UMBC

AN HONORS UNIVERSITY IN MARYLAND

Enabling Reproducibility of Scientific Data Flows with Provenance Equivalence

Curt Tilmes

Curt.Tilmes@umbc.edu

Committee:

Prof. Yelena Yesha Advisor

Prof. Milton Halem Advisor

Prof. Tim Finin

Prof. Anupam Joshi

Dr. Jim Smith

UMBC
AN HONORS UNIVERSITY IN MARYLAND

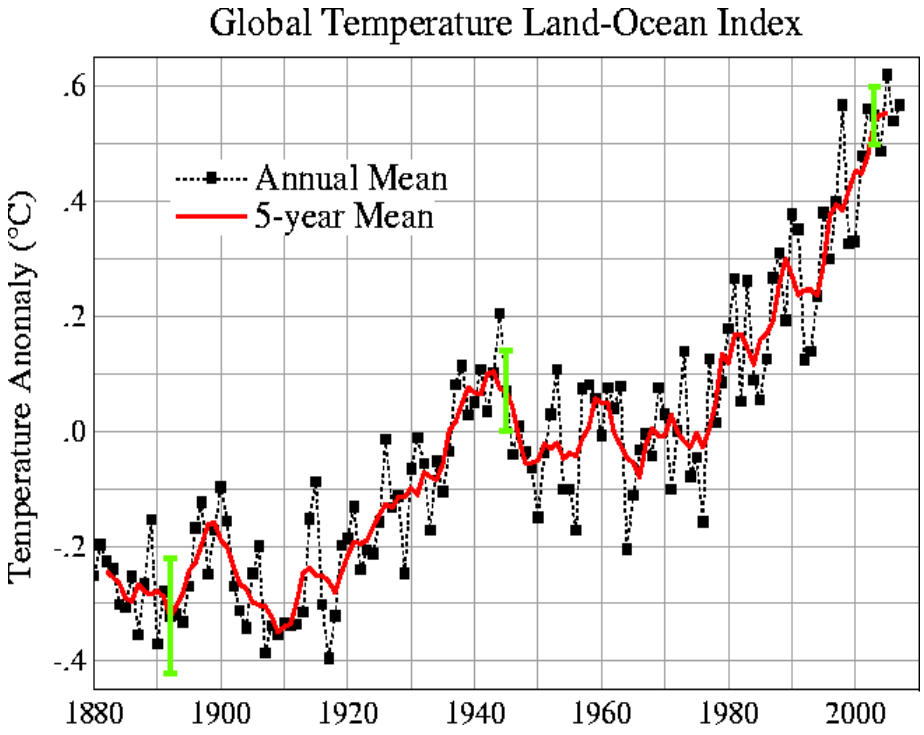
- ❑ Introduction
  - Background, Thesis Statement
- ❑ Related Work
- ❑ Contributions
  1. Data Model
  2. Equivalence and Reproducibility
  3. Dataset Instance Identification
  4. Provenance Equivalence Identification
  5. Dataset Provenance Equivalence Identification
- ❑ Evaluation
- ❑ Conclusions
- ❑ Future Work

Introduction

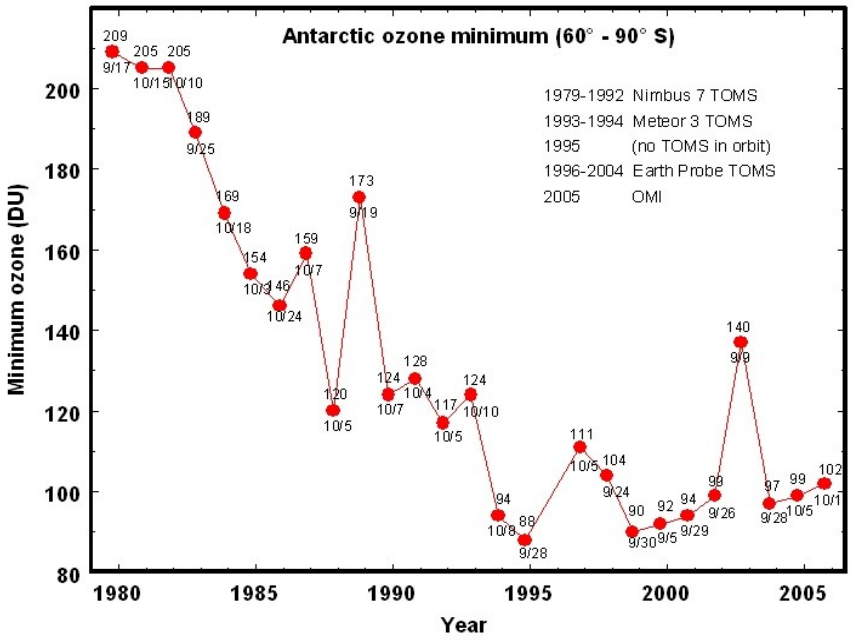
Science

- "An inherent principle of publication is that others should be able to *replicate* and build upon the authors' published claims. Therefore, a condition of publication in a Nature journal is that authors are required to make materials, data and associated protocols available in a publicly accessible database [...] or, where one does not exist, to readers promptly on request."
  - *(Guide to Publication Policies of the Nature Journals, 2007)*
- Science must be reproducible
  - *(or it isn't science...)*
- Traditionally, one could read a scientific paper, construct an identical experiment and confirm results
  - *(well, most of the time...)*
- *Reproducibility* yields *Credibility*

- Some modern scientific research is the result of lengthly computer analysis of a **very large** amount of data, building on the contributions of hundreds (thousands?) of individuals



http://data.giss.nasa.gov/gistemp/graphs/



http://macuv.gsfc.nasa.gov/ozone.md

Citations

- Current state of practice for citation of Earth Science Datasets is poor to non-existent
  - Some have acknowledgements
    - "Thanks to NASA/NOAA for data"
    - "Thanks to Fred who gave me some NASA data"
    - "Thanks to MODIS team for MODIS data"
  - Some reference specific data inline, with footnotes or in figure captions
    - * Used data from Terra MODIS instrument
    - * Used Collection 5 Land Surface Reflectance data from Terra MODIS
    - Used Collection 5 Land Surface Reflectance data from Terra MODIS downloaded on 2011-02-08
  - A few have started to actually include formal citations in references
    - Even those cite the dataset as a whole, not specific granules used in research

