On the Faceting and Linking of PROV for Earth Science Data Systems

AGU 2015 Fall Meeting
San Francisco, CA
Wednesday, December 16, 2015

Hook Hua, Gerald Manipon, Brian Wilson, Dexter Tan, and Michael Starch
NASA Jet Propulsion Laboratory, California Institute of Technology

Copyright 2015 California Institute of Technology.
Government sponsorship acknowledged.

PROV-ES

PROVENANCE FOR EARTH SCIENCE

Towards a Preservation Content Standard

- U.S. Earth Science Information Partners (ESIP) Data Stewardship Committee
  - January 2011 - Proposed development of "Provenance and Context Content Standard (PCCS)" enumerating all content items to capture completely the provenance and context
  - June 2011 - PCCS matrix developed based on U.S. Global Change Research Program Workshop Report (1998), experience with NASA mission instrument teams and NOAA's Satellite Products Documentation Guidelines
- NASA
  - November 2011 – Adopted PCCS matrix to develop NASA Earth Science Data Preservation Content Specification
  - In place since November 2011; current version dated January 2013
    - Requirement for new missions
    - Checklist for missions in operation
    - Has been used for data from several spaceborne instruments (e.g., HIRDLS on EOS Aura satellite, GLAS on ICESat-1)
- ESA
  - May 2011 – Long Term Data Preservation Preserved Data Set Composition
- "*Living document*" - subject to review and change via configuration management process
- Relevant to multiple **Data Stewardship Interest Area (DSIA) working groups** in NASA's Earth Science Data System Working Groups (ESDSWG)

Summary of PROV-ES

- *"If PCS is the **what**, then PROV-ES is the **how**"*

- An **interoperable** provenance specification for use in Earth Science Data Systems to enable:
  - Capturing the increasing amount of contextual processing information of Earth Science Data Records (ESDRs).
  - Improving transparency, understanding, and trust of ESDRs .
  - Providing an interoperable representation of provenance for NASA EOS data products.
  - Reducing the barrier of entry to infusing provenance capture and consumption in Earth Science Data Systems.
- Timeline
  - Developed first OPM/PROV approach from ACCESS 2009 project
  - Started as an ESDSWG 2013 Working Group
  - 2014 WG focused on taking 2013 results and developing proof of concept working implementations
  - 2015 WG focused on implementations by early adopters
- Infusion
  - Employed additional use cases from ESIP, NASA-funded projects, DAACs, and USGCRP/GCIS
  - Initial specification and implementations employed
  - Automatic PROV-ES generation infused into initial NASA data systems
  - Faceted search interface to explore PROV-ES records

ESDSWG 2015 PROV-ES Working Group Activity

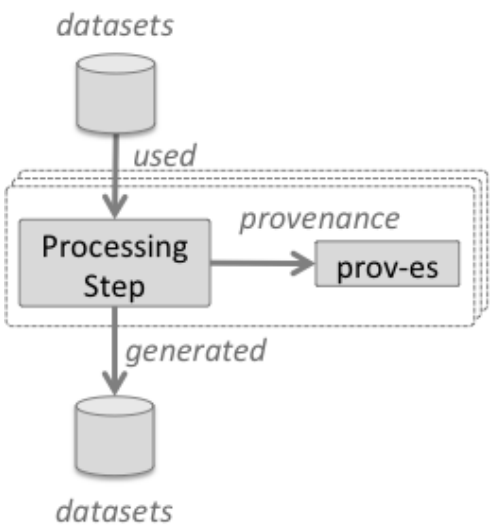
- 2015 activities
  - Additional **DAAC use cases**
    - E.g. Assess the provenance and approach needed for **supporting legacy datasets.**
  - Looking into **early adopter implementations**
  - Developing a "**provenance primer**"

- *What to capture to represent science publications for trust, scientific verifiability, and reproducibility?*
- *What is the provenance of the science analysis, findings, publication, citations, etc?*
- *How best to implement and infuse provenance into science data systems?*

Approach for PROV-ES

- Maximize **infusion potential**
- Leveraging **standards** (*W3C PROV*)
- Maximize **reuse** of open source tools and libraries supporting W3C PROV
- Minimize customizations / extensions
- *"Convention-over-Configuration"*
  - Preference for ***conventions***
  - Analogous to CF

- *Focuses on context of provenance in Science Data Systems*
- *Broadening for end-to-end life cycle*
  - *Follows PCS*



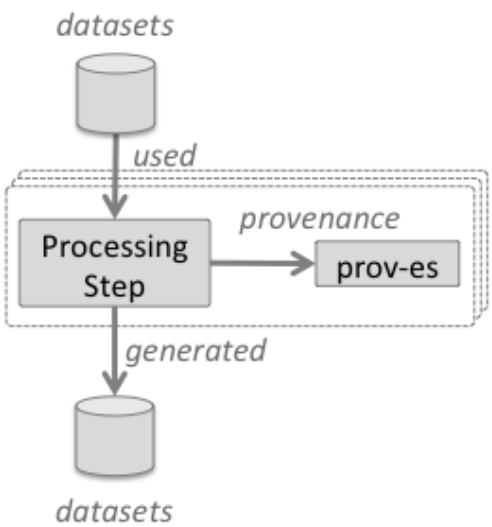
CONCEPTS

One Size Doesn't Fit All

- No single provenance specification to meet all NASA EOSDIS needs
  - Mission-specific adaptions vary
  - E.g GCIS information model effort making great progress on representing various reports and datasets
- Define a minimum basis set
  - *Use case*-driven
- Defined approach for easily adding project-specific *additional attributes*

Activities

- Examples of activities
  - Processing step
    - A running PGE with inputs and outputs data
  - Analysis
    - A science user performance an analysis with data generating plots for a publicaiton
- Capturing *internal provenance (eager)*
  - Details only known at time of activity within the activity
  - Capture appropriate provenance at time of activity
- Capturing *external provenance (lazy)*
  - Extrinsic characteristics such as data *prov:used* and data *prov:wasGeneratedBy*



