VRE: The Future of Science

Keith G Jeffery

keith.jeffery@keithgjefferyconsultants.co.uk

Agenda

- Who?
- VRE Past
- VRE Present
- VRE Challenges
- VRE Future

WHO?

Who

- Degree and PhD geology
  - In the sixties
  - Built an end-user friendly system for scientists: G-EXEC
- 1970s: Worked for British Geological Survey and led Natural Environment Research Council Central Computing Group
- 1980s: Worked at the UK National Science Lab: databases and office systems for science and administration
- 1990s: Led the IT Research Division
- Final position: Director, IT and International Strategy at STFC Rutherford Appleton Laboratory
  - 140 staff, 360000 users, large computers, all areas of research
  - Worked on setting up ESFRI with John Wood
- Retired 2013, now consultant
- Working on:
  - EPOS
  - ENVRIplus
  - ENVRIFAIR
  - MELODIC
- Co-chair of RDA VRE Interest Group
- Co-chair of RDA metadata groups

VRE PAST

Some definitions:

- e-I: e-Infrastructure: computers, networks, storage, sensors, lab equipment outside the RI and with multiple users
- e-RI: the electronic representation of an RI: its assets, software services, computing resources, sensors and lab equipment
- VRE: Virtual Research Environment
- SG: Science Gateway
- VL: Virtual Laboratory

VRE History

- 1990s Open University: an environment collecting the resources for distance learning and research
  - Use of internet for access to resources
- JISC 2004-2011: Definition / Proof of Concept / Embedding
  - For UK universities to have a similar environment for learning and research
- 2011-2014: concept of researcher workbench with access to:
  - e-literature
  - Data usually via portals maybe through services
  - Instruments/sensors and computing facilities
  - Teleconferencing facilities
  - Office facilities
  - Access to research administration tools (research grant systems, personal web pages, research reporting systems, personal bibliography…..
- 2014-present: VREs (Europe), SGs (Science Gateways) North America VLs (Virtual Laboratories) Australia

VRE PRESENT

Lifewatch

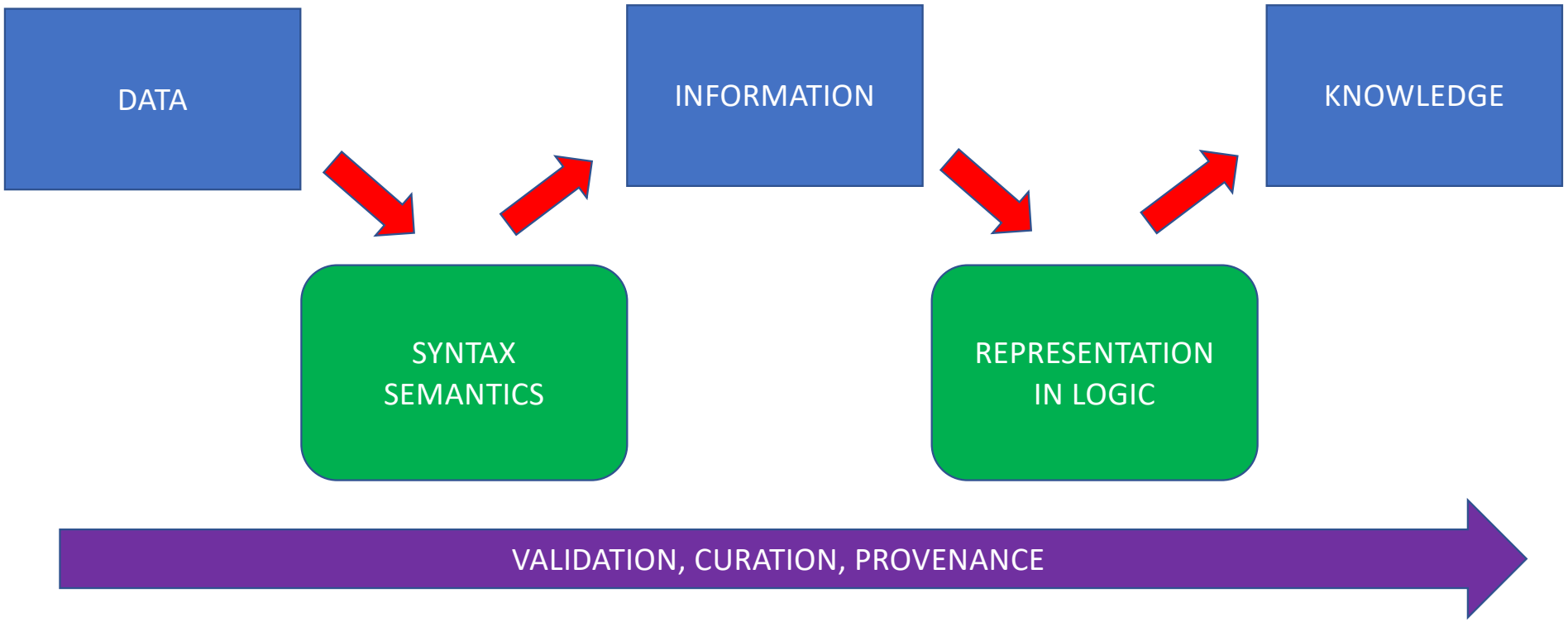
- 5 VREs/vLabs
  - **SWIRL: S**cenario-based **W**ater **I**nnovation & **R**esearch **L**aboratory
  - RvLab: use of R for analytics
  - RShiny data explorer: sensor data: Flemish
  - Phyto: Phytoplankton (uses Taverna)
  - AS: Alien Species
- 3 portals
  - IT
  - GR (catalog CIDOC model)
  - BE
- 16 services in catalog
- More advanced than many ENVRI RIs
  - Some of the presentation will be familiar / recognisable to at least some of you

The challenges to be addressed

Clearly, these require:

1. Multidisciplinary teams of researchers cooperating hence collaborative tools
2. FAIR (Finding, Accessing, Interoperating, Re-using)
   multidisciplinary assets: data, software services, instruments/sensors, e-infrastructure, literature including grey literature
3. production and sharing of results

The added value chain



Requirement: Researcher View

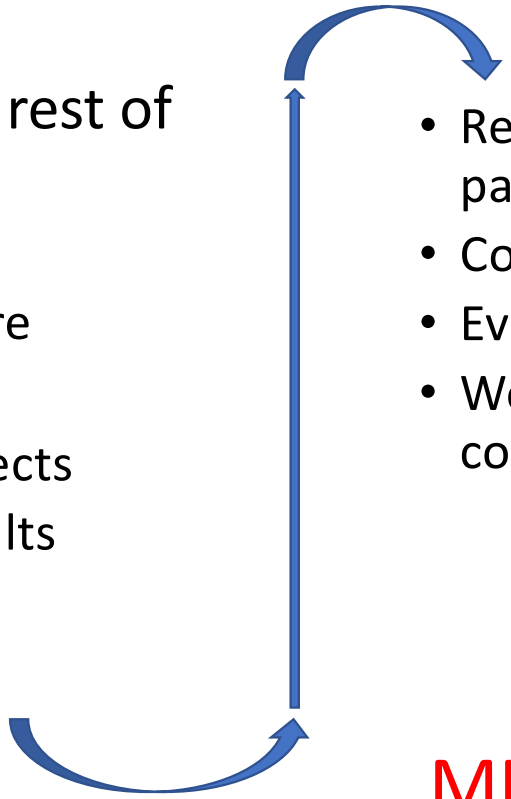
- User Request to system
  - Typing ➔ voice, static or mobile device
- System interacts to ensure understood
  - Clarification, improvement, disambiguate
  - Medium/format of output required
- System discovers relevant assets
  - Location, rights management, (costs), security, privacy, performance considerations
- System constructs workflow
  - Presents to user for validation /correction
  - Including all rights management, security, privacy, costs
  - Distributed, parallel
- System executes workflow
  - User has monitoring screen – can steer execution
  - Distributed, parallel (optimised for data locality)
- System returns results to end-user
  - In appropriate medium/format

- In essence same as requirement for G-EXEC 1968-71
- Main difference then
  - System did not interact to understand
  - User constructed the workflow
  - No user monitoring / steering (batch processing)
- Now possible thanks to virtualisation
  - Metadata
    - Datasets
    - Software
    - Publications
    - Persons
    - Organisations
    - Equipment, facilities
    - GRIDs, CLOUDs

METADATA IS THE KEY

Requirement: Researcher View

- plus assistance with the rest of the research lifecycle
  - Generating ideas
  - Researching the literature
  - Writing proposals
  - Managing research projects
  - Publicising research results
  - Maintaining online CV



- Reviewing proposals and scholarly papers
- Cooperating with other researchers
- Evaluating against other researchers
- Working on editorial boards and conference programme committees

METADATA IS THE KEY

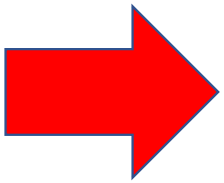
VRE Present

- EUROPE
- 6 EC-funded projects, 3 relevant
  - BlueBridge
    - ‘silo’ GUI to e-infrastructure
  - EVER-EST
    - Research objects
  - VRE4EIC
    - Reference architecture
    - toolset
- REST OF WORLD
- Science Gateways
  - Domain specific
  - Catalog of assets
  - ‘silo’ GUI to e-infrastructure
- Virtual Laboratories
  - Catalog of assets
  - “pick ‘n’ mix”

