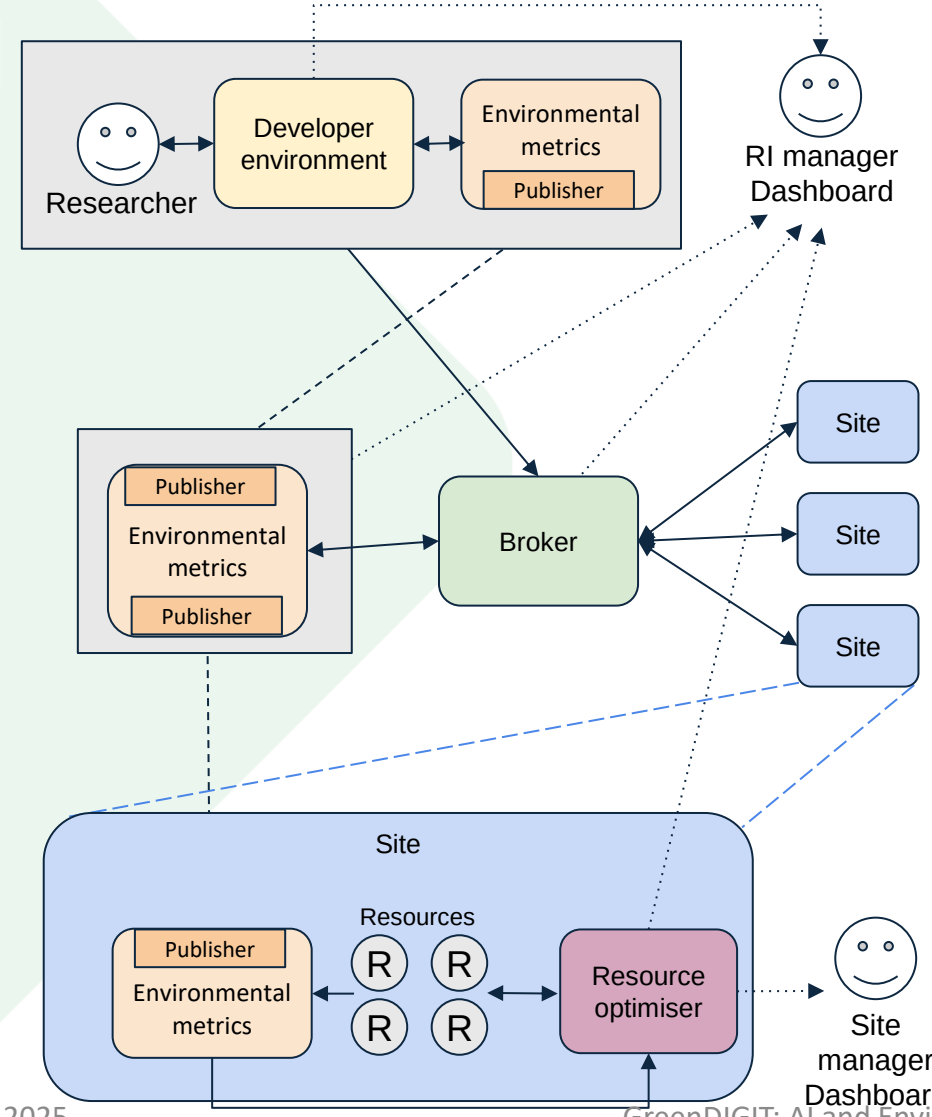
Pillar 5 - Sustainable Data Management and Energy-Efficient Storage Systems

