A Europe-wide Interoperable Virtual Research Environment
to Empower Multidisciplinary Research Communities and Accelerate Innovation and Collaboration

Why CERIF?

Keith G Jeffery Scientific Coordinator ERCIM

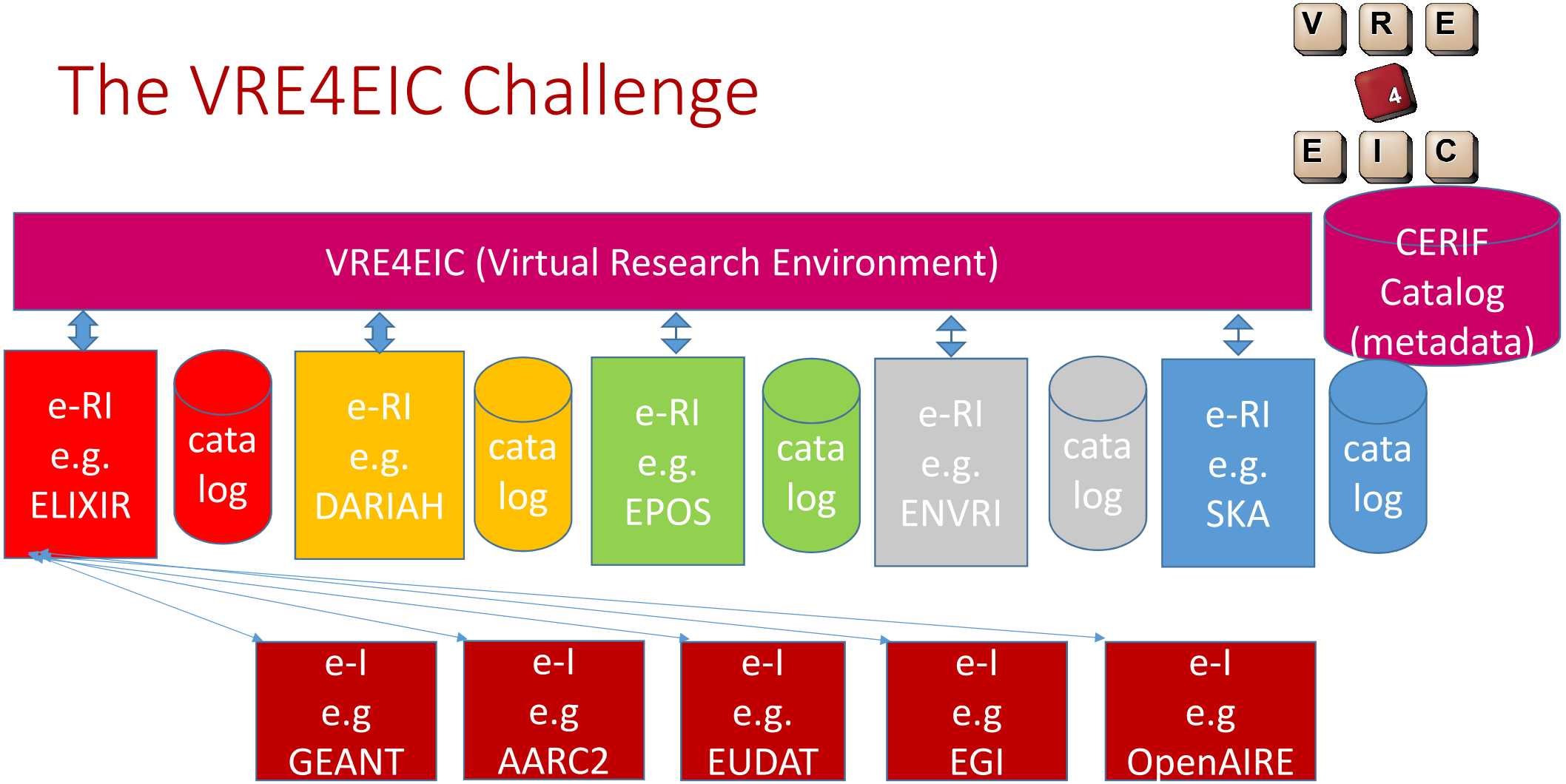
Anne Assserson euroCRIS

A brief history of Metadata

- Problem: making programmatic data structures persistent, externalised and virtualised
- Solution: database technology – schema - metadata
- Problem: needed richer metadata
  - Catalogues of data assets (library technology onwards)
  - Catalogues of software assets (software engineering)
  - Catalogues of computing resources (GRIDs then CLOUDs)
  - Catalogues of persons as users (AAAI)

