KNAW Humanities Cluster

# Introduction
# Provenance and Research Infrastructure

Daan Broeder - KNAW HuC DI

CLARIAH Provenance Workshop

DenHaag, Sept 3 2018

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

# The provenance of 'provenance'

KNAW Humanities Cluster

*Provenance refers to the sources of information, such as entities and processes, involved in producing or delivering an artefact**

- Originally primarily the chronology of ownership, custody or location of an historical object with main goals establishing ownership and authenticity
  - Art, Archives, Books/Manuscripts, Wines
- Science - adding goals of establishing usage and meaning of the artefacts
  - Archaeology, Paleontology, Anthropology
- (Research) Data provenance
  - Digital data and processing gives many extra challenges wrt the many ways that data can be produced and modified

- Database theory theory concepts are used as: *why-provenance*, *how-provenance*, *where-provenance* and provenance management strategies as *eager - and lazy provenance*. But this had only limited influence on our current usage

*from the W3C Wiki

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

# Why this CLARIAH provenance workshop?

KNAW Humanities Cluster

- Immediate cause is the discussion we had at the 'CLARIAH tech dag' about the WP3 VRE where tracing provenance by a VRE was demonstrated
  - Is it needed?
  - What use is provenance tracing for the researcher?
  - What is the priority for implementing this?
- Secondly provenance concept attracts attention addressing research reproducibility
- Many e-infrastructure and research infrastructure initiatives have given attention to provenance: DataOne, DataConservancy, EUDAT, RDA …, is it relevant to CLARIAH?
- CLARIAH WPs and tasks should collaborate on this subject

