Provenance

Keith G Jeffery, BGS/UKRI

Doron Goldfarb, EAA

What is Provenance?

Provenance

From French 'provenir' 'to come from'

/ˈprɒv(ə)nəns/

noun

1. the place of origin or earliest known history of something.

"an orange rug of Iranian provenance"

• the beginning of something's existence; something's origin.

"they try to understand the whole universe, its provenance and fate"

• a record of ownership of a work of art or an antique, used as a guide to authenticity or quality.

"the manuscript has a distinguished provenance"

ENVRI
FAIR

Provenience

� Distinguish provenance from provenience

� Provenience is where something has been found

� E.g. in archaeology or palaeontology

� Essentially the coordinates of the location

Provenance in Art: Painting

⪧ **Examples of fakes and forgeries in 2018**

⪧ In January, The Museum of Fine Arts in Ghent exhibited 26 fake works, which were on loan from a collector. The paintings, by 20th-century Russian artists Kazimir Malevich and Wassily Kandinsky, were thought to be fake by scholars who noted that the works were not included in any of the artists' catalog raisonnés.

⪧ _The Art Newspaper_ stated that "they have no exhibition history, have never before been reproduced in serious scholarly publications and have no traceable sales records." In addition, the museum did not carry out scientific analysis on the works because it is only standard policy for acquisitions, not loans.

ENVRI
FAIR

Provenance in Art: Painting

☞ What is missing is a historical record of:

☞ Exhibitions where the painting has been seen;

☞ Scholarly publications where the painting has been reproduced / discussed;

☞ Sale(s) of the painting

An example Kandinsky

Example of Provenance in Art

Drunken Satyr

Provenance	Exhibitions	Bibliography

1657	Found: villa of Domitian, Castel Gandolfo, Italy
1657 - 1693	Cardinal Flavio Chigi, 1631 - 1693, by inheritance to his son, Agostino Chigi
1693 - 1705	Agostino Chigi, Prince, Italian, 1634 - 1705, by inheritance to his son, Augusto Chigi, 1705.
1705 - 1728	Augusto Chigi, Prince, Italian, 1662 - 1744, sold to Frederik Augustus I as part of the Chigi collection, 1728.
1728 -	Frederick Augustus I, elector of Saxony and king of Poland, Saxon, 1670 - 1733, by inheritance within the family.
- 1918	Frederick Augustus III, king of Saxony, German, 1865 - 1932, taken during the revolution, 1918.
1918 - 1924	in the possession of the Freistaat Sachsen, returned to Haus Wettin (Saxon Royal House), July 21, 1924.
1924 - 1945/1948	Haus Wettin, taken by the Soviet government following World War II as part of the former royal collections.
1945/1948 -	Soviet government, ownership later transferred to the acting government (which became the Deutsche Demokratische Republik).
- 1999	Deutsche Demokratische Republik, returned to Haus Wettin, September 9th 1999.
1999 - 2001	Maria Emanuel, Margrave von Meissen (Haus Wettin), 1926 - 2012, sold to Alexander Rudigier, 2001.
2001 - 2002	Alexander Rudigier Ltd., sold to the J. Paul Getty Museum, 2002.

ENVRI FAIR

Acknowledgement Getty Center

Provenance in Art

- So, in this case, Provenance is a list of transactions
- But includes answers for each transaction to the questions:
- **When** : a date range
- **Where** : location where kept
    - including information about it e.g. House Wettin, Saxon Royal House
- **Who** : owned it having obtained it
    - including information about them e.g. Prince, Italian
- **Why** : it changed hands by inheritance, purchase or theft
    - Including more information e.g. during the revolution
- Had there been damage or modification then also the questions:
- **How** : has the asset changed
- **What** : changed it, e.g. a blow from a hammer

ENVRI

Provenance of Wine

Provenance: example

The following excerpt and slides are taken with permission from Moreau, L.
The Open Provenance Model: Towards inter-operability of Provenance Systems http://users.ecs.soton.ac.uk/lavm/talks/iam09.pdf

Example The **provenance of a bottle of wine** includes:

- Grapes from which it is made

- Where those grapes grew

- Process in the wine's preparation

- How the wine was stored

- Between which parties the wine was transported, e.g. producer to distributer to retailer

- Where it was auctioned

Ch Lafite Rothschild 1945

Provenance of Wine

- Note the additional information beyond recording the transactions in the example:
  - Grape from which the wine was made
- Commonly other information is also required for provenance for example:
  - Chateau/estate where grown
    - And linked to this the 'terroir', the owner…..
  - Winemaker
    - And her CV…….
  - Year of production
    - And linked to this the weather…..

