

# Semantically enhanced provenance capture for chamber model development with a master chemical mechanism

BY CHRIS J. MARTIN<sup>1,\*</sup>, MOHAMMED H. HAJI<sup>2</sup>, PETER M. DEW<sup>2</sup>,  
MICHAEL J. PILLING<sup>1</sup> AND PETER K. JIMACK<sup>2</sup>

<sup>1</sup>*School of Chemistry, and* <sup>2</sup>*School of Computing, University of Leeds,  
Leeds LS2 9JT, UK*

The development and maintenance of benchmark databases within scientific communities is reliant on interactions with database users. We explore the role of semantically enhanced provenance for computational modelling processes that make use of one such database: the master chemical mechanism, a key resource within the atmospheric chemistry community.

**Keywords:** provenance; Semantic Web; tropospheric chemistry; chamber models; chamber experiments

## 1. Introduction

The master chemical mechanism (MCM) is a benchmark database within the atmospheric chemistry community and provides a near-explicit description of the chemical reactions that take place in the troposphere. The MCM was built, and is revised, based upon the current state of evaluated chemical reaction data (rate constants and product yields) and chemical reaction data from the wider literature, according to the protocol of Saunders *et al.* (2003). The MCM can be incorporated into a number of types of atmospheric model including chamber models, seeking to understand the chemistry taking place in atmospheric simulation experiments. The MCM provides an essential link between laboratory measurements and atmospheric models. It has been described as the ‘gold standard’ against which application-specific mechanisms, e.g. the chemical mechanisms in global atmospheric models, can be evaluated.

Chamber model–experiment evaluation processes (CMEEPs), where comparisons between chamber models (incorporating the MCM) and chamber experiments are performed, are reported by the standard mechanisms of scientific publication or retained as a local resource of the data owner. This comparison activity is performed by a number of research groups across Europe, as part of the EUROCHAMP project (Barnes & Rudzinski 2006). Currently, the data and provenance produced by CMEEPs are recorded in an unstructured, ad

\* Author for correspondence (c.j.martin05@leeds.ac.uk).

One contribution of 24 to a Discussion Meeting Issue ‘The environmental eScience revolution’.

hoc manner using a disparate set of media such as text files (in a variety of formats), the laboratory notebook, word processor documents and annotations within data analysis files.

The CMEEPs conducted by the community have important implications for the ongoing development of the MCM; they can highlight aspects where further development is required or validate aspects of the MCM. The conclusions of CMEEPs often remain difficult for the MCM developers to access, and there is a lack of the associated provenance required to validate the conclusions. It is this issue that we seek to address, by designing, developing and evaluating an architecture to support the development of the MCM by leveraging the full value of the CMEEPs conducted by the community. In this paper, we discuss the first stage of this work: the capture of data and provenance for CMEEPs in a semantically enhanced form, using an electronic laboratory notebook (ELN), with the provenance represented and stored using Semantic Web technologies (RDF and OWL).

## 2. Capturing the provenance

We seek to encourage chamber modellers to capture the provenance for their CMEEPs, by highlighting the value of the provenance to support their own working practices, rather than to rely on the development of the MCM as a motivation. To this end, having adopted a scenario-based development methodology (Rosson & Carroll 2002), we developed scenarios to explore the requirements of a chamber modeller:

- the capture of provenance at model development time (case 1) and
- the use of provenance to help write a PhD thesis (case 2).

Case 1 was explored to understand the model development process that chamber modellers perform. We focused on the typical modelling process of a chamber modeller, with a relatively short time being spent on the model configuration (location of the chamber, date and time of experiment, etc.), followed by a more extensive experimentation with the chemical mechanism incorporated into the model. The experimentation involves iterations over the following processes: mechanism development, editing chemical reactions (where, for example, the MCM does not contain the latest literature value for a rate parameter), adding reactions to, and deleting reactions from the mechanism, etc; model execution; analysis of model output, drawing conclusions and making plans for future iterations. Case 2 was explored in order to understand the provenance requirements of, and the value of provenance to, the individual modeller after the time of capture, when returning to their own work and seeking to reinterpret results for publication.

Ontology that provides a vocabulary for structuring the provenance captured by the ELN was developed to describe atmospheric chemistry modelling experiments. Our ontology builds upon the CombeChem ELN ontology (Frey et al. 2004) for *in vitro* chemistry experiments. In the ontology, we add domain-specific concepts and develop a three-layer model for the provenance captured by the ELN. Each layer presents the model development process at a different level of abstraction, which is as follows.

