

# IJDC | *Peer-Reviewed Paper*

## Data Trajectories: Tracking Reuse of Published Data for Transitive Credit Attribution

Paolo Missier  
School of Computing Science  
Newcastle University, UK

### Abstract

The ability to measure the use and impact of published data sets is key to the success of the open data/open science paradigm. A direct measure of impact would require tracking data (re)use in the wild, which is difficult to achieve. This is therefore commonly replaced by simpler metrics based on data download and citation counts. In this paper we describe a scenario where it is possible to track the *trajectory* of a dataset *after* its publication, and show how this enables the design of accurate models for ascribing credit to data originators. A Data Trajectory (DT) is a graph that encodes knowledge of how, by whom, and in which context data has been re-used, possibly after several generations. We provide a theoretical model of DTs that is grounded in the W3C PROV data model for provenance, and we show how DTs can be used to automatically propagate a fraction of the credit associated with transitively derived datasets, back to original data contributors. We also show this model of *transitive credit* in action by means of a Data Reuse Simulator. In the longer term, our ultimate hope is that credit models based on direct measures of data reuse will provide further incentives to data publication. We conclude by outlining a research agenda to address the hard questions of creating, collecting, and using DTs systematically across a large number of data reuse instances in the wild.

*Received 29 June 2016 | Revision received 20 September 2016 | Accepted 21 September 2016*

Correspondence should be addressed to Paolo Missier, Claremont Road, Newcastle upon Tyne. Email: [paolo.missier@ncl.ac.uk](mailto:paolo.missier@ncl.ac.uk)

An earlier version of this paper was presented at the 11th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: <http://www.ijdc.net/>

Copyright rests with the authors. This work is released under a Creative Commons Attribution 4.0 International Licence. For details please see <http://creativecommons.org/licenses/by/4.0/>



## Introduction

The practice of publishing research data has been maturing rapidly, following increasing evidence that the combination of data sharing and emerging data citation practices represent new opportunities for extending the value chain of the data, rather than a threat to its owners (Piwowar & Vision, 2013). Reasons for publishing data, and scientific datasets in particular, include facilitating its re-use and enabling its validation. A plethora of data repositories are available where scientists can publish their datasets, assign a persistent identifier to them, and make them discoverable. Much less is known about the lifetime of those datasets *after* their publication, namely the knowledge of how, by whom, and in what contexts they have been re-used, and whether such instances of re-use have produced interesting derived data products, possibly after several generations. We refer to this new type of knowledge as the *trajectories* of published data (Data Trajectories, or DT). The main hypothesis that motivates our research is that knowledge of DTs makes it possible to quantify the impact and influence of research data through several generations of reuse and derivation, *transitively*. In turn, this will lead to new notions of *transitive credit* to data owners, which may inform and extend current data citation practices. We are aware of very few attempts at defining transitive credit in the context of data citation. Amongst these is (Katz, 2014), where the concept is not fully formalised nor made operational through metadata management and analysis.

### Challenges in Tracking Data Reuse and the Role of Data Citation

While counting data downloads from repositories is straightforward, tracking their usage in the wild is much more challenging. Data can be reused in endless ways through program logic, entirely or in part, on its own or combined with other datasets. Furthermore, such derivations can extend over several generations, and may take place on different, autonomous information systems and data processing environments.

(Robinson-García, Jiménez-Contreras & Torres-Salinas, 2015) describe data citation practices that go beyond simple download count as valid surrogates to direct tracking of data use. (Callaghan et al., 2012) recommend that data citation should be based on similar review stages as journal articles, as a necessary first step to treating data as a first class scientific object. However, recognising the complexity of data derivation, they also argue that further mechanisms are needed to facilitate data transparency and scrutiny. Even when data citation is primarily viewed as an extension of traditional article publication, tracking data citation requires different and more sophisticated processes than tracking data downloads (Mayernik, 2013).

Efforts in this direction include Thomson's data citation index<sup>1</sup>, as well as community efforts such as the Publishing Data Bibliometrics Working Group<sup>2</sup> at the Research Data Alliance (RDA)<sup>3</sup>; the Snowball Metrics project<sup>4</sup>; Altmetrics; and Elsevier's Metrics

<sup>1</sup> Web of Science Data Citation Index: [http://wokinfo.com/products\\_tools/multidisciplinary/dci/](http://wokinfo.com/products_tools/multidisciplinary/dci/)

<sup>2</sup> RDA/WDS Publishing Data Bibliometrics Working Group: <https://rd-alliance.org/groups/rdawds-publishing-data-bibliometrics-wg.html>

<sup>3</sup> Research Data Alliance: <http://rd-alliance.org>

<sup>4</sup> Snowball Metrics: <http://snowballmetrics.com>

Development Program<sup>5</sup>. In 2014, the NSF funded the “Make your data count” project, managed by the PloS Open Access journal in collaboration with DataONE<sup>6</sup> and the California Digital Library, to elicit ideas on data metrics from researchers. Earlier on, the MESUR project (Bollen, Van de Sompel & Rodriguez, 2008) focused on collecting evidence of usage through many types of events, but mostly those associated with references to articles. Organisations like DataCite<sup>7</sup> promote the use of persistent identifiers, like DOIs, for data, while the Publishing Data Services Working Group<sup>8</sup> at the RDA studies ways to link data and article publications.

## Contributions

This paper aims to contribute to the understanding of direct data reuse models, and its implications for the design of new credit models based on data reuse. Specifically, we make the following contributions.

