

A User-Orientated Approach to Provenance Capture and Representation for *in silico* Experiments: Explored within the Atmospheric Chemistry Community

Chris J. Martin^{1,2}, Mohammed H. Haji², Peter M. Dew², Michael J. Pilling¹, Peter K. Jimack²

¹ School of Chemistry, University of Leeds, Leeds, LS2 9JT, UK

² School of Computing, University of Leeds, Leeds LS2 9JT, UK

Abstract. We present a novel user-orientated approach to provenance capture and representation for *in silico* experiments, contrasted against the more systems-orientated approaches that have been typical within the e-Science domain. In our approach we seek to capture the scientist's reasoning in the form of annotations as an experiment evolves, whilst using the scientist's terminology in the representation of process provenance. Our user-orientated approach is applied in a case study within the atmospheric chemistry domain: where we consider the design, development and evaluation of an Electronic Laboratory Notebook, a provenance capture and storage tool, for iterative model development.

Keywords: Provenance, Semantic Web, Atmospheric Chemistry.

1 Introduction

Provenance, in relation to scientific data, can be defined as “the derivation history of a data product starting from its original sources.”[1]. Within the e-Science community capturing, representing and storing provenance for scientific experiments is an emerging field of research that has recently generated substantial interest. Research into this emerging field is motivated by the need to archive large quantities of data: provenance is required for the scientist to fully understand the scientific process they, and/or others, have executed when generating the data in question. The research presented in this paper explores a *user-orientated* approach to provenance capture and representation. We seek to place the scientist at the heart of the provenance capture process eliciting their scientific reasoning, as annotations, as they conduct an *in silico* experiment. Alongside these annotations we capture process provenance [2], structuring the provenance using terminology from the scientific domain. By adopting a user-orientated approach to provenance across a scientific community we suggest that a number of benefits can be realised by individual researchers and the wider research community, including: enabling researchers to reduce the amount of time they spend interpreting or re-interpreting archived data (produced either by themselves or third party researchers); facilitating novel sharing processes, enabled by the aggregation of provenance and data across a geographically distributed research community, such as developing benchmark community data and knowledge repositories [3] [4].

We evaluate our user-orientated approach to provenance by means of a case study, exploring the capture and representation of provenance for iterative computational modelling experiments in the atmospheric chemistry community. The atmospheric chemistry community relies on the complementary efforts of experimentalists and modellers seeking to develop a better understanding of the chemical processes taking place in the atmosphere. This understanding is used to construct chemical mechanisms, lists of elementary chemical reactions, that quantitatively describe atmospheric chemistry. Mechanisms can then be used, often in a reduced form, as components of predictive climate and air quality models. Mechanisms are grounded in experimental chemical kinetics and provide a critical link between fundamental experimental science and large-scale predictive models.

We use an Electronic Laboratory Notebook (ELN) [3] to capture both annotations and process provenance, implementing our user-orientated approach to provenance. The ELN is currently at a prototype stage and has been the subject of some preliminary user evaluations. The output of these user evaluations will inform

1
2
3 the design and development of a production quality, open source, ELN for use by the atmospheric
4 chemistry modellers across an international research community. Our ELN places annotation opportunities
5 within the scientific process executed by the computational modeller, in the form of prompts, whilst
6 seeking to minimise changes to the scientific process. The ELN monitors the processes executed by the
7 scientist to both capture provenance and drive the annotation prompts. By placing the annotation prompts
8 within the scientific process we seek to capture the modeller's reasoning as it takes place, mirroring the
9 current practices of a scientist making notes in their lab-book as they are going along. The process
10 provenance captured by the ELN is represented using terminology from the scientific domain of interest: in
11 this case study atmospheric chemistry. We seek to understand and capture the science taking place rather
12 than just recording the changes from a system orientation. For example what could be viewed from a
13 system orientation as a change to the last modified date of a model input file, is from a science-orientation a
14 change to the scientific nature of the computational model. The provenance captured by the ELN is
15 structured and stored using semantic web technologies, owl [5] and rdf [6], to enable the development of
16 provenance-consuming internet applications in our future work.

17
18 Section 2 of this paper discusses approaches to provenance, and outlines the characteristics of our user-
19 orientated approach, placing our approach in the context of related research. Section 3 provides an
20 introduction to the Electronic Laboratory Notebook (ELN) and its role within our user-orientated approach
21 to provenance. Section 4 introduces background information to the case study we use to evaluate our user-
22 orientated approach, discussing the atmospheric chemistry community and its computational modelling
23 processes. Section 5 presents our case study: the design, development and evaluation of our prototype
24 ELN, with particular reference to the interaction between the user and the ELN and the ontology used to
25 structure provenance. Section 6 provides our conclusion and an outline of our future work in this area.

27 2 Approaches to Provenance

28
29
30 **The Scientist's Approach to Provenance.** Scientists have been capturing provenance, alongside the
31 scientific data they generate, for centuries [7]. The traditional means of capturing provenance has been the
32 Laboratory Notebook (LN), used to capture both the experimental process (process provenance) and
33 annotations relating to the scientists reasoning (annotations). Whilst there are many drawbacks to capturing
34 provenance using a LN, the ways in which scientists use their LN suggest three important user
35 requirements for provenance capture and representation.

36
37 First, scientists capture provenance as they execute their experiments, we will refer to this as inline
38 provenance capture, in addition to capturing provenance before (pre-hoc) and after (post-hoc) their
39 experiments. Inline provenance capture is required to enable the scientist to capture process provenance and
40 reasoning annotations as the scientific process evolves, and decisions are made, not necessarily adhering to
41 an experimental plan. Secondly, scientists make annotations relative to different frames of reference,
42 dependent on the context of annotation. Frames of reference used include: the high level experiment where
43 a scientist may wish to provide annotations incorporating experimental goals and conclusions; individual
44 elements of the scientific workflow executed, e.g. the scientist may provide annotations incorporating
45 reasons for changing an individual experimental parameter; ad-hoc, aggregations of workflows or
46 workflow elements, for example a scientist may wish to define and annotate a set of sub-experiments that
47 have taken place under a single main experiment. It is important to note for each frame of reference
48 scientists make annotations with a different content, detail and structure, i.e. the annotation of an
49 experiment differs significantly from the annotation of changing a model parameter. Thirdly, scientists
50 capture provenance using scientific terminology. The use of scientific terminology, specific to the domain
51 of the experiment, enables a great deal of information to be recorded within the provenance in a concise
52 manner (relying on a common understanding of the terminology).