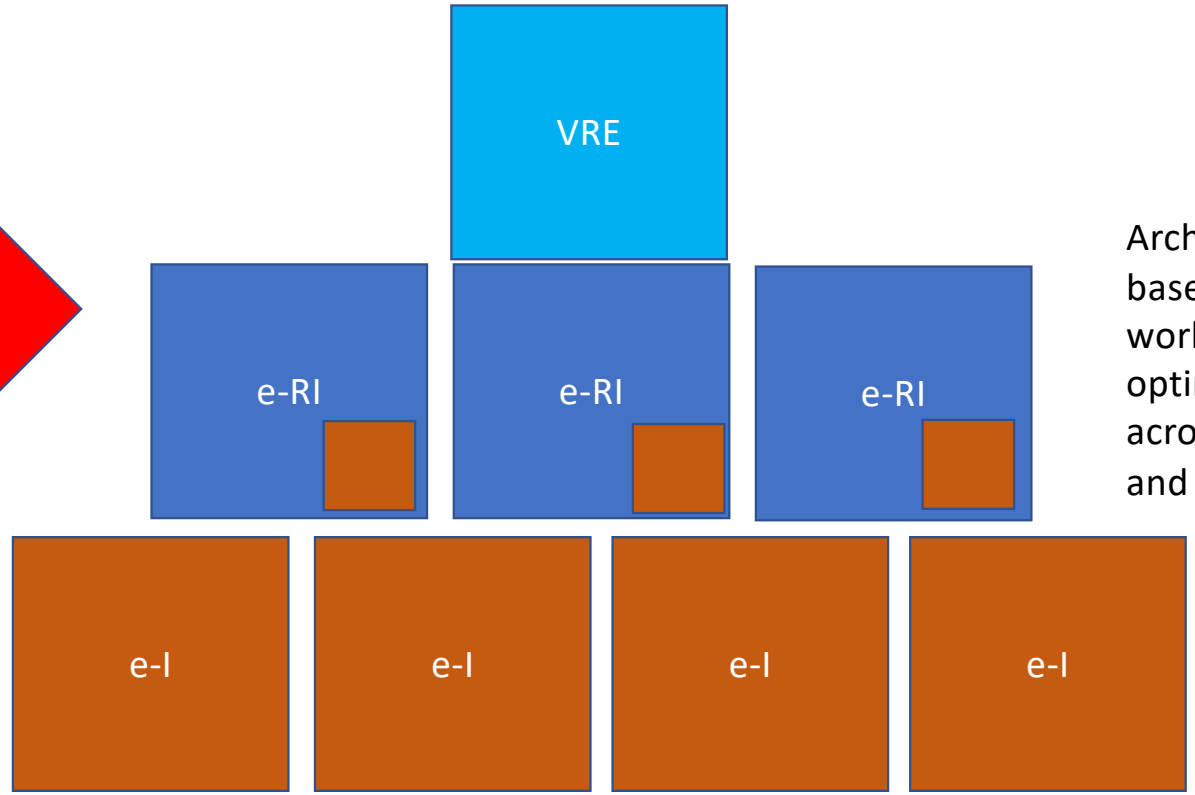
The evolution of VREs

Domain-specific

3-layer architecture : client / app server / data-compute server

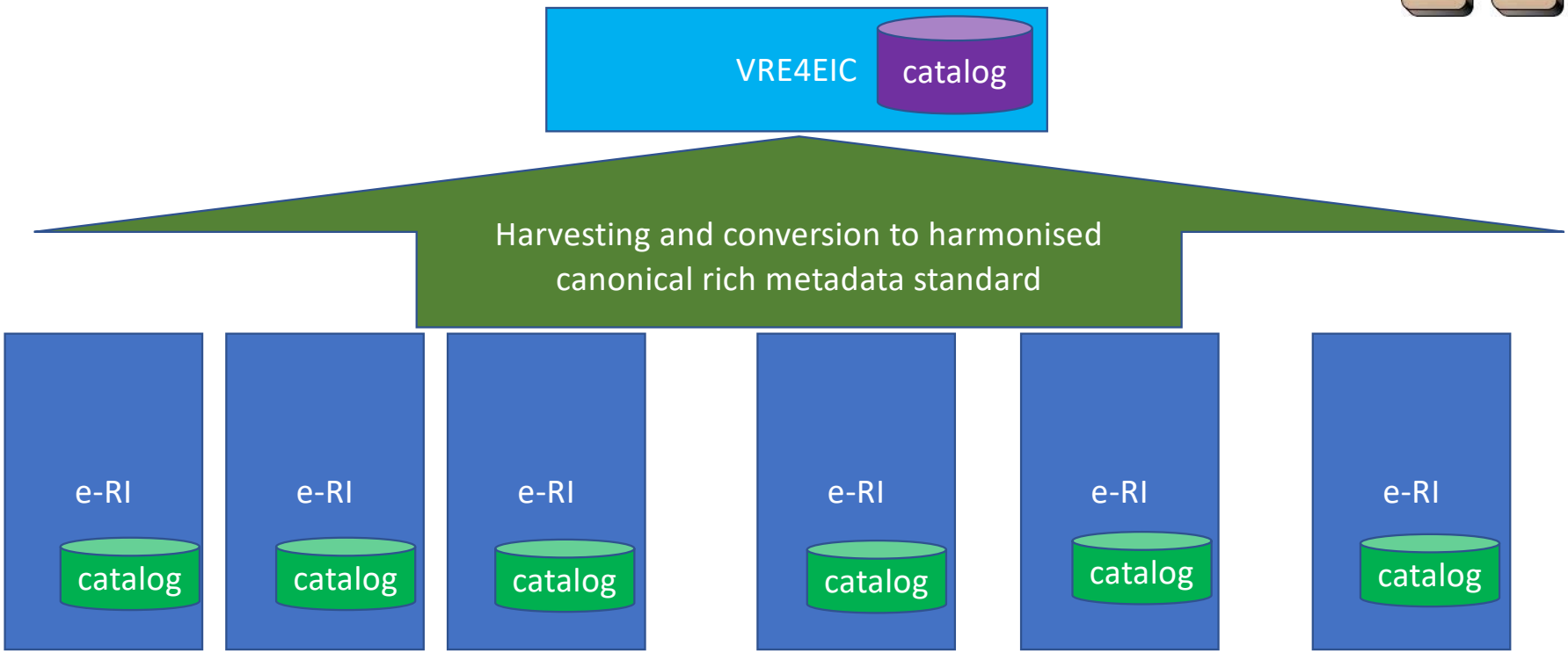


Multi-domain