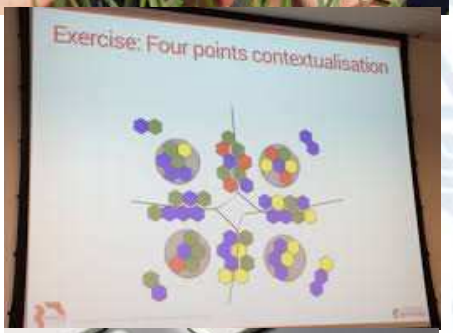
- All digital representations of objects are data
- The only difference between data and metadata is mode of use

The VRE4EIC Challenge



Need for Metadata

- Discovery
- Contextu
- alisation
- Action



FAIR Principles

Make your data:

- **F**indable
- **A**ccessible
- **I**nteroperable
- **R**eusable

Findable
- Descriptive metadata
- Persistent Identifiers

Accessible
- Determining what to share
- Participant consent and risk management
- Access status

Interoperable
- XML standards
- Data Documentation Initiative
- CDISC

Reusable
- Rights and licence models
- Permitted and non-permitted use

http://datafairport.org/

Data FAIRport
Find, Access, Interoperate & Re-use Data

Metadata Desiderata

- Formal syntax
  - Referential integrity
  - Functional integrity
- Declared semantics
  - Ontologies
  - multilingual

- To allow not only
  - Discovery
  - Contextualisation
    - relevance
    - quality
  - Action
    - Deployment
    - Execution
    - Interoperation
- by HUMANS
- but also by COMPUTERS

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.

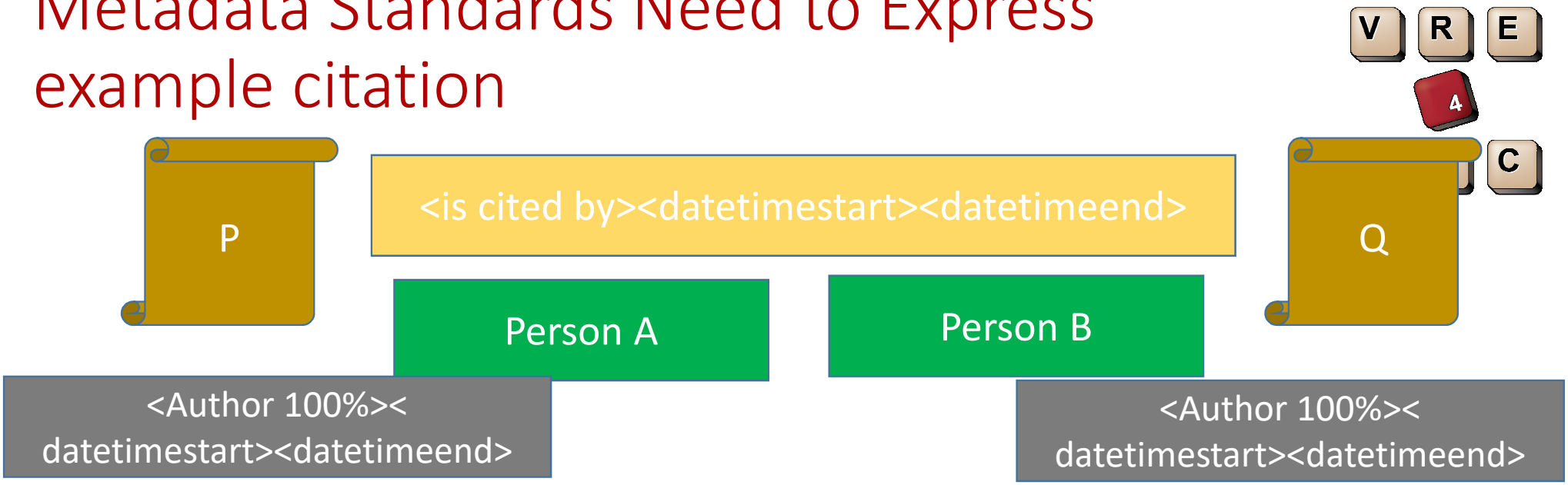
Senior lecturer is an academic rank.[1] In the United Kingdom, Republic of Ireland, New Zealand, Australia, and Switzerland, lecturer is a faculty position at a university or similar institution. The position is tenured and is roughly equivalent to an associate professor in the North American system. (Wikipedia)