Citations

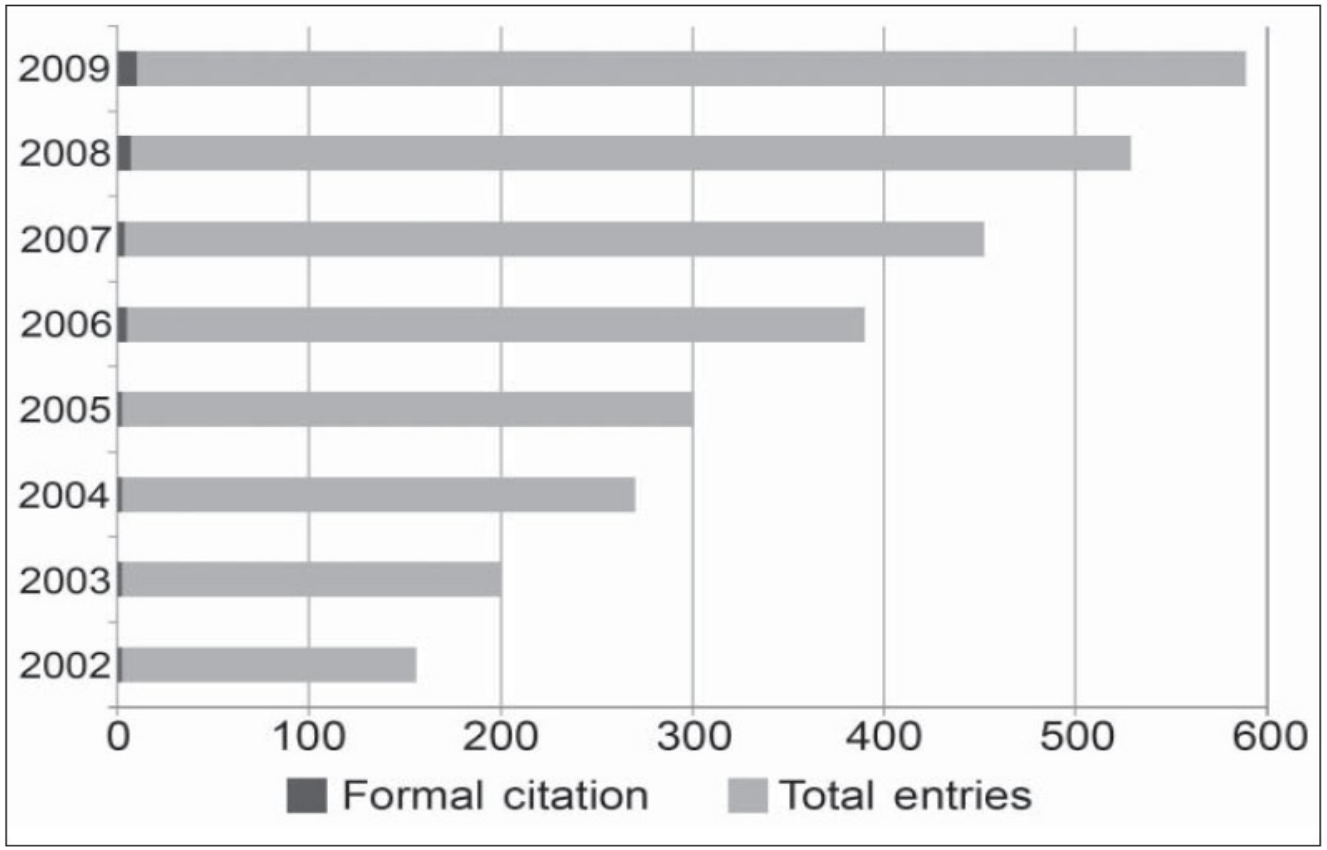


Fig 1. The National Snow and Ice Data Center distributes a variety of different snow cover products derived from the Moderate Resolution Imaging Spectroradiometer (MODIS). The results of a quick analysis of how many scientific papers mention use of "MODIS snow cover data" (according to Google Scholar™) and how often the data sets themselves are formally cited show a huge disparity, illustrating the infrequency of proper data citation in practice. Moreover, the lack of data citation standards introduces the possibility that informal references to data do not point to the data set actually used.

Parsons, et. al. "Data Citation and Peer Review", EOS, Transactions, AGU, 24 Aug. 2010.

"Data Citation in the Wild"

Valerie Enriquez, Sarah Walker Judson, Nicholas M. Weber, Suzie Allard, Robert B. Cook, Heather A. Piwowar, Robert J. Sandusky, Todd J. Vision, Bruce Wilson

"We found that few policies recommend robust data citation practices: in our preliminary evaluation, only one-third of repositories (n=26), 6% of journals (n=307), and 1 of 53 funders suggested a best practice for data citation. We manually reviewed 500 papers published between 2000 and 2010 across six journals; of the 198 papers that reused datasets, only 14% reported a unique dataset identifier in their dataset attribution, and a partially-overlapping 12% mentioned the author name and repository name. Few citations to datasets themselves were made in the article references section."

`http://openwetware.org/wiki/DataONE:Notebook/Summer_2010`

- "On the Utility of Identification Schemes for Digital Earth Science Data: An Assessment and Recommendations"
  - Addresses 4 use cases
    - Unique Identifier – UUID
    - Unique Locator – various schemes map to URL
    - Citable Locator – DOI
    - Scientifically Unique Identifier – No scheme is adequate

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								

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

Thesis Statement

- By developing a data model specifically for the kind of large datasets typical of the Earth and Space sciences, and creating identifier schemes for distinguishing data by its provenance equivalence,

  we enable precise references and citations of data used in scientific research.

- This foundational model and these provenance equivalence identifier schemes are key components that will ultimately enable the desired *reproducibility*.

Related Work

Related Work

- Replication and Reproducibility have been critical for science since it's beginning. Recently provenance as a research field has grown rapidly.
- Numerous models and representations of provenance have arisen:
  - Database – Buneman & Cheney, Tannen.
  - Workflow – Missier & Goble, Taverna; Sahoo & Sheth, Provenir; McGuiness et. al. PML; Zhao, Ouzo; Simmhan, Plale, Gannon, Karma2; PREMIS, SWAN, Changeset
  - Automated provenance collection - Frew & Bose UCSB ESSW, Seltzer & Braun Harvard PASS.
  - Open Provenance Model (OPM) – L. Moreau et. al.
  - W3C Provenance Incubator – Provenance Vocabulary Mappings
  - Ongoing activity of the ESIP Federation

❑ Geoscience Processing Data Models

- The CCSDS Reference Model for Open Archival Information Systems (OAIS) provides a very high level architecture, tying together data content with Provenance, Context, Reference and Fixity.
- NASA's Earth Observing System (EOS) Core System (ECS) Data Model provides some high level concepts.
- C. Tilmes and A. Fleig. “Provenance Tracking in an Earth Science Data Processing System.” In Provenance and Annotation of Data and Processes, volume 5272 of Lecture Notes in Computer Science, pp. 221-228. Springer Berlin / Heidelberg, 2008.
- B. Barkstrom. “A mathematical framework for earth science data provenance tracing.” Earth Science Informatics, 2010.
- C. Tilmes, Ye. Yesha and M. Halem, “Tracking Provenance of Earth Science Data.” Earth Science Informatics, 2010.

