# Generating Narratives from Provenance Relationship Chains

Heather S. Packer
University of Southampton
B32 University Road
Southampton, Hampshire, UK
hp3@ecs.soton.ac.uk

Luc Moreau
University of Southampton
B32 University Road
Southampton, Hampshire, UK
L.Moreau@ecs.soton.ac.uk

## ABSTRACT

Provenance data is a rich data structured source that has a similar role to narratives, since they can both provide an account of connected events. Consuming PROV data can be hard for both technical and non-technical users, because of its potential scale and the complexity of the relationships captured. Explicitly, it can be hard for users to follow and understand the chain of relationships connecting elements together. In this paper, we present an approach that generates narratives explaining chains of relationships and describe its nature with examples from a Ride Share application.

## Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

## General Terms

Design,Human Factors, Theory

## Keywords

Narrative Generation, Provenance, Relationship Chains

## 1. INTRODUCTION

The provenance standard (PROV) [13] can be used to capture information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness. Provenance implies a partial ordering between events and the entities it describes. The PROV elements typically follow a link data approach, where the elements identifier is a fully formed resolvable URI. The level of detail that provenance can captured makes provenance data very rich, which can support accountability for both users and systems. However, the more detail included in the provenance data, the harder it is for both technical and non-technical users to read. This difficulty increases with the scale of the provenance documents and complexity of the relationships captured. While the relationships between elements are explicitly defined, it can be hard for users to follow and understand the chain of relationships connecting one element to another. The provenance standard recommends offers provenance notation (PROVN) for human consumption [13]. While PROVN is more readable than other languages used to express PROV, we posit that a narrative better serves this function because they are an effective way to communicate information to humans [9, 6, 10].

Our aim is to enable users to follow and understand the chain of relationships connecting two elements, through narrative. In our approach, we: First transform the provenance data into a weighted graph and use Dijkstra's algorithm to find the shortest path between two PROV elements. Using a weighted graph enables us to use different weightings for different relationships, and thus we can prioritise describing relationships in the narrative that are more informative. For example, a document is attributed to an agent, that document was generated by an activity, and that activity was associated with the agent. When the weightings all have a value of 1 the shortest path between the document and the agent is through the attribution relationship, when the attribution relationship has a value of 3 and the others have a value of 1 then the shortest path describes the document's generation and association relationships. The additional relationships in the latter example provides details about which activity generated the document. There is however a trade-off between using less and more descriptive paths to generate narratives, concretely the greater number of relationships the longer the narrative will be, which could be excessive especially with long relationship chains; Second, use the relationships that connect two PROV elements to construct a narrative using a sentence generator. The sentence generator randomly selects a string of words from a predefined list, to describe the elements and connecting relationships. In this paper, we use a provenance data recorded from a ride share application to provide examples of narratives generated with our approach.

We describe related work in Section 2, which introduces provenance and narrative summarisation. Following that, in Section 3 we detail the algorithm we use to identify the shortest path between nodes and how we generate sentences using the chain of relationships. We then describe in Section 4 the ride sharing scenario and provenance data, and discuss some queries and their resulting sentences. Finally, we make our conclusions and explore future work in Section 6.

# 2. BACKGROUND

## 2.1 Provenance

Provenance has varied emerging applications: it may be used to make social computations accountable and transparent [19, 15]; provenance can help determine whether data or users can be trusted [4]; and provenance can be used to ensure reproducibility [11] of computations.

PROV is a recent set of recommendations of the W3C for representing provenance on the web. PROV is a conceptual data model (PROV-DM [13]), which can be mapped and serialized to different technologies. There is an OWL2 ontology for PROV (PROV-O [7]), allowing mapping to RDF, an XML schema for provenance [3], and a textual representation for PROV (PROV-N [14]).

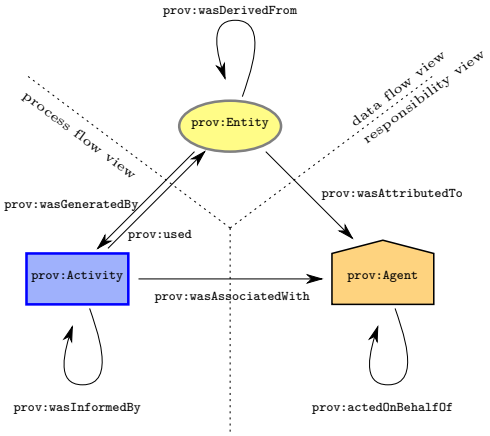


**Figure 1: Three Different Views of the Core of PROV. The figure adopts the PROV layout conventions: an entity is represented by a yellow ellipsis, an activity by a blue rectangle, and an agent by an orange pentagon. We note here that the diagram is a "class diagram" illustrating the classes that occur as domain and range of properties. Taken from [12].**

## 2.2 Narrative Summarisation

The role of narrative is to provide an account of connected events, which can be organised in a number of different categories. Provenance data has a similar role, it describes entities, activities and agents connected by relationships, and implicitly provides a sequence of their generation, performance and actions, respectively. Narratives are an effective way to communicate information, research has shown that it enables users to make sense of their data [9, 6, 10], and it is equally effective for both communities and individuals [1, 8].

