

# Sensitive Data Handling in the Social Sciences

Where do I start with FAIRification  
of sensitive data?

FAIR-IMPACT Workshop

25th June, 2024

Hervé L'Hours UK Data Service/CESSDA

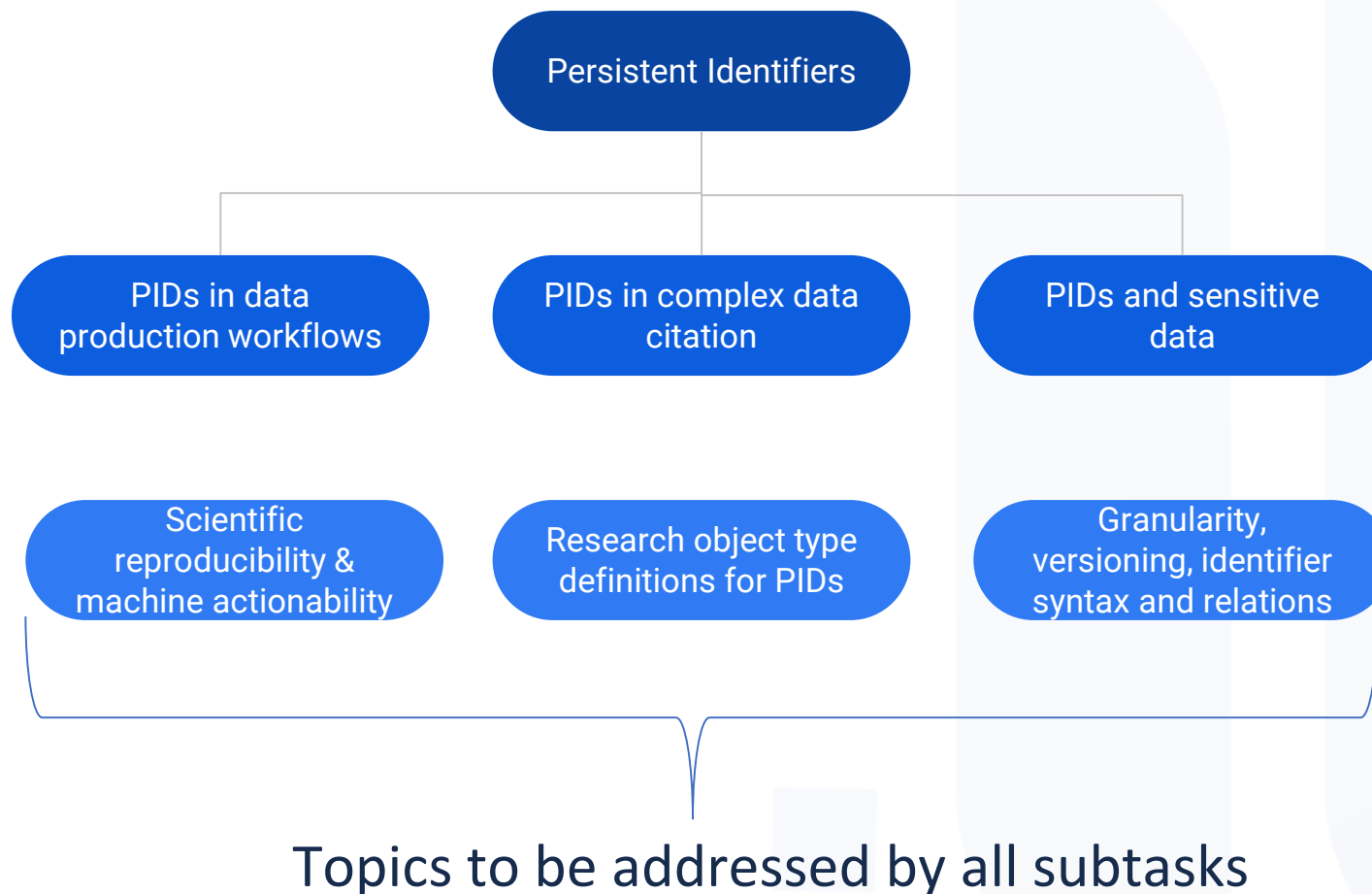
WP3 on Persistent Identifiers

# Use case topics & structure

Work Package

Subtasks

Cross-cutting themes

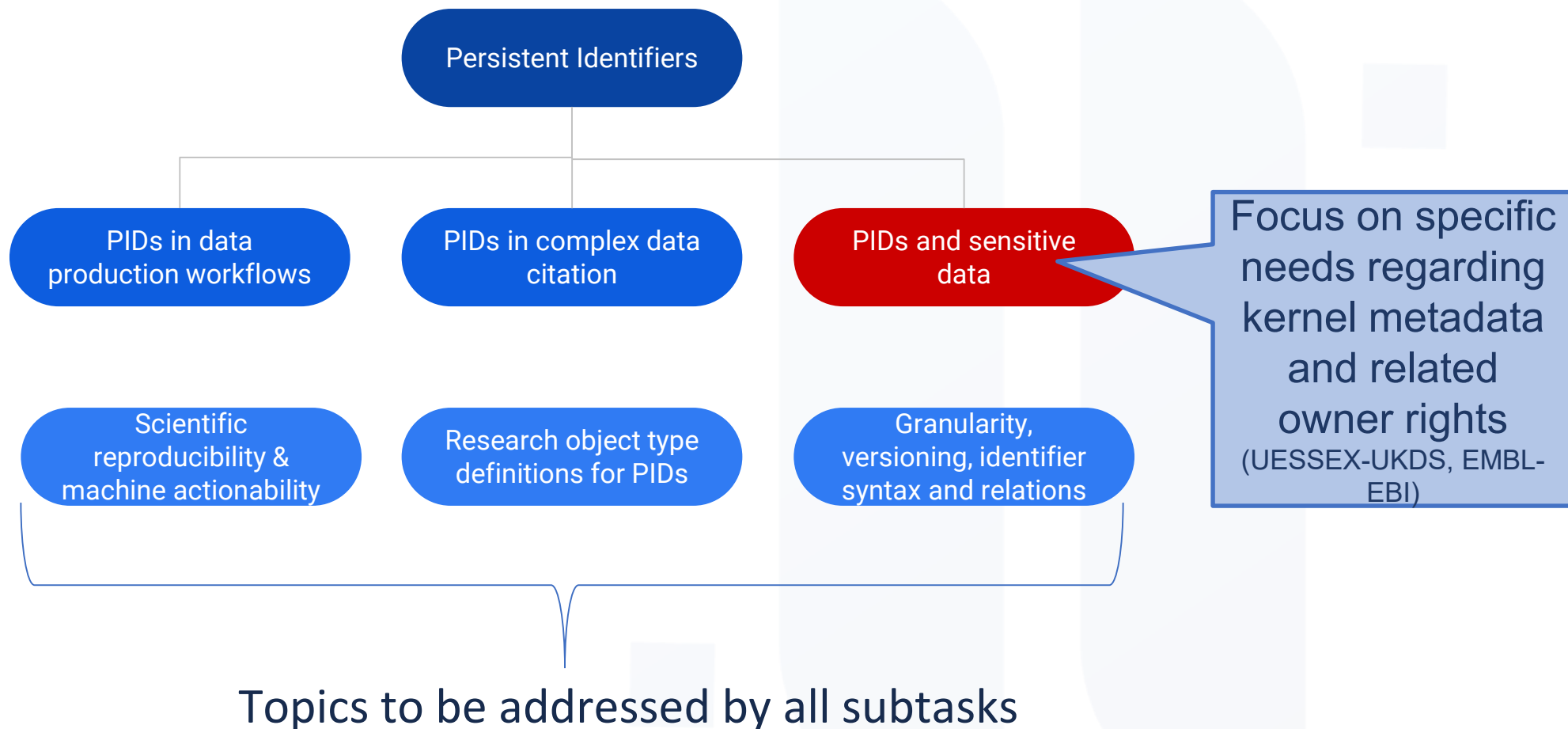


# Use case topics & structure

Work Package

Subtasks

Cross-cutting themes



## CESSDA UKDS - About

- CESSDA- Consortium of European Social Science Data Archives
  - 22 member countries and one observer
  - Explore use case specification across Service Providers
- UK Data Service
  - UK Service Provider to CESSDA
  - partnership between the Universities of Essex, Manchester, Southampton, UCL, Edinburgh and Jisc since 2012
- Lead partner: UK Data Archive
  - Originally Established 1967 at the University of Essex

## Context

- Data about humans in the Social Sciences
- Secure Remote Access
- Access via Safe Room
- Hybrid file-driven and RDF Linked Data Ecosystem
- ISO27000 Information Security
- Digital Economy Act UK (GDPR)
- EOSC PID Policy & evolving expectations of PID kernel metadata

