Dataset and Feature-Level Provenance Integration for Spatial Datasets

Nicholas J Car

Geoscience Australia, Symonston, ACT, Australia; Email: nicholas.car@ga.gov.au

SUMMARY

Large, multi-agency projects such as the Foundational Spatial Data Framework are interested in capturing the provenance of their spatial datasets as they are processed and combined to form products. Additionally, work is underway at the CRC for Spatial Information and elsewhere to track the provenance of the production of individual elements (features) within spatial datasets.

How can we reconcile these provenance situations, given the different levels of granularity? Can we relate the provenance from lower-level systems to higher levels? Can we use common tools and methodologies? This paper and talk present provenance modelling work that has taken place at Geoscience Australia and CSIRO to solve these issues. The differing levels of granularity can be related however, for interoperability, a standard must be used and we've used PROV.

Keywords: spatial dataset, provenance, multi-granularity, spatial data infrastructure, transparency

INTRODUCTION

Transparency of process and some measure of reproducibility are requirements for information hoping to engender a high degree of trust in its users. A system-independent, international, standard known as PROV [1], now exists to generically represent the provenance of *things* (i.e. anything that was produced) and can be used to describe the production of national spatial datasets. The use of such standards ensures the interoperability of provenance description across systems and the longevity of the understanding of such descriptions.

A presentation at a previous Locate conference by this author [2] demonstrated the standardised provenance representation of a single map's production, down to the 'layers' level using a formulation of PROV, PROV-O. More recent work by the Cooperative Research Centre for Spatial Information (CRC-SI) has represented individual geoprocessing toolkit actions undertaken to produce elements within spatial datasets using an extension to the PROV-O that they made, called GeoPROV [3]. Additionally, the Foundational Spatial Data Framework (FSDF) project[1] intends to use PROV to represent the overall information flow from base data to FSDF data products.

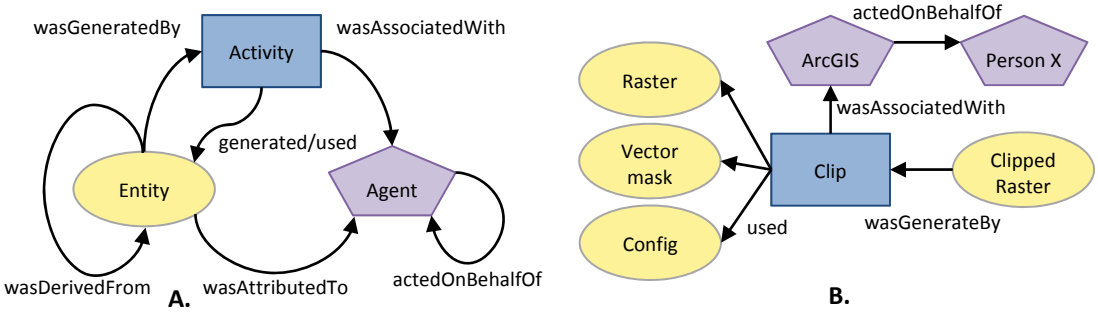

Figure 1. A: The basic PROV-O classes and their relationships. B: A simple implementation of PROV-O describing the clipping of a raster image using ArcGIS[2].

[1] http://www.anzlic.gov.au/foundation_spatial_data_framework
[2] https://esriaustralia.com.au/products-arcgis-software

These three bodies of work are all use PROV at different granularities and for slightly different purposes, however all three intend to enhance the transparency of the production of spatial products.

In this paper we will demonstrate how standardized provenance information recorded by different processes at different levels of granularity can be conceptually combined. Such combination is necessary in order to provide point-of-truth provenance information for data products.

USING THE PROV DATA MODEL

PROV-O provenance depiction

The PROV Data Model [1] consists of 3 main classes of concepts: *Entities* (things), *Activities* (events that act on Entities) and *Agents* (people or systems that trigger *Activities*). A diagram of these classes and their basic relationships is given in Figure 1A. An implementation of PROV-O for a simple geoprocessing task exhibiting a granularity similar to the examples in [2] is given in Figure 1B.

