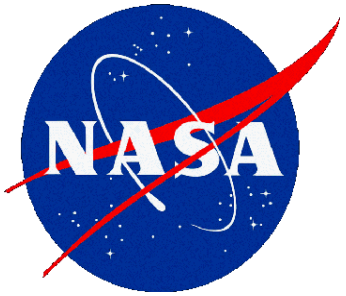


Towards an Earth Science Provenance / Context Ontology

ESIP Federation
Preservation and Stewardship
Cluster

ESIP FEDERATION

2011-01-05

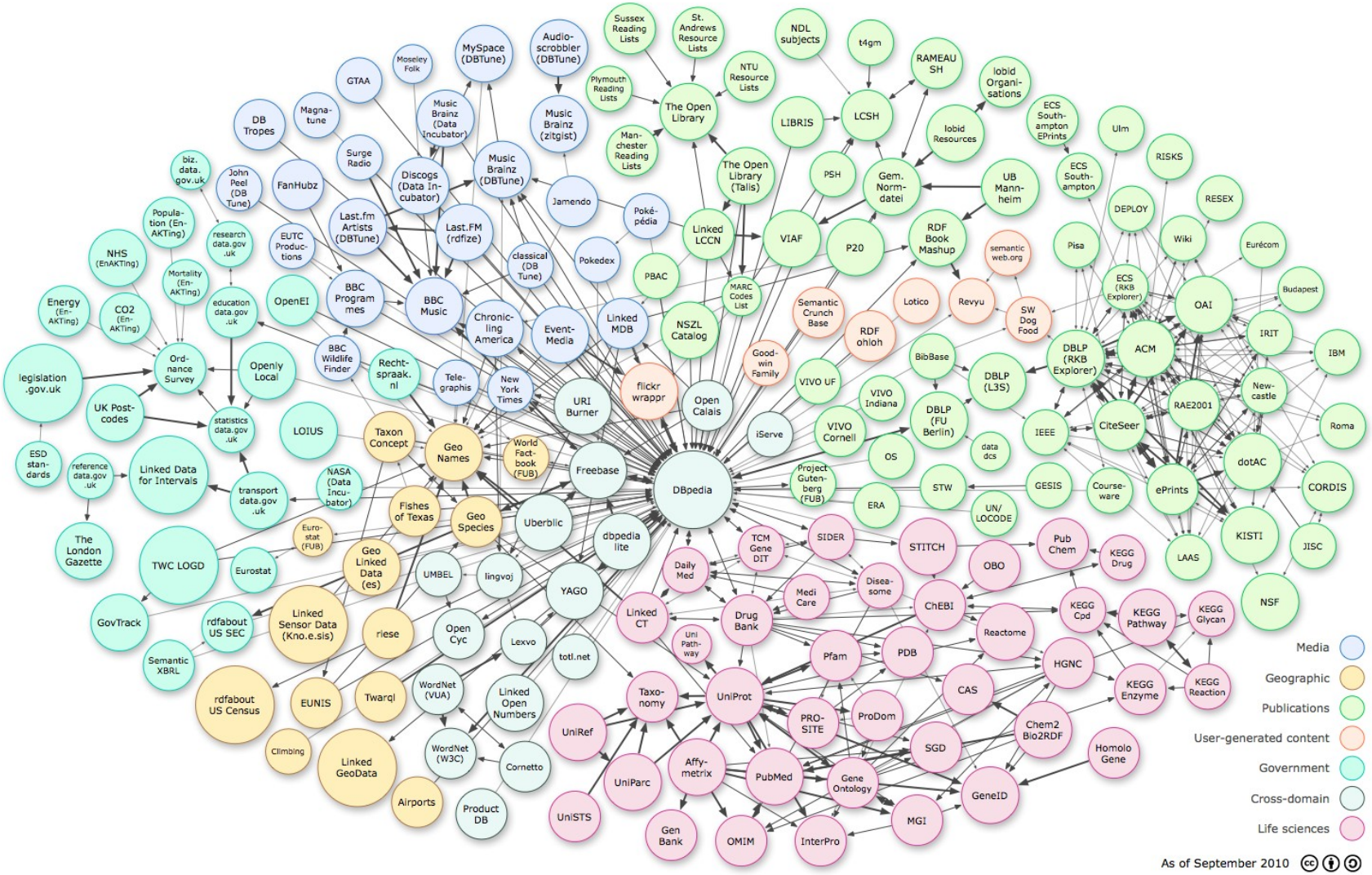
Provenance - *The information that documents the history of the Content Information. This information tells the origin or source of the Content Information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. Examples of Provenance Information are the principal investigator who recorded the data, and the information concerning its storage, handling, and migration.*