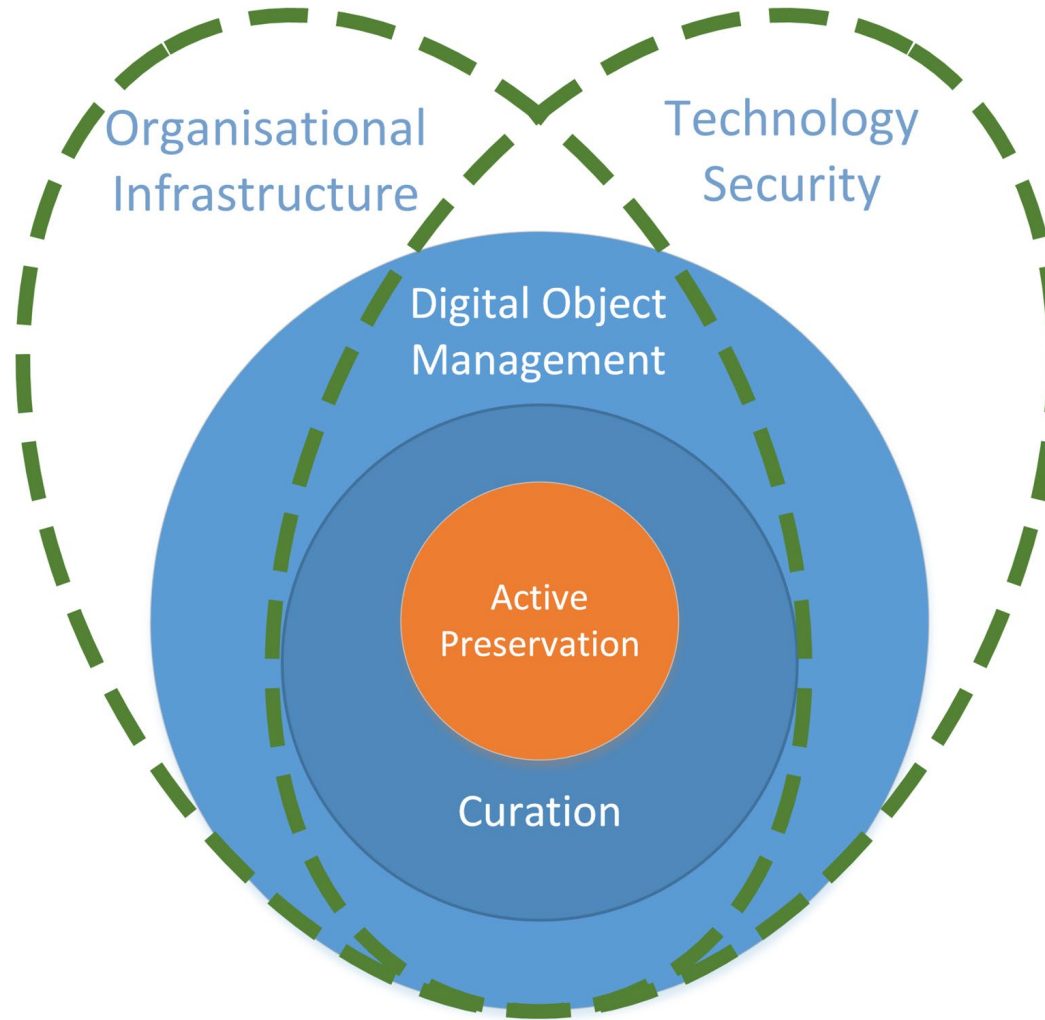
## Scope & approach

Digital Object Models (data & metadata), kernel metadata, rights management

Sensitive data and non-sensitive data classification and division

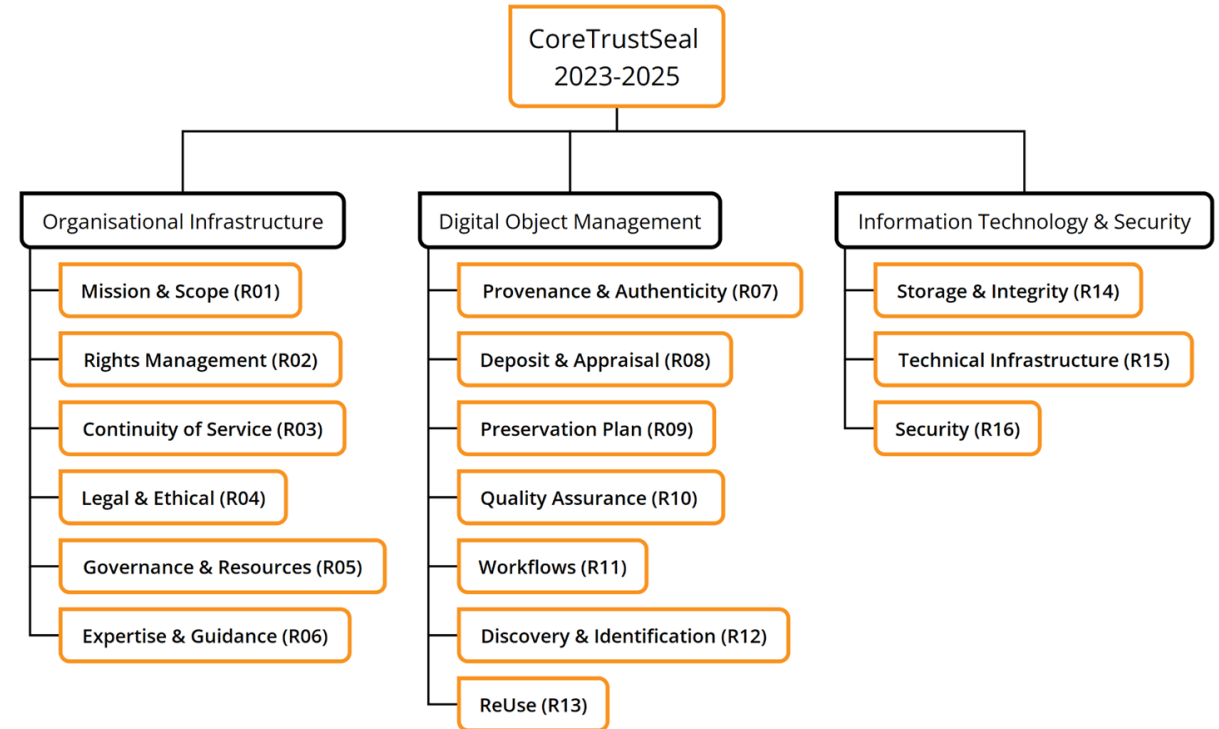
- At deposit
- During curation and later preservation (if sensitivity concerns change)
- Handling synthetic data, subsetting, outputs for checking etc

