# Influence Factor: Extending the PROV Model With a Quantitative Measure of Influence

Matthew Gamble    Carole Goble

University of Manchester

*first.last*@cs.manchester.ac.uk

## Abstract

A central tenet of provenance is to support the assessment of the quality, reliability, or trustworthiness of data. The World Wide Web Consortium's (W3C) PROV provenance data model shares this goal, and provides a domain-agnostic interchange language for provenance representation. In this paper we suggest that given the PROV model as it stands, there are cases where information relating to how one entity has influenced another falls short of that required to make these assessments. In light of this, we propose a simple extension to the model to capture a quantitative measure of influence.

To understand how provenance publishers use PROV to describe influence we have consulted the current Provbench datasets and evaluated the usage of the 13 sub properties of `wasInfluencedBy`. The findings suggest that publishers are willing to provide additional information about how an influencer affected an influencee beyond a simple wasInfluencedBy relation.

In the paper, we define influence factor as a quantitative measure of influence that one PROV entity, agent, or activity has had over another and introduce `influenceFactor` as property to enrich any qualified influence in the PROV model.

To demonstrate the use of the use of `influenceFactor` we have extended the Wikipedia-provenance dataset and tooling from ProvBench to capture a quantitative measure of influence between the provenance elements involved. We also briefly discuss how we have used the proposed influence factor to support the development of a probabilistic approach to information quality (IQ) assessment using Bayesian Networks.

***Keywords*** Provenance, PROV, Influence, Influence Factor, Quality, Trust

## 1. Introduction

In a distributed environment such as the Web of Data, provenance information is an important component of IQ assessment. As such the evaluation of quality, reliability and trustworthiness has been a primary use-case for the development of the PROV specification [6].



*TAPP '14*,  June 12–13, 2014, Cologne, DLR, Germany.

http://dx.doi.org/10.1145/

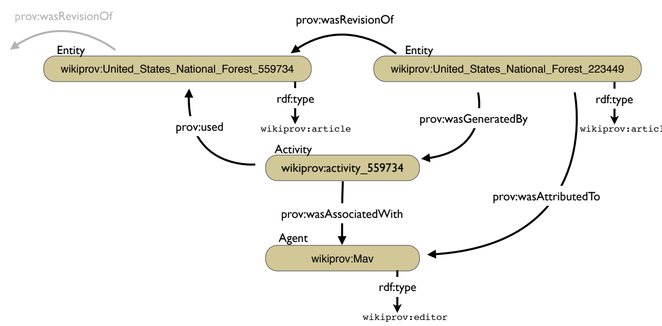


**Figure 1.** Example of a Provenance Graph from ProvBench wikipedia-provenance using sub-properties of wasInfluencedBy

The motivation to use provenance data for IQ assessment is that we do not always have quality related metadata, or a quality metric, that is suitable to directly assess the *intrinsic* quality of a particular Web resource. Previous work has demonstrated that metrics that make use of provenance data can increase the number of Web resources we can assess by considering the quality and trustworthiness of related resources in a provenance graph [9] and how their quality relates to the resource in question. Furthermore, a combination of both these *provenance-based* metrics and intrinsic metrics has been shown to improve performance beyond the application of either type of metric in isolation [2].

Specifically, we wish to exploit a common intuition that any resource that has influenced the production of another Web resource may have affected its likely quality. Therefore if we can evaluate the quality of these influencing resources, we can use that information to inform us of the *likely* quality of the resources that they have influenced.

This intuition that makes use of two types of provenance

- *lineage* provenance that describes the lineage of the web resource i.e. the other resources that were involved in its production.
- *how* provenance that describes to what extend those other resources contributed to its production.

Consider as an example a provenance graph for two revisions of the same article in Wikipedia (shown in Figure 1). We might have an expert review of the older article revision `wikiprov:United_States_National_Forest_559734`, providing us with an intrinsic measure of quality (as is the case with pages maintained by expert groups such as WikiProject Chemicals[1]). We might also have

[1] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Chemicals

an intrinsic measure of trust in the editor based upon their authorship status e.g. Administrator, Registered, Anonymous, or Blocked.

We can use the lineage provenance and intrinsic measures of quality and trustworthiness to make an estimate as to the likely quality of the new article revision. It is clear however that to make this assessment we require information detailing *how* the editor has influenced the new revision, and how much of the previous revision remains.

