Data Provenance for Distributed Data Sets

Curt Tilmes
NASA Goddard Space Flight Center
Code 614.5, Greenbelt, MD 20771

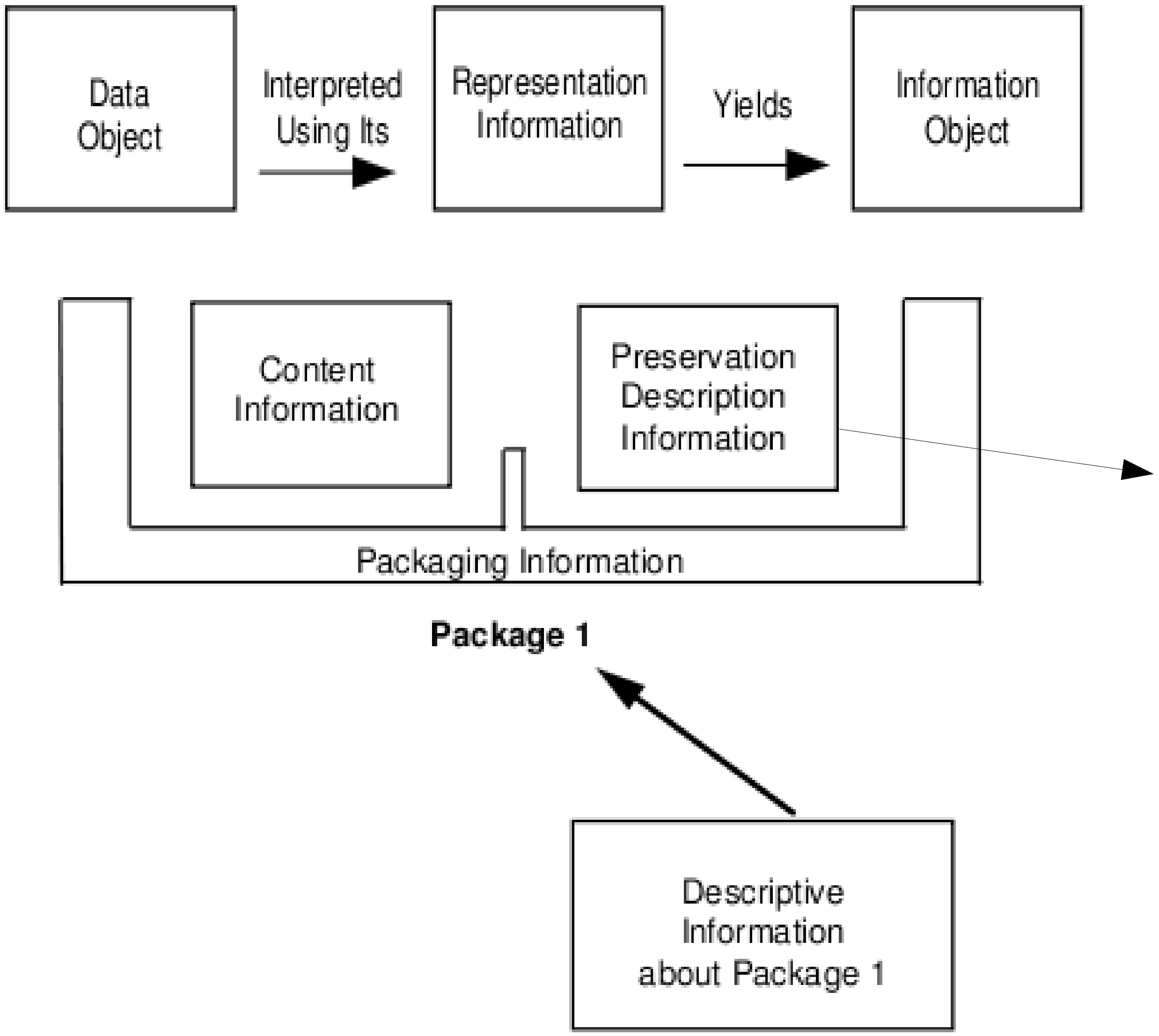
Curt.Tilmes@nasa.gov

NOAA EDMC 2011
2011-06-22

Vision

When scientific research is published, it should *reference* all data used in that research to a sufficient extent for *others* to *reproduce* that research and confirm the conclusions.