# Trustworthy Digital Repositories



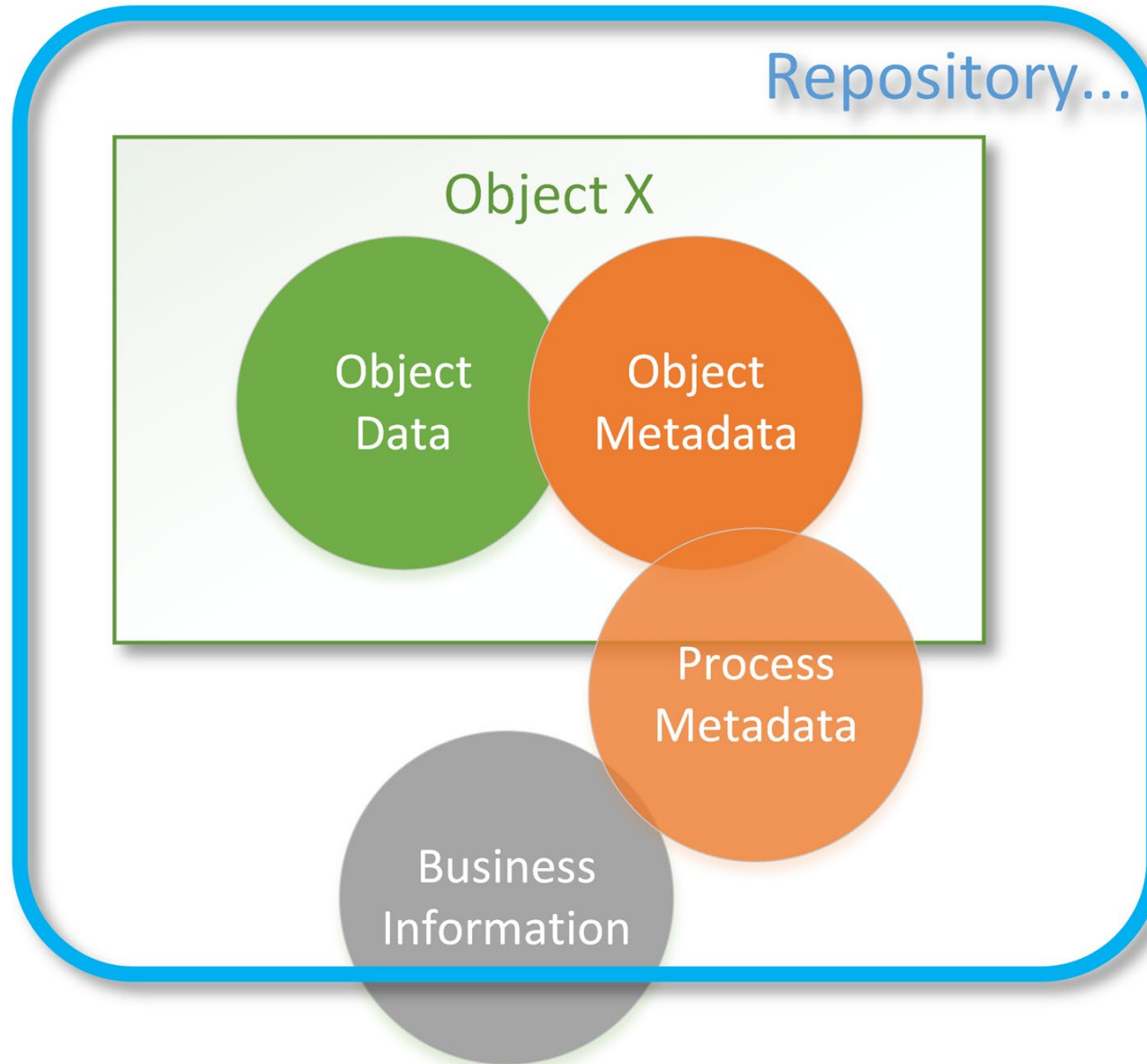
# CoreTrustSeal

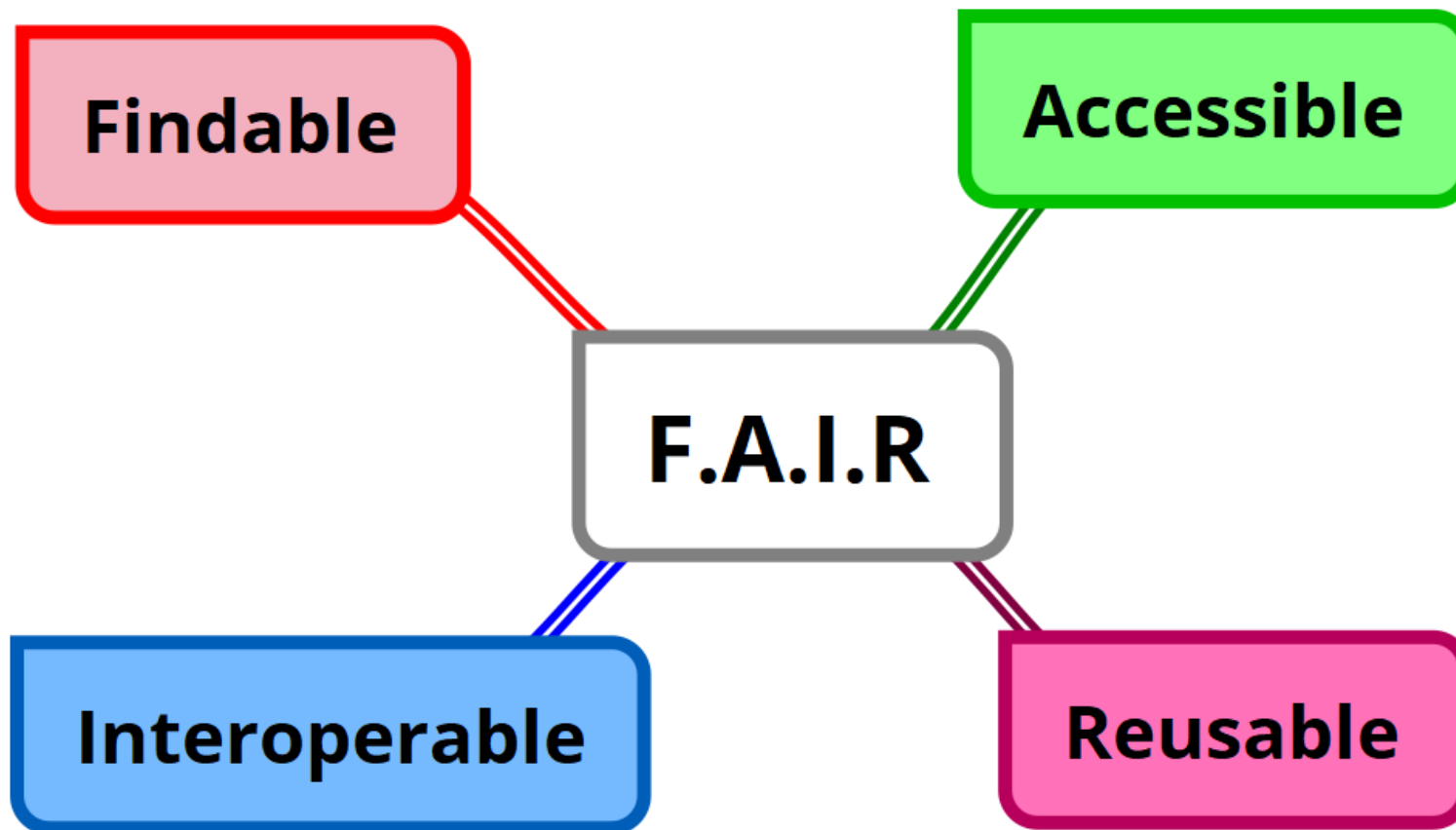
- “Core”
- Evidence aligned with artefacts necessary to deliver services
- Transparency > successful certifications public