53
54 **The Systems-Orientated Approach to Provenance for *in silico* Experiments.** Within the e-Science
55 domain, research into provenance capture, representation and storage for *in silico* experiments has been
56 tightly coupled with the workflow systems [8] [9] paradigm. For the purpose of comparison between the
57 workflow approach to provenance and our user-orientated approach we take the Taverna system [10] as an
58
59
60

1
2
3 exemplar from the workflow system paradigm. In reviewing the Taverna system we consider two key
4 characteristics.
5

6 First, Taverna [11], in common with many other workflow systems [12] [13], seeks to automatically
7 capture provenance for *in silico* experiments, minimising user involvement. Automatic provenance capture
8 is well suited to capturing process provenance, i.e. the structure and execution of the workflow, but
9 overlooks the importance of capturing the scientist's contribution to the scientific process (e.g. why they
10 used a given service, why they have re-run a workflow with a modification to the input parameters). Within
11 the Taverna workflow environment user involvement is limited to annotating a given workflow or
12 workflow component with a single high-level description, this annotation can be either pre-hoc (before
13 running the workflow) or post-hoc (after running the workflow). So Taverna can be seen to lack support for
14 inline annotations and provides limited support for annotating with respect to multiple frames of reference.
15 Secondly, the provenance captured by Taverna, as with many other workflow systems [14] [15], is
16 represented using domain independent semantics. So the scientific process (captured as a workflow/series
17 of workflows), of a given researcher, is represented independently of the particular scientific domain.
18 Whilst the use of domain independent semantics can be seen as an important factor in producing a domain
19 independent workflow system, that is deployable across scientific domains, domain independent semantics
20 remove the opportunity to leverage the informational content of the scientific terminology of a given
21 scientific domain. Given the key characteristics identified above, minimising user involvement in
22 provenance capture and using domain independent semantics to represent provenance, the Taverna
23 approach to provenance can be viewed as system orientated.
24

25 **A User-Orientated Approach to Provenance for *in silico* Experiments.** The differences between the
26 system-orientated (i.e. computer science driven) and the scientist's approaches to provenance, can be seen
27 to be a result of cultural differences between the two communities. Our work seeks to develop a user-
28 orientated approach to the capture of provenance, both process provenance and annotations, for *in silico*
29 experiments. We attempt to reconcile the scientist's and the system-orientated approaches to provenance
30 capture, discussed above. From the system-orientated approach we will seek to automate process
31 provenance capture, whilst adopting the key practices from the scientist's approach: inline annotation,
32 annotations with respect to multiple frames of reference and the use of scientific terminology in the
33 representation of provenance. So whilst we seek to minimise user involvement in the capture of process
34 provenance, we seek to engage the user in annotating their scientific process. By adopting this user-
35 orientated approach we can complement detailed process provenance, captured automatically, with a record
36 of the scientist's reasoning and leverage the informational content of the domain-specific scientific
37 terminology.
38

39 **Related Work.** The 1st Provenance Challenge [9] sought to understand how a number of provenance
40 systems address a benchmark provenance problem, with particular respect to: how provenance is
41 represented; the ability of the provenance system to answer queries; and what is considered in scope for
42 provenance capture. The MyGrid research group address the provenance challenge using Taverna plus a
43 knowledge template [16], which adds semantic annotation functionality. The knowledge template allows
44 users to create annotations to enrich the domain independent process provenance automatically captured by
45 Taverna with semantics from a specific scientific domain. This is in contrast to our approach where we
46 capture process provenance, using semantics from a specific scientific domain, automatically. The
47 VisTrails response to the first provenance challenge [17] adopts a change-based approach to provenance,
48 capturing the evolution of a workflow as a scientist conducts exploratory research. Provenance is captured,
49 and annotation enabled, at three layers: workflow evolution, the workflow structure and the workflow
50 execution. In our approach we take this one stage further, capturing changes to both the workflow and the
51 input data, using scientific terminology. A number of provenance systems, including Karma [18], applied
52 to the first provenance challenge, considered annotations beyond the scope of the provenance research
53 discipline. We view this as the extreme system-orientated perspective on provenance, completely
54 eliminating the role of the scientist in provenance capture, which runs the risk of capturing provenance of
55 limited value for the long-term archival of data. The extreme system-orientated approach produces
56 provenance that describes how a given data item was produced, but none of the critical scientific
57 information on why data was produced in a certain way that our approach seeks to capture.
58
59
60

1
2
3
4 The importance of the scientist's contribution to provenance has been recognised in the work of the
5 PolicyGrid project, where they seek to capture the scientist's intent as well as their method [19]. PolicyGrid
6 have taken the Kepler workflow environment [15], and added functionality to capture and structure
7 provenance that describes the intent of a scientist executing a workflow. This enables the scientist to
8 annotate a workflow, and structure these annotations with use of ontology, with goals, reasoning etc.,
9 whereas our approach seeks to capture annotations for the individual processes that composed a workflow
10 in a context sensitive fashion.

11 3 An ELN for Iterative Computational Modelling

12
13
14 Our user-orientated approach to provenance for *in silico* experiments makes use of an Electronic
15 Laboratory Notebook (ELN), and is evaluated in the context of the iterative development of computational
16 models in the atmospheric chemistry community. Iterative computational modelling can be defined, for the
17 purpose of this paper, as developing a computational model through a cycle of the following activities:
18 changing some aspect of the model; running the model; analysing the model output (where this analysis
19 informs the next change to the model).