Related Work

- Provenance and Context Content
  - The 1998 U.S. Global Change Research Program (USGCRP) workshop on “Global Change Science Requirements for Long-Term Archiving” Hunolt Report.
  - R. Duerr et. al. “Challenges in Long Term Data Stewardship”, 2004.
  - C. Tilmes, Ye. Yesha and M. Halem. “Provenance Artifact Identification in the Atmospheric Composition Processing System (ACPS)”, Proceedings of the 2nd Workshop on the Theory and Practice of Provenance, 2010.
  - Ongoing activity in the ESIP Federation

❑ Identifiers and Locators

- R. Duerr, R. Downs, C. Tilmes, B. Barkstrom, W. C. Lenhardt, J. Glassy, L. Bermudez, P. Slaughter. “On the Utility of Identification Schemes for Digital Earth Science Data: An Assessment and Recommendations” 2011 (to be published).
- C. Tilmes, Ye. Yesha and M. Halem. “Provenance Artifact Identification in the Atmospheric Composition Processing System (ACPS)”, Proceedings of the 2nd Workshop on the Theory and Practice of Provenance, 2010.
- Ongoing activity in the ESIP Federation and NASA ESDSWG.

❑ Citations

- M. Parsons, R. Duerr and J. Minster. “Data Citation and Peer Review” EOS Trans. AGU. 2010.
- V. Enriquez, S. Judson, N. Weber, S. Allard, R. Cook, H. Piwowar, R. Sandusky, T. Vision, B. Wilson. “Data Citation in the Wild”, 2010.
- M. Parsons, R. Duerr, C. Tilmes, B. Barkstrom. “Data Citation”. GeoData 2011, March, 2011.
- Ongoing activity in the ESIP Federation.

Related Work

- Content Equivalence
  - R. Cavanaugh and G. Graham. “Apples and apple-shaped oranges: Equivalence of data returned on subsequent queries with provenance information.” Data Provenance/Derivation Workshop, 2002.
  - A. Chapman and H.V. Jagadish, “Provenance and the Price of Identity”. Berlin, Heidelberg: Springer-Verlag. 2008.
  - Universal Numeric Fingerprint. Altman & King, 2007.
  - Barkstrom working in this area – upcoming paper to be published.
- Provenance Equivalence
  - C. Tilmes, Ye. Yesha and M. Halem “Distinguishing Provenance Equivalence of Earth Science Data”, Proceedings of The Intl Conf on Comp. Science, 2011. (*to be presented this summer*)

Contributions

Objectives

- We propose several specific identifier schemes that make identification, citation and comparison of provenance equivalence of reproduced data easier to accomplish.
  1. A general model of earth science processing, including some basic terminology and an organization of data for *large*, *dynamic* datasets.
  2. A discussion of data *scientific equivalence* and *reproducibility* and their relationship to one another. A taxonomy of equivalence concepts and terms. A notion of *essential provenance* as a way to distinguish the provenance needed for reproducibility.
  3. *Dataset Instance Identifiers* for referring to specific data granule membership in dynamic datasets, and an algorithm for calculating and maintaining them during changes to the dataset.
  4. *Provenance Equivalence Identifiers* as a proxy for a potentially large graph of a workflow leading to the creation of a data granule.
  5. A combination of DII and PEI concepts for identifying entire dynamic datasets not simply by their granule membership, but also by their provenance equivalence so that large dynamic datasets can be referred to and cited more precisely.

1. Data Model

Data Model

- Data are organized into discrete granules, described by their **Granularity**
- Each data granule is assigned a specific **DataType** that relates directly to: **Granularity**, algorithm, data format, **Granule Key**.
- Granules are assigned a **Version** or Collection
- A **Dataset** is comprised of Granules with same **DataType**,**Version (Dataset** = { **DataType**, **Version** })
- An **APP** uses production rules to determine **Runtime Parameters** and **Input Granules** within a **Dataset** by their **Granule Key**s
- A granule can be identified/distinguished by

  { **DataType**, **Version**, **Granule Key**, **Timestamp** }

Reproducibility

- In normal processing and reprocessing, we execute production rules to find the *best* input files
- For *reproducibility*, we need to find the *same* input files
  - Find the exact same granule that was previously used as an input, or if it is missing –
  - Find an *equivalent* granule that can be used in place of it.