- *Experiment*. The CMEEP is viewed as an *in silico* experiment, with a plan and a set of conclusions.
- *Iteration*. The CMEEP is viewed in terms of the intermediate plans and conclusions formed in the course of the iterative modelling process.
- *Modelling*. The CMEEP is viewed as a workflow of modelling processes, typically including the mechanism development, model execution and analysis stages discussed above.

The ELN is integrated with the existing modelling tools for developing chamber models using the MCM, which include a Fortran numerical model and a series of data manipulation scripts. Therefore, the user is presented with a standard interface to their modelling tools, similar to that which is typically employed, but with each stage of the modelling process enhanced as follows.

- *Mechanism development*. When the mechanism is changed by the user, they are automatically prompted to comment on the changes they have made, providing justifications and literature references when appropriate. All comments and changes to the mechanism are recorded in the provenance.
- *Model execution*. All model input and output files are automatically stored in a MySQL database. Performance metrics, such as model runtime, user comments and file locations, are recorded in the provenance.
- *Analysis*. The user is presented with a standard interface to record the data sources used in the analysis process and any conclusions and plans in the provenance.

### 3. Evaluation

We conducted an evaluation of the prototype ELN with potential users who are members of the atmospheric chemistry modelling group at the University of Leeds. The evaluation began with discussions about provenance capture, during which the evaluators identified the main potential barrier to the adoption of an ELN, or other provenance capture tools, as the amount of user input required at the time of modelling. The evaluation continued with a hands-on user test of our prototype ELN, during which the evaluators found that the amount of user input required by the prototype was not a burden to them. When asked whether they would use an ELN requiring a similar amount of user input to the prototype, the response was positive:

[Yes,] I think it would be a good thing. I don't think it is too much extra ... work.

The users intuitively grasped the benefits of recording provenance using an ELN and that the benefits would be realized after the time of modelling by a number of stakeholders:

if someone else wants to look at ... [your provenance], that's great because the person can see exactly what you have done, where you have been and where to go next. And for yourself, if you are writing up a PhD... [you can] ... see exactly what you've done whereas currently you have to rifle through lab-books to see exactly what you have done.

The main issue raised by the evaluators, during the user testing of the prototype, was a lack of flexibility in the ELN interfaces and functionality provided for user annotation:

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

#### 4. Conclusions and future work

The evaluation of our prototype ELN suggests that the amount of user input required by the ELN does not place excessive burden on the user, owing to the automation of much of the provenance capture. The evaluators could see sufficient value in the provenance captured by the ELN to envisage cases where it would be of benefit to themselves and other community members. Issues were also raised with respect to the flexibility of the prototype ELN; we will seek to understand and address these issues in the next iteration of prototype development.

Our future work will explore an architecture to support the MCM development process by leveraging the semantically enhanced provenance of CMEEPs. At the heart of this process is the ELN capture of CMEEP provenance, discussed in this paper. A scientist's personal ELN archive will store a complete record of provenance and data for a given scientist. A scientist can then choose to make their data and provenance available in a laboratory archive, which can be referenced in publications and accessed by the MCM development group. The MCM development group can use the information contained in laboratory archives, along with chemical reaction data, to revise and improve the MCM, which, in turn, can be used as an input for further CMEEPs.

We would like to thank Jeremy Frey and Nick Gibbons at the University of Southampton for their support and input, and also Andrew Rickard and Jenny Young at the University of Leeds and Roberto Sommariva at NOAA, Boulder, for providing experimental data and assistance with the use of the MCM.

#### References

- Barnes, I. & Rudzinski, K. J. 2006 The EUROCHAMP integrated infrastructure initiative. In *Environmental simulation chambers: application to atmospheric chemical processes*, vol. 62 (ed. P. Wiesen), pp. 295–299. Berlin, Germany: Springer.
- Frey, J. G., Hughes, G. V., Mill, H. R., Schraefel, M. C., Smith, G. M. & De Roure, D. 2004 Less is more: lightweight ontologies and user interfaces for smart labs. In *UK eScience All Hands Meeting, Nottingham, UK*.
- Rosson, M. B. & Carroll, J. M. 2002 *Usability engineering: scenario-based development of human-computer interaction*. San Francisco, CA: Morgan Kaufmann.
- Saunders, S. M., Jenkin, M. E., Derwent, R. G. & Pilling, M. J. 2003 Protocol for the development of the master chemical mechanism, MCM v3 (Part A): tropospheric degradation of non-aromatic volatile organic compounds. *Atmos. Chem. Phys.* **3**, 161–180.