Repository...









- F1. (meta)data are assigned a globally unique and persistent identifier.
- F2. data are described with rich metadata (defined by R1 below).
- F3. metadata clearly and explicitly include the identifier of the data it describes.
- F4. (meta)data are registered or indexed in a searchable resource.

#### Discovery & Identification (R12)



A1 (meta)data are retrievable by their identifier using a standardized communications protocol.

A1.1 the protocol is open, free, and universally implementable (vs context)

#### Technical Infrastructure (R15)

A1.2 the protocol allows for an authentication and authorization procedure, where necessary.

#### Security (R16)

A2 metadata are accessible, even when the data are no longer available.

#### Preservation Plan (R09)



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.

#### Reuse (R13)



R1. meta(data) are richly described with a plurality of accurate and relevant attributes.

#### Reuse (R13)

R1.1. (meta)data are released with a clear and accessible data usage license.

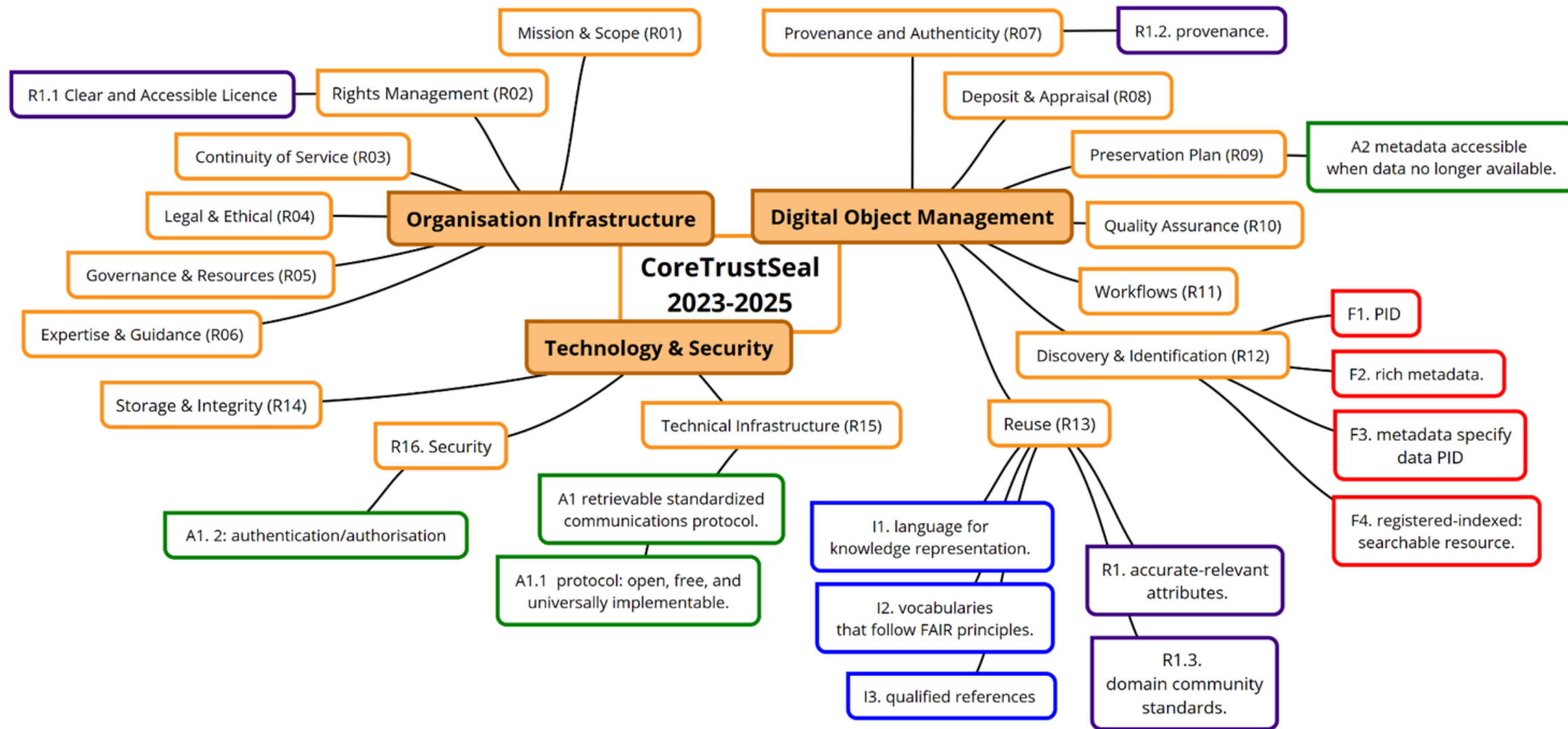
#### Rights Management (R02)