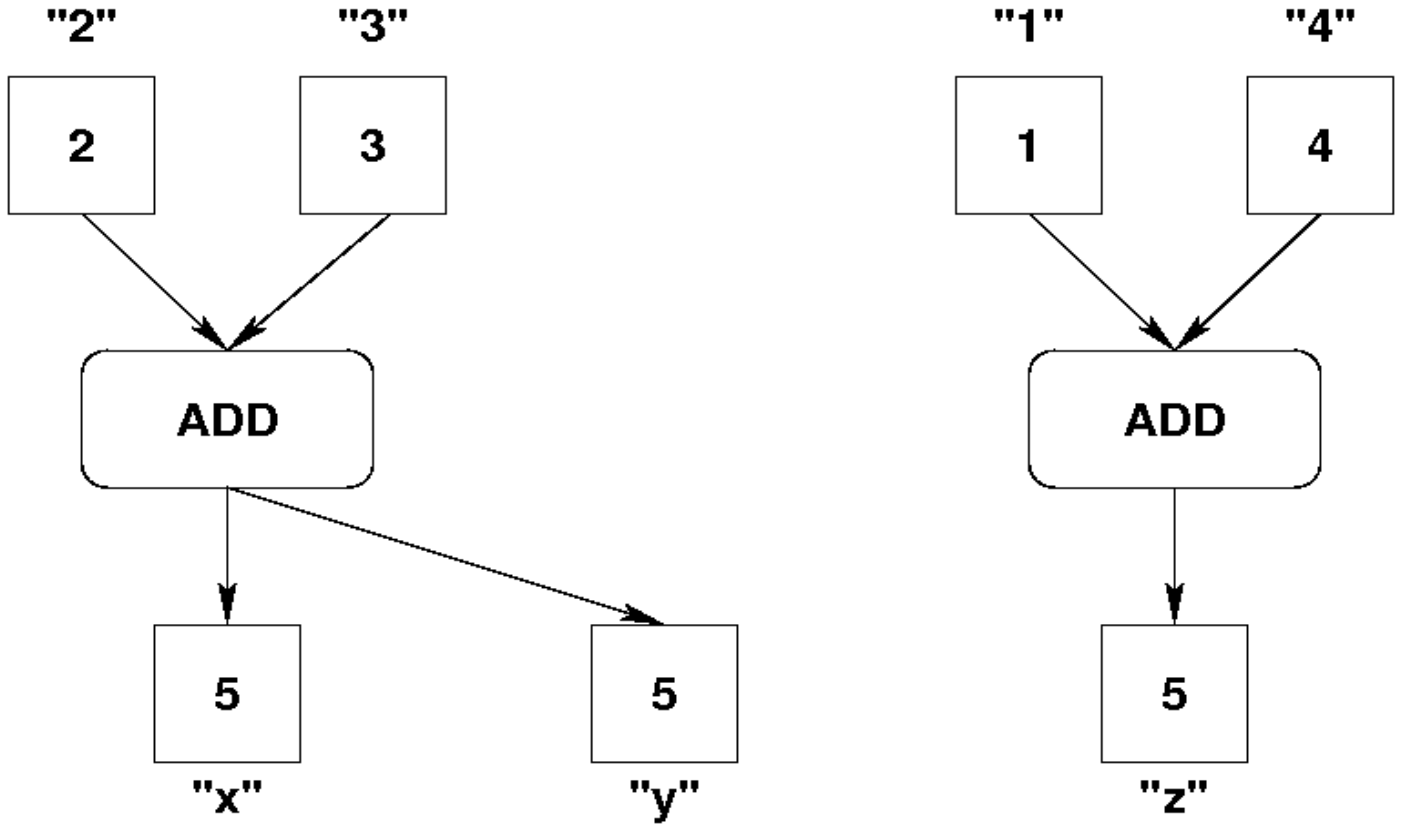
2. Equivalence And Reproducibility

Scientific Equivalence

- Proving perfect Scientific Equivalence in the general case is very difficult (impossible?), or at the least, very manual.
- There are two approaches for mechanically approximating this equivalence in a useful way:
  - Content Equivalence – Can I show that the content of two granules are sufficiently equivalent?
    - Others (and I) are working on this. Bruce B. claims it is impossible in the general case.
  - Provenance Equivalence – Can I show that two granules were made in *essentially* the same way?
    - This is the approach presented here.
    - Two ways to show this:
      - Explore the complete provenance graph that shows how each were produced, or
      - Create a Provenance Equivalence Identifier as a proxy of that graph, and just compare them directly

Equivalence of Scientific Data

- *Two granules sharing identical provenance are identical.*
- *Two granules with any aspect of provenance differing are distinct.*



Identical granules

- 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.

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.

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 possible 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*.

Reproducibility and Essential Provenance

- Consider all the elements of provenance for a process, 0..n, order them by the extent to which they contribute to the content:
  - $p_0, p_1, \ldots, p_j, p_{j+1}, \ldots, p_k, p_{k+1}, p_n$
- If the process is reproducible, there exists a point j where elements to the left (0..j) are essential for reproducibility. If the process is repeated with those elements the same, the resulting data granules will be *scientifically equivalent* to the original. If we can determine such a point (i.e. we can determine which provenance elements are essential – required for reproducibility) then the process is reproducible. If we can't determine the point (i.e. we don't know what information someone else must match), then the process is not reproducible.

3. Dataset Instance Identification

Static vs. Dynamic Datasets

- 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.
- (We've also called these “Open” vs. “Closed” datasets.)
- Identifiers for Static datasets are easy, we need a good identifier scheme for Dynamic datasets too..

Citations

- Based on our Data Model, there are two required fields for good citations:
  - Data Type or ESDT
  - Data Version or Collection
- For a static dataset, that is very useful, and today is a common way to identify a dataset
- For a *dynamic* dataset, we want to identify the *specific* granules that were part of that dataset at the time we accessed/downloaded the dataset. We had previously proposed using a timestamp to determine that granule list.
  - A timestamp can map to a set of granules, but many timestamps map to each set of granules. Can't compare citations with just the timestamp.
  - When considering two archive mirrors of the same dataset, the specific insert or removal timestamps for each granule are often different.

Dataset Instance Identifiers

1. The Dataset Identifier + Dataset Instance Identifier can be resolved by a dataset curator into a specific set of granule identifiers.
2. Two data citations will have the same DI + DII if and only if they are referring to exactly the same set of granules.
3. Two dataset mirrors will produce the same DII for the same specific sets of granules, regardless of the order of granule addition or removal.

DII Operations

1. Append additional granules to the dataset.

   Add granules with identifiers that sort higher than any other granule identifier in the dataset. This is by far the most common operation.

2. Remove existing granules.

   This could be any granule in the dataset, but practically most often occurs for recent granules.

3. Add older granules.

   Again, these could be at any granule position, but typically occur for recent granules.

DII Algorithm

- Calculate a Digital Signature / Hash of the first Granule Identifier
- When appending granules, take the previous hash, concatenate the next Granule Identifier and take the hash of the result.
- When changing older granules, roll back to the position in the granule identifier list and recalculate the list forward.
- Maintain a map of the DII to the time, and use that time to query the dataset map to determine the precise granule membership of the dataset at that time.

4. Provenance Equivalence Identification

Provenance Equivalence

- When comparing datasets, we are concerned with precise granule identification – whether or not a particular granule is the same granule (*Perfectly Identical*).
- *Provenance Equivalence* relaxes that to determine if two granules are *Scientifically Equivalent*.
- This is distinct from efforts that concentrate strictly on *Content Equivalence* and disregard provenance.
- We propose a *Provenance Equivalence Identifier* (PEI), 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*.

Provenance Equivalence Identifiers

- IF a process is reproducible, we can determine the essential provenance for the process.
- IF we repeat a reproducible process with identical essential provenance, we will get a scientifically equivalent granule.
- The PEI can be used as a proxy for the essential provenance graph that led to the creation of that data granule.
- Two granules with the same PEI will be scientifically equivalent to one another, even if their content varies slightly.