Previously, Semantic Web technologies have been used to generate narratives [18, 5, 2]. In more detail, Tuffield et al. [18] and Jewell et al. [5] describe the OntoMedia ontology, which supports the generation of narratives. The work presented in [18] discuss approaches to generate narratives from a vocabulary, the approaches included are based on character, plot and user modelling. The work presented in [5] describes how OntoMedia is used to annotate the vast collection of heterogeneous media. The work [2] use ontological domain knowledge to select and organise a narrative discourse on an interest topic to a user.

# 3. GENERATING EXPLANATIONS ABOUT CONNECTIONS

In order to generate an explanation about how two PROV elements are connected, we first identity the chain of relationships connecting them. We then generate a narrative explaining the chained relationships.

## 3.1 Identifying Connections between Elements and their Attributes

In order to identify whether a element has a connection to another element, we identify whether there is a chain of relationship that connects them. We first use a fragmentation algorithm, which generates a subset of PROV statements describing an element from a PROV document. The fragmentation algorithm we use is based on [16]'s basic fragmentation algorithm for ontologies to generate a fragment of a concept. This fragmentation approach allows large PROV documents to be queried without suffering from overhead costs associated with scale.

Second, we verify that this fragment references to two elements. Third, we use the Dijkstra's algorithm [17] to find the shortest path between two PROV elements, where for now all relationships connecting the nodes have a value of 1. Then using the shortest path we build a list of dictionaries, where each dictionary contains the name of the two connecting nodes, their relationship type, and the connecting relationship. For example, the following graph contains the triples: *a wasDerivedFrom b, b wasDerivedFrom d, a wasDerivedFrom d,* and *d wasDerivedFrom e.* Our approach created the following list describing how element *a* is connected to *e*:

```
[{
    'subject': 'a'
    'subject_type': 'entity',
    'relationship': 'wasDerivedFrom',
    'object': 'd',
    'object_type': 'entity',
},
{
    'object': 'd',
    'object_type': 'entity',
    'relationship': 'wasDerivedFrom',
    'subject': 'e'
     'subject_type': 'entity',
}]
```

The pseudocode for this algorithm is described in Algorithm 1.

## 3.2 Generating Sentences Describing Connected Elements

The narrative we generate is an account of how two elements connect, it has a linear structure of sentences based on the implied order of events recorded in the provenance. In order to generate sentences, we use the list generated by the chaining algorithm (described in the section above). Each

**Algorithm 1** This algorithm finds a path connecting the PROV elements subject_element and goal_element in a graph, within a max_distance, it returns a list of relationships connecting the two elements. Where the functions retrieveFragment returns a set of provenance statements, buildGraph returns a weighted graph given provenance data, and dijkstra returns a list of the shortest path between two elements.

```
1: procedure GetConnectingRelationships(graph, subject_element, goal_element, max_distance)
2:     prov_fragment = retrieveFragment(subject_element, max_distance)
3:     if prov_fragment.contains(subject_element, goal_element) then
4:         graph = buildGraph(prov_segment)
5:         shortest_path = dijkstra(graph, subject_element, goal_element)
6:         return shortest_path
7:     else
8:         return None
```

item in the list will be used to generate a single sentence, and each sentence is constructed in three parts. The first part of the sentence describes the subject, the second describes the relationship, and the third describes the object. For example, 'The *a* entity' 'was derived from' 'the entity *e*'. Each part of the sentence is randomly selected from a list of possible parts given their type or relationship, these are described in Tables 1, 2 and 3.

| Type | Sentence |
|---|---|
| agent | 'the {subject}'<br>'the {subject} is a {subject_type}<br>'the agent {subject}' |
| entity | 'the {subject}'<br>'the {subject} is a {subject_type}, which'<br>'the entity {subject}'<br>'the contents contained in {subject}'<br>'the contents of {subject}' |
| activity | 'the {subject}'<br>'the {subject} is a {subject_type}, which'<br>'the {subject} process'<br>'the activity {subject}' |