R1.2. (meta)data are associated with detailed provenance.

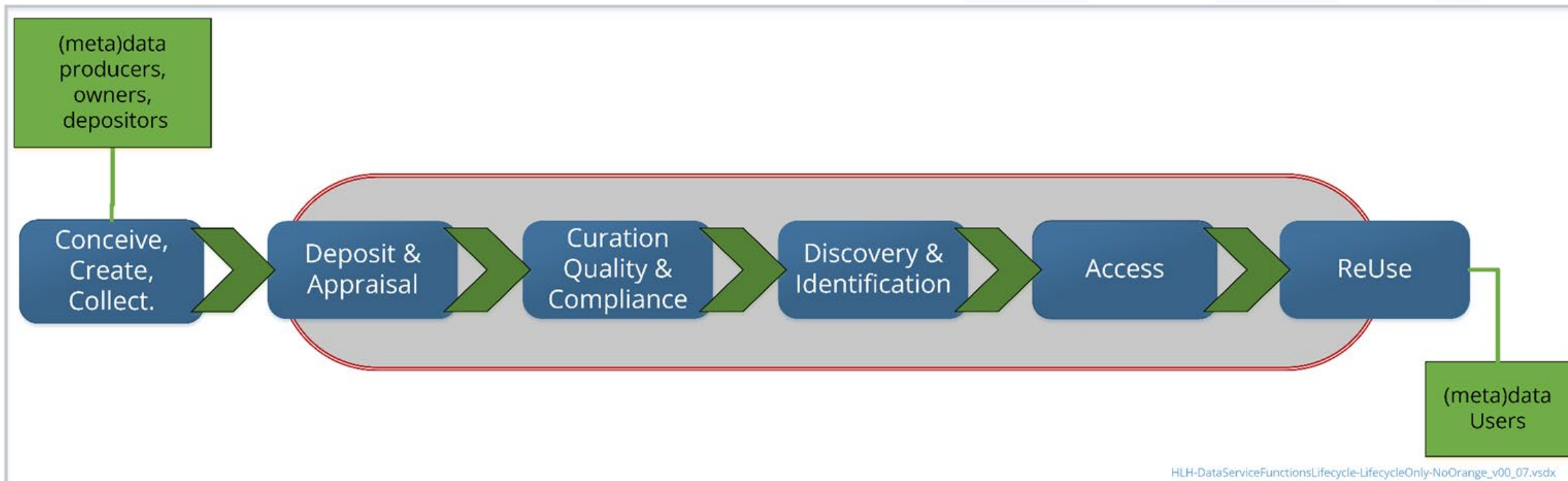
#### Provenance and Authenticity (R07)

R1.3. (meta)data meet domain-relevant community standards.