Architecture for Metrics and Optimisation

Level 1: RDE, Experiment

Level 2: Infrastructure Composition/ Scheduling

Level 3: Site, Resources Management



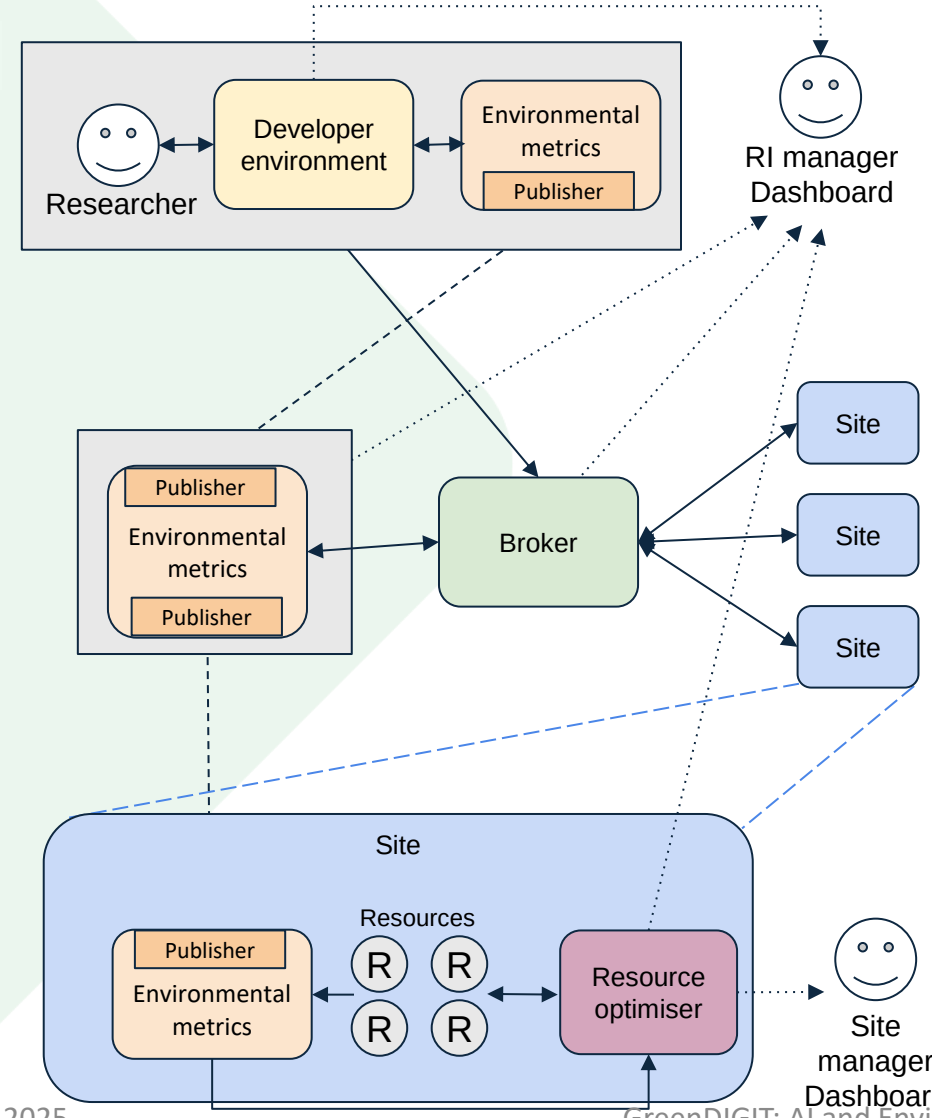
- **Level 1 – Developer tools for energy efficient software design**
  - Optimised code; Prevent re-execution of already conducted experiments
- **Level 2 – Energy efficiency aware schedulers**
  - Reshuffle jobs based on site carbon intensity attributes
- **Level 3 – Energy efficiency aware resources manager**
  - Delay execution based on electricity cost, regulations, carbon intensity, resources turning on-off
- **Metrics Infrastructure**
  - Gather relevant environmental impact metrics
  - Serve metrics to decision making systems (schedulers, researcher tools, VREs)
  - Present statistics and monitoring in dashboards

Architecture for Metrics and Optimisation

Level 1: RDE, Experiment

Level 2: Infrastructure Composition/ Scheduling

Level 3: Site, Resources Management



To be adopted and piloted on the EGI federated computing infrastructure

- **Level 1 – Developer tools for energy efficient software design**
  - Optimised code; Prevent re-execution of already conducted experiments
- **Level 2 – Energy efficiency aware schedulers**
  - Reshuffle jobs based on site carbon intensity attributes
- **Level 3 – Energy efficiency aware resources manager**
  - Delay execution based on electricity cost, regulations, carbon intensity, resources turning on-off
- **Metrics Infrastructure**
  - Gather relevant environmental impact metrics
  - Serve metrics to decision making systems (schedulers, researcher tools, VREs)
  - Present statistics and monitoring in dashboards

Digital RI Lifecycle Stages
and European Environmental Regulations Compliance (based on ESFRI Roadmap 2026 Guidelines)