OAIS-RM



Provenance

Context

Reference

Fixity

OAIS-RM Preservation Information

❑ Provenance

- Source of all Content Information
- Custody since Origination
- Processing History

❑ Reference

- Identifiers to uniquely identify Content

❑ Context

- How the Content relates to other information

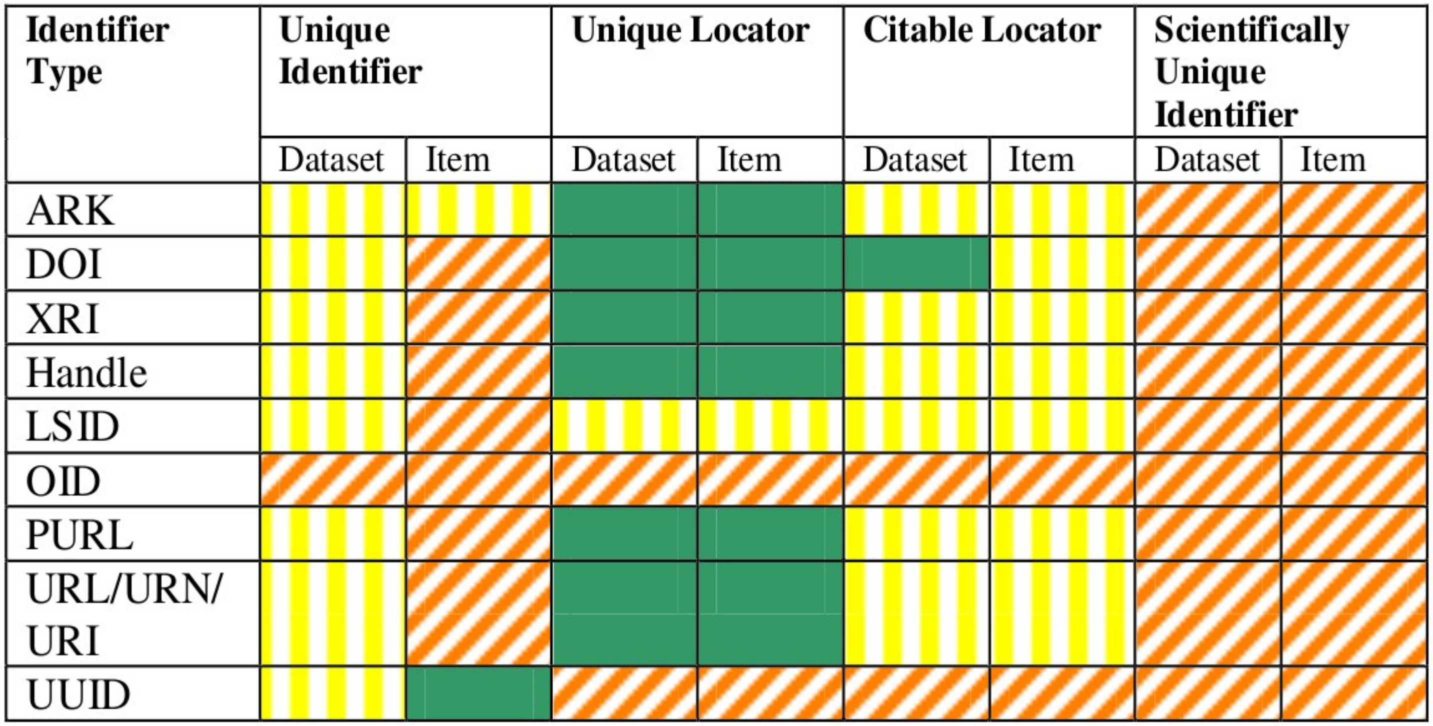
❑ Fixity

- Protection of Content from alteration
- Checksums or digital signatures

Reference – Identifiers

- “On the Utility of Identification Schemes for Digital Earth Science Data: An Assessment and Recommendations”

Table 2 Suitable Identifiers for Each Use Case where Solid Green Indicates High Suitability, Vertical Yellow Stripes Indicates Good to Fair Suitability; and Orange Diagonal Stripes Indicates Low Suitability.



Identifier Type	**Unique Identifier**		**Unique Locator**		**Citable Locator**		**Scientifically Unique Identifier**	
	Dataset	Item	Dataset	Item	Dataset	Item	Dataset	Item
ARK	▥	▥	🟩	🟩	▥	▥	▨	▨
DOI	▥	▨	🟩	🟩	🟩	▥	▨	▨
XRI	▥	▨	🟩	🟩	▥	▥	▨	▨
Handle	▥	▨	🟩	🟩	▥	▥	▨	▨
LSID	▥	▨	▥	▥	▥	▥	▨	▨
OID	▨	▨	▨	▨	▨	▨	▨	▨
PURL	▥	▨	🟩	🟩	▥	▥	▨	▨
URL/URN/URI	▥	▨	🟩	🟩	▥	▥	▨	▨
UUID	▥	🟩	▨	▨	▨	▨	▨	▨

Reference

- ❑ ESIP Fed
  - Identifiers Study and Testbed – *On the Utility of Identification Schemes for Digital Earth Science Data: An Assessment and Recommendations*
  - Developing Citation Standard – See draft
- ❑ Uniquely distinguish all data granules
  - Always assign a unique identifier, even when a granule is reproduced the 'same' way.
  - UUID gaining popularity
    - e.g. 7096bf5a-dec3-49fa-a54e-0404194b87d6
- ❑ Reference all Data Sets with identifiers suitable for citations
  - Actionable (Can someone do something with it?)
  - Persistent (What happens when you reprocess? What happens when a data set is transferred to a new organization?)
  - DOI is widely used for citations
    - doi:10.3334/ORNLDAAC/549
- ❑ Keep cited references valid, even if the data are obsolete/replaced
  - Provenance data can be used to reproduce the data, **if you keep it**!