Firstly, from the well-known notion of data provenance we derive a definition of the *trajectory* of a dataset  $D$ . Informally, this is the graph of all direct and transitive derivations from  $D$  to any other  $D'$ , such that there is a provenance graph that includes  $D$  and  $D'$ , and  $D$  is reachable from  $D'$  through derivation and usage/generation paths in the graph.

Secondly, we show how perfect knowledge of all such dependencies can be used to formally define *transitive credit*, and we are going to present one such credit model in detail as a concrete example. Transitive credit is based on the principle that any credit that is assigned to derivative work  $D'$  should propagate transitively “upstream” to every  $D$  such that  $D'$  is in the trajectory of  $D$ , i.e., to any  $D$  that contributed to the derivation of  $D'$ . Importantly, this model also accommodates any direct credit attribution that may be defined by the community, be it based on data citations, download counts, or other indirect criteria. Specifically, how much of  $D$ 's credit should be apportioned to  $D'$  is determined by the dependency relationships along the trajectory path from  $D$  to  $D'$ . Thus, we use the provenance of  $D'$  to assign fair credit to  $D$ , and thus to its publisher (the *Agent* responsible for  $D$ , in provenance parlance).

Thirdly, we present an instance of our credit model at work on a simulated data reuse scenario. With the understanding that many possible such models can be defined, we have implemented a *data reuse simulator*,<sup>9</sup> which we use as a research tool for exploring different credit models, and for understanding their implications for data publishers.

These contributions are designed to lay the foundations for further research in the area of data reuse analysis based on provenance. In this sense, we are aware that the concepts presented in the paper are still only theoretical. The reality of tracking data usage is a vision that presents many challenges because of the broad diversity of ways in which public data can be used without control, and the lack of metadata management infrastructure for generating and collecting provenance across many independent information systems. Implementing these ideas *in the wild* is therefore a long-term research proposition, for

<sup>5</sup> Metrics Development Program: <http://emdp.elsevier.com>

<sup>6</sup> DataONE: <http://dataone.org>

<sup>7</sup> DataCite: <http://datacite.org>

<sup>8</sup> Publishing Data Services Working Group: <https://rd-alliance.org/groups/rdawds-publishing-data-services-wg.html>

<sup>9</sup> Data Reuse Simulator: <https://github.com/PaoloMissier/DRS>

which simulated data reuse is only the beginning.

Thus, as our final contribution, we highlight some of the challenges and set out a research agenda for a practical realisation of our vision of pervasive tracking of published data through its lifetime.

## Provlets and Data Trajectories

We now outline an ideal, theoretical scenario where we assume that (i) published datasets are encapsulated as Research Objects (RO) (Bechhofer, De Roure, Gamble, Goble & Buchan, 2010), which are given unique and persistent identifiers through certified data repository managers, and (ii) complete provenance metadata is available, which describes each instance of RO reuse, at least at a high level. The research implications of relaxing these assumptions are discussed in the final section of the paper.

### Research Objects

Following emerging practice for data preservation, specifically for scientific datasets, it is now becoming realistic to assume that units of publishable data be represented as Research Objects. These are the main entities whose trajectories we want to track. ROs are encapsulations of data and metadata of any type, described by a Resource Map in ORE format. Metadata artifacts may include the description of the process (script, workflow) used to generate the data, the provenance of the data, and other metadata of varying types. Different vocabularies, or ontologies, can be used in the Resource Map to best describe such diverse metadata content. We also assume, following for example DataCite and FigShare practices amongst others, that data publishers assign unique persistent ID (PIDs), such as DOIs, to ROs upon publication, and that such PIDs are used consistently throughout the derivation chain. RO formats may vary, ranging from their original, complex, specification<sup>10</sup>, to the simpler notion of Data Packages as defined by the DataONE project<sup>11</sup>, to the even simpler but more radical notion of nanopublications (Mons et al., 2011).

### The PROV Model for Provenance

We use the PROV provenance model (Moreau et al., 2012) as a foundation for a formal and machine-processable definition of Data Trajectories. A recent book on PROV describes the W3C recommendation through a number of case studies (Moreau & Groth, 2013). Using PROV, we can express derivation dependencies of the form “*RO<sub>2</sub> wasDerivedFrom RO<sub>1</sub>*”, where *RO<sub>1</sub>*, *RO<sub>2</sub>* are PROV Entities, i.e., data or other artifacts to which we can associate a provenance. Further, if a program *P* is known to have *used RO<sub>1</sub>* as input, and have *generated* a new *RO<sub>2</sub>* as output, we can express the derivation of *RO<sub>2</sub>* from *RO<sub>1</sub>* through *P* using the following two PROV assertions:  $\langle P \text{ used } RO_1 \rangle$ ,  $\langle RO_2 \text{ wasGeneratedBy } P \rangle$ , which collectively form a (very basic) PROV document. Here, *P* is an example of an Activity, i.e., “*something that occurs over*

<sup>10</sup> See: <http://researchobjects.org>

<sup>11</sup> See: <https://releases.dataone.org/online/api-documentation-v1.2.0/design/DataPackage.html>

a period of time and acts upon or with entities” (Moreau et al., 2012).

We can also use PROV to explicitly associate both Entities and Activities with Actors, i.e., people but also, possibly, automated systems, who have been responsible for those Entities and Activities. The following PROV document extends the example above by including attribution annotations concerning two actors  $A_1, A_2$ :

$$\langle P \text{ used } RO_1 \rangle, \langle RO_2 \text{ wasGeneratedBy } P \rangle \quad (1)$$

$$\langle RO_1 \text{ wasAttributedTo } A_2 \rangle, \langle RO_2 \text{ wasAttributedTo } A_1 \rangle, \langle P \text{ wasAssociatedWith } A_1 \rangle \quad (2)$$

where assertions on line (1) describe dependencies amongst the ROs, and those on line (2) associate the ROs and the program  $P$  with Agents.