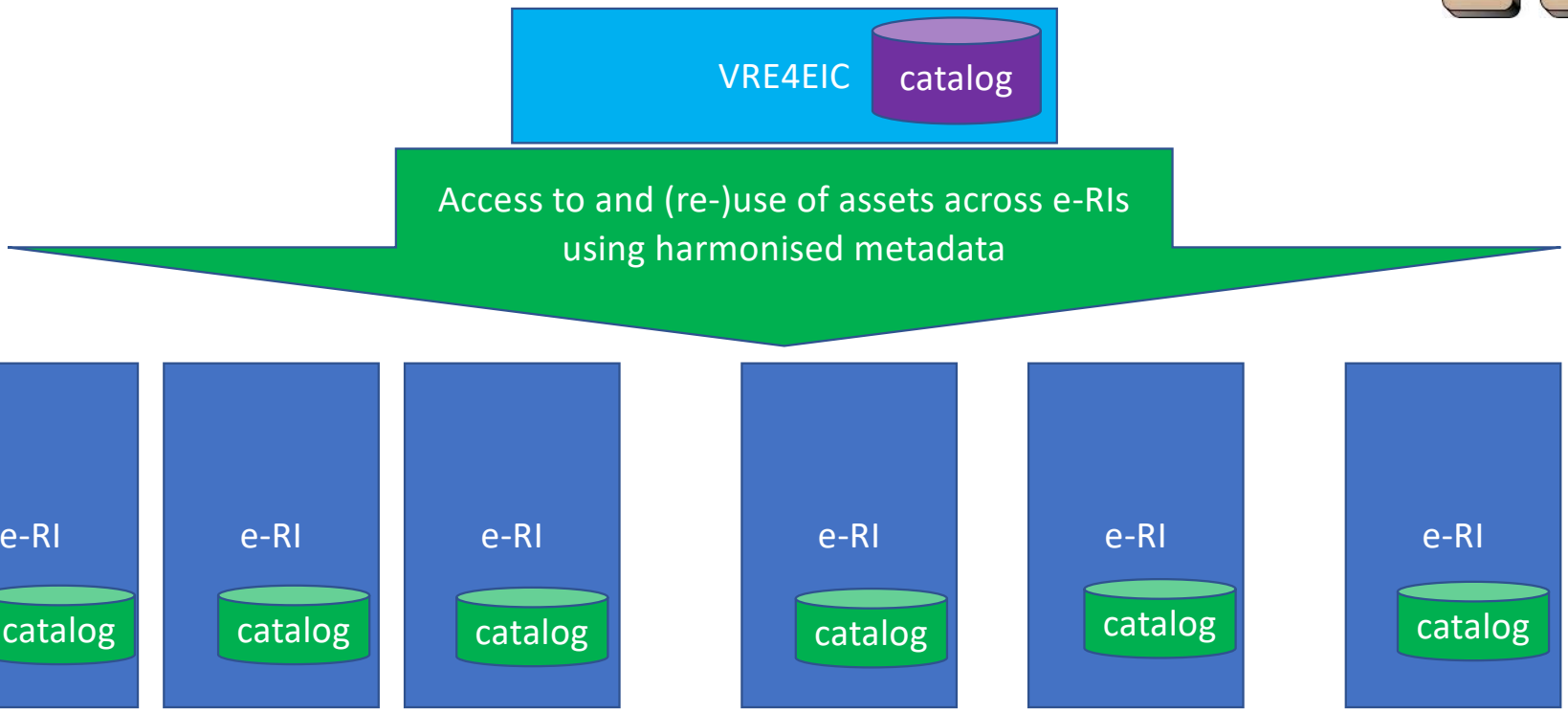


Architecture based on workflows optimised across e-RIs and e-Is

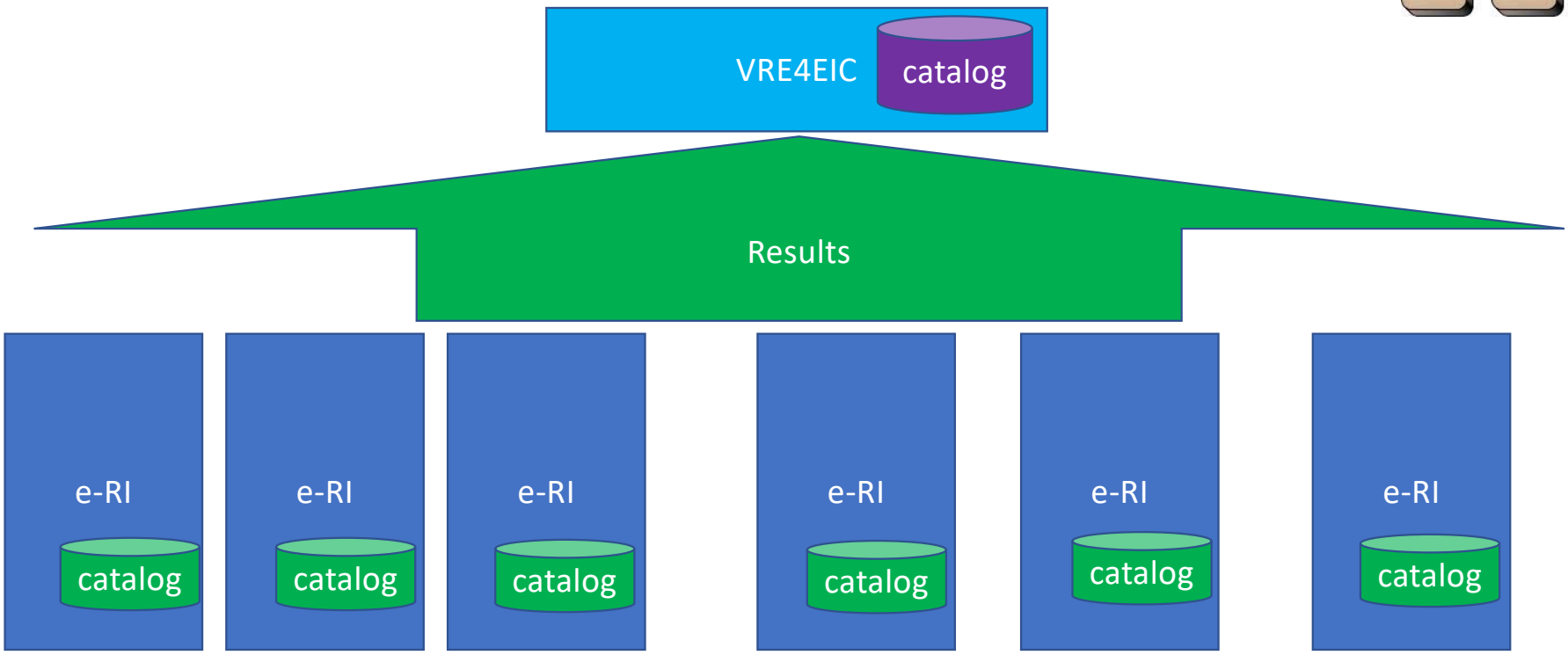
VRE4EIC Idea (Step 1)



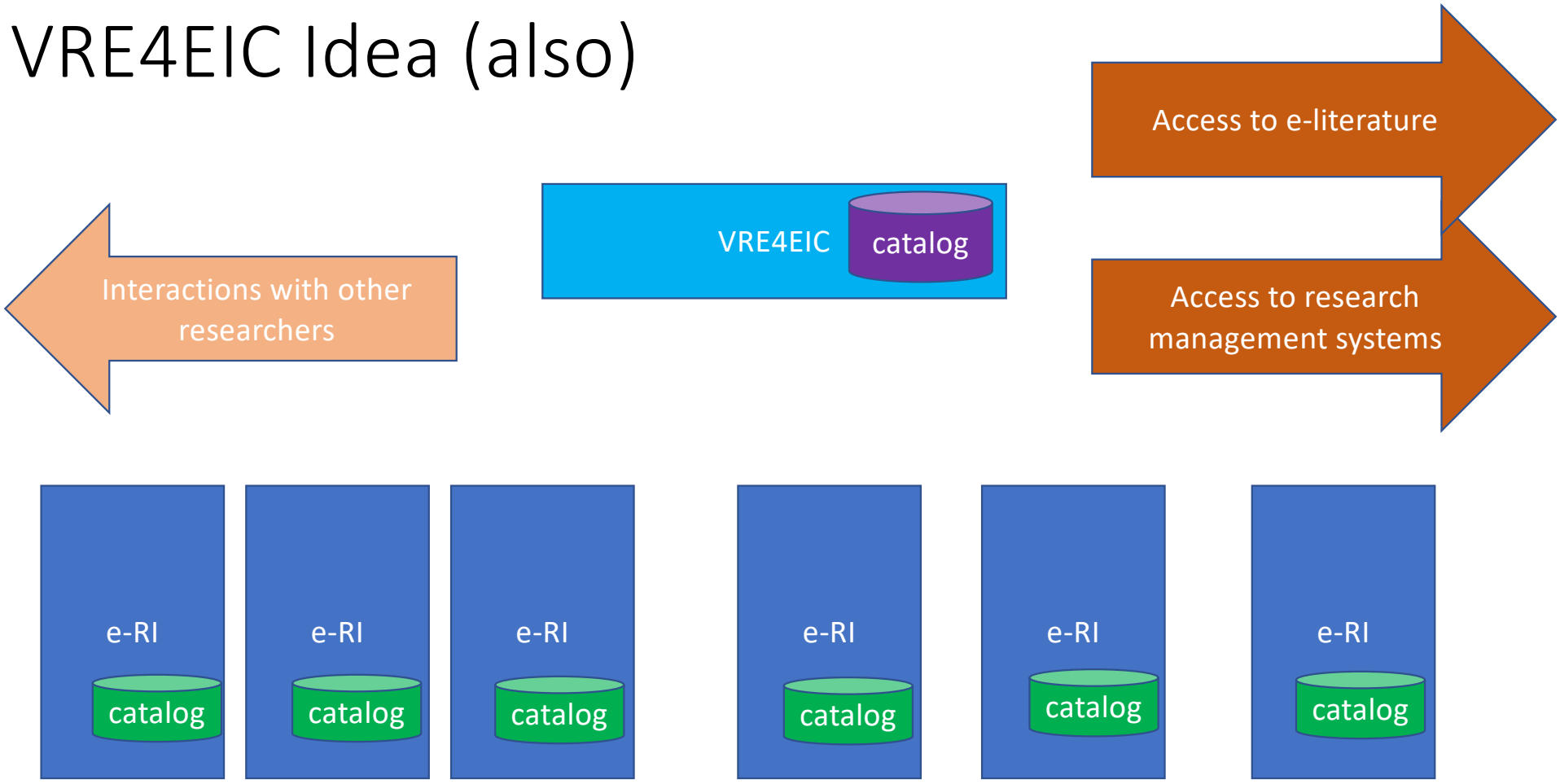
VRE4EIC Idea (Step 2)