**Table 1: Sentence parts for subject types**

# 4. RIDE SHARE

Ride Share is an application that enables car sharing for workplace workers, university students and similar large communities. The application allows both drivers and commuters to offer and request rides. These offers and ride requests include details about required travels, timing, locations, capacity, prices, and other details relevant for car sharing. It performs automatic matching of commuters to available cars, by considering origin and destination, routes, capacity and other available information. Incentives are used to influence participant behaviours and maximise the global system goals. Our ultimate motivation in this scenario is to describe how matches for rides and reputation reports are generated, because these types of processes are typically a black box to users. The execution of these processes is documented using PROV.

| Relationship | Sentence |
|---|---|
| wasDerivedFrom | 'was derived from'<br>'originates from'<br>'was sourced from' |
| wasAssociatedWith | 'was associated with'<br>'was performed by' |
| used | 'used'<br>'made use of' |
| wasAttributedTo | 'was attributed to'<br>'is connected to' |
| specializationOf | 'is a specialization of '<br>'is an instance of' |
| wasInformedBy | 'was informed by'<br>'was initiated because of'<br>'was performed after' |
| wasGeneratedBy | 'was generated by'<br>'was created by' |

**Table 2: Sentence parts for relationships**

| Type | Sentence |
|---|---|
| agent , entity, and activity | 'the {subject}'<br>'the {subject} {subject_type}' |

**Table 3: Sentence parts for object types**

The following two sections provide examples of sentences generated using our approach over the PROV data visualised as a graph in Figure 2. The provenance data in this figure was recorded by the Ride Share's UI, and shows a user logging in looking at their homepage, their ride offers, profile, reputation and submitting a ride request.

## 4.1 Ride Sharing Query 1

The first query we performed generates a narrative that connects a UI's login page to a ride plan, 'log:8069ee6a051f9ac83b2c647beddd723d' and 'rplan:85', respectively. The following structure was created by the finding the shortest path between the two entities (see Section 3.1).

```
[
 {'object_type': 'request',
 'subject_type': 'response',
 'relationship': 'wasDerivedFrom',
 'object: 'uuid:1baf0146-1711-4c56-8589-059ab29fcfca',
 'subject': 'rplan:85'}
 {'object_type': 'request',
 'subject_type': 'my_offer_page',
 'relationship': 'wasDerivedFrom',
 'object: 'moff:86b9df09090720c10279567105a499d2',
 'subject': 'uuid:1baf0146-1711-4c56-8589-059ab29fcfca'},
 {'object_type': 'my_offer_page',
 'subject_type': 'ride_requests',
 'relationship': 'wasDerivedFrom',
 'object: 'rreq:ima1',
 'subject': 'moff:86b9df09090720c10279567105a499d2'},
 {'object_type': 'response',
 'subject_type': 'home_page',
 'relationship': 'wasDerivedFrom',
 'object: 'uuid:55cf0887-77f7-46ed-9cc3-809fc701574e',
 'subject': 'rreq:ima1'},
 {'object_type': 'response',
 'subject_type': 'home_page',
```