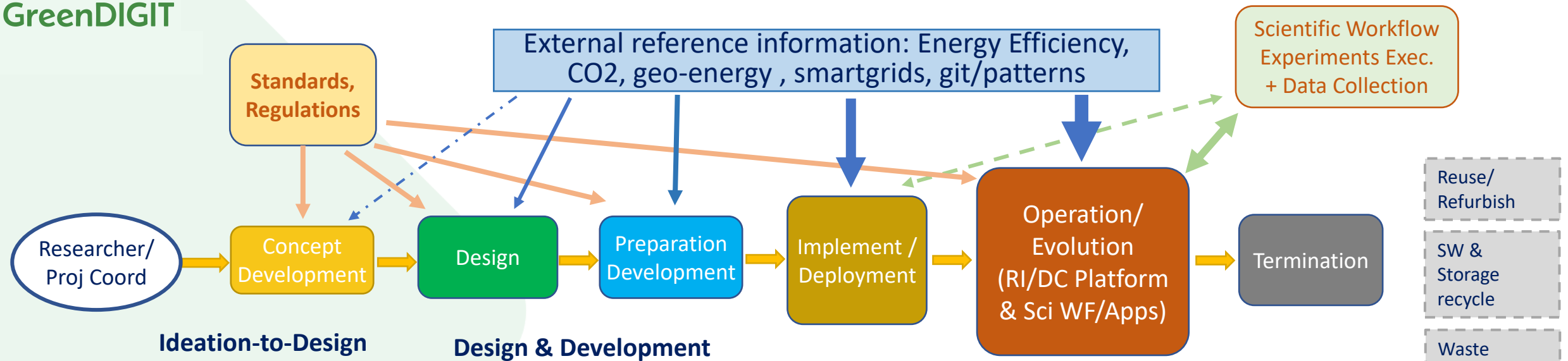
RI Lifecycle stages (ESFRI compliant)

- Concept Development
- Design
- Development
- Deployment and Pre-operation
- Operation & Evolution
- Termination/Decommissioning

Set of EU Regulations on Environmental Compliance

- ESRS/CSRD (European Sustainability Reporting Standard)
  - Required by ESFRI Roadmap 2026
- Ecodesign for Sustainable Products Regulation (ESPR)
- Energy Efficiency Directive (EED)
- Waste Electrical and Electronic Equipment (WEEE) Directive
- Lifecycle Analysis (LCA and ISO 14040/ISO 14044)

GreenDIGIT: RI Lifecycle Stages, Activities and Factors



Ideation-to-Design
- SotA
- Users/Actors
- Use cases
- Requirements Engineering
- Architecture
- Standards/Regulations