PROV-O representations of provenance are graph-based in structure. Graphs[3] by their nature, unlike relational databases, contain their schema within the data [4]. This allows for infinitely detailed and infinitely large representations of systems' provenance with the schema of the graph not limiting extensions of the information stored about items in it, or the links between items. Real limits on the information stored are only imposed by the ability of users to capture provenance information and for storage systems to physically cater for its management.

Additions to provenance graphs can be made by inserting new data into the graph, joining on appropriate prov:Activity[4], prov:Entity or prov:Agent nodes. Since PROV-O uses a Resource Description Framework (RDF)[5]-based graph, each node's identity is given as a URI[6], thus one just needs to discover the URI for a node and graph additions can be made.

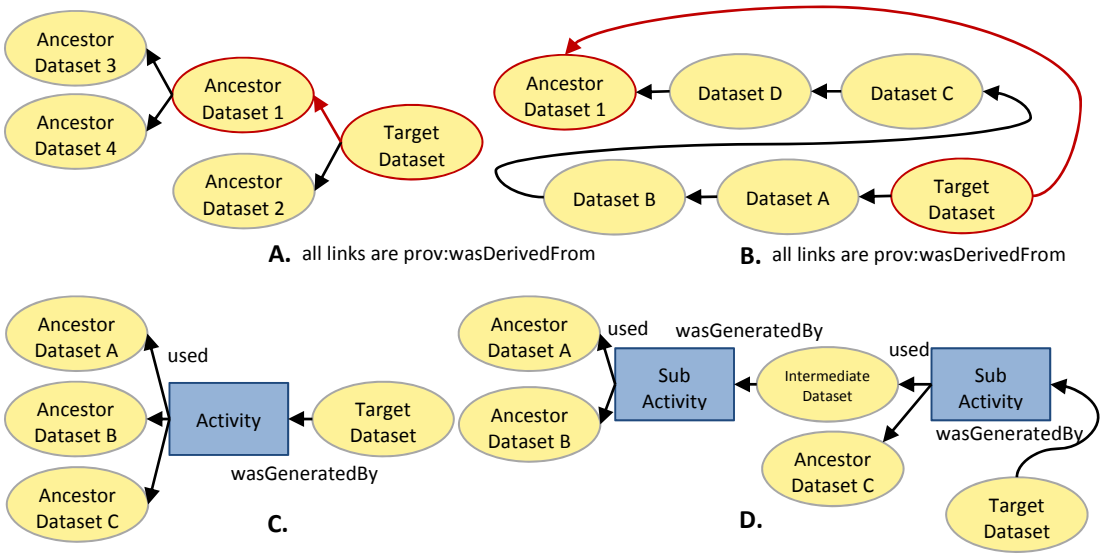

Figure 2. A: A high-level dataset provenance graph. B: Two datasets from A with intermediate datasets shown. C: A 'black box' Activity consuming 3 datasets and producing 1, D: The same datasets as C with the 'black box' broken down into two parts and an intermediate dataset shown.

PROV-O used at different levels of granularity

Detail insertion

If a system records the provenance of a dataset at a high level – perhaps just recording which datasets are a target dataset's ancestors (see Figure 2A) – and this information is stored, additions to that can

[3] https://en.wikipedia.org/wiki/Graph_(abstract_data_type)
[4] PROV-O objects are denoted prov:{CLASS_NAME}, e.g. a PROV Agent is denoted prov:Agent
[5] https://en.wikipedia.org/wiki/Resource_Description_Framework
[6] https://en.wikipedia.org/wiki/Uniform_Resource_Identifier

be made later that fill in intermediate steps (see Figure 2B). Additionally, if a process records high-level provenance noting an activity that has taken place and that consumes (*prov:used*) and produces (*prov:generated*) datasets (see Figure 2C) which is then stored, that too can be added to later by recording activities at a finer granularity and any intermediate datasets (these don't necessarily have to be persisted: their existence may only be represented) (Figure 2D).