PROV defines three types of sets: (i) Entities ( $En$ ), i.e., data, documents; (ii) Activities ( $Act$ ), which represent the execution of some process over a period of time, and (iii) Agents ( $Ag$ ), i.e., humans, computing systems, software. We are going to use the following subset of relations amongst these sets:

$$\text{usage:used} \subseteq Act \times En$$

$$\text{generation:wasGeneratedBy} \subseteq En \times Act$$

$$\text{derivation:wasDerivedFrom} \subseteq En \times En$$

$$\text{association:wasAssociatedWith} \subseteq Act \times Ag$$

$$\text{attribution:wasAttributedTo} \subseteq En \times Ag$$

Furthermore, to each activity  $a \in Act$  we associate a type,  $\tau_{act}(a)$ . Activity types are useful to describe properties that are common to a set of activities, such as the parameters used to compute transitive credit for ROs, as defined later. Finally, we represent a provenance document as a Directed Acyclic Graph (DAG), where nodes denote either Entities, Activities, or Agents, and an arc of the form  $x \xrightarrow{r} y$  denotes the directed relationship  $r(x, y)$ , where  $r$  is one of the relation types above.

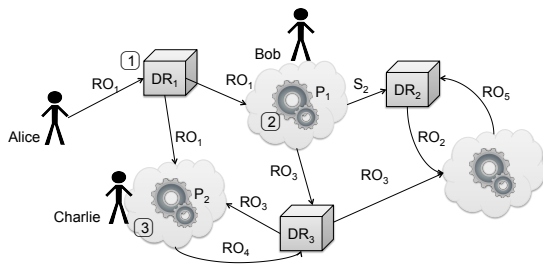
## Provlets

We have made the ideal assumption that complete provenance is available to describe each instance of data reuse. More precisely, each derivation/reuse event involving ROs is described by a small PROV document, such as those shown above. We have coined the term *provlets* to denote such documents. Although in reality each of these events may occur on a different Information System and at different times, we also assume that provlets, possibly created independently of each other, are available for each reuse event.

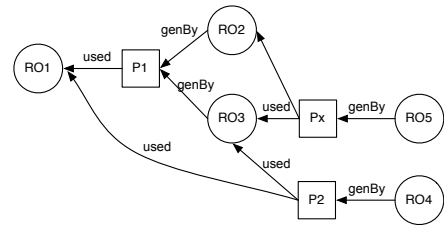
Taken individually, each provlet tells a limited story of an RO’s lifetime, as each is concerned with a single derivation step. However, as long as there is agreement amongst the system on consistently using the PIDs assigned to each RO, it is straightforward to combine a collection of provlets that contain references to the same RO, into a larger PROV document.

## A Publication-Reuse Scenario

We show the provlets idea on a simple RO publication-reuse scenario, depicted in Figure 1. The scenario involves an initial RO,  $RO_1$ , which is created and then published by Alice to data repository  $DR_1$ . This RO is later discovered, downloaded, and reused by Bob through a process  $P_1$ , and independently by Charlie through process  $P_2$ , resulting in derivative objects  $RO_2, RO_3$ , and  $RO_4$ , respectively. These new ROs may be published



**Figure 1.** A hypothetical sequence of publish-reuse actions.

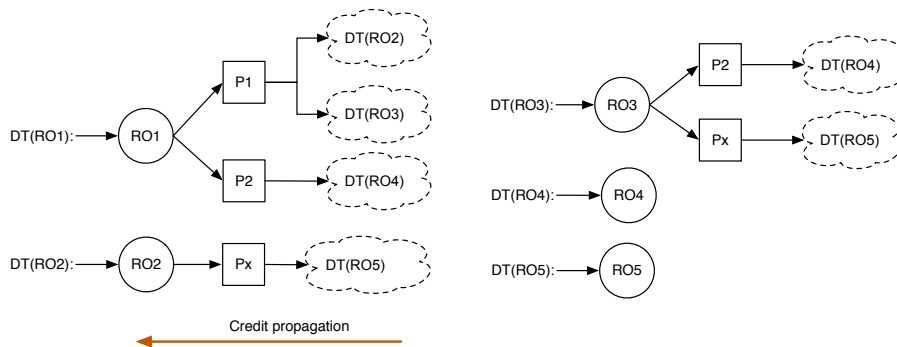


**Figure 2.** PROV graph for the sequence on the left.

into different and separate data repositories, e.g.  $DR_2$ ,  $DR_3$  as in the figure. Here Alice, Bob, and Charlie are modelled as PROV Agents, and  $P_1$ ,  $P_2$  as Activities. Not all details about a derivation are always available. For instance, in this example  $RO_2$  and  $RO_3$  are later themselves reused by some unknown Agent through some unknown Activity, generating  $RO_5$  as a result. Table 1 lists the RO reuse events for this scenario, along with the corresponding provlets in textual and graph form.