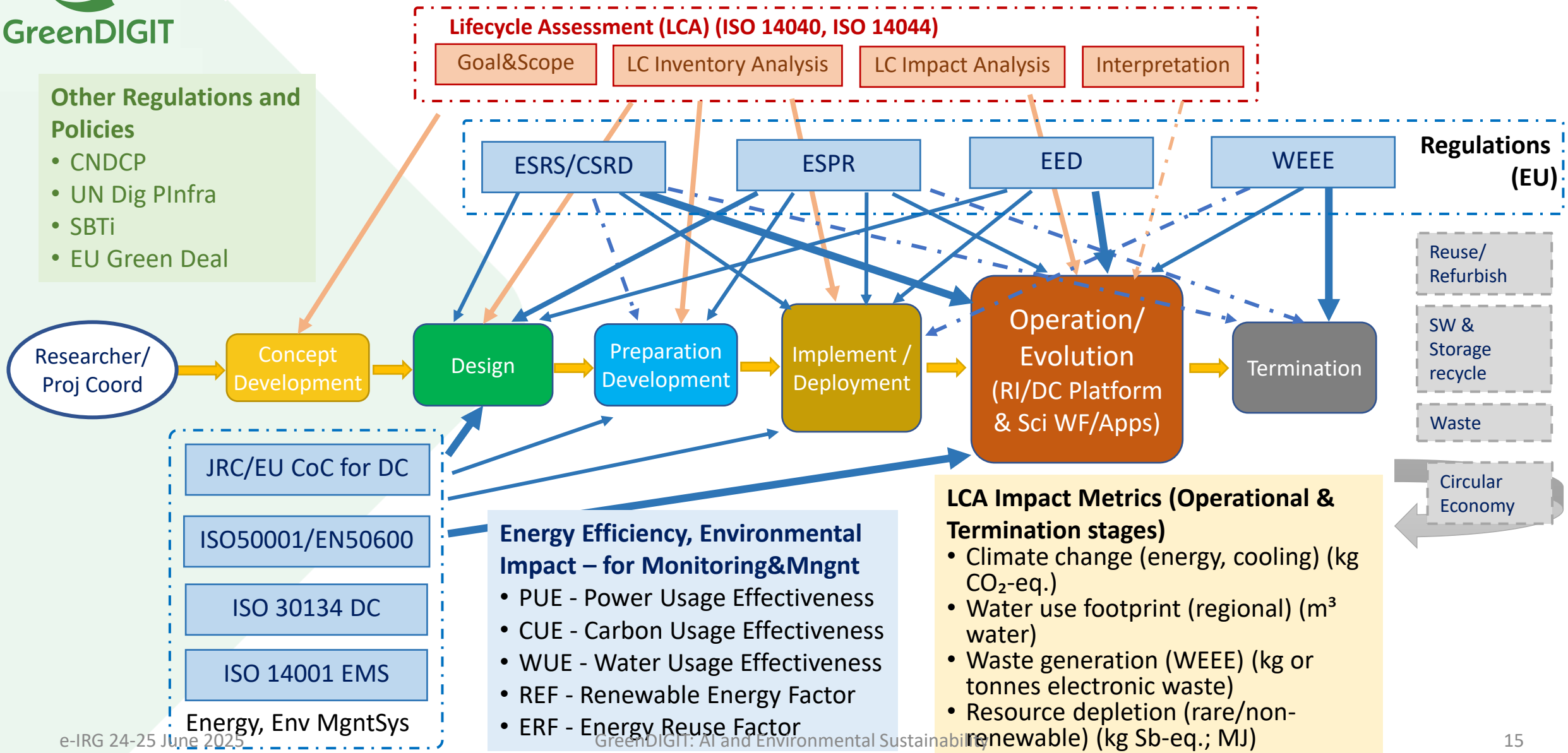
Design & Development
- Architecture & Design principles
- DevOps & DevSecOps
- Info data model
- Git, patterns, metrics, schemas, Repository
- API/WebAPI
- Testing, Logbook
- Project Management
- Tech Documentation
- Report, Compliance

Implementation
- CI/CD
- SW/HW & Implement Platform/Cloud
- Feedback to Dev

Operation Processes
- DevOps/SRE/DevGreenOps
- Monitoring, Metrics, KPI
- Maintenance
- Update/Upgrade/Procurement
- Audit, Certification
- Electricity, Water, Buildings

Research Infrastructure Lifecycle
- All stages and processes to be in compliance with standards and regulations
- Shared Responsibility model and User /operator tools Continuous monitoring

GreenDIGIT: RI Lifecycle Stages, Regulations, Impact Factors



GreenDIGIT: European Policies and Regulations for Development and Compliance

European Strategy Forum on Research Infrastructure (EFRI) Roadmap 2026:

- Environmental Sustainability of the European/ESFRI Research Infrastructures
- Compliance with ESRS (European Sustainability Reporting Standard) compliant with EU CSRD (Corporate Sustainability Reproting Directive)

European Code of Conduct for Data Centers

- JRC BCP Sections 2024 Best Practice Guidelines for the EU CoC on Data Centre Energy Efficiency
- Commission Delegated Regulation (EU) 2024/1364 of 14 March 2024 for Data centers

ISO/EN/ITU-T Standards Compliance – Basis for Certification and Audit

- ISO 50000, ISO 30134, ISO 14002
- ISO 14040, ISO 14044 – LifeCycle Analysis (LCA)
- EN 50600-4-1 and EN 50600-4-2 Datacenter and supporting infrastructure

ESFRI Roadmap 2026: Research Infrastructure Lifecycle Model and Requirements

- RILM stages – Compliant with ESFRI Lifecycle stages
- Reporting on Energy Efficiency and Environmental Sustainability
  - ESRS/CSRD and ESFRI RM2026
- Link to standards and regulations – Digital RI related
  - ESFRI Roadmap 2026 – To be analysed due to ESRS reporting specifics on Groups **E1 - Climate**, E2 - Pollution, **E3 – Water** and marine resources, E4 – Biodiversity and ecosystems, **E5 – Resource use and circular economy**
  - E1 group compliance:
    - E1-1, E1-2, E1-3 – Climate mitigation target, policies, actions; E1-4 - targets; E1-5 – Energy consumption and mix; E1-6 – Total GHG emission; E1-9 – Potential financial effect of measures
  - Compliance with LCA
- Roles and activities