Practically, we must also understand whether publishers of provenance information are motivated to publish additional provenance relating to how elements influenced each-other. To estimate this we have consulted the ProvBench provenance datasets [1]. There are currently 9 collections of provenance data available describing provenance in a range of domains including scientific Workflow systems, simulation experiments, and Web-based resources such as Wikipedia. A number of the ProvBench submissions also provide tooling to support the user in generating additional provenance data. In this paper we make particular use of the **wikipedia-provenance** dataset and tooling [8].

How provenance is captured in the PROV model in two ways: 1) using sub types of `wasInfluencedBy` and 2) qualified influences. PROV-DM provides 13 sub properties of `wasInfluencedBy` to better describe *how* the influencer influenced the influencee. Figure 1 illustrates the modelling the wikipedia-provenance, using the available sub properties of `wasInfluencedBy` to indicate the type of influence that each Web resource had.

For each dataset we have summarised the usage of the 13 sub properties of `wasInfluencedBy` based upon an analysis of the datasets and information from their supporting publications[2]. The findings in Table 1 suggest that publishers of provenance information are willing to provide information beyond a simple `wasInfluencedBy` relation and describe *how* the influencer affected the influencee. Indeed whilst there are currently PROV features that are not used in the ProvBench data, it is still the case that *all* datasets make use of between 3 and 7 of the more specific influence properties of PROV.

In addition to sub properties of `wasInfluencedBy`, PROV also provides *qualified influences*. Qualified influences use an N-ary relation to provide more detailed descriptions for influence relations. Each influence type has a corresponding qualified influence in PROV, and the model provides the properties `atTime`, `hadRole` and `hadPlan` to enrich qualified influences. Despite these qualified influences and additional properties we are still missing metadata that describes *to what extent* the influencer contributed to the influencee.

A PROV `Plan` provides detailed and specific information about a `qualifiedAssociation`, and might therefore provide this quantitative information. However for our purposes they have number of limitations. Firstly, they are restricted by the PROV specification to *only* be used with `qualifiedAssociations`, we instead want to be able to quantify *any* influencing relationship. Secondly, PROV Plans are not restricted by the PROV specification in their representation. As a result, whilst they may provide the quantitative information we desire, they might not be in a known representation, or even a machine readable representation. Therefore, we see the need for a vocabulary feature to enrich any qualified influence and provide the quantitative *how* provenance.

## 2. Influence Factor

We define influence factor as a *quantitative measure* of the influence that one PROV entity, agent, or activity has had over another. This information can be used to subsequently determine the quality or trustworthiness of a PROV element in terms of its influencers. This degree of influence is currently suggested with certain properties of the PROV vocabulary such as `wasQuotedFrom`, `wasGeneratedBy` and `hadPrimarySource`. With influence factor we are making this explicit.

For example, if an Activity generated an Entity, declared by `wasGeneratedBy`, and it is the only influencer described, then we might make the assumption that it had exclusive influence. For our two revisions of the Wikipedia article, the `prov:wasRevisionOf` relation between the two entities and the `qualifiedRevision` description falls short of fully describing the relationship between the two, specifically how much of the previous revision has remained. This is similarly the case for the `qualifiedAttribution`. If the author is not considered trustworthy then our belief in the likely quality of the resulting revision will differ depending on, for example, whether they have modified the whole page, or just contributed to a small part of it. In many cases in the production of data, it is possible to quantify this degree of influence. A mechanism for quantifying the difference between two revisions of an article in Wikipedia is a 'diff' between the two. A quantitative measure of this diff can be easily included as additional metadata.

Influence factor might not just reflect a physical attributes such as a diff. We might believe that some parts of a Wikipedia article are more important that others, for example Infobox data, or references section. Instead of an influence measure based on the size of contribution, we might weight it by where in the article that contribution is made.

A further example comes from the scholarly communications domain. When describing the creation of a scholarly artifact we can describe and attribute that creation to one or more creators indicating their role such as lead author, contributor, supervisor. Vocabularies for scholarly communications such as the Semantic Publishing and Referencing Ontologies (SPAR) suite of ontologies [3] capture this type of contribution description in the Publishing Roles (PRO) ontology, using classes such as editor, contributor, copy-editor etc. Whilst these categories of contribution are not numerical, they provide a spectrum of influence to which we can apply our own consistent weighting. Given these observations we believe that a mechanism for describing a *degree of influence* would increase the ability of the PROV vocabulary in its stated purpose to support the assessment of the quality, reliability and trustworthiness of data.