5. Dataset Provenance Equivalence Identification

Dataset Provenance Equivalence Identifier

- Consider the DII as one axis of the dataset, summarizing a dataset by a list of the granule membership of that dataset and the PEI as another axis, summarizing the provenance graph of each granule.
- Combine the two approaches to create a Dataset Provenance Equivalence Identifier (DPEI).
- Order PEIs by the Granule Key to maintain fast append operation.
- Citations can include DPEI to determine precise granule membership, but allow reproducibility to create identical identifiers for scientifically equivalent datasets.
- DPEI + PEI DAG gives two paths of exploration to discover differences between datasets. Compare sets of granules by their keys, to determine granule membership, then walk the tree of PEI to discover provenance differences.

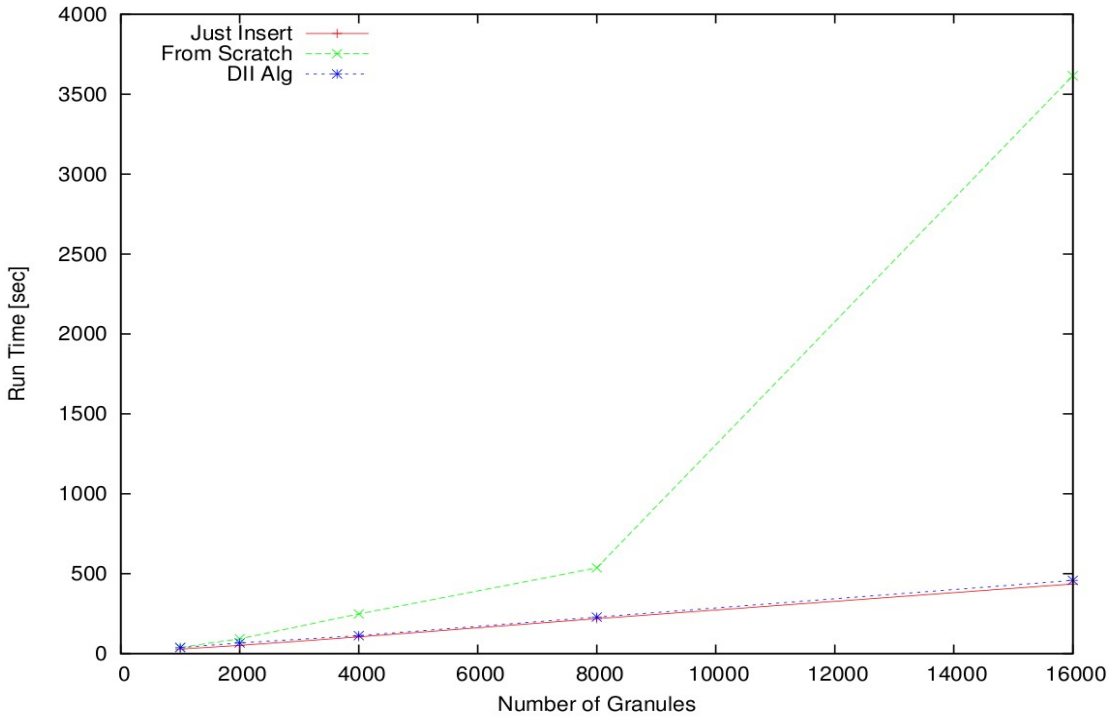
Evaluation

Dataset Instance Identifiers Algorithm Testing

- The DII algorithm was implemented in a dual core Intel Xeon 2.4GHz computer with 6GB of memory, running CentOS Linux 5.5 and PostgreSQL v. 8.4.1.
- We compared the "running total" DII scheme, optimized for our dynamic datasets against a simple alternative calculating a comparable DII from the hash of all the contents for each new dataset instance.
- The test was repeated for dataset sizes: 1000, 2000, 4000, 8000, 16000
- For each test size, three tests were performed:
  - A baseline, inserting the granule into the dataset, but not calculating any identifier
  - Calculating a DII from the list of identifiers
  - Our "optimized" DII "running total" approach

DII Algorithm Testing

Granules	Just Insert	From Scratch	DII Alg
1000	28.677	38.608	38.098
2000	51.924	93.408	67.105
4000	104.482	248.421	112.705
8000	218.844	536.626	228.591
16000	435.886	3615.314	458.022



DII Algorithm Testing evaluation

- The “running total” algorithm performs very well, imposing very little performance overhead on the database to maintain the identifiers.
- This test only proves the “best case” for the algorithm where granules are appended to the 'end' (the identifier can be sorted temporally higher than all other identifiers in the dataset).
- In the “worst case”, the algorithm would degenerate to the other line, (but that should be a rare occurrence)
- The assigned DIIs can be used to determine the precise granule membership of the dataset from the reference or citation.

PEI Experiments

- We have applied the PEI algorithm to assign identifiers to OMI data granules based on provenance information dumped from the production OMIDAPS database.
- 18,820,634 total files
- 2,788,858 total APP runs
- Assigned 4,435,351 leaf PEIs (not associated with an APP run)
- Produced 14,385,283 PEFs
- Found 4,865 “duplicate” PEIs in the operational database