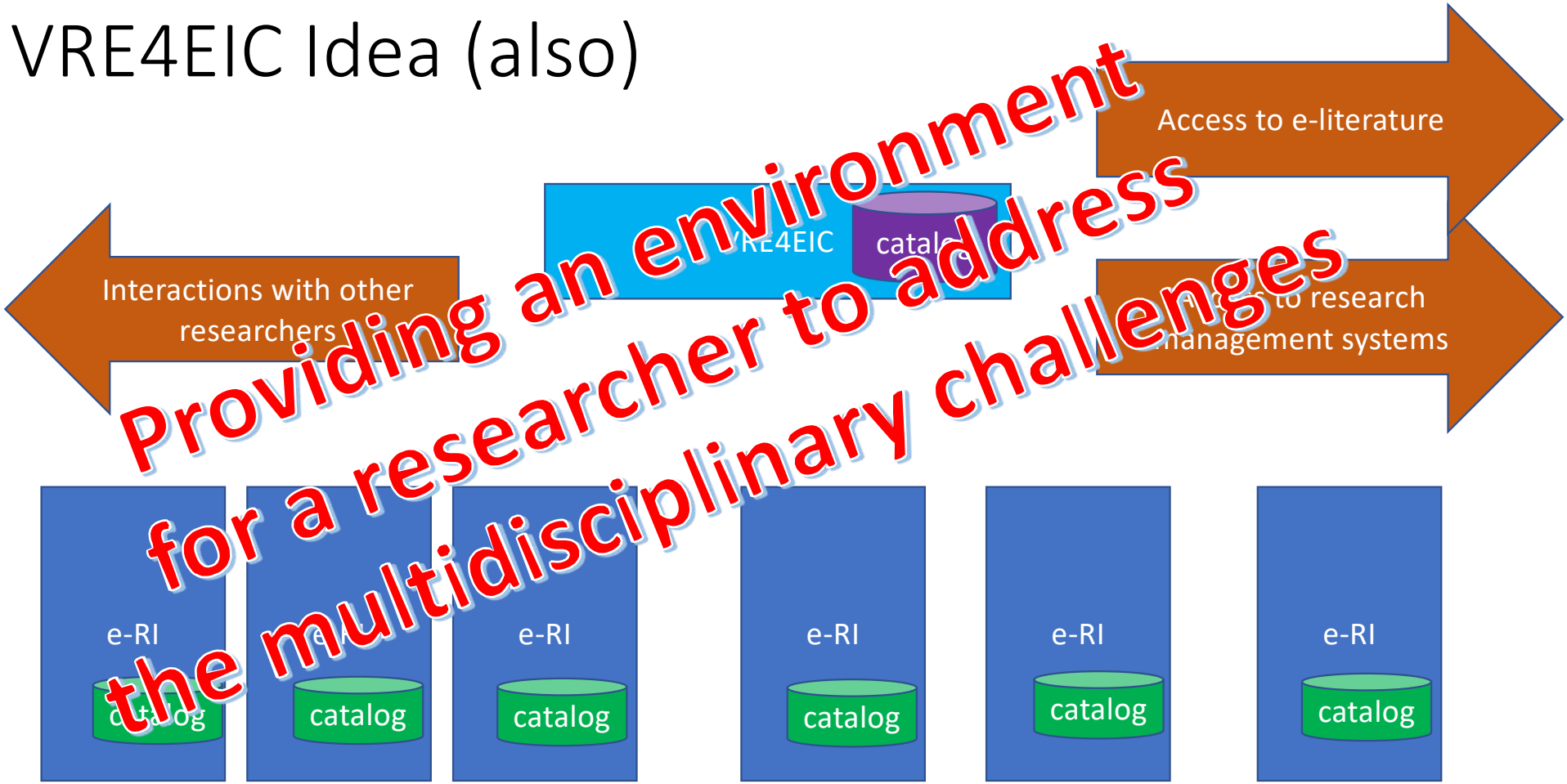
VRE4EIC Idea (Step 3)



VRE4EIC Idea (also)

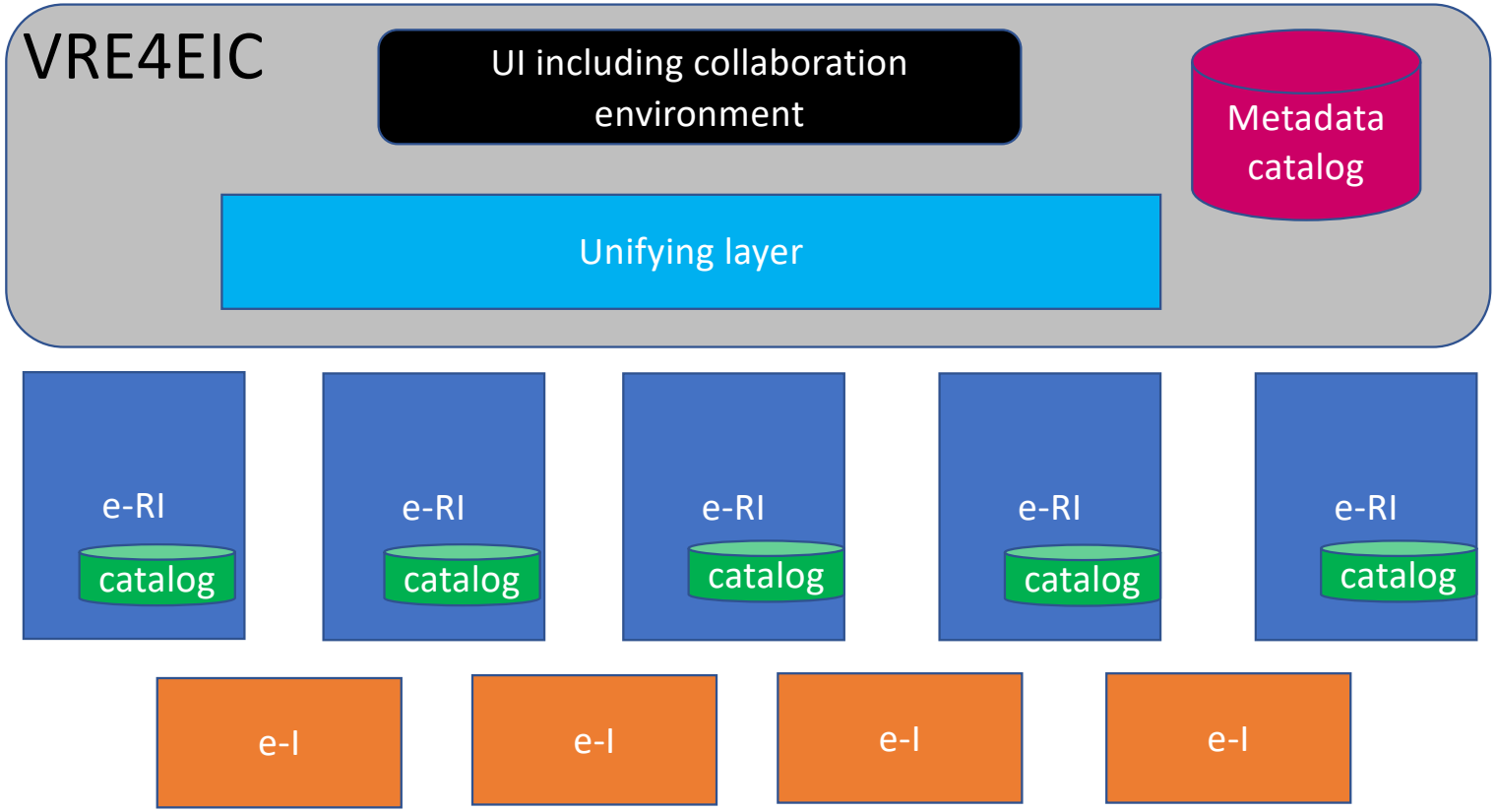


VRE4EIC Idea (also)

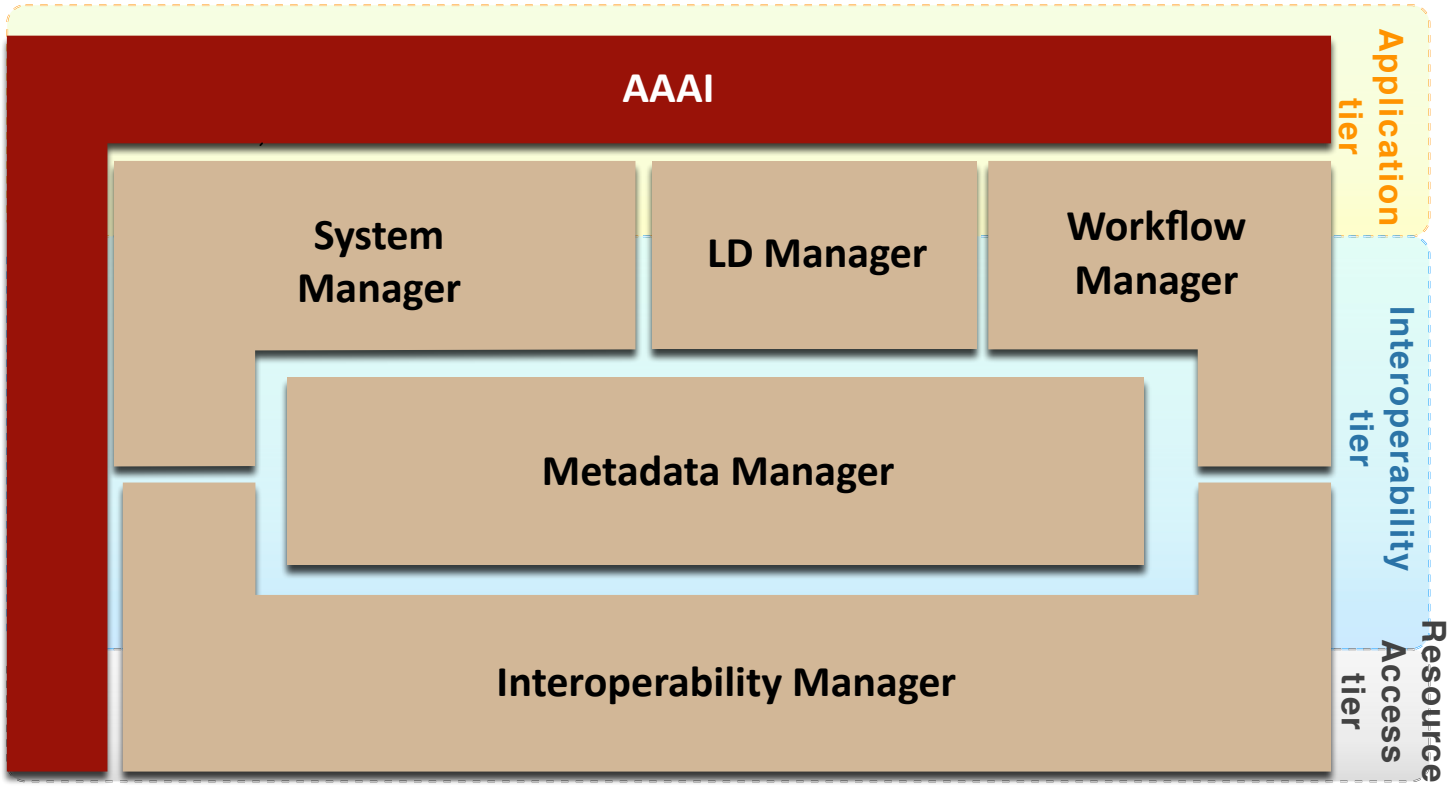


Providing an environment for a researcher to address the multidisciplinary challenges