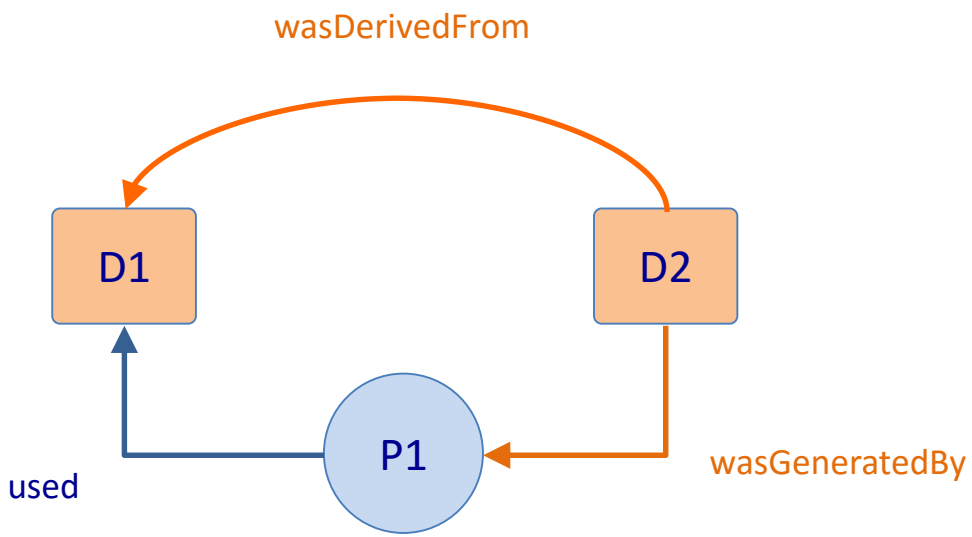
Entities

- Data resources are *prov:Entity*
  - Datasets
  - Collections
  - Granules
  - Software
- *prov:Entity* or *prov:Plan*
  - Algorithm
- *prov:Entity* or *prov:Agent*
  - Sensor
  - Platform

- *Agents have a role in generating new data*
- *Plan are a set of steps intended to be carried our by agents*

Level of Detail

- What level of detail provenance to capture?
  - *Source* provenance
  - *Transformative* provenance
- Should be *use case*-driven
  - *"A machine generate a massive amount of provenance, but a scientist can manually generate a small focused set that is highly useful."*

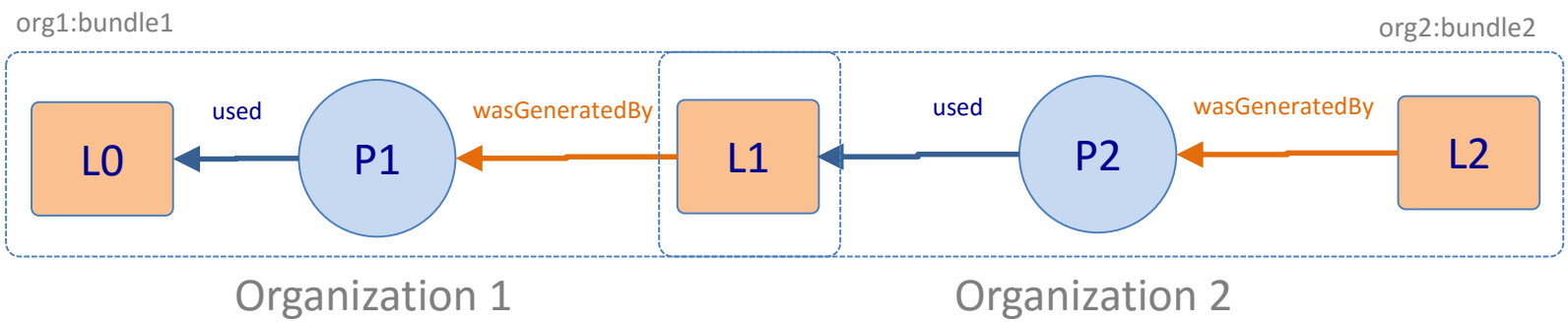


Linking Provenance Concepts

- Support real-world scenarios of **independent** provenance records
  - Status quo of limited use of coordinated end-to-end **identifiers and provenance bundles**
- Support loosely-coupled and **linked data** provenance resources
- **Ingest** distributed provenance traces
- Dynamically assemble provenance bundles
  - Generate a **reachability graph** from any given provenance resource
  - Via identifiers, with fallback to search terms