Additional Information

- AI/ML/LLM solutions to improve energy efficiency at algorithmic, workflow, domain specific applications layers
- GreenDIGIT Architecture based methodology
- Green competences and training for researchers and RI operators

ML/AI Algorithms Optimisation for Energy Efficiency

Designing Energy-Efficient Training and Inference Algorithms/Pipelines

- **Optimising models** training to avoid wasted resources and time:
  - Embracing mixed-precision training, gradient accumulation, adaptive learning rates and early stopping
- **Training or algorithmic side**, considering efficient model architectures like DistilBERT or adopting parameter-efficient fine-tuning strategies such as LoRA and Adapter Layers that may provide an energy efficiency effect
  - Train only a small portion of the model, cutting compute needs to achieve energy and time reduction
- **Inference side**, models can be quantised to smaller data formats (e.g., INT8 or INT4)
  - Model pruning, knowledge distillation, and conditional computation further streamline inference workloads, letting models make smart decisions about when and how much to compute based on the complexity of the input.

ML/AI Infrastructure Optimisation

Optimising Infrastructure for Sustainable AI shifting toward smarter, cleaner, and more adaptable compute environments.

- **Carbon-aware scheduling**, where workloads are timed to run when renewable energy is most available or carbon intensity is lowest.
  - Tools: Carbon Aware SDK being explored by tech giants such as Microsoft and Google.
- **Hardware-aware placement** ensures that each job is assigned to the most energy-efficient compute resources available - whether a choice of a GPU, TPU, or low-power accelerator like Jetson Nano.
  - Inference environments employing autoscaling and serverless computing platforms (such as AWS Lambda or Azure Functions)
- To **maximise reuse and avoid redundancy**, organisations are using on centralised model versioning (via MLFlow or Hugging Face Hub) and data preprocessing pipelines that cache tokenised inputs.
  - This helps eliminate repeated computations, especially in multi-tenant systems handling similar requests.

User Serving Infrastructure and Solutions

Intelligent Request Handling (at scale) is key to both performance and sustainability.

- The techniques include caching and reusing embeddings and intermediate states (like KV caches in transformers), large or complex request splitting to break a prompt into smaller sub-queries and using response combination methods, where partial outputs are merged using summarisation, voting, or Retrieval-Augmented Generation (RAG).
    - Tools like LangChain and AutoGen are providing usable solutions.
    - Platforms like vLLM and Hugging Face's Text Generation Inference (TGI) push this even further by introducing dynamic batching, where similar queries from different users are bundled into a single inference pass

Building Domain-Specific Contexts with Cached Intelligence

- The new domain oriented strategy involves embedding and caching responses at the user or group level. Over time, this will allow building a rich, domain-specific knowledge layer that reduces the need to recompute answers and enhances semantic relevance.
- Systems like Redis, Faiss, and SentenceTransformers are commonly used to support these contextual layers. They allow GenAI/LLM systems to "remember" prior interactions offering not only better performance but also a reduction in energy demand.
- Industry leaders like OpenAI, Google, Meta, Hugging Face, Azure, and AWS are already implementing many of these techniques

GreenDIGIT Architecture Definition Methodology

- General view on the RI Ecosystem Optimisation and Green IT
  - **Horizontal, vertical, lifecycle**
  - RI Operators and Researchers
  - RI continuum: (Research Object) – Sensor - (RAN) – Edge – Cloud – Workflow - Researcher
- Sustainable architecture design principles
  - As a basis for modelling and metrics for RI infrastructure operation and optimisation
  - **Shared Responsibility Model**: RI provider/operator and Researchers/Projects
  - **Sustainability by design** – A novel concept to be introduced to address different aspects and stages
- Linkage with existing Standards and Regulations to ensure Sustainable Architecture Design principles support compliance with the standards, regulations and audit
  - To provide the opportunity for RI/datacenter operation (and design) optimisation (through the whole lifecycle)
- System Engineering and Design (thinking) approach in Green research and technologies
  - **Sustainable (Durable) Architecture Design Principles**

