CHEMICAL MODELS FOR, AND THE ROLE OF DATA AND PROVENANCE IN, AN ATMOSPHERIC CHEMISTRY COMMUNITY

Chris James Martin

Submitted in Accordance with the Requirements for the Degree of Doctor of Philosophy

The University of Leeds
School of Chemistry
School of Computing

March 2009

The candidate confirms that the work submitted is his/her own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated overleaf. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Thesis content also presented in jointly-authored publications

The research presented in Chapters 5 to 10 of this thesis (describing the design, development and evaluation of an Electronic Laboratory Notebook) has been summarised and presented in the following jointly authored publications.

Martin, Chris J; Haji, Mohammed H; Dew, Peter M; Pilling, Michael J; Jimack, Peter K. *Semantically enhanced provenance capture for chamber model development with a master chemical mechanism*. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, vol. 367, pp. 987-990. 2009.

Martin, Chris; Haji, Mohammed H; Dew, Peter M; Pilling, Michael J; Jimack, Peter K. *Semantically-enhanced model-experiment-evaluation processes (SeMEEPs) within the Atmospheric Chemistry Community* in: Juliana Freire, David Koop & Luc Moreau (editors) **Provenance and Annotation of Data and Processes, Second International Provenance and Annotation Workshop, IPAW 2008**, pp. 293-308 Springer. 2008.

[to appear] Martin, Chris; Haji, Mohammed H; Jimack, Peter K; Pilling, Michael J; Dew, Peter M. *A User-Orientated Approach to Provenance Capture and Representation for in silico Experiments: Explored within the Atmospheric Chemistry Community.* in **Crossing Boundaries: Computational Science, E-Science and Global E-Infrastructure**. 2009: Philosophical Transactions of the Royal Society A.

The work in each of these papers was conducted primarily by Chris Martin, with technical software development contributions from Mohammed H. Haji. The major supervision for this work was provided by Peter M. Dew, with Peter K. Jimack providing supplementary supervision. Peter K. Jimack provided the main supervision for the development of the atmospheric chemical model (described in Chapters 3 and 4) that is closely coupled to the Electronic Laboratory Notebook (described in the papers). Atmospheric chemistry provides the application area for the work and the context of these papers. Michael J. Pilling provided the supervision for this aspect of the work, and members of his research group acted as users of the work. Chris Martin wrote the first drafts of each of these papers, organized subsequent drafts, is the first named author and is named as the author to whom correspondence should be addressed.

“If it can be broke then it can be fixed, if it can be fused then it can be split

It's all under control

If it can be lost then it can be won, if it can be touched then it can be turned

All you need is time

We promised the world we'd tame it, what were we hoping for?

A sense of purpose and a sense of skill, a sense of function but a disregard

We will not be the first, we won't”

From The Pioneers by Russell Lissack, Gordon Moakes, Kele Okereke and Matt Tong (2005).

Acknowledgements

So first and foremost, I would like to thank my supervisors Michael Pilling, Peter Jimack and Peter Dew. It has been tricky at times to balance the input and competing perspectives you have brought to the project, but hopefully I have struck the right balance and my thesis is all the better for resolving this creative tension. Thank you again for your energy, commitment, interest and support throughout the course of my studies. It has been a privilege to work with you over the last three years.

Thank you to all those who directly contributed to the work presented in my thesis: Mohammed H. Haji, Katarzyna Borońska, Monica Vazquez-Moreno and David Allen. I really appreciate the time you have invested in our research; and I have thoroughly enjoyed working with you all. Thank you to the many members of the atmospheric chemistry research community, at the University of Leeds, and out in the wider world, who informed and influenced my work, particularly Andrew Rickard, Jenny Stanton, Roberto Sommariva and Louise Whitehouse. Also, thank you to all the members of the Web Science and Scientific Computing research groups, at the University of Leeds, who provided support for, and alternative perspectives on, my research.

And finally on the work front, thank you to my friends and colleagues who made my days in the office more interesting and entertaining, particularly: Andrew, thank you for the music, showing me how to live the dream and talking to me like I understand chemistry; and Kelly, thank you for helping me to write emails and for being a civilising influence in the office!

I would also like to thank the people who encouraged me to start my academic odyssey: Alec McGuire, Dave Crane and Felicity Jay; it turned out to be a great decision, but for very different reasons to those I anticipated. Thank you to my friends, from outside the world of academia, for their refreshing lack of interest in my PhD and reminding me of more important things. I feel I must mention specifically Carmen (for only asking what I was studying after two and a half years of knowing me) and Hannah (I look forward to your thesis, I am sure it will be a more interesting read than mine!).

Thank you to everyone at Roundhay Lawn Tennis Club (particularly Chris Harper, Clive, Rob and everyone who has come to the drills sessions), for providing me with a refuge on court. I am so pleased that I have had the chance to contribute to the club[1] that has been such a fixture of my life for the past 15 years.

Mum, Ellen, Beth and Dad, as always you have been amazing throughout the last three and bit years. It has been fantastic to have had the chance to spend time with you and get to know you better, thank you for the meals, stories and banter.

And finally, Frin, thank you for all your love and support, for brightening my days, and knowing when to give me a good kick up the backside!

[1] Undoubtedly doing my PhD has given me the time and flexibility to make this contribution.

Abstract

This thesis presents research at the interface of the e-Science and atmospheric chemistry disciplines. Two inter-related research topics are addressed: first, the development of computational models of the troposphere (i.e. *in silico* experiments); and secondly, provenance capture and representation for data produced by these computational models. The research was conducted using an ethnographic approach, seeking to develop in-depth understanding of current working practices, which then informed the research itself. The research focused on the working practices of a defined research community; the users and developers of the MCM (Master Chemical Mechanism). The MCM is a key data and information repository used by researchers, with an interest in atmospheric chemistry, across the world.

A computational modelling system, the OSBM (Open Source Box Model) was successfully developed to encourage researchers to make use of the MCM, within their *in silico* experiments. Taking advantage of functionality provided by the OSBM, the use of *in situ* experimental data to constrain zero dimensional box models was explored. Limitations of current methodologies for constraining zero dimensional box models were identified, particularly associated with the use of piecewise constant interpolation and the averaging of constraint data. Improved methodologies for constraining zero dimensional box models were proposed, tested and demonstrated to offer gains in the accuracy of the model results and the efficiency of the model itself.

Current data generation and provenance related working practices, within the MCM community, were mapped. An opportunity was identified to apply Semantic Web technologies to improve working practices associated with gathering and evaluating feedback from *in silico* experiments, to inform the ongoing development of the MCM. These envisioned working practices rely on researchers, performing *in silico* experiments, that make use of the MCM, capturing data and provenance using an ELN (Electronic Laboratory Notebook). A prototype ELN, employing a user-orientation approach to provenance capture and representation, was then successfully designed, implemented and evaluated. The evaluation of this prototype ELN highlighted the importance of adopting a holistic approach to the development of provenance capture tools and the difficulties of balancing researchers' requirements for flexibility and structure their scientific processes.

Table of Contents

List of Tables

List of Figures

Glossary of Terms

ELN:	Electronic Laboratory Notebook, a computer-based tool for the capture of data and provenance.
EUROCHAMP:	integration of EUROpean simulation CHAMbers for investigating atmospheric Processes.
EXACT:	Effects of the oXidation of Aromatic Compounds in the Troposphere, a series of chamber experiments focussed on developing understanding of the degradation of aromatic compounds in the troposphere.
In silico experiments:	Experiments that simulate physical systems using computational resources (e.g. developing a computational model of the chemistry taking place in the troposphere).
In situ experiments:	Experiments that take place in the field (i.e. outside the controlled environment of the laboratory).
In vitro experiments:	Experiments that take place in a laboratory setting.
MCM:	The Master Chemical Mechanism, a quantitative description of the complex chemical processes taking in the troposphere. The MCM is a key information resource used across the atmospheric chemistry community.
OSBM:	Open Source Box Model, a modelling system for MCM users, the development of which is described in Chapter 3.
SMD:	Semantic MetaData, a description of some data (i.e. metadata) expressed using Semantic Web standards.

SOAPEX:	Southern Ocean Atmospheric Photochemical EXperiment, a field campaign focussed on developing understanding of the chemistry of clean air.
TORCH:	Tropospheric ORganic CHemistry experiment, a field campaign focussed on developing understanding of the chemistry of air polluted by anthropogenic emissions.

Chapter 1 Introduction

At the highest level the goal of my studies was to investigate the application of e-Science technologies, methodologies and approaches within the atmospheric chemistry research community. Where:

> "The term e-Science denotes the systematic development of research methods that exploit advanced computational thinking" [1].
>
> *And the atmospheric chemistry community consists of researchers (predominantly in the academic domain), developing understanding of the composition of, and chemical processes taking place within, the Earth's atmosphere.*

Achieving this goal required the adoption of a multidisciplinary approach, developing understanding of both the e-Science and atmospheric chemistry domains in parallel. Developing this understanding led to refinement of the high-level goal to one sufficiently well defined and constrained to be addressed within this thesis:

> *To develop tools that support better use of data in "in silico" experiments, within a specific atmospheric chemistry community. Addressing two issues: first, the use of "in situ" experimental data in computational models; and secondly, the capture and representation of provenance for data generated by computational models.*

This goal statement sets out the scope for the research presented in this thesis and is examined in further detail below.

- **Tool development:** This research focuses on the development of computational and information management tools to support the scientific activities of the atmospheric chemistry community.
- ***In silico* experiments:** The scope of the research was restricted to *in silico* experiments (i.e. computational modelling research). *In vitro* and *in situ* experimental experiments are not considered here.
- **A specific community:** The sub-community, within the wider atmospheric community, considered in this research is an MCM-centric community. This

community consists of the users and developers of the MCM (Master Chemical Mechanism), available from http://mcm.leeds.ac.uk/MCM/. The MCM is a core data and information resource, which provides a benchmark description of the chemistry taking place in the troposphere.

- **Incorporating experimental data into computational models:** *In vitro* and *in situ* experimental data play a critical role in the configuration of the computational models used in atmospheric chemistry research. Establishing and maintaining a link between the computational modelling and experimental domains, is a critical issue within atmospheric chemistry.
- **Provenance for data generated by computational models:** Provenance, more extensively defined in Chapter 5, can be considered as a description of how and why a given piece of data was created. The capture and representation of provenance for data generated by computational models is an active area of research in e-Science and, prior to this research, has not been extensively addressed in the atmospheric chemistry domain.

1.1 Research approach

This section describes the research approach that underpins the research presented in this thesis. The research approach is described in two parts: first, its multidisciplinary and ethnographic nature; and, secondly, the manner in which research objectives emerged. Prior to considering the research approach itself, my background is noted since this played an important role in determining and defining the research approach. My training and background has been based in computer science and information systems, with only a very basic understanding of the chemistry and atmospheric science domains.

A multidisciplinary, ethnographic approach: The research conducted was inherently multidisciplinary, seeking to make contributions to both the e-Science discipline and the atmospheric chemistry discipline. A prerequisite to making these contributions was to develop an understanding of: the breadth of research taking place across the atmospheric chemistry community; the language and terminology used by members of the community; and, the details of the processes involved in computational modelling research closely associated with the MCM. In order to develop this understanding of the atmospheric chemistry domain an ethnographic approach [2] was adopted. Ethnography is a holistic

approach, where the researcher embeds himself or herself within the community that they are studying; this enables the researcher to make use of first-hand experiences to inform his or her research. The benefit of adopting an ethnographic approach is that I was able to develop in-depth understanding of the processes, people and science that underpin the atmospheric chemistry domain considered in this thesis. This benefit came at the cost of developing a broader, more objective understanding of the problem domain than might have been developed if I had adopted the role of a more passive observer.

Emergence of research objectives: The research objectives, presented in the following sub-section, emerged over the course of my PhD. The study of provenance for data produced by computational models was motivated and informed by first-hand experiences and observations of issues with current working practices (in this case the use of the laboratory notebook to record provenance).

1.2 Research objectives

Having described the project and research approach above, the four research objectives addressed within this thesis are presented; each objective is described in turn below.

1. **Develop an open source modelling system:** to make it easier for atmospheric chemistry community members to develop computational models using the MCM (addressed in Chapter 3).
2. **Explore the role of experimental data in configuration of atmospheric chemistry models:** specifically the impact of the frequency of the experimental data, and the interpolation method used to determine the value of a variable in between data points (addressed in Chapter 4).
3. **Explore the role of provenance in current working practices:** mapping current data-generating working practices, and associated provenance capture practices, to identify opportunities to apply e-Science technologies to reengineer working practices and add value (addressed in Chapters 5 and 7).
4. **Design, develop and evaluate a tool to facilitate provenance capture:** based upon the opportunities, identified as part of objective 3, deliver a tool to support the capture and structuring of provenance for data generated by computational models (addressed in Chapters 8, 9 and 10).

1.3 Thesis structure

Chapter 2: Provides background to the chemistry research presented in this thesis, focusing on the relevant atmospheric chemistry and the role of computational modelling.

Chapter 3: Describes the design, development and testing of an Open Source Box Model (OSBM), a modelling system intended to make the MCM a more accessible and usable information source across the atmospheric chemistry community.

Chapter 4: Explores the impact of constraint implementation on modelled radical concentrations, where constraint implementation is a means of configuring a computational model using experimental data.

Chapter 5: Outlines data-generating working practices across the atmospheric chemistry community and identifies the capture of provenance for data produced by computational models, as an area where opportunities exist to re-engineer working practices and apply e-Science technologies to add value. An Electronic Laboratory Notebook (ELN), the subject of the following chapters, is proposed as a means of exploiting these opportunities. Background to this e-Science research is also provided.

Chapter 6: Describes the methodology used to develop and evaluate the ELN.

Chapter 7: Maps the current working practices of researchers; capturing provenance for data produced by computational models.

Chapter 8: Describes the design of the ELN, considering: the interactions between the ELN user and the ELN; and the design of the information structures used to represent the provenance captured by the ELN.

Chapter 9: Describes the implementation of the ELN, considering the technologies used to realise an ELN prototype.

Chapter 10: Describes the evaluation of the ELN prototype; exploring the responses of two members of the atmospheric chemistry community to the prototype, and identifying implications for the ELN design and implementation.

Chapter 11: Draws together the conclusions of the research presented in this thesis, and outlines potential future work that could build upon this research.

References

1. Atkinson, M. *e-Science*. [cited 19th February 2009]; Available from: http://www.rcuk.ac.uk/escience/default.htm.
2. 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. p. 175-197.

Chapter 2 An Introduction to Atmospheric Chemistry

As described in the first chapter, the research presented in this thesis is a result of a multi-disciplinary research project seeking to develop knowledge in the fields of atmospheric chemistry and e-Science. This research is firmly grounded in the atmospheric chemistry domain and this chapter provides general background to the research presented in this thesis, and detailed background for the model development research presented in Chapters 3 and 4. This chapter consists of six sections: first, the question "why study atmospheric chemistry?" is addressed; secondly, the structure of the atmosphere is described; thirdly, some key elements of the chemistry taking place in the atmosphere are described; next, an overview of the structure and core activities of the atmospheric chemistry community is presented; in the penultimate section, the computational models considered throughout this thesis are described; and finally, the Master Chemical Mechanism (MCM), a key information source for the atmospheric chemistry community is described.

2.1 Why Study Atmospheric Chemistry?

The goal of atmospheric chemistry research is to develop understanding of the composition of the atmosphere and the chemical processes taking place within it. The factors that motivate this goal emerge from the complex inter-relationship between humans and the composition of the atmosphere.

Many aspects of human life are impacted by the composition of the atmosphere including: human health, particularly in relation to the respiratory and cardiovascular systems; and, the social, political and economic landscape in which we live. Human health is impacted upon by the air quality we experience; and, the social, political and economic landscape in which we live is affected by the climate, and climate change trends (which are related to atmospheric composition, along with many other variables).

The composition of the atmosphere is impacted by many human activities, including: the emissions of chemical species through industrial processes, transportation etc.; and the way in which land is used and developed (e.g. cities create concentrated emission sources and change the atmospheric dynamics at a local level).

Having highlighted the complex inter-relationship between humans and the composition of the atmosphere as a source of factors that motivate atmospheric chemistry research, two of the primary motivating factors are examined in detail below.

Atmospheric chemistry, air quality and human health

Historically the study of atmospheric chemistry has been motivated by the occurrence of, and efforts to avoid, air quality episodes. Interest in air quality developed to a tipping point in the late 19^{th} and early 20^{th} century. The occurrence of two distinct types of air quality episode were critical in reaching the tipping point:

- London smog, referred to as "pea soup", occurred around the turn of the $19^{th}/20^{th}$ century where smoke and fog combined. The emissions from burning coal and the emissions from the early chemical industry were key contributors to these smog episodes.
- Los Angeles Smog, a photochemical, fog first formed in the 1930s, as a result of the interactions of hydrocarbons, nitrogen oxides and ozone in the presence of sunlight. The increase in these chemical species was driven by the increased use of automobiles.

The investigations of such smog episodes (particularly photochemical smog) have formed the basis of the tropospheric chemistry research discipline that exists today. This research has been motivated by the detrimental impact of smog episodes on public health [1], and the associated public interest.

Atmospheric chemistry and climate change

Climate change is driven by increases in anthropogenic emissions of greenhouse gases [2]; the most abundant of which, in the troposphere, are CO_2, CH_4, N_2O and O_3 [1]. Although these greenhouse gas (with the exception of O_3), at their current ambient troposphere concentrations, do not impact on human health [1], the impact of climate change is likely to be very significant. Understanding the chemical processes (formation and degradation) associated with these greenhouse gases, particularly O_3, is an important component of the atmospheric chemistry research agenda. Currently climate models, as evaluated by the IPCC (Intergovernmental Panel on Climate Change), do not typically include interactive descriptions of the chemistry taking place in the atmosphere [3], as

developed by atmospheric chemistry research. In the future, as improving computational resources enable ever more complicated climate models to be developed, it is likely that these interactive descriptions will be incorporated.

2.2 Atmospheric Structure

This section describes the structure of the atmosphere, in terms of the temperature profile with increasing distance from the Earth's surface. The atmosphere can be considered to be composed of four layers [4], as shown in the Figure 2.1, where a new layer is defined at a change in the sign of the temperature gradient[2]. The heights of the boundaries between layers vary, and are dependent on atmospheric conditions and latitude. The four layers are: the troposphere, the stratosphere, the mesosphere, and the thermosphere. The research in this thesis focuses on chemistry taking place in the troposphere, so only the structure of the troposphere is considered in detail below.

Temperature inversions: a temperature inversion is defined as a region of the atmosphere where the lapse rate is negative [4]. Where lapse rate is defined as the rate of decrease of temperature with altitude [4]. So, temperature inversions are layers of warmer air, on top of cooler air and are very stable with regard to vertical transport of air/matter.

2.2.1 Structure of the Troposphere

This sub-section describes the physical characteristics of the troposphere, where the majority of chemical material in the atmosphere resides. The troposphere itself can be considered to consist of three sub layers: the surface layer; the boundary layer; and, the free troposphere; each of these layers is described below.

[2] A positive temperature gradient sees the temperature of the atmosphere increase with increasing distance from the earth's surface. Conversely, a negative temperature gradient sees the temperature decrease with increasing distance.

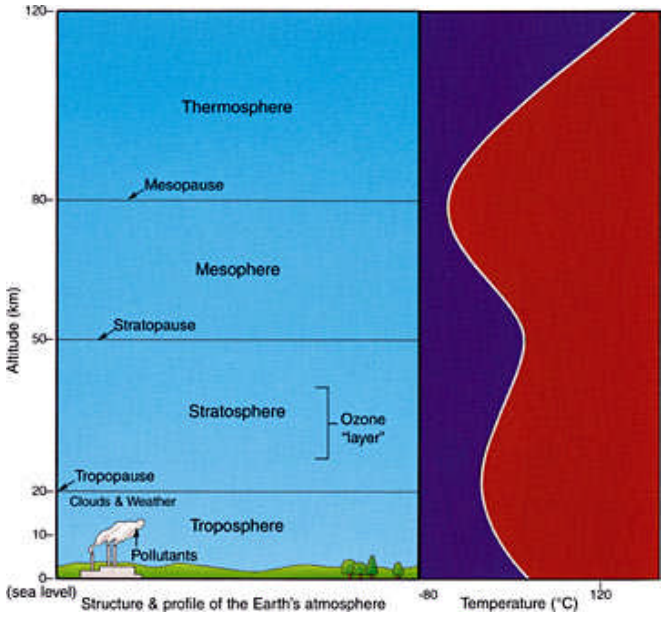


Figure 2.1: Atmospheric structure and temperature profile [5]. Showing the atmosphere divided into four sections: the troposphere; the stratosphere; the mesosphere; and the thermosphere.

The surface layer: A typical height for the surface layer is from the Earth's surface to between 50-300 m. The surface layer is characterised by the influence of the local landscape on its chemical composition and transport mechanisms [6]. The rough surface of the Earth causes turbulence, ensuring that the surface layer is well mixed. Heating takes place due to radiation (from the Earth's surface), convection and conduction. The surface layer is the most critical in terms of air quality, as it is the composition of this layer that results in population exposure and determines the health effects of air quality and pollution. During the night, due to the relative rates of cooling of the surface and the atmosphere, temperature inversions can occur at the boundary of the surface layer. Such inversions can prevent transport to the boundary layer, restricting the movement of pollutants. This phenomenon can lead to a build up of pollutants, severe winter episodes of this type have been experienced in the UK in 1991 and 2001 [7].

The boundary layer: The boundary layer typically occupies a region 300-3000 m above the earth's surface. Again this layer is well mixed, in this case due to convective mixing. The upper edge of the boundary layer is characterised by a small temperature inversion

during the day, ensuring that the transfer of chemical matter to the free troposphere is slow, trapping pollutants.

The free troposphere: The free troposphere typically occupies a region 3000-20000m above the Earth's surface, and is characterised by convective heating and a negative temperature gradient. The negative temperature gradient is a result of the reduced radiative heating effects of the Earth's surface with increasing height. The upper boundary of the free troposphere is the tropopause, at this point a temperature inversion occurs, with the temperature gradient becoming positive in the stratosphere.

2.3 Chemistry in the Troposphere

This section provides an overview of some important chemical processes that take place within the troposphere, focussing on the chemistry of the hydroxyl radical (OH), volatile organic compounds (VOCs)[3] [8], NO_x ($NO + NO_2$) and ozone. The overview provides background relevant to Chapter 4, where the computational modelling of hydroxyl radical concentrations on short timescales is explored. Three key components of the chemistry taking place in the troposphere are described below: first, the chemistry of the hydroxyl radical; secondly, the reactions involved in the degradation of a VOC; and thirdly, ozone generation processes.

2.3.1 Hydroxyl Chemistry

A key chemical species in atmospheric chemistry is the hydroxyl radical, OH. Although OH occurs in relatively small concentrations it drives many reactions in the atmosphere. This is because the vast majority of VOCs in the troposphere cannot be removed by deposition and none react with oxygen or nitrogen (the main constituents of the atmosphere). Given the absence of other reactions, and as OH is very reactive, it is involved in the initiation of many of the VOC degradation pathways [4]. A degradation

[3] VOCs (Volatile Organic Compounds) are ozone pre-cursors and comprise a wide range of chemical compounds including hydrocarbons (alkanes, alkenes, aromatics), oxygenates (alcohols, aldehydes, ketones, ethers) and halogen containing species. VOCs are emitted from anthropogenic sources (such as industry and transportation) and from biogenic sources (such as trees and other plants).

pathway is the mechanism by which a VOC is oxidised in the atmosphere; in the presence of NO_x, this mechanism leads to the formation of O_3.

OH is generated in the troposphere, primarily by the mechanism described below. Reactions R2.1 and R2.2 represent the two possible ways in which tropospheric ozone can be photolyzed to form ground state oxygen atoms, $O(^3P)$, or excited state oxygen atoms, $O(^1D)$. It is worth noting that this photolysis occurs only at wavelengths above 290nm, because light at wavelengths below 290nm has been absorbed by stratospheric ozone. $O(^3P)$ does not react to form OH, because it has insufficient energy, and leads to regeneration of O_3 via reaction with O_2. Equation R2.3 shows approximately 90% of $O(^1D)$ is converted to $O(^3P)$, energy being removed by collision with some other molecule, M, generally nitrogen or oxygen. The remaining 10% of $O(^1D)$ reacts with water vapour to form OH radicals, equation R2.4.

$$O_3 + h\nu \rightarrow O(^1D) + O_2 \qquad \text{(R2.1)}$$

$$O_3 + h\nu \rightarrow O(^3P) + O_2 \qquad \text{(R2.2)}$$

$$O(^1D) + M \rightarrow O(^3P) \qquad \text{for approximately 90\% of } O(^1D) \qquad \text{(R2.3)}$$

$$O(^1D) + H_2O \rightarrow 2OH \qquad \text{for approximately 10\% of } O(^1D) \qquad \text{(R2.4)}$$

2.3.2 VOC Chemistry

Having described the processes involved in the generation of OH, in the previous sub-section, this sub-section describes the degradation pathway of methane (as initiated by OH). As the simplest (in terms of chemical structure) VOC; this degradation pathway has been selected to act as an example of the wider set of degradation pathways for more complicated VOCs (such as alkanes, alkenes, aromatics, etc.). The important role of the hydroxyl radical in initiating these degradation pathways is shown in reaction R2.5.

Initiation

$$OH + CH_4 \xrightarrow{O_2} CH_3O_2 + H_2O \qquad \text{(R2.5)}$$

Propagation

$$CH_3O_2 + NO \rightarrow CH_3O + NO_2 \qquad \text{(R2.6)}$$

$$CH_3O + O_2 \rightarrow HCHO + HO_2 \qquad \text{(R2.7)}$$

$HO_2 + NO \rightarrow OH + NO_2$ (R2.8)

Termination

$HO_2 + HO_2 + M \rightarrow H_2O_2 + O_2 + M$ (R2.9)

$OH + NO_2 + M \rightarrow HNO_3 + M$ (R2.10)

Methane reacts with the hydroxy radical (OH), see R2.5, to form a methyl radical, which then reacts rapidly with O_2 to form a methyl peroxy radical (CH_3O_2). The methyl peroxy radical then reacts with nitric oxide (NO), reaction R2.6; this reaction highlights the importance of nitrogen oxides (NO_x) in the chemical mechanisms of the troposphere. The reaction of nitric oxide with the hydroperoxyl radical (HO_2) regenerates OH, reaction R2.8. The termination steps of the degradation pathway, reactions R2.9 and R2.10, lead to products that can be removed from the atmosphere. For example the nitric acid (HNO_3), can be removed by wet deposition. That is, the nitric acid dissolves in water in the atmosphere and is rained out. Alternatively dry deposition may occur, where the nitric acid is removed by interaction with the Earth's surface or aerosols.

2.3.3 Ozone Chemistry

The generation of ozone in the troposphere is important for three reasons: first, ozone is a major component of photochemical smog; secondly, ozone is damaging to human health [9, 10], and for this reason ozone concentrations are legally regulated and targets are set, to protect public health [11, 12]; thirdly, ozone is the greenhouse gas with the third greatest contribution to climate change.

The production of tropospheric ozone proceeds by the mechanism shown in R2.11 and R2.12 [13], and is driven by the photolysis of nitrogen dioxide at wavelengths less than 420 nm to produce an atom of ground state oxygen, $O(^3P)$. $O(^3P)$ then combines with an oxygen molecule (O_2) to form ozone (O_3). As the production of ozone is dependent on the concentration of nitrogen dioxide, it is tightly coupled with NO_x emissions (e.g. from transportation) and the degradation of VOCs in the presence of NO_x (as described above for methane).

$NO_2 + h\nu \rightarrow O(^3P) + NO$ (R2.11)

$O(^3P) + O_2 + M \rightarrow O_3 + M$ (R2.12)

Reaction R2.12 is the sole chemical source of ozone in the troposphere. Ozone reacts with NO to regenerate NO_2 (see R2.13). Reactions R2.11-2.13 form the basis of the photochemical cycle of NO_2, NO, O_3 [13].

$$O_3 + NO \rightarrow NO_2 + O_2 \qquad \text{(R2.13)}$$

2.4 Structure of the Research Community

This section describes the structure of the atmospheric chemistry research community. This description of the research community places the modelling research presented, in this thesis (Chapters 3 and 4), in context; and plays an important role in informing the provenance research presented in the later parts of this thesis (Chapter 5 onwards). An overview of the atmospheric chemistry research community is presented below, followed by more detailed descriptions of the main research activities taking place across the community.

2.4.1 Community Overview

Atmospheric chemistry is an inherently multi-scale 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 the global distribution of methane and ozone, which, after CO_2, are the trace gases with the greatest influence on climate change; and developing models which inform air quality policy. A central component of models investigating atmospheric chemistry on a global or regional scale is the chemical mechanism. Chemical mechanisms, part of the complex reaction scale of atmospheric chemistry research, consist of a coupled set of steps called elementary reactions in which chemical species are inter-converted (i.e. mechanisms are lists of chemical reactions).

Elementary reactions are investigated primarily in the laboratory; detailed chemical mechanisms are constructed from knowledge of these elementary reactions and their interactions (this activity is referred to as mechanism development). Mechanisms are used directly to construct models containing a very large set of ordinary differential equations, where the derivatives represent the rates at which the concentrations of species in the mechanism change with time. Such models are used for problems with modest fluid

dynamic requirements, e.g. local scale modelling of *in situ* measurements, in order to test the performance of the chemical mechanism. These mechanisms can contain a large number of elementary reactions, often in excess of 10000, and so are too computationally expensive to implement within global and regional models. In such cases, mechanisms of much lower dimension are used, ideally based on objective reduction and lumping of the detailed mechanisms, providing a link between the global and regional scale models, and fundamental chemical kinetics.

Having presented an overview of the structure of and activities conducted by the atmospheric chemistry research community, the remainder of this section provides details for each of the key community activities, starting with field studies.

2.4.2 Field Studies

The focus of a field study is to make *in situ* measurements in the atmosphere and generate understanding of the chemical processes taking place in the atmosphere, through the interpretation of these measurements. Field studies are conducted across the globe, in varying conditions and focussing on various chemical species. Field studies take place in both polluted and non-polluted environments. Studies in non-polluted environments, such as Mace Head (Ireland) [14] and Cape Grim (Tasmania) [15] [16], where air has travelled over oceans for a number of days, provide an opportunity to understand the chemistry taking place in very clean air. Studies in polluted environments, such the TORCH field campaign (Tropospheric Organic Chemistry experiment) [17], enable insight to be generated into the effects of anthropogenic emissions. Field studies in areas where there are significant biogenic emissions (usually VOCs), such the BEMA (Biogenic Emissions in the Mediterranean Area) project [18], enable insight to be generated into the role of biogenics in determining the composition of the troposphere. Computational modelling plays an integral role in the analysis of field study data. Feedback, between field studies and computational models, is provided in both directions with modelling helping to explain observed phenomena and field studies aiding the development of increasingly realistic models.

2.4.3 Laboratory Studies

Laboratory studies seek to determine rate coefficients and product yields for chemical reactions of atmospheric importance. An experimental technique often used in laboratory studies is flash photolysis [19]. Experiments are designed and executed to determine the effects of variables, including temperature and pressure, on the rate coefficient of a given reaction. Increasingly, laboratory studies are conducted in conjunction with computational model development, on an elementary reaction scale, using techniques such as the master equation [20]. The rate coefficients and product yields determined by lab experiments are incorporated into chemical mechanisms, describing the chemical processes taking place in the atmosphere (as described in Section 2.4.5).

2.4.4 Chamber Studies

Chamber studies lie between field and laboratory studies in the experimental domain, and typically focus on a subset of the chemical process taking place in the atmosphere, such as the NO_3 chemistry of aldehydes [21], or the photo-oxidation of aromatic species [22]. Chamber studies are conducted in large, controlled environments, which aim to recreate the characteristics of the real atmosphere whilst retaining experimental control. This allows a laboratory level of instrumentation which enables the study of mechanisms, reactions and species of the researchers' choosing. Chamber studies, and associated computational models, play a critical role in the development of chemical mechanisms (described below).

2.4.5 Mechanism Development

Mechanism development activities seek to develop chemical mechanisms that describe the chemical processes taking place in the atmosphere; e.g. the EXACT campaign (Effects of the oXidation of Aromatic Compounds in the Troposphere) [23], which explores the mechanisms of aromatic compounds. Mechanism development, will often focus on describing the degradation pathway of a given VOC, and involves three core activities.

- Determining the reactions taking place.
- Identifying and selecting the rate coefficients for the reactions, where possible, using data generated by laboratory studies.

- Testing the ability of the mechanism to predict the behaviour of the physical system; e.g. the mechanism is used within a computational model, and the model output data can be compared to chamber experiment data (in order to evaluate the mechanism's performance).

Having determined the mechanism for a given VOC, it can then be combined with the mechanisms for other VOCs to provide a description of the processes taking place in the atmosphere.

2.4.6 Computational Modelling

In each of the key community activities (field studies, laboratory studies, chamber studies and mechanism development), computational modelling and experimental science perform complementary roles in the pursuit of understanding and quantification of the chemical processes taking place in the atmosphere. In this sub-section the role of computational modelling is considered in greater detail. First, the question "why develop computational models?" is addressed, and then an overview of the types of models developed across the atmospheric chemistry community is presented.

2.4.6.1 Why Develop Computational Models?

The systems being investigated in atmospheric chemistry are inherently complex and it is impossible to accurately capture their complete state. So, any computational models developed are necessarily simplifications of systems they represent. The motivation for developing the computational models, within the atmospheric chemistry community, includes the following aspects.

- Capturing the essence of the physical system; if the computational model of current understanding accurately matches/predicts empirical measurements, then there is a good chance that the most important elements of the physical system have been incorporated in the computational model.
- Enabling researchers to leverage the computational resources now available to the scientific community, to explore larger numbers of more varied, scenarios than it is possible to explore experimentally.

- The output of a well-constructed and well-understood model is useful in guiding and supporting the experimental process. The agreement of model and experiment makes for a more convincing case than either alone can provide.

2.4.6.2 Computational Model Development Across the Atmospheric Chemistry Community

Having addressed the question "why develop computational models?", above, this sub-section describes the types of computational model developed across the atmospheric chemistry community. The types of computational models are described below according to the scale of phenomena they model.

- **At the elementary reaction scale:** models are developed to determine the characteristics (e.g. rate coefficients and branching ratios) of a given reaction (often conducted in conjunction with laboratory studies). Examples include the study of the reaction between methylglyoxal and OH/OD radical [20].
- **At the complex reaction scale:** models consider the chemical processes taking place at a given point in space; chamber and field experiments are often modelled in this way, using a so called "zero dimensional" box model. Examples include: modelling chamber experiments exploring the chemistry of aromatic compounds, EXACT [22]; and, modelling field experiments exploring the chemistry taking place in extremely clean environments, SOAPEX (Southern Ocean Atmospheric Photochemical Experiment) [16].
- **At the local/regional scale:** models consider both chemical processes and transport of chemical material (within a defined space), and are often linked to the prediction of air quality. Examples include: models of the distribution of pollutants within a street canyon (i.e. the space sets of tall buildings) [24]; and, models of the distribution of ozone across a city centre [25].
- **At the global scale:** models considering the global distribution of chemical species. For example the global distribution of methane is modelled [26] because it is an important greenhouse gas.

This section has provided a high-level overview of the structure of the atmospheric chemistry community and the activities that take place within it. The importance of performing computational modelling in conjunction with experimental investigations has

been highlighted. The next section provides additional details on the development of models at the complex reaction scale, as it is this type of model that is considered throughout the remainder of this thesis.

2.5 Box models

The complex reaction scale models considered throughout this thesis are zero dimensional box models, used to model field and chamber experiments. More complicated box models are not considered in the research presented in this thesis. This section consists of four components: first, a discussion of the nature of box models for field studies; secondly, a discussion of the nature of box models for chamber studies; thirdly, a description of the general mathematical specification for a box model; and, finally a description of the process of constraining box models to experimental data.

2.5.1 Box models for Field Experiments

Zero dimensional box models are so called because they consider the species within an air parcel to be uniform distributed, so all points within the box are equivalent (effectively reducing the model to a single, zero dimensional, point). Zero dimensional box models are often used for comparison with ground-level field campaign measurements, as there is a natural mapping between the static nature of the ground-based field study and a static box model. A zero dimensional box model is a single cube, with down-wind, cross-wind and vertical axes. Generally the box sits on the Earth's surface, on an area of research interest. A description of the structure of a field campaign box model is presented below.

Within the box:

- Chemical material is involved in chemical reactions defined by a mechanism;
- Some chemical reactions are driven by solar radiation entering the box.

Chemical material can only leave the box in the following ways:

- Deposition to the Earth's surface (i.e. out of the bottom of the box);
- Advective outflow due to wind (i.e. out of the side of the box);
- Detrainment due to the upwards movement of air (i.e. out of the top of the box).

Chemical material can only enter the box in the following ways:

- Emission from the Earth's surface (i.e. in from the bottom of the box);

- Advective inflow due to wind (i.e. in from the side of the box);
- Entrainment due to the downward movement of air (i.e. in from the top of the box).

The mathematical specification for a box model for a field study is shown as equation E2.1 [27].

$$\frac{\partial[i]}{\partial t} = \frac{u([i]_0 - [i])}{l} + S_i + C_i + \frac{w_v([i]_0^{\neq} - [i]) - w_{\alpha i}[i]}{h} \qquad \text{(E2.1)}$$

In equation E2.1, species *i* is present at concentration $[i]$ in a well-mixed square-based box of length *l* and height *h* with a fluid (i.e. air) velocity of *u.* S_i is the emission source term, $w_{\alpha i}$ is the surface deposition velocity of species *i*, w_v the ventilation velocity (describing the exchange with air above the box), $[i]_0^{\neq}$ the concentration of *i* above the box, C_i the chemical loss or production rate, and $[i]_0$ the upwind concentration of *i*. See "Atmospheric Change: An Earth System Perspective" [27] for further details.

2.5.2 Box Models for Chamber Experiments

Zero dimension box models are often used for comparison with the results of chamber studies. This comparison allows the performance of the mechanism (implemented within the box model) to be evaluated, for the restricted case being studied in the chamber experiment. The box is considered to sit within the chamber, and is bounded by the chamber walls. A description of the structure of a chamber study box model is presented below.

Within the box:

- Chemical material is involved in chemical reactions defined by a mechanism;
- Some chemical reactions are driven by radiation, either solar radiation or radiation from some source simulating solar radiation (e.g. a lamp).

Chemical material can leave the box by the following mechanisms:

- Though leakage from the chamber.

Chemical material can enter the box by the following mechanisms:

- Injection into the chamber by the experimentalist;
- Desorption from the walls of the chamber.

2.5.3 A General Mathematical Specification for a Box Model

Having described box models for field and chamber studies, in the preceding sub-sections, this sub-section presents the general mathematical specification that both types of box model adhere to. Atmospheric chemistry box models (for both field and chamber studies) can be considered as Ordinary Differential Equation (ODE) initial value problems, expressed mathematically [28] as:

$$\dot{y} = f(t, y), \quad y(t_0) = y_0 \qquad \text{(E2.2)}$$

Here $\dot{y} = \frac{\partial y}{\partial t}$, $y \in \Re^n$, n is the number of species being modelled and t is the independent variable (in this case time). It is assumed that the initial values for the concentrations of the species being modelled, y_o, is known. Relating the mathematical description to the physical system, the array y contains the concentrations of each chemical species at time t. In the atmosphere these concentrations are determined by the previous concentrations and the chemical reactions taking place. The array $\dot{y}$ contains the rate of changes of chemical concentration for each species at time t. Solving this set of ODEs is considered in greater depth in Chapter 3, which describes the development of a Open Source Box Model, for use by the atmospheric chemistry community.

2.5.4 The Role of Constraints on Box Models

It is standard practice within the atmospheric chemistry community to constrain photochemical box models to field data. A constrained box model seeks to develop and test the understanding of the chemistry taking place at a given location (e.g. where a field campaign has taken place). OH is often used as a target species (i.e. the species focussed on during comparisons of model output with *in situ* measurements) for constrained field models because it has a short atmospheric lifetime. The benefits of using a short-lived species as a target are: it is not affected by atmospheric transport (transport is not modelled in zero dimensional box models); and it responds rapidly to the constraint data.

Two types of constraints, serving different purposes, are used during model development; each constraint type is described below.

Environmental conditions: such as photolysis rates, temperature, relative humidity and solar declination, influence the chemistry taking place within the model. Within the model these conditions are implemented as variables, which take their values from the appropriate constraint dataset. The value of the variable at a given time is determined by the internal model time and either data interpolation methods (where the conditions are discrete) or a simple formula (where these are known). The purpose of environmental constraints (along with other parameters) is to place the model at the physical and temporal location of the field campaign.

Chemical constraints: The purpose of the chemical constraints is to provide the model with information about the air mass at the field campaign location at a given time. This eliminates the need to model constrained species entering or leaving the conceptual box. Chemical constraints can be implemented for any subset of chemical species, but are usually implemented for species such as Volatile Organic Compounds (VOCs), NO, NO_2, CH_4, CO_2, HCHO etc. [14] [16]. These constraints are effectively steering the computational model, based upon observations of the physical system, after the initial time.

This section has provided background on the use of box models, in conjunction with field and chamber studies, including a general mathematical specification for box models and the role constraints play in the modelling process. The discussion now progresses to consider the role of the MCM (Master Chemical Mechanism) in developing box models, in the final section of this chapter.