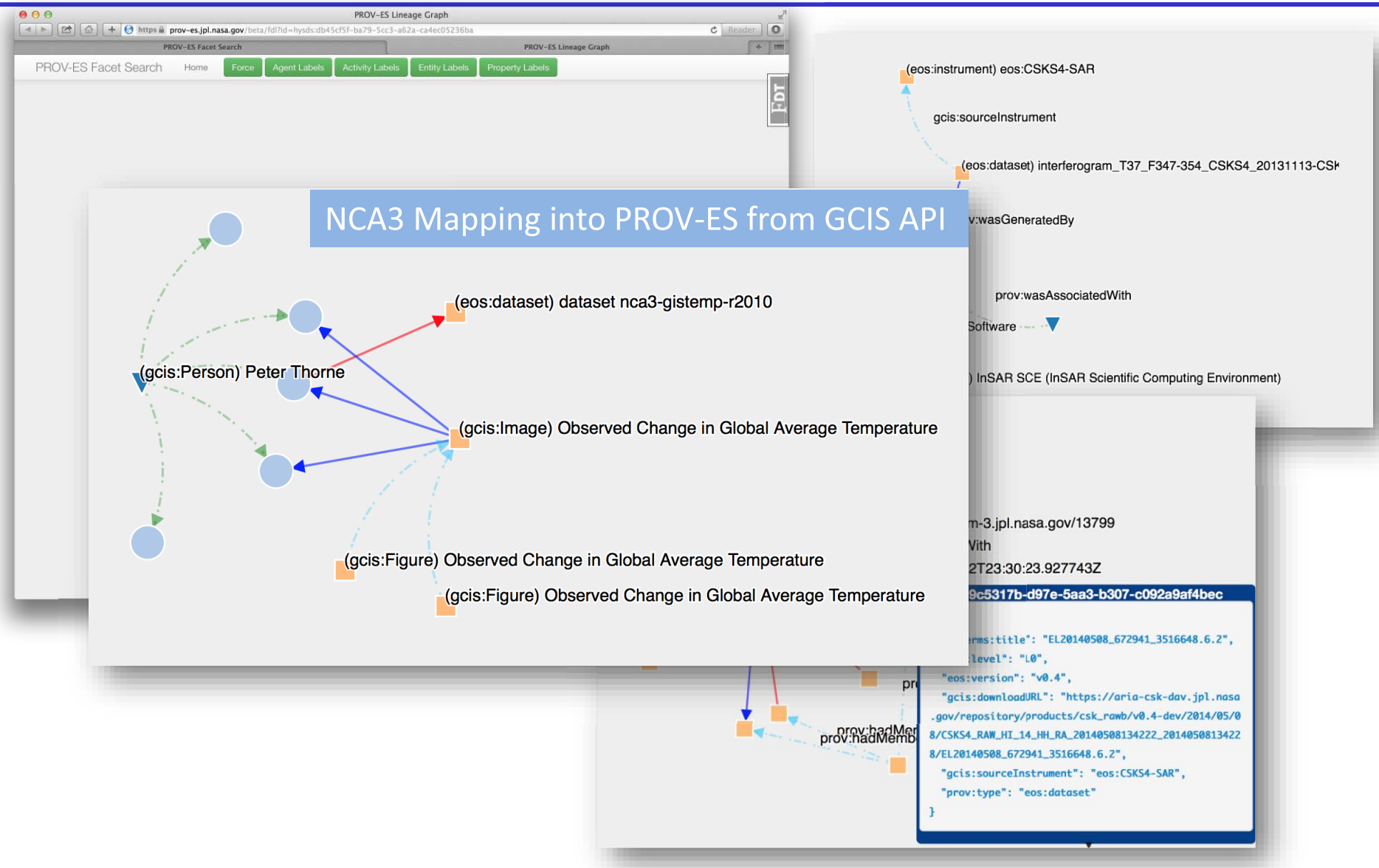
Linked Bundles

- Represent provenance in *linked* bundles
- Want each organization to generate their own provenance bundles
- Enable linked relationships across these bundles