As well as increasing the granularity of provenance graphs by filling in details, detailed provenance graphs can have their granularity decreased by querying. The SPARQL query protocol[7] is for RDF-based graph databases what SQL is for relational databases. It is able to skip over nodes in provenance graphs by using path-based, transitive queries. This skipping of intermediate nodes allows one to, for example, discover the ultimate ancestor of a dataset, despite there being any number of intermediate ancestors. For the scenario shown in Figure 2B, a path-based SPARQL query can tell the user that "Ancestor Dataset 1" is the ancestor of "Target Dataset".

Dataset Subsetting

Representing dataset subsetting is important for linking provenance at different granularities as subsetting can be the tie-in points for systems' reporting provenance at different scales.

There are a range of options regarding the recording of provenance for datasets that are subsets of other datasets. The PROV data model doesn't directly prescribe how one should represent subsetting of datasets or how a part of a dataset is related to the larger whole: such instructions require far more detail than the generic PROV data model can deliver. One method of representing detailed dataset subsetting is shown in Figure 3A. As per that diagram, a dataset subset is created via a prov:Activity subsetting procedure with instructions as to how the sub-setting was undertaken recorded in a prov:Plan class object which is a specialised prov:Entity used to denote methodology. The prov:Plan object could hold computer code, detailed manual methodology or other instructions.

Another method for representing subsetting is shown in Figure 3B. In this formulation, instructions for performing the subsetting are not given with additional input data but are described by typing the subsetting prov:Activity. An example could be a prov:Activity of a hypothetical class such as "TemporalExtentSubsetting" where the instances of such always subset the Large Dataset with some selection of a temporal extent. Sufficient metadata for the types subsetting activity, such as actual temporal extents, would need to be provided elsewhere (i.e. not in the provenance graph) in order to remove ambiguity from the action. One location for such metadata could be a register of typed activities maintained for use by a certain set of workflows. Figure 3C presents a combined formulation in which the typed prov:Activity demands that certain inputs to the subsetting action, in addition to the dataset from which a subset was taken, be represented in the provenance graph.

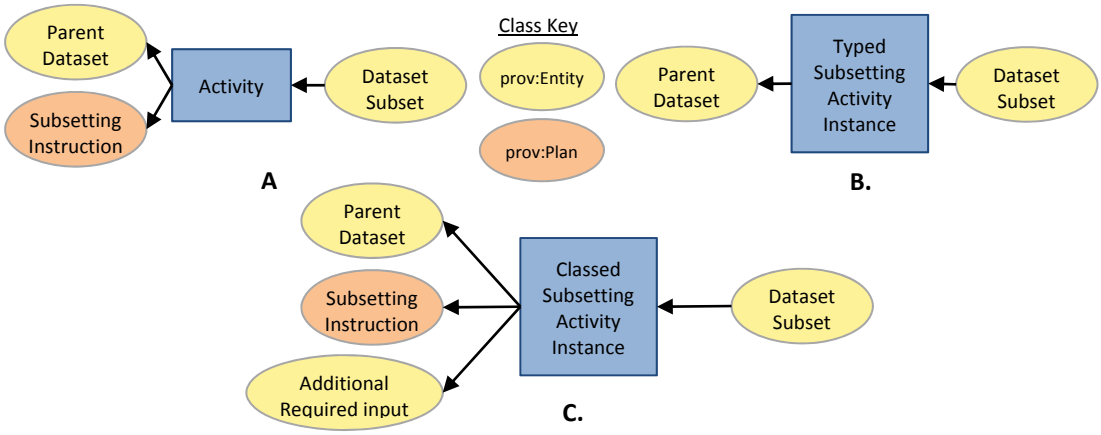

Figure 3. PROV-O Representations of subsetting actions. A: Using a prov:Plan object to hold subsetting instructions. B: By classifying the subsetting prov:Activity instance. C: Formulation combining A & B where required inputs are specified by the typed subsetting prov:Activity.

