A user-orientated approach to provenance capture and representation for \textit{in silico} experiments, explored within the atmospheric chemistry community

By Chris J. Martin$^{1,2,3,*}$, Mohammed H. Haji$^2$, Peter K. Jimack$^2$, Michael J. Pilling$^1$ and Peter M. Dew$^2$

$^1$School of Chemistry, and $^2$School of Computing, University of Leeds, Leeds LS2 9JT, UK
$^3$Environment Agency, Science Department, Lower Bristol Road, Bath BA2 9ES, UK

We present a novel user-orientated approach to provenance capture and representation for \textit{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, while 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: 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’ (Simmhan \textit{et al.} 2006). 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 that 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 \textit{in silico} experiment. Alongside these annotations, we capture process provenance (Braun \textit{et al.} 2006),

*Author and address for correspondence: Environment Agency, Science Department, Lower Bristol Road, Bath BA2 9ES, UK (christopher.martin@environment-agency.gov.uk).

One contribution of 16 to a Theme Issue ‘Crossing boundaries: computational science, e-Science and global e-Infrastructure II. Selected papers from the UK e-Science All Hands Meeting 2008’.

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 realized by individual researchers and the wider research community, including: enabling researchers to reduce the amount of time they spend interpreting or reinterpreting archived data (produced either by themselves or by third-party researchers); and 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 (Martin et al. 2008, 2009).

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, which 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; Martin et al. 2008) 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 the design and development of a production-quality, open-source, ELN for use by atmospheric chemistry modellers across an international research community. Our ELN places annotation opportunities within the scientific process executed by the computational modeller, in the form of prompts, while seeking to minimize changes to the scientific process. The ELN monitors the processes executed by the scientist both to capture provenance and to drive the annotation prompts. By placing the annotation prompts within the scientific process, we seek to capture the modeller’s reasoning as it takes place, mirroring the current practices of a scientist making notes in their laboratory book as they are going along. The process provenance captured by the ELN is represented using terminology from the scientific domain of interest: in this case study, atmospheric chemistry. We seek to understand and capture the science taking place rather than just recording the changes from a system orientation. For example, what could be viewed from a system orientation as a change to the last modified date of a model input file is from a science orientation a change to the scientific nature of the computational model. The provenance captured by the ELN is structured and stored using Semantic Web technologies, the Web Ontology Language (OWL; Horrocks et al. 2003) and the Resource Description Framework (RDF; Eric 1998), to enable the development of provenance-consuming Internet applications in our future work.

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

2. Approaches to provenance

(a) The scientist’s approach to provenance

Scientists have been capturing provenance, alongside the scientific data they generate, for centuries (Schraefel et al. 2004). The traditional means of capturing provenance has been the laboratory notebook (LN), used to capture both the experimental process (process provenance) and annotations relating to the scientists’ reasoning (annotations). While there are many drawbacks to capturing provenance using a LN, the ways in which scientists use their LN suggest three important user requirements for provenance capture and representation.

First, scientists capture provenance as they execute their experiments. We will refer to this as inline provenance capture. In addition, they capture provenance before (pre hoc) and after (post hoc) their experiments. Inline provenance capture is required to enable the scientist to capture process provenance and reasoning annotations as the scientific process evolves, and decisions are made, not necessarily adhering to an experimental plan.

Second, scientists make annotations relative to different frames of reference, dependent on the context of annotation. Frames of reference used include: the high-level experiment, where a scientist may wish to provide annotations incorporating experimental goals and conclusions; individual elements of the scientific workflow executed, e.g. the scientist may provide annotations incorporating reasons for changing an individual experimental parameter; and ad hoc, aggregations of workflows or workflow elements, e.g. a scientist may wish to define and annotate a set of sub-experiments that have taken place under a single main experiment. It is important to note for each frame of reference that scientists make annotations with a different content, detail and structure, i.e. the annotation of an experiment differs significantly from the annotation of changing a model parameter.