#### Reuse (R13)



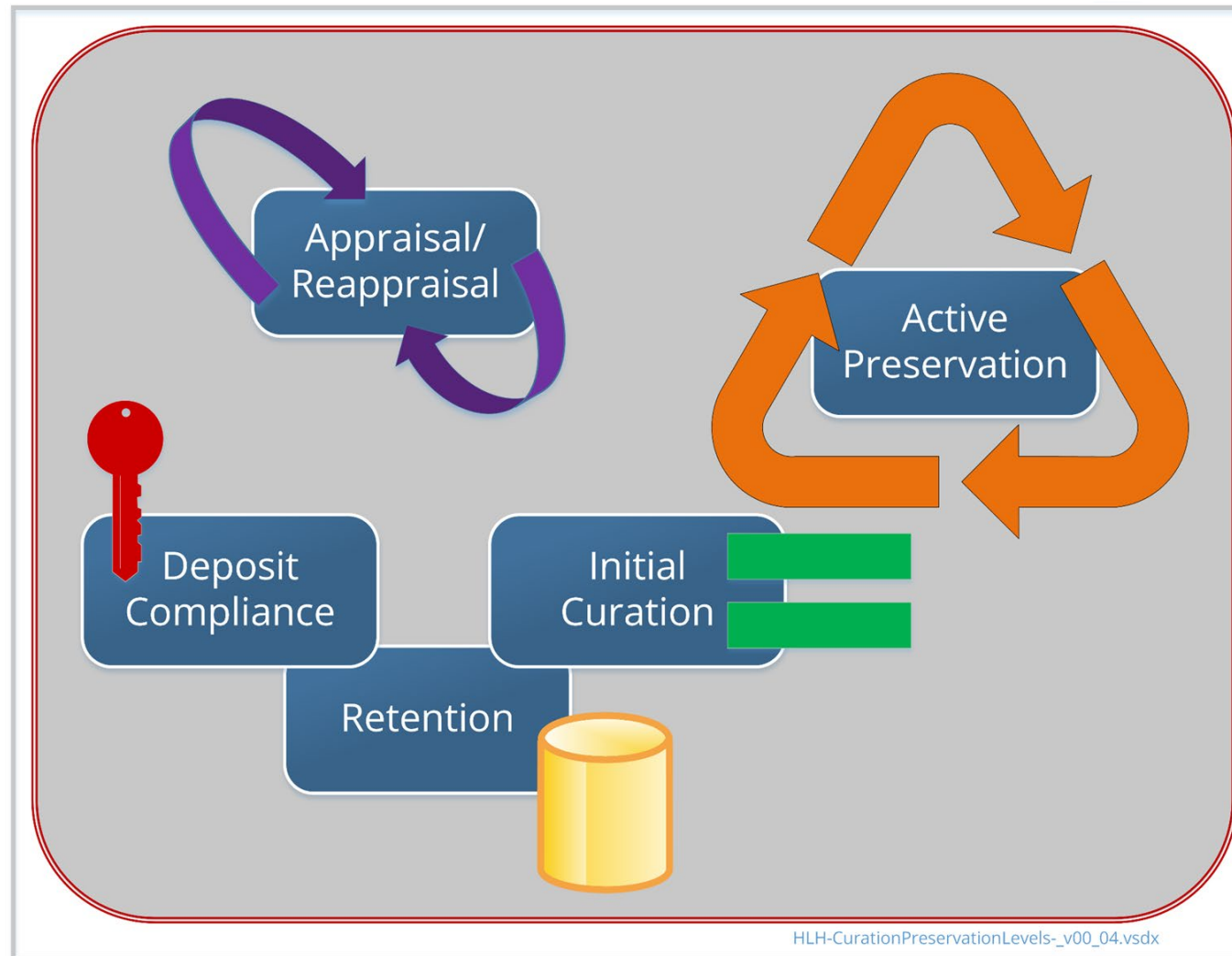
# Lifecycle Perspective



## Levels of Retention, Curation & Preservation

What is  
appraised & reappraised  
to be:

- retained,
- curated
- preserved



## +Sensitive Data

- Personal data, but also commercial, environmental, cultural and other drivers
- Restricted by design
- Sensitivity Classifications
  - Rights
  - Access Management



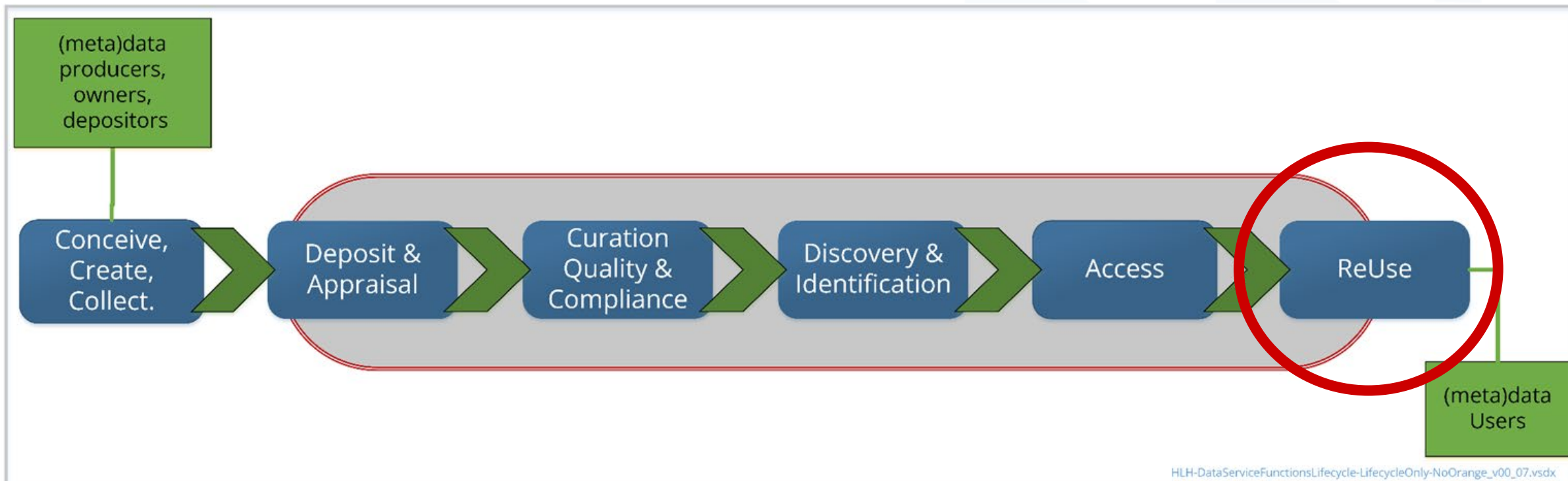


## Sensitive Data Implications

- The Core & TDR ‘security’ are not enough
- Inherent barriers to transparency
- Mediating ReUse



# Lifecycle Perspective: Enabling ReUse



## Scope & approach: Considerations

- **Classifications of Sensitivity**
  - Data vs. metadata
  - Personal data, commercially, environmentally, culturally sensitive
- **Assertions of Sensitivity**
  - Data owner-prescribed rights (self-declaration of sensitivity)
  - Intrinsic properties of the data (empirical run-time analysis)