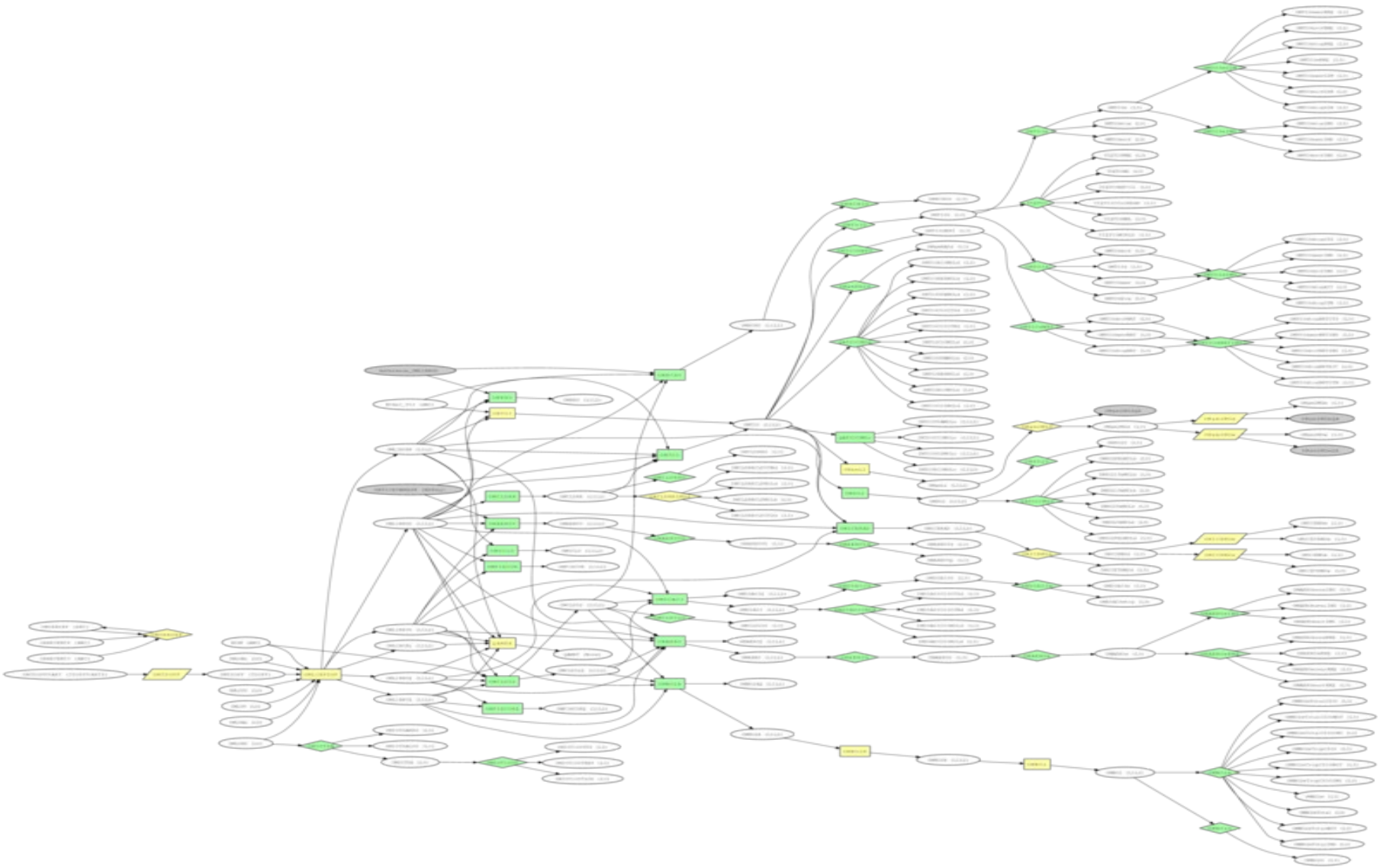
Duplicate PEI example

- Example: PEI `03e6b542baf71ae84fb3627f5d5ccb73` matches
  - `OMI-Aura_L2-OMCLDO2_2010m0929t1555-o33019_v003-2010m0929t215650.he5`
    - `MD5: 252cef7cf2401b8866c0e7bb106a555b`
    - `Produced: 2010-09-29 17:56:55.064829`
  - `OMI-Aura_L2-OMCLDO2_2010m0929t1555-o33019_v003-2010m0930t152929.he5`
    - `MD5: d0029a203e164287ac7634db8a15dd15`
    - `Produced: 2010-09-30 11:29:34.453187`
- `PEF:`
  -

```
EndTime: 2010-09-29T17:34:30.000000Z
OrbitNumber: 33019
PGE: OMCLDO2
PGEVersion: 1.1.1.3
Source: OMI
StartTime: 2010-09-29T15:55:37.000000Z
Inputs:
  - dada3d5d6c52e30863538416b2d2239c
  - 94616a80d9955ce1ee7d5f635c14fce5
  - b7a4ae7d11ac11d02ca9031a3c6a9df6
  - 3652235c40d929d303cae3dbfeac35b9
  - c57647f7fbb7e9b817cbee3566adc75f
Output: 2
```

UMBC
AN HONORS UNIVERSITY IN MARYLAND

OMI Data Flow



UMBC
AN HONORS UNIVERSITY IN MARYLAND

PEI Example

63505987a23317912a95b7a070808850

```
Date: 2010-02-23
Day: 054
DayOfYear: 054
EndTime: 2010-02-24T00:00:00.000000Z
PGE: OMTO3e
PGEVersion: 1.0.5.1
Source: OMI
StartTime: 2010-02-23T00:00:00.000000Z
Year: 2010
Inputs:
  - 642fefb516625dbce25658cbd091caef
  - e3db3533b8c384c813ffd1f4d137517b
  - 7a8e09e4de662f051083677f60bd8a58
  - 7ad12ff09d9fa3dd2cd50b1c02638e98
Output: 2
```

```
Date: 2010-02-22
Day: 053
DayOfYear: 053
EndOrbit: 29834
EndTime: 2010-02-23T00:00:00.000000Z
OrbitCount: 15
OrbitsProcessed: 15
PertinentOrbitCount: 15
PGE: OMTO3G
PGEVersion: 1.0.3.1
Source: OMI
StartOrbit: 29820
StartTime: 2010-02-22T00:00:00.000000Z
TotalOrbits: 15
Year: 2010
Inputs:
  - 642fefb516625dbce25658cbd091caef
  - 47eefbd4c6b09ac9bfeef6bc4a7af828
  - bdd16e7a62dfd4cd737ad59bed3b4c4c
  - ce4a5d94fc42ef9848694e4f2f2a7465
  - 9afd3ad9683d2f5e7e0ac1495d6ac8ef
  - 58314b59d8e63ac37b1ab68a9a1a12ae
  - 230b41ec843653b35301d0e036a096d8
  - 3c28b3aad8ebb338d5e83a4e1df8c9ec
  - 63f1bab8a30ed9bbaa8d10f92ed98c3c
  - 605f145e95ef00cb52baca12a6f9b3d8
  - bd72de93852dcc1d15378864fd40a191
  - dc16a1ddc6412aabb281a6b8e673fea5
  - a18f38557080f5accac17e098b13070b
  - 65955a5c3da9e333224974cbbc782984
  - f4159b4732ac67a2ea884a5730846f29
  - 685843445166ba47f425a0fb588f71fb
Output: 3

EndTime: 2010-02-22T01:23:33.000000Z
OrbitNumber: 29820
PGE: OMTO3
PGEVersion: 1.1.2.3
Source: OMI
StartTime: 2010-02-21T23:44:39.000000Z
Inputs:
  - 642fefb516625dbce25658cbd091caef
  - cd591c38637cb5a04e10148814d95006
  - e584200cebcc73bf7aa8609a3e3bf253
  - 61ed21683c1ab4bc6a4562b3c51e9e38
  - 960ddb17bb147d795e99de4621137746
Output: 3
```

```
EndTime: 2010-02-22T01:23:33.000000Z
OrbitNumber: 29820
PGE: OMCLDRR
PGEVersion: 1.6.0
Source: OMI
StartTime: 2010-02-21T23:44:39.000000Z
Inputs:
  - 642fefb516625dbce25658cbd091caef
  - 960ddb17bb147d795e99de4621137746
Output: 1
```

```
AscendingEquatorXingLongitude: -162.58
AscendingEquatorXingTime: 2010-02-22T00:36:17.000000Z
DescendingEquatorXingLongitude: 29.8
DescendingEquatorXingTime: 2010-02-21T23:46:46.000000Z
EndTime: 2010-02-22T01:23:33.000000Z
OrbitNumber: 29820
PGE: OML1BPDSP
PGEVersion: 1.1.3
Source: OMI
StartTime: 2010-02-21T23:44:39.000000Z
Inputs:
  - 642fefb516625dbce25658cbd091caef
  - d6fde623bda2468fb7b34f5b4ac44574
  - ae9903b82a80b3758100c7d70d983625
  - 6c7131e2e87ea88046cf95c267c282de
  - 57d06ec1d1f7e06c52d19ee8993b6e12
  - e405f7f6cbec7db9fba60ad29eb40523
  - cc8af9c6e671b1e8480b45cc3982f048
  - bd52df56d3385b4114b08873c424cf28
  - 97861acdd6dcfc630a3077c0ef6ff46c
  - de473731a55e49699170773b82b3e34c
  - a284535b04cd62eb7a3884170470de14
  - a07b5c59d8bbc4cd9a32f30d1789d441
  - 60df66b7faecb8e0018a59d3bfcdcce9
  - ea31d6a9f564ee978909f2a370212652
  - ee5a9baeb2d21d00edc49e310e0a8c6c
  - 4267464b3db8cf8b32a1148926297cf1
  - 0ede9c316e2dd2c0784d4f644aff7370
  - 89d2f8ae8b1c063078de02a3d99f00ab
  - d617d26113fe730c9cff3b1c2924c3d6
  - d0541165a0d8153dcfd0c9cefdade0c2
Output: 2
```