[7] https://en.wikipedia.org/wiki/SPARQL

Dataset Merging & Splitting

Dataset merging and splitting can be modelled like dataset subsetting with either prov:Plan objects or typed prov:Activities, or a combination of the two, providing the instructions the action. It follows that the representations of dataset merging & splitting are akin to that of dataset subsetting shown in Figure 3 but with multiple input (merging) or multiple output (splitting) datasets.

REPRESENTING FEATURE AND DATASET PROVENANCE

Limited sets of typed actions for features

Where the provenance of features manipulated via a limited set of actions is to be represented, the representation shown in Figure 3A or B may be used and then aggregated to dataset-level provenance. Figure 4 shows a representation of a hypothetical set of feature manipulation actions using the formulation given in Figure 3B: "selected", "not-selected", "merged", "split" and the generic "alter" typed prov:Activities are shown. These actions may have been carried out against features in one or more datasets and the results stored in a resultant dataset. They may be the result of specialized spatial tools, such as ArcGIS, certain actions of which are modelled using PROV-O in [3].

For a scenario in which features from one dataset (perhaps classes of vectors in a cadastral dataset) may be manipulated to form features in another dataset, such actions and their associated features may be represented as in Figure 4. Figure 4A shows feature-level manipulation and parts B, C & D dataset-level integration of feature-level provenance.

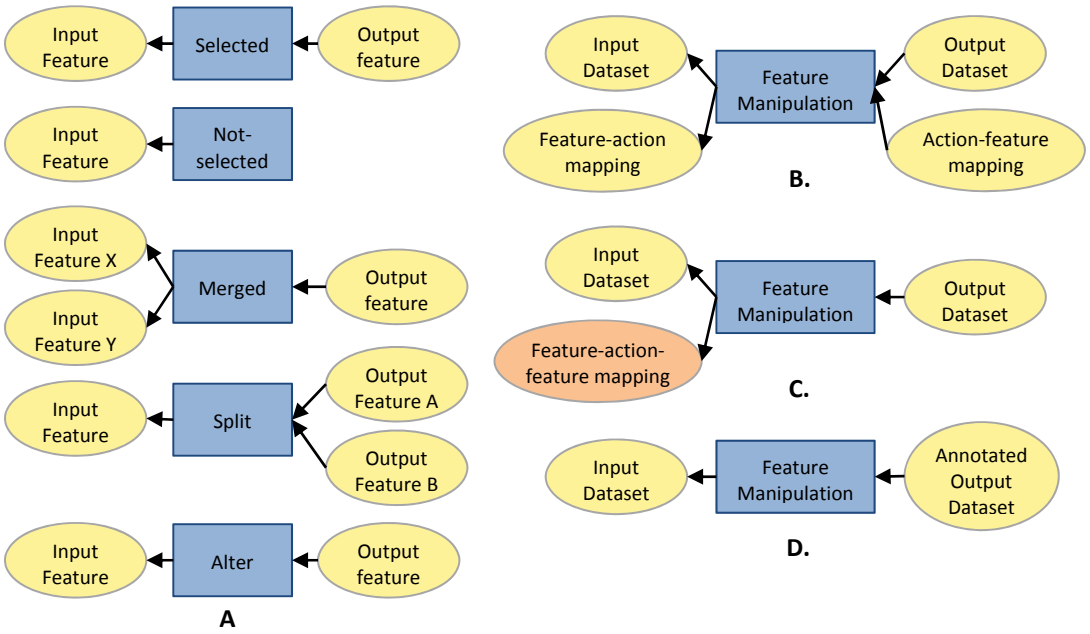

Figure 4. A: Feature manipulation actions as per Figure 3B. B: Aggregation of features manipulated into datasets with feature/action mappings preserved as prov:Activity inputs and outputs, C: Aggregation of features manipulated into datasets with feature/feature mappings preserved as a prov:Activity, prov:Plan input and, D: Aggregation of features manipulated into datasets with feature/feature mappings preserved by annotating output features with links to actions performed and features within the input dataset.