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

# Provenance discussion scope

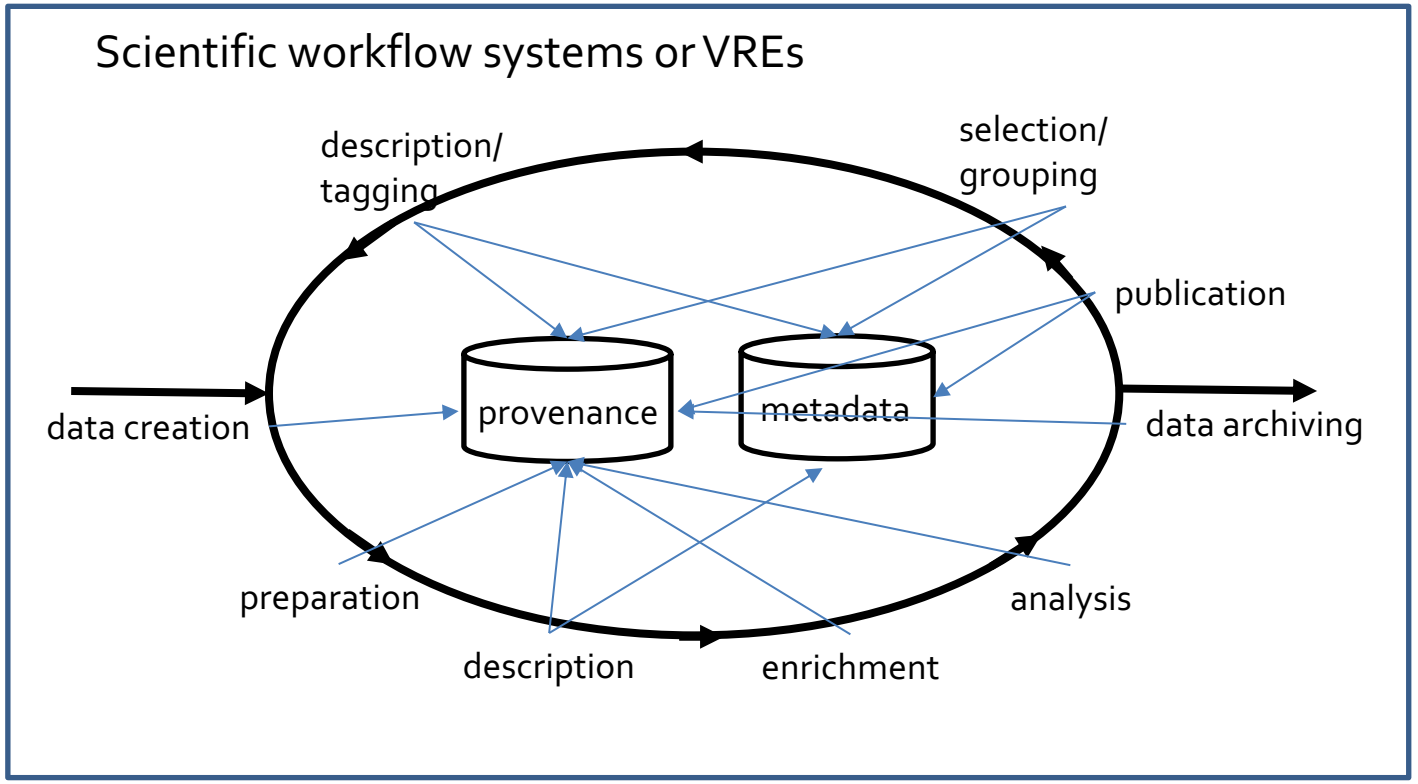
KNAW Humanities Cluster

- Our scope is research data and services in the whole data life cycle
  - Data-creation and enrichment: developer, researcher
  - Publication and archiving: data-manager, archivist
  - Citation: data-manager and researcher,
  - Sharing for reuse: researcher, funder
  - Verification and (peer-)review: researcher, funder
- Provenance discussion involves many already existing practices and implementations e.g. metadata, but imposes some extra requirements wrt. procedures and responsibilities
- Important to include are scalability and sustainability considerations
  - compromising between manageability and providing sufficient information
  - putting responsibility where it belongs and can be sustained

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

# Provenance and Research Infrastructure I

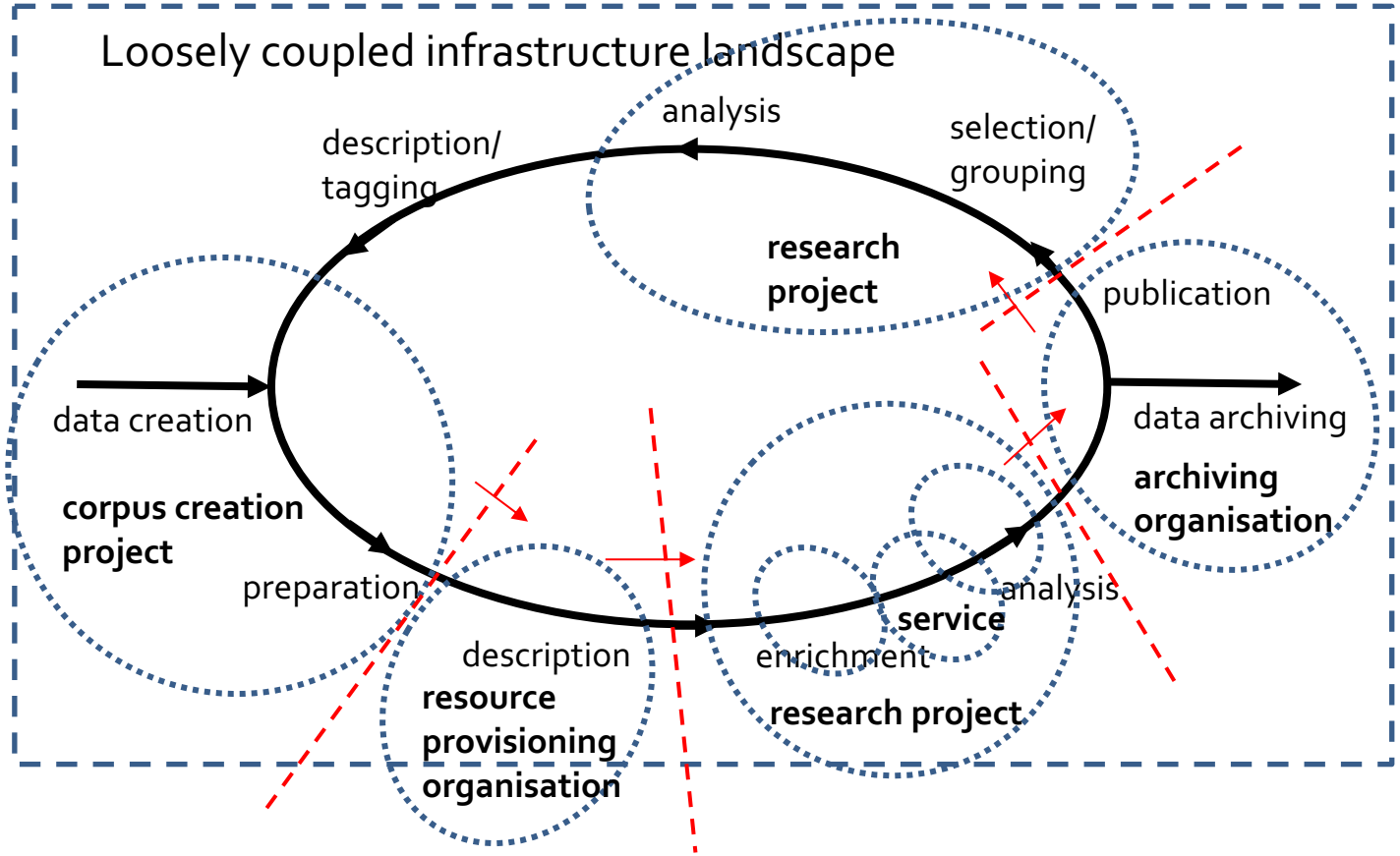
- Dedicated research environments as research workflow systems and VRE's would allow:
- consistent tracking and description of data processing
- support the researcher with automatic bookkeeping
- Using 'central' registries for provenance and metadata
- Enabling maximal granularity and specificity of provenance tracing