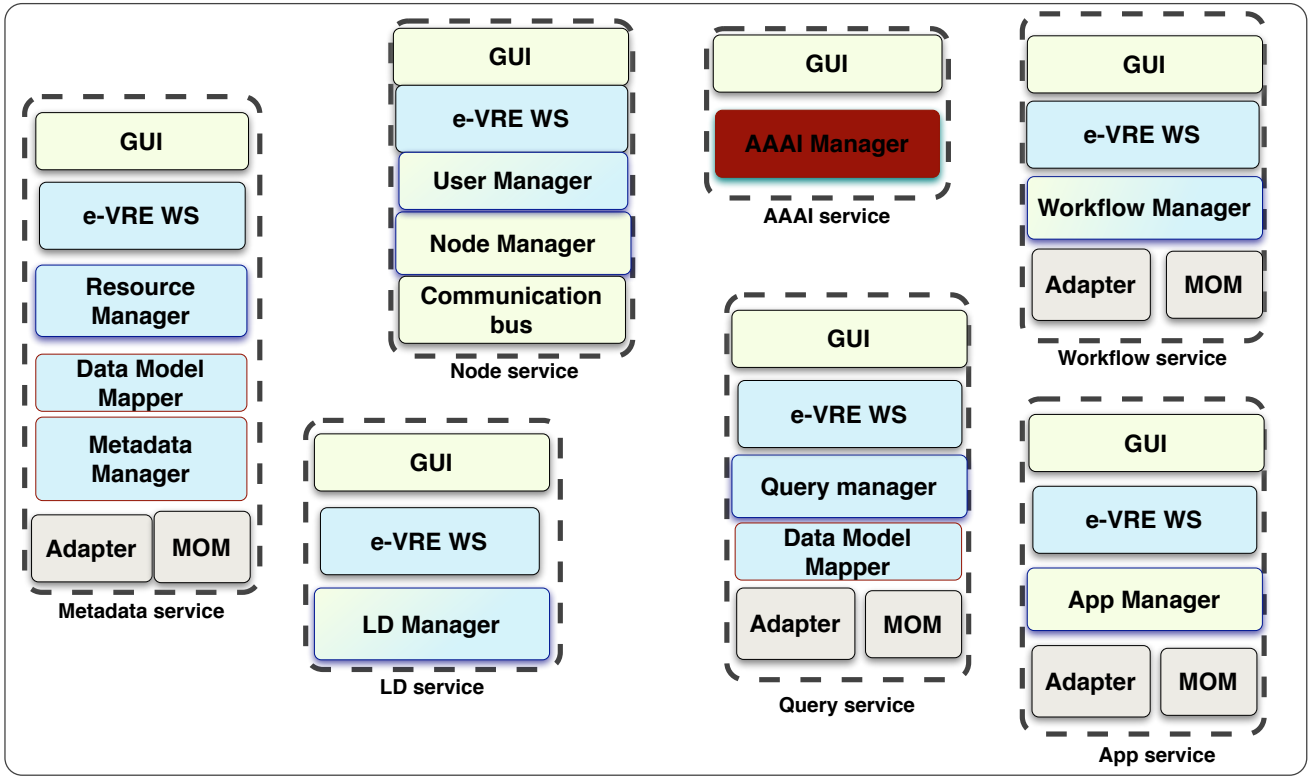
Interoperation across RIs



VRE4EIC Architecture Components



VRE4EIC Architecture Services & Microservices



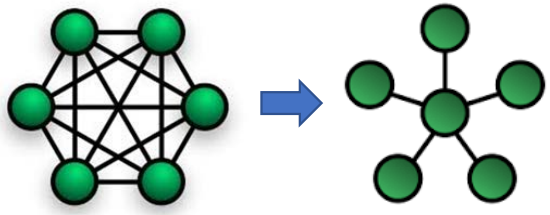
VRE CHALLENGES

VRE Challenges

- Interoperation
  - Integrating metadata
- AAAI
  - To manage trust, security, privacy with authorisation of resource usage
- Composing workflows
  - A script for accessing, integrating, re-using assets for the research purpose
- Deploying workflows
  - Appropriately across the e-infrastructure

Interoperation: Integrating Metadata

- The n squared problem: reduce to n by superset canonical rich metadata schema
  - Actually it is reducing (n*(n-1)) to (n)
- Need for formal syntax and declared semantics
  - Referential integrity
  - Functional integrity
  - Use of ontologies
  - Multilinguality
- Matching and mapping of metadata schemas
  - And then conversion of metadata records
  - Use of 3M technology from FORTH



If these properties are not present, there is likely to be ambiguity in matching, mapping and conversion

Metadata Desiderata: Example: Syntax

<ID><Title><Abstract><Author>

(a) May be >1 author (referential integrity)

(b) May be >1 title/abstract (multilinguality) (referential integrity)

(c) Author is not uniquely dependent on ID which refers to the digital object (functional integrity)

This was a simple example; it can get much more complex hence need for formal syntax

Metadata Desiderata: Example: Semantics

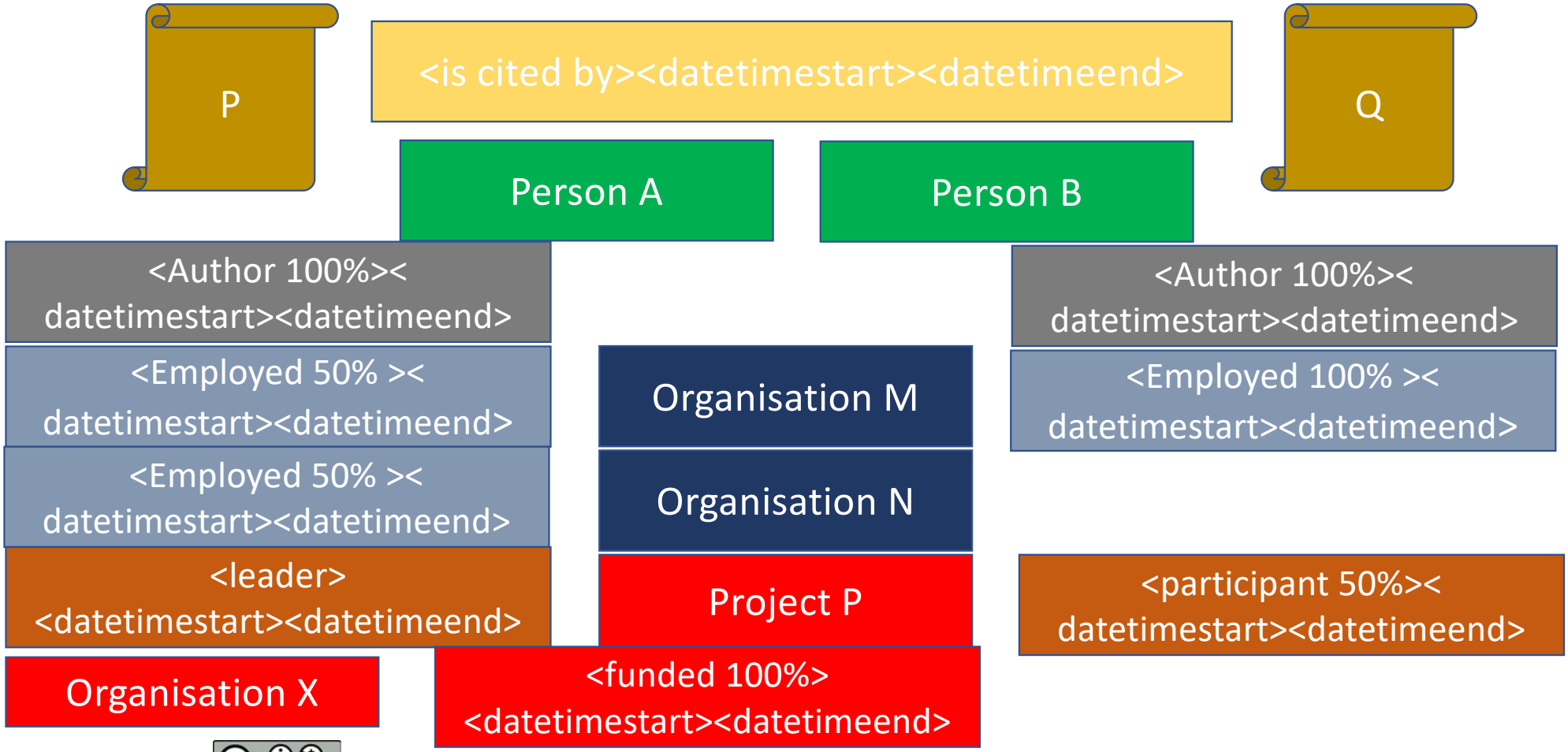
To obtain consistency it is essential that terms are:
(a) restricted to an agreed list; (b) related to other terms; (c) defined.