Third, scientists capture provenance using scientific terminology. The use of scientific terminology, specific to the domain of the experiment, enables a great deal of information to be recorded within the provenance in a concise manner (relying on a common understanding of the terminology).

(b) The systems-orientated approach to provenance for in silico experiments

Within the e-Science domain, research into provenance capture, representation and storage for in silico experiments has been tightly coupled with the workflow system’s (Simmhan et al. 2005; Luc et al. 2008) paradigm. For the purpose of comparison between the workflow approach to provenance and our user-orientated approach, we take the TAVERNA system (Oinn et al. 2004) as an exemplar from the workflow system paradigm. In reviewing the TAVERNA system, we consider two key characteristics.
First, TAVERNA (Zhao et al. 2004), in common with many other workflow systems (Foster 2003; Luda¨scher et al. 2006), seeks to automatically capture provenance for in silico experiments, minimizing user involvement. Automatic provenance capture is well suited to capture process provenance, i.e. the structure and execution of the workflow, but overlooks the importance of capturing the scientist’s contribution to the scientific process (e.g. why they used a given service, or why they have rerun a workflow with a modification to the input parameters). Within the TAVERNA workflow environment, user involvement is limited to annotating a given workflow or workflow component with a single high-level description; this annotation can be either pre hoc (before running the workflow) or post hoc (after running the workflow). Therefore, TAVERNA can be seen to lack support for inline annotations and provides limited support for annotating with respect to multiple frames of reference.

Second, the provenance captured by TAVERNA, as with many other workflow systems (Foster et al. 2002; Altintas et al. 2006), is represented using domain-independent semantics. So the scientific process (captured as a workflow/series of workflows), of a given researcher, is represented independently of the particular scientific domain. While the use of domain-independent semantics can be seen as an important factor in producing a domain-independent workflow system that is deployable across scientific domains, domain-independent semantics removes the opportunity to leverage the informational content of the scientific terminology of a given scientific domain. Given the key characteristics identified above, minimizing user involvement in provenance capture and using domain-independent semantics to represent provenance, the TAVERNA approach to provenance can be viewed as system orientated.

(c) A user-orientated approach to provenance for in silico experiments

The differences between the system-orientated (i.e. computer science driven) and the scientist’s approaches to provenance can be seen to be a result of cultural differences between the two communities. Our work seeks to develop a user-orientated approach to the capture of provenance, both process provenance and annotations, for in silico experiments. We attempt to reconcile the scientist’s and the system-orientated approaches to provenance capture, discussed above. From the system-orientated approach, we will seek to automate process provenance capture, while adopting the key practices from the scientist’s approach: inline annotation, annotations with respect to multiple frames of reference and the use of scientific terminology in the representation of provenance. So, while we seek to minimize user involvement in the capture of process provenance, we seek to engage the user in annotating their scientific process. By adopting this user-orientated approach, we can complement detailed process provenance, captured automatically, with a record of the scientist’s reasoning and leverage the informational content of the domain-specific scientific terminology.

(d) Related work

The first provenance challenge (Luc et al. 2008) sought to understand how a number of provenance systems address a benchmark provenance problem, with particular respect to: how provenance is represented; the ability of the provenance system to answer queries; and what is considered in scope for
provenance capture. The myGrid research group address the provenance challenge using TAVERNA plus a knowledge template (Zhao et al. 2008), which adds semantic annotation functionality. The knowledge template allows users to create annotations to enrich the domain-independent process provenance automatically captured by TAVERNA with semantics from a specific scientific domain. This is in contrast to our approach, where we capture process provenance, using semantics from a specific scientific domain, automatically. The VisTrails response to the first provenance challenge (Scheidegger et al. 2008) adopts a change-based approach to provenance, capturing the evolution of a workflow as a scientist conducts exploratory research. Provenance is captured, and annotation enabled, at three layers: workflow evolution; the workflow structure; and the workflow execution. In our approach, we take this one stage further, capturing changes in both the workflow and the input data, using scientific terminology. A number of provenance systems, including KARMA (Simmhan et al. 2008), applied to the first provenance challenge, considered annotations beyond the scope of the provenance research discipline. We view this as the extreme system-orientated perspective on provenance, completely eliminating the role of the scientist in provenance capture, which runs the risk of capturing provenance of limited value for the long-term archival of data. The extreme system-orientated approach produces provenance that describes how a given data item was produced, but none of the critical scientific information on why the data were produced in a certain way that our approach seeks to capture.