### 2.1 Modelling evident:influenceFactor

To capture the degree of influence one entity has on another, we have introduced `evident:influencefactor` as an additional property in our evident namespace[4] as an attribute for any of the PROV qualified influences. The property allows the provision of additional information quantifying the degree to which the influencing class has influenced the influenced class.

We have extended the `evident:influenceFactor` property to model two core types of influence factor, `evident:discreteInfluenceFactor` and `evident:continuousInfluenceFactor`. `evident:discreteInfluenceFactor` can be extended to model discrete states of influence such as those from the SPAR vocabularies.

Using `evident:continuousInfluenceFactor` we can model influence factor using a continuous numerical value. A sub property of `evident:continuousInfluenceFactor` that we include as part of our extension is `evident:normalInfluenceFactor`. This property describes a degree of influence as a real number

[2] For datasets that were not provided explicitly in PROV we consulted information from their supporting publication only

[3] http://opencitations.wordpress.com/2010/10/14/introducing -the-semantic-publishingand-referencing-spar-ontologies

[4] http://purl.org/net/evident

| Influencee | Property | Influencer | Taverna | Wings | Wiki | SRI | OBAMIO | CSIRO | Vis | Chiron | Swift |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Datasets | | | | | | | | |
| * | wasInfluencedBy | * | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Entity | wasGenerateBy | Activity | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Entity | wasDerivedFrom | Entity | ✓ | × | × | ✓ | ✓ | ✓ | × | ✓ | ✓ |
| Entity | wasAttributedTo | Agent | × | ✓ | × | × | ✓ | ✓ | × | ✓ | × |
| Entity | hadPrimarySource | Entity | × | ✓ | × | × | × | × | × | × | × |
| Entity | wasQuotedFrom | Entity | × | × | × | × | × | × | × | × | × |
| Entity | wasRevisionOf | Entity | × | × | ✓ | × | × | × | × | × | × |
| Entity | wasInvalidatedBy | Activity | × | × | × | × | × | × | × | × | × |
| Activity | wasInformedBy | Activity | ✓ | × | × | ✓ | ✓ | ✓ | × | ✓ | × |
| Activity | used | Entity | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Activity | wasAssociatedWith | Agent | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | × |
| Activity | wasStartedBy | Entity | × | × | × | × | × | × | × | × | × |
| Activity | wasEndedBy | Entity | × | × | × | × | × | × | × | × | × |
| Agent | actedOnBehalfOf | Agent | × | × | × | × | × | ✓ | × | × | ✓ |

**Table 1.** Summary of PROV Influence Type Usage in the ProvBench Datasets

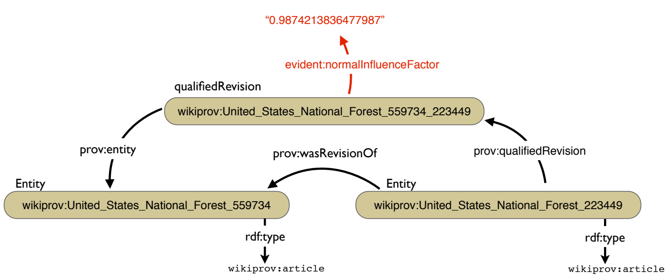


**Figure 2.** Using evident:influenceFactor in Wikipedia Provenance

on a scale $[0..1]$. Figure 2 illustrates the use of the `evident:normalInfluenceFactor` for the ProvBench Wikipedia revisions data to enrich a qualified revision. To quantify the influence the influencer `wikiprov:United_States_National_Forest_559734` has had on the influencee `wikiprov:United_States_National_Forest_223449`, we add the influence factor to the qualified revision description between the two.

We define two terms relating to influences that have been enriched with an influence factor.

- ***quantified influence***: To distinguish between a qualified influence that has been enriched with an influence factor and one that has not, we refer to a qualified influence that has been enriched as a *quantified influence*.
- ***quantified path***: We refer to any transitive path of influences between two entities in a graph such that at least one of the influences is quantified as a *quantified path*.

In the case of `normalInfluenceFactor` one might expect that by modelling influence factor on a scale of $[0...1]$ we should modify the conditions for provenance validity such that the sum of all influence factors that directly influence a given element should sum to 1. However, we believe that in a distributed publishing environment such as the Web of Data such a restriction would be prohibitively difficult for a data publisher to comply with. Instead we leave it to the consumer of the provenance to evaluate, and if needed, normalize any influence factors for a given entity.