2.6 A Master Chemical Mechanism

Research on elementary reactions and chemical mechanisms is conducted in laboratories throughout the world. The Master Chemical Mechanism (MCM) [29] [30] [31] is the leading detailed chemical mechanism and is used across the international research community. The MCM describes the chemistry occurring in the troposphere (i.e. 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. The box models considered in this thesis make use of the MCM to describe the chemistry taking place at a given field campaign location or within a chamber. Typically, a researcher will take the MCM and then tailor the mechanism to the specific requirements of the system being modelled.

The MCM is an explicit chemical mechanism developed by Jenkins *et al.* in 1997 [29], and subsequently updated in 2003 [30, 31], to reflect the advances in knowledge of atmospheric chemistry. The processes involved in constructing the MCM are described in detail in Chapter 5 of this thesis. The version currently available to the atmospheric chemistry community (and any other interested party) is MCM (v3.1) (http://mcm.leeds.ac.uk/MCM/). MCM (v3.1) includes the degradation schemes for 135 VOCs, describing their complete degradation (which ends with the final oxidation to CO_2 and H_2O).

The MCM was originally developed with support from the UK Department for Environment, Food and Rural Affairs (DEFRA). The goal of developing the MCM was to provide a chemical mechanism, incorporating the cutting edge of scientific knowledge, that describes the degradation of VOCs and the production of secondary photochemical pollutants (such as ozone). The target application for the MCM is inclusion in air quality models for the boundary layer over continental Europe and the UK. The models seek to derive scientific knowledge and inform policy decisions, such as emission regulations for a given chemical species.

Chapter Summary

This chapter has provided background to the research presented in this thesis; the fundamentals of atmospheric chemistry have been discussed, including the structure of the atmosphere and key chemical reactions taking place within it. The motivation for studying atmospheric chemistry and the research community that has evolved to study atmospheric chemistry have also been described. Detailed background has also been provided for Chapters 3 and 4, describing the role of zero dimensional box models and the Master Chemical Mechanism. The next chapter describes a modelling system that aims to facilitate the development of box models (for field and chamber studies) that make use of the MCM.

References

1. Jacobson, M.Z., *Atmospheric Pollution: history, science and regulation*. 2002, Cambridge: Cambridge University Press.
2. Bernstein, L., et al., *Climate Change 2007: Synthesis Report*. 2007, Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.
3. Randall, D.A., et al., *Climate Models and Their Evaluation*, in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, S. Solomon, et al., Editors. 2007, Cambridge University Press: Cambridge, United Kingdom and New York, NY, USA.
4. Wayne, R., *Chemistry of Atmospheres*. 2nd ed. 1991, Oxford: Clarendon Press.
5. *Structure of the atmosphere*. 23rd November 2006 [cited 14th June 2007]; Available from: http://www.partnersinair.org/en/images/curr_unit1a_bkgd_figure11.jpg.
6. Sommariva, R.C., *Understanding Field Measurements through a Master Chemical Mechanism*, in *School of Chemistry*. 2004, University of Leeds.
7. *Defra AQ Brochure 2004*. 2004 [cited 12th February 2008]; Available from: http://www.airquality.co.uk/archive/reports/cat05/0408161000_Defra_AQ_Brochure_2004_s.pdf.
8. Atkinson, R., *Atmospheric chemistry of VOCs and NOx*. Atmospheric Environment, 2000. **34**(12-14): p. 2063-2101.
9. Bernstein, J.A., et al., *Health effects of air pollution.* Journal of Allergy and Clinical Immunology, 2004. **114**(5): p. 1116-1123.
10. Bell, M., et al., *Climate change, ambient ozone, and health in 50 US cities.* Climatic Change, 2007. **82**(1): p. 61-76.
11. *The Air Quality Strategy for England, Scotland, Wales and Northern Ireland*, DEFRA, Editor. 2007.
12. *Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe*. 2008.
13. Seinfeld, J.H. and S.N. Pandis, *Chemistry of the Troposphere*, in *Atmospheric Chemistry and Physics - From Air Pollution to Climate Change (2nd Edition)*. 2006, John Wiley & Sons.
14. Heard, D.E., et al., *The North Atlantic Marine Boundary Layer Experiment (NAMBLEX). Overview of the campaign held at Mace Head, Ireland, in summer 2002.* Atmos. Chem. Phys., 2006. **6**(8): p. 2241-2272.
15. Monks, P.S., et al., *Fundamental ozone photochemistry in the remote marine boundary layer: the soapex experiment, measurement and theory.* Atmospheric Environment, 1998. **32**(21): p. 3647-3664.
16. Sommariva, R., et al., *OH and HO2 chemistry in clean marine air during SOAPEX-2.* Atmos. Chem. Phys., 2004. **4**(3): p. 839-856.
17. Emmerson, K.M., et al., *Free radical modelling studies during the UK TORCH Campaign in Summer 2003.* Atmos. Chem. Phys., 2007. **7**(1): p. 167-181.
18. Kesselmeier, J., et al., *Emission of monoterpenes and isoprene from a Mediterranean oak species Quercus ilex L. measured within the BEMA (Biogenic Emissions in the Mediterranean Area) project.* Atmospheric Environment, 1996. **30**(10-11): p. 1841-1850.
19. Blitz, M.A., et al., *Laser induced fluorescence studies of the reactions of O(1D) with N2, O2, N2O, CH4, H2, CO2 , Ar, Kr and n-C4H10.* Phys. Chem. Chem. Phys., 2004. **6**: p. 2162-2171.

20. Baeza-Romero, M.T., et al., *A combined experimental and theoretical study of the reaction between methylglyoxal and OH/OD radical: OH regeneration.* Physical Chemistry Chemical Physics, 2007. **9**(31): p. 4114-4128.
21. Bossmeyer, J., et al., *Simulation chamber studies on the NO3 chemistry of atmospheric aldehydes.* Geophys. Res. Lett., 2006. **33**.
22. Bloss, C., et al., *Evaluation of detailed aromatic mechanisms (MCMv3 and MCMv3.1) against environmental chamber data.* Atmos. Chem. Phys., 2005. **5**(3): p. 623-639.
23. Zádor, J., et al., *Measurement and investigation of chamber radical sources in the European Photoreactor (EUPHORE).* Journal of Atmospheric Chemistry, 2006. **55**(2): p. 147-166.
24. Baik, J., Y. Kang, and J. Kim, *Modeling reactive pollutant dispersion in an urban street canyon.* Atmospheric Environment, 2007. **41**(5): p. 934-949.
25. Ziomas, I.C., et al., *Ozone episodes in Athens, Greece. a modelling approach using data from the medcaphot-trace - an outline.* Atmospheric Environment, 1998. **32**: p. 2313-2321.
26. Bousquet, P., et al., *Contribution of anthropogenic and natural sources to atmospheric methane variability.* Nature, 2006. **443**(7110): p. 439-443.
27. Graedel, T.E. and P.J. Crutzen, *Atmospheric Change: An Earth System Perspective.* 1993, New York: W. H. Freeman and Company.
28. Cohen, S., D. and A. C. Hindmarsh, *CVODE: A Stiff/Nonstiff ODE Solver in C.* Computers in Physics, 1996. **10**(2): p. 138-143.
29. Saunders, S.M., et al., *World wide web site of a master chemical mechanism (MCM) for use in tropospheric chemistry models.* Atmospheric Environment, 1997. **31**(8): p. 1249-1249.
30. Saunders, S.M., et al., *Protocol for the development of the Master Chemical Mechanism, MCM v3 (Part A): tropospheric degradation of non-aromatic volatile organic compounds.* Atmos. Chem. Phys., 2003. **3**(1): p. 161-180.
31. Jenkin, M.E., et al., *Protocol for the development of the Master Chemical Mechanism, MCM v3 (Part B): tropospheric degradation of aromatic volatile organic compounds.* Atmos. Chem. Phys., 2003. **3**(1): p. 181-193.

Chapter 3 Developing an Open Source Box Model for use with the Master Chemical Mechanism

This chapter presents the development of an Open Source Box Model (OSBM) for use with the Master Chemical Mechanism (MCM). The motivation for developing the OSBM was to make it easier for atmospheric chemistry community members to develop models using the MCM. This chapter consists of eight sections:

1. An outline of the goals of developing the OSBM;
2. A review of existing modelling tools for developing atmospheric chemistry models;
3. The requirements specification for the OSBM;
4. The design of the OSBM;
5. The implementation of the OSBM;
6. The testing and benchmarking of the OSBM (with reference to two case studies);
7. A review of the progress made toward meeting the OSBM requirements specification is presented;
8. A discussion of future development work associated with the OSBM.

The work presented in this chapter is a combination of the efforts of Monica Vazquez-Moreno (CEAM, Valencia), Dr. Katarzyna Borońska (School of Computing University of Leeds) and the author. Monica Vazquez-Moreno designed and developed the graphical user interface for the OSBM. Dr. Katarzyna Borońska designed and developed the web service interface (see Section 3.4.1), and re-engineered the OSBM source code to a production quality. I undertook all other work presented.

3.1 Research Goal

The goal that motivated the development of the OSBM was to:

> *Encourage uptake and evaluation of the MCM by developing a generic box model that operates seamlessly with the MCM. The model should be free and easy to distribute, requiring minimal effort and experience on the part of the user to install and develop basic models (for both field and chamber experiments).*

This section examines the motivation for pursuing this goal and consists of two components: first, a discussion of the role the OSBM will play in improving the uptake of the MCM across the research community; secondly, a discussion of the EUROCHAMP project, and the role the OSBM will play within it.

3.1.1 Encouraging Uptake of the MCM Across the Research Community

Encouraging uptake of the MCM, is a desirable outcome because it will lead to improvements in: the quality of research across the atmospheric chemistry research community; and, the quality of the MCM itself. For example increased uptake of the MCM will:

- Ensure more computational models incorporate mechanisms with links to fundamental experimental science;
- Enable more feedback on MCM performance to be gathered, allowing the MCM to be incremental improved.

The MCM website can be viewed as a service accessed by the atmospheric research community, providing valuable resources that inform and facilitate the research taking place across the community. Currently the resources provided by the MCM website are informational, i.e. the mechanism itself. Freely available, high quality computational tools that enable scientific insight to be generated directly from the MCM are required to encourage use of the MCM. The OSBM is one such computational tool that could make the MCM easier to use. The target users for the OSBM cut across the atmospheric chemistry community include the following.

- *Experienced modellers and mechanism developers:* This group consists of researchers whose primary research interests lie in the domain of atmospheric chemistry modelling. Members of this group are likely to require a core set of modelling functionality plus the option to make extensive customisations to the OSBM.
- *Occasional modellers:* This group consists of researchers who conduct some modelling as part of their role, typically alongside *in vitro* or *in situ* experimental research. Members of this group are unlikely to be interested in the internal

workings of the model (i.e. the source code), and so will be happy to treat the OSBM as a 'black box'.

- *Novice modellers:* This group consists of researchers whose background and research interests lie firmly in the *in vitro* or *in situ* experimental domain, and are new to computational modelling. Members of this group are likely to require access to simple modelling functionality, which enforces a logical structure upon the modelling process that they execute.

Any potential OSBM user may not fall directly into one of the user categories identified above, but will fall somewhere on the continuum of modelling experience (from novice to experienced modeller).

3.1.2 EUROCHAMP

Having considered the role of the OSBM in encouraging uptake of the MCM across the research community in general, in the preceding sub-section, the role of the OSBM in the EUROCHAMP (Integration of European Simulation Chambers for Investigating Atmospheric Processes) project is now discussed. The EUROCHAMP project [1] consists of a consortium of 12 laboratories throughout Europe, each laboratory brings an atmospheric simulation chamber and associated experimental capability to the consortium. The aim of the project is to provide the 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. This has two benefits: facilitating the analysis of chamber experiment, to produce scientific knowledge; and ensuring that the performance of the MCM is frequently evaluated. The OSBM will form the core of this computational infrastructure.

3.2 Alternative Modelling Tools

In order justify the development of the OSBM it is useful to consider the existing modelling tools, already available to the atmospheric chemistry community. Three alternative modelling systems and their shortcomings (in terms of being able to encourage use of the MCM) are discussed in detail below: FACSIMILE [2]; KPP [3]; and ASAD [4].

Each of these modelling systems adopt the approach of offering generic functionality to develop models containing chemical mechanisms, allowing a specific model to be developed by researchers using this generic functionality. The main drawback of this approach is that the researcher must commit significant effort to learning how to use the modelling system and only then develop models, related to their specific research interests.

3.2.1 FACSIMILE

FACIMILE is a commercial software package distributed by ESM software. FACSIMILE enables scientists to develop computational models for "complex steady-state and time-dependent processes … it is especially suitable for solving chemical reactions with diffusion and/or advection" [2]. The MCM can be extracted in a FACSIMILE compatible format and a number of basic box models (that can be interpreted and executed by FACSIMILE) are openly available on the MCM website. The main drawback to promoting FACSIMILE, as a tool for developing models using the MCM, is that FACSIMILE licence fees are significant, and many potential MCM users are either unwilling or unable to pay these fees. The FACSIMILE mechanism/reaction format currently plays an extensive role in the use of the MCM. For this reason the OSBM was designed to be compatible with the FACSIMILE format, an example of this format is introduced below.

General reaction:

$reactants \xrightarrow{k} products$, where k is the rate co-efficient.

General reaction FACSIMILE format:

% k : $reactants$ = $products$;

Specific reaction:

$$OH + CH_4 \xrightarrow{9.65\times10^{-20}\times temp^{2.58}\times e^{(-1082/temp)}} CH_3O_2$$

Specific reaction FACSIMILE format:

% 9.65D-20*TEMP@2.58*EXP(-1082/TEMP) : OH + CH4 = CH3O2 ;

3.2.2 ASAD

A Self-contained Atmospheric chemistry coDe (ASAD) [4] has been developed, and is

supported, by the Atmospheric Chemistry Modelling Support Unit, at the University of Cambridge. ASAD can be viewed as a component of an atmospheric chemistry model and provides functionality to model chemical mechanisms. Other components required to realise a box model could include descriptions of photolysis processes, emissions, and transport of chemical material in to and out of the box. This component-based approach again leads to a significant model development overhead, which is likely to deter the less experienced researcher.

3.2.3 KPP

The Kinetics Pre-Processor (KPP) [3] is, in terms of functionality, similar to ASAD, providing functionality to model chemical mechanisms, and acting as a component within an atmospheric chemistry model. It therefore shares similar drawbacks.

Either KPP or ASAD could have been adopted as a starting point for the development of OSBM, but I decided to develop the OSBM from scratch. This decision was motivated by four key factors.

- First, starting from scratch provided the flexibility required to develop a modelling system customised for the requirements of a specialist research community (i.e. the MCM user community);
- Secondly, starting from scratch provided me with an opportunity to gain an in-depth understanding of the modelling process (later used to inform the provenance research presented in Chapters 5-11 of this thesis);
- Thirdly, starting from scratch, enabled the modelling process to be reviewed and reengineered (with minimal assumptions and restrictions);
- Finally, all three alternative systems lacked the flexibility to explore the role constraints play in the model's final solution (as explored in Chapter 4).

Having considered existing model development tools available to the atmospheric chemistry community, the next section of this chapter moves on to define the requirements for the OSBM.

3.3 Requirements

In this sub-section requirements are presented, detailing the way in which the user community wishes to use the OSBM. Two components are presented: first, the methodology used to capture requirements; and secondly, the requirements specification.

3.3.1 Requirements Capture Methodology

The requirements specification was first defined whilst I was becoming embedded within the atmospheric chemistry modelling community. The requirements specification evolved over the course of the OSBM development, as feedback was provided by potential users, to the state presented in the later part of this section. The specification for the OSBM was developed using two requirements capture methods: discussion with key members of the MCM user and development community; and inspection of existing models and modelling systems. The role of each these requirements capture methods is discussed in further detail below.

Discussions: Informal discussions were conducted with members of the University of Leeds Atmospheric Chemistry Modelling Research Group, in order to determine the key functionality required and to identify development priorities. These informal discussions took place with researchers and research group leaders, to ensure the requirements of these two stakeholder groups were understood. Both functional (i.e. the functionality the OSBM should present to the user) and non-functional requirements (such as ease of deployment, and technologies to be used) were captured during these informal discussions.

Inspection: By inspecting a number of models, implemented using FACSIMILE, an understanding was developed of the detailed functional requirements for the OSBM. The models inspected had generated published results and scientific insight, and had been archived by the research group in an informal file store. The owners of these models were not available to support their inspection, so a line-by-line walk-through of the model source code was required to develop a full understanding of a given model's features. The OSBM was then developed to support re-implementation of each of the models considered. The models inspected included: field models from the SOAPEX-2 [5] and TORCH-2 [6] campaigns; and chamber models from the EXACT [7] campaign.

3.3.2 Requirements Specification

Having described the requirements capture methodology in the preceding sub-section, the requirements specification is now presented.

1. **Functional scope**
   - 1.1. The OSBM should provide functionality to support the development of both chamber and field models, enabling:
     - 1.1.1. Mechanisms to be extracted from the MCM and be used directly as input;
     - 1.1.2. Species and environmental variable constraints to be implemented;
     - 1.1.3. Models to be configured without the need to edit the OSBM source code;
     - 1.1.4. Output to be obtained for species concentrations and the rates of reactions.
2. **Efficiency**
   - 2.1. The OSBM should be of comparable efficiency to FACSIMILE. The benchmark time for the execution of a field model containing the full MCM, simulating 2 days, is around 2 to 3 hours (dependent on model constraints and starting conditions).
3. **Usability**
   - 3.1. Installation: installing the OSBM should be possible without expert knowledge, on both Windows and Unix platforms.
   - 3.2. Example models should be provided.
   - 3.3. The OSBM should provide meaningful error messages, directing the user to the source of the error.
4. **Mathematical options**
   - 4.1. The OSBM should provide the option to use a variety of numerical methods. This provides the expert user with the ability to choose a solver and optimise it according the problem specification. This also provides users with the option to compare the results of several numerical methods, as a simple validity check.
5. **User interfaces**
   - 5.1. Source Code: The source code for the OSBM should be well commented and modular, to facilitate custom modifications by end users.

5.2. Graphical User Interface: The GUI should enable users, without programming skills, to access the OSBM, allowing key variables to be modified and perform basic model configuration to be performed.

5.3. Web Service: The web service should provide similar functionality to GUI.

6. **Documentation**

6.1. Comprehensive documentation should be provided enabling users to install and use the generic model with minimal effort.

This section has provided an overview of the requirements for the OSBM. The requirements specification presented above is not exhaustive, and is presented to provide an overview of the requirements of potential OSBM users.

3.4 OSBM Design

This section considers the design of the OSBM (as shown in Figure 3.1). The OSBM design was developed based upon the requirements specification presented in the preceding section of this chapter. The OSBM design emerged as the modelling functionality was iteratively developed and is presented below in its current state.

Figure 3.1 presents the OSBM architecture and consists of five components: first, a user interface layer; secondly, a model configuration layer; thirdly, a mechanism format conversion component; fourthly, a modelling logic layer, which translates the model configuration into a set of coupled ODEs (Ordinary Differential Equations); and finally, an ODE solver. Each component of the architecture is described in briefly below.

3.4.1 User Interface Layer

The user interface layer provides the user with three distinct interfaces to the OSBM; each of these interfaces is described in detail below.

Command line interface

The OSBM can be compiled and executed from the command line, with model input files edited in the user's choice of text editor. Full access is provided to the model source code, to allow an experienced modeller the opportunity to customise the OSBM to their specific requirements.

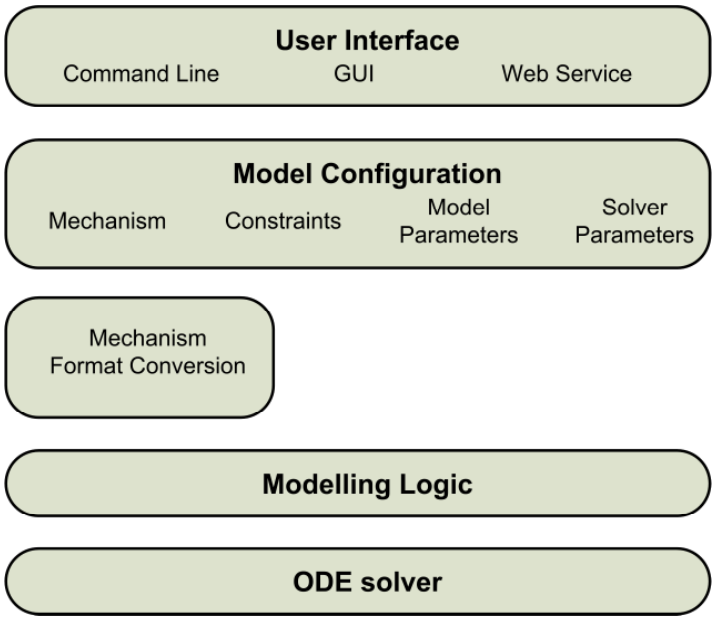


Figure 3.1: OSBM system architecture.

Graphical user interface

The OSBM can be accessed via a graphical user interface (GUI), allowing the model to compiled, executed and configured using a single application. The GUI provides a user with a simple, well-defined means of accessing the OSBM, but lacks the flexibility of the command line interface. An example of the GUI interface is presented below.

One of the main activities involved in using the OSBM is editing the chemical mechanism. The elements of the interface for developing a mechanism for a chamber model are shown below in Figure 3.2 (the interface for developing a mechanism for a field model is similar, but slightly simplified). The OSBM interface divides the chemical mechanism into six components; a tab for each component is shown across the top of the mechanism development interface screenshot Figure 3.2.

1. The main mechanism, describing the degradation of VOCs (typically a user will extract this mechanism from the MCM, and tailor it to their requirements).
2. The inorganic mechanism, describing the chemistry taking place between inorganic chemical species (again typically a user will take this mechanism from the MCM).

3. The auxiliary mechanism, describing reactions specific to the chamber being modelled (e.g. the reactions taking place on the wall of the chamber). The auxiliary mechanism will typically be determined by a set of independent, characterisation experiments within the chamber in question.
4. The dilution of stable species, describing the loss of chemical species from the chamber, due to leakage.
5. The RO_2 summation, defines a set of species (within the mechanism) as peroxy radicals [8] (e.g. CH_3O_2 and CH_3CO_3). This allows the model to output a sum of the concentrations of all the peroxy radicals, at a given time, a useful value when analysing model output data.
6. The NO_y summation, defines a set of species containing nitrogen and oxygen (e.g. $NO_2 + NO_3 + HNO_3 + HONO$). As with the RO_2 summation, the NO_y sum is useful in the analysis of model output data.

Each of these components are combined to provide a full representation of the mechanism, which the OSBM user can review prior to running the model, accessible from the "Full Mechanism" tab.

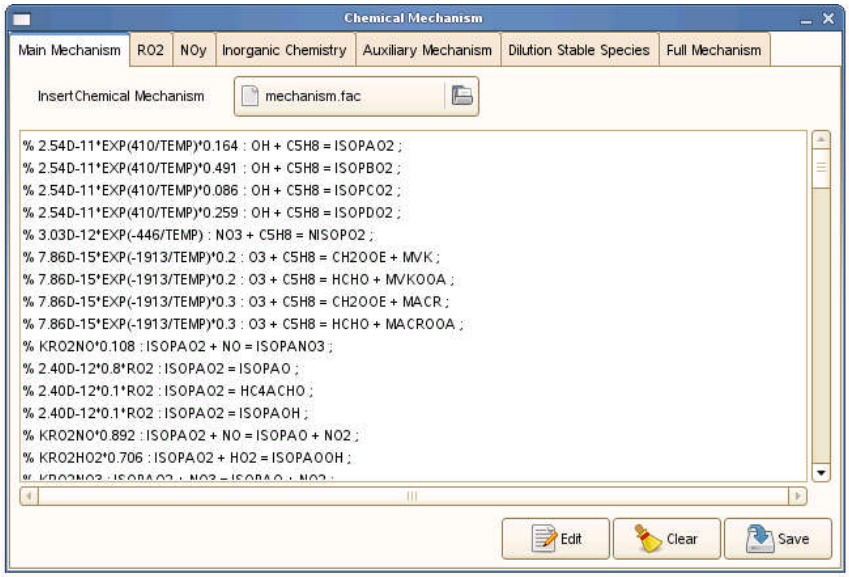


Figure 3.2: Mechanism development interface (Main Mechanism tab), allows the user to configure the mechanism describing the degradation of VOCs (within the model). The user can browse for and extract a mechanism (in a FACSIMILE format). This file will typically be downloaded from the MCM in the first instance (a section of the MCM Isoprene mechanism is shown in this case).

Web service interface

The OSBM can also be accessed using a web service, allowing an OSBM user to submit model configuration files, execute their model, and retrieve model results over the Internet. In this case the OSBM is hosted on a server; provided and maintained by the MCM support and development team. The key benefit of the web service interface is that it allows OSBM users to access modelling functionality without the overhead of installing and maintaining their own copy of the OSBM software.

3.4.2 Model Configuration Layer

The model configuration layer consists of a common representation of the model configuration, shared by the three user interfaces (as described above). As shown in Figure 3.1, the model configuration consists of four components, described below.

The mechanism: The mechanism is stored and represented in the FACSIMILE format. This format provides an intuitive representation of the mechanism that users can view and manipulate (using one of the interfaces described above). Using the FACSIMILE format for representing a chemical mechanism has the benefit of adhering to a *de facto* standard for representing mechanisms (as used by the MCM).

Constraints: The chemical species and environmental variables to be constrained are defined, alongside the constraining datasets.

Model parameters: Model parameters (other than the mechanism and constraint set) are grouped together. These model parameters include: parameters specifying the model output required; the model start and end time; the model location; etc.

Solver parameters: ODE solver parameters are grouped together, including: the type of solver to be used; and an array of parameters that determine the way in which the solver operates. It is anticipated that only experienced users will wish to make use of this functionality, so a default set of solver parameters will be provided. default set of solver parameters must be suitable for the vast majority of models, but will not necessarily deliver optimal solver efficiency.

3.4.3 Mechanism Format Conversion

The mechanism format conversion component of the architecture translates the mechanism from the FACSIMILE format to a custom numerical format. The modelling logic layer can read in this numerical format, where it is used in the construction of the set of coupled ODEs to be solved.

3.4.4 Modelling logic

The modelling logic layer combines the information contained in the model specification in order to generate a set of coupled ODEs that describe the system being modelled. This ODE system is then presented to the ODE solver interface.

3.4.5 ODE solver

The ODE solver takes the ODE system (presented by the modelling logic layer) and performs computations to determine the solution to the system (over the defined time period). This solution is then returned to modelling logic layer where it can be presented as model output via any of the user interfaces.

3.5 Model Implementation

Having described the OSBM system architecture, in the preceding section, this section describes the implementation of the OSBM; each component of the system architecture (see Figure 3.1) is revisited and its implementation described.

3.5.1 User Interface Layer

The implementation of each of the user interfaces is described briefly below.

- **Command Line:** The OSBM user has access to the model source code (i.e. the modelling logic layer), model input files, the mechanism format conversion code and a Makefile (used for compiling the model). All these resources can be edited in the user's choice of text editor.

- **Graphical user interface:** The GUI was developed by Monica Vazquez-Moreno, CEAM, using Anjunta (http://anjuta.sourceforge.net/); an integrated development environment for developing applications using C.
- **Web Service:** The web service interface is currently under development by Dr. Katarzyna Borońska, University of Leeds, School of Computing.

3.5.2 Model Configuration Layer

The model configuration is stored in a set of plain text files. These files can be edited by the OSBM interfaces, or by hand by the user. Plain text files were selected as the means of capturing and representing the model configuration, due to the simplicity of implementation. Using plain text files has a number of drawbacks including: the ease with which they can become corrupted; and their poorly defined structure. These drawbacks were accepted, in order to allow the core scientific functionality of the OSBM to be developed rapidly. An alternative approach would have been to develop an xml[4] representation of the model configuration. Using xml would have required additional development effort (compared to the use of plain text files), so the use of xml was deferred to form part of the future work.

3.5.3 Mechanism Format Conversion

The chemical mechanism is converted from the FACSIMILE format by a Python[5] script, to a numerical format, which can be read by the modelling logic layer. The conversion script accepts a restricted sub-set of FACSIMILE reaction representation.

3.5.4 Modelling logic

The modelling logic was implemented in Fortran90; the rationale for this decision is presented below.

- Modelling tools, previously developed within the University of Leeds, Atmospheric Chemistry Modelling Research Group, implemented some of the functionality required by the OSBM. This presented an opportunity to reuse

[4] XML (Extensible Markup Language) is a general-purpose specification for creating custom markup languages.
[5] Python is a scripting language, further details are available from http://www.python.org/

functionality and source code, to facilitate the development of the OSBM. The code selected for re-use was written in Fortran77.

- Fortran is widely used across scientific research communities, including the atmospheric chemistry modelling community (whilst other programming languages such as C, Java ect. are less widely used, for a variety of cultural reasons). So using Fortran to develop the OSBM will increase the chances of allowing users to transfer their existing knowledge, easing the transition to a new modelling tool.
- As Fortran is widely used across the scientific community as a whole, a wide variety of libraries exist to perform common numerical tasks (such as solving coupled ODE systems, of the type that describe an atmospheric chemistry model).

Therefore, although there are some significant drawbacks to the use of Fortran, mainly related to its relative age (compared to modern programming language such as Java and C#), a compelling case was presented for the use of Fortran90.

3.5.5 ODE Solver

The ODE solver used in the implementation of the OSBM was CVODE [9], part of the Sundials suite of solvers [10]. CVODE was selected for the following reasons.

- CVODE is specifically designed to solve coupled sets of ODE of the form a box model is translated to. This mathematical form is shown in equation E3.1.
- CVODE has interfaces in C and Fortran, allowing flexibility of language choice for future developments.
- CVODE is freely available software, with extensive support and documentation provided by the Centre for Applied Scientific Computing, Lawrence Livermore National Laboratory.
- CVODE is provided under The Sleepycat License (http://www.opensource.org/licenses/sleepycat.php). So CVODE can be freely redistributed as part of other applications (i.e. the OSBM), providing the source code for the application is made freely available.
- CVODE is well documented, allowing OSBM users to explore the configuration of the solver, if motivated to do so.

A box model can be considered as an ODE initial value problem. This is expressed mathematically as:

$$\dot{y} = f(t, y), \quad y(t_0) = y_0 \; ; \qquad \text{(E3.1)}$$

where $\dot{y} = \frac{\partial y}{\partial t}$, $y \in \Re^n$, n is the number of species being modelled and t is the independent variable (in this case time).

CVODE [9] uses a multi-step method with variable-step size and variable order, in the form:

$$\sum_{i=0}^{K_1} \alpha_{m,i} y^{m-i} + h_m \sum_{i=0}^{K_2} \beta_{m,i} \dot{y}^{m-i} = 0 \; . \qquad \text{(E3.2)}$$

where y^m are approximations to $y(t_m)$ and the step size $h_m = t_m - t_{m-1}$, the size of the m^{th} time step.

The choice of multi-step method is dictated by the stiffness of the problem being solved. Stiffness can be considered the property of a system with at least one rapidly damped mode, which has a small time constant relative to the system solution timescale. The observable manifestation of stiffness in a box model is that some chemical species have short atmospheric lifetimes, e.g. OH and HO_2, in the order of seconds; whereas the system solution timescale (i.e. the time period being modelled) is days or weeks (when modelling field campaigns).

More basic numerical methods for solving ODE systems, such as Runge-Kutta [11], cope badly with stiff systems. This is because the time step is determined by the rapidly damped mode, in physical terms the chemical species with the short atmospheric lifetimes. So the time step remains small relative to the system solution time-scale and the numerical method will prove inefficient.

This section has described the implementation of the OSBM. The OSBM source code[6] can be accessed on the CD associated with this thesis; the next section of this chapter moves on to discuss the testing of the OSBM.

3.6 Model Testing

This section presents the experiments conducted in order to gain confidence in the accuracy of the results produced by the OSBM. A number of comparisons have been performed between OSBM and FACSMILIE models, the results presented here consider two such comparisons: first, a simple field study model (SOAPEX-2); and secondly a more complicated field study model (TORCH-2). This section is presented in two parts: first background is provided for the two field campaigns in question; and secondly, the results of the model comparisons are presented.

3.6.1 SOAPEX-2 Background

The second Southern Ocean Atmospheric Photochemical Experiment (SOAPEX-2) took place in the austral summer, January 18th to February 18th 1999, at the Cape Grim Atmospheric Pollution Station, Tasmania (Australia). The core campaign objective was to study free-radical chemistry in the remote marine environment. A full description of the campaign site can be found in Roberto Sommariva's PhD thesis [12].

The model considered is used to simulate atmospheric chemistry taking place over February 7th and 8th, two of four baseline condition days that occurred during the campaign. Baseline conditions occur when the prevailing wind direction is West to South-West; calculated back trajectories show that the air reaching Cape Grim, on baseline days, had not travelled over land for at least five days [5]. On baseline days the lowest NO_x (where $[NO_x]= [NO] + [NO_2]$) and VOC concentrations of the campaign were measured. The model itself is a re-implementation (using the OSBM) of the simplified model used for exploring the OH and HO_2 chemistry for SOAPEX-2 [5]; the original model was implemented with FACSIMILE [2].

[6] This source code has been re-engineered by Dr. Katarzyna Borońska to improve its quality prior to the release of the OSBM as an open source project.

The model incorporates the inorganic mechanism, the CO and CH_4 oxidation mechanisms from the MCMv3 [5], heterogeneous loss reactions and dry deposition reactions. Full details of the mechanism can be found in Sommariva et. al. 2004 [5]. The model is constrained by field data for environmental conditions $J(O^1D)$, $J(NO_2)$, temperature, $[H_2O]$, declination angle) and species concentrations (NO, NO_2, HCHO, O_3, CO, CH_4). Here $J(O^1D)$ is the rate coefficient for photolysis reaction R3.1 and J (NO_2) is the rate coefficient for reaction R3.2.

$$O_3 + h\nu \rightarrow O^1D + O_2 \qquad \text{(R3.1)}$$

$$NO_2 + h\nu \rightarrow NO + O^3P \qquad \text{(R3.2)}$$

3.6.1.1 TORCH-2 Background

The TORCH-2 (Tropospheric ORganic CHemistry experiment) campaign took place during the April and May of 2004 at the Weybourne Atmospheric Observatory (http://weybourne.webapp2.uea.ac.uk/index.html). The main objective of the TORCH-2 campaign was to develop an understanding of: the composition of, and chemical processes occurring within, polluted air packets travelling from London across East Anglia to Weybourne. Unfortunately, during the campaign the prevailing wind direction was typically from the North-East, so the majority of air packets (including those modelled in this case study) came off the North Sea.

The model considered is used to simulate the atmospheric chemistry taking place over the 4^{th}-8^{th} May 2004. Again a model was developed using the OSBM, based upon a FACSIMILE model developed within the University of Leeds Atmospheric Chemistry Modelling Group. The mechanism incorporated in the model is substantially more complex than that considered in the SOAPEX case study (including the degradation pathways of 28 VOCs). The constraints of the model were also substantially more complex than those implemented in the SOAPEX model, including: 32 chemical constraints; 8 photolysis rate constraints; and, 4 environmental condition constraints. Further details of the model implementation and the mechanism used can be found in the PhD thesis of Jenny C. Stanton [6].

3.6.2 Results

For both OSBM-FACSIMILE comparisons, model output is compared for a set of key species (i.e. those a modeller is typically likely to be interested): OH, HO_2, CH_3O_2, HNO_3 and HONO. The model comparison focuses on the OH, HO_2 radicals because the objective of the original models was to develop understanding of the chemistry of these radicals. CH_3O_2, HNO_3 and HONO are also considered as examples of the wider set of species being modelled.

3.6.2.1 SOAPEX-2 Results

The results for the comparison of SOAPEX-2 model results are presented in this sub-section in three parts.

OH concentration comparison: The OH comparison is presented in Figure 3.3, and the OSBM and FACSIMILE results appear to be practically identical.

HO_2 concentration comparison: The HO_2 comparison is presented in Figure 3.4, and the OSBM and FACSIMILE results again appear to be practically identical.

Ratio comparison: The ratio comparison is presented in Figure 3.5 for OH, HO_2, CH_3O_2, HNO_3 and HONO. The OH and HO_2 ratios show that, whilst the values produced by FACSIMILE and the OSBM are not identical, they are indeed very similar: their agreement is within 1% throughout the course of the two day model run. The largest fractional difference in OH concentrations occurs at the start of the model run, the time when the model output is most sensitive to the configuration and behaviour of the ODE solver. The ratio for CH_3O_2 shows a profile very similar to the profiles of the OH and HO_2 ratios. The level of agreement for HNO_3 and HONO is lower, but remains within 2%. For HONO the level of agreement is better at night-time and worse during the day. This relates to the presence of higher HONO concentrations during the night (it is readily photolysed during the day) and very low concentrations during the day. Conversely, HNO_3 agreement is better during the day and worse during the night-time; this relates to the presence of higher HNO_3 concentrations during the day, and very low concentrations during the night. This model comparision demonstrates that differences between FACSIMILE and the OSBM results are greatest when species concentrations are at their

lowest; which is potentially due to differences in the solver tolerances and differences in the ODE solvers themselves.

Origin of differences between the OSBM and FACSIMILE models

The differences between the results are not significant (< 2%), given the considerable uncertainties involved in the models considered. Potentially causes of the differences include:

- Differences between the algorithms used to solve the ODEs;
- Differences between the compiler optimisation strategies, for each model, leading to differences in rounding errors;

Given the relative simplicity of the SOAPEX-2 model and the relatively small difference in the model results, it is unlikely (but possible) that the differences in model results are caused by differences in the model configurations.

3.6.2.2 TORCH-2 Results

The results for the comparison of TORCH-2 models are presented in this sub-section in three parts.

OH comparison: The OH comparison is presented in Figure 3.6. The OSBM and FACSIMILE results qualitatively appear to be very similar, but the agreement is not quite so good as the equivalent results for the SOAPEX-2 model.

HO_2 comparison: The HO_2 comparison is presented in Figure 3.7, and the OSBM and FACSIMILE results qualitatively appear to be very similar, but the agreement is again not quite as good as the equivalent results for the SOAPEX-2 model.

Ratio comparison: The ratio comparisons are presented in Figure 3.8: for OH, HO_2 and CH_3O_2, HNO_3 and HONO. The OH and HO_2 ratios show, that whilst the values produced by FACSIMILE and the OSBM are similar, the level of agreement is generally within 5%, but there are some outlying points for the OH ratio of up to 15%.

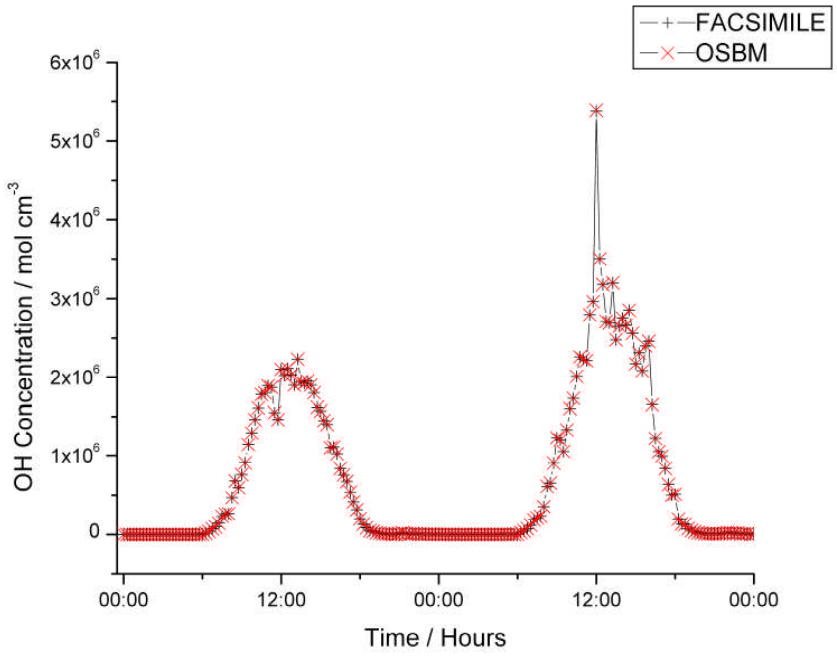


Figure 3.3: FACSIMILE-OSBM comparison of [OH], 7th-8th February 1999, Australian Eastern Standard Time (AEST).