The importance of the scientist’s contribution to provenance has been recognized in the work of the PolicyGrid project, where they seek to capture the scientist’s intent as well as their method (Pignotti et al. 2008). PolicyGrid has taken the Kepler workflow environment (Altintas et al. 2006), and added functionality to capture and structure provenance that describes the intent of a scientist executing a workflow. This enables the scientist to annotate a workflow, and structure these annotations with use of ontology, with goals, reasoning, etc., whereas our approach seeks to capture annotations for the individual processes that composed a workflow in a context-sensitive fashion.

3. An ELN for iterative computational modelling

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

ELNs have typically been used for the capture of provenance for in vitro experiments (Schraefel et al. 2004) and provide an electronic replacement for the traditional LN, in which a scientist is able to record their experimental process alongside their reasoning and thoughts. ELNs have been developed and deployed extensively in commercial settings (ChemOffice, http://www.camsoft.com/; SCRIP-SAFE, http://www.scrip-safe.com/laboratory_notebooks.htm), such as drug development, where they provide a stronger basis than a traditional LN, for
intellectual property claims. ELNs have also been researched and deployed in a variety of academic settings (Arnstein et al. 2002), including the COMBECHEM ELN (Schraefel et al. 2004), an important reference point for our research. The COMBECHEM ELN is used to capture provenance for organic chemistry synthesis experiments, where the scientist typically performs a sequential set of actions (mixing chemicals together, heating or cooling mixtures, etc.) in a laboratory setting. The response to a prototype COMBECHEM ELN by potential users has been positive (Schraefel et al. 2004), during initial usability trials, and a production-quality ELN is currently being engineered (J. Frey 2008, personal communication). Process provenance is captured from the plan of the experimental process (a mandatory safety requirement prior to commencing all experiments), with amendments to the experimental process and annotations made at experimental run-time. A key difference between iterative computational modelling and in vitro experiments is that when modelling there is no need for an experimental safety plan (or any detailed plan whatsoever), so we seek to capture process provenance automatically from the individual computational processes.

4. Case study background

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

(a) Atmospheric chemistry community

Atmospheric chemistry is an inherently multiscale science, incorporating a variety of field, in vitro and in silico experimental disciplines. At the global and regional scales, the atmospheric chemistry community is involved in a number of high-profile modelling activities including: modelling of global concentrations of methane and ozone, which, after CO$_2$, are the trace gases with the greatest influence on climate change; and developing models that inform the development of air quality policy. A central component of the models investigating atmospheric chemistry on a global or regional scale is the chemical mechanism. Chemical mechanisms, part of the molecular scale of atmospheric chemistry study, consist of a coupled set of steps called elementary reactions in which chemical species are interconverted (i.e. mechanisms are lists of chemical reactions). Each elementary reaction can be considered in the form:

\[ \text{reactants} \xrightarrow{k} \text{products}, \]