# Provenance and Research Infrastructure II

The existing research infrastructure landscape is loosely coupled:

- Organisations & projects are autonomous entities
  - own procedures and responsibilities
  - sometimes limited expertise
  - limited tooling
- Tools & services are from independent providers, with limited (provenance) interoperability
- Central provenance management is not scalable, no single responsible party
- Should rely on minimal provenance provisioning by the different components and organisations



# In reality…

- VRE scenario also has external dependencies
  - Archiving services
  - Resource provisioning
  - External processing services
- Distributed scenario can be improved
  - What to publish
  - Standardization, how to publish

- We can do both!
- … but I think there are priorities

Crucial need to:

1. **determine core provenance info for research** needs (see also existing metadata)
2. provide interoperability between different provenance descriptions via **shared vocabularies or registries** for:
   - People, organisations, data-sets, software, …

# Researcher requirements

Currently different metadata schema usually already contain some provenance information to answer:

- “When was it created”
- “Who is responsible”
- …

What more is needed or useful?

- What is of interest when using data?
- What is of interest when producing/modifying data
  - for sharing with others or
  - for reproducibility

# Identification …

We (kind of) understand the data domain

- PIDs for  persistent identification
- Checksums for verification
- Mature metadata schema, landing pages, versioning schemes
- Procedures for achieving long term persistency

Vocabularies for persons, organisations etc. already discussed within metadata scope

- But should provide adequate resources

Do we understand the software domain? It has unique aspects

- many external dependencies complicate stable source code references and fingerprinting
- influence of configuration parameters on service operations
- Pipe-lines that include many subcomponents

Nevertheless many new opportunities exist

- PIDs for container wrapped services
- self identification of services

# Results for today?

KNAW Humanities Cluster

- **Information exchange**: what are we doing wrt. provenance, what looks useful, what is required
- Start of **further engagement with the researchers** on the provenance topic to get their feedback and requirements
- We present a suitable cross-section of the DLC stakeholders to start **investigating strategies for provenance information exchange in CLARIAH**
- Find possible **common technology approaches**
  - Special topics also with bearing beyond CLARIAH provenance as **software service identification**

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

Thank you for your attention