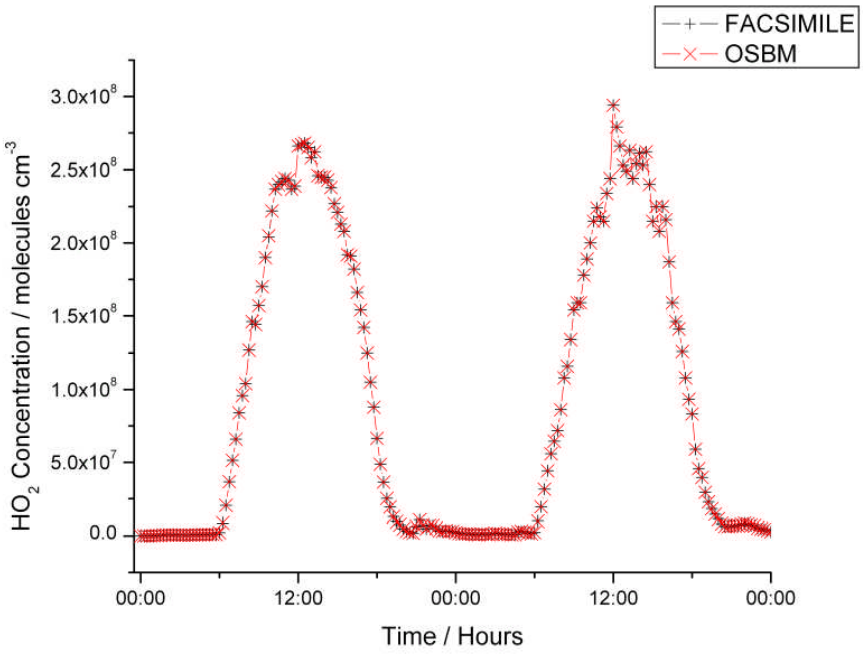


Figure 3.4: FACSIMILE-OSBM comparison of [HO_2], 7th-8th February 1999, AEST.

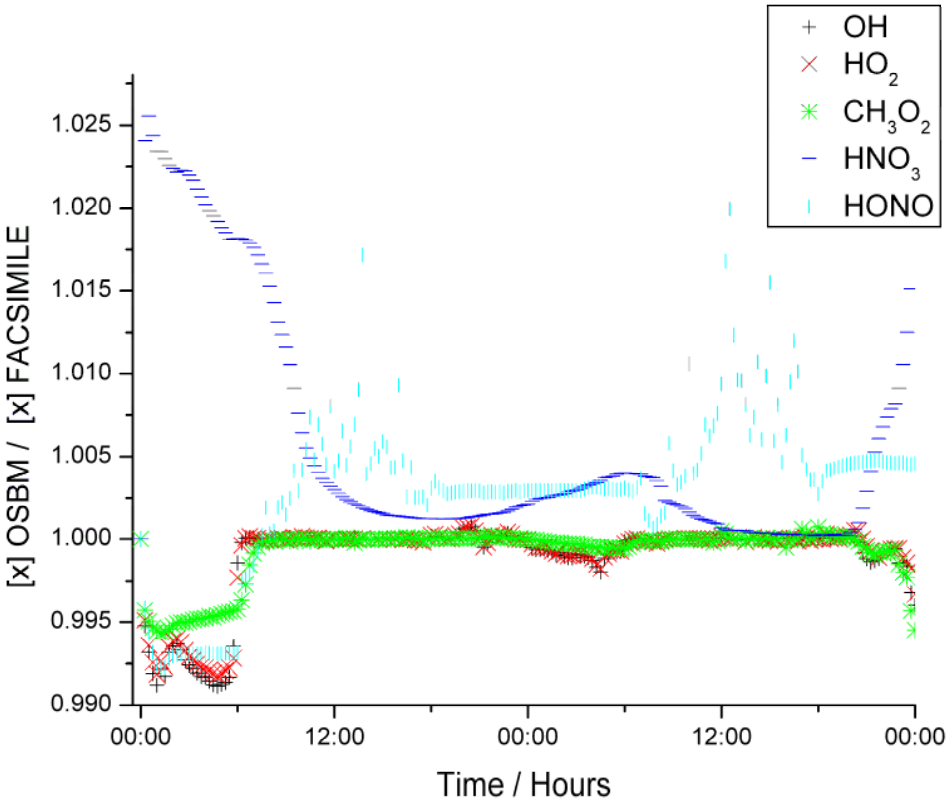


Figure 3.5: OSBM-FACSIMILE concentration ratio comparison of OH, HO_2, CH_3O_2, HNO_3, HONO, 7th-8th February 1999, AEST.

For both OH and HO_2 the ratios show better agreement during the day, than the night, again relating to low radical concentrations during the night. The level agreement for CH_3O_2, HNO_3 and HONO also typically remains within 5%; although there are a small number of outlying points up to 10%, particularly for CH_3O_2.

Origin of differences between the OSBM and FACSIMILE models: The differences between the results, although larger than the differences between the SOAPEX-2 models, are not significant (generally < 5%), given the uncertainties involved in the model considered. Potential causes of the differences, in addition to the potential causes identified for the SOAPEX-2 model, include differences between the model configurations in terms of the chemical mechanism and the constraint data used. The TORCH-2 model is significantly more complex than the SOAPEX-2 model, in terms of the chemistry used and the constraints applied.

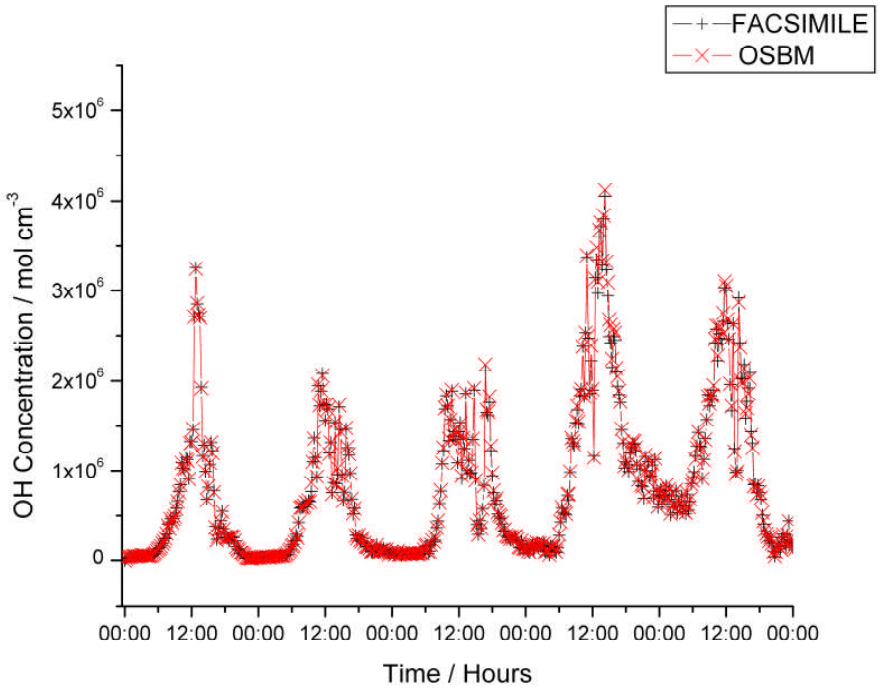


Figure 3.6: FACSIMILE-OSBM comparison of [OH], 4th-8th May 2004, British Summer Time (BST).

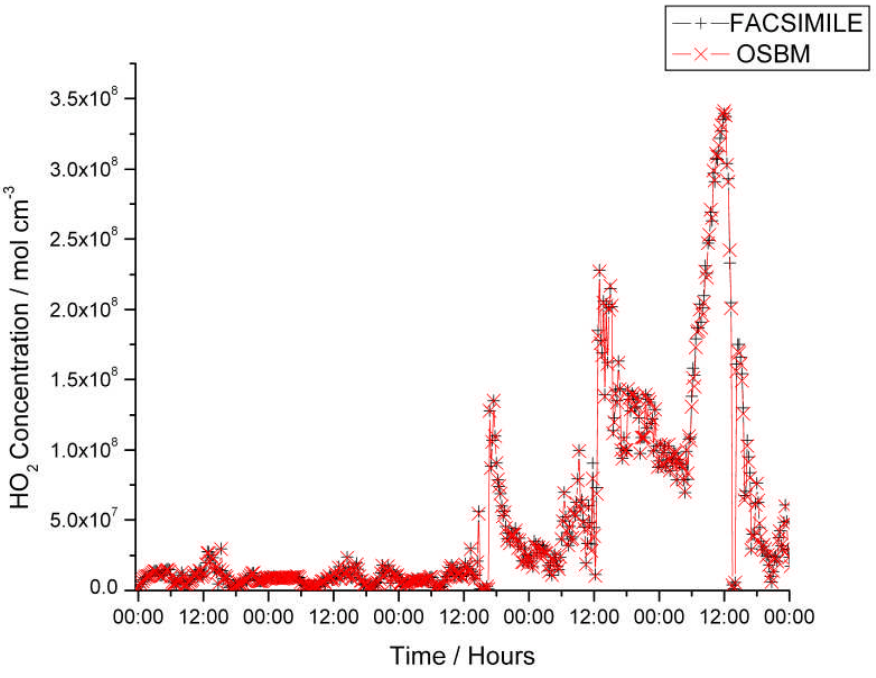


Figure 3.7: FACSIMILE-OSBM comparison of [HO_2], 4th-8th May 2004, BST.

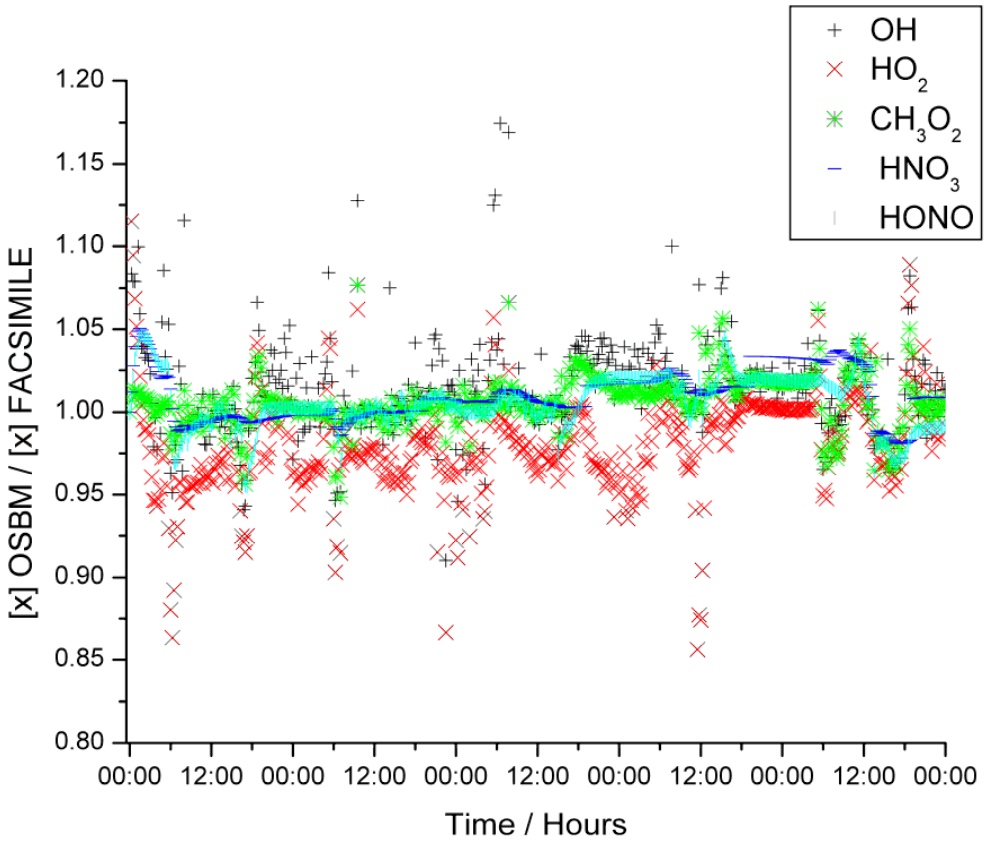


Figure 3.8: OSBM-FACSIMILE concentration ratio comparison of OH, HO_2, CH_3O_2, HNO_3, HONO, 4th-8th May 2004, BST.

Although every feasible effort was made to ensure that the FACSIMILE and OSBM models had identical configurations, it is likely that human error played a role in causing the differences between the model results. This highlights the difficultly in benchmarking modelling tools, for complex systems, where a substantial challenge is presented in producing identical model configurations.

This section has provided an overview of some of the testing conducted in order to establish confidence in the output of the OSBM. Two cases have been considered, field models of different complexity, benchmarking the OSBM against the well-used FACSIMILE system.

3.7 Progress Towards Meeting the OSBM Requirements Specification

The section briefly describes the progress made towards meeting the requirements specification, outlined in Section 3.3.2.

1. **Functional Scope**

   All core OSBM functionality has been implemented. A beta version of the OSBM is currently used by three members of the MCM-user community; who provide feedback in the form of suggested minor enhancements and fixes.

2. **Efficiency**

   For the SOAPEX-2 model the OSBM is approximately a factor of 10 slower than equivalent FACSIMILE model (model runtimes of approximately 300 vs. 30 seconds). For the TORCH-2 model the OSBM is approximately a factor of 2 slower than equivalent FACSIMILE model (model runtimes of approximately 20 hours vs. 10 hours). Further work is required to understand this difference in performance and optimise the OSBM accordingly.

3. **Usability**

   The OSBM can be installed, on both Windows and Unix platforms. Example models and error handling are currently in the early stages of development.

4. **Mathematical options**

   The OSBM provides the expert user with option to interact with CVODE, tailoring the solver configuration to meet their specific requirements.

5. **Interfaces**

   The source code interface has been successfully implemented and tested, the GUI and web service interfaces are currently being prototyped.

6. **Documentation**

   Comprehensive documentation has yet to be developed.

3.8 Future Work

This section provides an overview of potential future work associated with the development of the OSBM. In addition to work required to meet the currently unaddressed components of the requirement specification, there are three main areas of future work, as discussed below.

Photo-chemical trajectory model: A photo-chemical trajectory is similar to the static box models considered in this chapter, but rather than modelling the air packets arriving at a given location it models an air packet travelling from one point to another. So, the box follows a defined trajectory, and the contents of the box are determined by the chemical reactions taking place and the emissions of chemical species from the ground below (as defined by an emissions inventory). Examples of photo-chemical trajectory models that could be used to inform the design and development of an Open Source Photo-Chemical Trajectory Model include those developed by R.G. Derwent [14] [15], which incorporated the MCM.

Data analysis services: Currently MCM users typically take model output (from either FACSIMILE or the OSBM) and manually plot graphs (as an initial method of data analysis), using a package such as Microsoft Excel or Origin. Producing graphs in this manner is time consuming and error prone; so future work will develop tools that will integrate with the OSBM to automatically plot graphs and facilitate data analysis. Other methods of data analysis, including rate of production and loss analysis [5], are typically performed by customised scripts; future work will develop these scripts to ensure compatibility with the OSBM and make them sufficiently robust to be distributed with the MCM, for use by the community.

Composing workflows within in an e-Science environment: Once data analysis services have been developed, the possibility emerges of composing scientific workflows, using a workflow management tool such as Taverna [16]. This would further improve the researcher experience of the model development process, by enabling access to, and integration with, a variety of e-Science tools (e.g. MyExperiment [17]).

Chapter Summary

This chapter has presented an overview of the development of an Open Source Box Model, a tool designed to encourage uptake of the MCM across the atmospheric chemistry community; by providing an easy to use, flexible, open source model development tool that integrates with the MCM. The understanding of the model development process, that I gained during this work, formed the basis of the research presented in the remainder of

this thesis. In the next section the OSBM is used to explore the impact of constraint methodology, research that would have not been feasible with the tools that pre-date the OSBM.

References

1. Wiesen, P., *The EUROCHAMP Integrated Infrastructure Initiative Environmental*, in *Environmental Simulation Chambers: Application to Atmospheric Chemical Processes*. 2006. p. 295-299.
2. *ESM Software - Facsimile*. [cited 12th February 2009]; Available from: http://www.esm-software.com/facsimile/.
3. Damian, V., et al., *The kinetic preprocessor KPP-a software environment for solving chemical kinetics.* Computers & Chemical Engineering, 2002. **26**(11): p. 1567-1579.
4. Carver, G.D., P.D. Brown, and O. Wild, *The ASAD atmospheric chemistry integration package and chemical reaction database.* Computer Physics Communications, 1997. **105**(2-3): p. 197-215.
5. Sommariva, R., et al., *OH and HO2 chemistry in clean marine air during SOAPEX-2.* Atmos. Chem. Phys., 2004. **4**(3): p. 839-856.
6. Stanton, J.C., *Field and Modelling Studies of Volatile Organic Compunds in the Troposphere*, in *School of Chemistry*. 2006, University of Leeds.
7. Bloss, C., et al., *Evaluation of detailed aromatic mechanisms (MCMv3 and MCMv3.1) against environmental chamber data.* Atmos. Chem. Phys., 2005. **5**(3): p. 623-639.
8. Tyndall, G.S., et al., *Atmospheric chemistry of small organic peroxy radicals.* J. Geophys. Res. **106**.
9. Cohen, S.D. and A.C. Hindmarsh, *CVODE, a stiff/nonstiff ODE solver in C.* Comput. Phys., 1996. **10**(2): p. 138-143.
10. Hindmarsh, A.C., et al., *SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers.* ACM Trans. Math. Softw., 2005. **31**(3): p. 363-396.
11. Lambert, J.D., *Numerical methods for ordinary differential systems: the initial value problem*. 1991: John Wiley \& Sons, Inc. 293.
12. Sommariva, R.C., *Understanding Field Measurements through a Master Chemcial Mechanism*, in *School of Chemistry*. 2004, University of Leeds.
13. Heard, D.E., et al., *The North Atlantic Marine Boundary Layer Experiment (NAMBLEX). Overview of the campaign held at Mace Head, Ireland, in summer 2002.* Atmos. Chem. Phys., 2006. **6**(8): p. 2241-2272.
14. Derwent, R.G., M.E. Jenkin, and S.M. Saunders, *Photochemical ozone creation potentials for a large number of reactive hydrocarbons under European conditions.* Atmospheric Environment, 1996. **30**(2): p. 181-199.
15. Derwent, R.G., et al., *Photochemical ozone creation potentials for organic compounds in northwest Europe calculated with a master chemical mechanism.* Atmospheric Environment, 1998. **32**(14-15): p. 2429-2441.
16. Oinn, T., et al., *Taverna: a tool for the composition and enactment of bioinformatics workflows.* Bioinformatics, 2004. **20**(17): p. 3045-3054.
17. Goble, C. and D. De Roure, *myExperiment: social networking for workflow-using e-scientists*, in *Proceedings of the 2nd workshop on Workflows in support of large-scale science*. 2007, ACM: Monterey, California, USA.

Chapter 4 Exploring Constraint Implementations

This chapter explores the impact of constraint implementations on the modelling of radical concentrations, in zero-dimensional box models, at high time resolution[7]. The research presented in this chapter was conducting using the OSBM, as described in Chapter 3, and focuses on the impact of constraint implementation on the SOAPEX-2 box model, introduced in Chapter 3. This chapter consists of three sections: first, the way in which constraints are implemented is described; secondly, a series of tests, with simplified box models, are examined in order to understand the impact of constraint implementations; and thirdly, the impact of constraint implementation is examined for a complete box model.

4.1 Constraining Box Models

The role of constraints in the development of zero-dimensional box models was introduced in Chapter 2. This section provides a detailed description of how box models are constrained.

4.1.1 Implementation of Constraints

This sub-section discusses the implementation of constraints and establishes terminology for the domain. A constraint implementation has two components, discussed in detail below.

Constraint data frequency: This is the frequency of the time series for a given constraint used as an input to the model; it is worth considering an example to clarify this statement. The source experimental data for [NO], in the SOAPEX-2 campaign [1], has a frequency, limited by the measurement method, of 1 minute. These data could be used directly as constraint data, giving a constraint data frequency of 1 minute. Alternatively the source data could be averaged or sampled, to give data sets with lower frequencies. For example a 15 minute average/sampled dataset used as constraint input gives a constraint frequency

[7] producing modelled radical concentrations that exhibit realistic behaviour on timescales of less than 15 minutes

of 15 minutes. The frequency at which constraint datasets are measured varies, from 1 minute for [NO], to over an hour for some VOCs, dependent on the experimental technique used.

Constraint interpolation method: When a box model runs it requires values for constrained species or variables at times which are not included in the constraint dataset. For example if the [NO] constraint data frequency is 15 minute, with the dataset starting at 09:00, and the model requires [NO] at 11:10 the model must determine an appropriate value for [NO] based on the (time, value) pairs in the constraint data set. Determining this intermediate value, between data set points, is achieved by the constraint interpolation method.

4.1.2 Typical Constraint Implementation

The constraint implementation typically used in atmospheric chemistry box models [1, 2] is to average to 15 minute intervals. These 15 minute datasets are then interpolated, using piecewise constant interpolation, at model runtime to generate a value for the constrained parameter, at a given time (as determined by solver step size etc.). This leads to the stepped profile, as seen in Figure 4.1, between data points.

This typical constraint implementation loses a significant amount of information about how a physical quantity varies on a sub 15 minute timescale. This is particularly relevant for the rapidly changing and highly variable constraints, such as photolysis rates. An example of the source data and 15 minute averaged piecewise constant interpolation data is presented in Figure 4.1, for $J(NO_2)$ data from the SOAPEX-2 campaign, including the ratio of the two data sets at each minute interval.

Figure 4.1 shows that the differences for the source data introduced by averaging and interpolation, using the constraint implementation described above, are significant. In the 90 minute time interval considered the difference between the source and processed data is 20% or greater at 40 of the 90 minute points. And this difference is 50% or greater at 7 of the 90 minute points. Differences of this magnitude in the model constraints are likely to lead to significantly different model results. The remainder of this chapter seeks to demonstrate the impact of constraint methodology on model results.

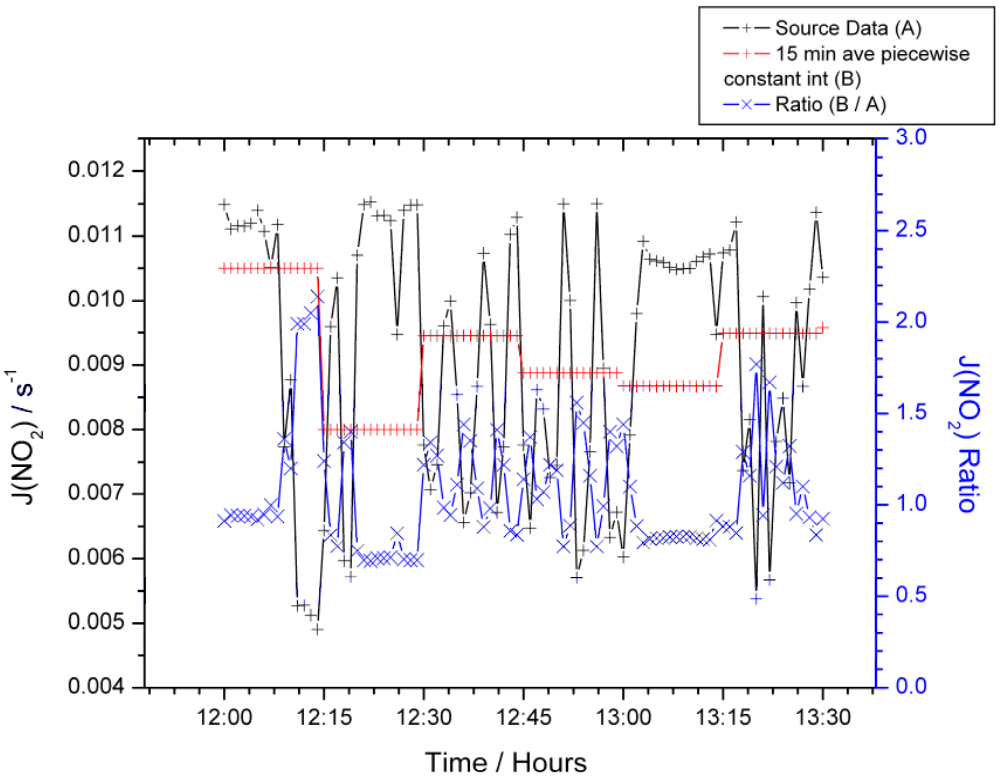


Figure 4.1: Constraint implementation example, data from the SOAPEX-2 campaign (January 18th 1999). The graph show $J(NO_2)$ source data (at a 1 minute interval), $J(NO_2)$ averaged to 15 minute interval with piecewise constant interpolation, and the ratio of theses two data sources (i.e. averaged data / source data).

4.1.3 Constraint Implementations to be Explored

This sub-section describes the constraint implementations to be explored, in this chapter, and consists of two components: first, a discussion of the constraint frequencies considered; and secondly, a discussion of the interpolation methods considered.

4.1.3.1 Constraint Frequencies

Source data for each of the SOAPEX-2 model constraints were retrieved from the British Atmospheric Data Centre (BADC). The BADC is responsible for the archiving of field campaign data, ensuring its availability to the atmospheric chemistry community. In the remainder of this chapter two constraint frequencies are considered: 15 minute averaged;

and source specific; the details of both of these constraint frequencies are described below. Figure 4.2 shows these constraint frequencies for $J(NO_2)$ constraint data for the SOAPEX-2 model.

- **15 minute averaged:** Where measurements are available more frequently than 15 minute intervals, an average (7 minutes forward, 7 minutes back) is calculated at each 15 minute interval. Where measurements are less frequent than 15 minutes, linear interpolation is used to generate 15 minute data points. Where known errors occur in the source data (and are flagged) they are disregarded from the averaging.
- **Source specific:** Datasets are used at the frequency at which they are measured; details for environmental and chemical constraints, for the SOAPEX-2 model, are given in Table 4.1. The frequency of measurement reflects the time resolution of the experimental technique used. Where known errors occur in the source data, gaps are left to be addressed by the interpolation method.

Species	**Source Data Time Interval (mins)**	**Environmental Conditions**	**Source Data Time Interval (mins)**
O_3	1	$J(O^1D)$	1
NO	1	$J(NO_2)$	1
NO_2	1	H_2O	15
CH_4	40	Temperature	1
CO	40		
HCHO	60		

Table 4.1: Measurement frequency for constrained species and environmental conditions

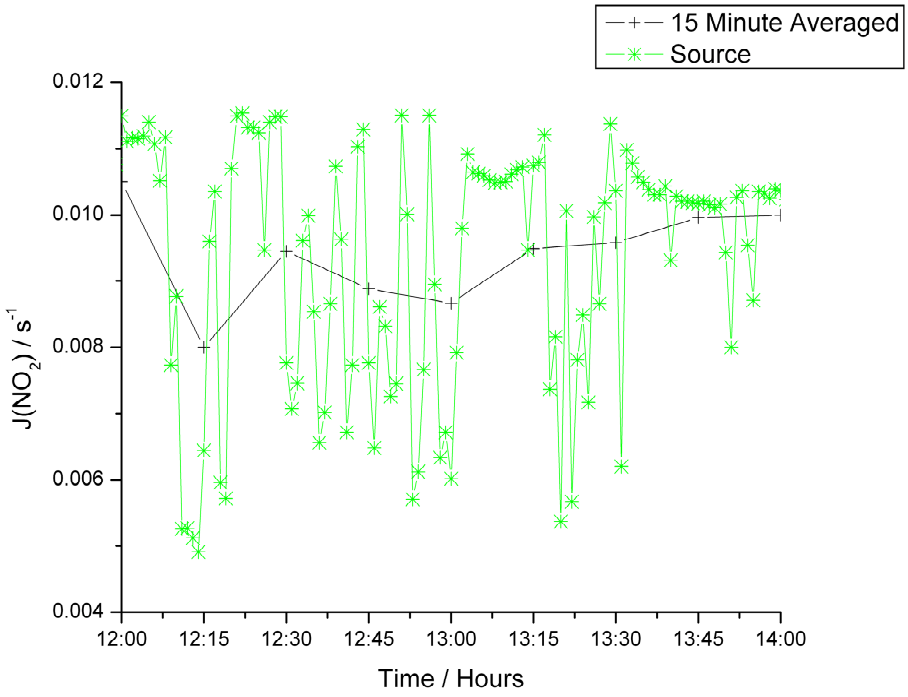


Figure 4.2: $J(NO_2)$ constraint frequency comparison (from SOAPEX-2, February 18th 1999); 15 minute averaged and source frequency (1 minute in this case). The variability in the source data is a result of clouds, passing over the measurement site, and absorbing solar irradiation.

4.1.3.2 Interpolation Methods

The impact of the following interpolation methods are investigated in this chapter: piecewise constant (as discussed in Section 4.1.2), piecewise linear, cubic spline [3]. Two varieties of cubic spline are investigated, one fitted through the source data and the other fitted through the natural logarithm of the source data (physical values are then given by the inverse natural log of the interpolated point). A cubic spline through the natural log of the source data (referred to a cubic spline (ln) in this chapter) is used as crude method of ensuring that only positive concentrations are interpolated (interpolation to give negative concentration leads to solver errors and a corruption of the physical system). Each of the interpolation methods is shown in Figure 4.3, for $J(NO_2)$ data taken from the SOAPEX-2 campaign. The plot shows the differences between the interpolation methods in terms of the value they return at a given time; these differences are particularly evident when comparing piecewise constant with the other methods.

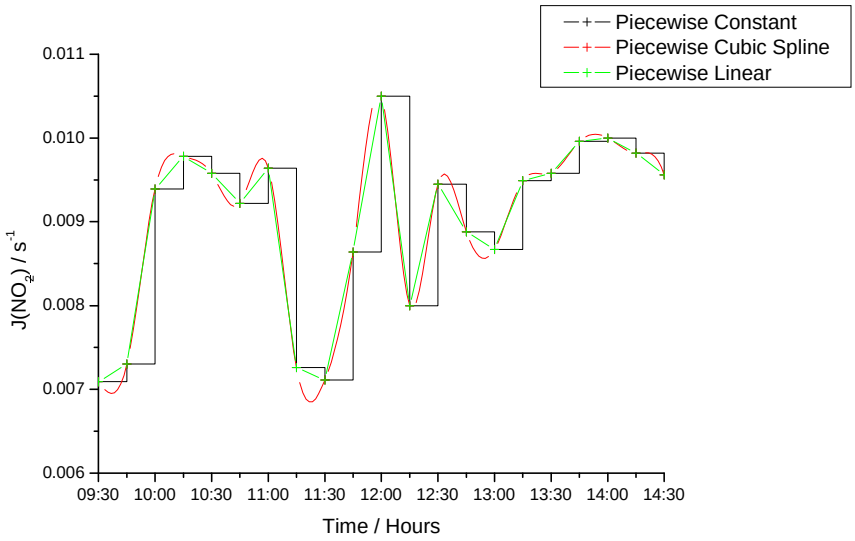


Figure 4.3: Three interpolation methods on 15 minute averaged $J(NO_2)$ constraint data (from SOAPEX-2, February 18th 1999); piecewise constant, cubic spline, piecewise linear.

Relating interpolation methods to the properties of the physical system

A key consideration in selecting an interpolation method is whether or not the interpolated points makes sense in the context of the physical system. It is worth examining this relationship for each of the interpolation methods in turn.

- ***Piecewise constant:*** Chemical concentrations and environmental constraint quantities are assumed to remain constant between data points and then change instantaneously. This contradicts the continuous nature of change for the physical and chemical quantities in question.
- ***Piecewise linear:*** This approach addresses the instantaneous change issue caused by piecewise constant interpolation; the interpolated points lie on the straight line (determined by the formula y = mx + c) between each enclosing pair of data points. Linear interpolation leads to the *rate of change* of the quantity changing discontinuously (at data points), which again contradicts (but to a lesser extent) the continuous nature of change of the physical quantities in question. The results later in this chapter suggest that this discontinuity in the rate of change does not have a significant impact on either model results or efficiency.

- ***Cubic spline:*** Addresses both the issues of instantaneous quantity change and discontinuities in rate of change. However, this approach allows the interpolated points to lie outside local/global observed data points, giving potential for unrealistic behaviour.

4.2 Solution Recovery Tests

This section describes the testing conducted in order to establish the impact of constraint implementation; solution recovery tests are considered for two simplified systems. A solution recovery test, as shown in Figure 4.4, consists of the following steps.

- An initial model run (with no species constraints), which produces a baseline concentration output dataset;
- Processing the baseline concentration output dataset, to form concentration constraint sets: for each species to be constrained; for each of the constraint frequencies being considered;
- Applying each of the concentration constraint sets to the original model and running the model with each of the interpolation methods being considered (i.e. a solution recovery run);
- The concentration output of each of the solution recovery runs can be compared with the baseline concentration output dataset to establish the performance of the constraint implementation used.

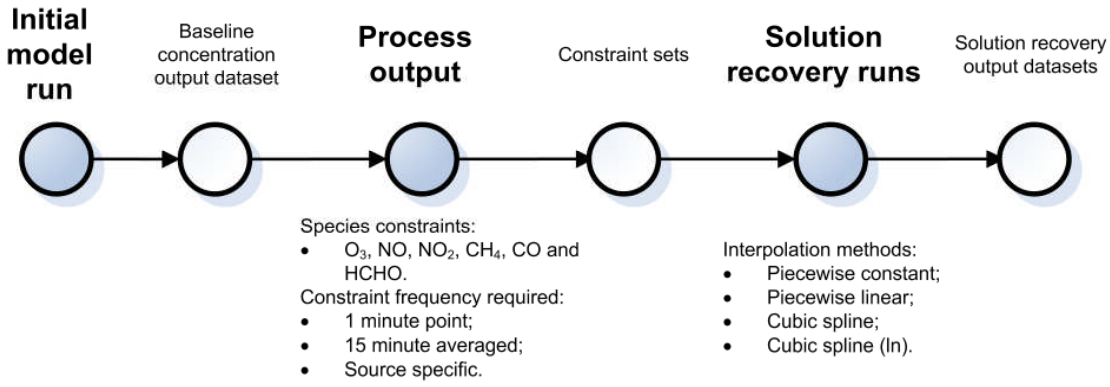


Figure 4.4: The process of executing a solution recovery test.

The solution recovery tests are now discussed in detail: first, the solution recovery test method is discussed; secondly, the results of the solution recovery tests are presented; and thirdly, the conclusions drawn from the solution recovery tests are presented.

4.2.1 Solution Recovery Test Method

The method for conducting solution recovery tests was outlined above. This sub-section provides additional detail on the solution recovery tests: first, the comparison of results from the initial model run and subsequent solution recovery model runs is described; secondly, solution recovery tests for a simplified, unconstrained version of the SOAPEX-2 model are described; thirdly, solution recovery tests for a more realistic, constrained version of the SOAPEX-2 model are described.

4.2.1.1 Unconstrained Model Tests

This sub-section describes the first set of solution recovery tests conducted for a simplified, unconstrained version of the SOAPEX-2 model.

Initial Run: The model used for the initial model run for was a simplified version of the SOAPEX-2 model. All constraints were removed, environmental conditions were calculated where necessary (photolysis rates, temperature) or assumed a fixed value (e.g. $[H_2O]$, declination), all chemical concentrations were solved for by the OSBM. The initial conditions of the model, initial chemical concentrations and fixed environmental conditions, were as the full SOAPEX model.

Processing of baseline concentration dataset: The baseline concentration dataset was processed to form constraint sets for the following chemical species: NO, NO_2, CH_4, CO, HCHO, O_3. For each of the following constraint frequencies: 15 minute averaged, 1 minute, source specific.

Solution recovery runs: The constraint datasets for each constraint frequency were applied, in turn, to the initial model. For each constraint frequency the model was then run for each of the following interpolation methods: piecewise constant, piecewise linear, cubic spline, and cubic spline (ln).

4.2.1.2 Condition-Constrained Model Tests

This sub-section describes the second set of solution recovery tests conducted. This second set of tests repeats the unconstrained model tests with one key difference: all model runs were constrained for a set of environmental conditions. Constraining environmental conditions adds complexity to the solution recovery tests, by producing a baseline concentration dataset with more variability (than the unconstrained case), so presenting a more challenging test to the constraint implementations being tested.

Initial Run: The initial model was constrained for $J(NO_2)$, $J(O^1D)$, $[H_2O]$ and temperature at a source specific constraint frequency and interpolation is performed using a cubic spline (ln). This constraint implementation for the environmental conditions was made based on the combination I thought provided the closest match to the physical system. The processing of the baseline concentration dataset and the solution recovery runs then proceeded as described for the unconstrained solution recovery tests.

4.2.1.3 Comparison of Model Output

The baseline concentration output dataset could be compared to the solution recovery output datasets in a number of ways. In this chapter results are compared using the OH and HO_2 concentrations, the concentrations of other species being modelled are not considered. Two factors played a key role in the selection of OH and HO_2 radical concentrations as the basis upon which to make comparisons: first, OH and HO_2 are amongst the species with the shortest atmospheric lifetimes, and so likely to respond to changes in constraint implementation over a short timescale; and secondly, radical chemistry is an important topic in atmospheric chemistry, which is often the focus of *in situ* experiments and model development (including the SOAPEX-2 field campaign and model considered in these solution recovery tests).

4.2.2 Solution Recovery Test Results

The results of the solution recovery tests are presented in three parts: first, the unconstrained model results, for radical concentrations; secondly, the conditions-constrained model results, for radical concentrations; and thirdly, the impact of the

constraint implementation on the model efficiency (i.e. how long the model takes to run) for both the unconstrained and the conditions-constrained model.

4.2.2.1 Unconstrained Model Test Results

The results of the unconstrained model solution recovery tests are shown below (see Figure 4.5 to Figure 4.10). Each of the graphs, presents solution recovery results for a given constraint frequency; with the ratio of the solution recovery results to the baseline results, plotted for each interpolation method. Four key results from the unconstrained model solution recovery results are presented below.

Dawn and dusk: In all tests the largest errors in solution recovery occur at dawn (approximately 06:00) and dusk (approximately 19:00). This is at a time when radical concentrations are changing most rapidly, and are relatively small (compared to peak values). Under these circumstances the impact of the interpolation method is accentuated, leading to larger relative error peaks (around 06:00 and 19:00) see Figure 4.5 to Figure 4.10.

Relative performance of interpolation methods: In all tests the piecewise constant interpolation is the least successful in recovering the original solution. The performance of the other interpolation methods is very similar during the day, with some variation in performance at dusk and dawn. During the day errors introduced by piecewise constant interpolation are typically an order of magnitude greater than the errors introduced by the other methods. For example, using 15 minute constraints, see Figure 4.6, piecewise constant day time errors are around 0.5% compared to 0.02% for piecewise linear and cubic spline interpolation.

Impact of constraint frequency: For both OH and HO_2 the impact of the constraint frequency is much greater than that of interpolation method. The peak errors at dusk for day 1 are an order of magnitude smaller for piecewise constant interpolation for the 1 minute compared with the 15 minute interval, see Figure 4.5 and Figure 4.6. The difference is two orders of magnitude for the equivalent comparison for the other interpolation methods. Using data on a 1 minute interval with a linear or cubic

interpolation method the error in recovering the original solution is in the order of 0.001 %.

OH vs. HO_2 Comparison: When comparing the OH and HO_2 errors from the same solution recovery run the relative errors for OH are greater. For example peak dusk error, approximately 19:00 on day 1, with piecewise constant interpolation on a 15 minute time interval is approximately 23% (see Figure 4.5), whilst for HO_2 the equivalent error is 18% (see Figure 4.8). Whilst the errors seen in this case study for OH and HO_2 are not significant (in the context of atmospheric chemistry modelling), these results establish the observable impact of constraint implementation even in a simplified/idealised system, such as the system used for these experiments.