### Data Trajectories

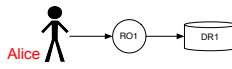
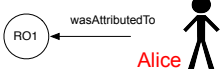
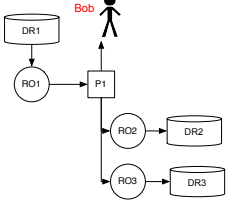
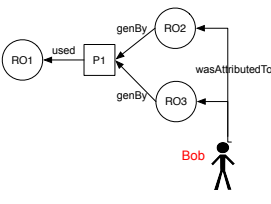
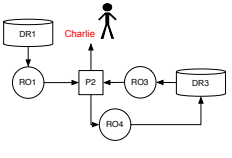
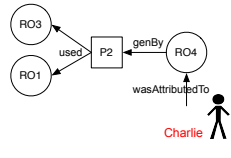
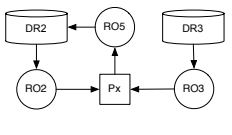
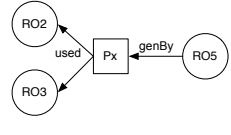
Given a provenance DAG  $p$ , consider the graph  $p'$  obtained by reversing the direction of the arcs in  $p$ . For each node  $RO$  of  $p'$ , we define the *trajectory*  $DT(RO)$  of  $RO$  to be the tree obtained by traversing  $p'$  starting from  $RO$ . We write  $DT.e(RO)$  and  $DT.a(RO)$  to denote the set of Entity (i.e. RO) nodes and Activity nodes, respectively, that appear in the  $DT(RO)$  tree. As an example, the trajectories of each of the ROs for the complete provenance graph in Figure 2 are presented in Figure 3. Note that this definition allows an RO to appear in the trajectory of another RO more than once, for instance  $RO_5$  appears twice in  $DT(RO_1)$ , because it is reachable from  $RO_1$  both through  $RO_2$  and  $RO_3$ .



**Figure 3.** Summary of trajectory trees for each of the ROs in the running example.

## From Data Trajectories to Transitive Credit for Data Owners

To illustrate how this simple notion of data trajectories provides a foundation for experimenting with models of *transitive credit*, we define one such model as an example.

Data reuse event	Prov fragment
Alice generates $RO_1$ 	$RO_1$ wasAttributedTo Alice 
Bob reuses $RO_1$ , generating $RO_2, RO_3$ 	$P_1$ used $RO_1$ , $RO_2$ wasGeneratedBy $P_1$ , $RO_3$ wasGeneratedBy $P_1$ , $RO_2$ wasAttributedTo Bob, $RO_3$ wasAttributedTo Bob, $P_1$ wasAssociatedWith Bob 
Charlie reuses $RO_1$ and $RO_3$ , generating $RO_4$ through $P_2$ 	$P_1$ used $RO_1$ , $P_2$ used $RO_1$ , $P_2$ used $RO_3$ , $RO_4$ wasGeneratedBy $P_2$ , $RO_4$ wasAttributedTo Charlie, $P_2$ wasAssociatedWith Charlie 
Unknown Agent reuses $RO_2$ and $RO_3$ , generating $RO_5$ through an unknown activity 	$P_x$ used $RO_2$ , $P_x$ used $RO_3$ , $RO_5$ wasGeneratedBy $P_x$ 

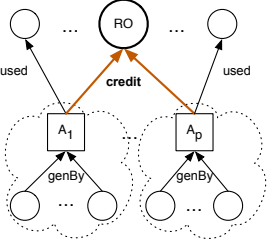
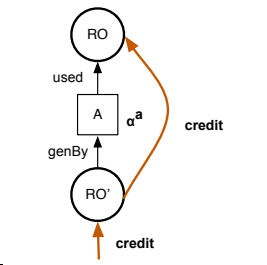
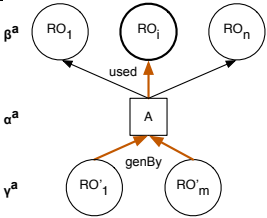
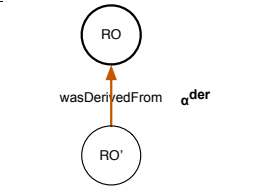
**Table 1.** RO reuse events and corresponding provlets for the running example.

The model is underpinned by a simple principle: when a derived data product  $RO'$  is credited, i.e. by the community, as a valuable research data contribution, then all of the other  $RO$ s that made  $RO'$  possible should receive some of that credit, in a proportion that depends on their importance on creating  $RO'$ . The more indispensable  $RO$  is perceived to  $RO'$ 's derivation, the more credit  $RO$  should receive. This principle applies transitively to account for multiple generations of reuse and derivation. We use Data Trajectories to determine how credit propagates “upstream” from derived  $RO$ s, possibly several steps removed from the original  $RO$ . We introduce a number of parameters, one for each of the types  $\tau_{act}(a)$  of activities  $a$  that account for the  $RO$  transformations, to quantify the notion of relative importance of the upstream  $RO$ s in the derivation process. Ultimately, credit transfers from the  $RO$ s to the Agents who are responsible for them, according to the Entity attribution assertions in the PROV document.

Following this rationale, we separate the total credit ascribed to  $RO$ , denoted  $cr(RO)$ , into two separate components. The first is the *external credit*, denoted  $cr_{ext}(RO)$ . This component accommodates any criteria that a community may decide to adopt for associating a score to a published  $RO$ , and which is independent on the reuse history of the  $RO$ . Such score may, for example, reflect emerging community practices on data citations in repositories. The second component of  $cr(RO)$  reflects the reuse history of  $RO$ . It allows each  $RO$  in the provenance graph to receive a fraction of the credit that is

ascribed to each “downstream”  $RO' \in DT.e(RO)$ . For the sake of the example, we assume that downstream credits combine linearly to provide credit to upstream nodes.