20
21 ELNs have typically been used for the capture of provenance for *in-vitro* experiments [7] and provide an
22 electronic replacement for the traditional laboratory notebook (LN) in which a scientist is able to record
23 their experimental process along side their reasoning and thoughts. ELNs have been developed and
24 deployed extensively in commercial settings [20, 21], such as drug development, where they provide a
25 stronger basis than a traditional LN, for intellectual property claims. ELNs have also been researched and
26 deployed in a variety of academic settings [22], including the CombeChem ELN [7] an important reference
27 point for our research. The CombeChem ELN is used to capture provenance for organic synthesis
28 chemistry experiments, where the scientist typically performs a sequential set of actions (mixing chemicals
29 together, heating or cooling mixtures, etc.) in a laboratory setting. The response to a prototype CombeChem
30 ELN by potential users has been positive [7], during initial usability trials, and a production quality ELN is
31 currently being engineered (personal communication Jeremy Frey, September 2008). Process provenance is
32 captured from the plan of the experimental process (a mandatory safety requirement prior to commencing
33 all experiments), with amendments to the experimental process and annotations made at experimental run
34 time. A key difference between iterative computational modelling and *in-vitro* experiments, is that when
35 modelling there is no need for an experimental safety plan (or any detailed plan whatsoever), so we seek to
36 capture process provenance automatically from the individual computational processes.

37 4 Case Study Background

38
39
40 In order to test our user-orientated approach to provenance we undertook a case study considering
41 provenance for the iterative development of computational models in the atmospheric chemistry
42 community. In this case study we focused on two aspects of the user-orientated approach to provenance:
43 inline annotation and the use of scientific terminology in provenance representation. Annotation is
44 considered only with respect to a single frame of reference; the annotation of individual workflow
45 components. This section provides background to the scientific community and modelling process involved
46 in the case study.

47
48 **Atmospheric Chemistry Community.** Atmospheric chemistry is an inherently multi-scale science,
49 incorporating a variety of field, *in vitro* and *in silico* experimental disciplines. At the global and regional
50 scales the atmospheric chemistry community is involved in a number of high profile modelling activities
51 including: modelling of global concentrations of methane and ozone, which, after CO₂ are the trace gases
52 with the greatest influence on climate change; developing models which inform the development air quality
53 policy. A central component of the models investigating atmospheric chemistry on a global or regional
54 scale is the chemical mechanism. Chemical mechanisms, part of the molecular scale of atmospheric
55 chemistry study, consist of a coupled set of steps called elementary reactions in which chemical species are
56 inter-converted (i.e. mechanisms are lists of chemical reactions). Each elementary reaction can be
57 considered in the form: $reactants \xrightarrow{k} products$, where the *reactants* are the set of chemical species that
58
59
60

1
2
3
4 react together to generate the *products* (another set of chemical species) and k is the rate co-efficient of
5 reaction. Elementary reactions are investigated primarily in the laboratory; detailed chemical mechanisms
6 are constructed from knowledge of these elementary reactions and their interactions. Mechanisms are used
7 directly to construct models containing a very large set of ordinary differential equations that represent the
8 rates at which the concentration of individual species in the mechanism change with time. Such models are
9 used for problems with modest fluid dynamic requirements e.g. local scale modelling, in order to test the
10 performance of the chemical mechanism. These mechanisms can contain a large number of elementary
11 reactions, often in excess of 10000, and so are too computationally expensive to implement within global
12 and regional models e.g. for aspects of climate change or regional air quality. In such cases, mechanisms of
13 much lower dimension are used, ideally based on objective lumping of the detailed mechanisms, providing
14 a link between the global and regional scale models, and fundamental chemical kinetics. Research on
15 elementary reactions and chemical mechanisms is conducted in research laboratories throughout the world.
16 The **Master Chemical Mechanism (MCM)** is the leading detailed chemical mechanism, used across the
17 international research community, and describes the chemistry occurring in the lower atmosphere. It is
18 used both directly in local scale models and to evaluate smaller lumped mechanisms used in global and
19 regional atmospheric models. Within the wider chemistry community a great deal of effort has been
20 committed to the development of schemas and ontology for representing chemical data, including the CML
21 [23, 24] and ChEBI [25] projects. Up until this point efforts have focused on describing the structural
22 properties of atoms and molecules, with neither of the aforementioned projects addressing the
23 representation of mechanisms or the processes involved in *in silico* atmospheric chemistry experiments, of
24 the type we consider in this paper.

25 **Atmospheric Chemistry Models.** Computational modelling takes many forms within the atmospheric
26 chemistry community, as described above. In this paper we focus on recording the provenance for one form
27 in particular, so-called zero-dimensional box models [26], where the aim of modelling is to develop an
28 understanding of the chemical processes taking place at a given location (i.e. the local scale). Field and *In*
29 *vitro* experiments at the local scale including: field campaigns that make *in situ* measurements at a single
30 location; and experiments in atmospheric simulation chambers; can be modelled using zero-dimensional
31 box models, incorporating the MCM. The output of these local scale models can then be compared to the
32 field or *in vitro* experiment data (as appropriate), in order to test the performance of the MCM. In this case
33 the modeller will make use of experimental data, various *in-situ* measurements of chemical concentrations,
34 and vary the configuration of the model comparing *in vitro* experimental data with model output data.
35 During this process the modeller will extensively experiment with the chemical mechanism implemented
36 within the model, adding, deleting or changing chemical reactions and testing the impact this has on the
37 model output (validated against the aforementioned *in-situ* measurements).