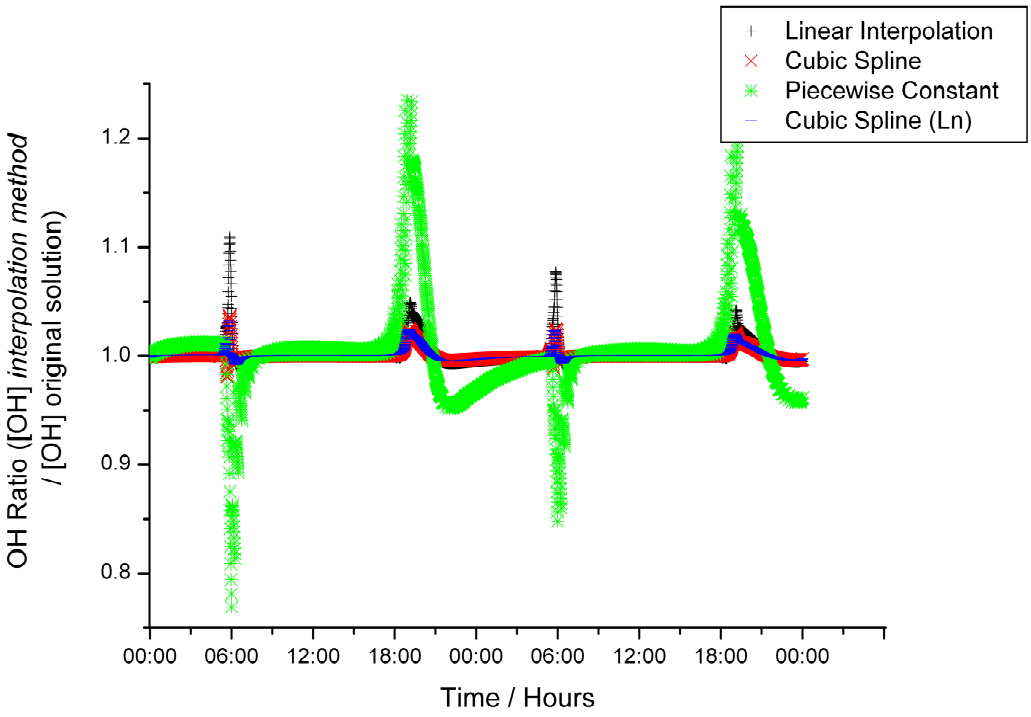


Figure 4.5: Unconstrained model, comparison of OH ratios for interpolation methods with constraint data at 15 minute frequency

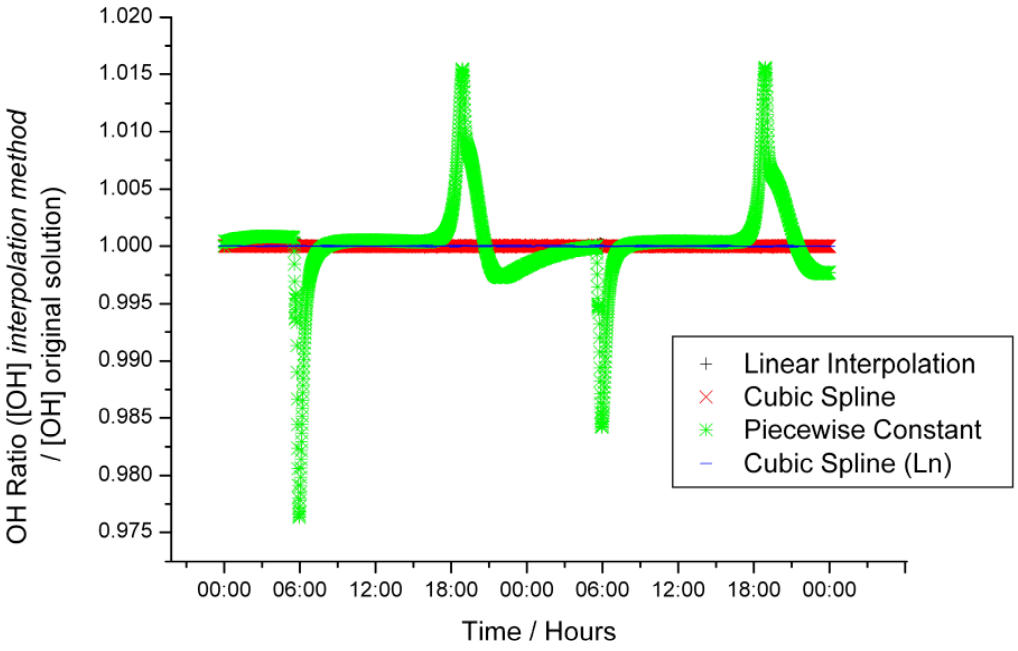


Figure 4.6: Unconstrained model, comparison of OH ratios for interpolation methods with constraint data at 1 minute frequency

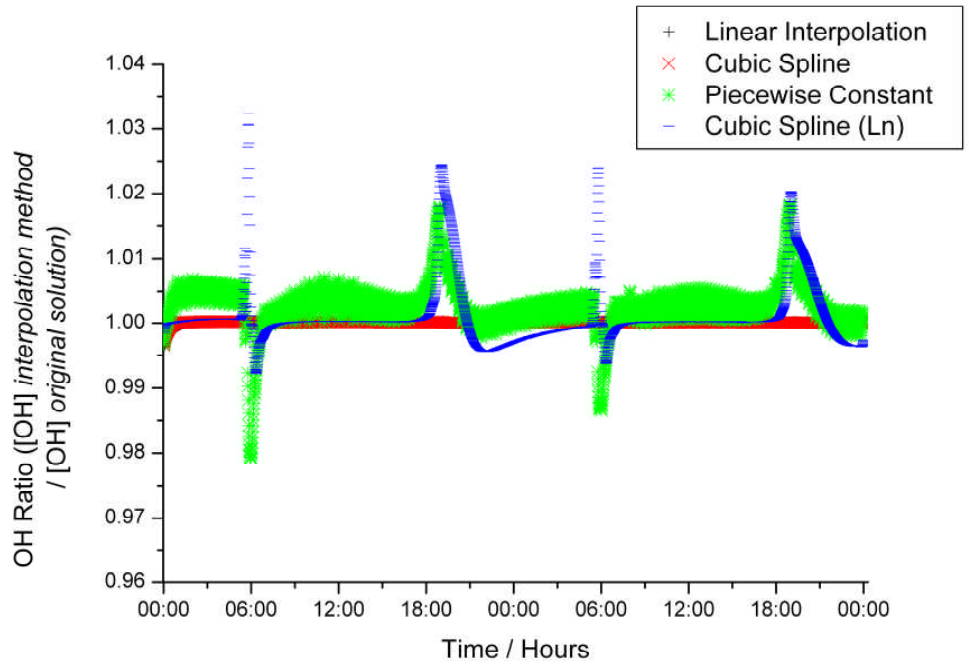


Figure 4.7: Unconstrained model, comparison of OH ratios for interpolation methods with constraint data at source specific frequency.

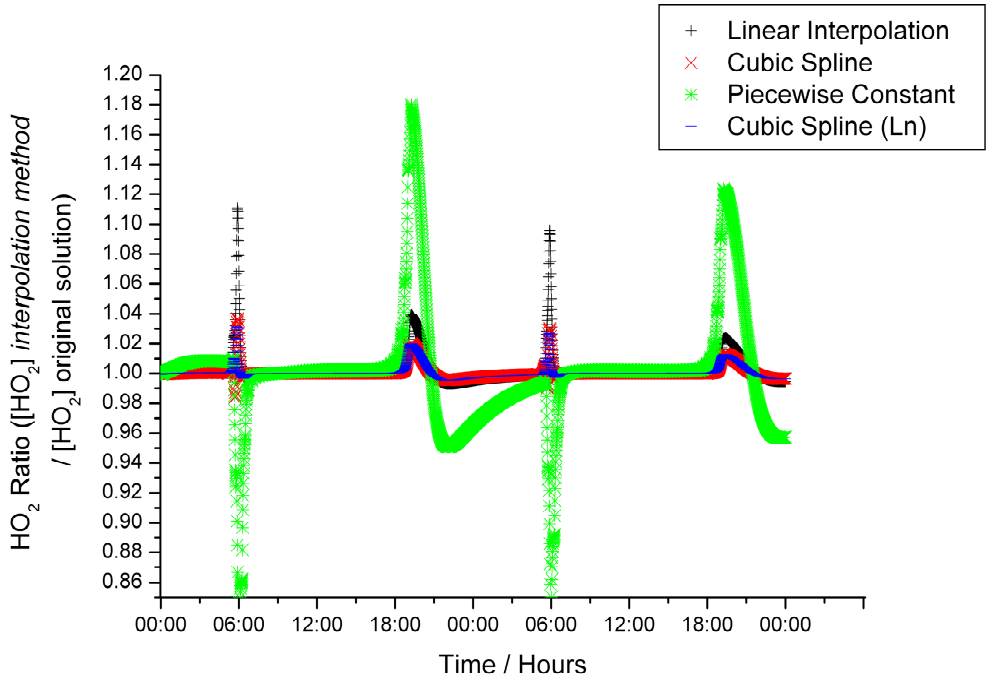


Figure 4.8: Unconstrained model, comparison of HO_2 ratios for interpolation methods with constraint data at 15 minute frequency

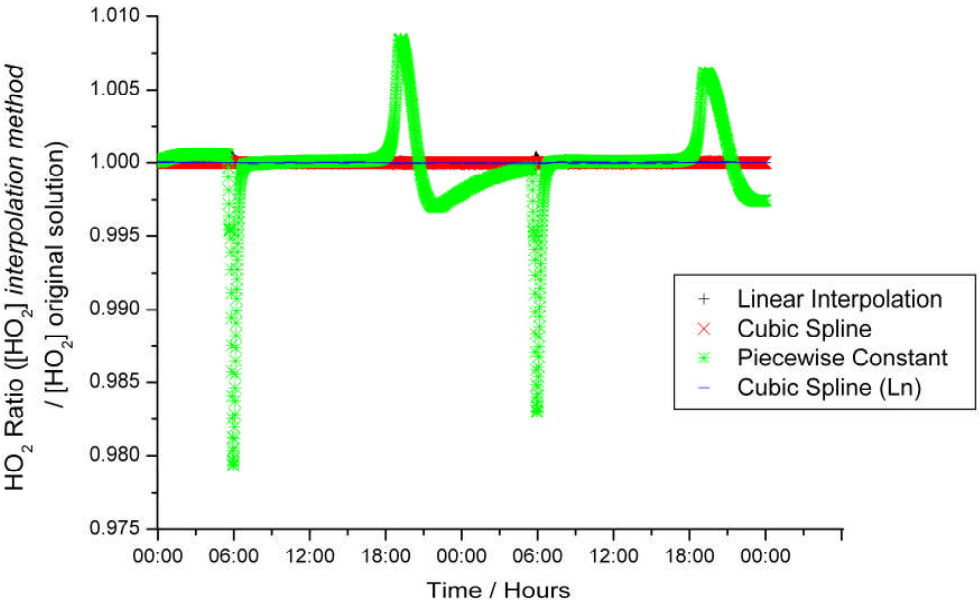


Figure 4.9: Unconstrained model, comparison of HO_2 ratios for interpolation methods with constraint data at 1 minute frequency

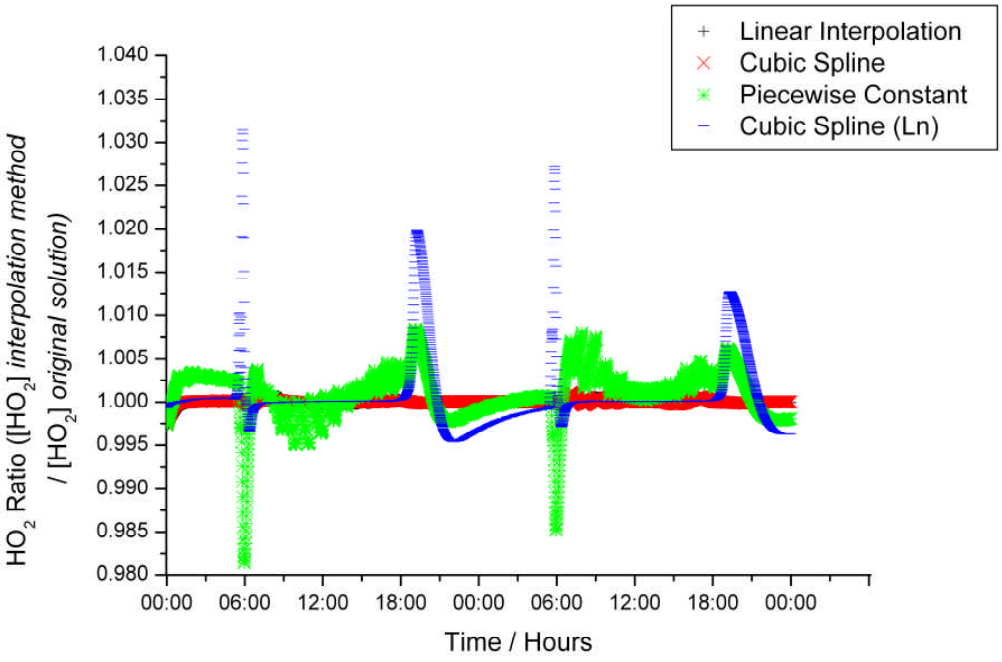


Figure 4.10: Unconstrained model, comparison of HO_2 ratios for interpolation methods with constraint data at source specific frequency

4.2.2.2 Condition-Constrained Model Test Results

The results of the condition-constrained model solution recovery tests are shown below (see Figure 4.11 to Figure 4.13). The results shown only consider the comparison of OH concentrations, as the HO_2 results are very similar, albeit with a smaller relative error magnitude (as in the unconstrained solution recovery tests discussed above). Three key results from the conditions-constrained model solution recovery results are presented below.

Magnitude of errors: Errors of up to 20% in OH concentration (see Figure 4.11) occur when using piecewise constant interpolation on 15 minute data. These peak errors are of similar magnitude to the peak errors in the unconstrained system. It is the difference in distribution of the errors that is telling; errors of greater than 5% for OH occurred throughout the day (where as in the unconstrained system errors greater than 5% occurred only at dawn and dusk). The total OH error (the sum of the magnitude of the relative errors in concentration at each minute) is approximately 2.5 times greater in the condition-constrained system compared to the unconstrained system (2.733×10^7 vs. 1.054×10^7

molecules cm^{-3}). The piecewise linear, and both cubic spline interpolants offer comparable performance at all three constraint frequencies, comfortably outperforming the piecewise constant interpolant.

Day time errors (photolysis): The discussion above hints at the distribution of errors across the day. In the case of the condition-constrained model errors occur for both OH and HO_2 throughout the daylight hours. This is a result of the variability introduced into the system by constraining photolysis rates on 1 minute time intervals, which in turn introduces variability over a 1 minute timescale in chemical constraints applied in the solution recovery tests.

Impact of constraint frequency: A notable feature of the solution recovery tests is that the impact of using data at a constraint frequency higher than 15 minutes is beneficial for all interpolation methods. The errors when using 1 minute or source specific constraints and an interpolation method other than piecewise constant are negligible (see Figure 4.12 and Figure 4.13). For piecewise constant interpolation, using 1 minute or source specific constraints reduces peak errors from approximately 20% with a 15 minute constraint frequency (see Figure 4.11) to approximately 5%.

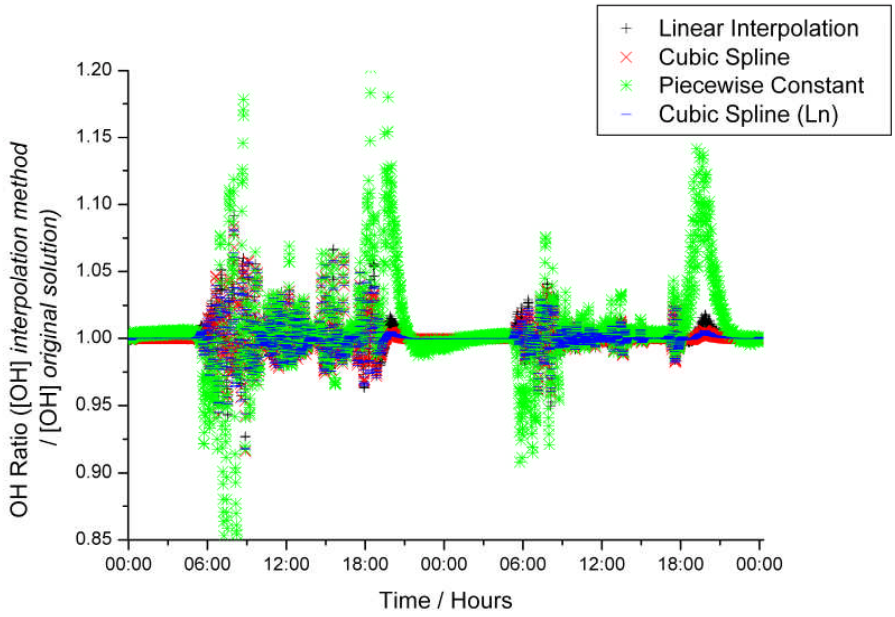


Figure 4.11: Condition-constrained model, comparison of OH ratios for interpolation methods with constraint data at 15 minute frequency

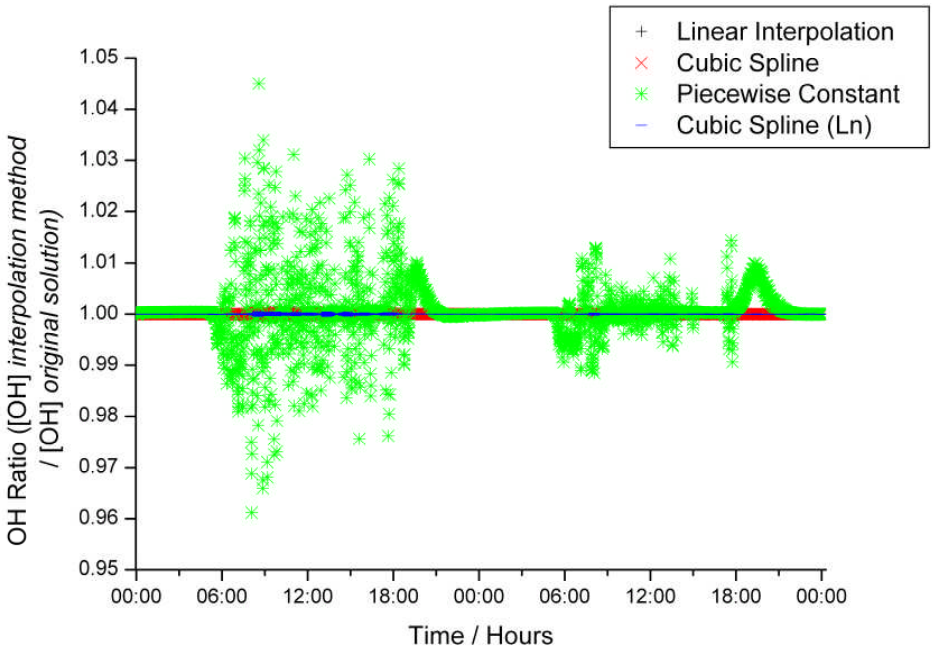


Figure 4.12: Condition-constrained model, comparison of OH ratios for interpolation methods with constraint data at 1 minute frequency

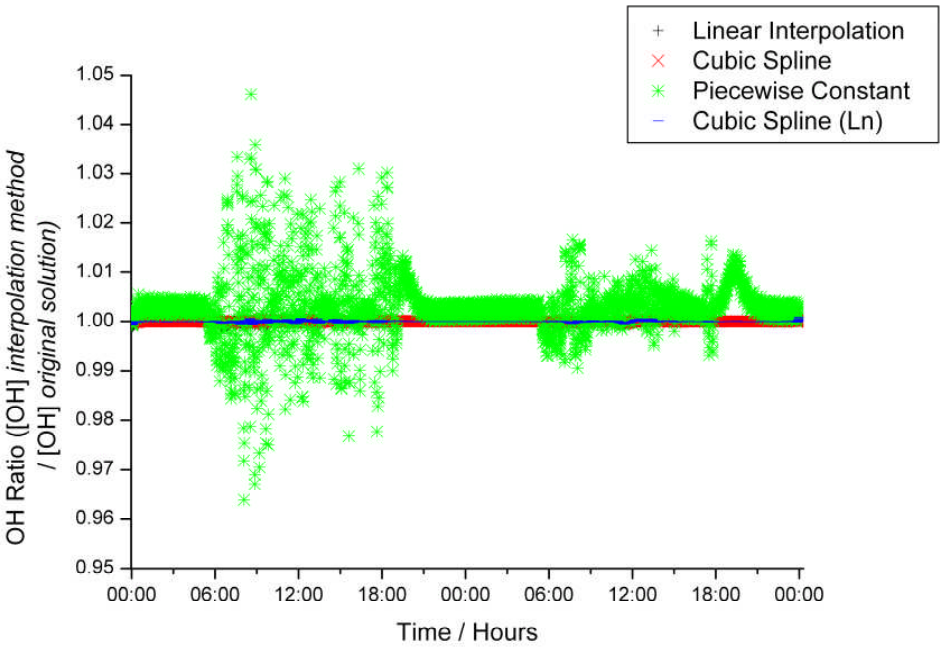


Figure 4.13: Condition-constrained model, comparison of OH ratios for interpolation methods with constraint data at source specific frequency

4.2.3 Model Efficiency

The preceding sub-sections have considered the impact of constraint implementation on model output. This section reviews the impact of constraint implementation on model efficiency (i.e. how long the model takes to run). Figure 4.14 shows the model runtimes for each of the unconstrained solution recovery tests, whilst Figure 4.15 shows the same results for the condition constrained solution recovery tests. These results demonstrate the independent impacts of the choice of interpolant and constraint frequency, each of these impacts are considered in the following subsections.

4.2.3.1 Impact of Constraint Frequency

For a given interpolant a clear relationship can be seen between the constraint frequency and the model runtime in both the unconstrained, see Figure 4.14, and condition-constrained tests, see Figure 4.15; the higher the constraint frequency the long the model runtime. For example, considering the linear interpolant in the unconstrained case:

- with 15 minute constraint frequency the model runtime is approx. 250 s;
- with source specific constraint frequency the model runtime is approx. 800 s;
- and, with 1 minute constraint frequency the model runtime is approx. 1500 s.

The relationship between constraint frequency and model runtime holds for all interpolants in both the constrained and unconstrained cases. This relationship exists because assimilating constraint data points ‘kicks’ (i.e. interrupts the normal operation of) the solver with two consequences: first, the solver time-step is reduced, so more steps are required to complete the model run; and secondly, the order of the solver is also reduced. So it can be seen that a trade off can be made between model runtime and the resolution of the model output.

4.2.3.2 Impact of Interpolant Choice

The impact of interpolant choice is less clear cut than the impact of constraint frequency. In both the unconstrained and condition constrained cases there is little to choose, in terms of model runtime, between cubic spline or linear interpolation. For example in the conditions-constrained case:

- with 15 minute constraint frequency the model runtime is approximately 250 s for both the cubic spline (ln) and linear interpolation methods;

- with source specific constraint frequency the model runtime is in the range 800 - 900 s for both the cubic spline (ln) and linear interpolation methods;
- and with 1 minute constraint frequency the model runtime is in the range 1700 – 1950 s for both the cubic spline (ln) and linear interpolation methods.

Using piecewise constant interpolation is substantially less efficient than cubic spline or linear interpolation, this can be seen in all solution recovery tests. It is particularly evident, when using constraint data at high frequency, for example in the unconstrained case with a constraint frequency of 1 min, using piecewise constant interpolation the model runtime is approximately 3200s, where as it is approximately 1500s using cubic spline or linear interpolation. The explanation for the relative inefficiency of piecewise constant interpolation relates to the discontinuities introduced at each constraint data point. These discontinuities cause a very rapid (instantaneous) change in the state of the underlying system of ODEs, so the ODE solver responds by reducing the solver step size and the order of the multi-step method to compensate and ensure solution accuracy.

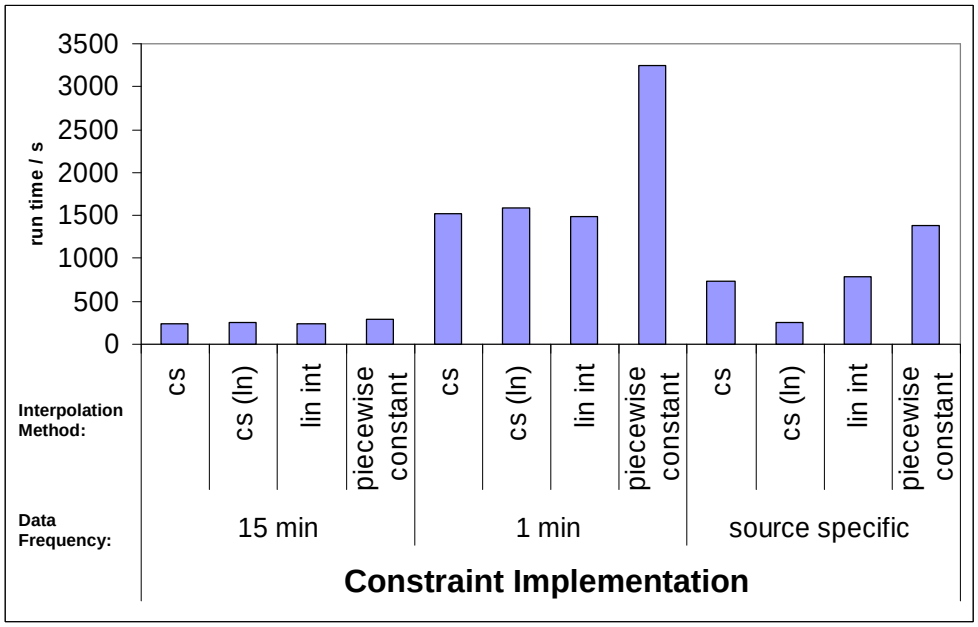


Figure 4.14: Model runtimes for unconstrained solution recovery tests.

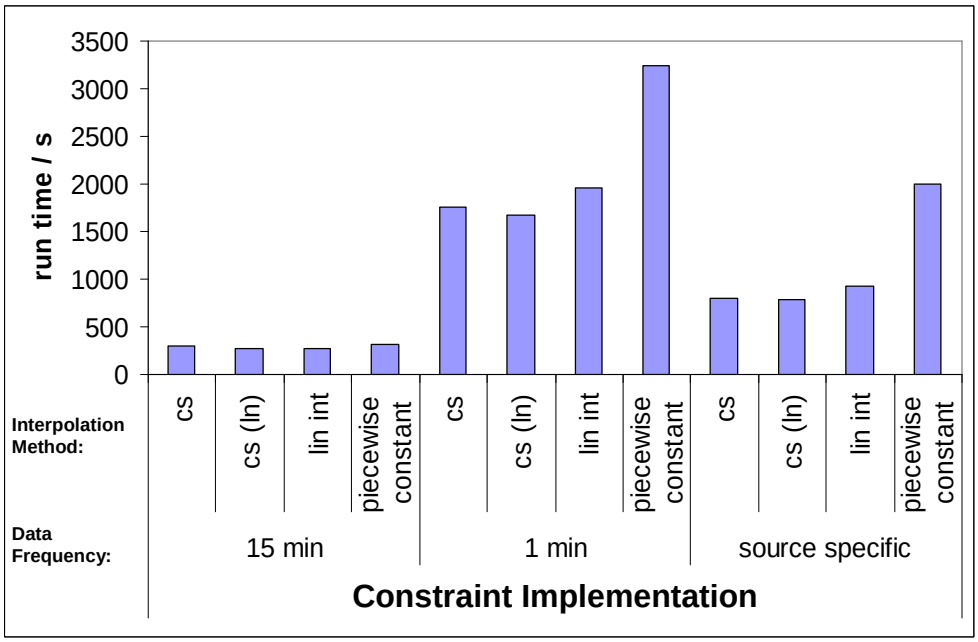


Figure 4.15: Model runtimes for condition constrained solution recovery tests.

4.2.4 Solution Recovery Test Conclusions

The preceding sub-section presented the results of a set of solution recovery tests designed to assess the impact of constraint implementation on model output (i.e. modelled values for species concentrations) and model efficiency (i.e. how long the model takes to run). The key conclusions of the results of these tests are presented below.

- Using a constraint frequency of higher than 15 minutes, brings substantially greater accuracy to the model output, but at the cost of reduced model efficiency;
- Of the interpolation methods tested, piecewise constant was by far the worst performing (in terms of accuracy of model output and model efficiency), with no significant difference between the performance of the other interpolation methods tested.

So as a result of the solution recovery tests I would recommend a constraint implementation consisting of: source specific constraint frequency (i.e. use as a high frequency as possible for each constraint, as determined by the experimental technique); and piecewise linear or piecewise cubic spline (ln) interpolation[8].

[8] Use of a cubic spline is not recommended, despite offering comparable performance to the linear and cubic spline (ln) interpolation, due to the potential to return negative values

4.3 SOAPEX-2 Model Tests

The solution recovery tests described in the previous section made use of simplified models in order to demonstrate, in quantitative terms, the impact of constraint implementation. This section progresses to consider the impact of constraint implementation on an atmospheric chemistry box model, as used to generate published scientific results. The model in question is the full SOAPEX-2 model, i.e. the simplified model used in the solution recovery tests with environmental conditions and species concentrations added. This section consists of three sub-sections: first, a description of the method used to test the impact of constraint implementation; secondly, a description of the results of the SOAPEX-2 tests; and finally, the conclusions drawn from the test results are outlined.

4.3.1 SOAPEX-2 model tests method

Source data for each of the constraints was retrieved from the BADC archive (http://badc.nerc.ac.uk/data/soapex/) where possible. It was not possible to retrieve data for the temperature and [H_2O] constraints from the BADC, so the required data was taken from Roberto Sommariva's[9] personal data archive, including: raw data for temperature and 15 minute averaged data for [H_2O]. In this section three constraint implementations are compared:

- 15 minute averaged data and piecewise constant interpolation; the constraint implementation used in the published simplified SOAPEX-2 model. Referred to as the baseline constraint implementation.
- Data at source specific intervals with piecewise linear interpolation. Referred to as the enhanced constraint implementation.
- Data at source specific intervals with cubic spline (ln) interpolation.

for constraints that are strictly positive (i.e. a negative concentration makes no sense in the physical system, but could be produced by a cubic spline).

[9] Roberto Sommariva developed the orginal SOAPEX-2 model, using the FACSIMILE modelling system.

4.3.2 SOAPEX-2 Model Tests Results

This sub-section presents the results of the SOAPEX-2 model tests: the section begins with a description of issues experienced with use of the cubic spline; results are then presented, comparing modelled OH and HO_2 concentrations, for the baseline and enhanced constraint implementations; and finally a comparison between modelled and measured OH concentrations is presented.

4.3.2.1 Unrealistic Cubic Spline Behaviour

Gaps in the constraining experimental datasets brought to light an issue not seen in the solution recovery tests. This issue is the unrealistic behaviour of the cubic spline interpolation method in cases where experimental data points are missing due to measurement errors, as seen in Figure 4.16. The peak value given by the cubic spline (ln) interpolation for [NO], in data gap (09:30-09:45), is 4 times greater than the surrounding data points. These interpolation issues, in data gaps, lead to unrealistic peaks in radical concentration. This can be seen in Figure 4.17 with outlying [OH] values, both maximum and minimum, around 9:30 and 12:00 on day 1 of the model. This erratic [OH] behaviour corresponds with gaps in NO and NO_2 data. There are number of potential methods for addressing this data gap issue, including:

- Filling the gaps with data to restrict the spline to more realistic behaviour. The fill data could be generated by linear interpolation for example.
- Using a more sophisticated interpolation algorithm, such as a member of the Shepard family of interpolants [4]. A Shepard interpolant is essentially a weight mean of basis functions.

Due to this unreliability in cubic spline interpolation the remainder of the section will focus on a comparison of the remaining two constraint implementations: first, 15 minute averaged constraint data, using piecewise constant interpolation (the baseline implementation); and secondly, source specific constraint data, using piecewise linear interpolation (the enhanced implementation).

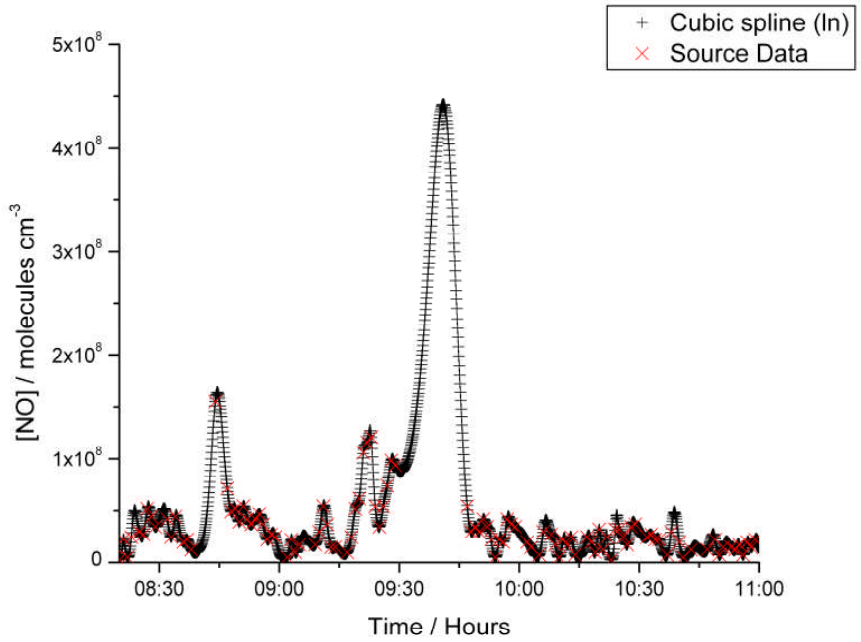


Figure 4.16: Unrealistic cubic spline behaviour over experimental data gaps for NO ((from SOAPEX-2, February 18th 1999).

4.3.2.2 OH Comparisons

Examining the [OH] profiles, see Figure 4.17, and ratio plot, see Figure 4.18, the baseline and enhanced constraint implementations can be seen to lead to significantly different results. The ratio plot shows that the differences between the enhanced case and the baseline case are commonly between 25% and 50%. The mean absolute percentage difference from the baseline case over the 2 day model run is 17%. From Figure 4.18, it can be seen that the errors tend to be greater at night than during daylight hours, so the mean absolute percentage difference maybe be unrepresentative and skewed by differences in the relatively small OH concentrations occurring at night.

Having reviewed the [OH] profile and comparison ratios over a two day period, it is worth examining the impact of each constraint methodology on a shorter timescale; a two hour period during the day on February 19th is shown in Figure 4.19. The striking feature of Figure 4.19 is that the [OH] for the enhanced constraint implementation almost instantaneously responding to the rapid change in $J(NO_2)$. This is in stark contrast to the

[OH] for the baseline constraint implementation which responds to $J(NO_2)$ only on the 15 minute input points. These observations of the relationship between $J(NO_2)$ and [OH], for each constraint implementation, hold throughout the daylight hours.

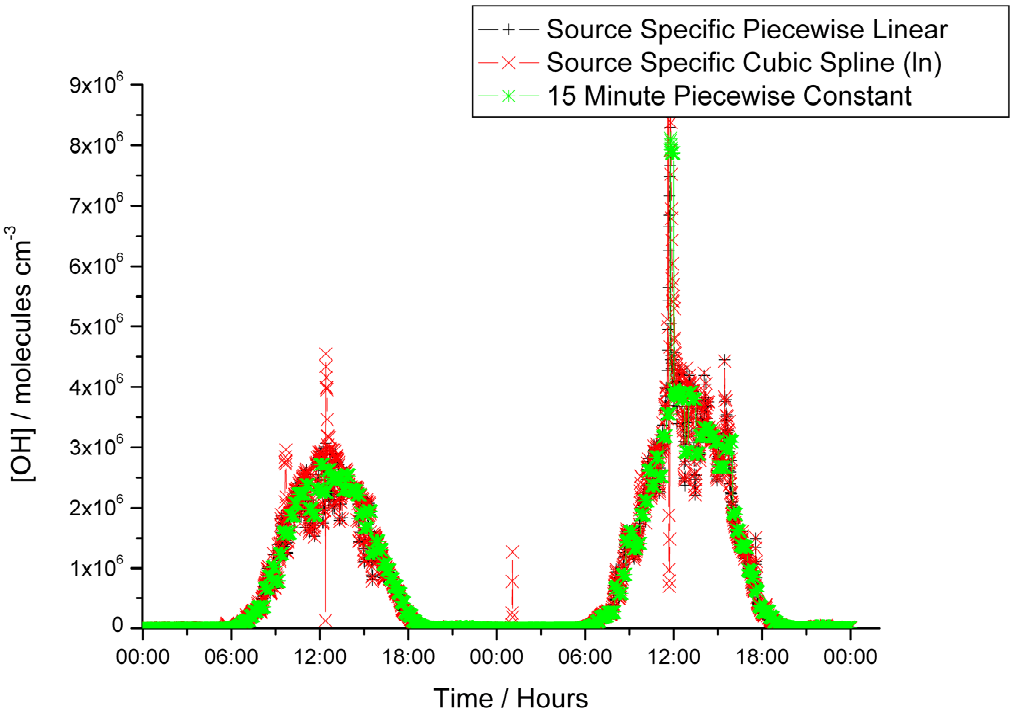


Figure 4.17: [OH] profile for February 18th-19th 1999 comparing constraint implementations

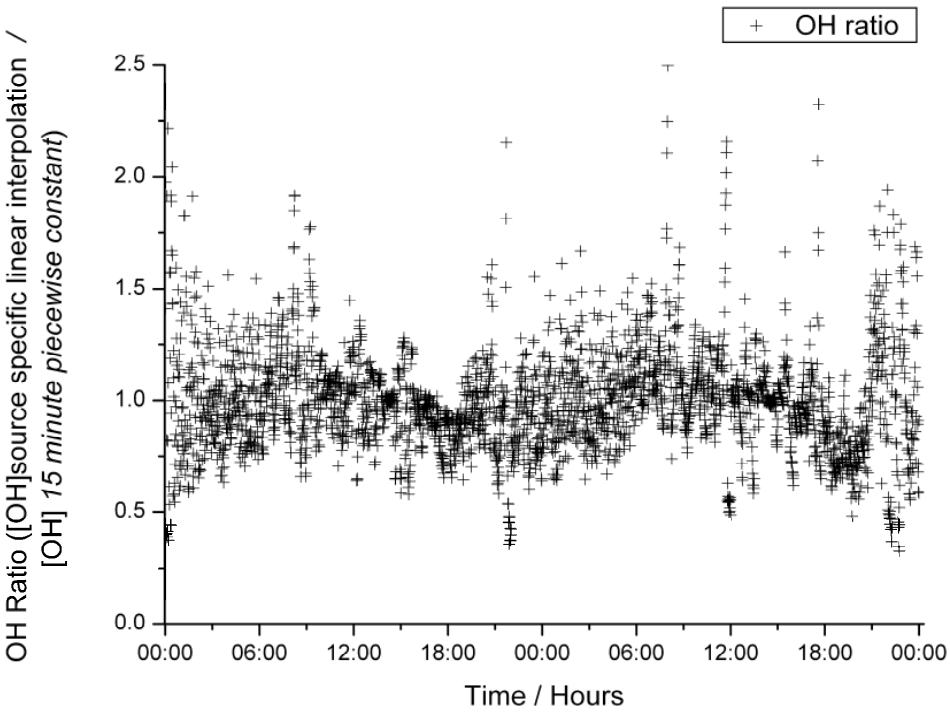


Figure 4.18: [OH] Ratios for February 18th-19th 1999

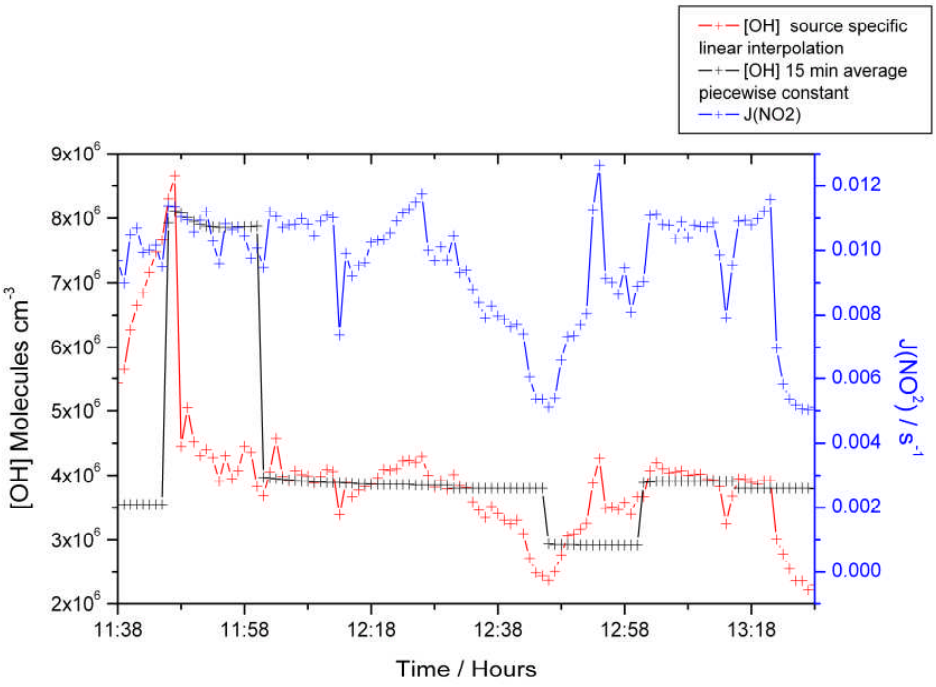


Figure 4.19: [OH] profile over midday February 19th 1999, comparing constraint implementations

4.3.2.3 Ratio Trend

An interesting feature in the OH ratio comparing the enhanced implementation to the baseline implementation is the generally decreasing ratio trend over the daylight hours, as seen in Figure 4.20. The spikes that occur in the ratio correlate very well with those in the $J(NO_2)$ profile. This is because under the baseline regime the spikes in photolysis rates (that are driving the OH spikes) are smoothed by 15 minute averaging, whereas the source specific linear interpolation implementation incorporates these variations on a minute time scale.

The explanation for this generally decreasing ratio can be identified by considering an ideal diurnal photolysis rate profile and the behaviour of the baseline constraint implementation relative to the profile. In the morning, see Figure 4.21, the baseline constraint implementation systematically underestimates the photolysis rate (J(..)) compared to the 1 minute linear interpolated profile. The impact of this can be seen in the ratio, in Figure 4.20, as it is ≥ 1 in the morning (ignoring spikes in the ratio correlated with spikes in photolysis rates). In the afternoon, see Figure 4.21, the baseline constraint implementation systematically over estimates the photolysis rate (J(..)) compared to the 1 minute linear interpolated profile. The impact of this can be seen in the ratio, as it is 1 or less in the afternoon (ignoring spikes in the ratio correlated with spikes in photolysis rates).

These systematic errors arise from a feature of the averaging and interpolation methods used in the original SOAPEX-2 model. If a notional data point at 10:00 for J(…) is considered: the value for J(…) is calculated as the average of data points (with a 1 minute frequency) between 09:53 and 10:07; during the model run this J(…) average will be applied (as a constant) over the 10:00-10:15 period. So the averaging period and application of the average in the model are misaligned, creating the systematic overestimation and underestimation shown in Figure 4.20. The systematic errors, and the correlation of the spikes in photolysis rate and comparison ratio, highlight the key limitations of the baseline constraint implementation. Firstly the baseline constraint implementation loses information about variation on sub 15 minute timescales. Secondly the piecewise constant interpolation makes unrealistic assumptions about the physical nature of the quantities being interpolated (i.e. a constant profile followed by a discontinuity).