- This is all contextual information (contextual metadata) used to assess the relevance and quality of the asset for the intended purpose

- The richer the contextual metadata, the more useful it is.

ENVRI
FAIR

What is Provenance in IT?

- A record of the history of an asset

- When, what, how, who, why, where
    - **When** was the asset changed (date/time start, date/time end)
    - **What** changed it (usually a software service/process)
    - **How** was it changed (what is the difference between the previous and current state)
    - **Who** caused it to be changed (relates to authentication and authorisation)
    - **Why** was it changed (the reason for the change – e.g. to correct an error)
    - **Where** was the change made (e.g. a computer centre)
- This is all additional information to the basic metadata (catalog) information about the asset
- Note: most existing provenance systems in IT are much simpler and more primitive

ENVRI FAIR

How does Provenance Relate to Research Digital Objects

Find:

- We may only be interested in finding assets that have gone through a given process, or been generated by a particular piece of equipment, or in a specific laboratory, or by a particular person or team...

Access:

- We may only be interested in accessing assets that are relevant, suitably licensed and of appropriate quality

Interoperate:

- We may wish to interoperate assets of equivalent quality (e.g. precision, accuracy)

Reuse:

- We may wish to reuse assets with a known history e.g. a data product derived from raw data through various calibrations, normalisations, conversions...

ENVRI
FAIR

How does Provenance Relate to Research Digital Objects

🌊Find:
- 🌊 We may only be interested in finding assets that have been through a given process, or been generated by a particular piece of equipment, or in a specific laboratory, or by a particular person or team…

🌊Access:
- 🌊 We may only be interested in accessing assets that are relevant, suitably licensed and of appropriate quality

🌊Interoperate:
- 🌊 We may wish to interoperate assets of equivalent quality (e.g. precision, accuracy)

🌊Reuse:
- 🌊 We may wish to reuse assets with a known history e.g. a data product derived from raw data through various calibrations, normalisations, conversions…

CONTEXTUALISATION by rich metadata

ENVRI FAIR

Why is Provenance Important

- **contextualisation**
  - relevance and quality of the asset – properties that may be used in many ways;

- **reproduction of results (scientific reproducibility – reuse)**
  - 'replaying' the workflow;

- **checkpoint/restart**
  - To avoid losing previous work in a workflow and starting from a known state;

- **audit/governance**
  - To record what has been done to an asset in its lifetime for later assessment of validity

ENVRI
FAIR

What does provenance relate to?

- ## catalog of assets
  - Ideally the provenance information should be integrated with the assets catalogued
  - Thus it can be used in finding and accessing (contextualising) assets;

- ## logging – and recovery
  - Since provenance records what has happened to an asset it can be used to restore to a previous state (utilising curation which is closely related) and then execute the successive workflow steps

- ## security and privacy
  - Provenance should record every access to an asset and thus can be used to check for security breaches (unauthorised access) or privacy breaches (access to protected person data)

- ## authentication, authorisation and access control
  - Provenance should record authorised access to an asset by an authenticated user or software and with what privileges (permissions, responsibilities) the access was made

ENVRI
FAIR

What does provenance relate to?

- Thus provenance records have to be linked to (with qualified references in FAIR terms i.e. semantically meaningful relationships including temporal validity) at least some of:

The digital object	Organisation	Person
Software service	Document	Project
E-Address	P-Address	Licence
Software services	Facility	Equipment
Location (coordinates)		
Indicators	Measurements	………

- and entities related to them e.g. CV for Person Funding for Project…..