Static vs. Dynamic Dataset Identifiers

- ❑ Earth science remote sensing missions often have very long lifespans.
- ❑ Move to measurement based datasets makes these even longer, spanning multiple missions.
- ❑ Static dataset – A bunch of data go into the dataset and stay there.
- ❑ Dynamic dataset – New granules are added to the 'end' of the dataset as time passes.
- ❑ For an operational mission, we also have operational issues that occasionally change older granules in the dataset.
- ❑ Identifiers for Static datasets are easy, Dynamic datasets are more difficult
  - Need to reference/cite a specific set of granules used
  - Date/timestamp is a start.

Provenance and Context

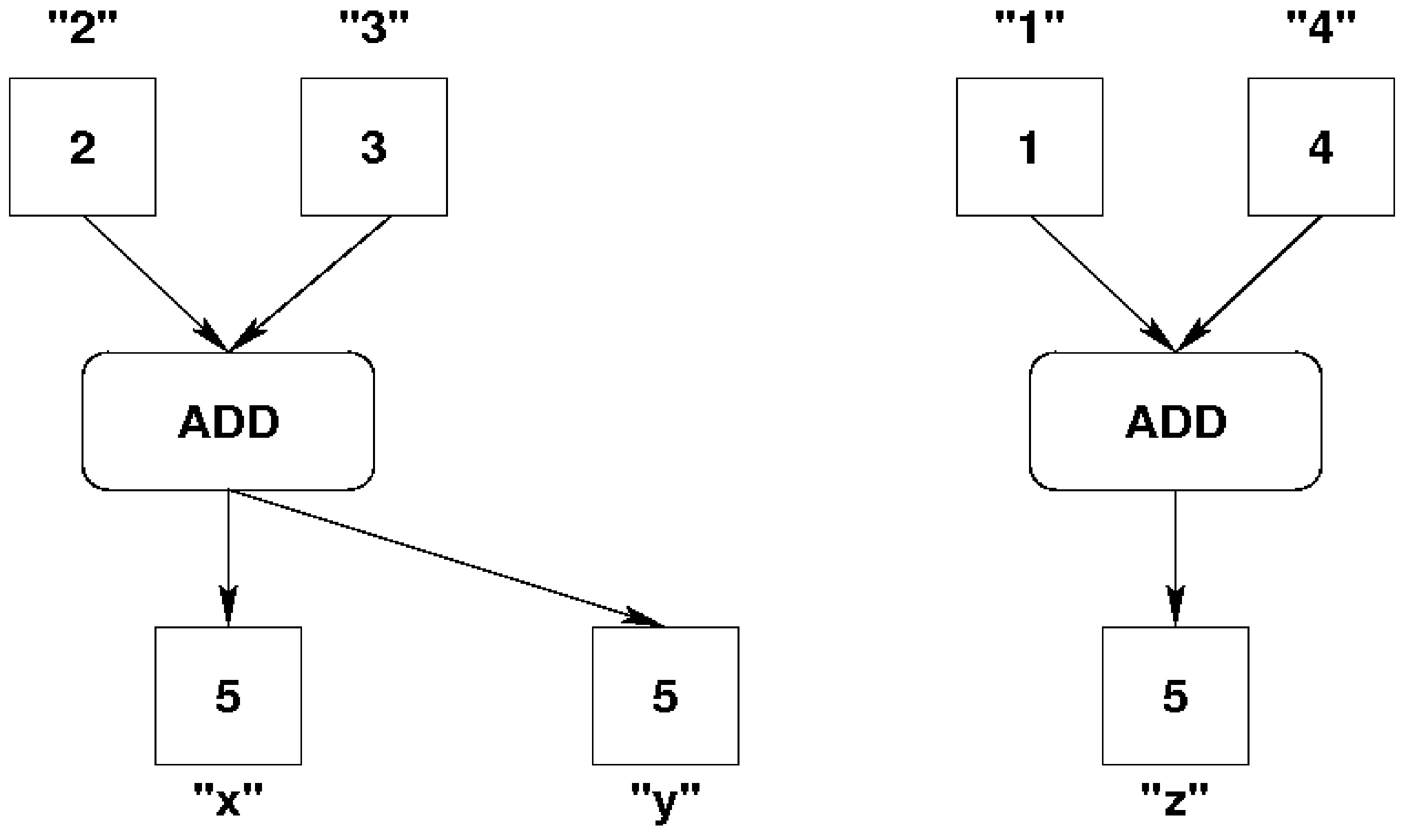
- ❑ ESIP Fed developing an “Earth Science Provenance and Context Content Standard” (PCCS)
  - The 1998 U.S. Global Change Research Program (USGCRP) workshop on *Global Change Science Requirements for Long-Term Archiving* Hunolt Report.
- ❑ Categories:
  - Preflight/Pre-Operations: Instrument/Sensor characteristics including pre-flight/pre-operations performance measurements; calibration method; radiometric and spectral response; noise characteristics; detector offsets
  - Products (Data): Raw instrument data, Level 0 through Level 4 data products and associated metadata
  - Product Documentation: Structure and format with definitions of all parameters and metadata fields; algorithm theoretical basis; processing history and product version history; quality assessment information
  - Mission Calibration: Instrument/sensor calibration method (in operation) and data; calibration software used to generate lookup tables; instrument and platform events and maneuvers
  - Product Software: Product generation software and software documentation
  - Algorithm Input: Any ancillary data or other data sets used in generation or calibration of the data or derived product; ancillary data description and documentation
  - Validation: Record and data sets
  - Software Tools: product access (reader) tools.