<senior lecturer (UK)> <equivalence> <associate professor (US)> <equivalence> <maitre des conferences (FR)> ……

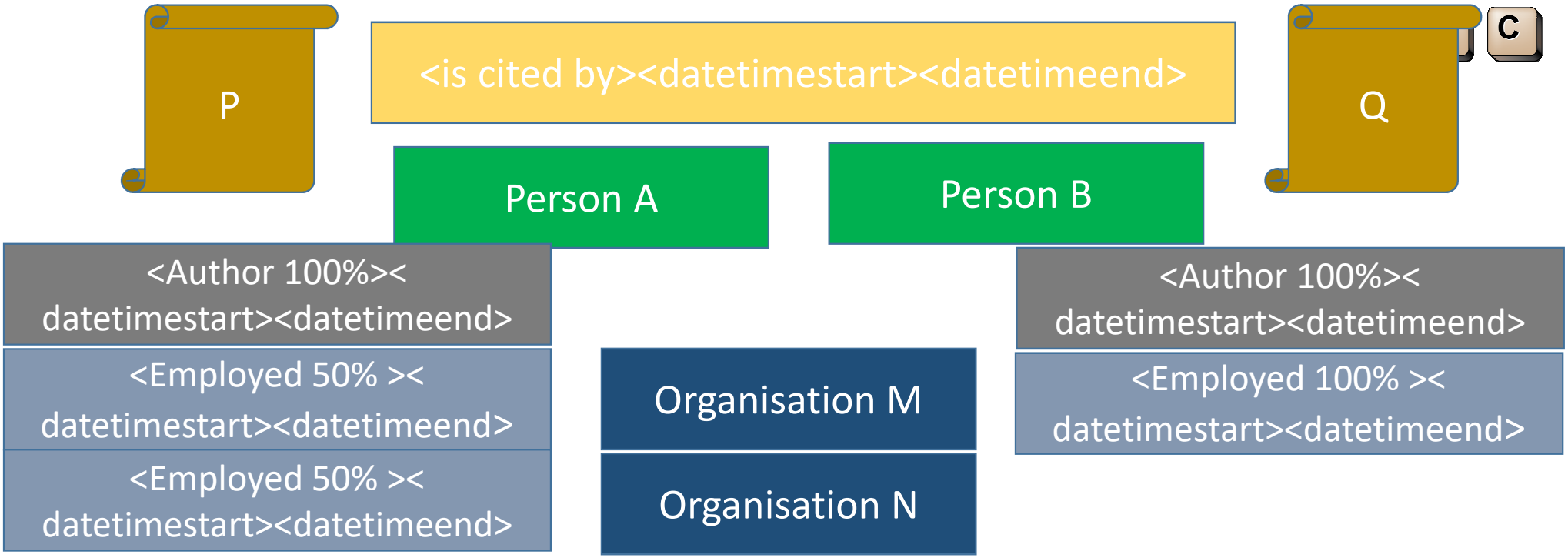
Particularly important for keywords and classification codes
e.g. Frascati ⇔ UDC ⇔ ……

Metadata Standards Need to Express example citation

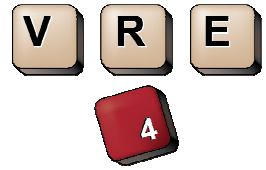
Metadata Standards Need to Express example citation