As with many modelling approaches, there is scope for human error when describing influence factor. One particular scenario we highlight is a case we define as *overstating influence*. This refers to a modeller attributing the same conceptual influence from one entity, agent or activity to more than one qualified influence. Consider the Wikipedia revision in Figure 3. To quantitatively capture the influence that the author agent had in the revision, the modeller has to decide where to describe the influence factor. The modeller

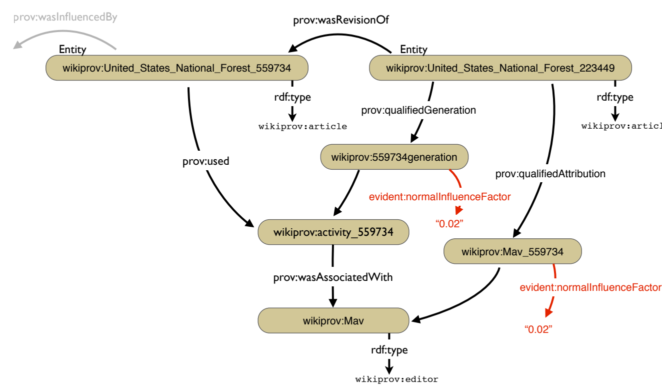


**Figure 3.** Overstating evident:normalInfluenceFactor in Wikipedia Provenance

could quantify either the qualified attribution between the author and revision entity, or the qualified generation between the activity and revision entity. The modelling approach shown in figure 3 would constitute overstating influence, where the same conceptual contribution to the revision is duplicated by quantifying both the qualified generation and qualified attribution. The overstating of influence is difficult to account for retrospectively by a data consumer because its occurrence is ambiguous. Given the example we would not know for example if the two quantified paths between author and entity captured the same conceptual influence, or two unrelated types of influence. This challenge of understanding the quality of provenance information is not however restricted to influence factor, but one that is relevant to all provenance metadata in general [3].

### 2.2 Extending wikipedia-provenance with :influenceFactor

To create PROV versions of the Wikipedia articles we have used the wikipedia-provenance tool[5] used to generate the existing ProvBench wikipedia-provenance data, and extended the tool[6] in two ways. Firstly we have extended the tool to generate RDF serialisations of the PROV data. These serialisations are constructed using the PROV Toolbox[7], a Java-based toolset provided and maintained by the Provenance community. We have also extended the tool to introduce a number of additional elements of metadata, including influence factor.

[5] https://github.com/PaoloMissier/wikipedia-provenance

[6] https://github.com/matthewgamble/wikipedia-provenance

[7] https://github.com/lucmoreau/ProvToolbox/

Influence factor is included in our data for two qualified influences using the property `evident:normalInfluenceFactor`:

- As part of the `qualifiedAttribution` between the author and the revision.
- As part of the `qualifiedRevision` between the current and previous article revision.

We have based the influence factor on the number of words contributed to a revision. To calculate it we use a popular open source word-based diff tool `wdiff`[8]. The tool calculates three values where comparing two revisions: the number of words in *common*, *changed*, and *inserted*. To represent the authors influence as a single quantified value we calculate the ratio between the number of words in common ($common$) with the total number of words in the previous revision ($previous$). The influence factors are calculated as follows:

- The influence factor for the author on the qualifiedAttribution is $1 - (common/previous)$.
- The influence factor for the qualifiedRevision is calculated as $(common/previous)$.

The listing below illustrates the inclusion of an influence factor for a qualified attribution:

```
wikiprov:Mav_559734 a prov:Attribution ;
        prov:agent wikiprov:Mav ,
        evident:normalInfluenceFactor "
            0.012578616352201255"^^xsd:double .
```

### 2.3 Using Influence Factor to calculate IQ

We have used this extended wikipedia-provenance dataset enriched with influence factor metadata as part of a broader investigation in to provenance-based IQ assessment [5]. Specifically we have used the data set to support the development of a procedure that can automatically generate Bayesian Networks suitable for IQ assessment. Our procedure makes use of three types of information: PROV provenance graphs, intrinsic quality measures, and influence factor annotations, to build Bayesian Networks. Influence factor is central to supporting the resulting IQ assessment and we have shown that we can successfully approximate the results of an existing metric for Wikipedia articles, and predict the a likely quality class for featured or cleanup articles.