Provenance and Context

- ❑ “Earth Science Provenance and Context Content Standard” (PCCS)
- ❑ Fields:
  - Item Number
  - Category
  - Content Name
  - Definition/ Description
  - Rationale (Why content is needed)
  - Criteria (How good the content should be)
  - Priority
  - Source
  - Project Phase for Capture
  - User Community
  - Representation*
  - Distribution Restrictions
  - Source identifying item

Equivalence of Scientific Data

- ❑ *Two granules sharing identical provenance are identical.*
- ❑ *Two granules with different provenance are distinct.*



Identical

- ❑ For two granules of data to be *Perfectly Identical*, they must not only have identical contents, but also identical identifiers and identical creation provenance.
  - *This is only meaningful if you really are talking about the same granule, or two 'copies' of the same granule.*

- ❑ Two granules are *Scientifically Identical* if the data contents are the same, even if the identifiers of the granules, or the provenance of the granules are different. We also call this *Equal Content*. It doesn't matter how the content came to be – each such granule can be used in the same analysis and would result in the same results/conclusions.
  - *Digital signatures can show this.*

Equivalent

- ❑ Two granules have *Scientifically Equivalent Content* if the use of those granules in every possible scientific analysis will lead to the same results or conclusions.

- ❑ This definition allows 'slight' differences in the content – as long as they are close enough not to affect any analysis in a scientifically meaningful way.

- ❑ This is what we usually end up with for Process-On-Demand.

- ❑ Proving perfect Scientific Equivalence in the general case is very difficult (impossible?), or at the least, very manual.

Reproducibility

- ❑ *Scientifically Reproducible* refers to a process which is capable of reproducing granules that are *Scientifically Equivalent* to the original granules. *Scientific Reproducibility* is the extent to which a process is *Scientifically Reproducible*.

- ❑ Some processes are chaotic in that very slight differences in processing are compounded producing possibly drastically different results. We can apply sensitivity analyses to assess this characteristic and help determine if the process is suitably reproducible.

- ❑ If a process is unable to reliably reproduce data granules that are *scientifically equivalent*, we would claim that the process is not *reproducible*.

Scientific Equivalence

❑ There are two primary approaches for mechanically approximating this equivalence in a useful way:

- Content Equivalence – Can I show that the contents of two granules are sufficiently equivalent?

- Provenance Equivalence – Can I show that two granules were *created* in *essentially* the same way?

    - A *Provenance Equivalence Identifier* (PEI) can created with a digital signature from a canonical serialization of the *essential* provenance of the granule.
    - Each granule sharing a PEI is made in a sufficiently similar manner (they share all *essential provenance* elements) that they are *scientifically equivalent*.

ESIP Fed Data Preservation Activities

❑ The Federation of Earth Science Information Partners (ESIP) Preservation and Stewardship Cluster is working in a number of related areas:

- **Data Identifiers** – Publishing recommendations, Test Bed, working on longer term schemes
- **Data Citations** – Recommendations, best practices, standards
- **Provenance and Context Content Standard** – What artifacts should be preserved? Why? How should they be represented?
- **Earth Science Preservation Ontology** – An Earth Science domain profile built on the Open Provenance Model.

http://wiki.esipfed.org/index.php/Preservation_and_Stewardship

Thank You!