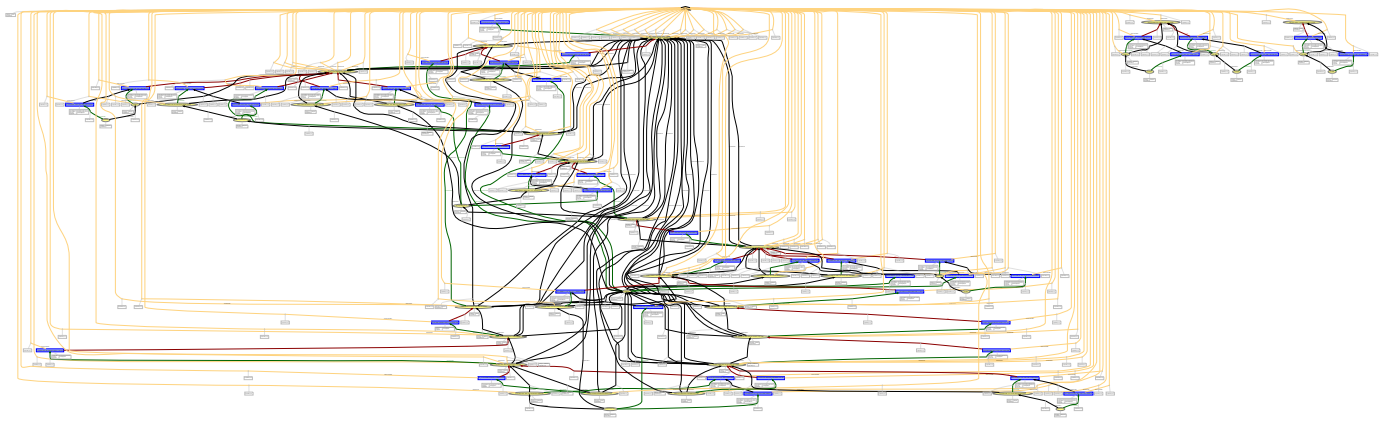

**Figure 2: This figure is used for illustrative purposely only, for scalable detail please use this link: `https://provenance.ecs.soton.ac.uk/store/documents/84089/`**

```
  'relationship': 'wasDerivedFrom',
  'object: 'home:b51fa37a247bbdd13202103b87381b16',
  'subject': 'uuid:55cf0887-77f7-46ed-9cc3-809fc701574e'},
  {'object_type': 'home_page',
  'subject_type': 'login_page',
  'relationship': 'wasDerivedFrom',
  'object: 'log:8069ee6a051f9ac83b2c647beddd723d',
  'subject': 'home:b51fa37a247bbdd13202103b87381b16'},
]
```

The above structure is used to create the following sentence, which are generating using the approach described in Section 3.2:

*The contents contained in rplan:85 was derived from the uuid:1baf0146-1711-4c56-8589-059ab29fcfca. The contents of uuid:1baf0146-1711-4c56-8589-059ab29fcfca was sourced from the moff:86b9df09090720c10279567105a499d2. The moff:86b9df09090720c10279567105a499d2 is a my_offer_page, and was derived from the rreq:ima1 request. The contents of rreq:ima1 was derived from uuid:55cf0887-77f7-46ed-9cc3-809fc701574e request. The uuid:55cf0887-77f7-46ed-9cc3-809fc701574e was derived from the home:b51fa37a247bbdd13202103b87381b16. The home_page home:b51fa37a247bbdd13202103b87381b16 orginates from the log:8069ee6a051f9ac83b2c647beddd723d login_page.*

## 4.2 Ride Sharing Query 2

The second query generates a narrative that connects a UI's new offers page and a profile page, 'noff:f9de6bdea5b4c849b884aa44ef6b8700', and 'prof:b7f89504e98004d9b60d8785e53c468b', and produces the following:

*The noff:f9de6bdea5b4c849b884aa44ef6b8700 new_offer_page is a specialization of the log:5e0454ca474ccdadc8d2580f58b26581 log_offer_page. The log:5e0454ca474ccdadc8d2580f58b26581 was derived from the prof:b7f89504e98004d9b60d8785e53c468b profile_page.*

# 5. DISCUSSION

While our approach generates a linear narrative describing the connection between two PROV elements, however lacks a natural flow present in discourse. This lack of flow stems from the repetitive use of connecting sentence phrases, and long identifiers used for PROV elements. In order to improve the narrative, we have identified key three areas:

1. Summarisation: The narrative presented in Section 4.1 illustrates the repetitive sentence structure for PROV elements that are connected by the same relationship. Chains of elements that are connected via the same relation can be shortened using lists. For example:

   *The contents contained in rplan:85 was derived from the following entities: the request uuid:1baf0146-1711-4c56-8589-059ab29fcfca, the my_offer_page moff:86b9df09090720c10279567105a499d2, the response rreq:ima1, the request uuid:55cf0887-77f7-46ed-9cc3-809fc701574e, the home_page home:b51fa37a247bbdd13202103b87381b16, and the log:8069ee6a051f9ac83b2c647beddd723d login_page.*

   While this shortens the narrative, it does not improve its readability or the reader's interest in the subject because there are no descriptions of the identifiers which are non-descriptive. This issue is more prevalent with longer chains of elements connected by the same relationship.

2. Element Path: The shortest path PROV elements might not provide the user with most informative narrative. For example, PROV uses a single specialisation relationship to connect two elements, however a longer path could explain who was responsible for its creation or which activities used it. This is highlighted in the narrative presented in Section 4.2.

3. Narrative Generation: The vocabulary used to generate the narratives is limited, using a different approach might be able to expand the vocabulary. Increasing the vocabulary used to describe elements could improve the readability, however there is a trade-off between using a broad vocabulary where the narrative could

become less specific, and a narrow vocabulary where the narrative becomes repetitive.