Metadata Standards Need to Express example citation



Metadata Standards Need to Express example citation



P

<is cited by><datetimestart><datetimeend>

Q

Person A

Person B

<Author 100%>< datetimestart><datetimeend>

<Author 100%>< datetimestart><datetimeend>

<Employed 50% >< datetimestart><datetimeend>

Organisation M

<Employed 100% >< datetimestart><datetimeend>

<Employed 50% >< datetimestart><datetimeend>

Organisation N

<leader> <datetimestart><datetimeend>

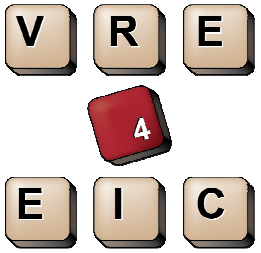
Project P

<participant 50%>< datetimestart><datetimeend>

Organisation X

<funded 100%> <datetimestart><datetimeend>

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

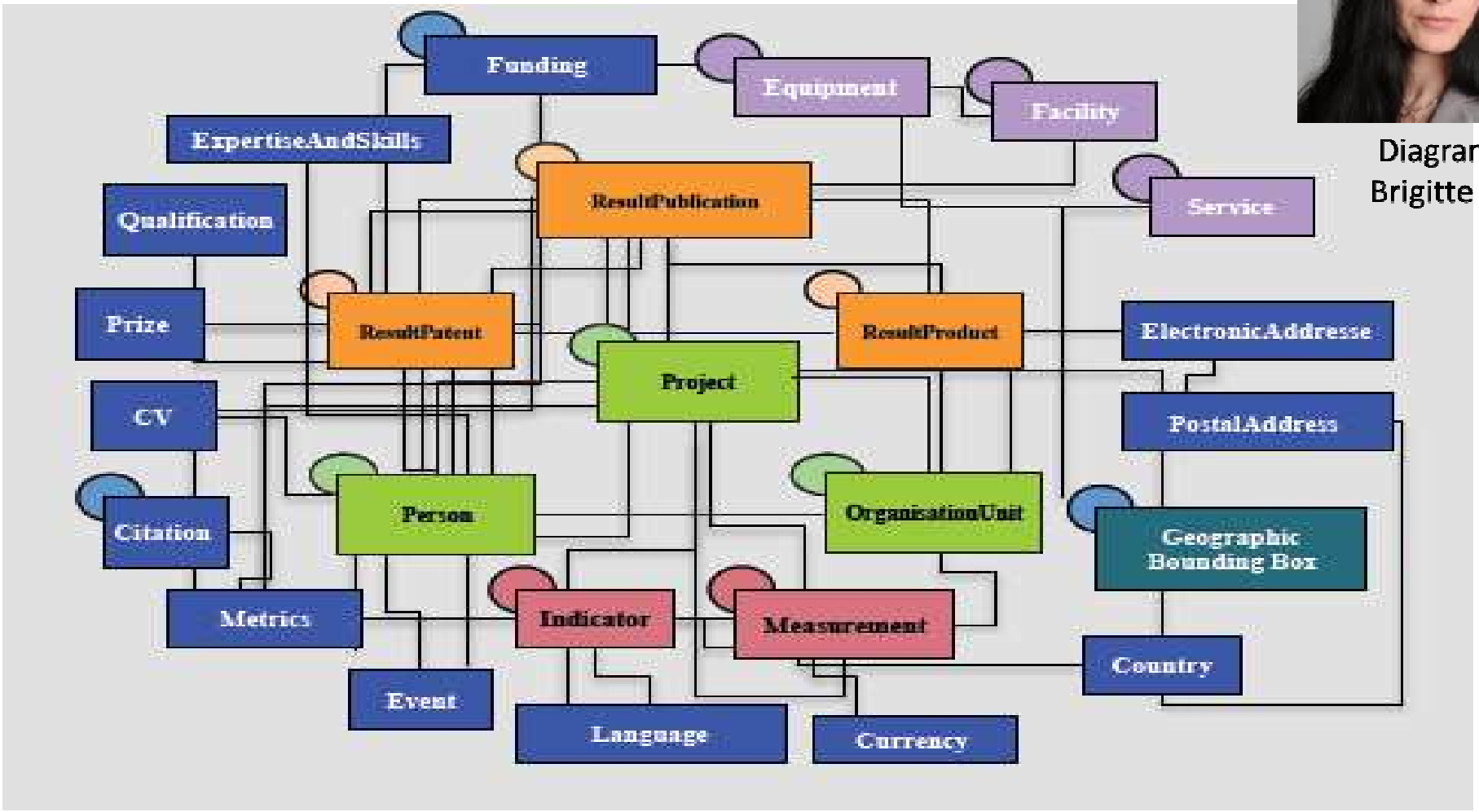
- This is essentially expressions in (decorated) first order logic – the elements in red are linking semantic relations
- Therefore can do **deduction** (facts from rules) and **induction** (rules from facts): this
  - reduces the effort of input,
  - increases the quality by consistency validation
  - and ensures actionable

