Provenance (for Earth science data)

DKRZ-Seminar, Oct 15 2012

Tobias Weigel
Deutsches Klimarechenzentrum (DKRZ)

Agenda

- What is provenance? Why do we care?
- Alternative Provenance definitions
- Gathering and representing provenance information
- Further resources

What is provenance?

And what does it mean
in our context?

Provenance: Definition

- For data produced by computer systems:
  - „The provenance of a piece of data is the process that led to that piece of data.“ (Moreau 2010)
- This is a generic base definition.
- Other terms: data lineage, history

- (provenance also applies to food ingredients, works of art, ...)

Our context...

- What is our context?
  - digital-born ESM output data
  - observational data, e.g. remote sensing imagery
  - various processed derivates
- What are its characteristics?
  - complex, non-standardized toolchain
  - various processing steps by various actors
  - no single infrastructure

Use Cases (1)

- Quality of scientific data
  - The processing history of a data object forms an important part of its scientific context.
  - Users who did not create a data product must be able to understand the implications that went into its creation.
  - Data may be reused many years after creation.

Use Cases (2)

- Reproducibility
  - If processing steps are recorded in detail, a future user may reproduce them to get the exact same results
- May be impossible for ESM output data in all its depth
  - We can‘t archive the supercomputer itself
- Yet, try to capture as much as possible

Use Cases (3)

- Attribution
  - Give credit to the original data producer
  - Citing a DataCite DOI may not be enough
  - Who is using data that is generated with DKRZ‘s resources?
- Provenance can enable anyone to trace back to the original source and producer

Data-intensive science

- The scenarios grow more important with data-intensive science
  - Data is shared across scientific communities
  - Focus shifts from data production to data analysis

The Fourth Paradigm (2009)

Provenance and the data life cycle

- Provenance may cover the whole data life cycle
- Here: focus on earlier parts
  - data generation
  - data processing

Alternative provenance definitions

There‘s more than one!

The task

- The task here: Develop an understanding of provenance that is specific and pragmatic in our context.

Provenance definitions

1. Why-Provenance
2. Where-Provenance

} Database context

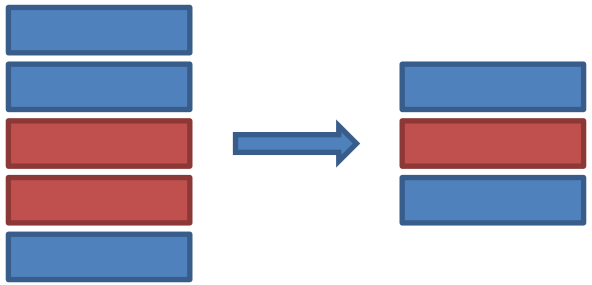
3. Provenance as a process
4. Provenance as a Directed Acyclic Graph

(there are more...)

Moreau (2010)

Why-Provenance

- Context of database queries
- Why-Provenance: „tuples whose presence justifies a query result“
- „Why is X part of the result?“
  - „Because the queried input data contains tuple A“



Moreau (2010), Buneman et al. (2001)

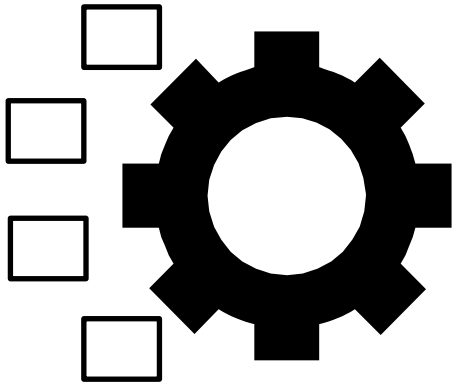
Where-Provenance

- A Website displays a typo in a menu entry
- What is the database field this string comes from?
  - This may not be the database directly connected to the website, but e.g. a citation database maintained elsewhere and queried by the site
- Helps to illuminate the copying of information across databases.

Moreau (2010), Buneman et al. (2001)

Provenance as a process (1)

- The computation that resulted in the data
- Any
  - data
  - event
  - user action



- that can be connected to the data through a computational process potentially belongs to its provenance

Moreau (2010)

Provenance as a process (2)

- ESM execution: The context can get very vast.
  - model source code
  - all parameters, model conditions, forcings
  - user name, libraries, OS version, parallel architecture
  - ...