RI/Systems Sustainability by Design Components/Aspects

- Architecture for Sustainability by Design
  - Functional components, layers, API, Requirements **<= Sustainable Architecture Design Principles (SADP)** + Energy efficiency tactics (VU)
  - **RI continuum: (Research Object) – Sensor - (RAN) – Edge – Cloud – Workflow - Researcher**
- Software and application components that can be optimised during design and controlled during operation
  - Corresponding optimisation model to be proposed
  - Supported with VRE and IDE with sustainability awareness
  - Sustainability design patterns (software, application, middleware, infrastructure) **<= Architecture styles**
- Link and interaction between components via APIs
  - **Green aware API** including necessary energy, performance, environment information
  - Based on well defined Information Model
- Common information/data model and metadata (naming)
  - Including Requirements, KPI, Metrics
  - Verified with existing standards
  - Compliant with PUE and other KPI
- Relations between the system/RI and software sustainability aspects to be defined
  - System related aspects: Lifecycle, Operation and Governance, metrics on energy and GHG
- RI and applications lifecycle
  - Scientific workflow and research data lifecycle to be aligned with sustainability policy and monitoring

Architecture Framework Pillars – Technical – Pillars 1-5 - Details

Pillar 1 –Layered and Modular Architecture for Sustainable Digital Infrastructure

Benefits of layers definition:

- Benefit of layered architecture is that it makes it easier to support dynamic energy-aware orchestration across layers (actually using the same API/infomodel)

Pillar 2 - Software and Scientific Applications Design Patterns for Energy Efficiency and Environmental Impact Reduction

Goal and benefits:

- Primarily focus on scientific workflows and user-driven applications for energy efficiency awareness.
- Allows to benefit from (general) green software engineering and (research focused) reproducible (experimental) science.
- Links application design to infrastructure energy-aware execution.

Pillar 3 - Lifecycle-Oriented Sustainability Management of Research Infrastructures and Scientific Applications

Goal and benefits:

- Lifecycle stages define different actors and requirements to RI components and services
- Incorporate **LCA methodologies** to evaluate and mitigate environmental impacts throughout the RI's operational life

Pillar 4 – Interoperable/common Information Models for Sustainability Metrics and Monitoring

Goal and benefit:

- Ensures semantic interoperability for API definition and machine-actionability.
- Simplifies infrastructure metrics telemetry with regulatory reporting needs.
- Cross-infrastructure benchmarking and optimization

Pillar 5 - Sustainable Data Management and Energy-Efficient Storage Systems

Goal and benefits:

- Efficient data management practices contribute to the overall sustainability of RIs
- Implement tiered storage architectures and data compression techniques to reduce energy consumption associated with data storage and retrieval
- Linking FAIR data principles for sustainability metadata: Important for RI/EOSC liaison but needs some additional definitions

Architecture Framework Pillars (Additional) – Pillars 6-7 – Focused on ESFRI RM2026

Additional pillars to extend the sustainability architecture of future DRIs and scientific applications and address critical gaps and emerging needs in areas such as governance and policy integration, which are essential to achieving environmental sustainability goals through the DRI lifecycle (ensuring alignment with ESFRI Roadmap 2026):

Pillar 6 – Governance, Policy Integration, and Compliance Framework

- Establish clear governance models that incorporate sustainability objectives, ensuring accountability and continuous alignment with ESFRI's evolving requirements.
- Provide a basis for interoperability and energy-aware infrastructure federation, adopt interoperable standards for data exchange across federated infrastructures.