Take the 4 character lexical term ‘bond’

- Chemistry: e.g. ionic bond, covalent bond
- Economics: a government ‘share’
- Sociology: a special relationship – the bond between mother and child
- Entertainment: 007 James Bond

Particularly important for keywords and classification codes
e.g. Frascati ⇔ UDC ⇔ ……

Metadata Standards Need to Express: example citation



Metadata Standards Need to Express: example citation

<{part of} article P> <is cited {positively | negatively} {start date,time-end date,time} by> <{part of} article Q>

<Article P> <{100%}authored {start date,time-end date,time} by> <Person A>

<Person A> <employed {50%} {start date,time-end date,time} by> <Organisation M>

<Person A> <employed{50%} {start date,time-end date,time} by> <Organisation N>

<Person A> <is {start date,time-end date,time} leader of> <Project P>

<article Q> <{100%} authored by {start date,time-end date,time} > <Person B>

<Person B> <employed {100%} {start date,time-end date,time} by> <Organisation M>

<Person B> < {start date,time-end date,time} participates{50%} in> <Project P>

<Project P < {start date,time-end date,time} is funded 100%} by> <Organisation X>

Same is true for citation of datasets or software

Advantage of Formal Expression

- This is essentially expressions in (decorated) first order logic – the elements in red (previous slide) are linking semantic relations
- Therefore can do **deduction** (facts from rules) and **induction** (rules from facts): this
  - reduces the effort of input of metadata
  - increases the quality by consistency validation
  - improves precision and recall of query to find assets
  - can manage user preferences
  - ensures metadata actionable (as in FAIR principles)

Current Metadata Standards

- DC (Dublin Core)
  - Text, html, XML, RDF
- DCAT (Data Catalog Vocabulary)
  - XML, RDF
- ISO19115/INSPIRE
  - XML, RDF
- CKAN (Comprehensive Knowledge Access Network)
  - RDF

} all suffer from lack of Referential integrity & Functional integrity

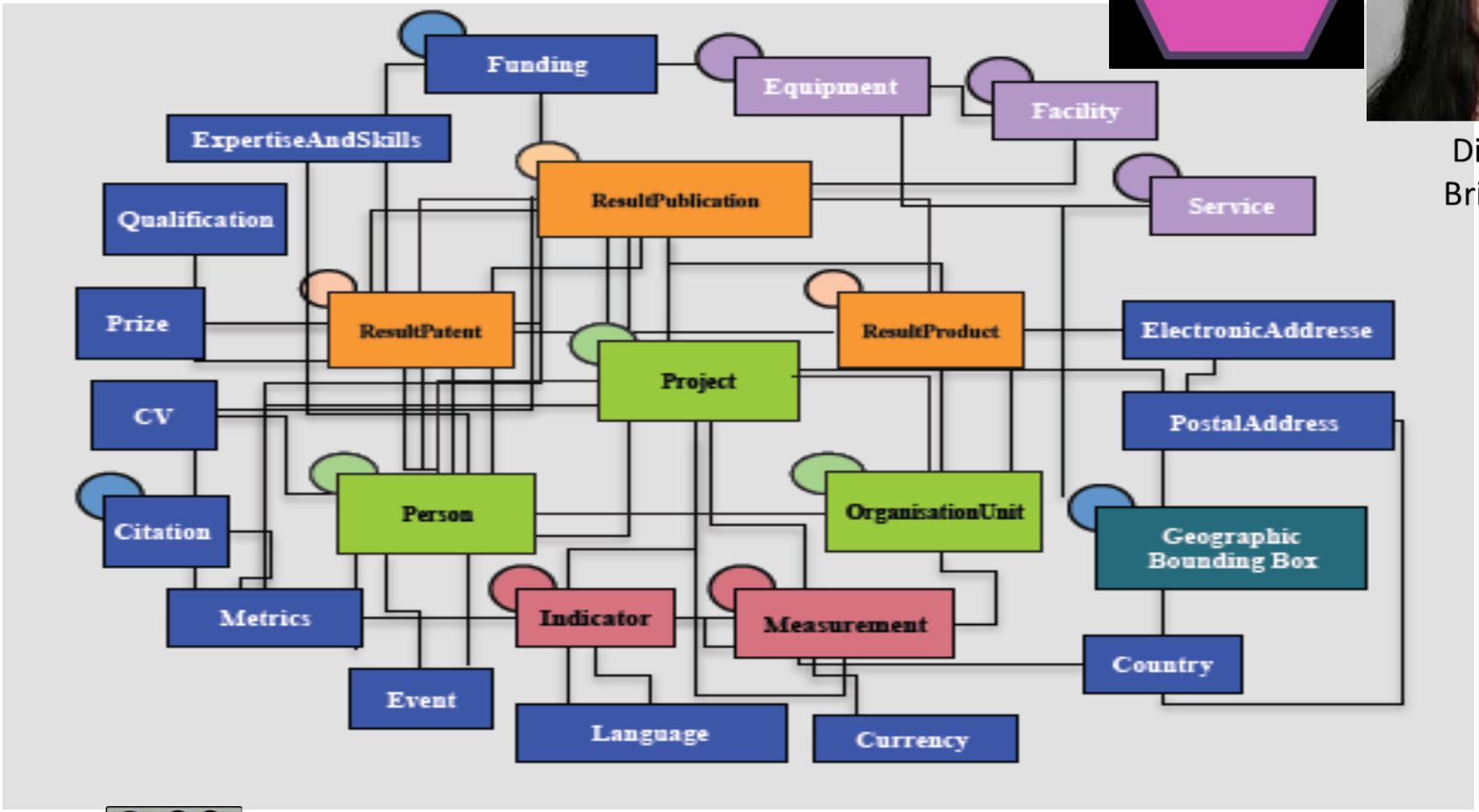
- Move to RDF – provides formal syntax and semantics – over last 10 years
- But RDF is triples; need many triples to express complex role-based, temporal relationships

Metadata Standards: CERIF

- **C**ommon **E**uropean **R**esearch **I**nformation **F**ormat
  - Expert group: 1 representative per Member State
  - Data Model for exchange and storage of information about research
  - EU Recommendation to Member States
- CERIF91 (1987-1990) quite like the later Dublin Core (late 1990s)
  - Tested and found to be inadequate (as predicted by Jeffery)
- CERIF2000 (1997-1999) used full E-E-R-T modelling
  - formalised by Jeffery & Asserson
  - Base entities
  - Linking entities with role and temporal interval (i.e. decorated FOL)
  - Can be implemented in almost any system environment (relational, object-oriented, hypermedia, logic……
- 2002 EC requested euroCRIS to maintain, develop and promote CERIF www.eurocris.org
  - Now in use in 43 countries and national standard for research information in 10
  - 6 SMEs providing CERIF systems , 2 bought up by Elsevier and Thomson-Reuters respectively

Contextual Metadata: CERIF

Diagram by Brigitte Jörg



Characteristics of CERIF 1

- (a) it **separates clearly base entities from relationships** between them and thus represents the more flexible fully-connected graph rather than a hierarchy;
- (b) it has **generalised base entities with instances specialised by role** (for example <person> rather than <author>), the role specialisation is in the linking entities;
- (c) it handles **multilinguality by design** (multiple language representations are linked with role (e.g. machine or human-translated) and temporal information (so representing versions of the translation) to the appropriate attribute treated as an entity – example <title> linked to <publication>;
- (d) **temporal information is in the link entities** not the base entities (example employment between two dates is in the linking relation between <person> and <organisation> and not an attribute of either of the base entities;

Characteristics of CERIF 2

- (e) the temporal information in linking entities **provides provenance and versioning** recording e.g. versions of datasets and – in the associated role attribute – the method of update or change;

- (f) CERIF **separates the semantics into a special 'layer'** which is referenced from CERIF instances. The sematic layer includes permissible values for roles in any linking entity (e.g. <person> <author|editor|illustrator|reviewer…> <publication>) and also permissible values for controlled values of attributes in base entities (e.g. ISO country codes). Thus semantic terms are stored once and referenced many times (preserving integrity). The semantic layer – like the syntactic layer - consists of base entities (e.g. the valid values for an attribute or valid roles for a linking entity) and linking entities thus allowing relationships between vocabularies and relationships between individual terms to be represented i.e. an ontology. Thus CERIF provides formal syntax and declared (multilingual) semantics.

- (g) CERIF has a **formal review and update process** controlled by euroCRIS and so can evolve. However, it is designed to evolve by accretion (there is a method of adding additional entities and appropriate linking entities to existing base entities) so that the core of CERIF remains constant for interoperability among CERIF installations.