38
39 The model development process we consider in this paper is iterative, with the changes made to the
40 mechanism, determined by the conclusions drawn when comparing the model output to experimental data.
41 Typically the modeller does not form a detailed plan of action, instead working in an exploratory manner
42 drawing on their own knowledge and experience, in conjunction with the conclusions they draw from the
43 comparison of model output and experimental data. This method of working has a significant implication
44 for provenance capture: it places a premium on capturing the modeller's reasoning and thoughts alongside
45 the details of the modelling workflow. We seek to address this within our user-orientated approach to
46 provenance.

47 48 **5 Case Study**

49 50 **5.1 Requirements Capture and Design Methodology**

51
52 Given the focus of our work on adopting a user-orientated approach to provenance, an ethnographic
53 methodology [27] was adopted to ensure that the requirements and motivations of modellers within the
54 atmospheric chemistry community could be understood. An author, CJM, was embedded within the
55 atmospheric chemistry modelling group at the University of Leeds. Prior to, and throughout, the
56 development of the ELN he worked on atmospheric chemistry modelling projects, seeking to deliver
57 atmospheric chemistry research while developing personal insight into the scientific processes, motivation
58 and provenance requirements of atmospheric chemistry modellers.

1
2
3
4 Capturing the modelling process used by atmospheric chemistry modellers was the first phase of
5 developing the ELN prototype. The process capture was facilitated by considering a modelling case study
6 based on the development of a model for a field campaign that took place in Tasmania, SOAPEX [28]. The
7 SOAPEX field campaign made measurements of: free-radical species concentrations including OH, HO₂;
8 environmental conditions including photolysis rates, temperature and pressure; concentrations of other
9 important chemical species including O₃, CO, NO, NO₂, and a variety of hydrocarbons. The campaign
10 took place at Cape Grim, Tasmania, in extremely clean air conditions ([NO]₃ ppt). Subsequently 0-
11 dimensional box models were developed to enable model-experiment comparisons for HO_x radicals with
12 insight developed into the chemistry of HO_x radicals in clean air. The model in the case study was
13 relatively simple, but also retained all the key characteristics of more complex models. The process for
14 developing the SOAPEX model was then mapped, at the finest granularity of task description possible, to
15 produce a process description for the case study. The importance of capturing process at such fine
16 granularity is that only with this level of detail is it possible to repeat an experiment (either modelling or
17 laboratory based). The case study process description was then examined to develop a provenance
18 specification. This provenance specification was developed from an end user perspective, in the form of a
19 set of provenance reports for the case study modelling process. The subsequent design and implementation
20 of the prototype was guided by this provenance specification.
21

22 **5.2 Prototype Implementation**

23 In this paper we consider the two aspects of the prototype implementation: first, the scientific terminology
24 used in the representation of the provenance, in the form of ontology; secondly, the interaction patterns
25 between the user and the ELN during inline annotation. Further details of the design and implementation of
26 the prototype ELN are provided elsewhere [3].
27

28 **5.2.1 The use of scientific terminology in provenance representation**

29 As a starting point to the development of our ontology we took the CombeChem ELN ontology [7],
30 designed to structure provenance for *in vitro* chemistry experiments. Our ontology shares the same set of
31 top-level concepts, from which all other concepts inherit, with the CombeChem ontology. The top-level
32 concepts, in both ontologies, are processes and materials, below this level we have developed domain-
33 specific and domain independent elements of ontology as required during the development of the prototype
34 ELN. In this section we explore the ontology developed to capture the changes made to the chemical
35 mechanism within a zero-dimensional box model. Our ontology is expressed using owl, with the
36 provenance generated as rdf conforming to the ontology.
37

38 High-level processes are used to describe elements of the iterative model development process, and can be
39 linked together to describe the scientific workflow executed during iterative modelling. A typical fragment
40 of workflow, as shown in Fig. 1, would incorporate model development, model execution and data analysis
41 processes, linked together in series to form a single iteration of model development. The spine of the
42 workflow is composed of process-material pairs, as in CombeChem, so the materials (in this case data
43 products) provide the glue that holds the workflow together.
44

45 The three high-level processes shown in Figure 1 can be seen as system-orientated concepts for capturing a
46 computational modelling workflow using domain independent concepts. We developed the ontology
47 further through a lower conceptual level to incorporate scientific terminology from the atmospheric
48 chemistry domain. Taking the “model development” process as an example, we identified a number of
49 types of model development: mechanism development; developing the environmental conditions (e.g. the
50 input data for setting the temperature profile over time); and, developing the solver configuration (i.e.
51 tuning the numerical integrator for the particular problem being solved). These concepts are sufficiently
52 specific for the atmospheric chemistry modellers to relate to, but can be decomposed further to allow more
53 scientific detail to be included in the process provenance capture by our ELN. So we continued to develop
54 the ontology at lower conceptual levels.
55

56 Taking for example the decomposition of the “mechanism development” process, the modeller can perform
57 a wide variety of operations on the mechanism, see Figure 2, including adding, deleting and editing
58
59
60

reactions. The ontology also includes the decomposition of the edit reaction process (edit reactants, edit products, edit rate coefficient). Where a modeller has performed a number of operations on a mechanism during one modelling iteration, each operation is captured individually (the implications for annotation are considered below). Capturing this level of detail in the provenance, if it is appropriately annotated by the modeller, provides the potential to enable user-orientated queries. In the next section we consider the ELN interface that enables the capture of user annotation.

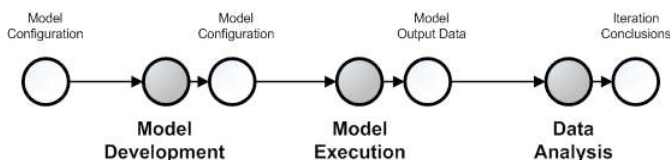


Figure 1: The high level workflow typically executed by an atmospheric chemist, performing an *in silico* experiment. The model configuration is edited in some way, during the model development process, the model configuration is then realised, within the model execution process. Model output data, is produced by the model execution process, and is an input to the data analysis process, which outputs a set of conclusions about the impact of the change to the model configuration.