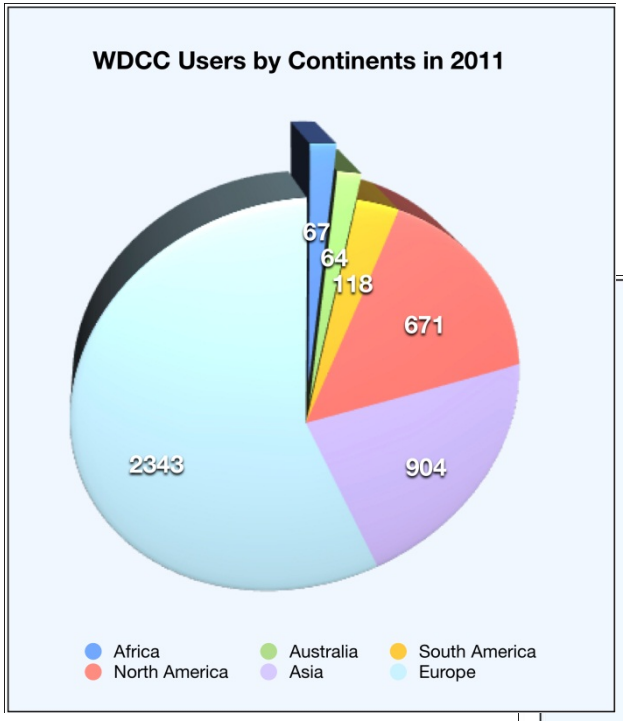
Moreau (2010)

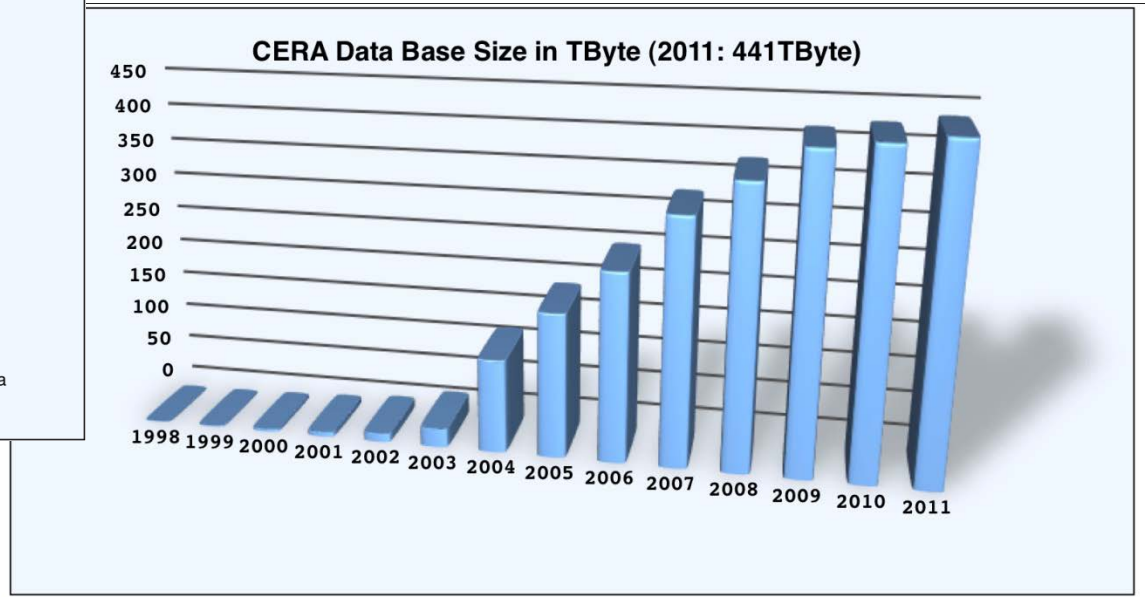
Provenance as a DAG

- What is a Directed Acyclic Graph?
- ... What is a graph?

What is not a graph?

- These are no (mathematical) graphs.