Metadata Standards: CERIF

- **C**ommon **E**uropean **R**esearch **I**nformation **F**ormat
  - Data Model for exchange and storage of information about research
- 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 modelling (formalised by Jeffery & Asserson)
  - Base entities
  - Linking entities with role and temporal interval (i.e. decorated FOL)
- 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
  - SMEs providing CERIF systems , 2 bought up by Elsevier and Thomson-Reuters

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

- 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 e.g.
- And in leading-edge CLOUD project PaaSage

The Challenge

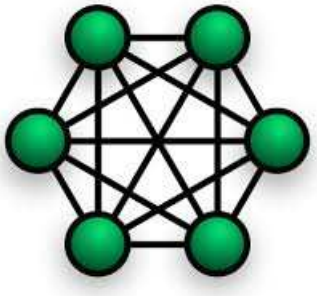
- Metadata Standards must be a good thing
  - there are so many of them
  - And commonly several dialects of each
- So to do research we need to interoperate to gain access to/re-use of assets described by metadata such as:
  - Datasets
  - Software modules
  - Services
  - Workflows
  - Instruments/detectors
  - Computing resources
  - Persons (experts)
  - Publications

So we need to interoperate across metadata standards

But…

- Metadata standards have different:
  - Elements (attribute names, types)
  - Syntax (structure within and between elements)
  - Semantics (meaning of values of elements or of components of elements)
  - Language
  - (I used to also say character set but Unicode has more-or-less solved this)

The n squared problem



- If we have *n* metadata standards
- Then to interoperate we need *n*(n-1)* mappings/convertors (almost *n**2*)
- ➔ define a superset canonical metadata standard and use that as a kind of 'switchboard'
- Then have *n* mappings/convertors (a considerable saving!)
- ➔ clearly this canonical metadata standard has to be
  - A superset of the elements from the *n* metadata standards to be interoperated
  - Richer in syntax and semantics than any of the *n* metadata standards to be interoperated
  - Interoperable with each of the *n* metadata standards to be interoperated
- CERIF mappings exist⇔ DC, DCAT, eGMS, CKAN, INSPIRE…..

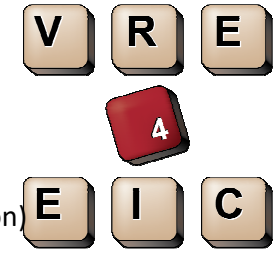
Research Data Alliance

- Metadata Interest Group
  - Metadata Standards Catalog Working Group
  - Data in Context Interest Group
  - Research data provenance interest group

- Data Formats & Types Working Group
- PIDs Working Group
- Data Fabric Interest Group
- (and many more)

RDA Element set

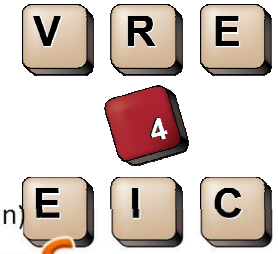
- Collected use cases
  - Across many domains
- Analysed for commonality
- Proposed element set (Jeffery & Koskela)
- Minor changes suggested by RDA attendees
- Now checking applicability across domains
- Next detail the elements
- Then decide representation



- 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 element set (Jeffery & Koskela)
- Minor changes suggested by RDA attendees
- Now checking applicability across domains
- Next define elements
- Then decide representation



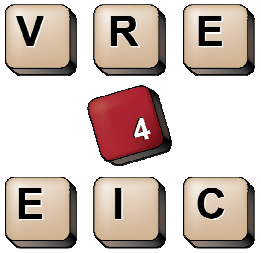
- Unique Identifier (for later use including citation)
- Location (URL)
- Description
- Keywords (terms)
- Temporal coordinates
- Spatial coordinates
- Originator (organisation(s) / person(s))
- Project
- Facility / equipment
- Quality
- Availability (licence, persistence)
- Provenance
- Citations
-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