Identifier handling

The three feature-level provenance integration strategies presented in Figure 4B, C & D all rely on feature identification in order to link input and output features to their manipulation actions and each other. All three strategies are therefore dependent on either a mechanism for minting IDs for features that, although they are part of a dataset, are referenceable from outside that dataset or a feature register

that records feature identity independently from any particular dataset. The first case is implementable by URI patterns in accordance with Linked Data[8] principles where the feature-level URIs are mapped to a higher level dataset-level URI via a relative, logical path. The second case requires a master feature register that can mint identifiers for features which can be referred to by any dataset containing them. Such a register may provide access to authoritative copies of their data, but this is not necessary.

In addition to the requirements listed above, the part B scenario also relies on the identification of, and storage of, the instance of each typed prov:Activity in order to preserve feature-level provenance since the feature linking is not directly coupled – it is in two parts: input feature(s) → action then action → output feature(s). The part C scenario conceptualizes the input and output feature mapping as a prov:Plan object for such a mapping if it contains feature-to-action-to-feature mappings that act as the entire instructions for the "Feature Manipulation" prov:Activity.

The part D scenario annotates each feature in the output dataset with the identity of its relevant manipulation actions instance as well as the input features manipulated. Such a formulation is also dependent on the identification and storage of the instance of each typed prov:Activity, as per part B, but it also has a shortcoming not present in parts B & C: actions that result in no output feature, such as feature non-selection, will not be identifiable in the annotated output dataset.

FSDF DATASET PRODUCTION CASE STUDY

Detail insertion, dataset subsetting, aggregating and splitting actions, as described two sections above, can easily be used in specific spatial data scenarios. Feature-level action recording and feature/action mapping as outlined in the section above can be applied to spatial datasets if the feature manipulation systems are able to record it and if the dependencies, also outlined above, are met.

Figure 5 shows the processing of two hypothetical FSDF source datasets (A & B) into an FSDF product. Part A shows simple dataset-level provenance, part B shows dataset-level provenance but with more details PROV-O formulation, as per Figure 3A. 5C implements many of the techniques described above, specifically:

- The whole of 5C shows detail insertion (Figure 2D);
- The path from Source Dataset A to Intermediate X shows detail addition (3A) and either 4B or 4C formulation, depending on whether feature-action + action-feature mapping (4B) or feature-action-feature mapping (4C) is used;
- The Intermediate X to Intermediate Y path shows typed prov:Activity formulation (3B) and could use annotated output dataset (4D) mapping;
- Intermediate Y plus Source Dataset B fusing to form the FSDF product could be a 3C-type exercise where the types prov:Activity, "Merging" specifies two input datasets and an feature mapping prov:Plan which preserves feature origin knowledge. This formulation is also a feature-action-feature mapping (4C).

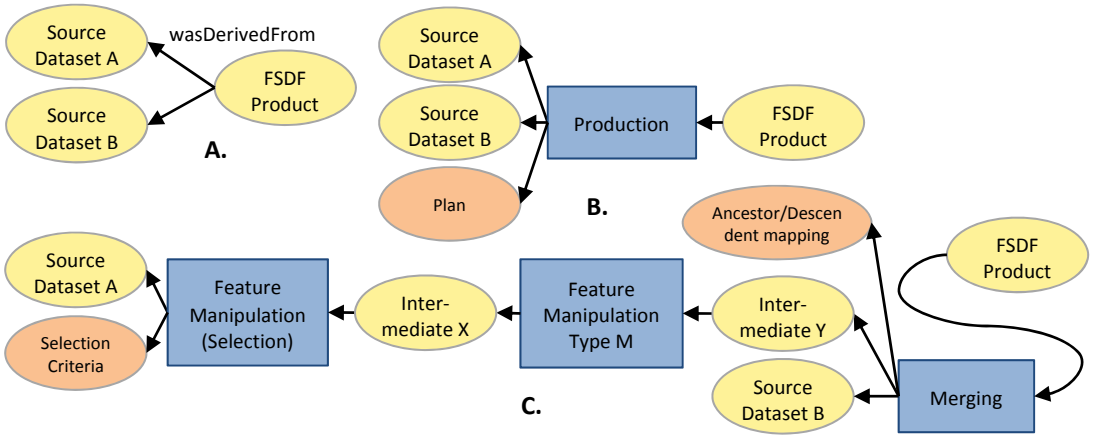

Figure 5. A hypothetical FSDF product generation scenario modelled with different amounts of detail and at different levels of granularity.