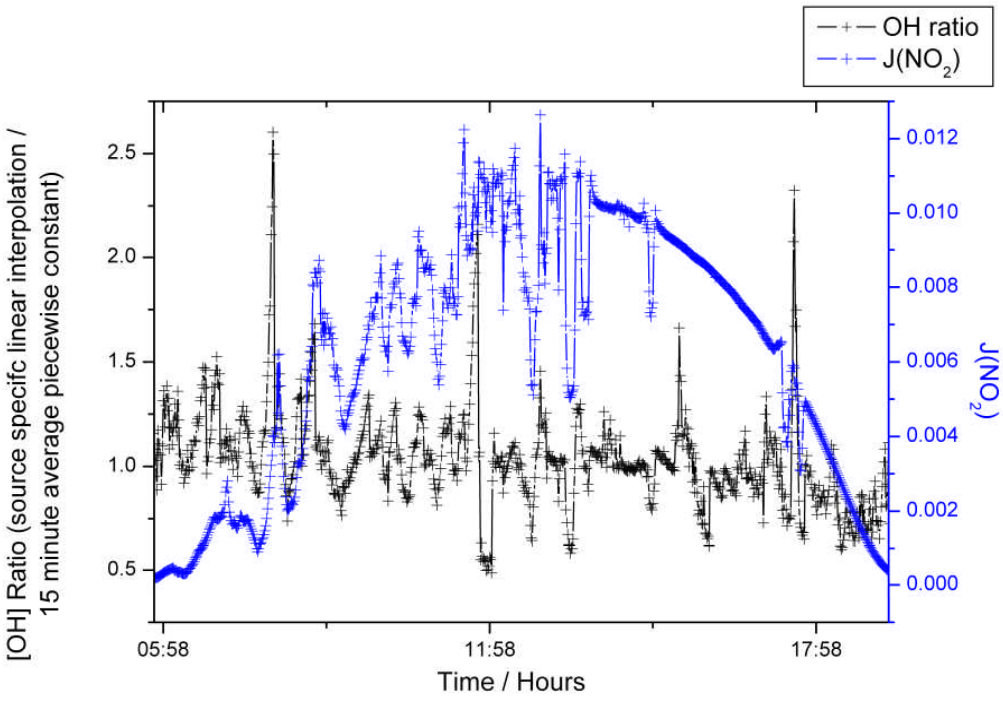


Figure 4.20: OH Ratio and $J(NO_2)$ profile for 19th February 1999

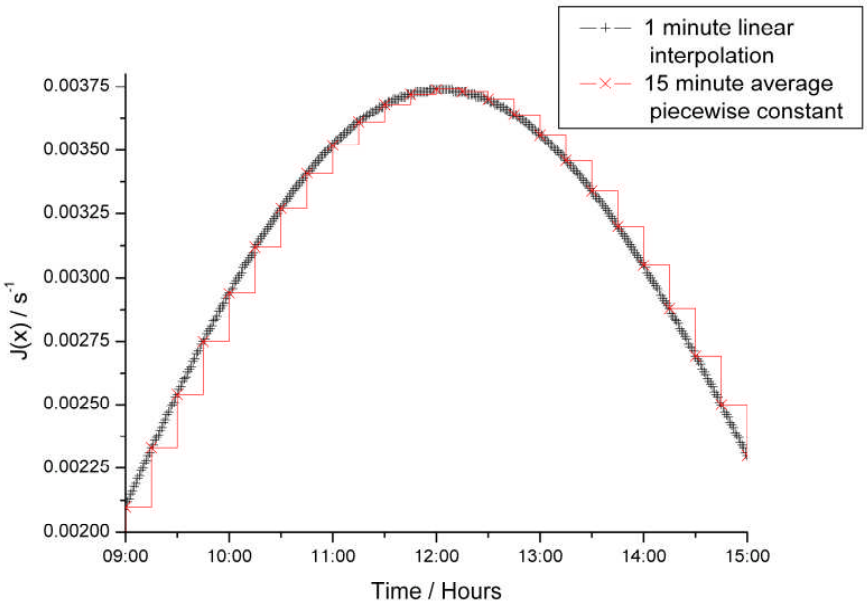


Figure 4.21: Systematic underestimation and overestimation of photolysis rates given an idealised diurnal photolysis profile

4.3.2.4 Comparison at 15 Minute Points

So far comparisons of constraint implementations have considered model output on a 1 minute interval. This subsection looks at comparisons on a 15 minute interval because model output is typically analysed on a 15 minute interval (as in SOAPEX [1] and NAMBLEX [2]). Figure 4.22 demonstrates that each constraint implementation produces significantly different OH output, on 15 minute points.

The mean absolute percentage difference for [OH], when comparing the enhanced implementation and the baseline implementation at 15 minute intervals, over the 2 day model run is 24%. This difference is greater than the average of 17% if the one minute output is compared. A possible explanation for this result is that in the baseline case output and constraint input take place on the same 15 minute intervals. This means that when output takes place it is 15 minutes (the longest possible period) since the constraint data has last been processed.

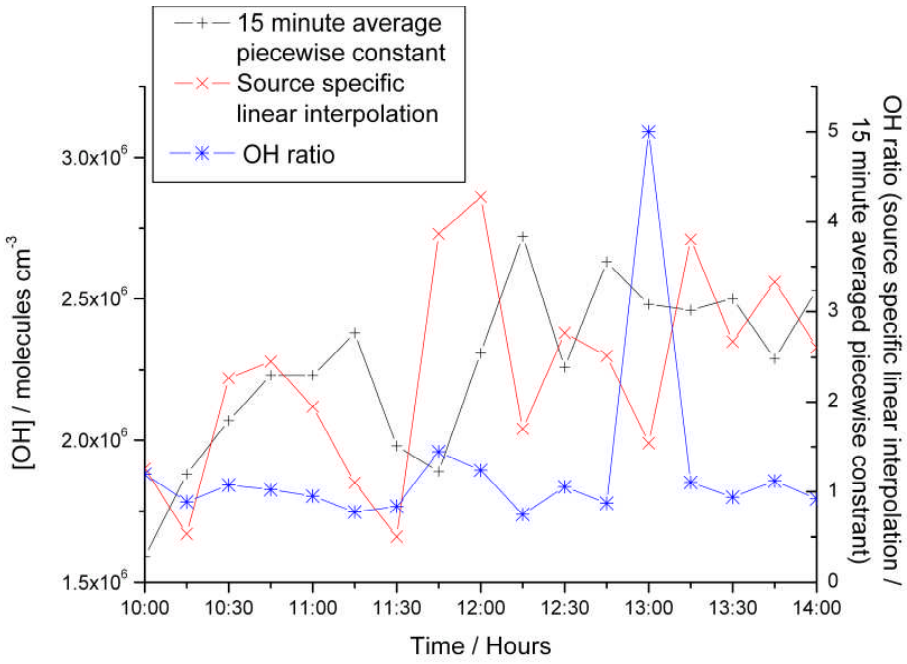


Figure 4.22: Comparison, on 15 minute model output, of OH profile over midday 18th February 1999.

4.3.2.5 HO_2 Comparisons

The impact of constraint implementation on HO_2 concentrations is very similar to that observed with OH (as in the solution recovery tests), with the exception of the relative observed errors being of a smaller magnitude. The mean absolute difference, for one minute output, between the enhanced and baseline constraint implementations for [HO_2] is 6% (compared to 17% for [OH]). So the key observations made in the preceding sections and reiterated below, hold for [HO_2] as well as [OH]:

- Declining comparison ratio trend over the day time (Figure 4.23).
- Correlation between radical concentration and photolysis rates on a one minute time frame with the enhanced constraint implementation (Figure 4.24).
- Larger mean error when comparing 15 minute point data (8% compared to 6%).

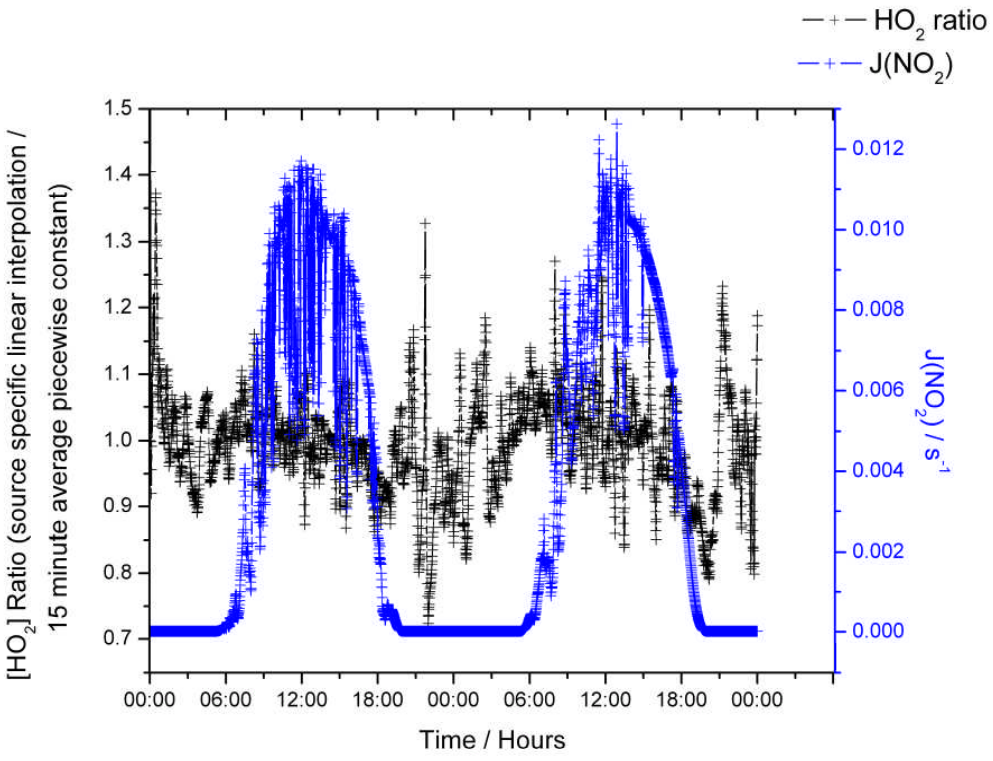


Figure 4.23: [HO_2] Ratio and J(NO_2) profile for 18th-19th February 1999.

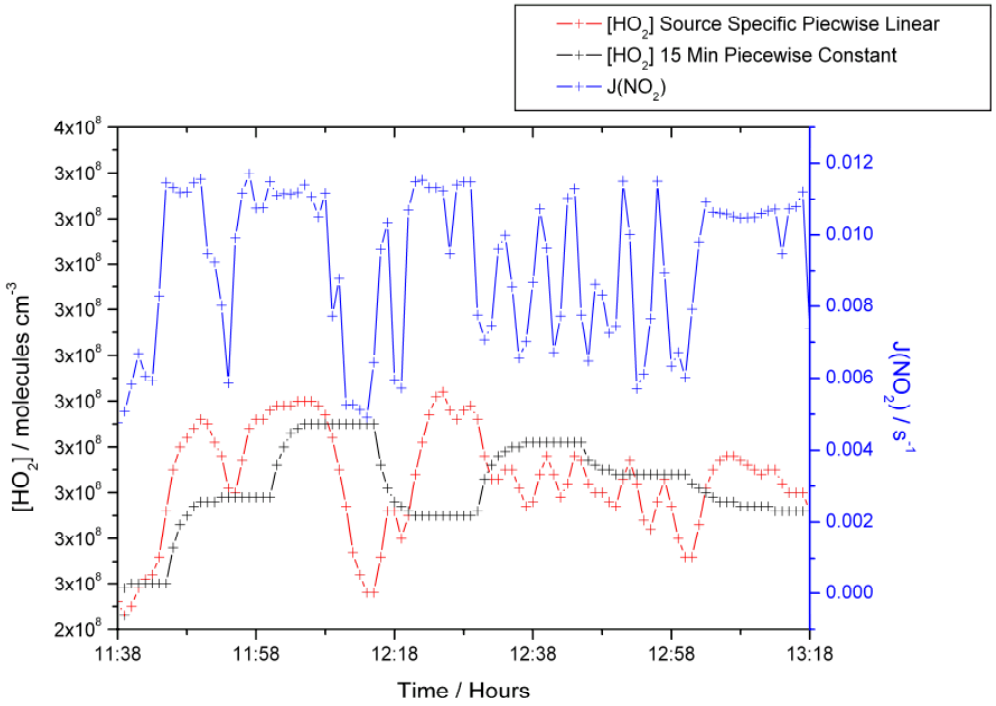


Figure 4.24: $[HO_2]$ and $J(NO_2)$ profiles over midday 18th February 1999, comparing baseline and enhanced constraint implementations

4.3.2.6 Comparisons with Measured FAGE data

This sub-section looks at the agreement between the modelling (presented above) and the [OH] measurements made using the FAGE technique [5] [6] during the SOAPEX-2 campaign. The baseline and enhanced constraint implementations are compared to the experimental measurements. The [OH] measurements clearly vary, as shown in Figure 4.25, on a sub 15 minute timescale. There is some correlation between this variation and the photolysis rate, so the source-specific constraint methodology performs much better than the baseline methodology in capturing the essence of these local variations. In contrast the baseline methodology produces [OH] profiles that look very much like the constraint data used as model input: [OH] remains constant for 15 minutes and then a discrete change in concentration takes place. Figure 4.26 show a model-measurement plot for [OH], both the baseline and enhanced constraint implementation overestimate the measured [OH] over the two day period modelled. Although the enhanced constraint implementation was expected to deliver better agreement with the experimental

measurements; there is little to choose between the two constraint implementations, in terms of their ability to deliver agreement with the experimental measurements.

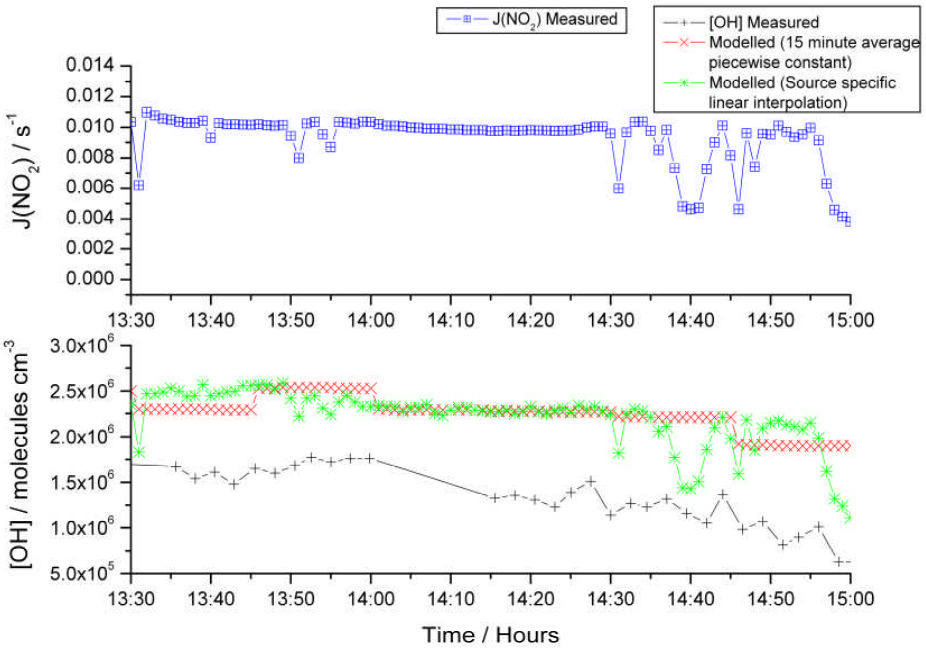


Figure 4.25: OH profile model-measurement comparison, over midday 18th February 1999

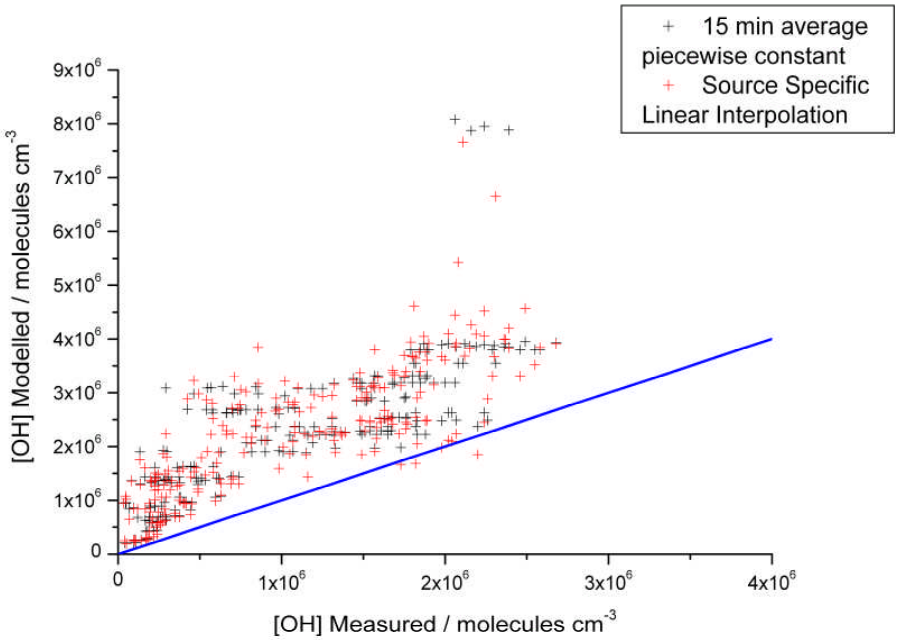


Figure 4.26: OH Model-Measurement plot for baseline and enhanced constraint implementations (18th-19th February 1999)

4.3.3 SOAPEX-2 Model Tests Conclusions

This section has presented tests designed to determine the impact of constraint implementation on modelled concentrations of radical species (OH and HO_2), for a fully constrained box model. The SOAPEX-2 model was selected for these tests and the following conclusions can be drawn.

- The constraint implementation has a significant impact on modelled radical concentrations;
- Using constraint data at a high frequency (where possible a 1 minute frequency) and using linear interpolation, allows radical behaviour to be modelled at high resolution (1 minute);
- Use of the cubic spline interpolation is not appropriate, where constraint data contains gaps;
- Using constraint data at a high frequency (where possible 1 minute frequency) and using linear interpolation does not provide significantly better agreement with the experimental measurements (in the case of the SOAPEX-2 model), than the baseline constraint implementation.

4.4 Conclusions and future work

The solution recovery tests demonstrate that generally the higher the frequency of the constraint data the smaller the error in recovering the original model solution. This suggests that in real modelling applications, as presented in section 3 of this chapter, greater accuracy for radical output can be achieved by using a constraint frequency as high as possible. It is important to note that this recommendation relies on the constraint data being presented in a form that required no further averaging to eliminate noise and errors. The use of high frequency constraint data also reduces the impact of the interpolant choice on solution quality.

The improved solution accuracy delivered by using high frequency constraint data comes at the cost of model efficiency. In the case study in this chapter (i.e. SOAPEX-2) the runtime issue is not particularly significant, as the model is relatively small and runtimes are typical in the order of minutes. This efficiency issue should be investigated in future work as it is likely to be more significant in larger models, with thousands of species rather than tens of species.

Piecewise constant interpolation, as used in the original SOAPEX-2 modelling, leads to the least accurate modelled radical concentrations, as established in the solution recovery tests. It is also the least efficient choice of interpolation method, due to the discontinuities it introduces in constraint values. So given both these drawbacks, and the ready availability of a robust alternative in the form of linear interpolation, where possible it is best to avoid the use of piecewise constant interpolation.

Linear interpolation can be seen to perform comparably to cubic spline interpolation in the solution recovery tests, in terms of result accuracy and model efficiency. The benefit of using linear interpolation is seen in the SOAPEX-2 model tests, Section 4.3, as it deals with gaps in experimental data in a simpler more effective manner than the cubic spline. Given that the benefits of cubic spline interpolation are unclear and there is an overhead required to ensure cubic splines behave realistically over gaps in experimental data; linear interpolation is recommended as the most suitable interpolation method, pending future work on more sophisticated interpolation methods.

Although not presented with the chapter, the solution recovery tests have been repeated for the TORCH2 model (introduced in Chapter 3). The TORCH 2 model is significantly more complex than the SOAPEX-2 model, in terms of the chemistry taking place and constraints applied to the model. The results of the solution recovery tests on the TORCH2 model are entirely consistent with those presented in this chapter.

Future Work

Having summarised the conclusions of this chapter, the opportunities for future work are outlined below. First and most obviously, the findings of this research need to be validated, by repeating the tests on a variety of models (of differing complexity). This research could also be extended by investigating the use of smooth non-oscillating interpolants, such as a member of the Shepard family of interpolants [4]. Further analysis of model output could be conducted to understand radical behaviour on short timescales, including Rates of Production and Loss Analysis [7] for OH and HO_2. In the research presented in this chapter the experimental data used to constrain box models is treated as a series of exact data points. In reality each data point, in a constraint set, has an uncertainty associated with it. Incorporating this uncertainty into the constraint implementation is

likely to prove a challenging task, leading to potentially interesting and important scientific findings.

Chapter summary

This chapter has investigated the modelling of radical concentrations on short time scales, and the associated impact of constraint implementation. The next chapter of this thesis draws upon my experience developing the OSBM and the models required to perform the constraint implementation research (in this chapter); to address an important e-Science issue.

References

1. Sommariva, R., et al., *OH and HO2 chemistry in clean marine air during SOAPEX-2.* Atmos. Chem. Phys., 2004. **4**(3): p. 839-856.
2. Sommariva, R., et al., *OH and HO2 chemistry during NAMBLEX: roles of oxygenates, halogen oxides and heterogeneous uptake.* Atmos. Chem. Phys., 2006. **6**(4): p. 1135-1153.
3. Ahlberg, J.H., E.N. Nilson, and J.L. Walsh, *The Theory of Splines and Their Applications*. 1967, New York: Academic Press.
4. Shepard, D., *A two-dimensional interpolation function for irregularly-spaced data.* Proceeding of the 1968 23rd ACM national conference, 1968: p. 517- 524.
5. Creasey, D.J., et al., *Implementation and initial deployment of a field instrument for measurement of OH and HO2 in the troposphere by laser-induced fluorescence.* Faraday Transactions, 1997. **93**: p. 2907-2913
6. Creasey, D.J., et al., *Measurements of OH and HO2 concentrations in the Southern Ocean marine boundary layer.* Journal of Geophysical Research, 2003. **108**: p. 4475.
7. Sommariva, R. and e. al., *OH and HO2 chemistry in clean marine air during SOAPEX-2.* Atmospheric Chemistry and Physics Discussions, 2004. **4**(1): p. 839-856.

Chapter 5 Data and Provenance within the MCM Community

This chapter provides the link between the computational modelling based research, presented in Chapters 3 and 4, and the e-Science based (i.e. provenance based) research, presented in the remainder of this thesis. An outline of the background to and the motivation for investigating the role of provenance within the atmospheric chemistry community is presented. The rationale for developing a provenance capture tool (an Electronic Laboratory Notebook) for *in silico* experiments, emerges from issues identified with current provenance related working practices. The structure of this chapter is as follows: first, an overview of the provenance research in this thesis is presented; secondly, current working practices across the atmospheric chemistry community are described, and drawbacks of current practices are identified; thirdly, envisioned working practices that make use of e-Science technologies are described; and finally, related work is discussed.

5.1 Provenance and the Atmospheric Chemistry Community

This introductory section outlines the motivation for, and nature of, the provenance research presented in the following chapters of this thesis. The questions "what is provenance?" and "why capture provenance for scientific data?" are addressed. The approach to the provenance-based research presented in this thesis is then outlined.

5.1.1 What is Provenance?

The research, in the following chapters, is concerned with capturing and representing data provenance within the atmospheric chemistry community, specifically provenance for data produced by *in-silico* experiments that make use of the Master Chemical Mechanism (MCM). Multiple definitions for data provenance are used across the e-Science community. For example Simmhan et al. define data provenance "as information that helps determine the derivation history of a data product, starting from its original sources" [1]. Greenwood and co-workers [2] consider data provenance to be composed of two components: first, the "derivation path" which records the scientific workflow used to generate the output data, this includes the processes executed and input data used; and

secondly, annotations, i.e. additional information attached to the processes and data in the scientific workflow by the scientist. The definition of Greenwood et al. [2] has been adopted in this work, as it recognises the two-dimensional nature of the provenance; the scientific process executed and the scientists' reasoning associated with the process executed.

5.1.2 Why Capture Provenance for Scientific Data?

In this sub-section two key motivators for capturing provenance for scientific data are presented. These motivators apply across the atmospheric chemistry community, and to the wider scientific community (as a whole).

Supporting the principles of the scientific method

The capture of provenance for scientific data has been recognised as an important issue at the core of the scientific process, since the birth of the scientific method. A key tenet of the scientific method is to ensure that experimental results are reproducible; capturing provenance describing the experimental process can ensure the reproducibility of scientific results, or at least the repeatability of the experiment. As scientific fields of enquiry have developed, the complexity of experimental processes has grown, making it difficult to provide the detailed provenance required to ensure the reproducibility of results using standard methods of publications (i.e. journal articles). So, a two layer scientific model has developed, where: scientific results are published alongside a summary of the experimental process; and, the detailed provenance required to reproduced published results is managed and stored by the publishing research group, to be made accessible upon request.

Capturing provenance is a good working practice

Capturing provenance is a good working practice to adhere to during the generation of scientific results and insight, bringing benefits in terms of the quality and efficiency of the science conducted. For example, by capturing and archiving provenance:

- A researcher can reduce the amount time they need to spend reacquainting themselves with an experiment, when returning to a piece of research after a break.

- The continuity of research (within a given research group) can be ensured, as a new group member can continue the research of former group member, with a sound understanding of previous work.
- A researcher can answer questions about the process and reasoning underpinning published results with confidence; and can point to a provenance record as a form of evidence.
- A researcher can gain a better understanding of the results of other researchers (than if a provenance record was not available).
- Data can be shared and reused, in unanticipated ways, to generate scientific insight.
- Ownership of data can be transferred, e.g. from an individual data producer to an organisation (such as a research group or an institutional archive), as interpretation of the data is not reliant on the tacit knowledge of a researcher.

5.1.3 Drawbacks of Current Approach to Provenance Capture

Across the atmospheric chemistry community provenance is captured using a variety of manual and automated techniques and archived using a variety of storage media. Many approaches to provenance capture consist of the local application of manual, *ad hoc* techniques, with the laboratory notebook often playing a central role in the provenance capture process; the drawbacks of this type of approach include those listed below.

- The reliance on manual processes for provenance capture, can lead to incomplete provenance records, of varying quality;
- Provenance with an *ad hoc* structure can be difficult to interpret, particularly for anyone other than the researcher who originally captured the provenance;
- Provenance is often stored in a single location, e.g. the laboratory notebook, as an analogue artefact, making provenance difficult to share;
- Provenance archives are often fragile, with a single point of failure, e.g. the laboratory notebook, so can be easily corrupted or lost (e.g. a researcher leaves a research group and takes a lab book with him or her);

The drawbacks, identified above, are a result of a complex set of interacting factors including: researchers having insufficient time to capture and archive high quality provenance records; the lack of appropriate tools to support provenance capture and archiving; the academic reward and recognition systems not recognising provenance

capture as an important activity; and in some cases a lack of interest and inclination on the part of the researcher responsible for capturing provenance.

5.1.4 A User-Orientated Approach to Provenance

In the course of addressing the drawbacks of the current approach to provenance, set out above, I sought to develop a user-orientated approach to provenance. This user-orientated approach is a defining characteristic of the research presented in the remainder of this thesis, and is described in this sub-section.

Definition

I have defined a user-orientated approach to provenance as:

> *Developing a provenance capture tool seeking to retain the beneficial features of current provenance-related working practices, whilst addressing the drawbacks of current provenance-related working practices using methodologies from e-Science/provenance research. Retaining the beneficial features of current provenance-related working practices is prioritised over the application of e-Science/provenance methodologies and theory.*

Motivation

This user-orientated approach can be seen to require an in-depth understanding of the atmospheric chemistry domain, rather than an in-depth understanding of the logic and formalisations that underpin provenance research, and prioritise meeting the requirements of system users over making advances in e-Science and provenance research. I believe that investing time in understanding the problem domain is more likely to develop understanding that leads to usable, rapidly adopted provenance systems that deliver benefits across scientific communities. And as a result of this user-orientation, I suggest advances in the theory of provenance and e-Science will emerge. The user-orientated approach naturally aligns with the overall approach that guided the research presented in this thesis. The ethnographic, interdisciplinary research approach, described in Chapter 1, led me to become embedded within the atmospheric chemistry community, experiencing the processes being studied first hand.

5.1.5 Links to Computational Modelling Research

My first-hand experiences of the model development process (as outlined in Chapters 3 and 4), played an important role in identifying provenance capture for *in silico* experiments as a research topic. My experiences then played an important role in informing the design (see Chapter 8), implementation (see Chapter 9), and evaluation (see Chapter 10) of the ELN; this role is examined below, in three parts.

Understanding provenance issues

During the development of the Open Source Box Model (OSBM), benchmarking the OSBM against an alternative modelling system proved to be a challenge. A number of models were used as benchmarks, including the existing SOAPEX-2 [3] and TORCH-2 [4] models (described in Chapter 3). The challenge presented was fully understanding the FACSIMILE version of the model, in order to allow the model to be re-implemented using the OSBM.

When seeking to understand the implementation of the FACSIMILE models the resources available included: thesis chapters briefly describing the implementation of the model and extensively describing the use of the model; a limited number of comments in the model source code (although many of these were misleading or inaccurate); publications containing the results used from the models in question. The resources unavailable included: the laboratory notebook containing provenance relating to the development of the model; the insight of the researcher who originally developed the model (who did not have sufficient time to provide a commentary on the process of developing the original models). Piecing together information from the available resources was a time consuming process; and led to an incomplete, fragmented understanding of the model, which was supplemented by a line-by-line examination of the model source code. This experience highlighted issues with the current practices for the capture and archival of provenance, and provided a motivation on personal level to pursue research into the role of provenance within the atmospheric chemistry community.

Understanding modelling processes: Developing a modelling system for MCM users (i.e. the OSBM) required an in-depth understanding of the *in silico* experimental processes employed by the MCM users, in the course of their research. This understanding informed and facilitated the development of a tool to capture provenance for these *in silico*

experiments (i.e. the ELN); allowing the focus of research to look beyond mapping current model development processes.

Understanding atmospheric chemistry terminology: Being embedded within the atmospheric chemistry community and working closely with members of the community to develop the OSBM, enabled me to gain an understanding of the terminology used within the community. This understanding was critical in enabling a user-orientated approach (described above) to be adopted.

5.2 Current Practice

This section reviews current working practices across the atmospheric chemistry community; focussing on data generation and provenance capture. The review concludes by identifying an opportunity to apply e-Science technologies to deliver benefits across the community. Three sub-sections are presented addressing the following topics: first, the multiple-scale nature of atmospheric chemistry research is described; secondly, the concept of a community evaluation activity is introduced; and finally, an example of a community evaluation activity within the atmospheric chemistry community is outlined.

5.2.1 Atmospheric Chemistry as a Multi-Scale Science

In this sub-section a conceptual model of the atmospheric community activity is explored, shown in Figure 5.1. This conceptual model has been developed based upon my observations and experiences within the community. The conceptual model describes atmospheric community activity at three scales: the elementary reaction scale; the complex reaction scale; and the application scale; the nature of research activity at each scale is described below.

The Elementary Reaction Scale: At this scale atmospheric chemists are interested in understanding the characteristics of elementary chemical reactions of atmospheric relevance. The reaction characteristics of interest include the: products, rate coefficients, product yields.

The Complex Reaction Scale: At this scale atmospheric chemists are interested in the coupled chemical processes that take place within the atmosphere. These coupled chemical processes can be represented as chemical mechanisms, essentially a list of the reactions taking place. When developing a mechanism a scientist will draw on the reaction characteristics determined by research at the reaction scale to ensure the mechanism is grounded in established scientific knowledge. Atmospheric chemical mechanisms can be used within computational models, enabling the chemistry predicted by the mechanism/model to be compared to *in-situ* experimental measurements.

The Application Scale: Chemical mechanisms can be taken as is, or reduced in size (by a variety of mechanism reduction methods), and used in scientific applications requiring a description of the chemistry taking place in the atmosphere. Examples of such applications include: computational models of the local/regional/global distribution of chemical species, these models include atmospheric dynamics and other earth system components.

The research in this thesis is rooted at the mechanism scale, with some interaction with the reaction scale, the application scale is beyond the scope of my research so not considered further.

5.2.2 Community Evaluation Activities

At both the reaction and mechanism scales Community Evaluation Activities (CEAs) take place. CEAs can be defined as involving a group of expert researchers evaluating research outputs across the community seeking to develop a shared, gold-standard, understanding of the current state of knowledge, that can then be used across the community. The benefits of the current state of knowledge being available for use across the community are significant: for example, individuals do not need to survey the vast literature to find the best value for a given parameter, in the atmospheric chemistry case the best reaction characteristics or mechanism. CEAs are typically comprised of the following activities.

Aggregation. Aggregating the data, information and knowledge produced across the community. This aggregation process is enabled by the expert nature of the CEA panel, i.e. the experts know where to look for the appropriate publications.

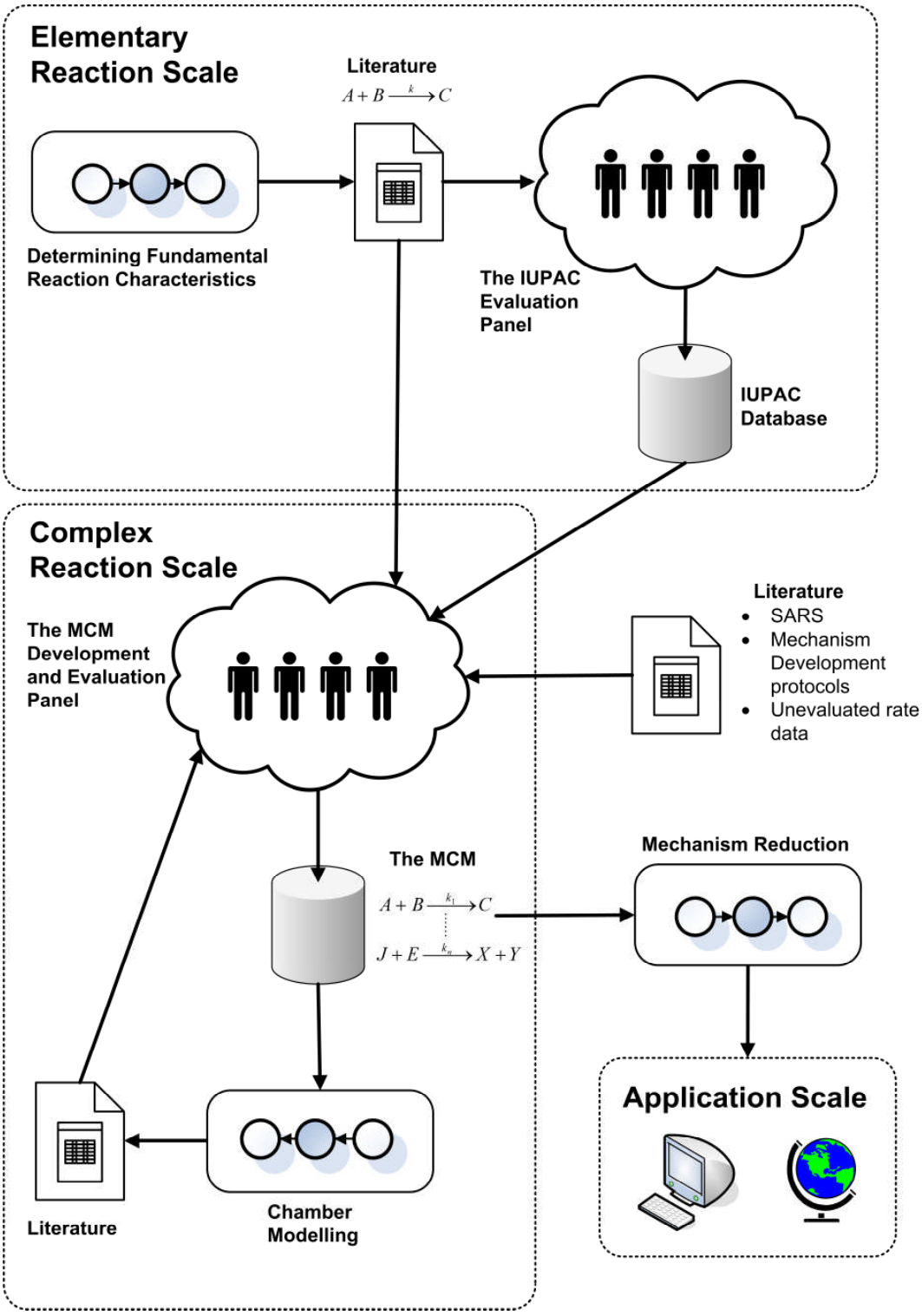


Figure 5.1: A conceptual model of an atmospheric chemistry community activity at the elementary reaction, complex reaction and application scale.

Apply quality rating. Each piece of research and resulting knowledge can have a quality rating applied to it. Determining quality ratings is an inherently subjective activity, relying on the ability of the expert panel to understand and judge quality. The definition of quality is not clear-cut, but is related to the quality of the experimental process (how carefully was it conducted? how accurate is the experimental technique? etc.) and how useful the knowledge will be to a potential user (e.g. a precise rate co-efficient with narrow error bounds is useful to a mechanism developer).

Develop knowledge landscape. Once the knowledge has been aggregated and had a quality rating applied to it a knowledge landscape can be composed. Where multiple knowledge items are competing for the same space within the landscape (e.g. different measurements of the same rate co-efficient) the superior item can be selected or items of similar quality can be combined. Within the knowledge landscape areas of weakness can be highlighted, stimulating future research.

Update knowledge base. Once the knowledge landscape has been developed, its essence can be captured and placed within a knowledge base, which can then be made available to the community for subsequent use.

Given this general description of CEAs this section continues to consider the CEAs within the atmospheric chemistry community research activity.

IUPAC evaluation

The IUPAC CEA [5] provides recommended reaction characteristics (products, rate co-efficients, etc.) for reactions of atmospheric importance. Currently the IUPAC CEA only reviews research from within the public domain, i.e. as appearing in journal publications, and has access to only the journal articles themselves.

MCM development and evaluation panel

Developing the MCM consists of two complementary CEAs.

- Mechanism development; which is a periodic, structured update of the MCM based on the latest available data (generated at the reaction scale).

- Gathering and evaluating feedback from *in silico* experiments, at the mechanism scale (i.e. comparisons between chamber experiments and computational models that make use of the MCM), prompting amendments and corrections to the MCM.

5.2.3 Gathering Feedback from *In Silico* Experiments

Having described the nature of community evaluation activities, in general, this sub-section describes current working practices related to a CEA within the MCM community; gathering and evaluating feedback from *in silico* experiments. This description is presented in three parts: first, the *in silico* experiments of interest are described; secondly, the CEA process is described; and thirdly, the drawbacks of the current CEA process are outlined.

5.2.3.1 *In Silico* Experiments of Interest

For the purpose of this discussion an *in silico* experiment can be considered to consist of the following elements:

- Development of a computational model (for a given experiment), using the MCM to describe the chemical processes taking place within the model;
- Comparing model output with experimental results;
- Developing insight about the performance of the MCM and causes of the observed experimental results.

These three steps are likely to be iterated over multiple times, in order to reach a set of conclusions. *In silico* experiments relating to chamber experiments provide the most information relevant to the ongoing development of the MCM because chamber experiments often have the expressed goal of developing understanding of chemical mechanism.

5.2.3.2 Gathering and Evaluating Feedback from *In Silico* Experiments

Figure 5.2 shows current working practices for gathering and evaluating feedback from *in silico* experiments. The core elements of these working practices and Figure 5.2 are described below.

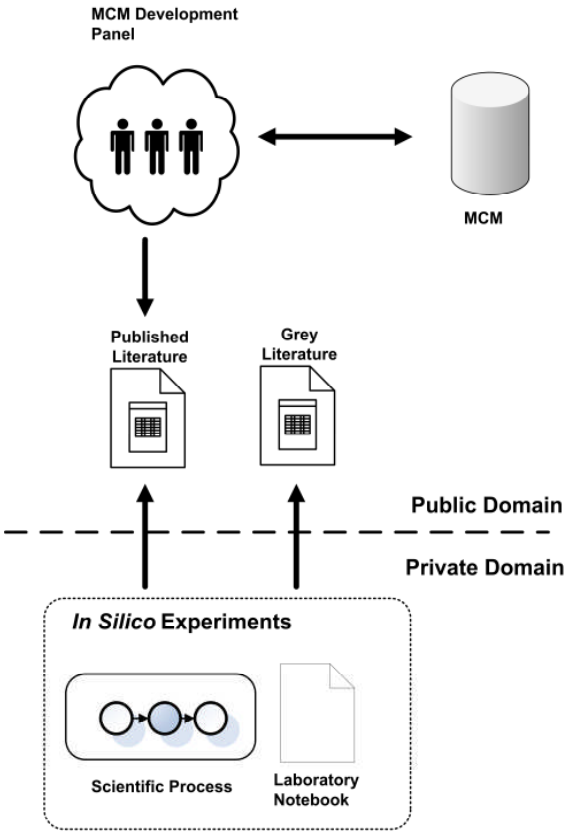


Figure 5.2: Current working practice for gathering and evaluating feedback from *in silico* experiments.