Foster provenance representation in linked form for reuse

Interactive Visualization for Linked Relationships



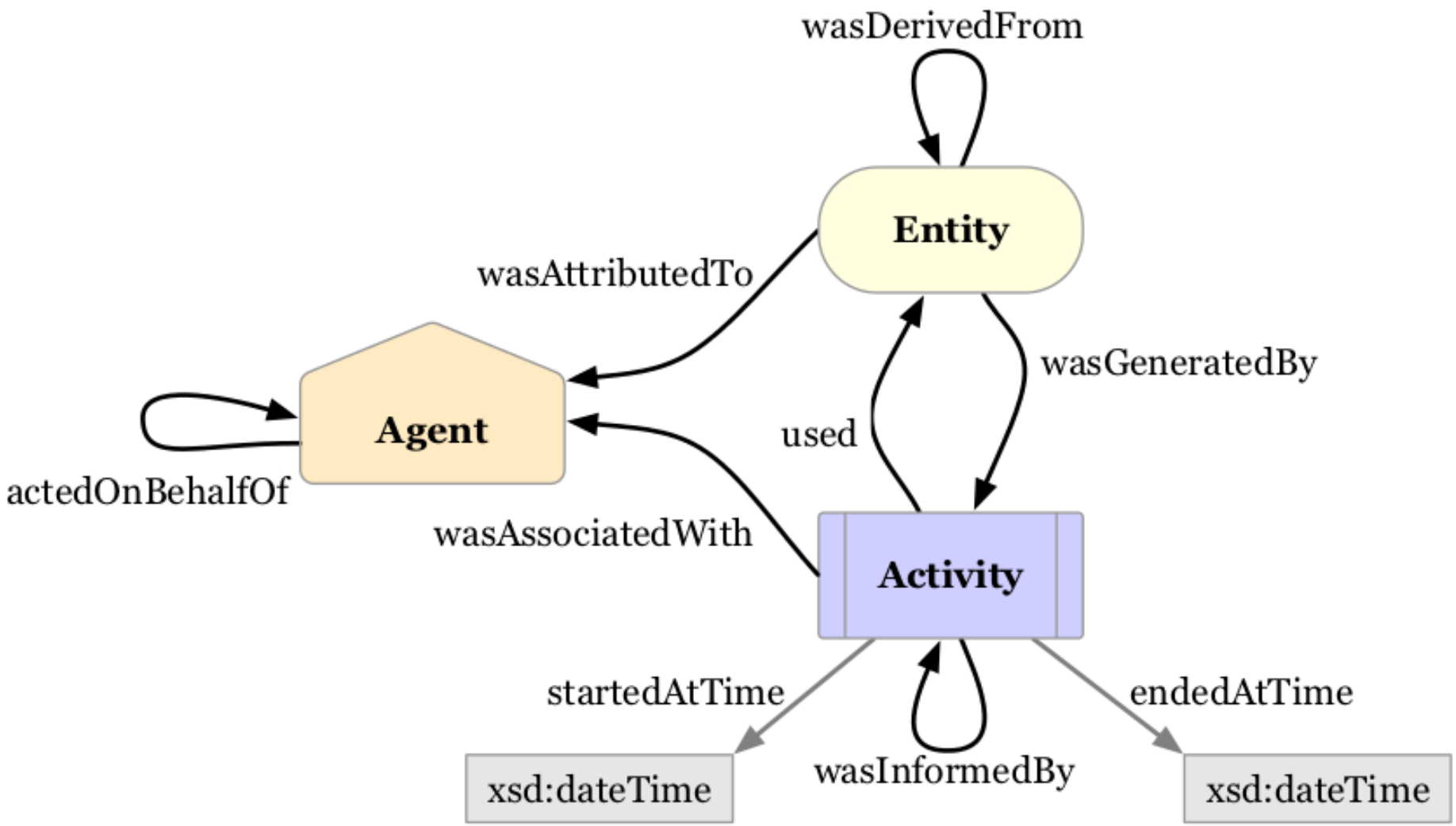
PROV-ES

APPROACH

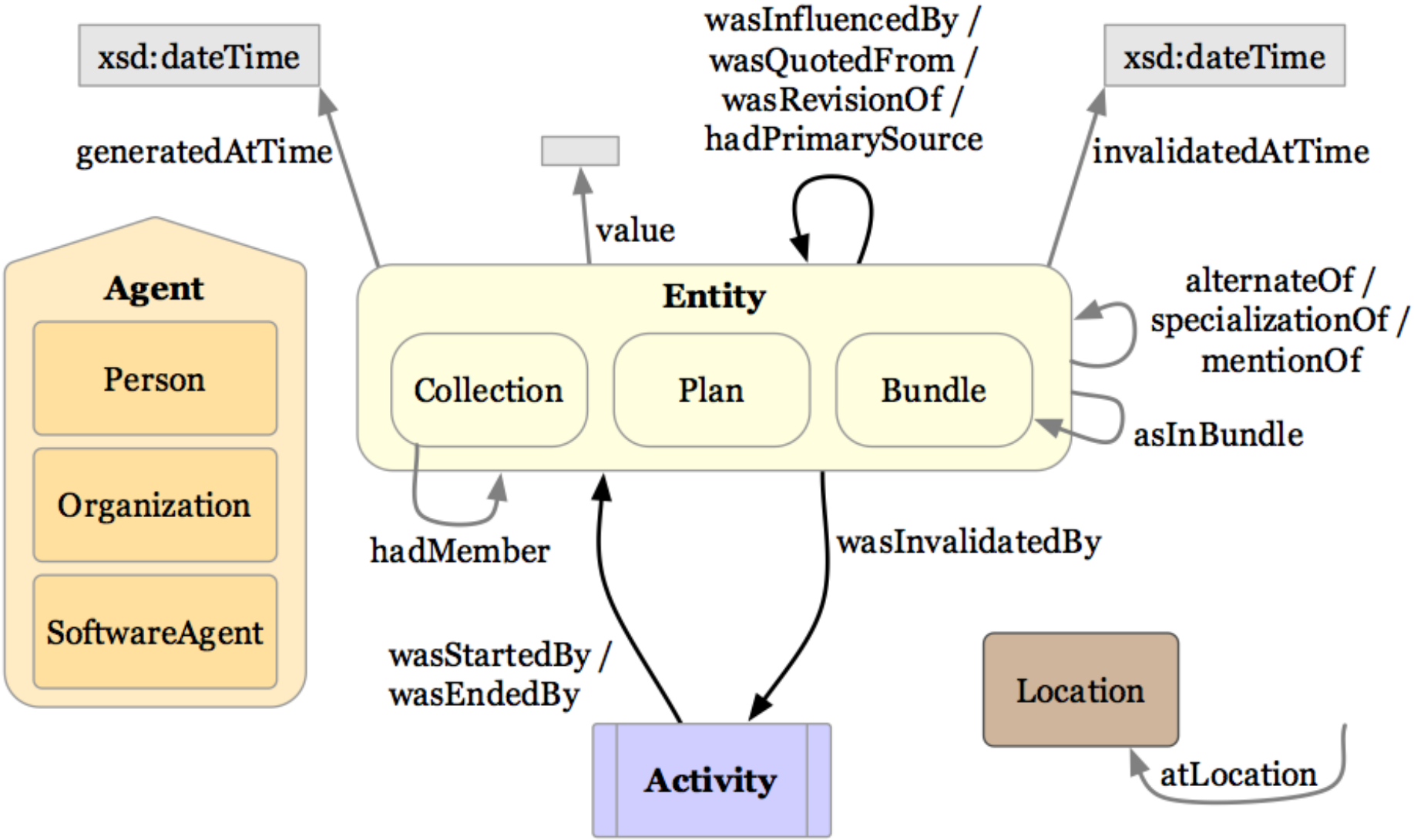
W3C PROV Family

- Model
  - PROV-DM (Recommendation)
    - PROV data model for provenance
  - PROV-CONSTRAINTS (Recommendation)
    - a set of constraints applying to the PROV data model
  - PROV-SEM (Note)
    - declarative specification in terms of first-order logic of the PROV data model
  - PROV-LINKS (Note)
    - mechanism to link across bundles
- Encodings
  - PROV-N (Recommendation)
    - Notation for provenance aimed at human consumption
  - PROV-O (Recommendation)
    - PROV ontology, an OWL2 ontology allowing the mapping of the PROV data model to RDF
  - PROV-DC (Note)
    - mapping between PROV-O and Dublin Core Terms
  - PROV-XML (Note)
    - XML schema for the PROV data model
  - PROV-JSON (Submission)
    - JSON representation for the PROV Data Model
- Access
  - PROV-AQ (Note)

W3C PROV – Starting Points

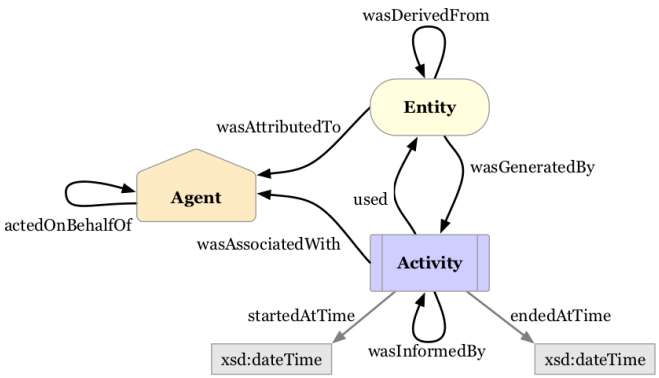


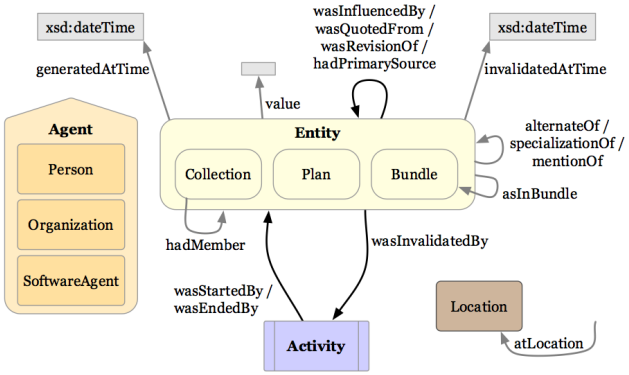
W3C PROV – Additional Concepts



PROV-ES Implementation

- **Interoperability** with W3C PROV specification
  - Maximize reuse of available tools interoperable with PROV
    - Examples include (but not limited to): ProvToolbox (Java), Prov (Python), Prov-O-Viz (Javascript)
- Support **additional attributes** to W3C PROV to encode PROV-ES concepts
  - PROV Earth Science mapping to existing W3C PROV terms
- Support domain-specific extended concepts
- Provenance capture **instrumented** into processing steps
  - Entities, activities, and agents are kept loosely-coupled to enable more versatile provenance curation
  - E.g. instrumentation Python API for PROV-ES
  - Serialization into **PROV-JSON** format
- Ease **infusion** through (re)use of existing interoperable W3C PROV tools