Pillar 7 – User Awareness, Behavior, and training for Sustainability by Design

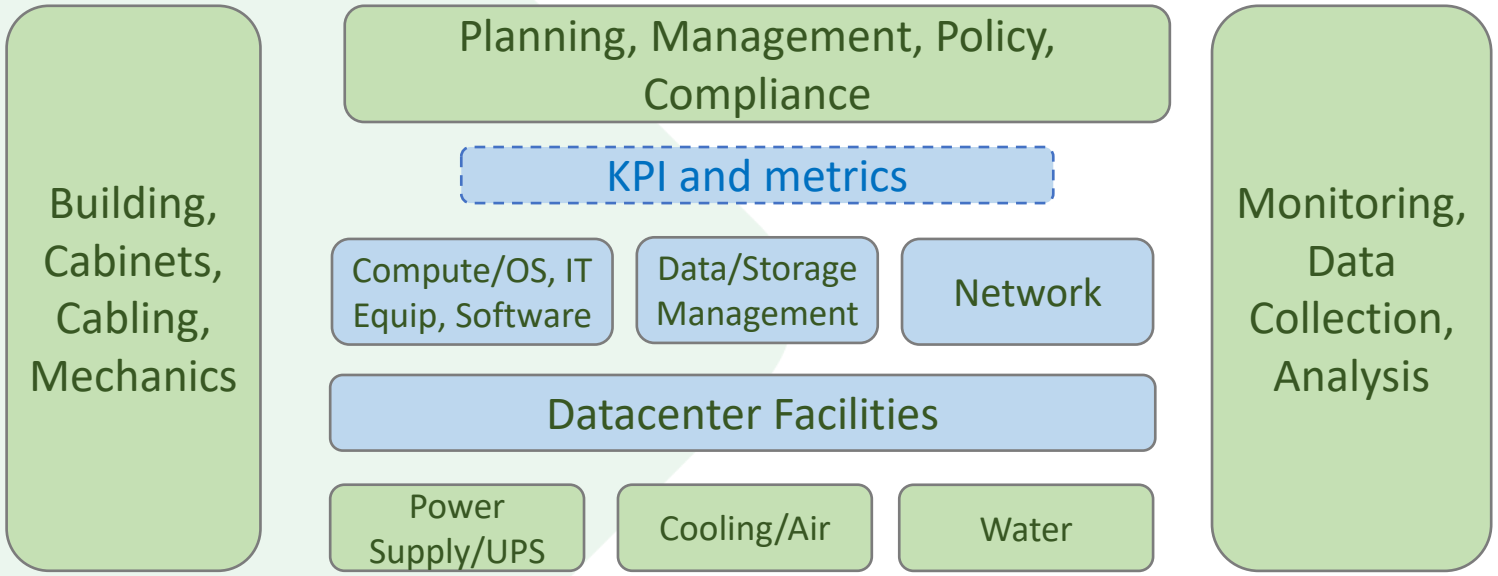
- Develop training programs and user interfaces that encourage energy-efficient behaviors and provide feedback on the environmental impact of user activities

Policies and Regulations for Energy Efficiency and Environmental Sustainability (GreenDIGIT Milestone Report MS4/MS8.1 – 50 pp.)

- Policies related to environmental sustainability and energy efficiency UN Sustainable development goals (SDG)
  - UN Digital Public Infrastructure for Environmental Sustainability
  - SBTi: The Science Based Targets Initiative
  - The European Green Deal
  - ESFRI Roadmap 2026 Public Guide
  - Climate Neutral Data Centre Pact (CNDCP)
- Regulations and Directive (to be checked)
  - JRC Data Center Code of Conduct
  - EU Delegated Regulation (EU) 2024/1364 of 14 March 2024
  - Corporate Sustainability Reporting Directive (CSRD) and European Sustainability Reporting Standards (ESRS)
  - Ecodesign for Sustainable Products Regulation (ESPR)
  - Energy Efficiency Directive (EED)
  - Waste Electrical and Electronic Equipment (WEEE) Directive
  - Lifecycle Analysis (LCA)
  - Environmental Management Standards (EMS)
    - Energy Management Systems ISO 50001, ISO 50002, ISO 14001, ISO 30134
    - European standards series EN 50600: EN 50600-4-1 and EN 50600-4-2 Datacenter and supporting infrastructure

Best Practice for the EU CoC on Data Centre Energy Efficiency: Main Infrastructure components – RI/Datacenter Operator Perspective

Also supported in the EC Delegated Regulation (EU) 2024/1364 of 14 March 2024



Areas of responsibility/ management

- Physical building
- Mechanical and electrical plant
- Data floor and air floor
- Cabinets and cabinets airflow
- Metrics and operation measurement points
- **IT equipment**
- **Operating systems and virtualisation**
- **Software**
- Business Practices

Different roles of participants/ stakeholders

- Operator
- Colocation provider
- Colocation customer
- Managed services provider
- Managed services provider in colocation space

[ref] https://e3p.jrc.ec.europa.eu/sites/default/files/documents/publications/jrc136986_2024_best_practice_guidelines.pdf