Note that this is a definition by induction, following the tree structure of  $DT(RO)$ . The base case is that of a  $RO'$  that has not been reused at all. In this case, only the external, baseline credit component  $cr_{ext}(RO')$  applies. For the induction, we now distinguish several PROV patterns of reuse. A summary of these patterns, along with their corresponding credit propagation rules and the trajectory patterns, is depicted in Figure 4.

RO reused p times		$cr(RO) = \sum_{k:1}^p cr_{a_k}(RO)$
single-input, single-output activity		$cr_a(RO) = \alpha^{(a)} \cdot cr(RO') + cr_{ext}(RO)$
many-input, many-output activity		$cr_a(RO) = \alpha^a \cdot \beta_i^a \cdot \sum_{j:1}^m \gamma_j^a \cdot cr(RO'_j) + cr_{ext}(RO)$
RO derivation with unknown activity		$cr(RO) = \frac{\alpha^{der}}{n} \cdot cr(RO') + cr_{ext}(RO)$

**Figure 4.** RO reuse patterns, trajectories, and credit propagation rules

To begin, consider the most general case, where we assume that  $RO$  has been reused by  $r$  different activities,  $a_1 \dots a_r$ , possibly at different times, as in Figure4(a). Following the structure of  $DT(RO)$  from Figure 3, we define  $cr(RO)$  to be the sum of  $r$  distinct credit components,  $cr_{a_1}(RO) \dots cr_{a_r}(RO)$ , each due to one activity  $a_k$  that has reused  $RO$ :

$$cr(RO) = \sum_{k:1}^r cr_{a_k}(RO) \quad (3)$$

We now progressively build up to a general definition of  $cr_a(RO)$ , for a generic activity  $a$ . We begin with the simplest case where  $RO$  is used by  $a$  to generate a single new RO,  $RO'$ , as in Figure4(b). As mentioned, we want  $RO$  to receive a fraction of  $RO'$ 's credit. To model the extent to which credit propagates through  $a$ , we introduce a *credit transfer parameter*  $\alpha^{(a)}$ , with  $0 \leq \alpha^{(a)} \leq 1$ . To explain its function, recall that the idea of



credit propagation through a reuse pattern  $\langle a \text{ used } RO \rangle$ ,  $\langle RO' \text{ wasGeneratedBy } a \rangle$  is based upon the intuition that  $RO'$  owes its value to both  $RO$ , and the transformation  $a$ . Introducing  $\alpha^{(a)}$  allows us to explicitly model the value contribution due to the transformation  $a$ , relative to that of its input data  $RO$ . For instance, consider a data cleaning algorithm that takes noisy data  $RO$  and produces a cleaner version,  $RO'$ , of the same data. One may argue that much of the value in  $RO'$  is due to the algorithm, rather than to the data. We model this by only transferring a small portion of  $cr(RO')$  credit back to  $RO$ , i.e., by setting a low value for  $\alpha^{(a)}$ . Note that discussing specific criteria for setting the values of this and other parameters introduced in the model is beyond the scope of this paper and left for further research, as mentioned in the last section of the paper.

Formally, we define the credit propagation rule for the graph pattern in Figure4(b) as:

$$cr_a(RO) = \alpha^{(a)} \cdot cr(RO') + cr_{ext}(RO) \quad (4)$$

where  $cr_a(RO)$  is defined inductively in terms of  $cr(RO')$ , with the external credit  $cr_{ext}(RO')$  as the base case.

Next, we extend Equation (4) to the case where  $RO$  is only one of  $n > 1$  inputs used by  $a$ . This new pattern is shown in Figure4(c). In this scenario, in addition to the transfer parameter  $\alpha^{(a)}$ , we also account for the relative importance of each of the  $n$  inputs  $RO_1 \dots RO_n$ . We therefore introduce  $n$  new factors,  $0 < \beta_i^{(a)} \leq 1, i : 1 \dots n$ , subject to:

$$\sum_{i:1}^n \beta_i^{(a)} = 1$$

and define:

$$cr_a(RO_i) = \alpha^{(a)} \cdot \beta_i^{(a)} \cdot cr(RO') + cr_{ext}(RO_i) \quad (5)$$

With this new definition,  $RO$  accrues a proportion of the total credit of  $RO'$ , which accounts for its perceived importance in computing  $RO'$  using  $a$ . Note that when there is only one input, Equation (5) reduces to Equation (4) as expected, and when all inputs to  $a$  are equally important, i.e.  $\beta_i^{(a)} = \frac{1}{n}$  for all  $i$ , Equation (5) becomes

$$cr_a(RO_i) = \frac{\alpha^a}{n} \cdot cr(RO') + cr_{ext}(RO_i) \quad (6)$$

Finally, we extend Equation (5) one more time, to account for the most general pattern where not only is  $RO$  only one of the inputs, but also,  $a$  generates  $m > 1$  outputs, as shown in Figure 4(d). In this situation,  $RO$  receives credit from each of the outputs  $RO'$ , which are all part of  $DT(RO)$ . Again, we model the different importance ascribed to each of these derived data products by introducing  $m$  new factors  $\gamma_j^{(a)}$ , subject to

$$\sum_{j:1}^m \gamma_j^a = m$$

and define:

$$cr_a(RO_i) = \alpha^a \cdot \beta_i^a \cdot \sum_{j:1}^m \gamma_j^a \cdot cr(RO'_j) + cr_{ext}(RO_i) \quad (7)$$