W3C PROV Data Model Types and Relations

Type or Relation Name	Representation in the PROV-N notation	Component
Entity	**entity(id, [attr1=val1, ...])**	Component 1: Entities/Activities
Activity	**activity(id, st, et, [attr1=val1, ...])**	
Generation	**wasGeneratedBy(**id;**e,a**,t,attrs**)**	
Usage	**used(**id;**a,e**,t,attrs**)**	
Communication	**wasInformedBy(**id;**a2,a1**,attrs**)**	
Start	wasStartedBy(id;a2,e,a1,t,attrs)	
End	wasEndedBy(id;a2,e,a1,t,attrs)	
Invalidation	wasInvalidatedBy(id;e,a,t,attrs)	
Derivation	**wasDerivedFrom(**id; **e2, e1**, a, g2, u1, attrs**)**	Component 2: Derivations
Revision	... prov:type='prov:Revision' ...	
Quotation	... prov:type='prov:Quotation' ...	
Primary Source	... prov:type='prov:PrimarySource' ...	
Agent	**agent(id, [attr1=val1, ...])**	Component 3: Agents, Responsibility, Influence
Attribution	**wasAttributedTo(**id;**e,ag**,attr**)**	
Association	**wasAssociatedWith(**id;**a,ag**,pl,attrs**)**	
Delegation	**actedOnBehalfOf(**id;**ag2,ag1**,a,attrs**)**	
Plan	... prov:type='prov:Plan' ...	
Person	... prov:type='prov:Person' ...	
Organization	... prov:type='prov:Organization' ...	
SoftwareAgent	... prov:type='prov:SoftwareAgent' ...	
Influence	wasInfluencedBy(id;e2,e1,attrs)	
Bundle constructor	bundle id description_1 ... description_n endBundle	Component 4: Bundles
Bundle type	... prov:type='prov:Bundle' ...	
Alternate	alternateOf(alt1, alt2)	Component 5: Alternate
Specialization	specializationOf(infra, supra)	
Collection	... prov:type='prov:Collection' ...	Component 6: Collections
EmptyCollection	... prov:type='prov:EmptyCollection' ...	
Membership	hadMember(c,e)	

Source: http://www.w3.org/TR/prov-dm/#prov-dm-types-and-relations

PROV-ES Concepts and Properties

ES concept	Base PROV concept	Additional Attributes			Notes
		key name	value type	cardinality	
eos:collection	prov:Entity				
		prov:type	*eos:collection*		literal
		prov:label	xsd:string		display name. typically same as short name.
		prov:location	xsd:anyURI	[0..1]	A URL to the collection. e.g. landing page.
		eos:shortName	xsd:string		
		eos:longName	xsd:string		human-readable long name of dataset
		bibo:doi	doi	[0..1]	DOI for the collection
		gcis:sourceInstrument	eos:instrument	[0..*]	A dataset was collected by using one or more instruments.
		eos:level	xsd:string	[0..1]	http://science.nasa.gov/earth-science/earth-science-data/data-processing-levels-for-eosdis-data-products/
		eos:version	xsd:string	[0..1]	
eos:granule	prov:Entity				
		prov:type	*eos:granule*		literal
		prov:label	xsd:string		human-readable name
		prov:location	xsd:anyURI	[0..1]	A URL to the granule.
		eos:partOfCollection	eos:collection	[0..1]	a granule is part of a collection
		gcis:sourceInstrument	eos:instrument	[0..*]	A dataset was collected by using one or more instruments.
		eos:level	xsd:string	[0..1]	http://science.nasa.gov/earth-science/earth-science-data/data-processing-levels-for-eosdis-data-products/
		eos:version	xsd:string	[0..1]	
eos:product	prov:Entity				
		prov:type	*eos:product*		literal
		prov:label	xsd:string		human-readable name
		prov:location	xsd:anyURI	[0..1]	A URL to the granule.
		eos:partOfCollection	eos:collection	[0..1]	a granule is part of a collection
		gcis:sourceInstrument	eos:instrument	[0..*]	A dataset was collected by using one or more instruments.
		eos:level	xsd:string	[0..1]	http://science.nasa.gov/earth-science/earth-science-data/data-processing-levels-for-eosdis-data-products/
		eos:version	xsd:string	[0..1]	
eos:file	prov:Entity				
		prov:type	*eos:file*		literal
		prov:label	xsd:string		human-readable name
		prov:location	xsd:anyURI	[0..1]	A URL to the file.
eos:platform	prov:Entity				
		prov:type	*eos:platform*		literal
		prov:label	xsd:string		human-readable name
		gcis:hasInstrument	eos:instrument	[0..*]	
eos:instrument	prov:Entity				
		prov:type	*eos:instrument*		literal
		prov:label	xsd:string		human-readable name
		gcis:inPlatform	eos:platform		
		gcis:hasSensor	eos:sensor	[0..*]	
		gcis:hasGoverningOrganization	prov:Organization	[0..*]	
eos:sensor	prov:Entity				
		prov:type	*eos:sensor*		literal
		prov:label	xsd:string		human-readable name
		gcis:inInstrument	eos:instrument		
eos:software	prov:Entity				
		prov:type	*eos:software*		literal
		prov:label	xsd:string		human-readable name
		gcis:implements	eos:algorithm	[0..*]	
		eos:version	xsd:string	[0..1]	
eos:algorithm	prov:Entity				

W3C PROV-JSON serialization

- http://www.w3.org/Submission/2013/SUBM-prov-json-20130424/

- Design goals
  - ***Interoperability***:
    - *PROV-JSON aims to map closely the PROV Data Model in order to ensure the interoperability between applications adopting the PROV recommendations.*
    - *Therefore, provenance information represented in PROV-JSON can be converted from and to its counterparts PROV-N (the Provenance Notation used by PROV-DM [PROV-N]), PROV-O (The PROV Ontology [PROV-O]), and PROV-XML (The PROV XML Schema [PROV-XML]).*
  - ***Lightweight***:
    - *native JavaScript data types are used where possible.*
  - ***Extensibility***:
    - *application-specific terms and data can be introduced via the use of namespaces and custom data types.*
  - ***Ease of use***:
    - *The design of PROV-JSON aims to follow the Convention-over-Configuration approach. Although PROV-JSON offers flexibility in extending the core specification to accommodate application-specific requirements, it is possible to produce simple PROV-JSON structures without unnecessary overhead, such as namespace declarations.*