Conclusion

- So Provenance needs to be:
  - Integrated in/with the asset catalog
    - For contextualisation
  - Represented by a formal syntax
    - For autonomic processing
  - With referential and functional integrity
    - For reliable processing
  - With declared semantics
    - For consistent processing

ENVRI
FAIR

SUMMARY

WHAT IS REQUIRED

- History of an asset
  - Dataset
  - Service
  - Workflow
- And at each step what else is involved:
  - Person
  - Organisation
  - Process
  - E-Infrastructure resources

Classical Provenance in IT systems provides ONLY Data Provenance

ENVRI

SUMMARY

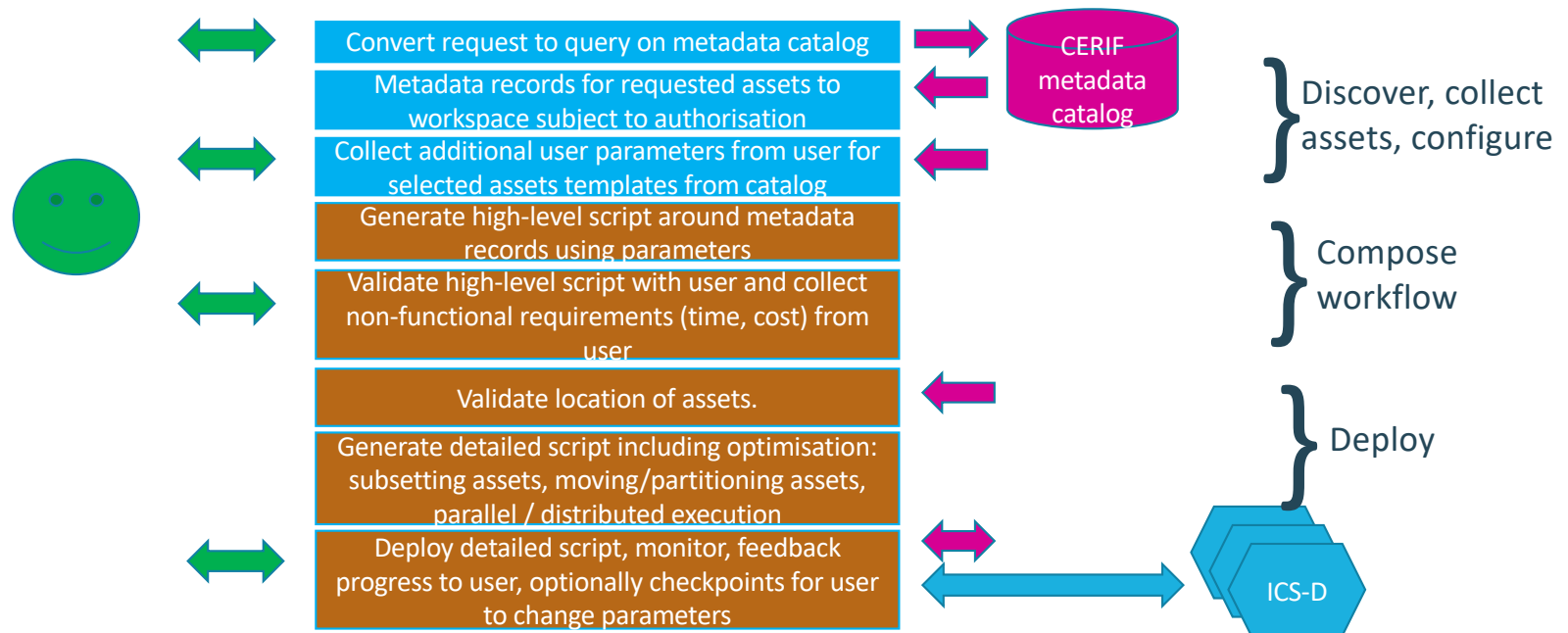## WHAT IS REQUIRED

- History of an asset
  - Dataset
  - Service
  - Workflow
- And at each step what else is involved:
  - Person
  - Organisation
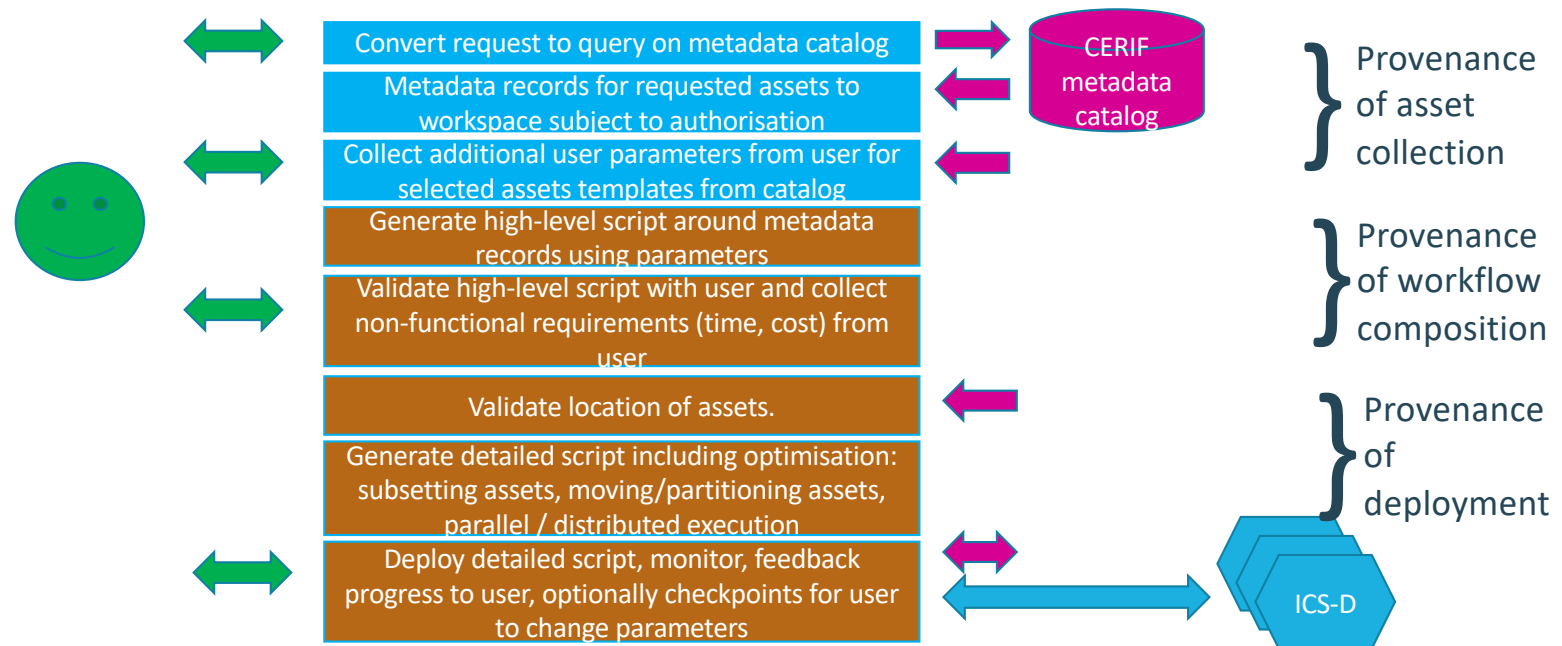 - Process
  - E-Infrastructure resources

CERIF Provides
- Data (and other asset) Provenance
- User activities (logging activities on the system)
- Editing metadata with the (versioning)

EPOS Workflow Design

EPOS-ICS-C

EPOS
EUROPEANPLATEOBSERVINGSYSTEM

| Convert request to query on metadata catalog |
| Metadata records for requested assets to workspace subject to authorisation |
| Collect additional user parameters from user for selected assets templates from catalog |

CERIF metadata catalog

} Discover, collect assets, configure

| Generate high-level script around metadata records using parameters |
| Validate high-level script with user and collect non-functional requirements (time, cost) from user |

} Compose workflow

| Validate location of assets. |

| Generate detailed script including optimisation: subsetting assets, moving/partitioning assets, parallel / distributed execution |
| Deploy detailed script, monitor, feedback progress to user, optionally checkpoints for user to change parameters |

} Deploy

ICS-D

Not yet implemented fully: work in progress

ENVRI
FAIR

EPOS Workflow Design

EPOS ICS-C

EPOS
EUROPEANPLATEOBSERVINGSYSTEM

Convert request to query on metadata catalog

Metadata records for requested assets to workspace subject to authorisation

Collect additional user parameters from user for selected assets templates from catalog

Generate high-level script around metadata records using parameters

Validate high-level script with user and collect non-functional requirements (time, cost) from user

Validate location of assets.