We conclude by adding the special case where the activity that accounts for the RO reuse is unknown. In this case, we use the generic data derivation relationship:

$$RO' \text{ wasDerivedFrom } RO \quad (8)$$

where of course more than one  $RO'$  may have been derived from  $RO$ . According to the PROV constraints document (Cheney, Missier & Moreau, 2012), from pattern (8) we can infer the existence of an activity  $a$ , such that both assertions  $\langle a \text{ used } RO \rangle$ ,  $\langle RO' \text{ wasGeneratedBy } a \rangle$  hold. We introduce a final credit transfer parameter,  $\alpha^{der}$ , to model credit propagation due to derivation. In this case, when there are  $n$  known derivations of  $RO$ , rule (4) becomes:

$$cr(RO) = \frac{\alpha^{der}}{n} \cdot cr(RO') + cr_{ext}(RO) \quad (9)$$

Finally, we stipulate that the Agents  $Ag$  that are mentioned in the PROV document accrue a credit  $cr^{ag}(Ag)$  that is simply the sum of every credit associated to the ROs they are responsible for:

$$cr^{ag}(Ag) = \sum_r cr(r) \text{ over all RO } r \text{ s.t. } \langle r \text{ wasAttributedTo } Ag \rangle \text{ holds.} \quad (10)$$

## Model Summary

We have shown how a formal notion of a *data trajectory*  $DT(RO)$ , derived from a composition of multiple, independently generated provlets, can be used to apportion credit to data publishers. As an example, we have presented a model that consists of three main elements:

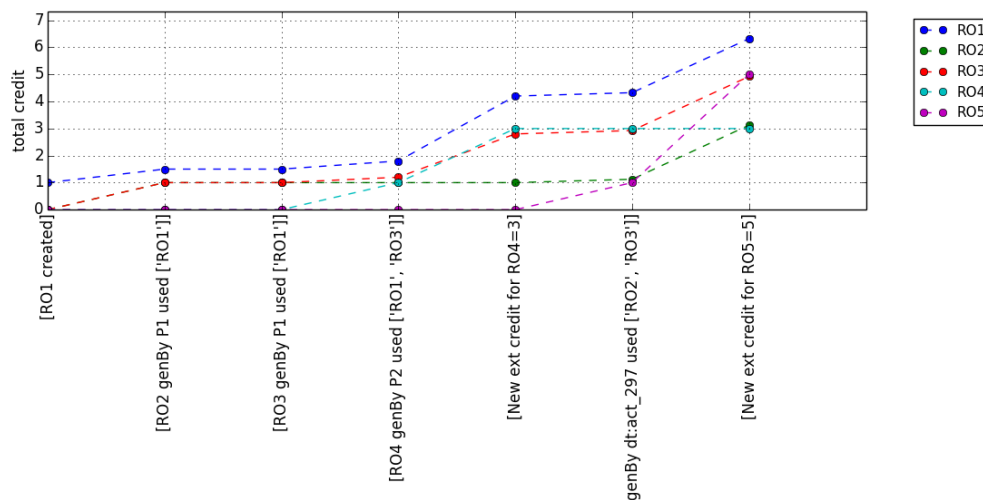
- An *external credit* function,  $cr_{ext}(RO)$ , which associates a value to each  $RO$  that appears in the compound provenance graph. Such value can follow any community-based scoring scheme of data relevance;
- A set of credit propagation rules (3) through (9) that are computed inductively from  $DT(RO)$  and which formalise the notion of *transitive credit*,  $cr(RO)$ ;
- A set of credit transfer parameters, which account for the nature of the activities involved in the trajectory of  $RO$ , including, where this information is available, the relative importance of each of its inputs and outputs.

## Simulating Data Trajectories and Credit Propagation

Realising an information management infrastructure that is capable of generating data trajectories for all instances of data reuse is a long-term, challenging research proposition, which we articulate in the final section of this paper. As a starting point for the research, we have implemented a *Data Reuse Simulator*, which we use as a tool for experimenting with various assumptions regarding the completeness of data trajectories, and with

different credit models.<sup>12</sup> The simulator is capable of generating two types of events: (i) new instances of data reuse and derivation, and (ii) updates to the external credit of one or more of the ROs, on the assumption that community-ascribed credit may change over time. Data reuse events cause the generation of more ROs, the creation of the corresponding provlets, and the update of data trajectories to reflect the new derivation and usage/generation relationships, as shown in the example of Figure 3. They also trigger the propagation of the initial external credit associated with new ROs, backwards along each of the relevant trajectories. The second type of events, changes to external credit, also triggers the propagation of the credit updates.

The simulator can be used to explore many scenarios of possible trajectory structures and credit propagation dynamics, through the generation of random interleavings of events, with some user control. Here we show the simulator in action, to reproduce the scenario in Figure 1. We have also presented a more complex data reuse scenario in the Appendix, to provide a better intuition for the simulator's capabilities. The plot in Figure 5 shows how credit changes for the ROs, in response to key events in our example, shown at the bottom. Initially, all new ROs have the same external credit value 1. Following the reference scenario, these values propagate through activities P1 and P2, as well as through a third unknown activity.



**Figure 5.** Total credit changes to ROs following reuse and external credit adjustment events.

In the simulator, we make the simplifying assumption that all inputs to an activity  $a$  are equally important, i.e. we use Equation (5) where  $\beta_i^{(a)} = \frac{1}{n}$  for all  $i$ . Similarly, we use a single value  $\gamma^{(a)} = m$ , the number of inputs to  $a$ . With these assumptions, we can express the type  $\tau_{act}(a)$  of an activity  $a$  as a triple  $\tau_{act}(a) = [\alpha, \beta, \gamma]$ . In the example, we have used  $\tau_{act}(P1) = [0.5, 1, 0.5]$ , and  $\tau_{act}(P2) = [0.8, 0.5, 1]$ . The implicit activity  $dt:act_{297}$  is assigned  $\tau_{act}(P1)$  by default.