PROV-JSON example: collection

prov concept

identifier

prov attributes

additional attributes

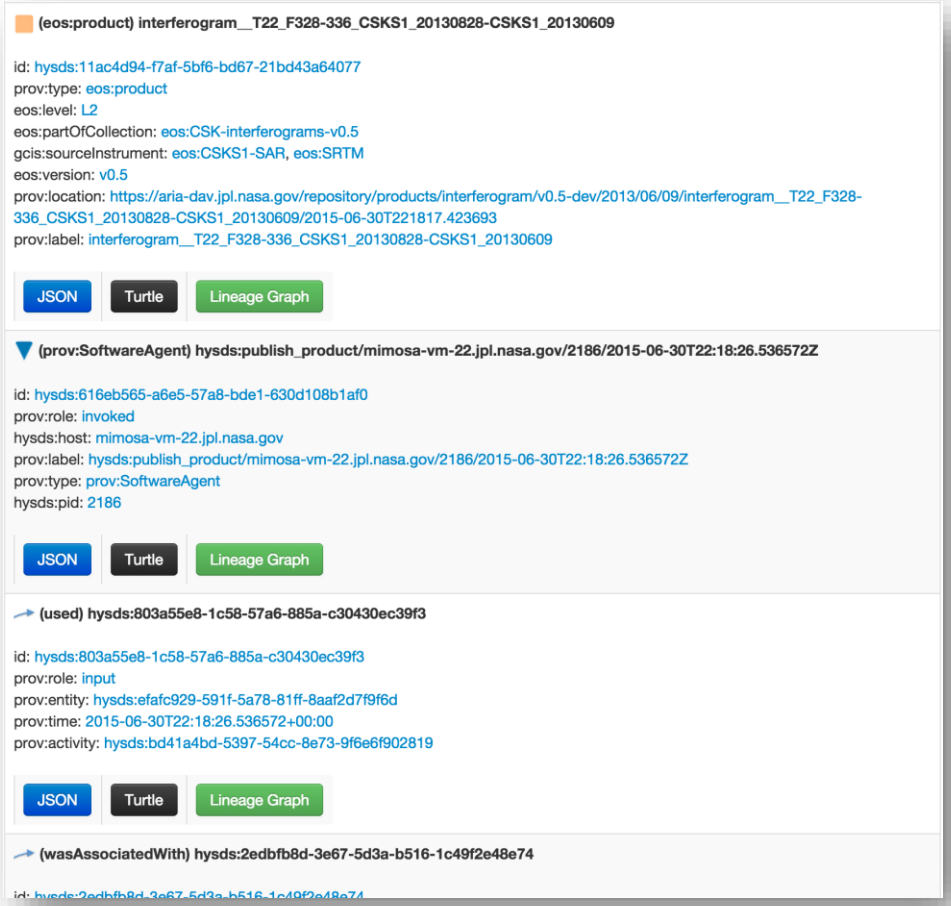
```
{
  "entity": {
    "eos:CloudSat-CPR-2B-CLDCLASS-R04": {
      "prov:type": "eos:collection”,
      "prov:label": "CloudSat CPR 2B-CLDCLASS",
      "prov:location": [
        "http://www.cloudsat.cira.colostate.edu/data-products/level-
2b/2b-cldclass"
      ],

      "eos:level": "2B",
      "eos:longName": "CloudSat CPR 2B-CLDCLASS",
      "eos:shortName": "2B-CLDCLASS",
      "eos:version": "R04",
      "gcis:sourceInstrument": [
        "eos:CloudSat-CPR"
      ]
}
  },
```

FACETED PROVENANCE

Provenance for Discovery

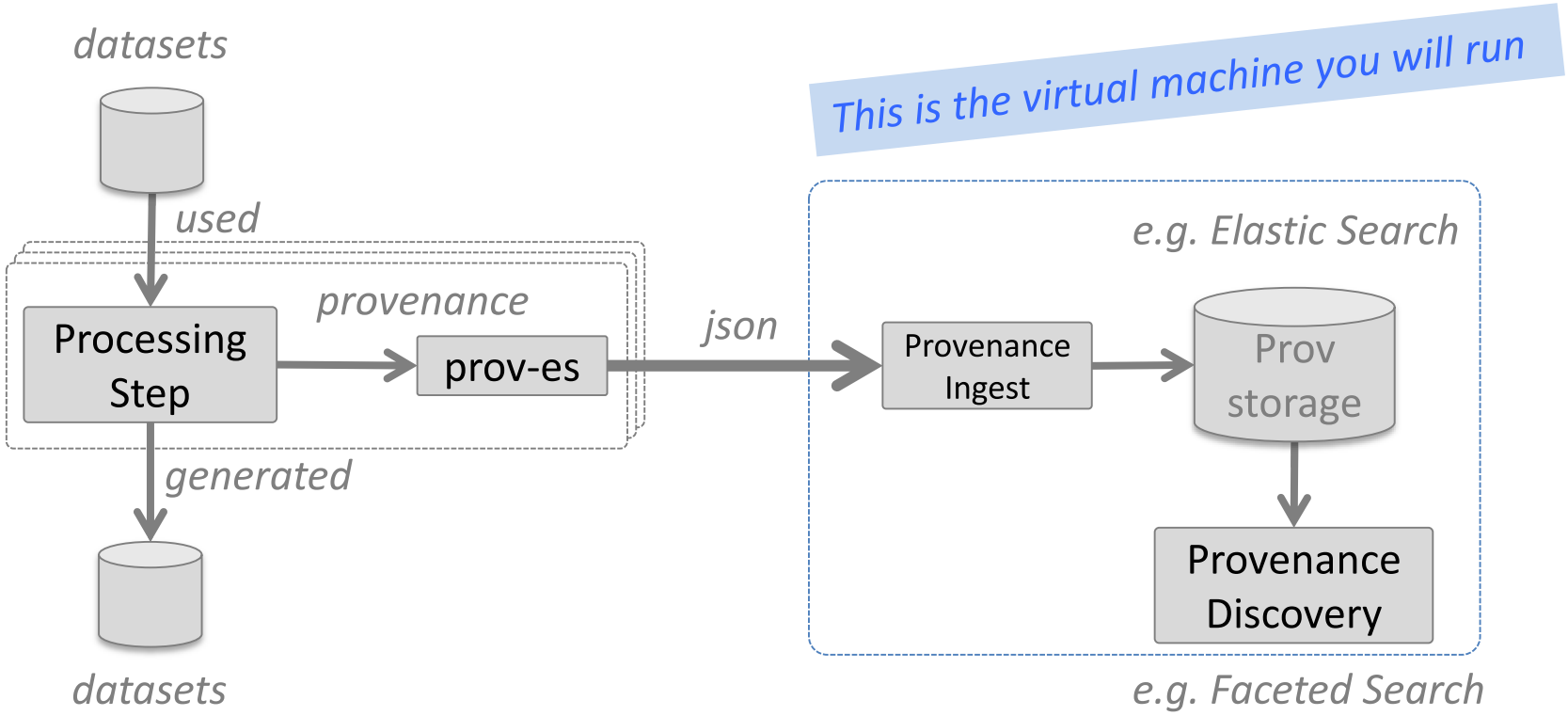
- What should the provenance "***atomic unit***" be for discovery purposes?
  - Search result set
    - As provenance bundles
    - As provenance triple
- Preference to handle provenance discovery at **unit of "triples"** instead of entire bundles
  - Entities
  - Activities
  - Agents
  - Relationships



Example result set from PROV-ES Faceted Search

From Instrumentation to Discovery

- PROV-ES "log lines" generated in JSON
- Distributed PROV-ES traces crawled and ingested into Elastic Search datastore
- Support discovery of multi-dimensionality of provenance through **Faceted Search**
  - Enables drilling down across end-to-end provenance concepts such as datasets, processing steps, algorithms, documentation, instruments, sensors, platforms, etc.