where the reactants are the set of chemical species that react together to generate the products (another set of chemical species) and $k$ is the rate coefficient of the reaction. Elementary reactions are investigated primarily in the laboratory;
detailed chemical mechanisms are constructed from knowledge of these elementary reactions and their interactions. Mechanisms are used directly to construct models containing a very large set of ordinary differential equations that represent the rates at which the concentrations of individual species in the mechanism change with time. Such models are used for problems with modest fluid dynamic requirements, e.g. local-scale modelling, in order to test the performance of the chemical mechanism. These mechanisms can contain a large number of elementary reactions, often in excess of 10,000, and so are too computationally expensive to implement within global and regional models, e.g. for aspects of climate change or regional air quality. In such cases, mechanisms of much lower dimension are used, ideally based on objective lumping of the detailed mechanisms, providing a link between the global and regional scale models, and fundamental chemical kinetics. Research on elementary reactions and chemical mechanisms is conducted in research laboratories throughout the world. The Master Chemical Mechanism (MCM) is the leading detailed chemical mechanism, used across the international research community, and describes the chemistry occurring in the lower atmosphere. It is used both directly in local-scale models and to evaluate smaller lumped mechanisms used in global and regional atmospheric models. Within the wider chemistry community, a great deal of effort has been committed to the development of schemas and ontology for representing chemical data, including the Chemical Markup Language (CML; Wakelin et al. 2005; Holliday et al. 2006) and Chemical Entities of Biological Interest (ChEBI; Degtyarenko 2007) projects. Up until to this point, efforts have focused on describing the structural properties of atoms and molecules, with neither of the aforementioned projects addressing the representation of mechanisms nor the processes involved in in silico atmospheric chemistry experiments, of the type we consider in this paper.

(b) Atmospheric chemistry models

Computational modelling takes many forms within the atmospheric chemistry community, as described above. In this paper, we focus on recording the provenance for one form in particular, the so-called zero-dimensional box models (Sportisse 2001), where the aim of modelling is to develop an understanding of the chemical processes taking place at a given location (i.e. the local scale). Field and in vitro experiments at the local scale, including field campaigns that make in situ measurements at a single location, and experiments in atmospheric simulation chambers, can be modelled using zero-dimensional box models, incorporating the MCM. The output of these local-scale models can then be compared with the field or in vitro experiment data (as appropriate), in order to test the performance of the MCM. In this case, the modeller will make use of experimental data and various in situ measurements of chemical concentrations, and vary the configuration of the model to compare in vitro experimental data with model output data. During this process, the modeller will extensively experiment with the chemical mechanism implemented within the model, adding, deleting or changing chemical reactions and testing the impact that this has on the model output (validated against the aforementioned in situ measurements).

Phil. Trans. R. Soc. A (2009)
The model development process we consider in this paper is iterative, with the changes made to the mechanism determined by the conclusions drawn when comparing the model output with experimental data. Typically, the modeller does not form a detailed plan of action, instead working in an exploratory manner, drawing on their own knowledge and experience, in conjunction with the conclusions they draw from the comparison of model output and experimental data. This method of working has a significant implication for provenance capture: it places a premium on capturing the modeller’s reasoning and thoughts alongside the details of the modelling workflow. We seek to address this within our user-orientated approach to provenance.

5. Case study

(a) Requirements capture and design methodology

Given the focus of our work on adopting a user-orientated approach to provenance, an ethnographic methodology (Blomberg 1995) was adopted to ensure that the requirements and motivations of modellers within the atmospheric chemistry community could be understood. One author, CJM, was embedded within the atmospheric chemistry modelling group at the University of Leeds. Prior to, and throughout, the development of the ELN, he worked on atmospheric chemistry modelling projects, seeking to deliver atmospheric chemistry research while developing personal insight into the scientific processes, motivation and provenance requirements of atmospheric chemistry modellers.

Capturing the modelling process used by atmospheric chemistry modellers was the first phase of developing the ELN prototype. The process capture was facilitated by considering a modelling case study based on the development of a model for a field campaign that took place in Tasmania, SOAPEX (Sommariva et al. 2004). The SOAPEX field campaign made measurements of: free-radical species concentrations, including OH and HO$_2$; environmental conditions, including photolysis rates, temperature and pressure; and concentrations of other important chemical species, including O$_3$, CO, NO, NO$_2$ and a variety of hydrocarbons. The campaign took place at Cape Grim, Tasmania, Australia, in extremely clean air conditions ([NO]$<3$ ppt). Subsequently, zero-dimensional box models were developed to enable model–experiment comparisons for HO$_x$ radicals with insight developed into the chemistry of HO$_x$ radicals in clean air. The model in the case study was relatively simple, but also retained all the key characteristics of more complex models. The process for developing the SOAPEX model was then mapped, at the finest granularity of task description possible, to produce a process description for the case study. The importance of capturing process at such fine granularity is that only with this level of detail is it possible to repeat an experiment (either modelling or laboratory based). The case study process description was then examined to develop a provenance specification. This provenance specification was developed from an end-user perspective, in the form of a set of provenance reports for the case study modelling process. The subsequent design and implementation of the prototype were guided by this provenance specification.
Prototype implementation