The figure illustrates the different ways that the total credit of each RO progresses, at a faster or slower pace than that of others, depending on the amount of reuse and the type of

<sup>12</sup> The current version of the simulator software is available at <http://github.com/PaoloMissier/DRS>. It is implemented in Python and makes use of the Southampton provenance suite: <http://provenance.ecs.soton.ac.uk>

activity that consumes the RO. As expected, the oldest RO,  $RO_1$  acquires the highest credit as its trajectory extends over time, and as its descendents acquire recognition through additional external credit. Note that credit can be transferred from ROs to the agents that are responsible for them, by using the *attribution* and *association* PROV relationships.

## Data Trajectories in Practice: Challenges and Research

The data trajectories and the transitive credit model illustrated in this paper are both theoretical. In reality, because of the broad diversity of ways in which public data can be used without control, the vision of tracking data usage *in the wild* faces many challenges. We conclude by highlighting some of these challenges, and set out a research agenda for realising transitive credit in practice.

Trajectories are compositions of independently created provlets, which must be systematically generated by multiple, diverse, autonomous information systems, to the extent possible through observation of data transformation processes. This is not unrealistic, as provenance recorders exist for languages like Python (Murta, Braganholo, Chirigati, Koop & Freire, 2014) and R (Liu & Pounds, 2014; Lerner & Boose, 2014), as well as for many workflow management systems including Taverna, eScience Central, SciCumulus, Pegasus, Kepler. However, no system today systematically harvests these traces in a central place, where trajectories can be computed. This is a long-term infrastructure problem, requiring concerted efforts across data repositories organisations. Also, the granularity at which provenance is recorded varies, depending on the systems' provenance capture capabilities. Further, provlet composition requires the consistent use of data identifiers across instances of data reuse and across systems. This is by no means the norm today, although standards for data PIDs, like those promoted by DataCite, are gaining acceptance in forums like the Digital Curation Centre in the UK<sup>13</sup>, and more globally, the RDA. However, even when identifiers are available data consumers have no obligation to acknowledge their primary source of data. This is particularly problematic in the so-called *long tail of science* (Wallis, Rolando & Borgman, 2013), where consumers are less likely to record reuse in any systematic way. Credit management is further complicated when ROs are only *partially* reused, as this violates the assumption that ROs are atomic data entities.

To some extent, these issues can be addressed through a long-term plan to develop infrastructure to support the notion of data trajectories across the broad research science community. More fundamentally, however, we should assume that trajectories are always bound to be fragmented and incomplete representations of actual data reuse, leading in turn to unrealistic credit assignments. Our suggested research agenda is therefore focused on addressing the following key research questions.

- Firstly, under what circumstances it is possible to estimate the likelihood of some of the missing derivations (for instance, using machine learning and predictive analytics techniques)?
- Secondly, to what extent can the resulting probabilistic provenance graphs and trajectories be used to support useful, fair, and credible transitive credit models?

<sup>13</sup> Digital Curation Centre: <http://dcc.ac.uk>

- Thirdly, when using a credit model that relies on credit transfer parameters, as we have shown, how are these determined? Can they be learnt, or adjusted dynamically following feedback from the community?

## References

- Bechhofer, S., De Roure, D., Gamble, M., Goble, C. & Buchan, I. (2010). Research objects: Towards exchange and reuse of digital knowledge. *Nature Precedings*. doi:10.1038/npre.2010.4626.1
- Bollen, J., Van de Sompel, H. & Rodriguez, M. A. (2008). Towards usage-based impact metrics. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries - JCDL '08* (p. 231). New York, New York, USA: ACM Press. doi:10.1145/1378889.1378928
- Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., . . . Wright, D. (2012). *Making data a first class scientific output: Data citation and publication by NERC's environmental data centres*. doi:10.2218/ijdc.v7i1.218
- Cheney, J., Missier, P. & Moreau, L. (2012). *Constraints of the provenance data model* (Tech. Rep.). Retrieved from <http://www.w3.org/TR/prov-constraints/>
- Katz, D. S. (2014). Transitive credit as a means to address social and technological concerns stemming from citation and attribution of digital products. *Journal of Open Research Software*, 2(1), e20.
- Lerner, B. S. & Boose, E. R. (2014). Collecting provenance in an interactive scripting environment. *Proceedings of TAPP'14*.
- Liu, Z. & Pounds, S. (2014). An R package that automatically collects and archives details for reproducible computing. *BMC Bioinformatics*, 15(1), 138. doi:10.1186/1471-2105-15-138
- Mayernik, M. S. (2013). *Bridging data lifecycles: Tracking data use via data citations workshop report* (Tech. Rep.). National Center for Atmospheric Research.
- Mons, B., van Haagen, H., Chichester, C., Hoen, P.-B. T., den Dunnen, J. T., van Ommen, G., . . . Schultes, E. (2011). The value of data. *Nature Genetics*, 43(4), 281–3. doi:10.1038/ng0411-281
- Moreau, L. & Groth, P. (2013). Provenance: An introduction to PROV. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 3(4), 1–129.
- Moreau, L., Missier, P., Belhajjame, K., B'Far, R., Cheney, J., Coppens, S., . . . Tilmes, C. (2012). *PROV-DM: The PROV data model* (Tech. Rep.). World Wide Web Consortium. Retrieved from <http://www.w3.org/TR/prov-dm/>
- Murta, L., Braganholo, V., Chirigati, F., Koop, D. & Freire, J. (2014). noWorkflow: Capturing and analyzing provenance of scripts. *Proceedings of IPAW'14*.