Generate detailed script including optimisation: subsetting assets, moving/partitioning assets, parallel / distributed execution

Deploy detailed script, monitor, feedback progress to user, optionally checkpoints for user to change parameters

CERIF metadata catalog

ICS-D

} Provenance of asset collection

} Provenance of workflow composition

} Provenance of deployment

Not yet implemented fully: work in progress

ENVRI
FAIR

CERIF and Provenance

Keith G Jeffery, BGS/UKRI

Doron Goldfarb, VUT

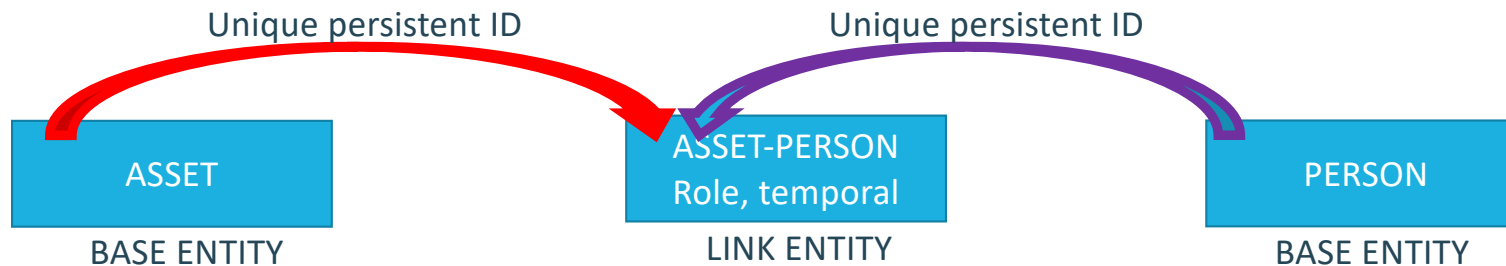# CERIF: Common European Research Information Format

- EC Expert Group
  - 1987-1991 ➜ CERIF91: like the later Dublin Core. Tested, failed.
  - 1997-2000 ➜ CERIF2000: full extended temporal entity-relationship model
    - Formal syntax, referential and functional integrity, declared semantics
- EU Recommendation to Member States
- Used extensively in research-funding and research-producing organisations
- Used outside Europe (all continents)
- Inside products of Elsevier and Thomson-Reuters
- Used by European Research Council
- Used in OpenAIRE and ORCID
- Has convertors to/from many other metadata standards because:
  - Superset, canonical, rich metadata standard
  - Can be used as a 'switchboard' for interoperation

https://www.eurocris.org/services/main-features-cerif

ENVRI
FAIR
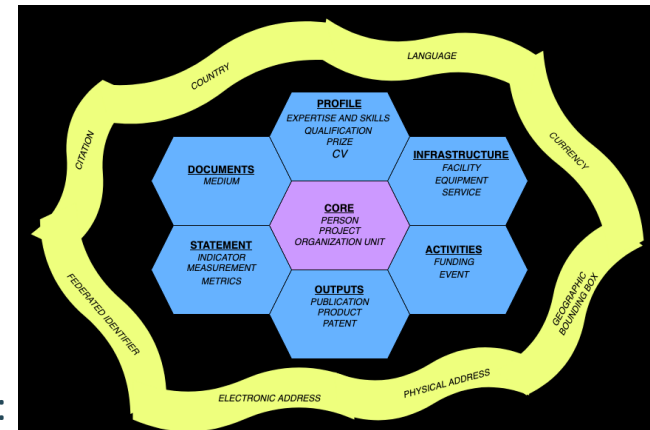
CERIF: the basic model

- Base entities: objects in the real world e.g. Asset, Person which have attributes
- Link Entities: relationships between base entities with attributes including 'role' and 'temporal duration'