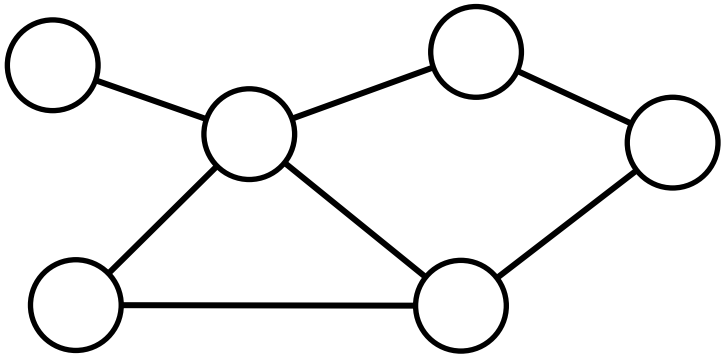
What is a graph?

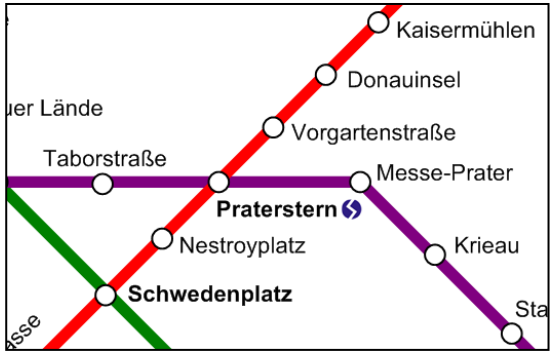
- This is a graph. Graphs are everywhere.

Wikipedia

What is a graph?

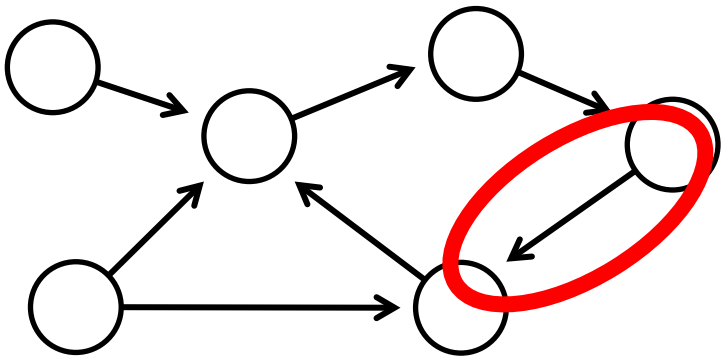
- Graph theory: A graph consists of a set of nodes (vertices) and a set of edges
  - $G = (V, E)$
  - any $e \in E$ is an unordered set (v_1, v_2); $v_1, v_2 \in V$





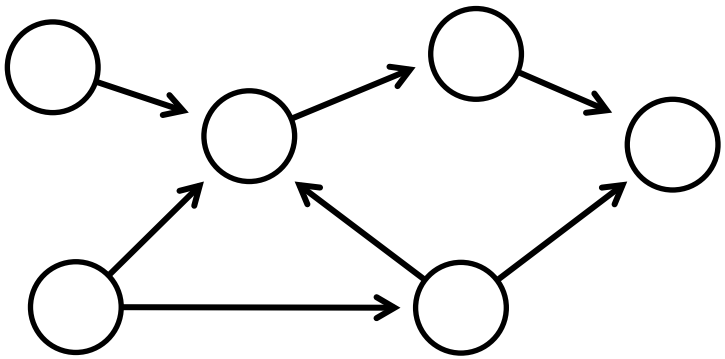
What is a directed graph?

- Directed graph
  - directed edges
  - set (v_1, v_2) is ordered; the edge is directed from v_1 to v_2