The MCM development panel: MCM developers evaluate the findings of the *in silico* experiments to determine if the findings have any implications for the MCM. Minor revisions to the MCM can then be made where appropriate; revisions can be either made directly to the publicly available version of the MCM or queued awaiting the release of a new version of the MCM. Relevant findings of an *in silico* experiment could include:

- Deficiencies in the current MCM, for example the MCM over-predicts ozone concentrations, when considering the degradation of a certain VOC in a chamber experiment;
- Evaluations of the current MCM, in terms of its ability to predict concentrations of radicals, ozone, etc. For example the MCM is performing well in predicting the radical concentrations associated with of the degradation of a certain VOC in a chamber experiment;
- Suggested improvements to the MCM, having identified deficiencies with the performance of the MCM a researcher may go one step further, test a set of potential mechanism amendments to provide a recommended set of changes to the publicly available version of the MCM.

The literature: The MCM development panel review the published literature to identify the applications of the MCM, to be evaluated. Published literature is the main source of data and information (i.e. provenance) available to the development panel. Supplementary data and information can be acquired by personal communication with the researchers associated who conducted the MCM application. Grey literature, i.e. *in silico* experiments in the public domain but not yet peer reviewed, is not typically used due to difficulties accurately evaluating it.

In silico experiments: Across the MCM user community researchers use the MCM in *in silico* experiments, using custom modelling tools (including the OSBM) and *ad hoc* provenance capture methods (such as the laboratory notebook), as described in Section 5.1.3. So detailed provenance is generally not available to the MCM development panel.

Gathering and evaluating feedback from *in silico* experiments (in practice)

This approach, outlined above, to gathering feedback from *in silico* experiments was applied when developing the MCM from version 3.0 to version 3.1 [6], and led to improvements in the MCM's ability to describe the chemistry of several organic compounds (including toluene). Figure 5.3 shows the resulting improvement in model-measurement agreement, between v3.0 and v3.1, for the Toluene mechanism, including: for toluene, O_3, NO_2 and NO.

5.2.3.3 Drawbacks of Current Practice

A number of drawbacks are experienced in the current practices for gathering and evaluating feedback from *in silico* experiments. These drawbacks prevent the full value of *in silico* experiments being realised and are listed below.

- Unpublished data and provenance from *in silico* experiments are typically retained, within the private domain of the researcher; so valuable insight may not come to the attention of the MCM development panel.
- The MCM development panel operates with a limited amount of information, as they do not have access to data and provenance underpinning published results.
- Searching for MCM applications in the published literature is a time consuming process, and relies on the tacit knowledge of the MCM developers.

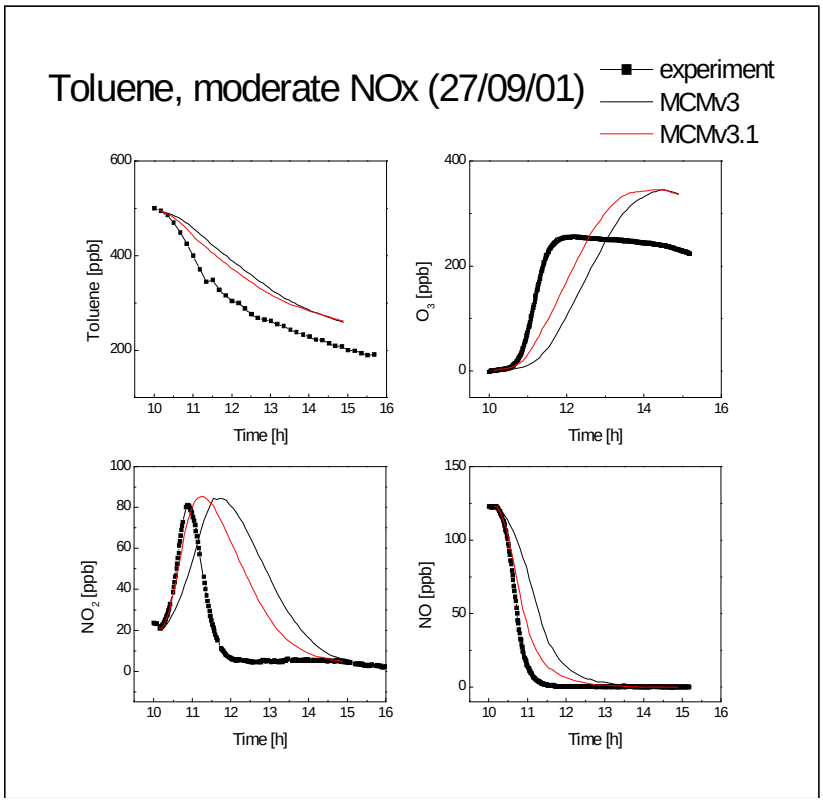


Figure 5.3: Model-measurement comparison for toluene, O_3, NO_2 and NO in toluene photosmog experiment of 27th September 2001 [6].

Given these drawbacks, an opportunity exists to systematically develop the processes for gathering and evaluating feedback from *in silico* experiments; to exploit advanced computational thinking (i.e. to conduct some e-Science research [7] [10]).

5.3 Envisioned Working Practices for Gathering and Evaluating Feedback from In Silico Experiments

In the preceding section gathering and evaluating feedback from *in silico* experiments was identified as an area where e-Science technologies and methodologies could be applied to deliver benefits. In this section I present a vision for a transformed *in silico* experiment CEA, leveraging semantic web technologies to enable improvements in the quality of the CEA output. In this context, quality of the CEA output is related to two key factors: first, the MCM developer's ability to aggregate all the relevant available *in silico* experiments; secondly, the MCM developers ability to accurately judge the quality of the *in silico* experiment they are evaluating. The later factor is tightly coupled to the availability of provenance and data from beyond the standard sources (i.e. journal articles and other publications). Envision working practices for the CEA are presented in Figure 5.4, and discussed below, preceded by a discussion of the role of the semantic web in realising these envisioned working practices.

5.3.1 The Role of the Semantic Web

The current World Wide Web (www) can be thought of as a very large set of documents, the content of the web (i.e. data) is primarily readable by humans. In general, computers cannot understand and process the data that make up the www. So a problem arises; the development of applications that can process the vast quantities of online data, to deliver benefits to www users, is inhibited. The Semantic Web [8] seeks to address this problem, by making the data on the www readable by computers; this requires common formats and standards to enable the integration of data from diverse online sources. Semantic metadata (SMD) describes data, available on the www, and conforms to Semantic Web standard to enable data integration.

[10] "The term e-Science denotes the systematic development of research methods that exploit advanced computational thinking"

5.3.2 Gathering and Evaluating Feedback from *In Silico* Experiments on the Semantic Web

Having outlined the nature of the semantic web, above, gathering and evaluating feedback from *in silico* experiments can be reframed as a Semantic Web problem, as shown in Figure 5.4, where:

- For each *in silico* experiment, conducted by a researcher, the associated data and provenance is made available on the Semantic Web;
- And the MCM developers use to Semantic Web tools to aggregate the available data (i.e. *in silico* experiments) in order to inform the development of the MCM.

Envisioned working practices for gathering and evaluating feedback are described below in four parts.

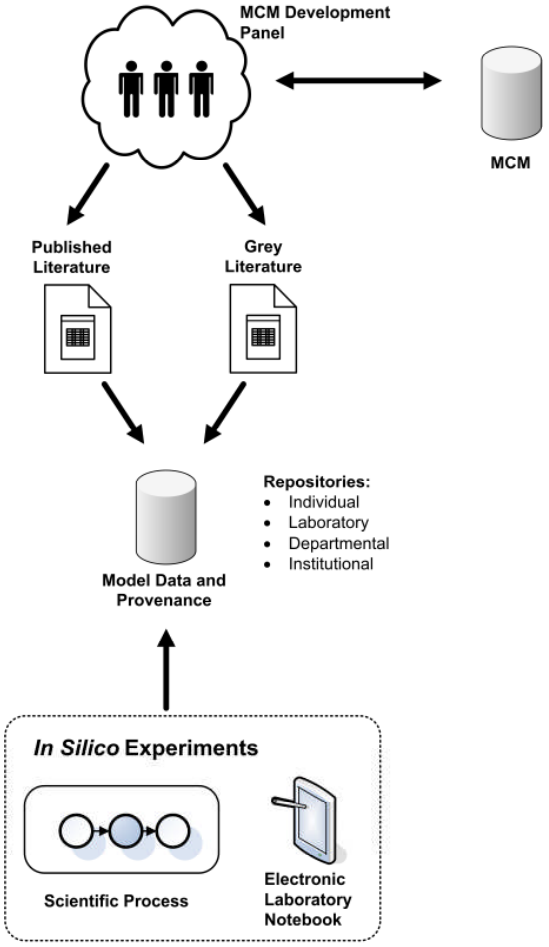


Figure 5.4: Envisaged *in silico* experiment feedback Community Evaluation Activity (CEA).

The MCM development panel

When gathering feedback from across the community, the MCM development panel have access to two information sources: published literature (as in current practice) and grey literature (i.e. *in silico* experiments not yet with the published domain). Data aggregation tools are provided to enable MCM developers to quickly aggregate *in silico* experiments of relevance, from across the MCM community.

The literature

Published and grey literature is described with semantic metadata, which supports the data aggregation tools used by the MCM development panel. Links to data and provenance that underpins the literature are provided with two key benefits: first, reviewers and readers of an article can drill down into additional experimental detail provided by the data and provenance; secondly, MCM developers can also benefit, having found research of interest within the literature they can obtain further detail and explicitly link changes in the MCM back to the source data.

Repositories

Data and provenance from *in silico* experiments will be stored in a variety of repositories including: personal, laboratory, departmental and institutional repositories. Data and provenance will be shared as freely and openly sharing as possible, across the MCM user community. The security and access rights for these repositories are acknowledged as a significant issue, but are not addressed further within this thesis. Common standards for both provenance and data will be required to enable interoperability across distributed repositories.

***In silico* experiments**

In silico experiments are typically performed iteratively, e.g. a researcher amends the chemical mechanism, runs the model, analyses the model output and then reflects on their findings before editing the mechanism again. Data and provenance from *in silico* experiments will be captured using an Electronic Laboratory Notebook (ELN), and stored in an appropriate archive. Provenance captured by the ELN will be represented using Semantic Web standards, and underpins the envisioned *in silico* experiment CEA.

5.3.2.1 Benefits of Envisioned Practice

The benefits of adopting the envisioned working practices for the *in silico* experiments CEA include the following.

- MCM developers can access both published and unpublished research, to support the development of the MCM;
- MCM developers can operate with access to extensive information resources for each *in silico* experiment, including associated laboratory notebook entries;
- Semantic Web technologies are applied to automatically aggregate the data and provenance required by the MCM developers.

In this section an envisioned CEA, for gathering and evaluating feedback from MCM applications, has been described. In order to realise this envisioned CEA provenance and data will be captured for the *in-silico* experiments taking place across the MCM community, using ELNs. It is the design, implementation and evaluation of an ELN, that operates within the envisioned MCM CEA, that is explored during the remainder of this thesis (Chapters 6 – 10). The final section, of this chapter, places the envisioned CEA in the context of related work.

5.4 Related Work

The final section of this chapter positions the envisioned *in silico* experiment CEA, in the context of related e-Science research. Three topics are considered: first, the first class object within an e-Science environment; secondly, electronic laboratory notebooks; and thirdly, system orientated approaches to provenance for *in silico* experiments.

5.4.1 First Class Objects

The nature of the first class object within an e-Science architecture can often be used as a distinguishing characteristic; it essentially defines the type of object (e.g. data, workflow etc.) the system is developed to support (i.e. what is it that system users are primarily interested in sharing or exchanging). In order to provide the MCM developers with a holistic view of the scientific activity taking place within a community a science-based perspective must be adopted, treating scientific experiments themselves as first class objects within the community model. Treating scientific experiments as first class objects

with an e-Science architecture, is a novel approach; alternative approaches are discussed below.

Workflows as first class objects

The MyExperiment project aims to develop a Virtual Research Environment (VRE) where scientists, across all scientific domains, can share their scientific workflows [9]. In this case the term workflow means composition of services, composed in order to perform a scientific process. MyExperiment can in many ways be seen to be similar to social networking sites such as facebook (www.facebook.co.uk), with a variety of functionality to support the specific requirements of scientific researchers. Adopting workflows as the first class objects within a VRE raises a significant question: do workflows map well to scientific processes executed by scientists? The workflow concept clearly maps well to the scientific process within the bioinformatics domain, as much of the research relating to scientific workflows has taken place in this domain [10] [11] [12]. Workflows are typically linear and do not map well to iterative model development processes (as used across the MCM user community), so adopting workflows as the first class object for the *in silico* experiment CEA would not be appropriate.

Data as first class objects

The CombeChem project (http://www.combechem.org/) adopted an alternative approach, by treating data as a first class object. CombeChem is a diverse project addressing e-Science issues across a variety of chemistry research areas including: provenance capture for *in vitro* organic chemistry experiments [13, 14]; data and provenance publication for the crystallography research community [15]. Treating data as a first class object would be inappropriate with the *in silico* experiment CEA because the MCM developers are interested in not just: the data produced by researchers using the MCM; but also the knowledge and scientific insight generated by researchers using the MCM.

Digital resources as first class objects:

The CARMEN project seeks to develop a Virtual Laboratory Environment (VLE), a specific type of VRE, for the Neurophysiology scientific community [16] with the aim of promoting the dissemination, reuse and sharing of digital resources. Digital resources, including data and computational models, are considered the first class objects within the

VLE. Treating digital resources as a first class object is not appropriate for the MCM CEA because; the MCM developers are primarily interested in the data, information and knowledge generated by the scientific processes of researchers using the MCM not digital resources (i.e. models, code etc.).

The laboratory notebook as a first class object

Open Notebook Science (http://drexel-coas-elearning.blogspot.com/2006/09/open-notebook-science.html) (ONS), is an emerging paradigm with the field of e-Science that represents a revolution in the scientific process. In ONS a researcher conducts their scientific process, publishing their data and laboratory notebook entries as the experimental process takes place. So the research process takes place in the open allowing any interested parties to make use of data or contribute to the research process as collaboration opportunities emerge. ONS takes the Laboratory Notebook, the traditional means of capturing data and provenance in a laboratory setting, as a metaphor and creates an openly accessible online Laboratory Notebook. So as the Laboratory Notebook is being shared the first class object within the system can seen to Laboratory Notebook itself. Treating the laboratory notebook as a first class object is similar to the approach adopted for the *in silico* experiment CEA. A laboratory notebook can be considered composed of a collection of experiments, with each experiment composed of an aggregation of experimental data and provenance, it is these experiments that an MCM developer is interested in accessing and reviewing.

5.4.2 Electronic Laboratory Notebooks

ELN have been developed and deployed in a number of commercial and academic settings. The majority of ELN are targeted at scientists performing *in vitro* experiments, with Schraefel et al. [13] identify four categories of ELN, each outlined below.

- **Replication:** Replication systems, such as SCRIP-SAFE [17], allow users to digitise the contents of their paper laboratory notebook (using scanning systems); and are often used to protect intellectual property claims. Such systems suffer the drawback of having limited search functionality.
- **Supplement:** Supplementary systems, such as Labscape Lab Assistant [18] capture provenance whilst anticipating the continued use of the paper laboratory

notebook. Such systems suffer the drawback of producing a fragmented provenance record.

- **Replacement:** Replacements systems**,** such as the Comb*e*Chem ELN [13, 14] described above, seek to replace the paper laboratory notebook, by recreating the experience of using the paper lab-book with a tablet PC. This approach retains the flexibility of the paper laboratory notebook whilst adhering to a structured provenance representation format.
- **Augmentation:** Augmentation systems, such as a-Book [19], seek to capture paper lab-book entries as they are made. For example a tablet PC is placed underneath the paper lab-book, and captures the hand written note of the user.

The replacement strategy has been adopted in the ELN development considered in the following chapters, seeking to take advantage of the fact that the scientific process takes place at the computer. So, the ELN can integrate directly with the scientific process to capture provenance and data in a well-structured manner.

5.4.3 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 systems [20] [21] paradigm. For the purpose of comparison between the workflow approach to provenance and the user-orientated approach (described in this thesis) I take the Taverna system [10] as an exemplar from the workflow system paradigm.

Taverna [10], in common with many other workflow systems [22] [23], seeks to automatically capture provenance for *in silico* experiments, minimising user involvement. Automatic provenance capture is well-suited to capturing 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, why they have re-run a workflow with a modification to the input parameters, etc.). 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). Secondly, the provenance captured by Taverna, as with other workflow

systems [23] [24], is represented using domain independent semantics. So the scientific process (captured as a workflow/series of workflows), is represented independently of the particular scientific domain. Whilst the use of domain independent semantics can be seen as an important factor in producing a domain independent workflow system, that is deployable across different scientific domains, domain independent semantics remove the opportunity to leverage the informational content of the scientific terminology of a given scientific domain. Given the key characteristics identified above, minimising user involvement in provenance capture and using domain independent semantics to represent provenance, the typical workflow approach to provenance can be viewed as system orientated.

The First Provenance Challenge [21] 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 to be within scope for provenance capture.

The MyGrid research group addresses the provenance challenge using Taverna plus a knowledge template [25], 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 the user-orientated approach where process provenance is captured, using semantics from a specific scientific domain, automatically.

The VisTrails response to the first provenance challenge [26] 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 to both the workflow and the input data, using scientific terminology.

A number of provenance systems, including Karma [27], applied to the first provenance challenge, considered annotations beyond the scope of the provenance research discipline. I 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 data was produced in a certain way that the user orientated approach seeks to capture.

The importance of the scientist's contribution to provenance has been recognised in the work of the PolicyGrid project, where they seek to capture the scientist's intent as well as their method [28]. PolicyGrid have taken the Kepler workflow environment [24], 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 (with goals, reasoning, etc.), and structure these annotations with the use of ontology.

Chapter Summary

In this chapter I have made the case for capturing provenance for *in silico* experiments within the atmospheric chemistry community. The capture of provenance for *in silico* experiments is motivated by the vision of radically re-engineering community evaluation activities; required to deliver a step-change in the quality of the output of these community evaluation activities. An Electronic Laboratory Notebook for *in silico* experiments is required to support the envisioned CEA. It is the design, implementation and evaluation this ELN that is the topic for the remainder of this thesis. The next chapter describes the methodology used to guide the ELN development.

References

1. Simmhan, Y., B. Plale, and D. Gannon, *A survey of data provenance in e-science.* ACM SIGMOD Record, 2005. **34**(3): p. 31-36.
2. Greenwood, M., et al., *Provenance of e-Science Experiments - experience from Bioinformatics*, in *UK e-Science All Hands Meeting 2003*. 2003: East Midlands Conference Centre, Nottingham.
3. Sommariva, R., et al., *OH and HO2 chemistry in clean marine air during SOAPEX-2.* Atmos. Chem. Phys., 2004. **4**(3): p. 839-856.
4. Stanton, J.C., *Field and Modelling Studies of Volatile Organic Compunds in the Troposphere*, in *School of Chemistry*. 2006, University of Leeds.
5. Atkinson, R., et al., *Ox, HOx, NOx, SOx reactions: summary of currently recommended data.* . 2006, IUPAC Subcommittee for Gas Kinetic Data Evalaution.

6. Bloss, C., et al., *Evaluation of detailed aromatic mechanisms (MCMv3 and MCMv3.1) against environmental chamber data.* Atmos. Chem. Phys., 2005. **5**(3): p. 623-639.
7. Atkinson, M. *e-Science*. [cited 19th February 2009]; Available from: http://www.rcuk.ac.uk/escience/default.htm.
8. Shadbolt, N., T. Berners-Lee, and W. Hall, *The Semantic Web Revisited.* IEEE Intelligent Systems, 2006. **21**(3): p. 96-101.
9. Carole Anne, G. and R. David Charles De, *myExperiment: social networking for workflow-using e-scientists*, in *Proceedings of the 2nd workshop on Workflows in support of large-scale science*. 2007, ACM: Monterey, California, USA.
10. Oinn, T., et al., *Taverna: a tool for the composition and enactment of bioinformatics workflows.* Bioinformatics, 2004. **20**(17): p. 3045-3054.
11. Carrere, S. and J. Gouzy, *REMORA: a pilot in the ocean of BioMoby web-services.* Bioinformatics, 2006. **22**(7): p. 900-901.
12. Peleg, M., I. Yeh, and R.B. Altman, *Modelling biological processes using workflow and Petri Net models.* Bioinformatics, 2002. **18**(6): p. 825-837.
13. 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.
14. Schraefel, m.c., et al., *Making Tea: Iterative Design through Analogy*, in *Designing Interactive Systems, 2004*. 2004: Cambridge Massachusetts, USA.
15. Coles, S.J., et al., *An E-Science Environment for Service Crystallography from Submission to Dissemination.* Journal of Chemical Information and Modeling, 2006. **46**(3): p. 1006-1016.
16. Watson, P. *e-Science in the Cloud with CARMEN*. in *Parallel and Distributed Computing, Applications and Technologies, 2007. PDCAT '07. Eighth International Conference on*. 2007.
17. Baeza-Romero, M.T., et al., *A combined experimental and theoretical study of the reaction between methylglyoxal and OH/OD radical: OH regeneration.* Physical Chemistry Chemical Physics, 2007. **9**(31): p. 4114-4128.
18. Arnstein, L., et al., *Labscape: a smart environment for the cell biology laboratory.* Pervasive Computing, IEEE, 2002. **1**(3): p. 13-21.
19. Wendy, E.M., et al., *The missing link: augmenting biology laboratory notebooks*, in *Proceedings of the 15th annual ACM symposium on User interface software and technology*. 2002, ACM: Paris, France.
20. Simmhan, Y., B. Plale, and D. Gannon, *A survey of data provenance in e-science.* SIGMOD Rec., 2005. **34**(3): p. 31-36.
21. Luc Moreau, et al., *Special Issue: The First Provenance Challenge.* Concurrency and Computation: Practice and Experience, 2008. **20**(5): p. 409-418.
22. Ludäscher, B., et al., *Scientific workflow management and the Kepler system.* Concurrency and Computation: Practice and Experience, 2006. **18**(10): p. 1039-1065.
23. Foster, I., et al. *Chimera: a virtual data system for representing, querying, and automating data derivation*. in *Scientific and Statistical Database Management, 2002. Proceedings. 14th International Conference on*. 2002.
24. 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.
25. Zhao, J., et al., *Mining Taverna's semantic web of provenance.* Concurrency and Computation-Practice & Experience, 2008. **20**(5): p. 463-472.

26. Scheidegger, C., et al., *Tackling the Provenance Challenge one layer at a time.* Concurrency and Computation-Practice & Experience, 2008. **20**(5): p. 473-483.
27. 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.
28. 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.

Chapter 6 Software Development Methodology

This chapter describes the software development methodology adopted during the development of an Electronic Laboratory Notebook (ELN), to support provenance capture and sharing across the MCM community. A software development methodology is a framework, adopted in order to structure and control the process of developing a software system. When developing the ELN a hybrid methodology was adopted, based on Scenario Based Design (SBD) [1], but also encompassing elements of task analysis [2]. The hybrid methodology adopted consists of 4 development phases: analysis of current practice; design of the ELN; implementation of the ELN; evaluation of the ELN. These components are addressed by chapters 7, 8, 9 and 10 of my thesis, respectively. The methodology presented in this chapter, can be used as a guide to the structure and content of the remainder of this thesis. Although presented as a linear progression through the 4 development phases, development of the ELN involved multiple iterations over each of the phases informed by feedback gathered from ELN stakeholders. The content of the following chapters represent the current state of each of the development phases.

This chapter starts with a discussion of two approaches to representing current and envisioned working practices: first, scenarios, essentially high level stories; secondly, task analysis, essentially a detailed, structured model of the activity in question. The benefits and drawbacks of each of these approaches are then discussed. The chapter concludes with a description of the hybrid software development methodology, consisting of components of scenario based design and task analysis.

6.1 Scenarios

In this section scenarios, the basic components of scenario-based design, are described and discussed. This section breaks down into four sub-sections: first, the complex and challenging nature of the software development process and the emergence of scenarios as a means of addressing this complexity; secondly, the question “what is a scenario?” is addressed; thirdly, the benefits of using scenarios in software development processes are considered; finally, the drawbacks of using scenarios in software development processes are considered.

6.1.1 Software Development is a Complex and Challenging Process

Developing a software system, such as the ELN, is a complex process, Carroll identifies six characteristics of the software development process [3] that make it challenging to successfully execute.

- At the outset of the software development process the situation in which the software will be deployed is not fully understood or completely specified. So the first step of the development process is to map the current situation that will be modified by the software development process.
- There is little guidance on what design moves exist and which of the number of possible design moves to adopt. Given a mapping of the current situation, a design can generate many possible 'design moves' by reasoning about how to improve the situation, in the context of the artefact being designed.
- The goal (i.e. the final piece of software) of a software development process cannot be known in advance. The goal of the development emerges from the development process itself (and is dependent on a large number of variables and subjective decisions).
- Tradeoffs must be made amongst the complex and interrelated components of the software being developed. Assuming the project has a fixed set of resources, components of the software being developed will compete for these resources.
- Developing software draws upon a wide variety of skills and knowledge. The stakeholders of the development must be engaged in the development; as no one individual has the knowledge and skills to required to complete the analysis and design of software system.
- The software being developed will impact on working practices in a wide variety of ways, over an extended period of time. Software systems affect the way in which people interact with their environment, often in ways that are unforeseen at design-time.

The six challenging characteristics of the software development process, demonstrate the fluidity and complexity of the development process. Scenarios embrace the complexity and fluidity of the design process as an understanding of the problem domain is developed [1].

6.1.2 An Introduction to Scenarios

What is a scenario?

Scenarios are stories, usually represented in text form, about people and the activities they are involved in, including details of their interactions with each other and with information systems [3]. So scenarios can be used within software development to capture current working practices (during the analysis phase) or envision future practice (during the design phase). Scenarios typically include the four following components [3].

- **A setting:** provides the context, such as the physical setting (e.g. an office, a research laboratory etc.) and the background of the people involved (e.g. James is a PhD student in his first year).
- **Actors:** Within a scenario there will be one or more actors, i.e. people, performing activities and interacting with each other and information systems.
- **Goals:** Each actor within a scenario will have a goal or set of goals they are seeking to achieve, within the context provided.
- **Actions and Events:** The sequence of actions and events presented within the scenario, form a plot or storyline, actions and events maybe conducive to achieving an actor's goals, or maybe disruptive to achieving an actor's goals.

Why use scenarios?

In order to understand the motivation behind using scenarios it is useful to consider software applications from a socio-technical systems perspective [4, 5], where computer hardware and software make up the technical component of the system and the associated people, groups and communities make up the social component. Given this perspective the nature of a software application can be seen to inevitably impact human activity, and conversely human activity presents a set of conditions that the software application must fulfil. So, in order to successfully develop software applications the constraints presented by human activity can be used to develop the set of conditions the information system should satisfy (i.e. the requirements specification). The purpose of using scenarios is to capture the constraints presented by human activity, in order to enable these constraints to inform the design of a software application. This purpose makes scenarios inherently user-orientated, and aligns well with the goals of this research (i.e. to develop a user-orientated approach to provenance within an atmospheric chemistry community).

How to develop scenarios?

As stated in the preceding discussion, scenarios are used to capture or envision working practices and interactions with software applications. As scenarios are inherently user-orientated, developing scenarios requires significant input from potential users and other stakeholders. Typically an IT analyst will work in conjunction with users and stakeholders in order to create a first draft of a scenario, and then iteratively refine the scenario through discussions with users. Scenarios are intended to be used in order to explore the problem domain, and so changes to scenarios are welcomed as understanding of the problem domain develops.

6.1.3 The Benefits of Using Scenarios

Having answered the question "what is a scenario?", in the preceding sub-section, the benefits of using scenarios are now discussed. In particular the mapping between the benefits of using scenarios and the key challenges of the ELN development is highlighted.

- Each scenario developed is a concrete, tangible object that a stakeholder can engage with; stakeholders understand scenarios as low-fidelity simulations [1] of the activity in question. This benefit helped to bridge the gap in terminology between the informatics and atmospheric chemistry stakeholders in the ELN development project.
- Scenarios are flexible, easy to develop and disposable (if necessary). This feature of scenarios was important in the early phases of ELN development as I explored, with members of the atmospheric chemistry community, a previously unmapped domain (the role of provenance in the atmospheric chemistry community).
- Scenarios are focused on user activity. This feature of scenarios ensured that the focus of the development methodology was aligned with the focus of the project as a whole, to adopt a user-orientated approach to provenance within the MCM user community.
- Using scenarios engages stakeholders at the outset of a development project; allowing them to influence the vision and assumptions that drive the project.

6.1.4 The Drawbacks of Using Scenarios

Having considered the benefits of using scenarios, it is appropriate to also consider the drawbacks of scenarios; in order to ensure an understanding of the limitations of scenarios and the potential to mitigate these limitations.

- There are a large number of potential scenarios that could be developed during any given software development project. Carroll [3] suggest that ten to twelve representative scenarios is an appropriate number of scenarios to develop during a project. This suggestion raises questions [6] such as: “how is it ensured that the scenarios are representative and there is no bias in the scenario selection criteria?”; and, “what factors inform the scenario selection criteria?”.
- Scenarios can prematurely commit to design decisions during the analysis phase of a project [6]. If the premature commitment to design decisions is not desirable, that scenario must be crafted carefully, negating a key benefit of scenarios; they are cheap and quick to develop.
- Scenarios can lack sufficient detail to inform the design process [6]. Scenario-based design proponents would suggest that this lack of detail, is important during the design of a system to stimulate debate and discussion about the problem space.

6.2 Task Analysis

Having introduced scenarios, as a means of capturing or envisioning working practices, in the preceding section, in this section, task analysis, as a complementary means of capturing or envisioning working practices, is introduced. Three sub-sections are presented: first, an introduction to task analysis; secondly, the benefits of using task analysis in software development processes; thirdly and finally, the drawbacks of using task analysis in software development processes.

6.2.1 An Introduction to Task Analysis?

What is a task analysis? The goal of task analysis [2] [7] is to develop detailed and structured representations of the activities and cognitive processes taking place in current or envisioned working practices. A task analysis consists of two core components: first, a description of the world, i.e. the domain being studied (similar to the setting in a

scenario); secondly, a description of how tasks are performed within the world [2] (similar to actions and events in a scenario). Task analysis does not explicitly address the actors and their goals that are integral components of a scenario. So scenarios can be seen to be broader in scope and more orientated to the human aspects of a problem, than task analysis.

Why use task analysis? Task analyses of current practice can be used to inform the design of a software application, by capturing the detail of the activities the software application must support [7]. Task analyses of envisioned practices, can be used to ensure that the envisioned practices are coherent and can be used as a means of communication between software developers and users (at the design stage).

How to develop task analysis? Developing a task analysis, for a given activity, requires the activity in question to be identified and then a hierarchical decomposition of the activity to be defined [8]. Typically this activity will involve an IT analysts working in conjunction with individuals or groups that execute the task in question.

6.2.2 Benefits of Task Analysis

Having answered the question "What is task analysis?", the benefits of using task analysis within a software development methodology are described in this sub-section.

- Task analysis ensures that existing working practices are well understood, and that the important details of existing working practice are captured. These details are required to inform the design of the software application and envisioned working practices.
- Defining envisioned working practices using task analysis guides the design of the software application and provides clear criteria against which the application can be tested/evaluated against.

6.2.3 Drawbacks of Task Analysis

Having considered the benefits of task analysis, in the preceding subsection, the drawbacks of task analysis are described below.

- "Traditional Task analysis assumes that there is a correct and complete symbolic description of user tasks" [1]. This assumption, that it is possible to develop a "complete and correct" description of the tasks executed by system users and other system stakeholders, can make task analysis a time consuming process, with no guarantee of an appropriate outcome being developed.
- Carroll [1] suggests that task analysis is a "mechanical process of articulating and implementing 'correct' representations", focusing on the capture of task descriptions rather than informing the design process.

6.3 Hybrid Methodology

Having introduced scenarios and task analysis as tools for describing working practices, this chapter concludes with a description of the hybrid development methodology adopted for the development of the ELN. The hybrid methodology draws together elements of scenarios, scenario based design and task analysis, and other custom elements. This section begins by discussing the alignment between scenario based design and task analysis, highlighting where each approach can deliver value in the development of the ELN. The hybrid methodology is then outlined, with each of the key phases (analysis of current working practice, ELN design, ELN implementation and ELN evaluation) discussed in detail.

6.3.1 Aligning Scenario Based Design and Task Analysis

A hybrid methodology, for development of the ELN, was required due to the combination of two factors: first, a vague, poorly understood problem domain (i.e. the role of provenance within the atmospheric chemistry community); and secondly, the need to capture detailed provenance in a rigorous and formal manner requiring an in-depth understanding of current and envisioned working practices.

Scenarios, with their strengths including: capturing contextual factors; being cheap to develop; and being easy, for users, to understand and relate to; are well suited to addressing the first issue, a vague, poorly understood problem domain. The lack of detail in scenarios makes them unsuitable for addressing the second issue, developing an in-depth understanding of working practices. Task analysis, on the other hand, is well suited

to addressing the second issue, developing an in-depth understanding of working practices (current or envisioned).

So in crude terms the strengths and weaknesses of task analysis and can be seen to be complementary. The inclusion of elements of both scenario based design and task analysis, within the hybrid methodology, ensures that both the key issues development issues; a vague, poorly understood problem domain and developing an in-depth understanding of working practices; can be addressed.

6.3.2 Hybrid Methodology Outline

The hybrid methodology takes the core elements of the scenario-based design methodology, and adds task analysis activities and other activities to tailor the methodology to the development of the ELN. The hybrid methodology, shown in Figure 6.1, consists of four phases: analysis of current working practices; ELN design; ELN implementation; and ELN evaluation. Figure 6.1 shows the iterative nature of the hybrid methodology, with the scope for feedback between each of the phases. For example during the design phase it often became clear that the understanding of current practices was insufficiently developed, so a return to the analysis phase was required. In the remainder of this section each of the development phases is described.

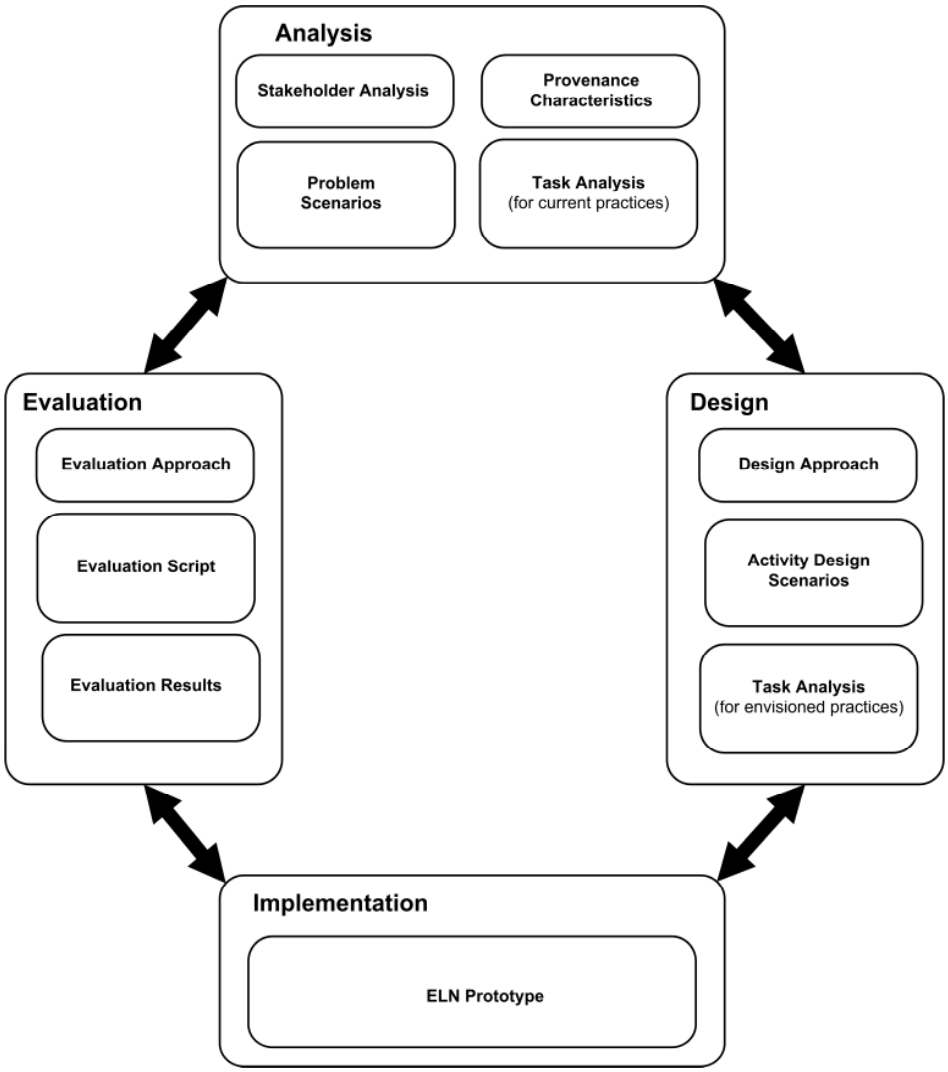


Figure 6.1. The hybrid software development methodology adopted for the development of the ELN. Four phases are depicted: analysis of current practice; ELN design; ELN implementation; evaluation of the ELN. Within each phase the key artefacts (e.g. scenarios, designs etc.) are identified. The diagram depicts exploratory nature of the methodology, with the scope to iterate over the 4 development phases (Analysis, Design, Implementation and Evaluation) and the scope to return to previous phases to conduct additional work.

6.3.3 Analysis of Current Working Practices

Analysis is the process of understanding the current environment and activities, in to which the software being developed with be implemented. The analysis phase of the hybrid software development methodology, consisted of four elements.

- A stakeholder analysis, seeking to understand the individuals and groups that have an interest in the development of the ELN.
- A set of problem scenarios[11]. Each scenario was developed iteratively, in conjunction with members of the atmospheric chemistry community, to capture current working practices and provide additional context complementary to the stakeholder analysis.
- A set of provenance characteristics, that capture the essence of current practices, based upon an analysis of the problem scenarios.
- A task analysis of a scenario, to capture the detail of current working practices with laboratory notebooks.

6.3.4 ELN Design

Design is the process of envisioning new software artefacts and new working practices, which can be implemented in subsequent phases of the software development methodology. In the design phase, three types of scenario are used, each described in the remainder of this section.

Activity design scenarios [3]: Are the first concrete description of the new software functionality and new working practices, to be generated in the scenario based design methodology. Each activity design scenario will map to a problem scenario, i.e. envisioned practices in the activity design scenario map to a set of current practices in a problem scenario. There maybe multiple activity design scenarios generated for each problem scenario, mapping to multiple possible approaches to improving current practices.

[11] Problem scenarios tell the story of current practice, within the domain in question. It is important to capture current practice, as within scenario-based design current practice informs the design of new artefacts and activities. The contents of problem scenarios must be considered carefully, in order to ensure that the scenario is understandable by all the project stakeholders and highlights the elements of current practice that have implications for the design process. Problem scenarios are named in reference to their role in developing understanding of the problem domain, rather than in reference to a particular problem being addressed.

The design phase of the hybrid software development methodology, consisted of three elements. First, a design approach for the ELN, based upon the provenance characteristics identified in the analysis phase. Secondly, a set of activity design scenarios (each corresponding to a problem scenario, from the analysis phase) were developed to describe envisioned working practices with an ELN. Thirdly, where appropriate a task analysis of the activity design scenarios was performed to inform the interaction and information design of the ELN.

6.3.5 ELN Implementation

The implementation phase of the hybrid software development methodology consisted of the ELN prototype. A vertical prototype was developed, i.e. a prototype with limited user functionality, but a full realisation from user-interface to back-end database.

6.3.6 ELN Evaluation

Based upon the scenarios developed in the earlier phases of the methodology, a prototype or production quality software application can be developed. Evaluation of this software application can take two forms: summative evaluation, judging the quality of and benefits delivered by the software application at the end of the development project; or formative evaluation, judging the quality of and benefits delivered by the software application during the development project, in order to improve the requirements specification or understanding of the problem domain. Following formative evaluation a new iteration over the development phases can commence, in order to further refine the prototype being developed.

The evaluation phase of the hybrid software development methodology was formative and consisted of three elements. First, an evaluation approach, which defined a high-level strategy for evaluation of the ELN prototype and the user-orientated approach to provenance. Second, an evaluation script was developed to structure the user evaluations of the ELN. Thirdly, the evaluation results.

Chapter Summary

This chapter has presented an overview of the software development methodology adopted during the development of the ELN, a hybrid methodology composed of elements of scenario based design and task analysis. The remainder of this thesis follows a structure defined by the hybrid methodology, so the next chapter considers the analysis of current practices, Chapter 8 describes the design of the ELN, Chapter 9 the implementation of the ELN and Chapter 10 the evaluation of the ELN.