Use of CERIF

euroCRIS
Current Research Information Systems

- Originally intended for CRIS (Current Research Information Systems)
  - Research institutions to manage their portfolios, publish their research information (web pages), offer services, interoperate with others
  - Research funding institutions to manage their portfolios, assess funded research
  - Industry to discover relevant research to be used for wealth creation
  - Government to discover relevant research to be used for policymaking
  - Publishers to check their own catalogs, to find new potential authors, to track emerging domains
  - Media to publish research 'stories'

- Now used in large e-Research infrastructures /projects

RDA Metadata Principles

- The only difference between metadata and data is mode of use
- Metadata is not just for data, it is also for users, software services, computing resources
- Metadata is not just for description and discovery; it is also for contextualisation (relevance, quality, restrictions (rights, costs)) and for coupling users, software and computing resources to data (to provide a Virtual Research Environment)
- Metadata must be machine-understandable as well as human understandable for autonomicity (formalism)
- Management (meta)data is also relevant (research proposal, funding, project information, research outputs, outcomes, impact…)

RDA Element set

- Collected use cases
  - Across many domains
- Analysed for commonality
- Proposed original element set (Jeffery & Koskela)
- Minor changes suggested by RDA attendees
- Now checking applicability across domains
- Next detail the elements
- Then decide representation

- Unique Identifier (for later use including citation)
- Location (URL)
- Description
- Keywords (terms)
- Temporal coordinates
- Spatial coordinates
- Originator (organisation(s) / person(s))
- Project
- Facility / equipment
- Quality
- Availability (licence, persistence)
- Provenance
- Citations
- Related publications (white or grey)
- Related software
- Schema
- Medium / format

RDA Element set

- Collected use cases
  - Across many domains
- Analysed for commonality
- Proposed original element set (Jeffery & Koskela)
- Minor changes suggested by RDA attendees
- Now checking applicability across domains
- Next detail the elements
- Then decide representation

- Unique Identifier (for later use including citation)
- Location (URL)
- Description
- Keywords (terms)
- Temporal coordinates
- Spatial coordinates
- Originator (organisation(s) / person(s))
- Project
- Facility / equipment
- Quality
- Availability (licence, persistence)
- Provenance
- Citations
- Related publications (white or grey)
- Related software
- Schema
- Medium / format

Elements are NOT simple attributes
They have syntax (structure) and semantics

VRE Challenges

- Interoperation
  - Integrating metadata
- AAAI
  - To manage trust, security, privacy with authorisation of resource usage
- Composing workflows
  - A script for accessing, integrating, re-using assets for the research purpose
- Deploying workflows
  - Appropriately across the e-infrastructure

AAAI : authentication, authorisation, accounting infrastructure

- Controls user access
- Controls permissions for one entity to access and use another
  - Person utilises laboratory equipment
  - Person executes (web) service
  - (web) service accesses dataset [(read, update, delete)(time interval)(conditions of licence)…]
- Depends on information in the catalog
  - To authenticate user for the VRE (and the e-RIs available through the VRE) by multiple factors
  - To provide - to access control software - the parameters for authorisation permissions
  - To record activity and resource usage
    - Links to versioning, provenance, curation
- Ongoing European projects: AARC2

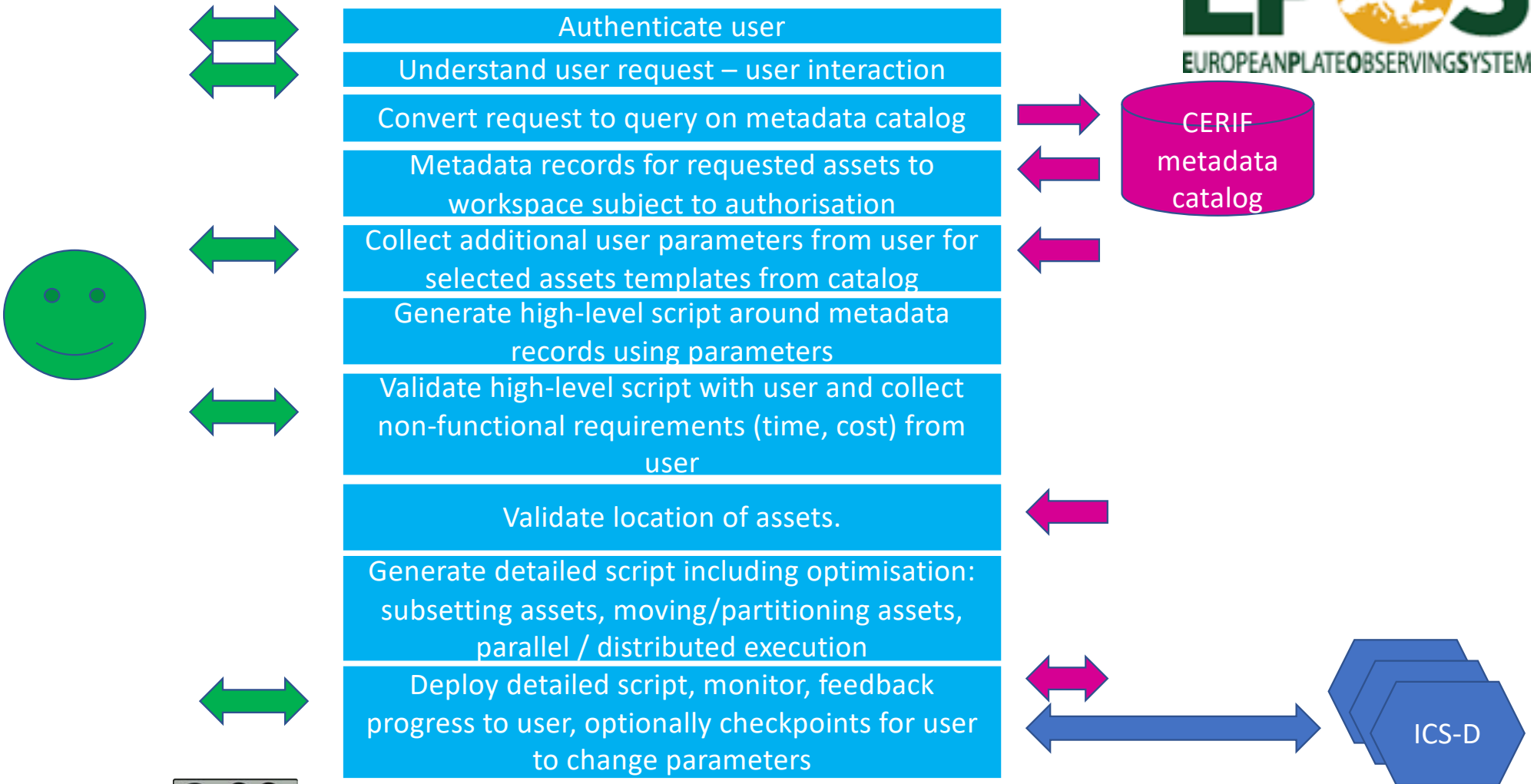
VRE Challenges

- Interoperation
  - Integrating metadata
- AAAI
  - To manage trust, security, privacy with authorisation of resource usage
- Composing workflows
  - A script for accessing, integrating, re-using assets for the research purpose
- Deploying workflows
  - Appropriately across the e-infrastructure

Composing workflows

- Gather into a workspace metadata records selected by query from the catalog i.e. the assets to be utilised
- Arrange in order noting where sequential and where parallel (dependencies)
- Pass to workflow engine / environment
  - E.g. Taverna, Kepler, more recently Jupyter Notebooks
- Manual or autonomic?
  - How much user interaction with the composition/orchestration process?

EPOS ICS-C



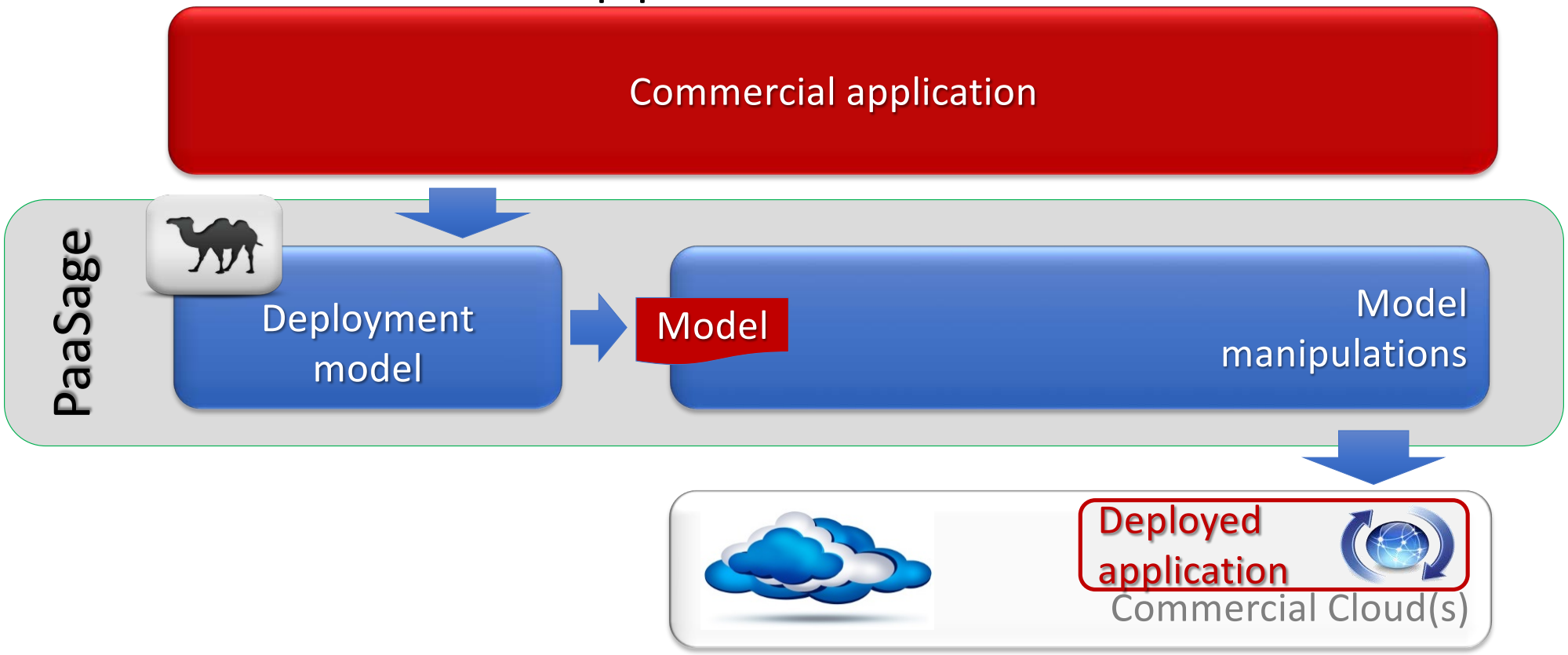
VRE Challenges

- Interoperation
  - Integrating metadata
- AAAI
  - To manage trust, security, privacy with authorisation of resource usage
- Composing workflows
  - A script for accessing, integrating, re-using assets for the research purpose
- Deploying workflows
  - Appropriately across the e-infrastructure