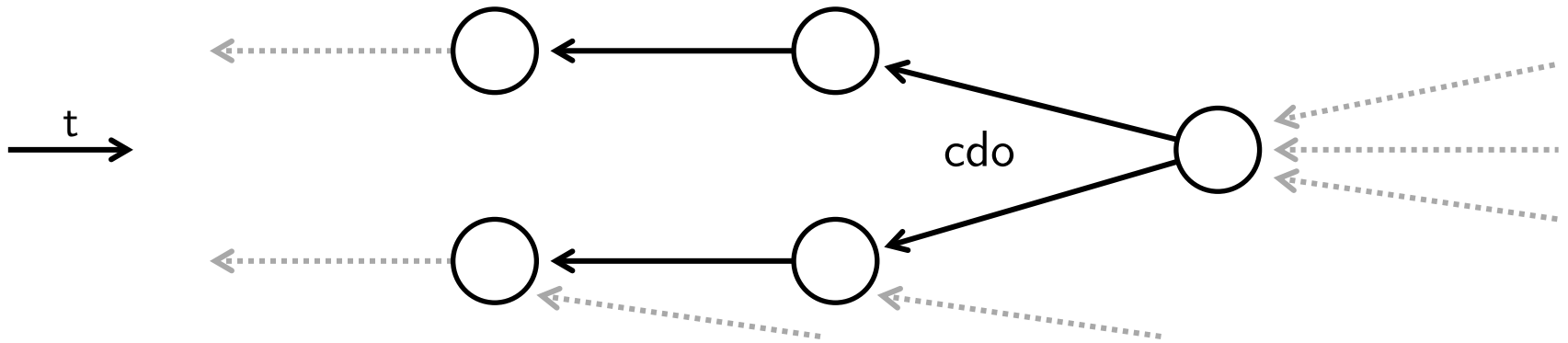
What is a Directed Acyclic Graph?

- Directed Acyclic Graph (DAG)
  - directed edges
  - no cycles allowed!



Provenance as a Directed Acyclic Graph

- Simplified, data-centric view
- Nodes represent data items
- Edges represent derivative operations
  - „predecessor“, „successor“, „derived-from“, ...
  - uni- or bidirectional
- Level of detail depends on the use cases



Moreau (2010)

Provenance information is part of the metadata

- Provenance information is part of the metadata
- Curating this metadata is tedious and pragmatically impossible
- It is agreed that provenance gathering must be automated
- View a provenance record as something created on the fly, rather than a stored document
  - **provenance is the result of a query over process assertions** (Moreau 2010)

Applications

What is out there to gather, represent, exploit provenance?

Gathering provenance information

- Many tools exist to capture provenance through an embracing system (particularly workflow systems)
  - Lots of research and academic prototypes
  - A list is available at http://www.openprovenance.org
- Provenance information may be aggregated in a specific database *(provenance store)*

Gathering provenance: workflow systems

- Scientific workflow systems
  - e.g. Taverna, Kepler, VisTrails, ...
- Advantages
  - Potentially good coverage
  - Improves collaboration and knowledge transfer
- Disadvantages
  - all or nothing
  - high migration costs
  - do not match user‘s traditional workflow

Gathering provenance information - alternative

- For us: no overarching system possible in the mid-term
- Alternative idea: capture proveance in small pieces by enhancing the existing tools of the research environments

Gathering provenance: in small steps

- Advantages
  - results scale well with implementation effort
  - potentially small change to user workflow
- Disadvantages
  - fragmentary coverage
  - incoherent, potentially chaotic information
  - mandates strict standardization

Representation of provenance information

- Provenance information can be represented in many formats
  - human-interpretable: human-adressed log files, free text
  - machine-interpretable: traversable graphs
    - simple (A derived from B)
    - complex (semantic graph, Open Provenance Model)

Machine-interpretable representation required?

- Machine-interpretable representation of provenance: what is the desired level of detail?
  - more detail → more sophisticated representation language required
- So the core question is: does your use case require sophisticated machine-interpretable representation?
  - remember: machine-interpretability is for tools, not for humans