References

1. Carroll, J.M., *Making use is more than a matter of task analysis.* Interacting with Computers, 2002. **14**(5): p. 619-627.
2. Diaper, D., *Understanding Task Analysis for Human-Computer Interaction* in *The Handbook of Task Analysis for Human-Computer Interaction* D. Diaper and N. Stanton, Editors. 2003, CRC Press. p. 5-48.
3. Carroll, J.M., *Making Use: Scenario-Based Design of Human-Computer Interactions*. 2000, Cambridge, Massachusetts: The MIT Press.
4. Emery, F.E. and E. Trist, *Socio-technical systems*, in *Systems thinking*, F.E. Emery, Editor. 1969, Penguin: Harmondsworth, England.
5. Sutcliffe, A.G., et al., *Supporting scenario-based requirements engineering.* Software Engineering, IEEE Transactions on, 1998. **24**(12): p. 1072-1088.
6. Diaper, D., *Scenarios and task analysis.* Interacting with Computers, 2002. **14**(4): p. 379-395.
7. Hackos, J.T. and J.C. Redish, *User and task analysis for interface design*. 1998: John Wiley & Sons, Inc. 488.
8. Annett, J. and N. Stanton, *Task Analysis*. 2000: CRC Press. 242.

Chapter 7 Analysis of Current Practice

This chapter describes the analysis of the current working practices of researchers performing *in silico* experiments, i.e. developing computational models that make use of the MCM. This corresponds to the analysis phase of the hybrid software development methodology (presented in chapter 6). There are four main sections within this chapter. First, a stakeholder analysis is presented, considering stakeholders with a vested interest in the use of ELNs across the MCM user community. Secondly, based upon the stakeholder analysis, a set of problem scenarios are presented and analysed, in order to identify characteristics of the provenance captured when using current working practices. Thirdly, current working practices for capturing data and provenance are explored with reference to a task analysis of a computational model development process. In-depth task analysis maps the processes involved in developing a model, in order to develop an understanding of the provenance required to completely describe a given modelling process.

7.1 Background information

This section provides background to the chapter, defining some terminology and describing some resources, which are referred to throughout the chapter.

Terminology used

Process Provenance: records the process executed to produce a given piece of data.

Scientific rationale: the scientific reasoning behind executing a given process (or set of processes), i.e. why a researcher conducted his or her research in a given way.

Scenario background

Each scenario, in this chapter, refers to an actor (a fictional character, in many ways representative of MCM user community members). Helen is conducting research, during her PhD, using the MCM to develop chamber and field models, further information regarding the actor is presented in the background section of this introduction.

Actor: Helen is a PhD student in an atmospheric chemistry modelling group.

Research Interests: Developing models and chemical mechanisms for field and chamber models, making use of the MCM as the starting point for chemical mechanism development.
Priorities: Deriving results of interest to the wider atmospheric chemistry community, for publication in papers or thesis. Development of a modelling framework that enables model reuse and helps reduce duplication of modelling effort.

Provenance documentation

In the analysis of current practice a variety of existing provenance documentation was reviewed; with the EXACT provenance documentation playing leading role. The EXACT campaign [1] consisted of series of chamber experiments conducted in order to facilitate the development of mechanisms for several aromatic compounds. The EXACT campaign provenance documentation [Claire Bloss, personal communication, January 2008] consists of a number of word processor documents that record the development of the various computational models and associated mechanisms used throughout the modelling of the EXACT campaign. The provenance documentation was developed by Claire Bloss, a former member of the University of Leeds, Atmospheric Chemistry Modelling Research Group, and was made available to use in this research.

The EXACT provenance documentation was developed for use by a small group of researchers associated with the development of the MCM. This group used the provenance documentation to enable model output data to be shared and interpreted within the group, with the aim of developing new understanding of the atmospheric mechanisms of aromatic compounds. As the audience for the provenance documentation was restricted, to a defined group, the reader was assumed to understand the context of the documentation, including:

- The overall goals of the research being conducted;
- The nature of the chamber experiments being modelled;
- The goals of each *in silico* experiment;
- A set of terminology associated with the MCM.

As this context is assumed, it is not incorporated with the provenance documentation. The analysis of the EXACT provenance documentation, in this chapter, takes the documentation out of context. I assess the provenance documentation from the perspective of a member of the wider atmospheric chemistry community, with little understanding of

the context required to interpret the provenance documentation. I adopted this perspective because a key goal of developing the ELN is to capture provenance that can be used across the atmospheric chemistry community. So any shortcomings of the provenance documentation identified below are not intended to reflect on the provenance documentation's fitness for its original purpose; but reflect opportunities to enhance provenance documentation of this type.

The EXACT provenance documentation can be considered an atypical example of provenance documentation, within the MCM user community, as it is comprehensive and was developed with the goal of ensuring that the EXACT campaign modelling data could be effectively archived. A more typical example of provenance documentation was also reviewed, the SOAPEX campaign [2]. In this case no formal provenance documentation was available (the laboratory notebook of the researcher who conducted the research was not available[12]), so an incomplete provenance had to be aggregated from a number sources: comments within the model code, the contents of a research paper [2], and relevant PhD thesis chapters [2].

Where excerpts of the EXACT provenance documentation are provided within this chapter, they are included to support the analysis of the scenarios and the determination of provenance characteristics. The scientific content of the provenance excerpts is not particularly significant, in terms of determining provenance characteristics, so explanation of the excerpts (within the text) focuses on the provenance content rather than the scientific content.

7.2 Stakeholder Analysis

In this section, stakeholders across the atmospheric chemistry community, with a potential interest in the development of an ELN for computational modelling, are analysed. The purpose of this stakeholder analysis was to establish which stakeholders to focus on during the development of the ELN, i.e. the requirements of those stakeholders with significant interest and influence guided the development of the ELN. The stakeholder

[12] This highlights one of the main issues of the traditional laboratory notebook, when a researcher leaves a research group the notebook either stays with the group or leaves with the researcher. The ELN will address this issue by allowing both parties to access a digital provenance record.

analysis was informed by my experiences in and interactions with the atmospheric chemistry community; with feedback sought from members the stakeholder groups during the iterative development and refinement of the analysis.

Each stakeholder is analysed with respect to two dimensions: first, the stakeholder's interest in the development of the ELN; and secondly, the stakeholder's ability to influence adoption of the ELN across the atmospheric chemistry community. Five stakeholder groups are analysed in this section: the MCM developers, research group leaders, the researchers performing *in silico* experiments, the research councils, and scientific publishers. This section concludes with a review of the implications of this stakeholder analysis, for the development of the ELN. It is assumed that a given individual may be a member of more than one of the stakeholder groups considered. For example a research group leader, may also perform the researcher role (i.e they may perform *in silico* experiments as well as supervising other researchers). Or a researcher may also have a role within the MCM development team. The stakeholder groups are considered, in turn, in a sub-section below.

7.2.1 MCM Developers

Stakeholder description: MCM developers are members of the MCM development panel, a small group of experts responsible for the development and maintenance of the MCM. The development of the MCM is described in full in Chapter 5.

Interest: MCM developers have a substantial interest in the use of an ELN across the MCM user community, as described in chapter 5. Gaining access to the data and associated provenance produced by researchers performing *in silico* experiments using the MCM, would bring significant benefits, e.g. allowing MCM developers to find and review detailed reports on the performance of the MCM and factor the findings of these reports into the ongoing development of the MCM.

Influence: MCM developers have limited influence over the uptake of the ELN across the MCM-user community. Researchers could be encouraged to use the ELN by making the software freely available, from the MCM website, and packaging the ELN with existing modelling tools provided for use with the MCM. Making use of the ELN a condition of

using the MCM is not a reasonable option, as it could be viewed as too directive and inflexible and so discourage potential users of both the MCM and ELN.

7.2.2 Research Group Leaders

Stakeholder description: Research group leaders (i.e. Professors and Lecturers) are responsible for providing leadership and guidance to researchers (i.e. PhD students and post-doctoral researchers), performing *in silico* experiments. Research group leaders typically hold permanent positions and have a set of established (but often involving) research interests.

Interest: For research group leaders, whose research group members perform *in silico* experiments with the MCM, the use of an ELN could bring a variety of benefits including: the ability to monitor the progress of researchers; enabling research group members to work in a more efficient and rigorous manner, so producing more, or higher quality, research outputs for the research group; ensuring it is possible to archive data produced by the research group, enabling continuity of research (as the membership of the research group changes over time).

Influence: Research group leaders typically have the some influence, but not complete influence, over the working practices of researchers within their group. So research group leaders could require members of their group to use the ELN to manage data and provenance capture. There is also the potential for research group leaders to link ELN use with existing reporting methods; e.g. ELN records could be linked to the reporting of the progress of PhD students (as defined by institutional regulations).

7.2.3 Researchers

Stakeholder description: Researchers (i.e. PhD students and post-doctoral researchers), are directly responsible for performing *in silico* experiments. Researcher typically hold fixed term positions and are developing their research interests. For researchers aiming to develop an academic career, publication of their research is a priority (to build their research profile).

Interest: Researchers will be responsible for performing data and provenance capture with the ELN, and so have the greatest interest, of all the stakeholders, in the details of the design and implementation of the ELN. It is likely that the researchers' interest in the impact of the ELN on their day-to-day working practices will dominate their motives.

Influence: Researchers have the greatest influence, of all the stakeholders, over the adoption of the ELN. The ELN must deliver sufficient benefits to the researcher in the course of their every-day work, to motivate the researcher to overcome whatever learning curve and changes to working practices are associated with adopting the ELN. As academic research tends to be a fairly independent activity, other stakeholders have the potential to influence the adoption of the ELN across the MCM user community, but if the researcher finds the ELN difficult to use or the ELN burdens users with additional work (without sufficient benefits being clearly delivered) they will not make use of the ELN.

7.2.4 Research Funders and Sponsors

Stakeholder description: Research councils, particularly NERC (The Natural Environment Science Research Council) within the UK, are responsible for funding the majority of research across the atmospheric chemistry community.

Interest: Research councils and other research funding bodies are increasingly interested in ensuring that data, produced by the research that they fund, is effectively archived to ensure data reusability. Data is viewed as a valuable, strategic resource and efforts must be made to ensure its long-term sustainability by developing archives of research data. This view is outlined in NERC data policy handbook [3].

Influence: By defining and enforcing data provenance policies, necessarily at a generic level (rather than a domain specific level), the research councils can raise the profile of data provenance on the researchers' agenda, hopefully ensuring that capturing provenance to enable archiving of research data becomes an integral part of the researchers' role. So whilst research councils are unlikely to be directly interested in an ELN for a relatively small research community, their emerging interest in data and provenance may help to bring about the cultural shift that increases the likelihood of the ELN being adopted across the MCM community.

7.2.5 Publishers

Stakeholder description: Publishers provide mechanisms for distributing research findings, typically in the form of physical or online journals.

Interest: As with the research councils, publishers are taking an increasingly active interest in the role of data and its provenance within research communities. Publishers' interest is motivated by the potential for providing value-added services to their readers, based upon data and associated provenance. For example Project Prospect [4], provides enhanced access to some RSC (Royal Society of Chemistry) journal articles, linking paper content such as chemical species names to online databases to enable the reader to access information about the species in question. Another area of interest for publishers is connecting journal articles to the provenance of the underpinning scientific process (e.g. the provenance captured by the ELN).

Influence: As with research councils, publishers will influence the adoption of the ELN across the MCM user community, by prompting a cultural shift where the value of data and provenance is fully recognised.

7.2.6 Implications of Stakeholder Analysis

Having considered five stakeholder groups, in the analysis presented above, this sub-section identifies key implications for the development of the ELN. Researchers, who will potentially use the ELN, have the greatest interest in the detail of the design of the ELN. Also within an academic research environment, researchers typically select the tools that match their individual requirements, so researchers will play an important role in decision-making processes that determine if the ELN is adopted. Given the primary importance of satisfying the requirements of researchers, the remainder of this thesis focuses on the researcher and their relationship with the ELN. The requirements of research group leaders will also be considered, as they have a longer-term perspective, than researchers, in terms of archiving data. The requirements of MCM developers, described in chapter 5, in terms of provenance representation will also, where possible, be addressed. The requirements of publishers and research councils will not be addressed, as at the current time the role of these two stakeholders with regard to data archival and provenance is evolving significantly and far from clear.

7.3 Scenarios and Scenario Analysis

This section introduces and analyses problem scenarios (as introduced in Chapter 6) used to capture the current working practices of researchers performing *in silico* experiments. This section consists of five subsections:

1. First, the way in which the two scenarios presented in this thesis were selected and developed.
2. Secondly, Scenario 1, capturing data and provenance as a model is developed, is presented and subsequently analysed.
3. Next, scenario 2, a researcher reviewing their own data for inclusion in a publication, is presented and analysed.
4. This section then concludes with a summary of the provenance characteristics identified in the course of analysing the scenarios.

A distinction is drawn between the provenance generation and provenance use scenarios, as they are analysed in different ways. The provenance generation scenario, scenario 1, is analysed by reviewing a set of provenance records. The provenance use scenario, scenarios 2, is analysed by determining a set of queries the researcher may wish to ask in the context of the scenario in question, and testing the ability of a set of provenance records to answer these queries.

7.3.1 Developing and Selecting the Scenarios

This sub-section consists of two components: first; a discussion of how the two scenarios presented in this chapter, were selected from a wider set of scenarios; and secondly, a discussion of the information sources used to develop the scenarios.

Presented in the following sub-sections are two problem scenarios. These scenarios were selected from a wider set of scenarios, listed below, considered during the analysis phase of the ELN development. In the list below the scenarios selected for presentation in this thesis are marked (S). This initial list of scenarios was generated in conjunction with atmospheric chemists, in order to develop an understanding of the problem domain. The scenario list is not intended to be exhaustive, but representative of the types of provenance-related activities taking place across the MCM-user community.

1. A researcher capturing data and provenance when performing *in silico* experiments; (S)

2. A supervisor conducting a routine (weekly, monthly, etc.) review of the provenance records of researchers (to ensure the quality of the provenance records in question);
3. A researcher reinterpreting their own data for inclusion in a publication; (S)
4. A supervisor reviewing the work conducted within their research group, with a view to identifying potential fruitful research directions;
5. A researcher interpreting the data of another scientist to build on his or her work;
6. A researcher conducting a peer review of a publication, where the reviewer wishes to access provenance underpinning data presented in the publication;
7. An MCM developer gathering feedback on the performance of the MCM by reviewing publications where the MCM is applied;
8. A researcher reviewing the data they included within a publication, in order to answer the referees comments;
9. A researcher reinterpreting data (either own data or the data of a third party), adding new annotations and interpretations of the *in silico* experiment.

The scenarios identified above fall into three categories: first, an individual creating provenance; secondly, an individual reviewing their own data and provenance; and thirdly, an individual reviewing the data and provenance of a third party. The two scenarios (scenarios 1 and 3 in the list above) presented in this chapter were selected based upon two criteria: ensuring coverage of two of the three scenario categories (identified above); and allowing the relationship between the researcher, performing *in silico* experiments, and the ELN to be explored in depth.

Information sources

The scenarios presented below were developed iteratively drawing on three information sources: first, provenance documentation from the problem domain; secondly, interviews and informal discussions with members of the atmospheric chemistry community; thirdly, my personal experience performing *in silico* experiments. Each of these information sources is described in further detail in the remainder of this section.

Provenance documentation

Provenance documentation, as described in the background section of the chapter, was reviewed in order to inform the development of the scenarios.

Interviews and informal discussions

During the iterative development and analysis of the scenarios, a number of informal interviews were conducted with members of the atmospheric chemistry community. PhD students and post-doctoral researchers, and more senior, experienced academics were involved in the interviews and discussions to ensure that perspectives from across the community were considered. The interviews and discussions took place in two phases: first, stakeholders identified important provenance related issues and potential scenarios of interest; and secondly, stakeholders were presented with draft scenarios and asked for feedback, allowing the scenarios to be iteratively refined as understanding of the domain developed.

Personal experience

Having adopted an ethnographic approach [5] throughout the course of my PhD, my personal experience played an important role in developing and analysing the scenarios. Three elements of my personal experience where involved in the scenario development:

- Benchmarking the OSBM (as described in Chapter 3) required the development of a complete understanding of how a number of benchmark datasets were created, which relied on the insight drawn from available provenance records;
- Developing and testing the OSBM I created and then made subsequent use of provenance records, experiencing first-hand some of the issues with current provenance capture practices;
- Observing the provenance capture practices adopted by a number of researchers, and the associated provenance issues that have arisen within the group. The use of my personal experience in developing and analysing the scenarios, inherently reduced the breadth of the analysis of current working practices, but also enabled a depth of insight to be developed beyond that typically accessible by observational methodologies (i.e. non-ethnographic approaches, where the research maintains a distance from their subject in order to ensure an objective perspective is developed) [6] [7] .

In the following sub-sections of this chapter two scenarios are presented along with analyses that identify the provenance characteristics associated with each scenario.

7.3.2 Scenario 1: Capturing Data and Provenance

In this sub-section problem scenario 1, capturing data and provenance during the development of a computational model using the MCM, is presented. Scenario 1 is then analysed, with reference to provenance documentation from the EXACT campaign (as described in Section 7.1).

7.3.2.1 Current Practice (Problem Scenario)

> Helen is developing computational models of a set of chamber experiments. The models are developed iteratively. A modelling iteration typically involves model development, running the model, and analysing the model output to identify the appropriate model development for the next iteration. The goal of her piece of modelling research is to obtain a good agreement between model output and experimental measurements, deriving some insight into the chemical mechanism in the process.
>
> Helen records the iterative modelling process in her lab-book, she also takes back-ups of the model at various points (providing a snap shot of model development). The description of the modelling process Helen records in her lab-book, captures the essence of the process but is *ad hoc* in nature and incomplete. Details such as the exact sequence of editing the mechanism and locations for output files are often not captured. Helen finds the prospect of fully documenting her modelling process in her lab-book unpalatable as it is time consuming, of limited value (at the time the task is undertaken) and, she feels it detracts from the scientific work that is her focus.

Problem Scenario 1: capturing data and provenance during the development of a computational model using the MCM.

7.3.2.2 Analysis of Current Practice

In this section problem scenario 1, presented above, is analysed with reference to the EXACT provenance documentation. The analysis seeks to determine the characteristics of the provenance captured by researchers developing models using the MCM. In this case two distinct types of provenance characteristics are considered: first, informational

characteristics, these are characteristics related to the information content of the provenance captured by a performing *in silico* experiments; secondly, functional characteristics, i.e. characteristics related to the way in which provenance is captured, are addressed. Informational characteristics are examined with direct reference to excerpts from the EXACT provenance documentation, whereas functional characteristics are based upon my observations and discussions with members of the atmospheric chemistry community.

Informational characteristics of provenance capture

Characteristic 1. Capturing provenance for human and computational processes

Within process provenance researchers typically capture their scientific process in terms of what they do (e.g. added a reaction to the mechanism), along side what the computer does (e.g. which model was run when). Provenance excerpt 1, shown below, records that the researcher has updated the rate coefficients of three reactions (a human process). Also recorded is the name of the model run (Pxylene019w.fac), referring to a computational process, which can be used to find the associated model output within the researcher's modelling archive. The provenance excerpt shows that the amount of provenance captured for human processes is much greater than that captured for computational processes. This has two potential implications. First, the researcher in question perceives greater value in a provenance record of human processes (i.e. changes to the mechanism) than computational process (i.e. where and when the model was executed). Secondly, the computational processes are assume to be self-describing (i.e. the model code is archived, so could be run again by any interested party).

Pxylene019w.fac

Change pxyol + NO_3 rate *coefficient* following *experimental* information from John Wenger.

```
G833 % 3.48D-11*0.39          : PXYLOL + NO3 = PXY1O + HNO3       +S833;
G834 % 3.48D-11*0.51          : PXYLOL + NO3 = NPXYOLO2           +S834;
G835 % 3.48D-11*0.10          : PXYLOL + NO3 = PXYOLO2 + HNO3     +S835;
```

Provenance Excerpt 1: Taken from EXACT documentation (pxylene_modelversions.doc), shows the updated rate-coefficients for three reactions and an associated annotation.

N.B. In the provenance documentation the species are referred to by their MCM name (a unique identifier in the MCM database); in the case of this excerpt it is worth noting that pxyol refers to 2,5-dimethylphenol.

Characteristic 2. The reasoning behind the scientific process is important

Within the provenance records, the researcher's reasoning for adopting a given scientific process and the scientific process itself are often paired. In provenance excerpt 2, the branching ratios for product channels of the benzene plus hydroxyl radical + (O_2) reaction, have been edited by the researcher from the original branching ratios (as stated in the MCM) to reflect the findings of a recent paper [8]. The reference provided (Volkamer et al.) gives very limited information, presumably the reference would be obvious to the researcher who conducted the modelling, but for third parties some effort is required to resolve the reference to the appropriate publication.

benbox105w.fac (23/04/02)
Change initial product channel branching ratios in accordance with work of Volkamer et al.

```
G47 % 3.58D-12*EXP(-280/TEMP)*0.352  : BENZENE + OH = BZBIPERO2    +S47;
G48 % 3.58D-12*EXP(-280/TEMP)*0.118  : BENZENE+ OH = BZEPOXMUC+ HO2+S48;
G49 % 3.58D-12*EXP(-280/TEMP)*0.53   : BENZENE + OH = PHENOL + HO2 +S49;
G50 % 3.58D-12*EXP(-280/TEMP)*0.00   : BENZENE + OH = BZPERO2      +S50;
```

Provenance Excerpt 2: Shows provenance for updating branching ratios for the benzene plus hydroxyl radical + (O_2) reaction, plus an annotation making reference to an associated publication.

In a small number of cases, including Provenance Excerpt 3 where a pair of photolysis rates are increased, provenance records show processes taking place with no associated scientific rationale. There are several possible reasons for the lack of scientific rationale within provenance records: first, there was no scientific rationale for making the change to the mechanism, this raises a number of questions in itself (such as should a researcher be changing the mechanism without any scientific rationale? Does the researcher have a hunch based on experience about why this change should be made, but considers this rationale too speculative to record?); secondly, the researcher did not have time to record the scientific rationale; thirdly, the researcher did not consider the scientific rationale to be important enough to record.

benbox113w.fac (from benzene_modelversions.doc)
Increase photolysis rate of epoxide

```
G59 % J<4>*0.1*0.5     : BZEPOXMUC = MALDIAL + HO2 + CO + HO2 + CO +S59;
G60 % J<4>*0.1*0.5           : BZEPOXMUC = C5DIALO2 + HO2 + CO     +S60;
```

Provenance Excerpt 3: Shows provenance for increasing the photolysis rates of a species (BZEPOXMUC).

Characteristic 3. Annotations are with respect to two frames of reference

Researchers make annotations with respect to multiple frames of reference; i.e. annotations made related to different concepts within the *in silico* experimental domain. Annotations with respect to two, distinct frames of reference can be identified in the EXACT provenance documentation.

- Researchers annotate elements of their scientific process, for example why they changed a given model parameter. Provenance Excerpts 1, 2, and 3 show annotations attached to the scientific processes (i.e. recording exactly how the mechanism was changed and why).
- Researchers annotate *in silico* experiments, as entities in its own right, for example their experimental goals and the conclusions of a given experiment. Provenance Excerpt 4 shows annotations made at an experimental level, considering the comparison of three models (105,106 and 106a). The experiment in question compares three (very similar) models to ensure consistency of the model output. The results of the experiment are also recorded (with only minor differences in model output identified, and these differences can be rationalised).

Toluene mechanism – versions 105, 106 and 106a 6/12/01

All these have the new cresol chemistry from Mike Jenkin.

105 – here the "creso2" lines 873-881 occurs twice.
106 – (i.e. previous 105a) has lines 873 – 881 commented out to solve the above problem.
106a – as 106 but rearranged with new chemistry at the end and all reactions numbered.

Tested all these models – base model is toluene experiment euph221097, and also initialised for cresol experiment (041001). Looking at the cresol concentration in each case 106 and 106a are essentially the same (as expected) and 105 has minor difference, up to ~ 1% for the toluene case and only ~ 0.1% for the cresol experiment

Provenance Excerpt 4: Shows provenance for the testing of three similar models, focussing on the cresol chemistry.

Characteristic 4. Scientist's uses domain specific scientific terminology

When recording process provenance and annotations, researchers make use of domain specific scientific terminology. For example in Provenance Excerpt 3, a researcher refers to updating photolysis rates rather than updating a model input file. The use of domain specific scientific terminology allows information to be recorded quickly, relying on the informational content of the terminology. The use of domain specific terminology can be seen throughout all the provenance excerpts presented in this section.

Functional Characteristics of approach to provenance capture

Characteristic 5. Provenance capture is interleaved with the scientific process

Researchers typically do not perform provenance capture as a separate activity, either before or after executing their scientific process. Provenance capture is typically interleaved with the execution of the scientific process, as ideas occur and the need to record provenance becomes evident to the researcher. For example a researcher would change a reaction within the mechanism, make a note of the change in their laboratory notebook, run the model, perform some analysis on the model output, making notes on the analysis as the pertinent points occurred.

Characteristic 6. Provenance is captured and stored in multiple media

A number of media are used for the capture of provenance including: laboratory notebook, word processor documents (as used for the EXACT campaign provenance documentation), annotations in spreadsheets, file and directory names, and model output files.

Characteristic 7. Provenance capture is a manual activity

Capturing complete and detailed provenance records is a time consuming activity. So researchers keep partial provenance records, attempting to predict where data will be reused and so predict where provenance is required. One of the benefits of using word processor documents, as used in the EXACT campaign provenance documentation, is that model components can be copied and pasted from the model code to the provenance record, reducing the amount of manual activity that the researcher must complete to record their provenance.

Characteristic 8. Provenance capture has a significant learning curve:

The value of provenance is not immediately perceived by new researchers. A learning curve is involved, whereby a researcher comes to value provenance by experience, typically the frustrating experience of not recording provenance and returning to data at a later date only to be unable to interpret it.

7.3.3 Scenario 2: Interpreting or Re-interpreting Data for Publication

In this sub-section scenario 2, where a researcher reviews her own data and provenance records during the preparation of a publication, is presented and analysed. The result of this analysis is a set of provenance characteristics, which at the conclusion of this section will be reconciled with the provenance characteristics associated other scenario to form a broad picture of the current practice.

7.3.3.1 Current Practice (Problem Scenario)

Helen is about to commence writing up her thesis. Over the course of the previous three years she has been involved in modelling a number of chamber experiments and field campaigns. Her early experiments have been written up as part of her first year and 18 month reports, the experiments are also informally documented in her laboratory notebook.

Even with access to these two information resources (and the memories of doing the work) it is a painstaking process to piece together the modelling process and re-interpret the model output for presentation in her final thesis. The reports provide a good overview of the modelling process and her laboratory notebook provides details on experiments and her scientific reasoning. The utility of the laboratory notebook is limited by the *ad hoc* structure of the information it contains making it difficult for Helen to reconcile it with other information sources. The laboratory notebook provides an incomplete record of the experiments; important comments and data locations are omitted (as they were not seen to be significant at the time of modelling), this leads to Helen repeating previous work in order to make results interpretable. Helen finds this duplication of effort unproductive and that it impacts on her motivation during a critical stage of her PhD.

Problem Scenario 2: Researcher reviews her own data and provenance records during the preparation of a publication.

7.3.3.2 Analysis of Current Practice

The analysis, of problem scenario 2, takes the form of considering a number of queries the researcher in the scenario, Helen, would like to ask when reviewing her data and associated provenance during the preparation of her PhD thesis. The queries were devised in conjunction with the members of the MCM user community. The ability of a sub-set of

the EXACT provenance documentation[13] (referred to as the cut-down EXACT provenance documentation), shown in Provenance Excerpt 5, to answer these queries was then tested. The results of these tests, successfully or otherwise, were then used to infer characteristics of the current approach to provenance.

benbox105w.fac (23/04/02)
Change initial product channel branching ratios in accordance with work of Volkamer et al.

```
G47 % 3.58D-12*EXP(-280/TEMP)*0.352  : BENZENE + OH = BZBIPERO2    +S47;
G48 % 3.58D-12*EXP(-280/TEMP)*0.118  : BENZENE+ OH = BZEPOXMUC+ HO2+S48;
G49 % 3.58D-12*EXP(-280/TEMP)*0.53   : BENZENE + OH = PHENOL + HO2 +S49;
G50 % 3.58D-12*EXP(-280/TEMP)*0.00   : BENZENE + OH = BZPERO2      +S50;
```

benbox106w.fac
correction to reaction 69

```
G69 % KRO2NO*0.082           : BZPERO2 + NO = NO2 + BZPERNO3       +S69;
```

Now

```
G69 % KRO2NO*0.082           : BZPERO2 + NO = BZPERNO3       +S69;
```

benbox107w.fac
phenol + NO_3 branching ratios from work by Wuppertal

```
G571 % 3.78D-12*0.742    : PHENOL + NO3 = C6H5O + HNO3            +S571;
G572 % 3.78D-12*0.0     : PHENOL + NO3 = PHENO2 + HNO3           +S572;
G573 % 3.78D-12*0.258    : PHENOL + NO3 = NPHENO2                 +S573;
```

benbox108w.fac
Updated inorganic chemistry and correction of photolysis rates as tolbox120w.fac

benbox109wA.fac
No reaction of maleic anhydride with NO_3.

benbox109wB.fac
test of butenedial + HO_2 reaction added.

benbox113w.fac
Increase photolysis rate of epoxide

```
G59 % J<4>*0.1*0.5     : BZEPOXMUC = MALDIAL + HO2 + CO + HO2 + CO +S59;
G60 % J<4>*0.1*0.5           : BZEPOXMUC = C5DIALO2 + HO2 + CO     +S60;
```

Provenance Excerpt 5: EXACT Provenance Documentation: Cut down case study.

Three query types are addressed: first, “history of” queries; secondly, “scientific object” queries; thirdly, “in silico experiment” queries.

[13] The text itself has not been edited, unnecessary content has just been omitted to form a concise provenance document.

"History of" queries

"History of" queries focus on understanding how something (e.g. the chemical mechanism) in model changed during the model development process. Three "history of" queries are presented below.

Query 1

Show me the history (changes plus annotations) of the mechanism in this series of experiments?

A partial answer this query can be obtained (see Provenance Query Result 1), with several pieces of information missing. Where reactions have been edited by the researcher, the previous state of the reaction has not been recorded in all cases, potentially because it is assumed that the previous state is as per the MCM (although this is not stated). In some cases (Change 6) the edit to the mechanism is not specifically recorded, an annotation just refers to an added reaction.

Query 2

Show me the history of reaction R7.1.

$$BZPERO_2 + NO \longrightarrow BZPERNO_3 \qquad (R7.1)$$

A full answer to this query can be obtained (see Provenance Query Result 2).

Query 3

Show me the history of all the photolysis reactions.

A partial answer this query can be obtained (see Provenance Query Result 3). Again the issue of a lack of the provenance regarding the initial state of the reactions is seen. Also the reactions themselves are not typed, the reaction type was retrieved from the annotation. Photolysis rates can also be identified as their rate coefficient including J<*n*>, where *n* is an integer. This query raises questions about how types are allocated to a given reaction including: Should the type be part of the annotation or modelled as a separate conceptual entity? How should types be determined, automatically, manually, or some appropriate combination of the two? Can reactions have multiple types, if multiple types are allowed which type take precedence?

Initial State

Initial mechanism: Not recorded

Change 1

Before editing: Not recorded

After Editing:

```
G47 % 3.58D-12*EXP(-280/TEMP)*0.352  : BENZENE + OH = BZBIPERO2    +S47;
G48 % 3.58D-12*EXP(-280/TEMP)*0.118  : BENZENE+ OH = BZEPOXMUC+ HO2+S48;
G49 % 3.58D-12*EXP(-280/TEMP)*0.53   : BENZENE + OH = PHENOL + HO2 +S49;
G50 % 3.58D-12*EXP(-280/TEMP)*0.00   : BENZENE + OH = BZPERO2      +S50;
```

Annotation: Change initial product channel branching ratios in accordance with work of Volkamer et al.

Change 2

Before editing:

```
G69 % KRO2NO*0.082           : BZPERO2 + NO = NO2 + BZPERNO3       +S69;
```

After Editing:

```
G69 % KRO2NO*0.082           : BZPERO2 + NO = BZPERNO3       +S69;
```

Annotation: correction to reaction 69

…

Change 6

Add reaction: not recorded

Annotation: test of butenedial + HO2 reaction added.

…

Provenance Query Result 1: Result for query 1 "show me the history (changes plus annotations) of the mechanism in this series of experiments?"

Change 1

Before editing:

```
G69 % KRO2NO*0.082           : BZPERO2 + NO = NO2 + BZPERNO3       +S69;
```

After Editing:

```
G69 % KRO2NO*0.082           : BZPERO2 + NO = BZPERNO3       +S69;
```

Annotation: correction to reaction 69.

Provenance Query Result 2: Result for query 2"show me the history of reaction $BZPERO_2 + NO \longrightarrow BZPERNO_3$"

Change 1

Before editing: Not recorded

After editing:

```
G59 % J<4>*0.1*0.5     : BZEPOXMUC = MALDIAL + HO2 + CO + HO2 + CO +S59;
G60 % J<4>*0.1*0.5           : BZEPOXMUC = C5DIALO2 + HO2 + CO     +S60;
```

Annotation: Increase photolysis rate of epoxide.

Provenance Query Result 3: Result for query 3 "Show me the history of all the photolysis reactions."

"Scientific object" queries

Scientific object queries focus on understanding the state of a scientific object (e.g. the set of constrained species). Two "scientific object" queries are presented below.

Query 4

Show me all the model-experiment comparisons for OH.

It is not possible to answer this query with reference to the cut-down EXACT provenance documentation, as model-experiment comparisons[14] are not addressed. Model-experiment comparisons were performed during the research, and stored in data analysis documents. No links to data analysis documents were provided in provenance documentation, so it provided difficult to retrieve the correct model-experiment comparisons.

Query 5

Show me all the NO_2 producing reactions that I added or edited.

An answer to this query can be obtained (see Provenance Query Result 4), it is possible that this answer is incomplete; if changes to the mechanism only described by annotations (e.g. benbox109wA.fac and benbox109wA.fac in Provenance Excerpt 5) relate to NO_2 producing reactions.

[14] A model-experiment comparison could be in the form of graph comparing model and experimental data or some statistical measure of the match between model and experimental data.

Change 1

Before editing:

```
G69 % KRO2NO*0.082          : BZPERO2 + NO = NO2 + BZPERNO3       +S69;
```

After Editing:

```
G69 % KRO2NO*0.082          : BZPERO2 + NO = BZPERNO3       +S69;
```

Annotation: correction to reaction 69.

Provenance Query Result 4: Result for query 5 "Show me all the NO2 producing reactions that I added or edited".

In silico experiment queries

In silico experiment queries focus on understanding the high level goals, conclusions etc. of the experiments that took place, three *in silico* experiment queries are presented below.

Query 6

What were the goals of this piece of research?

Query 7

What were the conclusions of this piece of research?

Query 8

Are there any related pieces of research? (Preceding, Follow-on, Branches, Dead ends)

The EXACT campaign provenance documentation, provides very limited provenance at an *in silico* experiment level. So it is not possible to answer any of the three queries presented above directly, but the queries can be answered indirectly with reference to the associated publications. *In silico e*xperiment provenance would be of great value when navigating and searching provenance records, but is poorly dealt with in the EXACT provenance documentation. I suggest that the poor quality of experimental level provenance is a result of the fact that current provenance records are typically manually captured and reviewed, so it makes no sense to invest effort in provenance to aid search-ability;

7.3.3.3 Provenance Characteristics

Based upon the queries relating to scenario 2, a set of query characteristics can be identified. Five query characteristics are presented below.

Query characteristic 1. Queries seek to understand human (as well as computational) processes

The queries presented in the analysis of scenario 2, refer to human activities within the model process (particularly mechanism development). This again highlights the importance of capturing human processes alongside computational processes when capturing provenance.

Query characteristic 2. Queries seek to uncover the reasoning behind the scientific process

In queries presented in the analysis of scenario 2, equal importance is often placed on annotations and process provenance. For example in queries 1, where the history of the mechanism is being sought, the history is considered to include both the process of how the mechanism evolved and also the associated scientific reasoning (i.e. annotations).

Query characteristic 3. Queries are made with respect to (at least) two frames of reference

Provenance documentation is queried with respect to two frames of reference in the analysis above: first, the scientific process is queried (in queries 1-8); and secondly, the *in-silico* experiment is queried (to answers queries 6-8).

Query characteristic 4. Queries are constructed using domain specific scientific terminology

Queries are constructed using domain specific scientific terminology (e.g. reaction, mechanism, 'photolysis'). In order to answer any of the queries 1-5 it is necessary for the semantics used within the provenance to include domain specific scientific terminology, in this case atmospheric chemistry modelling concepts.

Query characteristic 5. Queries about experimental level provenance are difficult to answer

Queries 5-8 demonstrates the lack of experimental level provenance with the provenance documentation. Some of this experiment level provenance can be obtained from publications (when available), but it difficult to reconcile the publication content with provenance documentation content.

Query Characteristic 6. Queries about data analysis processes are difficult to answer

Query 4 highlight the lack of provenance for data analysis processes, within the case study provenance documentation. This lack of provenance makes it difficult for a provenance user to identify and understand the impact of changes made to the model.

7.3.4 A Summary of Provenance Characteristics

In Table 7.1, presented below, the provenance and query characteristics from the two scenarios are aligned where possible to compile an aggregated list of the MCM modeller's approach to provenance. This aggregated set of characteristics is not intended to be exhaustive, but rather indicative of the diverse set of characteristics that exists; based upon my experiences of, and interactions with, the atmospheric chemistry community. The aggregated list of characteristics of the approach to provenance (adopted by researchers developing models using the MCM), will be revisited in the following chapter, where the implications of these characteristics for the design of the ELN will be addressed.

Scenario 1: Provenance Characteristics	**Scenario 2: Query Characteristics**	**Aggregated characteristics of the MCM modellers' approach to provenance**
1. Provenance is captured for human and computational processes.	1. Queries seek to understand human (as well as computational) process.	1. Provenance is captured for human and computational processes.
2. The reasoning behind the scientific process is important.	2. Queries seek to uncover the reasoning behind the scientific process.	2. The reasoning behind the scientific process is important.
3. Annotations are made with respect to two frames of reference.	3. Queries are made with respect to (at least) two frames of reference.	3. Annotations are made with respect to two frames of reference.
4. Scientist's use domain specific scientific terminology.	4. Queries are constructed using domain specific scientific terminology.	4. Scientist's use domain specific scientific terminology.
5. Provenance capture is interleaved with the scientific process.		5. Provenance capture is interleaved with the scientific process.
6. Provenance is captured and stored in multiple media.		6. Provenance is captured and stored in multiple media.
7. Provenance capture is a manual activity.		7. Provenance capture is a manual activity.
8. Provenance capture has a significant learning curve.		8. Provenance capture has a significant learning curve.
	5. Queries about experimental level provenance are difficult to answer.	9. Experimental level provenance is generally poorly addressed.
	6. Queries about data analysis processes are difficult to answer.	10. Provenance for data analysis processes is generally poorly addressed.

Table 7.1: This table lists, aligns and aggregates the characteristics (of the approach of a researcher developing models to provenance) identified during the analysis of two problem scenarios.

7.4 Task Analysis of a Model Development Process

Having considered the ELN stakeholders and mapped the current working practices of potential ELN users (at a high-level using scenarios), in the preceding sections of this chapter scenario 1, capturing data and provenance, is considered in greater depth. A task analysis, i.e. a detailed and structured representation of the activities and cognitive processes taking place when executing a task, is used to capture the detail of current working practices. I made decision to focus on the capture data and provenance (rather than provenance use), on a pragmatic basis; the first task the ELN must perform is to capture data and provenance, once this task has been understood it is then appropriate to consider provenance use. Further definitions and details regarding task analysis can be found in chapter 6, ELN development Methodology. The remainder of this section consists of three components: first, an introduction to the task analysis; secondly, the task analysis itself; and thirdly, a discussion of the implications of the task analysis.

7.4.1 An Introduction to the Task Analysis

This introduction address three questions: first, "what tasks were analysed"; secondly, "why develop a task analysis?"; and finally, "how was the task analysis developed". This introduction then concludes with some background information, before the next sub-section presents the task analysis itself.

"What tasks were analysed": In order to develop a detailed understanding of current working practices a task analysis, as described in chapter 6, was developed for a model development case study. The case study considers the development of a model of the SOAPEX field campaign [2] (discussed in Chapter 3). This particular piece of modelling was selected due to its relative simplicity and my familiarity with it; having studied the SOAPEX model in detail during the benchmarking of OSBM. The task analysis provides a description, at the finest granularity of task description possible, of a representative sub-set of the activities required to develop the SOAPEX model.

"Why develop a task analysis?": The task analysis of the SOAPEX model development case study was required in order to develop an in-depth understanding of the activities involved in developing a model using the MCM. This understanding will enable, as described in subsequent chapters, the ELN to be developed to fulfil its key aim; capturing

provenance that enables an *in silico* experiment to be completely understood or re-implemented.