PROV-ES Faceted Search

PROV-ES Facet Search

Home API

PROV-ES Facet Search faceted search interface for earth science provenance

search term

eos:dataset ✖ L2 ✖

.. 1 – 10 of 78 next »

(entity - eos:dataset) interferogram_T37_F347-354_CSKS4_20140406-CSKS4_20140217_2015-03-23T052003.484355
id: hysds:db45cf5f-ba79-5cc3-a62a-ca4ec05236ba
prov:type: eos:dataset
gcis:downloadURL: https://aria-dav.jpl.nasa.gov/repository/products/interferogram/v0.4-dev/2014/02/17/interferogram__T37_F347-354_CSKS4_20140406-CSKS4_20140217/2015-03-23T052003.484355
gcis:sourceInstrument: eos:CSKS4-SAR
eos:level: L2
eos:version: v0.4

JSON Turtle Lineage Graph

(entity - eos:dataset) interferogram_T37_F347-354_CSKS4_20131113-CSKS4_20130809_2015-03-23T032113.464621
id: hysds:e69e8d85-859f-5930-b082-89560d1d1f4e
prov:type: eos:dataset
gcis:downloadURL: https://aria-dav.jpl.nasa.gov/repository/products/interferogram/v0.4-dev/2013/08/09/interferogram__T37_F347-354_CSKS4_20131113-CSKS4_20130809/2015-03-23T032113.464621
gcis:sourceInstrument: eos:CSKS4-SAR
eos:level: L2
eos:version: v0.4

JSON Turtle Lineage Graph

(entity - eos:dataset) interferogram_T37_F347-354_CSKS4_20131012-CSKS4_20130910_2015-03-23T020718.420966

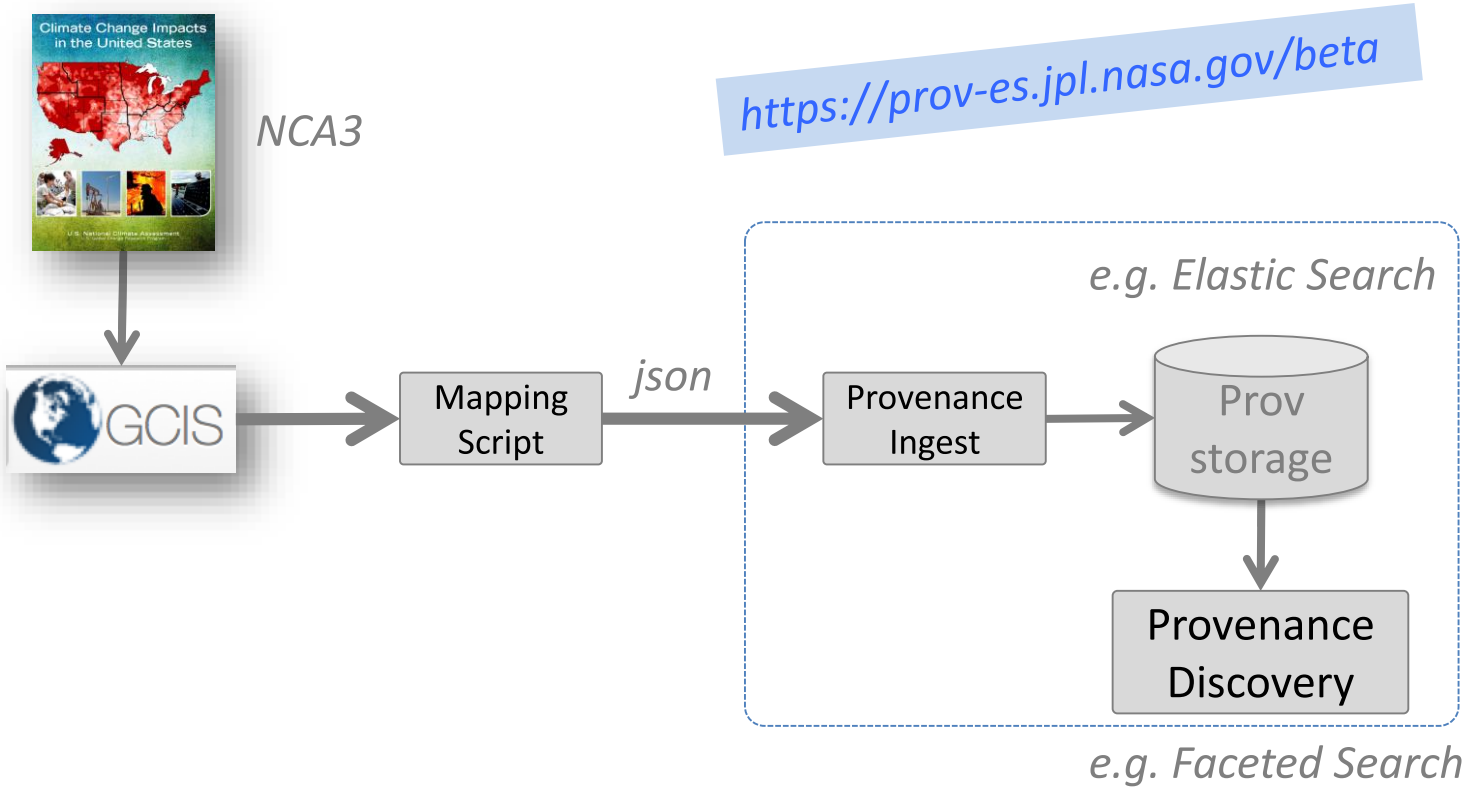
INFUSION

PROV-ES Infusion

- AIST on **geodetic InSAR processing**
  - PROV-ES production in NASA AIST's Advanced Rapid Imaging & Analysis for Monitoring Hazards (ARIA-MH) Hybrid Cloud Science Data System (HySDS)
- NASA OCO-2 Level-2 Full Physics in AWS
  - Streaming data into Amazon cloud for large-scale processing
- MEaSUREs 2012 on **A-Train atmospheric science data fusion**
  - PROV-ES production in NASA MEaSUREs's "A Multi-Sensor Water Vapor, Temperature and Cloud Climate Data Record"
- **Mapping NCA3 provenance from GCIS** into PROV-ES

Mapping From Other Sources

- Support Faceted Search of NCA3 Report provenance by mapping from GCIS into PROV-ES