Representation formats

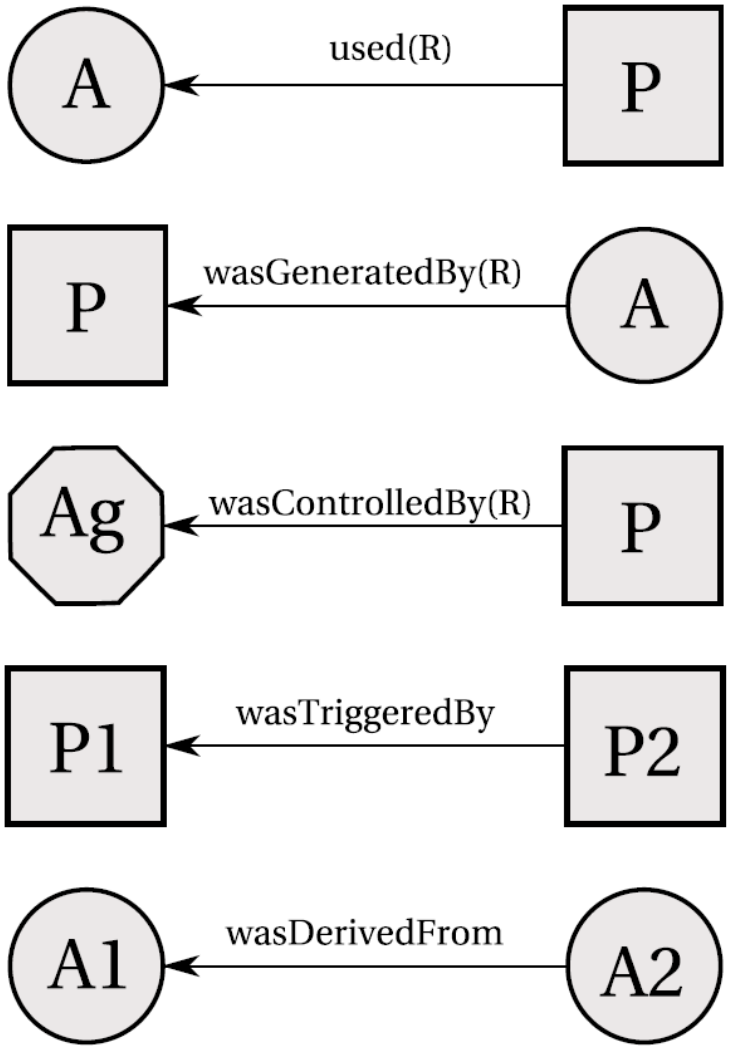
- Machine-interpretable representation of gathered information
- Designed to span across systems
- Standardized representation formats
  - 2010: Open Provenance Model (OPM)
  - 2012: W3C PROV (Draft status)
- The same set of people are involved

OPM: Agents, Processes and Artifacts

- OPM and W3C are both graph-based representations
- In the following: The Open Provenance Model (OPM) in brief

Open Provenance Model: base elements

- Agents
  - cdo, user
- Processes
  - calculate monthly means
- Artefacts / Entities
  - input and output data, log file
- These are modelled *in the past*.



Moreau et al. (2010b)

Motivations for W3C PROV

- W3C PROV continues the work of OPM
- Roughly: align it to RDF/OWL and other Semantic Web standards

Querying and viewing provenance

- Exploit encoded provenance information?
  - visualization
  - querying

Summary: what is provenance?

- To summarize this particular view:
  - Provenance is the result of a query over process assertions.
  - Such assertions can in their simplest form be represented through an (ever growing) DAG.
  - Including more details requires a process view that embraces a larger context.
  - Provenance information is subject to LTA

Bottom-up approach

- Suggestion: Start small and simple.
- Collect small pieces of information
  - automatically, infrastructure task, do not burden data producer
- Provide tools to gather intelligence from this heap of information
- DAG-view has obvious simple querying model (tree) and is easy to understand and explain
- Build the DAG as a base layer, then attach richer context to the nodes or edges

And then some...

- Construct a provenance graph using Persistent Identifiers?
  - PhD topic

- DKRZ-Seminar on Persistent Identifiers
  - Wednesday, 17 Oct
  - 14-16h
  - Same place (R34)

Handle System®

EUDAT

Further reading

- Luc Moreau: The Foundations for Provenance on the Web (2010)
  - main influence is Web science
  - summarizes the research field very well
  - includes an extensive bibliography

- OPM specification: http://www.openprovenance.org
- W3C PROV: http://www.w3.org/TR/prov-primer/

The End.

Thank you for your attention.

References

- Moreau (2010): The Foundations for Provenance on the Web, doi:10.1561/1800000010
  - pre-print: http://eprints.soton.ac.uk/268176/
- Moreau et al. (2010b): The Open Provenance Model Core Specification (v1.1), doi:10.1016/j.future.2010.07.005
- The Fourth Paradigm, 2009, Microsoft Research
- Buneman et al. (2001): Why and Where: A characterization of Data Provenance, doi:10.1007/3-540-44503-X_20