In this paper, we consider two aspects of the prototype implementation: first, the scientific terminology used in the representation of the provenance, in the form of ontology; second, the interaction patterns between the user and the ELN during inline annotation. Further details of the design and implementation of the prototype ELN are provided elsewhere (Martin et al. 2008).

(i) The use of scientific terminology in provenance representation

As a starting point for the development of our ontology, we took the COMBECHEM ELN ontology (Schraefel et al. 2004), designed to structure provenance for in vitro chemistry experiments. Our ontology shares the same set of top-level concepts, from which all other concepts are inherited, with the COMBECHEM ontology. The top-level concepts, in both ontologies, are processes and materials; below this level we have developed domain-specific and domain-independent elements of ontology as required during the development of the prototype ELN. In this section, we explore the ontology developed to capture the changes made to the chemical mechanism within a zero-dimensional box model. Our ontology is expressed using OWL, with the provenance generated as RDF conforming to the ontology.

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

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, and the model configuration is then realized within the model execution process. Model output data are produced by the model execution process, and are an input to the data analysis process, which outputs a set of conclusions about the impact of the change to the model configuration.

(b) Prototype implementation

In this paper, we consider two aspects of the prototype implementation: first, the scientific terminology used in the representation of the provenance, in the form of ontology; second, the interaction patterns between the user and the ELN during inline annotation. Further details of the design and implementation of the prototype ELN are provided elsewhere (Martin et al. 2008).
These concepts are sufficiently specific for the atmospheric chemistry modellers to relate to, but can be decomposed further to allow more scientific detail to be included in the process provenance capture by our ELN. So we continued to develop the ontology at lower conceptual levels.

Taking for example the decomposition of the ‘mechanism development’ process, the modeller can perform a wide variety of operations on the mechanism (see figure 2), including adding, deleting and editing reactions. The ontology also includes the decomposition of the ‘edit reaction’ process (edit reactants, edit products and 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.

(ii) 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 LN.

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

$$\text{CH}_3\text{OH} + \text{OH} + (\text{O}_2) \underset{k}{\text{\rightarrow}} \text{HO}_2 + \text{HCHO} + \text{H}_2\text{O},$$

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

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

(iii) 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 3 and the discussion below.

(iv) 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 4 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.
(v) 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’.

(vi) Upon completion of the prompt, the model runs within the modelling environment.

Figure 3 shows a simplified representation of the semantic provenance generated for the sequence above. The 'add reaction' process has two inputs, a chemical mechanism (mechanism\textsubscript{pre-update}) and a reaction (R1), and one output, a revised chemical mechanism (mechanism\textsubscript{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 names for each of the chemical species involved in the reaction are compared with a reference database, enabling an International Chemical Identifier (InChI; Heller \textit{et al.} 2005; non-proprietary identifier for chemical substances) to be used within the provenance. The rate coefficient is captured as a text string. Work in progress on a related project looks at representations of chemical rate coefficients using mathML (Lv \& Yan 2007) and CML (Holliday \textit{et al.} 2006); once this task is completed, it will be considered for incorporation into our work. Each mechanism is identified by a Uniform Resource Identifier (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.

(c) Prototype evaluation methodology

To evaluate the ELN prototype system, we adopted an approach that draws on the scenario-based development paradigm (Rosson \& Carroll 2002). 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 (Scriven 1996), 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.

(d) Prototype evaluation results

(i) 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; second, 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 would not 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 LN. 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.

(ii) 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 LN. 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.
It was also noted that the associated literature reference field would need to be optional, as on some occasions the user may be editing a reaction based on their own experience and knowledge rather than based on literature information.