PEI Experiments

- We can follow the provenance equivalence through multiple layers of production.
- Indexing the database on the PEI allows the system to locate equivalent granules.
- When portions of the data are removed, we can determine use the metadata and provenance database to determine the “essential provenance” using equivalence of predecessor files rather than requiring the exact files (like other provenance models).

Scientific Community Evaluation

- Our data model and equivalence and reproducibility concepts have been presented in multiple forums and published in peer-reviewed scientific literature.
- These ideas are being honed and revised based on community feedback.
- These papers are gaining acceptance and being cited and used as a basis for other research.

Conclusions

- We proposed several specific identifier schemes that make identification, citation and comparison of provenance equivalence of reproduced data easier to accomplish.
  1. A general model of earth science processing, including some basic terminology and an organization of data for *large*, *dynamic* datasets.
  - A discussion of data *scientific equivalence* and *reproducibility* and their relationship to one another. A taxonomy of equivalence concepts and terms. A notion of *essential provenance* as a way to distinguish the provenance needed for reproducibility.
  1. *Dataset Instance Identifiers* for referring to specific data granule membership in dynamic datasets, and an algorithm for calculating and maintaining them during changes to the dataset.
  2. *Provenance Equivalence Identifiers* as a proxy for a potentially large graph of a workflow leading to the creation of a data granule.
  3. A combination of DII and PEI concepts for identifying entire dynamic datasets not simply by their granule membership, but also by their provenance equivalence to that large dynamic datasets can be referred to and cited more precisely.

Future Work

- Federation of Earth Science Information Partners (ESIP) Preservation and Stewardship Cluster is working on a number of related areas:
  - Data Citations – Short term recommendations, working on longer term identifier schemes such as those proposed here.
  - Provenance and Context Content Standard – What artifacts should be preserved, why? How should they be represented?
  - Earth Science Data Processing Ontology – An earth science domain profile built on the Open Provenance Model.
- Extending 'industrial model' processing to more community accessible e-Science model. Use identifier schemes such as those proposed here to convey datasets between systems and trace provenance graphs across organizations and system.
- Provide Cloud Computing / Social e-Science system for mechanical reproduction of properly referenced datasets.

Thank You!

Backup

Granularity

- Dealing with data at the extremes of granularity is awkward:
  - All data from all places for all times
  - A single measurement of some property for a single place at a single instant in time.
- Convention breaks down data into “granules” where neither the size of a single granule nor the total number of granules in a dataset are overwhelming.
- Sometimes this is called an “archival unit” or the smallest individual unit of data to be archived.
- Granules are related to Files, but different. You can have multiple files that are part of a single granule.
- There are also ways to pull even smaller bits of data out of a granule.

Data Type

- We need a controlled vocabulary for distinguishing different types of data.
- Consider an example:
  - One of the MODIS products is “Surface Reflectance”
  - We define a more precise identifier for the type of that product with the identifier **MOD09A1**.
- EOS uses the term “Earth Science Data Type” (**ESDT**) for this more precise data type identifier.
- It identifies more than the broad type of data in the dataset:
  - A specific algorithm (with published Algorithm Theoretical Basis Document 'ATBD')
  - A specific data format
  - A specific data **Granularity** which includes:
    - A consistent granule definition (spatial/temporal/other)
    - A **Granule Key** that can uniquely identify a granule in a dataset.

Versions

- Basic configuration management works well for software.
- Anytime the software is changed, we tag a snapshot with a revision number (v. 1.2.3) through our CM tools.
- We can go back and check out that version of the software, compare versions, etc.
- Data versioning is more complicated. The direct predecessors and the software that produced a given granule could have the same version, but due to changes 'up-stream' in the workflow, the data are different.

EOS Collection Model

- Scientists don't like things that change too frequently.
- We do “major” reprocessing in collections, batching up a bunch of changes at once.
- Could involve new calibration, new formats (hopefully minor changes..), new software versions throughout the chain.
- “MOD09A1.004” and “MOD09A1.005” are two different collections from MOD09A1.

Provenance and Context Artifacts

- All of the "artifacts" involved or related to the scientific result:
  - Data
  - Algorithms, Processes, Configuration Tables, Runtime Parameters ("Workflow Provenance")
  - Documentation (ATBDs, Design Docs, Commented Source)
  - Sensors/Instruments/Instrument platforms
  - People/Organizations (reputation)
  - Published scientific papers (add to credibility and understanding)
  - Computer systems, Hardware, OS, Libraries, Software
  - Abstract things like "a data transformation event," "Software Build Event" or "a validation experiment"
  - An ephemeral execution of a web service
  - Versions from all of the above: Rigorous Configuration Management.
  - Specific relationships between all the artifacts.
- Things that increase *understanding* and enable *reproducibility*.
- ESIP Federation developing a "Provenance and Context Content Standard"

Essential Provenance

- What aspects of the provenance are “essential” for reproducibility?
- Can't record “Big Bang” provenance
  - the “butterfly effect”
- Some things are definitely “essential”
  - Workflow artifacts – inputs, runtime parameters
- Some things are definitely “non-essential”
  - Name of processing host
  - These are useful for auditing and increase credibility of provenance.
- Some things aren't so clear
  - Heinrich Hertz testing Maxwell's Equations – didn't report the size of the room he worked in – turned out to be “essential”
  - Compiler Flags? Library Versions? OS architecture?