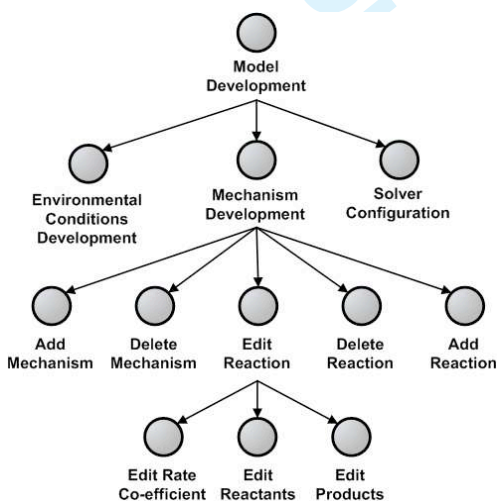


Figure 2: Domain-specific terminology for the “model development” process, an example from provenance captured by our prototype ELN for *in silico* atmospheric chemistry experiments. The figure provides a hierarchical decomposition of the model development process, considering developing the chemical mechanism and editing a reaction within the chemical mechanism as exemplar processes.

5.2.2 Capturing Inline Annotations

Continuing the discussion of a modeller iteratively developing a chemical mechanism within a box model we now consider the general pattern of interaction between the user, the model and the ELN. In the interaction sequence described below annotation is placed inline within the scientific process, mirroring how a scientist would make annotations as they go along when using their laboratory notebook.

1. The interaction begins with the modeller editing the chemical mechanism using a text editor, provided within the modelling environment. For example adding the reaction:



where k is the rate coefficient and equal to $6.01 \times 10^{18} \times (T / K)^2 \times e^{(170K/T)}$

2. The user then runs the model to test the impact of adding this reaction, by accessing functionality within the modelling environment.

3. The ELN then compares the submitted mechanism, with the preceding mechanism (retrieved from a local database) to determine how the mechanism has been changed. In this example reaction R1 has been added. Semantic provenance is then generated; see Figure 4 and the discussion below.
4. The changes in the mechanism then drive a prompt to appear within the ELN user interface. The user must address this prompt before the model runs. In the example the prompt shown in Figure 3 would be presented to the user, here the reaction is represented using a notation specific to the atmospheric chemistry community involved in the case study.
5. The user then enters their annotation in the text field within the prompt. In the example the annotation could be “Add initial oxidation reaction for Methanol. This reaction had been omitted from the original mechanism in error.”
6. Upon completion of the prompt the model runs within the modelling environment.

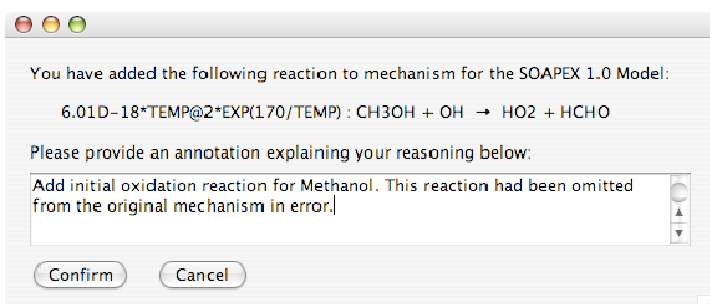


Figure 3: A prompt generated by the ELN in response to a user adding a reaction to the chemical mechanism. The prompt provides the users with an opportunity to record the scientific reasoning that underpins the change, in the form of a free text annotation.

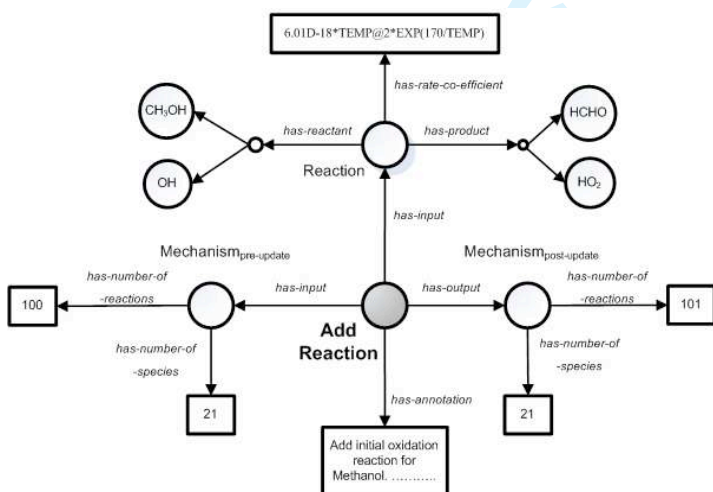


Figure 4: A graphical representation of the rdf generated, for an “add reaction” process, by the our prototype ELN. Domain-specific terminology is used to record the change to mechanism, e.g. “reaction”, “mechanism” etc. The annotation captured, by the ELN prompt (see figure 3), is also represented.

Figure 4 shows a simplified representation of the semantic provenance generated for the sequence above. The “add reaction” process has two inputs, a chemical mechanism (mechanism_{pre-update}) and a reaction (R1), and one output, a revised chemical mechanism (mechanism_{post-update}). The ELN parses the reaction added to the mechanism, enabling a structured representation of the reaction to be recorded in the provenance. The text name for each of the chemical species involved in the reaction are compared to a reference database enabling an InChi [29](non-proprietary identifier for chemical substances) to be used within the provenance. The rate co-efficient is captured as a text string - work in progress on a related project looks at representations of chemical rate co-efficients using mathML [30] and CML [23], once this task is

completed it will be considered for incorporation into our work. Each mechanism is identified by a URI generated when the mechanism is submitted to the ELN. The ELN also generates some simple metadata for the mechanism including the number of reactions and chemical species within the mechanism.