(iii) **Flexibility in annotation interface**

The evaluators identified the lack of flexibility in the annotation interfaces as a significant drawback to using the ELN.

[The ELN prototype is] not tailored to what you want to write, some people might not find it as useful as other people.

In order to provide additional flexibility in the annotation interface, the evaluators felt that it would be beneficial to complement inline annotation of the scientific process by: allowing *post hoc* annotation of the scientific process; enabling annotations in forms other than text including digital objects (graphs, etc.); and enabling the user to customize the annotation interface.

(iv) **Provenance terminology**

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

6. **Conclusions and future work**

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

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

Our work to date has focused on the capture and representation of provenance, while we have postponed work developing functionality to query (using SPARQL) and leverage value from provenance records. In our future work, we will develop provenance query functionality, based on a set of queries and scenarios specified by members of the atmospheric chemistry community. It will be here that the value of storing the provenance captured by our ELN using Semantic Web technologies will be the most apparent. We will also explore the implications of integrating the provenance captured, using our user-orientated approach, with elements of the wider Semantic Web and knowledge ecosystem. We will address the integration of in silico experiment provenance with the metadata associated with journal publications (e.g. the Project Prospect, http://www.rsc.org/Publishing/Journals/ProjectProspect/index.asp, a Royal Society of Chemistry project to provide enhanced semantic content for journal publications). Another possibility is to link in silico experiment provenance with semantic representations of scientists, e.g. Friend of a Friend (Li et al. 2005), and address issues of building communities of interest.

The EUROCHAMP project (Wiesen 2006) consists of a consortium of 12 laboratories throughout Europe, each laboratory bringing an atmospheric simulation chamber and associated experimental capability to the consortium. The aim of the project is to develop the in vitro experimental, computational modelling and data-archiving infrastructure required to enable pressing issues in atmospheric chemistry to be addressed by developing understanding of specific chemical mechanisms. The EUROCHAMP computational modelling infrastructure seeks to ensure that, for each chamber experiment, a computational model is developed using the MCM, which has two benefits: facilitating the analysis of in vitro experimental data, to produce scientific knowledge; and ensuring that the performance of the MCM is frequently tested. The computational modelling infrastructure is currently being developed, and includes a modelling and data analysis environment and a modelling Web service. Provenance, for data generated by computational models, will be captured using a re-engineered version of the current ELN prototype. In order to facilitate sharing model output data and the associated provenance, i.e. the contents of the ELN, we will implement a provenance and knowledge management architecture. We envisage that each researcher using an ELN will be able to make sections of their ELN available to the community; the security and sharing models for the ELN have yet to be determined. The provenance and knowledge management architecture will enable querying across the geographically distributed ELNs, and browsing of available ELN content, subject to the data owner’s security settings. We envisage that adopting ELNs and sharing user-orientated provenance across the EUROCHAMP
community will improve existing practices and enable novel processes that deliver a wide variety of benefits. These benefits include: enabling individual researchers to better manage their data archives, so reducing the time spent searching for or repeating misplaced research; enabling researchers to search across their community, composing queries in their own scientific terminology, for relevant \textit{in silico} experiments that could inform their current research; and improving the quality of modelling taking place across the community, both by providing better access to information and by encouraging best practice using inline annotation prompts. In a wide-ranging application of our user-orientated approach to provenance, MCM developers will be able to review, in detail not possible with current publication methods, the performance of the MCM by reviewing provenance records and data stored in ELNs across the EUROCHAMP community; this case is considered in our associated publications (Martin \textit{et al.} 2008, 2009).

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

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

The authors wish to thank Jeremy Frey and Nick Gibbons at the University of Southampton for their support and input; Andrew Rickard and Jenny Young at the University of Leeds and Roberto Sommariva at NOAA, Boulder, Colorado, for providing experimental data and assistance with the use of the MCM; and David Allen at Leeds University Business School for assistance with the ELN evaluation methodology.

References