Deploying workflows

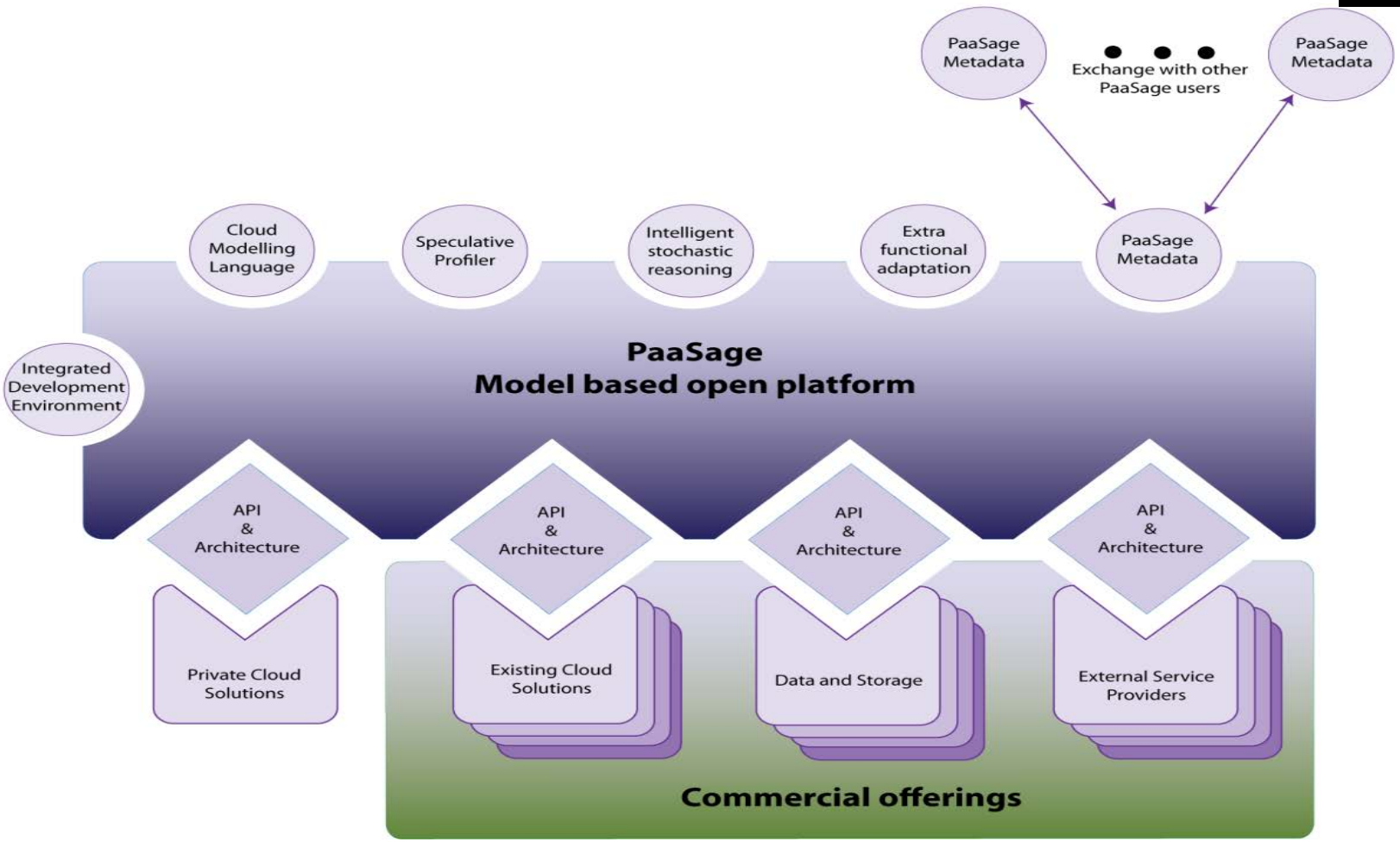
- Independent of target platform
  - Multiclouds, fog/edge, IoT (sensor networks)
  - EOSC (European Open Science Cloud)
- Optimised for
  - Performance | cost | security |privacy | elapsed time
- Note: data location / locality may be the major factor
  - Cost of transport and latency
  - Transport to less secure location
- So:
  - Data management to reduce datasets to minimum required locally / distributed
  - Process (software to the data) as much as possible locally / distributed
  - Transport derived data to location for final processing (analytics, visualisation, simulation)

Model-Driven Approach



uses

The Architecture



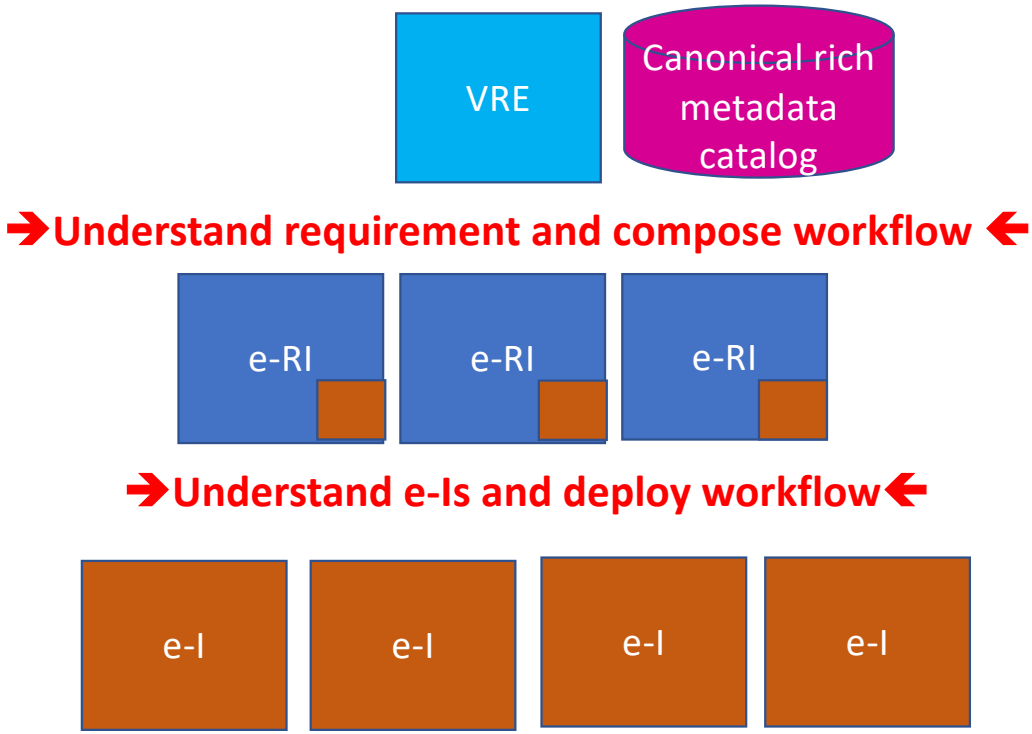
Data-Aware Deployment

- PaaSage as basis
- Improved CAMEL for
  - Data location / locality
  - Security / privacy
- Aimed at Big Data
  - HDFS Hierarchic Distributed File Systems e.g. Hadoop
- Ongoing project – real industry participation

VRE FUTURE

VRE Future: Take-home messages

- Architecture with clean interfaces
  - VRE / RI with assets / e-Infrastructure
  - Allows flexibility and choice of
    - Asset source
    - Processing offer
- Rich Metadata
  - To describe the 'world of interest' for the researcher
  - To assist in AAAI
  - To assist discovery / contextualisation / execution
    - FAIR +R (reproducibility) +R (resolvable UUPID)
  - To assist in workflow composition/orchestration/deployment



Acknowledgements

- CERIF https://www.eurocris.org/
- VRE4EIC https://www.vre4eic.eu/
- EPOS https://www.epos-ip.org/
- ENVRIPlus http://www.envriplus.eu/
- PaaSage https://paasage.ercim.eu/
- MELODIC http://melodic.cloud/
- RDA https://rd-alliance.org/
- The author acknowledges EC support for these activities directly or via projects; the EC grant references are within each project website