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

RDA Element Set : Considerations

- The RDA element set is not unlike DDI
- It is much richer than DC, DCAT, CKAN, INSPIRE
- Its representation will be an interesting discussion
  - Likely follow FAIR principles
  - Formal syntax and declared semantics

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
  - Distributed, parallel
- 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 (batch processing)
- Now possible thanks to virtualisation
  - Metadata
    - Datasets
    - Software
    - Publications
    - Persons
    - Organisations
    - Equipment, facilities
    - GRIDs, CLOUDs

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

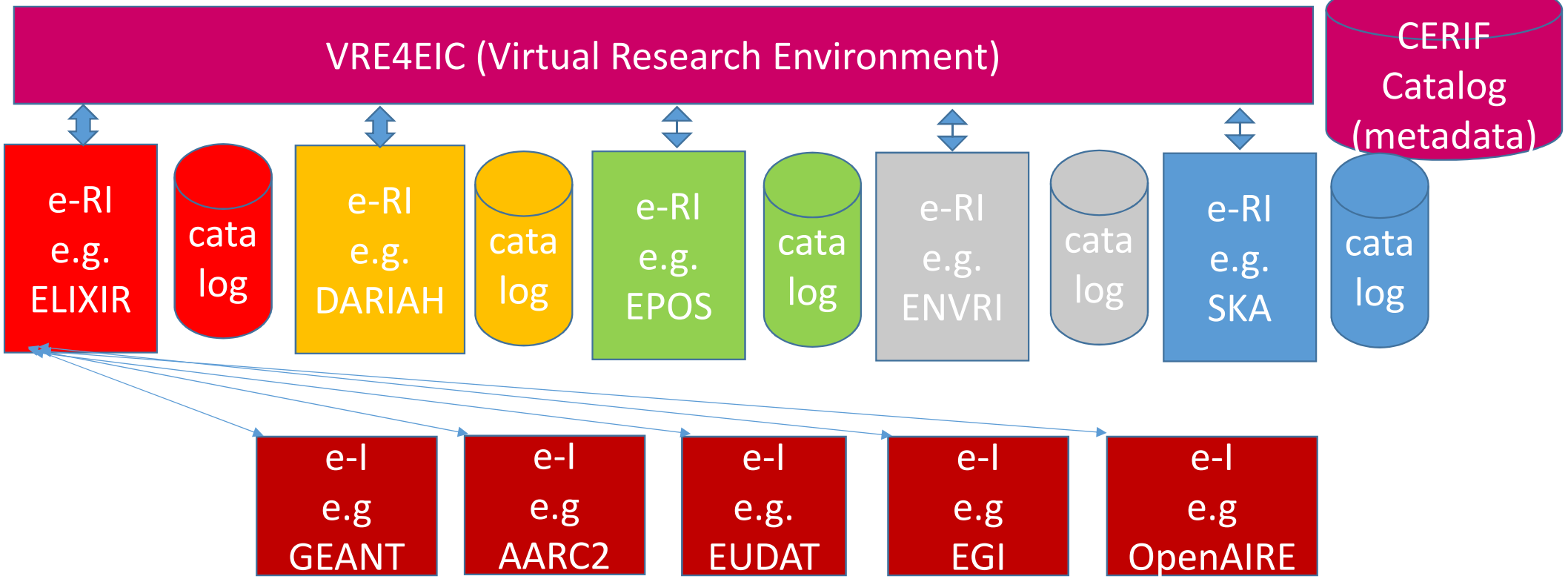
- Reviewing proposals and scholarly papers
- Cooperating with other researchers
- Evaluating against other researchers
- Working on editorial boards and conference programme committees

Requirement Metadata

- Metadata to describe (virtualise)
  - Datasets
  - Software components
  - Publications (white and grey)
  - Equipment/facilities
  - Computing resources
  - Workflows
  - Persons
  - Organisations
  - etc.

- FAIR principles
- Managed rights (ideally toll-free open access and utilisation)
- Suitable for
  - Discovery
  - Contextualisation (relevance, quality)
  - Action
- Re-use for new research
- Re-use for reproducibility
- Formal syntax, declared semantics

VRE4EIC Architectural Positioning



Questions

?

Acknowledgment

The VRE4EIC project has received funding
from the European Union's Horizon 2020 research and innovation
programme under grant agreement No 676247