5.3 Prototype Evaluation

5.3.1 Evaluation Methodology

To evaluate the ELN prototype system we adopted an approach that draws on the Scenario Based Development paradigm [31]. The goal of the evaluation was to elicit responses that can inform the design of a production quality ELN for use by the wider community. Two members of the atmospheric chemistry research group at the University of Leeds evaluated the prototype ELN; both evaluators regularly develop atmospheric chemistry models using the MCM. The evaluators of the ELN were not involved at any point during the design and development of the prototype ELN, so came to use and evaluate the ELN with minimal prior knowledge or preconceptions. The mode of evaluation was very much formative [32], seeking to elicit user responses on topics including: the efficacy of the ELN prototype, the benefits and drawbacks of using an ELN and ways in which provenance could be used once captured by an ELN. The evaluation explored the provenance capture and use scenarios, as well as the ELN prototype itself, using elements of semi-structured interview, discussion, prototype demonstration and user exploration of the prototype. This approach attempted to strike a balance between the interviewer's ability to respond to user feedback as it occurs and providing a structure that ensures important topics are addressed. In this paper we focus on the findings of the evaluation with regard to user-orientated provenance, in particular the mode of capturing annotations.

5.3.2 Evaluation Results

Prompting encourages good practice. During the design of the ELN the decision to implement inline annotation by prompting the user had caused two concerns: first, users may find the prompts an unwelcome interruption from getting on with their scientific process; secondly, would it be possible to design and implement the prompts to be sufficiently context sensitive to be useful to the modeller? The overall response to the prompts used in the prototype was positive:

"I think ... [prompting is] ... a good way of ... [capturing annotations] ... because otherwise you won't do it. It would be nice to be prompted when you are doing [the] analysis [of model output data]"

In the quote above the evaluator suggests that inline annotation prompts will encourage users to adopt good practice in their provenance capture, being driven by the prompts to record their annotations more frequently and in a more structured manner than with a traditional lab notebook. The inline annotation prompts were also perceived to encourage good practice in the modelling process itself, by encouraging the modeller to consider and record a justification for each change they make to the model:

"[The inline annotation prompts] will prompt you to change ... [the chemical mechanism] in an iterative [manner], ... [and make those changes in a] logical order; therefore ...[you] think in a more scientific way as well. Therefore speeding up the modelling process."

More Structure in annotations. The inline annotation prompts provide a single text field to enable annotation of changes to the chemical mechanism. Presenting a single text field to the user was intended to provide a flexible means of annotation, that mimicked the traditional lab-book. The feedback during the evaluation suggested that this minimal structuring of the annotation is not in line with the requirements of users. A number of suggestions were made regarding adding structure to the inline annotation prompts, including separate annotation fields for the scientific rationale for changing a given reaction and an associated literature reference:

"[It would be useful to have] Two text boxes, one [requesting a] ... justification and one [requesting a] ... reference."

1
2
3
4 It was also noted that the associated literature reference field would need to be optional, as on some
5 occasions the user maybe editing a reaction based on their own experience and knowledge rather than
6 based on literature information.

7 **Flexibility in annotation interface.** The evaluators identified the lack of flexibility in the annotation
8 interfaces as a significant drawback to using the ELN.

9 *"[The ELN prototype is] not tailored to what you want to write, some people might not find it as*
10 *useful as other people"*
11

12 In order to provide additional flexibility in the annotation interface, the evaluators felt it would be
13 beneficial to complement inline annotation of the scientific process, with: post-hoc annotation of the
14 scientific process; enabling annotations in forms other than text including digital objects (graphs etc.);
15 enabling the user to customise the annotation interface.

16
17 **Provenance Terminology.** The scientific terminology used in the provenance was well received by the
18 evaluators, who saw no need to amend any of the terminology or its mode of use. The terminology was
19 evaluated indirectly: the evaluators were presented with a series of provenance reports, for a predefined
20 experiment, and asked to review them. It proved difficult to engage the evaluators in discussion of the
21 relative merits of using terminology from their scientific domain versus domain independent terminology,
22 as the evaluators found the concept of domain independent terminology within their provenance records
23 difficult to relate to.
24

25 **6 Conclusions and Future Work**

26
27 In this paper we have presented a user-orientated approach to the capture of provenance for *in silico*
28 experiments. We have argued that the limitations of workflow systems in capturing provenance, for *in*
29 *silico* experiments, can be in part be addressed by learning from the current practices of scientists (who
30 have been involved in the capture of provenance for centuries) and the development and adaptation of the
31 ELN concept to the *in silico* domain. Elements of this user-orientated approach have been evaluated in a
32 case study that investigates provenance capture and representation, using an ELN, for the iterative
33 development of computational models in the atmospheric chemistry community. The user responses to the
34 our user-orientated approach were generally positive: inline annotation of the scientific process was well
35 received with the users perceiving benefits in terms of the quality of provenance captured and encouraging
36 good practice in iterative modelling development. The use of scientific terminology in the representation of
37 the provenance proved difficult to evaluate directly but the response to indirect evaluation of the
38 terminology was generally positive.
39

40 In light of the evaluation results, presented above, we will further develop the ELN prototype in the
41 following areas: first, rather than adopt a minimal approach to the structuring of annotations prompts, as in
42 the ELN development to date, more structure will be added to the annotation prompts to enable a finer
43 grain of information to be captured. Secondly, develop functionality to enable the user to add pre and post
44 hoc annotations, in addition to inline annotation, and explore how scientists make use of this combination
45 of annotation functionality. Thirdly, develop functionality to enable users to annotate their experiments
46 with respect to multiple frames of reference, and explore how scientists make use of this functionality.
47 Given the difficulty we had evaluating the use of scientific terminology in provenance representation we
48 will also perform a comparative evaluation, with members of the atmospheric chemistry community, of the
49 provenance records generated by the system-orientated approach of workflow systems and our user-
50 orientated approach.
51