Context - *The information that documents the relationships of the Content Information to its environment. This includes why the Content Information was created and how it relates to other Content Information objects.*

What is an ontology and why do we need one?

- Previous discussions have concentrated on the content of provenance and context.
- This discussion will focus on the representation of that provenance and context information.
- We are looking for interoperability, at least among ourselves, but hopefully with the rest of the world too.
- To speak with the rest of the world, we need to figure out what language they are using.
- *Linked Data is about using the Web to connect related data that wasn't previously linked, or using the Web to lower the barriers to linking data currently linked using other methods. More specifically, Wikipedia defines Linked Data as "a term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF." – linkeddata.org*

Linked Data



"Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/"

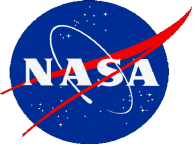
- Wikipedia: A group of methods and technologies to allow machines to understand the meaning – or "semantics" – of information on the World Wide Web.
- The traditional web describes information for people. People are smart – they can figure out what you are talking about, fill in the blanks, resolve ambiguities, etc.
- Computers are dumb (well, until the AI natural language folks catch up with humans, anyway).
- We want to describe things with a clear, distinct, unambiguous vocabulary describing facts about things and relationships between things.
- W3C semantic web organizes these facts and relationships with triples:
  - (Subject, Predicate, Object)

- Dictionaries catalog meanings of terms.
- “Folksonomies” - ad-hoc tagging of things to organically capture common terms to encourage re-use of them.
- Ontologies organize entities and concepts into common hierarchies and precisely describe relationships between those entities and concepts.
- Ontologies themselves have relationships and hierarchy.
- W3C Semantic Web ontologies can build on other ontologies and share concepts – *adopt, adapt, develop*.

- Identifiers are key.
- Data is our big problem, covered extensively earlier.
- Every other entity we want to reference also needs a good identifier.
- Consider a scientist, what can be used as a globally unique, persistent identifier?
  - Names aren't unique, can change
  - Organizations change name and address
  - Email addresses change, even in the same organization
- Semantic web uses URIs as identifiers. Linked data principles require they be resolvable.
- When two entities reference something that is the “same” (semantically equivalent), they should use the same identifier (or assert their equivalence).

- Various systems and disciplines have developed ontologies that represent certain types of information relevant to their purpose.
- Through a series of “Provenance Challenges”, commonalities were discovered and distilled into

  The Open Provenance Model
  - (Cue Hook!)



Provenance Vocabulary Mappings

- W3C Provenance Incubator Group

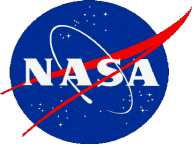
  `http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Vocabulary_Mappings`

- One of their activities has been to examine the existing provenance vocabularies and ontologies and construct a mapping between common terms.
  - Open Provenance Model
  - Provenir ontology
  - Provenance Vocabulary
  - Proof Markup Language
  - Dublin Core
  - PREMIS
  - WOT Schema
  - SWAN Provenance Ontology
  - Semantic Web Publishing Vocabulary
  - Changeset Vocabulary
- Mappings based on concepts from Simple Knowledge Organization System (SKOS)