# 6. CONCLUSION

In this paper, we present an approach to generate a narrative describing the connection between two PROV elements. It describes how two PROV elements are connected by identifying a chain of relationships connecting them. We then randomly selected terms to describe the subject, its relationship, the element to which it is connected, to form a simple sentence structure. We then discuss areas which could improve readability of the generated narrative.

For future work, we will develop an approach to summarise chains of elements connected by the same relationships to improve readability. We will also investigate whether using different vertex weightings for different relationships could output more interesting narratives. In order to evaluate this work we will deploy a user study which will allow users to query how their data is connected to other entities in the ride share application.

# 7. ACKNOWLEDGMENTS

The research leading to these results has received partially funding from the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement n. 600854 Smart Society: hybrid and diversity-aware collective adaptive systems: where people meet machines to build smarter societies `http://www.smart-society-project.eu/`.

# 8. REFERENCES

[1] O. Ferret, B. Grau, and N. Masson. Thematic segmentation of texts: two methods for two kinds of texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 392–396. Association for Computational Linguistics, 1998.

[2] J. Geurts, S. Bocconi, J. Van Ossenbruggen, and L. Hardman. *Towards ontology-driven discourse: From semantic graphs to multimedia presentations*. Springer, 2003.

[3] H. Hua, C. Tilmes, S. Zednik (eds.), and L. Moreau. PROV-XML: The PROV XML Schema. W3C Working Group Note NOTE-prov-xml-20130430, World Wide Web Consortium, Apr. 2013.

[4] T. D. Huynh. Trust and reputation in open multi-agent systems. June 2006.

[5] M. O. Jewell, K. F. Lawrence, M. M. Tuffield, A. Prugel-Bennett, D. E. Millard, M. S. Nixon, N. R. Shadbolt, et al. Ontomedia: An ontology for the representation of heterogeneous media. In *In Proceeding of SIGIR workshop on Mutlimedia Information Retrieval*. ACM SIGIR, 2005.

[6] A. Kuchinsky, C. Pering, M. L. Creech, D. Freeze, B. Serra, and J. Gwizdka. Fotofile: a consumer multimedia organization and retrieval system. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 496–503. ACM, 1999.

[7] T. Lebo, S. Sahoo, D. McGuinness (eds.), K. Behajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao. PROV-O: The PROV Ontology. W3C Recommendation REC-prov-o-20130430, World Wide Web Consortium, Oct. 2013.

[8] C. Lévi-Strauss. *Structural anthropology*. Basic Books, 2008.

[9] O. Mallett and R. Wapshott. The challenges of identity work: Developing ricoeurian narrative identity in organisations. *ephemera*, page 271, 2011.

[10] Y. Matsuo and M. Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(01):157–169, 2004.

[11] L. Moreau. Provenance-based reproducibility in the semantic web. *Web Semantics: Science Services and Agents on the World Wide Web*, 9(2):202–221, July 2011.

[12] L. Moreau and P. Groth. *Provenance: An Introduction to PROV*. Morgan and Claypool, September 2013.

[13] L. Moreau, P. Missier (eds.), K. Belhajjame, R. B'Far, J. Cheney, S. Coppens, S. Cresswell, Y. Gil, P. Groth, G. Klyne, T. Lebo, J. McCusker, S. Miles, J. Myers, S. Sahoo, and C. Tilmes. PROV-DM: The PROV Data Model. W3C Recommendation REC-prov-dm-20130430, World Wide Web Consortium, Oct. 2013.

[14] L. Moreau, P. Missier (eds.), J. Cheney, and S. Soiland-Reyes. PROV-N: The Provenance Notation. W3C Recommendation REC-prov-n-20130430, World Wide Web Consortium, Oct. 2013.

[15] H. S. Packer, L. Drăgan, and L. Moreau. An auditable reputation service for collective adaptive systems. In *Social Collective Intelligence*, pages 159–184. Springer, 2014.

[16] J. Seidenberg and A. Rector. Web ontology segmentation: analysis, classification and use. In *Proceedings of the 15th international conference on World Wide Web*, pages 13–22. ACM, 2006.

[17] S. Skiena. Dijkstra's algorithm. *Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica, Reading, MA: Addison-Wesley*, pages 225–227, 1990.

[18] M. M. Tuffield, D. E. Millard, and N. R. Shadbolt. Ontological approaches to modelling narrative. 2006.

[19] D. J. Weitzner, H. Abelson, T. Berners-Lee, J. Feigenbaum, J. Hendler, and G. J. Sussman. Information accountability. *Communications of the ACM*, 51(6):82–87, 2008.