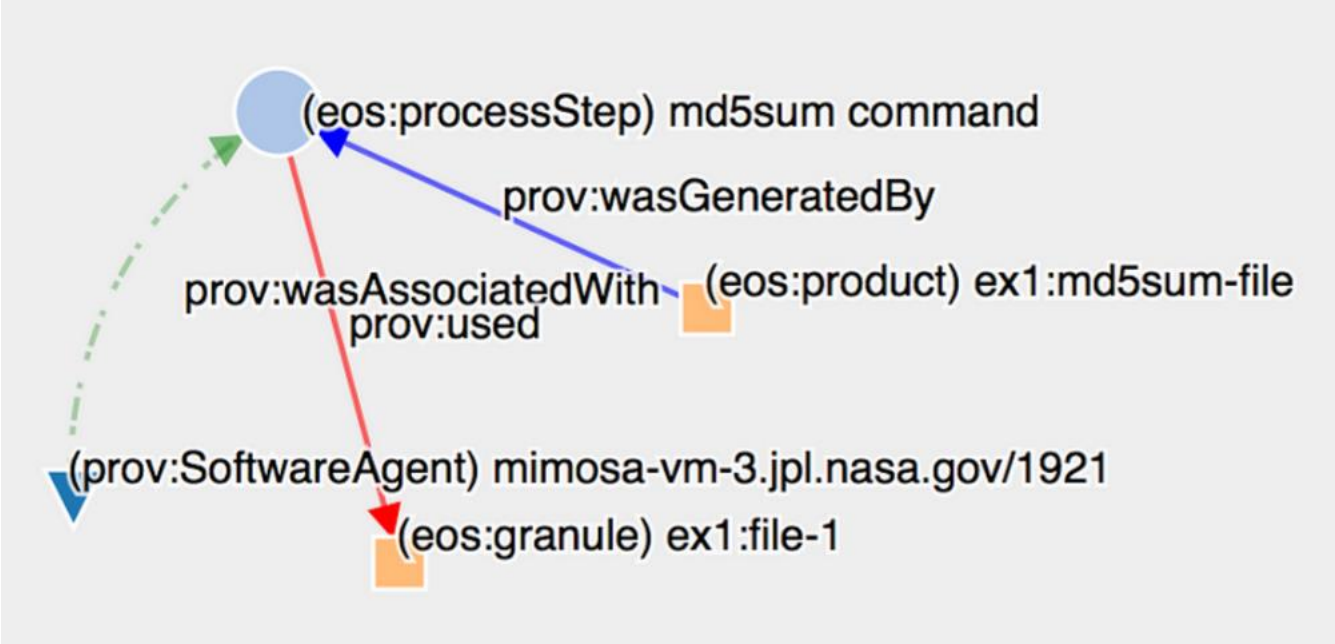
Resources

- W3C PROV
  - W3C PROV-DM
    - http://www.w3.org/TR/2013/REC-prov-dm-20130430/#prov-dm-example
  - PROV-JSON serialization
    - http://www.w3.org/Submission/2013/SUBM-prov-json-20130424/
- PROV-ES Google Drive Share
  - https://drive.google.com/folderview?id=0B1Q95ca89UmYfmNFMFZYblQxR01vbEdUcnBzMFlJTkVLelN6a1hpdkIyVTRCM2VSVzRQOUk&usp=sharing#list
- PROV-ES Faceted Search and API
  - https://prov-es.jpl.nasa.gov/beta/
- ESDSWG PROV-ES Working Group
  - Wiki
    - https://wiki.earthdata.nasa.gov/display/ESDSWG/Provenance+for+Earth+Science+(PROV-ES)+Working+Group (login required)
  - Mailing list
    - esdswg-prov-es@lists.nasa.gov
    - Subscribe at https://lists.nasa.gov/mailman/listinfo/esdswg-prov-es
  - Telecon notes
    - https://docs.google.com/document/d/1mw0ROVcW4-7W3IfZk2pizM1k6sHM2hgmRTvChrYtpO8/edit?pli=1
  - Use Case Mapping
    - https://docs.google.com/spreadsheet/ccc?key=0AlQ95ca89UmYdEJXX05NVDRhRGRWV09hcmNENnFROHc#gid=0
  - PROV-ES Mapping to W3C PROV
    - https://docs.google.com/spreadsheets/d/1gsyuKY-xD1yQaHD6QVB6vJ3YmDWL9WKPiSopW8BtuB8/edit#gid=0
- NASA Earth Science Data Preservation Content Specification (NASA ESD PCS)
  - http://earthdata.nasa.gov/sites/default/files/field/document/423-SPEC-001_NASA%20ESD_Preservation_Spec_OriginalCh01_0.pdf
- Provenance and Context Content Standard
  - http://wiki.esipfed.org/index.php/Provenance_and_Context_Content_Standard

BACKUP

Example: MD5 checksum

- Applying an MD5 checksum command on a file to get a checksum file



Example: MD5 checksum (PROV-JSON)


```
{
 "agent": {
    "ex1:my-software-agent": {
      "hysds:host": "mimosa-vm-3.jpl.nasa.gov",
      "prov:type": {
        "type": "prov:QualifiedName",
        "$": "prov:SoftwareAgent"
      },
      "hysds:pid": "1921"
    }
  },
  "entity": {
    "ex1:file-1": {
      "prov:location": "http://path/to/my/input-file",
      "prov:type": {
        "type": "prov:QualifiedName",
        "$": "eos:granule"
      }
    },
    "ex1:md5sum-file": {
      "prov:location": "http://path/to/my/output-file",
      "prov:type": {
        "type": "prov:QualifiedName",
        "$": "eos:product"
      }
    }
  },
  "activity": {
    "ex1:my-md5sum-activity": {
      "prov:wasAssociatedWith": "ex1:my-software-agent",
      "prov:label": "md5sum command",
      "prov:startTime": "2015-03-22T14:55:43.906447+00:00",
      "prov:type": {
        "type": "prov:QualifiedName",
        "$": "eos:processStep"
      },
      "prov:endTime": "2015-03-22T14:56:43.906447+00:00"
    }
  },
```

```
{
"wasAssociatedWith": {
    "hysds:my-activity-agent-association": {
      "prov:role": "softwareAgent",
      "prov:agent": "ex1:my-software-agent",
      "prov:activity": "ex1:my-md5sum-activity"
    }
  },
 "used": {
    "ex1:used-file-1": {
      "prov:role": "input",
      "prov:time": "2015-03-22T16:07:05.195235+00:00",
      "prov:entity": "ex1:file-1",
      "prov:activity": "ex1:my-md5sum-activity"
    }
  },
  "wasGeneratedBy": {
    "ex1:generated-md5sum-file": {
      "prov:role": "output",
      "prov:time": "2015-03-22T14:56:43.906447+00:00",
      "prov:entity": "ex1:md5sum-file",
      "prov:activity": "ex1:my-md5sum-activity"
    }
  }
}
```