Unique persistent ID   Unique persistent ID

ASSET	ASSET-PERSON Role, temporal	PERSON
BASE ENTITY	LINK ENTITY	BASE ENTITY

- Example Link entity instance:

-    ASSET   ROLE   TEMPORAL INTERVAL   PERSON
- <Paper X>[<authored by> <19970128 00:00:00><19980812 00:00:00>]<Person P>

CERIF: syntactic and semantic layers

- Syntactic layer: referential and functional integrity

- Semantic Layer: semantic consistency and interoperability

- In fact 'authored by' is not represented as a text string in the role attribute of the link entity but as a UUID which points to the semantic layer where a role 'authored by' is stored in a base entity and can be related to textual terms in the same or other base entities by link entities thus allowing full ontological capability including multilinguality (see "scritta da")
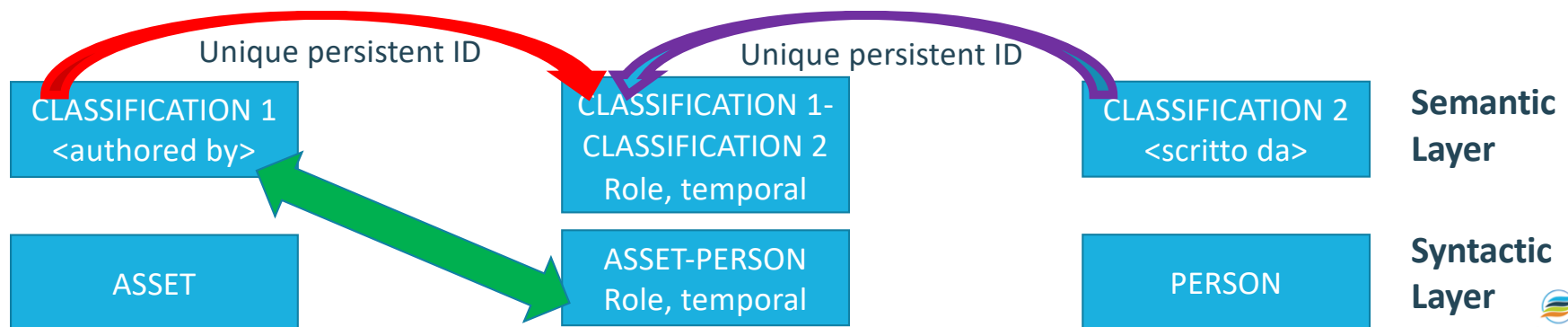
CERIF: syntactic and semantic layers

- Syntactic layer: referential and functional integrity

- Semantic Layer: semantic consistency and interoperability

- In fact 'authored by' is not represented as a text string in the role attribute of the link entity but as a GUID which points to the semantic layer where a role 'authored by' is stored in a base entity and can be related to textual terms in the same other base entities by link entities thus allowing full ontological capability including multilinguality

CERIF Provenance

- Provenance in CERIF is recorded in the link entities and is thus ***intrinsic*** to the catalog

- Link entity ASSET-PERSON
  - <Asset A>[<created by> <20200110 09:01:00><20200110 09:55:00>]<Person P >
  - <Asset A>[<edited by>   <20200224 15:05:00><20200225 18:22:00>]<Person Q>
  - <Asset A>[<accessed by><20200301 09:03:00><20200301 09:05:00>]<Person R>
- And so on.

- Querying the Link entity ASSET-PERSON for all instances of <Asset A> provides the asset provenance (when ordered by either start date/time or end date/time)

- Querying the Link entity ASSET-PERSON for all instances of <Person R> provides the user provenance (when ordered by either start date/time or end date/time)

CERIF Provenance with Contextualisation

- Provenance in CERIF is recorded in the link entities and is thus *intrinsic* to the catalog

- Because <Asset A> is an instance in a base entity, and has relationships expressed as instances in link entities to other base entities, for example:
  - <Asset A><is licensed>   <20200110 09:01:00><null><CC-BY-NC>
  - <Asset A><is documented><20200110 09:01:00><null><Document D>
- We have:
- a rich fully connected graph with all information related to the asset
- linked by qualified relationships in FAIR terms
- available for all the purposes of provenance
- because it is all within the metadata catalog in an integrated structure

ENVRI FAIR