DII Algorithm

CALCULATENEWDATASETINSTANCEIDENTIFIER(D, I, i, t)

```
1  i ← i − 1
2  while i ≥ 0 and D[i].DELETED ≠ NULL
3        do i ← i − 1
4  if p < 0
5     then h ← ''
6     else  h ← GETDATASETINSTANCEIDENTIFIER(I, D[p].INSERTEDTIME)
7  for i ← p to D.MAX
8        do if D[i].DELETEDTIME = NULL
9              then h ← HASH(CONCATENATE(h, D[i].ID))
10 APPEND(I, {t, h})
```

ADDGRANULE(D, I, a, t)

```
1  i ← SEARCH(D, a)
2  INSERT(D, i, {a, t, NULL})
3  CALCULATENEWDATASETINSTANCEIDENTIFIER(D, I, i, t)
```

DII Example

- Add some granules to a dataset:

Granule ID	Time
$x1$	1
$x2$	2
$x3$	3

(a) Dataset Table

ID	insertedtime	deletedtime
x1	1	NULL
x2	2	NULL
x3	3	NULL

(b) DII Table

Time	DII
1	$\textsc{Hash}(x1)$
2	$\textsc{Hash}(\textsc{Hash}(x1) + x2)$
3	$\textsc{Hash}(\textsc{Hash}(\textsc{Hash}(x1) + x2) + x3)$

DII Example (2)

- At time 4, remove granule x2

(a) Dataset Table

ID	insertedtime	deletedtime
x1	1	NULL
x2	2	4
x3	3	NULL

(b) DII Table

Time	DII
1	$\textsc{Hash}(x1)$
2	$\textsc{Hash}(\textsc{Hash}(x1) + x2)$
3	$\textsc{Hash}(\textsc{Hash}(\textsc{Hash}(x1) + x2) + x3)$
4	$\textsc{Hash}(\textsc{Hash}(x1) + x3)$

DII Example (3)

- Consider a mirror of this dataset that adds the granules out of order. At time 1, granules 1 and 3 are added, at time 2 granule 2 is added, and at time 3 granule 2 is removed.

(a) Dataset Table

ID	insertedtime	deletedtime
x1	1	NULL
x2	2	3
x3	1	NULL

(b) DII Table

Time	DII
1	$\textsc{Hash}(\textsc{Hash}(x1) + x3)$
2	$\textsc{Hash}(\textsc{Hash}(\textsc{Hash}(x1) + x2) + x3)$
3	$\textsc{Hash}(\textsc{Hash}(x1) + x3)$

Leaf Granules

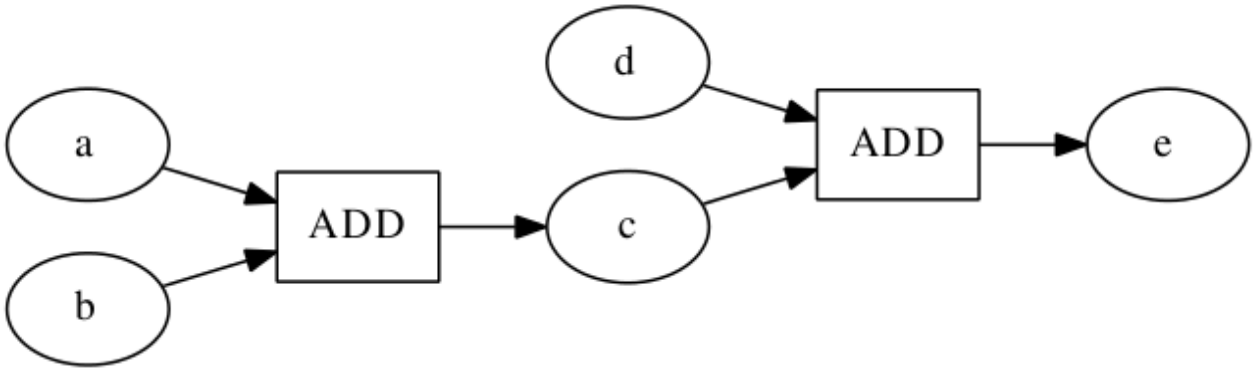
- Some granules come from 'outside' our processing system's scope. If they already have a PEI assigned to them, great, if not, we need to 'prime the pump'.
- Calculate a digital signature / hash of the content of the granule, and use that as the PEI for that granule.
- Independent systems that get the same granule will produce the same PEI for that granule.

PEI Calculation

- The PEI for each subsequent data granule is a hash of a canonical serialization of the essential provenance for that granule.
- For our demonstration implementation, and the examples here, we simplify to three things:
  - Runtime Parameters – these can change the manner of execution of the APP, environment variables, command line arguments, APP identifier, APP version
  - Input Granules – the PEIs of all other input files to the process. The order must be the same.
  - Output Granule Distinguisher – If there are more than one output file, we use a serial number to guarantee a distinct PEI.

PEI Example

- Simple workflow adding some numbers.



- a,b,d are leaf granules:

$$\mathbf{PEI}(a) = \texttt{401b30e3b8b5d629635a5c613cdb7919}$$

$$\mathbf{PEI}(b) = \texttt{009520053b00386d1173f3988c55d192}$$

$$\mathbf{PEI}(d) = \texttt{e29311f6f1bf1af907f9ef9f44b8328b}$$

PEI Example (2)

- Construct a Provenance Equivalence File (PEF) to calculate the PEI of c:

```
APP: ADD
APPVersion: 1.0
Inputs:
  - 401b30e3b8b5d629635a5c613cdb7919
  - 009520053b00386d1173f3988c55d192
Output: 1
```

$$\mathbf{PEI}(c) = \texttt{a84c0efc1873b527e6d25f380da7bcf1}$$

PEI Example (3)

- Construct a PEF and calculate the PEI of e:

```
APP: ADD
APPVersion: 1.0
Inputs:
  - a84c0efc1873b527e6d25f380da7bcf1
  - e29311f6f1bf1af907f9ef9f44b8328b
Output: 1
```

$$\mathrm{PEI}(e) = \texttt{cbedcb426502400ecf4f40a7dd7de89f}$$