As part of this work we have also begun to explore strategies to manage and normalize continuous influence factor for a number of different scenarios where influence factor has, for example, been overstated or omitted.

### 2.4 Related Work

We are not the first to recognize the need to annotate provenance with further quantitative information to support the computational tractability of quality assessment. Hartig et al. [7] proposes a type of annotation called *impact values* in their work using provenance to assess the timeliness of Web Data. In contrast to our influence factor, the authors use the term impact values to refer *any* type of metadata that informs a quality assessment. What is considered an impact value is therefore contextual, and tied to the particular type of quality assessment being performed. For example an impact value might be the *creation time* of a resource for a timeliness assessment, or a data creators *credibility* for a believability assessment. Our influence factor is instead a general mechanism for capturing a quantitative influence from one resource to another. We therefore see influence factor and impact values being complimentary, where influence factor can characterized as a certain class of impact value, depending on the quality assessment.

Dai et al [4] propose a general provenance-based approach to assess the quality of data in a distributed data system that takes into account the trustworthiness of data sources and intermediate agents. The authors define two types of interaction that an intermediate agent can have with data, PASS or INFER. PASS indicates that an agent simply passed the data on, INFER indicates that the agent inferred new knowledge from some input data. These actions impact resulting trustworthiness score differently and can be seen to be types of discrete influence factor.

### 2.5 Discussion

Provenance plays a key role in the assessment of information quality. In this paper we have motivated the need for a quantitative measure of influence that extends the scope of *how* provenance that can be recorded with the PROV model to support these assessments. We have demonstrated a practical approach to achieving this within the PROV model using `influenceFactor`, a property that can be used to enrich any PROV qualified influence. In particular we have demonstrated the use of `normalInfluenceFactor`, introducing it into the wikipedia-provenance ProvBench data.

We are interested in developing the concept and application of influence factor further, and identifying common classes of influence factor. A quantified diff for example could also be used in other scenarios that involve textual information, including this years Provenance Reconstruction challenge for version-controlled source code. We also recognise that there are cases where it is less obvious whether influence factor is readily quantifiable. In the domain of scientific workflows it is less clear how to quantify the influence that an input to a service has had on an output, which may depend upon additional domain expertise, as well as the level of granularity at which the provenance metadata in being captured.

In summary we present influence factor as a step towards a practical mechanism for broadening the scope of the PROV model for quality and trustworthiness assessment.

## References


[1] K. Belhajjame, J. Zhao, J. Manuel Gomez-Perez, and S. Sahoo, editors. *Provbench Workshop*, 2013. ACM, Proceedings of the Joint EDBT/ICDT 2013 Workshops.

[2] D. Ceolin, P. T. Groth, W. R. van Hage, A. Nottamkandath, and W. Fokkink. Trust Evaluation Through User Reputation and Provenance Analysis. In *URSW*, pages 15–26, 2012.

[3] Y.-W. Cheah and B. Plale. Provenance Analysis: towards Quality Provenance. In *E-Science (e-Science), 2012 IEEE 8th International Conference on*, pages 1–8. IEEE, 2012.

[4] C. Dai, D. Lin, E. Bertino, and M. Kantarcioglu. An Approach to Evaluate Data Trustworthiness ased on Data Provenance. In *Secure Data Management*, pages 82–98. Springer, 2008.

[5] M. Gamble. *Modelling and Computing the Quality of Scientific Information on the Web of Data.* PhD thesis, University of Manchester, 2014.

[6] P. Groth and L. Moreau. An Overview of The Prov Family of Documents. `http://www.w3.org/TR/prov-overview/`, April 2013. [accessed 20/09/2013].

[7] O. Hartig and J. Zhao. Using Web Data Provenance for Quality Assessment. In *Proceedings of the First International Workshop on the role of Semantic Web in Provenance Management at ISWC2009*, Washington D.C., 2009. URL `http://ceur-ws.org/Vol-526/paper_1.pdf`.

[8] P. Missier and Z. Chen. Extracting PROV Provenance Traces from Wikipedia History Pages. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, pages 327–330. ACM, 2013.


[8] Gnu Wdiff: http://www.gnu.org/software/wdiff/

[9] I. Zaihrayeu, P. P. Da Silva, and D. L. McGuinness. IWTrust: Improving User Trust In Answers From The Web. In *Trust Management*, pages 384–392. Springer, 2005.