[8] http://www.w3.org/TR/ld-bp/

PROVENANCE DATA MANAGEMENT

It's also not possible to write in generalities about provenance data collection or generation – in-depth knowledge of specific systems is required in order to make sensible descriptions – and collecting provenance data in standardised formats is far harder than managing and storing it [5, see Discussion]. Once collected however, there are a range of generic tools available to manage and manipulate it. The PROMS family of tools and their associated methodology [6][9] allow any number of systems to report PROV-O-based provenance information and have it stored in a graph database. The system will automatically join provenance graphs where the same node URIs are used, thus detail insertion, as per Figure 2, can easily be achieved. Similarly, the joining of small provenance graphs into larger super-graphs can be achieved which allows independent systems to assemble continuous graphs across their individual processes, as long as they can share dataset or feature identifiers in order to report against them. Most RDF-based graph database allow querying via SPARQL thus the abstraction of detailed graphs into simpler ones can take place when detail insertion or multi-process reporting has taken place. Installations of PROMS Server make the SPARQL endpoint of its underlying RDF graph database available for such use thus allowing fine to coarse granularity translation out of the box.

CONCLUSIONS

We have presented a range of PROV-O-based modelling formulations (ontology design patterns) to help provenance data managers meld provenance information at varying levels of granularity. We focused on dataset and feature level provenance, as these are the two obvious granularities for spatial data products, but the principles could apply to information at other granularities. We have presented alternative methods for the integration of provenance information of different granularities and pointed out some of the logical and system dependencies that certain patterns require. We have given a very brief FSDF case study implementing many of the techniques and also finally described several aspects of provenance data management referencing a particular tool.

ACKNOWLEDGEMENTS

This paper is published with the permission of the CEO, Geoscience Australia.

REFERENCES

[1] Moreau, L. & Missier, P. (eds.) PROV-DM: The PROV Data Model. W3C Recommendation 30 April 2013 W3C (2013). Online at http://www.w3.org/TR/prov-dm/. Accessed 2015-12-08.
[2] Car, N.J. Map data lineage: provenance concepts, tools and future shared infrastructure. Locate2015 Conference presentation (2015).
[3] Sadiq, M.A., West, G., Arnold, L., McMeekin, D.A. and Moncrieff, S. Spatial data supply chain provenance modelling for next generation spatial infrastructures using semantic web technologies. MODSIM2015, Gold Coast, Australia, 29th Nov – 4th Dec, 2015. (2015) Online at http://mssanz.org.au/modsim2015/.
[4] Robinson, I., Webber, J. & Eifrem, E. (2013) Graph Databases. O'Reilly Media. ISBN 978-1-4493-5626-2. Online at http://graphdatabases.com. Accessed 2015-12-11.
[5] C. Wise, N. J. Car, R. Fraser and G. Squire. Standard Provenance Reporting and Scientific Software Management in Virtual Laboratories. MODSIM2015, Gold Coast, Australia, 29th Nov – 4th Dec, 2015. (2015) Online at http://mssanz.org.au/modsim2015/.
[6] Nicholas J Car, Matt Stenson, Mick Hartcher, Simon Cox, Peter Fitch, and David Lemon. A provenance management methodology and example architecture for science projects containing heterogeneous automated and manual processes. In HIC 2014 – 11th International Conference on Hydroinformatics, page 8, New York, USA, 2014. International Water Association. URL http://academicworks.cuny.edu/cc_conf_hic/57/. Accessed 2015-12-11.

[9] See http://promsns.org for up-to-date information on the PROMS family of provenance tools