- Provenir ontology properties for inclusion in table of provenance terms:
  - **provenir:part_of** – This property is used to represent parthood relation between entities (both class and instance-level).
  - **provenir:contained_in** - This property is used to represent containment relation between entities.
  - **provenir: adjacent_to** - Spatial proximity is represented by this property.
  - **provenir:transformation_of** – This property is similar to the ro:transformation_of property that is asserted between two entities that preserve their identity between the two transformation stages.
  - **provenir:preceded_by** - This property is used define a temporal ordering of processes, which may or may not be modeled be linked by a common artifact (such as in OPM:triggered_by).
  - **provenir:located_in** - An instance of data or agent is associated with exactly one spatial region that is its exact location at given instance of time.
  - **provenir:has_temporal_value** - This property is used to explicitly associate temporal value with individuals of Provenir classes.

- provenir:process is mapped to opm:Process using skos:broadMatch
  - provenir:process allows modeling of processes that may or may not result in creation of new entities (provenir:data). opm:process is defined as "...actions resulting in new artifacts."
- provenir:data is mapped to opm:Artifact using skos:relatedMatch
  - OPM does not define the relationship between opm:Artifact and opm:Account. Specializations (sub-class) of the provenir:data can be used to model information entities represented by both opm:Artifact and opm:Account. Further, opm:Artifact are "immutable piece of state" whereas provenir:data allows representation of both immutable entities as well as entities that can undergo change or modification without losing their identities (for example, an organism retains its "identity" from its birth to its death).

- provenir:derives_from is mapped to opm:wasDerivedFrom using skos:relatedMatch
  - provenir:derives_from property represents the derivation history of data entities as a chain or pathway. Unlike opm:wasDerivedFrom, provenir:derives_from may or may not represent an existential relationship between entities.
- provenir:has_participant is mapped to opm:used using skos:broadMatch
  - provenir:has_participant property describes the participation of provenir:data entities in a provenir:process. Unlike opm:used, provenir:has_participant may or may not represent an existential relationship between the provenir:data and provenir:process, in other words the provenir:process may or may not require the existence of the provenir:data to initiate/terminate.
- provenir:has_participant is mapped to opm:wasGeneratedBy using skos:broadMatch
  - opm:wasGeneratedBy can be interpreted as an inverse property of opm:used. provenir:has_participant allows modeling of more types of relationships between data and process, in addition to the existential relationship modeled by opm:wasGeneratedBy.



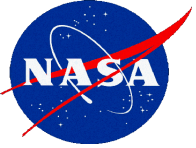
Provenance Vocabulary

- prv:Execution is narrower than opm:Process
  - Both terms refer to a specific execution of a process. However, while the definition of opm:Process only requires that this execution must have started in the past, prv:Execution explicitly refers to executions that also have already been completed.
- prv:Artifact is similar to opm:Artifact
  - prv:Artifact is anything that can be the input to the execution of a process or (one of) the result(s) of such an execution. Hence, the Provenance Vocabulary does not understand artifacts as an "immutable piece of state" as OPM does. (Question: is prv:Artifact broader than opm:Artifact?)
- prv:File is narrower than opm:Artifact
  - prv:File is a special kind of artifacts represented by opm:Artifact.
- prv:precededBy is narrower than opm:wasDerivedFrom
  - Deriving something (a prv:DataItem in the case of the Provenance Vocabulary) from a preceding version of it is a special kind of deriving something from something else.
- prv:usedData is narrower than opm:used
  - Since prv:DataItem is a special kind of opm:Artifact using a data item for the execution of a process is a special kind of using an opm:Artifact for the process.