Piwowar, H. A. & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, 1, e175. doi:10.7717/peerj.175

Robinson-García, N., Jiménez-Contreras, E. & Torres-Salinas, D. (2015). Analyzing data citation practices using the data citation index. *Journal of the Association for Information Science and Technology*. doi:10.1002/asi.23529

Wallis, J. C., Rolando, E. & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS ONE*, 8(7), e67332. doi:10.1371/journal.pone.0067332

## **Appendix: A More Complex Instance of Simulated Data Reuse**

Figure 6 shows a more complex simulated data reuse scenario, which includes 15 ROs, managed by nine Data Operators (the Agents at the top of the figure), with a random combination of ten derivation and usage/generation events. These are (randomly) interleaved with ten external credit update events. The resulting progression of total credit over time is shown in Figure 7.

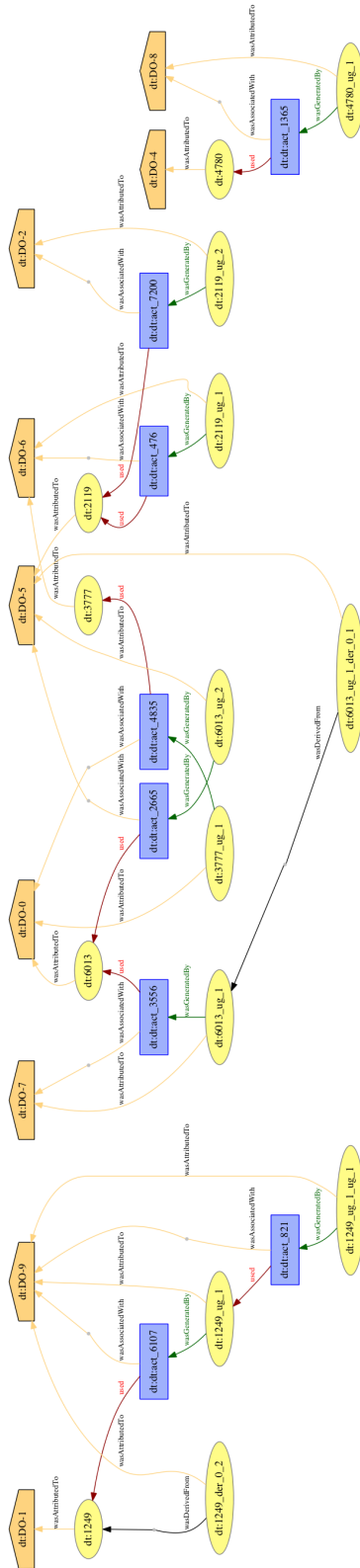


Figure 6. The global provenance graph for the entire reuse history.

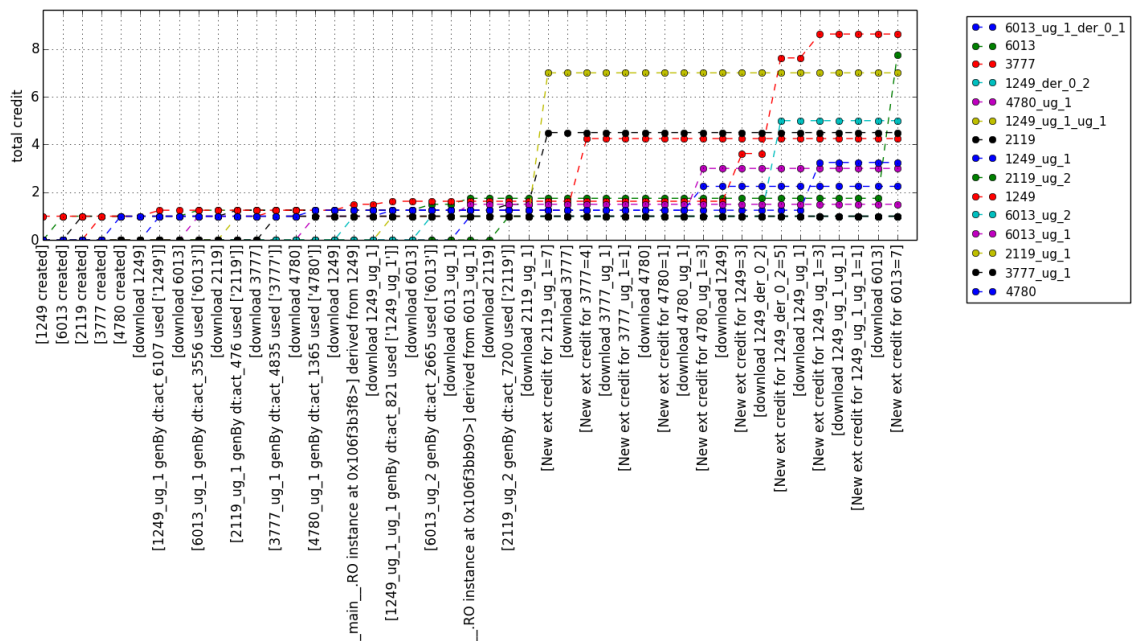


Figure 7. RO total credit progression for the data reuse scenario of Figure 6.