52 Our work to date has focused on the capture and representation of provenance, whilst we have postponed
53 work developing functionality to query, using SPARQL, and leverage value from provenance records. In
54 our future work we will develop provenance query functionality, based on a set of queries and scenarios
55 specified by members of the atmospheric chemistry community. It will be here that the value of storing the
56 provenance captured by our ELN using semantic web technologies will be most apparent. We will also
57 explore the implications of integrating the provenance captured, using our user-orientated approach, with
58
59
60

1
2
3 elements of the wider semantic web and knowledge ecosystem. We will address the integration of *in silico*
4 experiment provenance with the metadata associated with journal publications (e.g. the Project Prospect
5 [33] a Royal Society of Chemistry project to provide enhanced semantic content for journal publications).
6 Another possibility is to link *in silico* experiment provenance with semantic representations of scientists,
7 e.g. Friend of a Friend [34], and address issues of building communities of interest.
8

9 The EUROCHAMP project [35] consists of a consortium of 12 laboratories throughout Europe, each
10 laboratory brings an atmospheric simulation chamber and associated experimental capability to the
11 consortium. The aim of the project is to develop the *in vitro* experimental, computational modelling and
12 data archiving infrastructure, required to enable pressing issues in atmospheric chemistry to be addressed
13 by developing understanding of specific chemical mechanisms. The EUROCHAMP computational
14 modelling infrastructure seeks to ensure that for each chamber experiment a computational model is
15 developed using the MCM, this has two benefits: facilitating the analysis of *in vitro* experimental data, to
16 produce scientific knowledge; and ensuring that the performance of the MCM is frequently tested. The
17 computational modelling infrastructure is currently being developed and includes a modelling and data
18 analysis environment, and a modelling web service. Provenance, for data generated by computational
19 models, will be captured using a re-engineered version of the current ELN prototype. In order to facilitate
20 sharing model output data and the associated provenance, i.e. the contents of the ELN, we will implement a
21 provenance and knowledge management architecture. We envisage that each researcher using an ELN will
22 be able to make sections of their ELN available to community, the security and sharing models for the ELN
23 have yet to be determined. The provenance and knowledge management architecture will enable querying
24 across the geographically distributed ELNs, and browsing of available ELN content, subject to the data
25 owner's security settings. We envision that adopting ELNs and sharing user-orientated provenance across
26 the EUROCHAMP community will improve existing practices and enable novel processes that deliver a
27 wide variety of benefits. These benefits include: enabling individual researchers to better manage their data
28 archives, so reducing the time spent searching for or repeating misplaced research; enabling researchers to
29 search across their community, composing queries in their own scientific terminology, for relevant *in silico*
30 experiments that could inform their current research; improving the quality of modelling taking place
31 across the community, both by providing better access to information and by encourage best practice using
32 inline annotation prompts. In a wide-ranging application, of our user-orientated approach to provenance,
33 MCM developers will be able to review, in detail not possible with current publication methods, the
34 performance of the MCM by reviewing provenance records and data stored in ELNs across the
35 EUROCHAMP community; this case is considered in our associated publications [3] [4].
36

37 Our ELN will be reengineered for use within the EUROCHAMP project, in order to provide the user
38 community with robust, production quality software. The ELN will then be disseminated from the MCM
39 website (<http://mcm.leeds.ac.uk/MCM/>), alongside a set of complementary modelling and data analysis
40 tools, to the full MCM user community. We anticipate that the reengineered ELN software will be available
41 from early 2010, and will continue to be developed as open source software, by the interested members of
42 the MCM user community. In the longer term we hope to integrated the software associated with the MCM
43 (i.e. our ELN, the modelling and data analysis tools), into an integrated modelling environment tailored to
44 the needs of the MCM user community. Once the ELN is embedded within the MCM user community we
45 will perform an in-depth evaluation of the adoption and benefits of the ELN and our user orientated
46 approach to provenance.
47

48 Beyond the atmospheric chemistry domain, we suggest that our user-orientated approach is widely
49 applicable to computational science led projects involving provenance. Where the core elements of our
50 user-orientated approach; the use of scientific terminology in provenance representation (in place or in
51 addition to generic, computationally orientated terminology), the use of inline provenance capture to
52 encourage researcher to record annotations, placing equal importance on the capture and representation of
53 process provenance and the associated scientific rationale; can be applied to ensure scientists actively
54 engage in and benefit from the provenance captured in e-Science applications. Transferability of our user-
55 orientated approach to provenance will therefore need to be evaluated across other scientific communities.
56

57 Acknowledgements

58 Thank you to Jeremy Frey and Nick Gibbons at the University of Southampton for their support and input.
59
60

Thank you also to Andrew Rickard and Jenny Young at the University of Leeds and Roberto Sommariva at NOAA, Boulder for providing experimental data and assistance with use of the MCM, and to David Allen at Leeds University Business School for assistance with the ELN evaluation methodology.

References:

1. Simmhan, Y., B. Plale, and D. Gannon, *A Framework for Collecting Provenance in Data-Centric Scientific Workflows*, in *Proceedings of the IEEE International Conference on Web Services*. 2006, IEEE Computer Society.
2. Braun, U., et al., *Issues in automatic provenance collection*, in *Provenance and Annotation of Data*. 2006, Springer-Verlag Berlin: Berlin. p. 171-183.
3. Martin, C., et al. *Semantically-Enhanced Model-Experiment-Evaluation Processes (SeMEEPs) within the Atmospheric Chemistry Community*. in *Provenance and Annotation of Data and Processes, Second International Provenance and Annotation Workshop*. 2008. Salt Lake City, UT, USA: Springer
4. Martin, C.J., et al. *Semantically enhanced provenance capture for chamber model development with a master chemical mechanism*. in *The environmental eScience revolution*. 2008: Philosophical Transactions of the Royal Society A.
5. Horrocks, I., P.F. Patel-Schneider, and F. van Harmelen, *From SHIQ and RDF to OWL: the making of a Web Ontology Language*. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2003. **1**(1): p. 7-26.
6. Eric, M., *An Introduction to the Resource Description Framework*. *Bulletin of the American Society for Information Science and Technology*, 1998. **25**(1): p. 15-19.
7. Schraefel, M.C., et al., *Breaking the Book: Translating the Chemistry Lab Book into a Pervasive Computing Lab Environment*, in *CHI 2004*. 2004, ACM Press: Vienna, Austria.
8. Simmhan, Y., B. Plale, and D. Gannon, *A survey of data provenance in e-science*. *SIGMOD Rec.*, 2005. **34**(3): p. 31-36.
9. Luc Moreau, et al., *Special Issue: The First Provenance Challenge*. *Concurrency and Computation: Practice and Experience*, 2008. **20**(5): p. 409-418.
10. Oinn, T., et al., *Taverna: a tool for the composition and enactment of bioinformatics workflows*. *Bioinformatics*, 2004. **20**(17): p. 3045-3054.
11. Zhao, J., et al., *Semantically linking and browsing provenance logs for e-science*, in *Semantics of a Networked World: Semantics for Grid Databases*. 2004, Springer: New York. p. 158-176.
12. Foster, I., *The virtual data grid: A new model and architecture for data-intensive collaboration*, in *Ssdmb 2002: 15th International Conference on Scientific and Statistical Database Management*, S. Nittel and D. Gunopulos, Editors. 2003, Ieee Computer Soc: Los Alamitos. p. 11-11.
13. Ludäscher, B., et al., *Scientific workflow management and the Kepler system*. *Concurrency and Computation: Practice and Experience*, 2006. **18**(10): p. 1039-1065.
14. Foster, I., et al., *Chimera: A virtual data system for representing, querying, and automating data derivation*, in *14th International Conference on Scientific and Statistical Database Management, Proceedings*, J. Kennedy, Editor. 2002, Ieee Computer Soc: Los Alamitos. p. 37-46.
15. Altintas, I., O. Barney, and E. Jaeger-Frank, *Provenance collection support in the Kepler Scientific Workflow System*, in *Provenance and Annotation of Data*. 2006, Springer-Verlag Berlin: Berlin. p. 118-132.
16. Zhao, J., et al., *Mining Taverna's semantic web of provenance*. *Concurrency and Computation-Practice & Experience*, 2008. **20**(5): p. 463-472.
17. Scheidegger, C., et al., *Tackling the Provenance Challenge one layer at a time*. *Concurrency and Computation-Practice & Experience*, 2008. **20**(5): p. 473-483.
18. Simmhan, Y.L., B. Plale, and D. Gannon, *Query capabilities of the Karma provenance framework*. *Concurrency and Computation-Practice & Experience*, 2008. **20**(5): p. 441-451.
19. Pignotti, E., et al., *Enhancing workflow with a semantic description of scientific intent*, in *Semantic Web: Research and Applications, Proceedings*, S. Bechhofer, et al., Editors. 2008, Springer-Verlag Berlin: Berlin. p. 644-658.
20. *ChemOffice*. [cited 20th August 2008]; Available from: <http://www.camsoft.com/>.
21. *SCRIP-SAFE*. [cited 20/08/2008]; Available from: http://www.scrip-safe.com/laboratory_notebooks.htm.
22. Amstein, L., et al., *Labscape: a smart environment for the cell biology laboratory*. *Pervasive Computing*, IEEE, 2002. **1**(3): p. 13-21.
23. Holliday, G.L., P. Murray-Rust, and H.S. Rzepa, *Chemical Markup, XML, and the World Wide Web. 6. CMLReact, an XML Vocabulary for Chemical Reactions*. *Journal of Chemical Information and Modeling*, 2006. **46**(1): p. 145-157.
24. Wakelin, J., et al., *CML tools and information flow in atomic scale simulations*. *Molecular Simulation*, 2005. **31**(5): p. 315-322.
25. Degtyarenko, K., et al., *ChEBI: a database and ontology for chemical entities of biological interest*. *Nucl. Acids Res.*, 2007: p. gkm791.
26. Sportisse, B., *Box models versus Eulerian models in air pollution modeling*. *Atmospheric Environment*, 2001. **35**(1): p. 173-178.
27. Blomberg, J., *Ethnography: aligning field studies of work and system design*, in *Perspectives on HCI: Diverse approaches*, A.F. Monk and N. Gilbert, Editors. 1995, Academic Press: London. p. 175-197.
28. Sommariva, R., et al., *OH and HO2 chemistry in clean marine air during SOAPEX-2*. *Atmos. Chem. Phys.*, 2004. **4**(3): p. 839-856.
29. Heller, S.R., S.E. Stein, and D.V. Tchekhovskoi, *InChI: Open access/open source and the IUPAC international chemical identifier*. *Abstracts of Papers of the American Chemical Society*, 2005. **230**: p. U1025-U1026.
30. Lv, T. and P. Yan, *An introduction to MathML and its applications*, in *2007 International Symposium on Computer Science & Technology, Proceedings*, S. Zhang, Editor. 2007, American Scholars Press: Marietta. p. 603-605.
31. Rosson, M.B. and J.M. Carroll, *Usability Engineering: Scenario-Based Development of Human-Computer Interaction*. 2002, San Francisco: Morgan Kaufmann.
32. Scriven, M., *Types of Evaluation and Types of Evaluator*. *American Journal of Evaluation*, 1996. **17**(2): p. 151-161.

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
33. *Project Prospect*. 2008 [cited 8th December 2008]; Available from: <http://www.rsc.org/Publishing/Journals/ProjectProspect/index.asp>.
34. Li, D., et al. *How the Semantic Web is Being Used: An Analysis of FOAF Documents*. in *System Sciences, 2005. HICSS '05. Proceedings of the 38th Annual Hawaii International Conference on*. 2005.
35. Wiesen, P., *The EUROCHAMP Integrated Infrastructure Initiative Environmental*, in *Environmental Simulation Chambers: Application to Atmospheric Chemical Processes*. 2006. p. 295-299.

For Review Only