- Three ontologies: Provenance, Justification, Trust.
- pmlj:InferenceStep is a relatedMatch to opm:Process
  - Both terms refer to a specific execution of a process. While the term InferenceStep might seem to imply a subtype of step, it is used broadly to apply to many types of mathematical/computational process executions as well as logical inference and thus appears to be a match for opm:Process
- pmlp:Information is a closeMatch to opm:Artifact
  - Information "supports references to information at various levels of granularity and structure" and is used in examples to represent text strings and scientific data files and thus appears to be a close match to the opm:Artifact concept
- pmlp:Source is related to opm:Artifact
  - Source appears to be used both for things that would map to opm:Artifact (i.e. Documents, web pages) as well as opm:Agents (i.e. an agent/person). Sources are associated with Information that comes from them (hasSourceUsage), which, as discussed later, appears to be a form of opm:wasDerivedFrom relation where the 'usage' process is not described.
- pmlp:hasSourceUsage is narrower than opm:wasDerivedFrom
  - PML doesn't appear to have a general causal connection between pmlp:Information instances but does provide such a link between Sources (which can be documents) and Information (i.e. a text string from that document).

- dcmitype:Event is related to opm:Process
  - dcmitype:Event represents a non-persistent, time-based occurrence. An opm:Process is similarly an individual non-persistent occurrence, though with a causation-based rather than time-based identity. dcmitype:Event could also denote a future occurrence, while opm:Process refers to past occurrences only.
- dct:replaces, dct:source, dct:hasPart, and dct:references are narrower than opm:wasDerivedFrom
  - Each of the Dublin Core terms listed relates an artifact to another from which it is in some way derived, so is a kind of opm:wasDerivedFrom.
- dct:requires, dct:isRequiredBy is narrower than opm:used
- dct:source is broader than opm:wasGeneratedBy

- premis:Event is mapped to opm:Process using skos:relatedMatch
  - A premis:Event describes any event applied to a premis:Object (bitstream, file, representation). This event may or may not change the premis:Object. Examples are a file format migration or an MD5 check. The premis:Event is timebased, that is why it is related to opm:Process.
- premis:Object is mapped to opm:Artifact using skos:narrowMatch
  - A premis:Object can only be a bitstream, file or aggregation (representation). It does not refer to metadata, which is the reason for the narrow match.
- premis:relatedObjectIdentification is mapped to opm:wasDerivedFrom using skos:broadMatch
  - A premis:relatedObjectIdentification relates two premis:Objects to each other. The relationship can be structural (a premis:Object as part of another premis:Object) or a derivation (a premis:Object can be migrated from another premis:Object). --> broader match.
- premis:relatedEventIdentification is mapped to opm:wasGeneratedBy using skos:broadMatch
  - a premis:relatedEventIdentification relates a premis:Object to a premis:Event. it is broadly matched to opm:wasGeneratedBy because the relationship between the premis:Object and premis:Event can be broader than just causal. The premis:Object could be used, e.g., as input for the premis:Event.



Some thoughts

- We want to go beyond the simple workflow provenance to encompass formal description of all of the elements of *provenance* and *context* previously described.
- We want more precise terms not just “artifact” or “file”, but distinguishing data granules, data levels, calibration, ancillary data, validation data, etc.
- We need good common, globally unique and distinct, persistent, identifiers for all our artifacts.
- It should be possible to follow our graphs across our own organizations and elsewhere into the “linked data” cloud.

- ❑ Develop use cases for Provenance/Context applications
- ❑ Grow in parallel with Provenance/Context content standard
- ❑ Analyze existing work
  - Provenance vocabularies and OPM
  - SWEET (Semantic Web for Earth and Environmental Terminology)
  - VSTO (Virtual Solar-Terrestrial Observatory)
  - GeoBrain
  - Giovanni
  - MMI (Marine Metadata)
  - GeoSciML
  - Tetherless World Constellation / InferenceWeb
  - U of Al, GMU, RPI, etc.
- ❑ Work with experts over in the next room!
- ❑ Take advantage of ESIP Test Bed to try things out