## Scope & approach: Considerations

- **Subsets, linked data, statistical disclosure control**
- **PIDS Resolving & Routing to...**
  - Multiple serialisations e.g. XML, JSON etc. or RDF (RDF/XML, JSON-LD, Turtle etc.)
  - vs html landing pages & catalogue records
  - cf: machine actionability
- **Granularity & Version Management**
  - Sub-“object”: variables (statistics), questions (surveys) etc
- **PID Syntax and Semantics**
  - RDA PID Kernel recommendation, & FAIRCORE4EOSC

## Sensitive Data Requirements

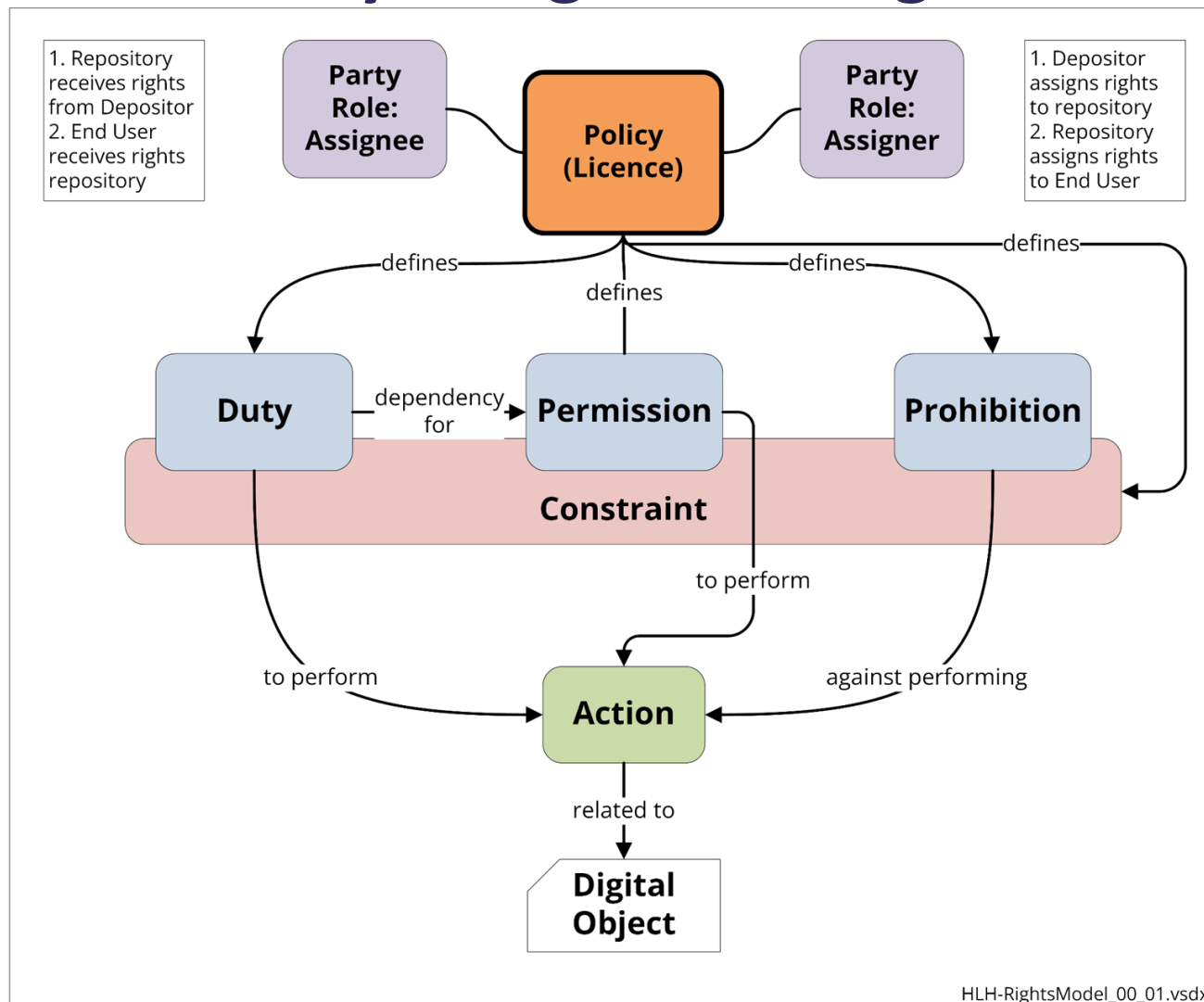
- As Open as Possible as  
**Protected** as Ne
- 5 Safes
  - Safe **Data**
  - Safe **Projects**
  - Safe **People**
  - Safe **Settings**
  - Safe **Outputs**



## Repository + + Additional...

- Infrastructure (Safe Settings)
  - Trusted Research Environment (TRE), Safe Room, Secure Remote Access
- Supporting Services (Training, Approval, SDC)
  - Resource & expertise
- Objects (subsets, linked and outputs)
  - Levels of Retention, Curation & Preservation

# Sensitivity & Rights Management



## Expected outcome & added-value

- Practical guidance on optimal PID usage (access & management) for sensitive data
- Consistent handling of sensitive data within and between secure environments
- More efficient use of PIDs for sensitive data to benefit research & deliver societal and economic impact
- High level kernel metadata on sensitivity & access restrictions
- Steps towards machine readable and machine actionable rights





@fairimpact\_eu /company/fair-impact-eu-project



Funded by  
the European Union