"How was the task analysis developed?": During the development of the task analysis two resources were drawn on heavily: first and primarily, my personal experience of developing the SOAPEX model referred to in the case study; secondly, the input of members of the University of Leeds, Atmospheric Chemistry research group, who provided feedback in order to iteratively refine the task analysis.

7.4.2 SOAPEX Model Development Task Analysis

This sub-section presents the task analysis of the SOAPEX model development case study and consists of two core components: first, a description of the setting; secondly, a description of how tasks are performed within this setting. This task description separates the modelling process into three types of activity: first, model development, i.e. changing the configuration of the model; secondly, model execution, i.e. running the model on an appropriate computational resource; and thirdly, data analysis, i.e. interpreting the data produced by the last model run in conjunction with other data resources. Upon completing the data analysis process, a researcher will then return to the model development activity, forming an iterative loop over the three activities. Each activity in the task analysis is described using a standard format, immediately below.

1. **Process Type** (i.e. model development, model execution, data analysis)

   Process Description: A simple description of the process that took place.

   Process Metadata: A set of metadata, that describes the process that took place.

   Associated Lab-book entry:

   > The notes made in a lab-book, or word processing documentation, represent the type of notes that a researcher could be expected to make when executing the activity in question.

   Comment: My comments regarding the nature of the process and the associate lab-book entry.

7.4.2.1 Setting

Location: Research laboratory, University of Leeds, School of Chemistry.

Task owner: Chris Martin.

Tools used: OSBM, desktop computer, the MCM, Laboratory Notebook, Microsoft Word.

7.4.2.2 Task description (step-by-step)

1. **Model Development**

   Process description: Add-Intial-Mechanism

   Download an initial mechanism, to be used as input to the SOAPEX model, from the MCM.

   Process metadata:

   Primary VOCs selected: CH_4

   Extraction format selected: FACSMILE

   MCM version: 3.1

   Extraction date: 26/09/2008

   Extracted by: Chris Martin

   Number of chemical species: 29

   Number of chemical reactions: 70

   Associated laboratory notebook entry:

Add MCM v3.1 methane mechanism to establish baseline model.

2. **Model Execution**

   Process description: Run model on desktop machine, using FACSIMILE

   Process metadata:

   Model executed: C:\\run\SOAPEX_129

   Execution location: CHMIBM719

   Model runtime: 30 seconds

   Model output location: C:\\run\SOAPEX_129\run1

   Associated laboratory notebook entry: none.

   Comments: Each time the model runs, model output is placed in the same directory (overwriting the output of previous runs). If a researcher wants to retain

the model output for a given model run then his or she stores the model output within an *ad hoc* file structure.

3. Data analysis

Process description: Analyse output of the initial model, plot a set of concentration graphs using Microsoft Excel.

Process Metadata:

Species concentrations plotted: OH, HO_2

Data sources compared: Model output (from step 2) and experimental data from the BADC field campaign database.

Associated laboratory notebook entry:

> Baseline model established, modelled/measured radical concentrations in the same order of magnitude.

Comments: With current methods of model output archiving (i.e. outputting to the same directory and copying to an ad hoc personal file system) it is difficult to identify the data sources being compared. A description of the experimental data origin and pre-processing is not addressed in the laboratory notebook entry.

4. Model development

Process description: Add two reactions to mechanism, to characterise elements of chemistry taking place during the night.

Process metadata:

Reaction added: %2.50d-22*H2O : N_2O_5 = HNO_3 + HNO_3**;**

Reaction added: %1.80d-39*H2O*H2O: N_2O_5 = HNO_3 + HNO_3**;**

Lab-book Entry:

> Add night-time N_2O_5 reactions.
>
> ```
> %2.50d-22*H2O : N2O5 = HNO3 + HNO3;
> %1.80d-39*H2O*H2O: N2O5 = HNO3 + HNO3;
> ```

Comments: Potential additional annotation could include the source of each reaction.

5. Model execution

Process description: Run model on desktop machine, using FACSMILE

Process metadata:

Model executed: C:\\run\SOAPEX_129

Execution location: CHMIBM719

Model runtime: 30 seconds

Model output location: C:\\run\SOAPEX_129\run1

Lab-book entry: none.

6. Data analysis

Process description: Analysis output of model (from step 5), plot a set of concentration graphs using Microsoft Excel.

Process metadata:

Species Concentrations Plotted: OH, HO_2

Data Sources Compared: Model output from steps 2 and 5

Associated laboratory notebook entry:

> Little difference in radical concentrations as result of addition of N_2O_5 reactions.

Comments: The data sources compared and the location of the analysis spreadsheet are not recorded in laboratory notebook.

7. Model development

Process description: Edit O^1D quenching reaction

Process metadata:

Reaction before editing: % 1.80d-11*N2*exp(107/TEMP) : O^1D = O ;

Reaction after editing: % 2.10d-11*N2*exp(115/TEMP) : O^1D = O ;

Associated laboratory notebook entry:

> Update the mechanism to reflect the latest available experimental data, including the redetermination of the rate coefficient for the reaction of $O(^1D)$ with N_2, Ravishankara et al., 2002 [9].

Comments: The laboratory notebook entry refers to the change made but does not provide a complete description of the change that took place (i.e. the updated rate coefficient).

.

8. **Model execution**

Process description: Run model on desktop machine, using FACSIMILE

Process metadata:

Model executed: C:\\run\SOAPEX_129

Execution location: CHMIBM719

Model runtime: 30 seconds

Model output location: C:\\run\SOAPEX_129\run1

Annotations: none.

9. **Data analysis**

Process description: Analysis output of model (from step 8), plot a set of concentration graphs using Microsoft Excel.

Process metadata:

Species concentrations plotted: OH, HO_2

Data sources compared: Model output from steps 2, 5 and 8.

Associated laboratory notebook entry:

> The effect of the new rate coefficient is to decrease the OH concentration by approx. 10% and HO2 by approx. 2%. Reduction brings modelled concentrations closer to measurements.

Comments: Again the data sources used in the analysis are not recorded.

7.4.3 Discussion of Task Analysis

The task analysis has identified a number of key processes (within current working practices) and the metadata required to describe these processes. Understanding of these processes is drawn upon during the design of the ELN, as described in the next chapter of this thesis. The task analysis itself has not been comprehensive. For example the task analysis considers only the main mode of model development (developing the mechanism). There are in fact many other modes of model development including: developing the constraint set; editing model parameters (e.g. model start and end time); etc.. Also, model output data can be analysed in a number of ways (beyond a simple comparison of species concentrations), including: analysing rates of production and loss

for a given species; mechanism visualisation; and sensitivity analysis. I made the decision to restrict the scope of the task analysis, and so the development of the ELN, in order to ensure that the scope of the research remained manageable. Further discussion of this scoping decision can be found in the next chapter.

Chapter Summary

This chapter has presented an analysis of the current working practices of researchers developing computational models, using the MCM. Particular attention has been paid to the working practices associated with provenance capture and use. This analysis has employed three techniques: stakeholder analysis; problem scenarios; and task analysis; to ensure an understanding of the problem domain from an abstract to a concrete level. Two key outputs from this chapter will be carried forward to the next chapter: first, the set of provenance characteristics, identified during the analysis of the problem scenarios; and secondly the SOAPEX case study task analysis. The provenance characteristics are analysed in the next chapter, to generate a high-level design statement (that describes the design principles adopted in the development of the ELN). The SOAPEX case study provides the detailed understanding of *in silico* experimental processes and provenance capture required to design the user-ELN interactions and the information structure for the provenance captured by the ELN.

References

1. Bloss, C., et al., *Evaluation of detailed aromatic mechanisms (MCMv3 and MCMv3.1) against environmental chamber data.* Atmos. Chem. Phys., 2005. **5**(3): p. 623-639.
2. Sommariva, R., et al., *OH and HO2 chemistry in clean marine air during SOAPEX-2.* Atmos. Chem. Phys., 2004. **4**(3): p. 839-856.
3. Natural Environment Research Council. *NERC Data Policy Handbook*. 2002 [cited 12th March 2009]; Available from: http://www.nerc.ac.uk/research/sites/data/documents/datahandbook.pdf.
4. *Project Prospect*. 2008 [cited 8th December 2008]; Available from: http://www.rsc.org/Publishing/Journals/ProjectProspect/index.asp.
5. Blomberg, J., M. Burrell, and G. Guest, *An ethnographic approach to design*, in *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications*. 2003, L. Erlbaum Associates Inc. p. 964-986.
6. Schroeder, R. and J. Fry, *Social Science Approaches to e-Science: Framing an Agenda*. Journal of Computer-Mediated Communication, 2007. **12**: p. 563-582.

7. Scott, S. and W. Venters, *The Practice of e-Science and e-Social Science*, in *Virtuality and Virtualization*. 2007. p. 267-279.
8. Volkamer, R., et al., *OH-initiated oxidation of benzenePart I. Phenol formation under atmospheric conditions.* Physical Chemistry Chemical Physics, 2002. **4**(9): p. 1598-1610.
9. Ravishankara, A.R., et al., *Redetermination of the rate coefficient for the reaction of $O(^1D)$ with N_2.* Geophysical Research Letters, 2002. **29**(15): p. 35-1.

Chapter 8 Design of the ELN

This chapter takes the understanding of current working practices, captured in the previous chapter, and uses them to inform the design of the ELN. The first section outlines the design approach that guided the ELN prototype development. The high-level ELN design is then presented from two perspectives; first, from a user perspective, envisioned working practices (i.e. capturing data and provenance using the ELN) are described; secondly, from a system perspective, the system architecture of the ELN is presented. This chapter then progresses to describe the detail of the ELN design: first, the interaction design (i.e. how the user interacts with the ELN); and secondly, the information design (i.e. how the provenance captured by the ELN is structured.

8.1 Implications of the MCM Modellers' Approach to Provenance

The section consists of two components: first, the characteristics of the MCM modellers' approach to provenance (identified in the previous chapter) are analysed to determine their implications for the design of the ELN; and secondly, based upon these implications, the high level design approach adopted during ELN development is presented.

8.1.1 Analysis of Provenance Characteristics

1. Provenance is captured for human and computational processes

In current working practices modellers describe their scientific processes in terms of the human processes (e.g. adding reactions to a mechanism) and computational processes (e.g. running a model). By capturing provenance in these terms, two benefits are realised: first, a complete, unified provenance record can be captured (for a given *in silico* experiment); secondly, the provenance for human processes provides valuable information about the nature of the research taking place; so it is desirable to retain this characteristic in the provenance capture by the ELN.

2. The scientific reasoning behind the scientific process is important

In current working practices modellers record both their scientific reasoning and their scientific process in provenance documentation, either as process-reasoning pairs[15] or individually. These working practices highlight the importance of being able to answer a pair of questions when seeking to understand a given dataset: first, "how was the dataset produced?"; secondly, "why was the dataset produced that way". Being able to answer both these questions, using one source of provenance documentation, would be beneficial to a provenance user, so in the design of the ELN equal importance will be placed upon capturing the scientific process and the associated scientific reasoning.

3. Provenance is recorded with respect to two frames of reference

In current working practices, provenance is recorded with reference to two frames of reference: first, the scientific process being executed; and secondly, at a higher conceptual level, the *in silico* experiment (being implemented by the scientific process). Capturing provenance with respect to the detailed scientific process allows a given dataset to be interpreted and understood. Capturing provenance with respect to the *in silico* experiment allows the high-level goals of the scientific process to be understood. Capturing provenance with respect to these two frames of reference is complementary, and the ELN will be designed to retain this characteristic.

4. Scientists use domain-specific scientific terminology

When recording provenance researchers make use of domain-specific terminology, such as "adding a reaction to a mechanism". The use of this set of terminology has significant benefits, as domain-specific terminology contains a great deal of informational content (for the individual conversant with the terminology in question); so the ELN will be designed to retain this characteristic.

5. Provenance capture is interleaved with the scientific process

Provenance capture does not take place as an isolated activity, it is interleaved within the scientific process (i.e. developing a model using the MCM). This characteristic enables a researcher to record provenance as the need occurs; reducing the likelihood of provenance being captured at some point after the scientific process takes place, which could

[15] e.g. "I added reaction X to mechanism (process) because the findings of paper Y suggest this will improve the performance of the mechanism (reasoning)."

potentially lead to a low quality of provenance due to difficulties recalling the exact nature of the process. Again this characteristic can be seen to be beneficial to provenance users, so will be incorporated in the design of the ELN.

6. Provenance is captured and stored in multiple media

In current working practices modellers capture and store provenance and data in a number of media, including: lab-books, word processor documents, file names, annotation in data analysis spreadsheets etc. This fragmentation makes provenance records difficult to "piece together" and interpret (particularly for anyone other than the original creator of the provenance). So the ELN will seek to capture a single provenance record for a given scientific process, where this is not possible due to feasibility constraints (e.g. resources/time available) then the ELN design will seek to minimise fragmentation of the provenance record.

7. Provenance capture is a manual activity

Current provenance capture practices are entirely manual, this makes provenance capture a time consuming activity. The time consuming nature of provenance capture is a contributory factor to the incomplete nature of provenance records (discussed under characteristic 9), so the ELN will be designed to automate provenance capture where possible (this is particularly appropriate for process provenance) in order to reduce the effort required by the researcher to maintain provenance records. By automating process provenance capture, the ELN will allow the researcher to focus on recording their scientific reasoning.

8. Provenance capture has a significant learning curve

Capturing provenance is an activity that a researcher learns to do over a period of time, often through experiencing the consequences of failing to capture provenance. The ELN will be designed to minimise the learning curve of both the ELN (as a tool) and the provenance capture process.

9. Provenance capture is generally incomplete

Here two characteristics from the original list, of the characteristics of the modellers approach to provenance, are aggregated under one heading (provenance capture is generally incomplete). The characteristics aggregated are: experimental level provenance

and provenance for data analysis processes. A general discussion of the implications of incomplete provenance capture is presented next, followed by a discussion of the implications of each of the aggregated characteristics.

General discussion. Much of the provenance captured by MCM model developers focuses on the changes made to the chemical mechanism; this is understandable as the key goal of model development is to generate insight in to the chemistry taking place. Provenance for mechanism development is still often incomplete, with other aspects of the provenance poorly addressed or even completely neglected. The consequence of incomplete provenance are limitations in terms of the ability of the provenance to aid interpretation of the associated dataset; so the ELN will be designed with a focus on capturing complete provenance for mechanism development.

Experimental level provenance is generally poorly addressed. Characteristic 3, above, identifies that provenance is captured with reference to the *in silico* experiment taking place; however often this provenance is absent or incomplete and of poor quality. Provenance with respect to the *in silico* experiment (due to its high level nature) will play an important role in enabling archives of provenance documentation to be queried and navigated. So the ELN will be designed to enable high quality provenance to be captured with respect to the *in silico* experiment.

Provenance for data analysis processes is generally poorly addressed. A key factor in the typically low quality of provenance for data analysis processes is the *ad hoc*, manual, unstructured nature of the data analysis processes in question. The ELN design will seek to overcome this issue to enable high quality provenance to be captured for analysis processes.

8.1.2 Design Approach Overview

This sub-section provides an overview of the design approach adopted during the development of the ELN and breaks down into a three components: first, a re-iteration of the high-level goal of the ELN (as introduced in Chapter 5); secondly, a statement of the scope of the ELN development; and thirdly, a list of the design principles employed during the ELN development.

Goal

The high level goal of the ELN development is to:

> *Enable the capture of provenance for the process of developing models using the MCM, whilst minimising the burden on ELN users. The provenance captured should be, where possible, complete.*

Here, a complete provenance is defined as the provenance required to re-implement a given model, recreate a given dataset and understand why the model was implemented in a given way.

Design scope

The design scope, which constrained the development of the ELN to a specific problem space, was informed by the development resources available, and the perceived resource requirements of the development tasks. The description of the design scope is presented in terms of three scoping statements below.

Retain existing model development processes and tools. The ELN will integrate with existing model development processes and tools. This scoping statement was adopted for two reasons: first, limiting the changes to the working practices of the researcher, so increasing the chances of the ELN being adopted across the community; and secondly, embedding the ELN within real working practices (as analysed in Chapter 7).

Address the capture of provenance for model development. Development of ELN functionality was limited to addressing Scenario 1, capturing data and provenance for the model development process. So scenarios 2, which refers to querying provenance and data archives remain outside the scope of this design chapter. This constraint focuses the research on the issues of provenance representation and capture.

Address the most frequently occurring modes of core activities. The computational modelling process consists of three core activities: Model Development; Model Execution; Data Analysis. For each of these core activities there are a number of possible modes. Taking model development for example, modes could include editing the chemical mechanism; constraining datasets; model start and end time; mathematical parameters that

control the ODE solver. Given this diversity, I selected a single mode for each core activity to explore during the development of the ELN, to act as an exemplar of the wider set of modes. So for model development activity, mechanism development was selected; for model execution, model execution on a local machine was selected; and for data analysis, the comparison of concentration data from various data sources using Microsoft Excel was selected. These modes were selected as they are the most frequently use modes occurring in current working practices.

Design principles

The final component of the design approach is a set of principles employed during the design of the ELN. These principles are distilled from the analysis of the characteristics of current working practices (see Section 8.1.1). So the high-level design goal (described above) will be achieved, within the specified design scope (described above), by adopting the following design principles.

- Provenance will be, where possible, captured automatically;
- Provenance captured will be represented, stored and queried using the terminology of the atmospheric chemistry domain;
- Provenance will be captured for both human and computational processes;
- Equal importance will be placed on capturing process provenance (generally automatically) and scientific reasoning (requiring the ELN user to make annotations recording their scientific reasoning);
- Provenance will be captured with respect to two frames of reference: first, the scientific process; secondly, the *in silico* experiment;
- Provenance capture will be interleaved with the modelling process (referred to as "inline provenance capture" in the remainder of this thesis);
- The learning curve for the ELN (as a tool) and provenance capture (as a process) will be minimised.

Throughout the remainder of this design chapter where design decisions are discussed, adherence to this set of design principles will be highlighted.

8.2 Envisioned Working Practices

The section presents envisioned working practices of a researcher developing computational models using the MCM and an ELN. As stated in the design approach (see

section 8.1.2), the ELN design is restricted to addressing the capture of provenance for mechanism development processes. The envisioned working practices are presented in the form of an activity design scenario, as described previously in chapter 6. Following on from the activity design scenario, the key design decisions committed to in this scenario are highlighted. The purpose of this section is to give a high-level description of how the ELN will be used to capture provenance, providing a foundation for the in-depth description of the interaction and information design that follows.

8.2.1 Activity Design Scenario

The activity design scenario, presented below, describes envisioned working practices for capturing data and provenance using the ELN and corresponds to problem scenario 1 (which describes current working practice for capturing data and provenance, see Chapter 7).

Helen is developing models of a set of chamber experiments; the models are developed iteratively. A modelling iteration involves mechanism development, running the model (on a local machine), and analysing the model output (by comparing graphs of various species concentrations with experimental data). The goal of her piece of modelling research is to obtain a good agreement between model output and experimental measurements and derive some insight into the chemical mechanism.

The modelling process conducted by Helen is, where possible, captured automatically by the ELN. For example changes to the chemical mechanism are automatically captured and model output is automatically archived in a database and referenced in the ELN. Where it is not possible for the modelling process to be captured automatically, a standard interface is presented to enable Helen to quickly record the relevant metadata, for example when comparing data sources using spreadsheet/graphing software. As the ELN captures the experimental process, Helen is prompted to make annotations ensuring that she considers and records the scientific reasoning associated with her modelling processes. After completing a number of model development iterations, Helen uses the ELN to make some high-level notes about the overall goals and findings of her modelling research, and links her current experiment to other related work.

Activity Design Scenario 1: envisioned working practices for capturing data and provenance using the ELN

8.2.2 Key Design Decisions

The activity design scenario commits to three key design decisions; in this sub-section the nature of, and rationale for, these decisions is described.

Automatic archiving of model output data

Current working practices rely on *ad hoc* data archiving solutions, which leads to difficulties linking data to associated provenance documentation. In order to address this issue model output will be automatically archived by the ELN, as described in the activity design scenario "model output is automatically archived in a database and referenced by the ELN". Automatic archiving of model output data aligns with the overall ELN design

goal; *Enable the capture of provenance for the process of developing models using the MCM, whilst minimising the burden on ELN users.*

Prompting the user to record their scientific rationale

The activity design scenario states "As the ELN captures the experimental process, Helen is prompted to make annotations ensuring that she considers and records the scientific reasoning associated with her modelling processes". The use of prompts, based upon on the actions of the ELN user, seeks to encourage ELN users to record their scientific reasoning (alongside the automatically capture process provenance). Further discussion on the use of prompting and alternative methods of capturing scientific reasoning can be found in Section 8.4. Prompting the user to record their scientific rationale aligns with two design principles: *equal importance will be placed on capturing process provenance* (generally automatically) *and scientific reasoning* (requiring the ELN user to make annotations); secondly, *provenance will be captured inline within the scientific process*.

Light-touch provenance capture for data analysis

The activity design scenario states "Where it is not possible for the modelling process to be captured automatically, a standard interface is presented to enable Helen to quickly record the relevant metadata, for example when comparing data sources using spreadsheet/graphing software". This light-touch approach to capturing provenance for data analysis processes was adopted, despite contradictions with efforts to capture process provenance automatically in a complete form, due to the difficulty of automatically capturing provenance for *ad hoc* data analysis using proprietary software (such Microsoft Excel). A light touch approach to the capture of provenance for data analysis aligns with the design scoping statement: *retain existing model development processes and tools*. Rather than change existing working practice to make provenance capture easier (e.g. automating standard data analysis processes), the ELN has been designed to integrate with existing *ad hoc* data analysis processes.

8.3 System Architecture

In the third section of this chapter an overview of the ELN system architecture is presented, see Figure 8.1. The architecture consists of five components, the purpose of each of these architectural components are described in detail below.

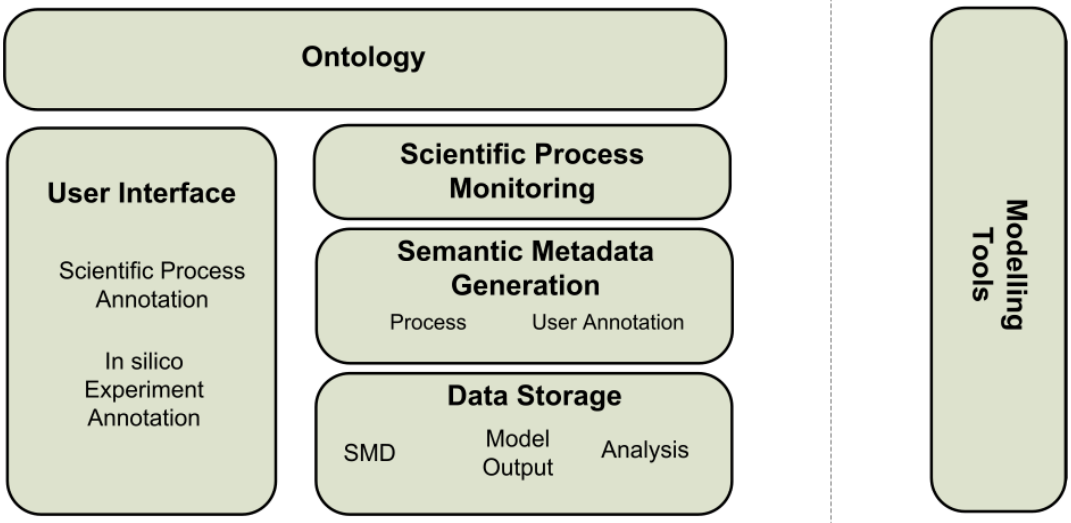


Figure 8.1: The ELN system architecture. This figure presents the ELN architecture, consisting of five components: the ontology, providing structure for the provenance captured; the user interface, providing the modeller with functionality to record annotations; scientific process monitor, automatically capturing process provenance; SMD generation, converting the provenance captured by the ELN into a semantic metadata representation; and data storage.

Ontology

The ontology provides a control vocabulary of terms and relationships used to structure the provenance captured by the ELN. This vocabulary consists of terminology from the atmospheric chemistry modelling community (i.e. it is domain specific). The ontology itself is used by the SMD generation component, when converting provenance captured by the ELN in to a semantic metadata representation.

User interface

The user interface provides the ELN user with two perspectives: scientific process; and *in silico* experiment. The scientific process interface, discussed in detail in Section 8.4, enables the ELN user to capture their scientific rationale inline with their scientific process, when prompted. The *in silico* experiment perspective enables the ELN users to view their scientific processes at a more abstract level (i.e. as *in silico* experiments) and annotate them according (i.e. describe the goals and conclusions of a set of scientific

processes). The provenance captured by the user interface is passed to the SMD generation component.

Scientific process monitoring

This component of the architecture interacts with the modelling tools used to develop and execute models, in order to capture data from, and provenance for, the scientific process being executed. The data captured is submitted to the model output database and the provenance captured is passed to the semantic metadata generation layer.

Semantic metadata generation

The SMD generation component takes provenance from the user interface (i.e. scientific rationale and *in silico* experiment annotations) and the scientific process monitoring component (i.e. automatically captured process provenance). It then combines the provenance, from these two sources, generating a provenance representation in form of SMD conforming to the ontology. The resulting SMD is then passed to the data storage component for archival.

Data Storage

Provides database storage for: model output data; semantic metadata representations of the provenance; and data analysis documents.

8.4 Interaction Design

This section considers the design of the interaction patterns (between the ELN and ELN user) and it consists of two cases: first, the interaction pattern for the capture of scientific process provenance is addressed; secondly, the interaction pattern for the capture of *in silico* experiment provenance is addressed.

8.4.1 Capture of Provenance with Respect to the Scientific Process

This sub-section addresses the interaction pattern (between the ELN and user), for the capture of scientific process provenance and consists of two components: first, a description of the general interaction pattern and some justification adopting it; secondly, a sub-set of the SOAPEX model development case study (introduced in Chapter 7) is

revisited, and an interaction specification is presented for a set of specific actions by the ELN user.

8.4.1.1 Interaction Approach

The section describes the general interaction pattern (between the ELN and an MCM modeller), for the capture of scientific process provenance. The general interaction pattern adopted is that the researcher performs some action (using the available modelling tools), the ELN responds to this action by prompting the researcher to record their scientific rationale for the action in question. The interaction specification, shown below, demonstrates this general interaction pattern.

10. **User performs a generic action (as part of their model development process)**

    User action: Perform action A.

    ELN action: Display context dependent prompt (prompt interrupts the modelling process, i.e. the modelling process can not continue until the user has addressed the prompt) see Figure 8.2.

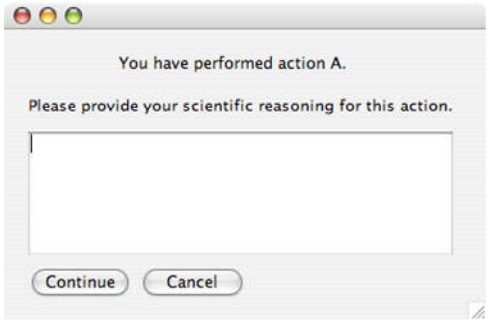


Figure 8.2: Generic ELN prompt. The ELN prompt interrupts the scientific process, to encourage the researcher record their scientific reasoning. The prompt is driven by the action performed by the researcher.

User action: Provide scientific reasoning and click continue.

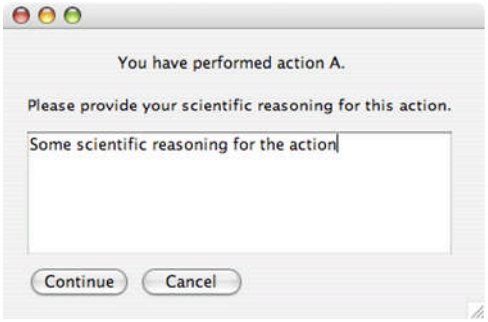


Figure 8.3: Completed generic ELN prompt. As Figure 8.2, but with scientific reasoning provided by the ELN user: "Some scientific reasoning for the action".

ELN action: Returns user to appropriate modelling process.

Why use prompts for inline capture of scientific reasoning?

The rationale for adopting this general approach consists of three components.

- First, the use of prompts is intended to encourage researchers to capture their scientific reasoning as (or just after) it takes place. The prompt provides a visual cue to remind the researcher to consider and record their scientific reasoning, conveying the message that; capturing provenance (particularly scientific reasoning) is a core part of the scientific process (not an optional extra if time is available).
- Secondly, the alternative approach (to the use of prompts) is to allow users to record scientific reasoning at their discretion. Current practices, centred about the laboratory notebook, allow complete discretion in terms of what provenance should and should not be captured, and can be seen to lead to incomplete or absent provenance records.
- Thirdly, the use of prompts supports current working practices, in terms of enabling inline provenance capture.

The drawbacks of using prompts for inline capture of scientific reasoning

Having considered the rationale for adopting prompts as a means of encouraging researchers to record their scientific rationale it is appropriate to consider the potential drawbacks of this approach. Two key drawbacks are identified below.

- First, the prompts may be ignored (i.e. the user just clicks continue without considering making an annotation). This case may occur when a researcher has

been using the ELN for an extended period of time, and becomes "immune" to prompts. Although the potential for prompts to be ignored is concerning, if prompts are ignored the researcher is deciding to take a more discretionary approach to recording their scientific reasoning (which in some cases will not be possible to avoid).

- Secondly, and of greater concern, the user may view the prompts as interrupting their real work (i.e. generating scientific understanding). If this is the case, then the researcher is unlikely to adopt or use the ELN.

8.4.1.2 SOAPEX Case Study

The preceding sub-section considered the generic approach to interaction, for the capture of scientific reasoning associated with a given scientific process. This sub-section moves on to provide a detailed interaction specification for a specific scientific process. The process in question was introduced as the SOAPEX case study, in Chapter 7. A subset of the SOAPEX case study is considered in the interaction specification, steps 4 – 8, where: two reactions are added to a mechanism; the model is executed; the output data is analysed; and finally a single reaction (within the mechanism) is edited. The sub-set considers one full modelling iteration, and part of a second iteration to provide a flavour of the experience of using the ELN. The second iteration is only considered in part because the interactions with the ELN associated with model execution and data analysis processes are similar across any given set of model development iterations.

Interaction specification

4. **Model development**

   User action: Add two reactions to the chemical mechanism (by editing the model's mechanism input text file, using a text editor of choice), to characterise elements of chemistry taking place during the night. The user commits the changes to the mechanism by running the model (using the command line).

   ELN action: Interrupt model execution. Display prompt for scientific rationale, see Figure 8.4.

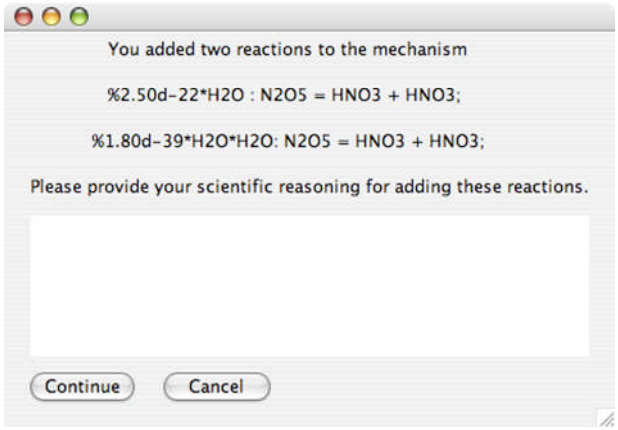


Figure 8.4: ELN prompt, generated by the modeller adding two reactions to the chemical mechanism.

User action:

Complete prompt text box: "Add night-time N_2O_5 reactions".

Click: Continue.

ELN action: Close prompt window, return user to model execution interface (i.e. the terminal).

5. **Model execution**

Model run completes

ELN action: Prompt the user to record any comments about the model execution (e.g. slow model run due to other jobs running on the machine).

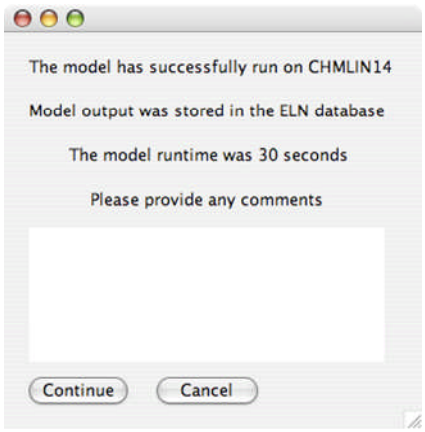


Figure 8.5: ELN prompt, generated upon completion of a model run, requesting any comments on the model execution. The prompt highlights the success (in this case) or otherwise of the model execution.

User action:

Complete prompt text box: "n/a".

Click: Continue.

ELN action: Prompt for user to record any provenance regarding the data analysis process. Lock model (i.e. prevent model from running until analysis interface has been completed).

6. Data analysis

User action: Complete data analysis processes comparing the concentrations for OH and HO_2 (for model output data and field experiment data), using data analysis software, not integrated with the ELN (e.g. Microsoft Excel).

User action: Complete data analysis prompt, see Figure 8.6.

Add data source information 1:

Data location: http://badc.nerc.ac.uk/browse/badc/soapex/data/in-situ/as990118.ho2

Select data type: Field Experiment

Data description: Experimental HO_2 Data

Add data source information 2:

Click: Add New Data Source

Data location: http://badc.nerc.ac.uk/browse/badc/soapex/data/in-situ/as990118.oh

Select data type: Field Experiment

Data description: Experimental OH Data

Add data source information 3:

Click: Add New Data Source

Click: Browse (model output database displayed in browser)

Select: Latest model run output

Data location and data type: Automatically populate

Data description: Data from latest model run

Add comments on the data analysis process:

Complete text field: As expected little difference in radical concentrations as result of addition of N_2O_5 reactions.

Click: Save comments

Attach data analysis documentation:

Click: browse (file browser for local machine presented)

Select: Appropriate data analysis document

Click: save

Click: Continue

ELN Action: Close data analysis prompt, release lock on model.

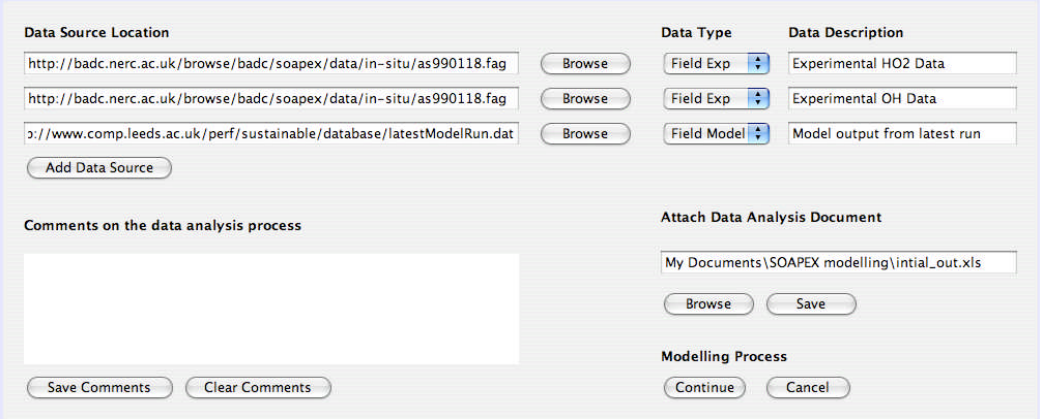


Figure 8.6: ELN prompt to capture provenance for data analysis performed by the modeller, generated following a success model execution. The prompt enables the modeller to record the data sources used in the analysis, add any comments about the analysis process and store any associated data analysis documents.

7. Model development

User action: Edit O^1D quenching reaction and run the model.

from: % 1.80d-11*N2*exp(107/TEMP) : O^1D = O ;

to: % 2.10d-11*N2*exp(115/TEMP) : O^1D = O ;

ELN Action: Interrupt model execution, display prompt for scientific rationale.

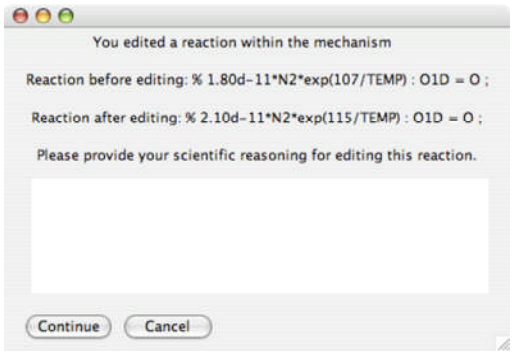


Figure 8.7: ELN prompt, generated by the modeller editing a reaction within the chemical mechanism. The reaction pre and post editing is displayed and the associated scientific reasoning for editing the reaction is requested.

User action:

Complete prompt text box: "Update the mechanism to reflect the latest available experimental data, including the redetermination of the rate coefficient for the reaction of $O(^1D)$ with N_2, Ravishankara et al., 2002."

Click: Continue

Key design decisions

Within the interaction specification presented two key design decisions are committed to. Each design decision is addressed in turn below.

Design decision 1: Where possible the annotation interface is structured in a minimal fashion. The prompts for model development and model execution present a single text field for the ELN user to complete as they see fit. This simplicity and flexibility mimics current working practice, where researchers can record provenance unhindered by any enforced structure. This design decision has the two drawbacks: first, that no specific details (e.g. references to the literature) are required within a given annotation, so may be omitted or incomplete; and secondly, the lack of structure makes it difficult to search for specific information (e.g. a given reference) within a set of annotations, due to the lack of a standard representation.

Design Decision 2: Process provenance for data analysis activities is not captured automatically, but relies on the ELN user completing data analysis prompt. Whilst it

would be desirable to automatically capture process provenance in this case, it not feasible due to *ad hoc* data analysis processes that make use of complex, proprietary software (such as MS Excel). This design decision has the drawback of reducing the quality of provenance captured for data analysis processes.

This sub-section has presented the interaction approach (between the ELN and ELN user) taken for the capture of provenance for the computational modelling process executed by a researcher developing a model using the MCM. Both the general approach, using prompts to capture the researcher's scientific rationale inline (with their scientific process), and an interaction specification for a model development case study have been addressed.

8.4.2 Capturing Provenance with Respect to an *In Silico* Experiment

In this sub-section the interaction patterns (between the ELN and ELN user) are addressed for the capture of provenance about *in silico* experiments themselves. These interaction patterns are considered from two perspectives: first, the general approach to provenance capture; and secondly, the interfaces designed to support this general approach.

In contrast to the interaction patterns for the capture of provenance with respect to the scientific process, described in the preceding sub-section, the interactions (between the ELN user and the ELN) in this sub-section are described in a rather vague fashion. For example when considering the scientific process, a precise interaction specification is presented; whereas when considering *in silico* experiments, a set of ELN interfaces are presented. The reason for this difference is that the modelling process itself, and the capture of associated provenance, is well understood based upon the analysis of current practice. Whereas, because provenance captured with respect *in silico* experiments is addressed poorly or not addressed at all in current practices, how researchers should interact with the ELN to capture this provenance is more difficult to define. Developing ELN functionality to capture provenance with respect *in silico* experiments is likely further development iterations.

The interface design presented in this sub-section was initially informed by my personal experiences using and capturing provenance during the development of the OSBM.

Drawing on these experiences I developed a series of paper prototype interface designs, which facilitated a discussion with members of the University of Leeds, Atmospheric Chemistry Modelling Research Group about their requirements for capturing provenance with respect to *in silico* experiments. Based upon these discussions interface prototypes were developed and further feedback was sought, and the interface prototypes were refined (to the state seen in this sub-section).

8.4.2.1 General interaction approach

Provenance captured with respect *in silico* experiments is composed of the researchers' high-level thoughts and reasoning about a given experiment, and so can only be captured by engaging the ELN user (i.e. this type of provenance can not be captured automatically). The ELN allows provenance with respect to *in silico* experiments to be captured in two ways: *pre hoc* annotation, i.e. the researcher define their goals, experimental method, etc. prior to executing a given scientific process; and *post hoc* annotation, i.e. the researcher executes a scientific process, and then defines the goals, experimental method, etc. For any given *in silico* experiment a combination of both *pre hoc* and *post hoc* annotation may take place.

8.4.2.2 Interface design

When designing the ELN interface for the capture of provenance with repect to an *in silico* experiment, the goal was to provide a flexible interface, that retains enough structure to capture useful provenance. The ELN user is free to navigate the interface as they see fit, the user can flick between different elements of the interface using the (always visible) navigation panel on the left of the interface (see Figure 8.8). All fields are optional and the provenance entered in the interface is saved when clicking on the "Save Experiment Description" button. Having completed pre-hoc annotation the ELN user can then begin executing their scientific process. This subsection considers the interface design in the context of the two modes of annotation: *pre hoc* and *post hoc*.

Pre-hoc provenance capture with respect to an *in silico* experiment

When making annotations prior to executing an *in silico* experiment two particular interfaces are likely to be used by a modeller: first, the basic information interface (see Figure 8.8); and secondly, the experimental method interface (see Figure 8.9).

Basic information interface

This basic information interface (see Figure 8.8) enables a researcher to record a high-level description of the *in silico* experiment they are conducting. The interface consists of seven components, each of which is described in detail below.

Experiment name: A free text field allowing a researcher to define a name for the experiment. When the provenance is saved to the ELN database, the experiment name will be checked against existing experiment names, to avoid duplication.

Experiment description: An unrestricted text field enables the ELN user to record a description of the experiment they plan to conduct. In Figure 8.8 an example description is provided, for the SOAPEX relate *in silico* experiment discussed in Chapters 4 of this thesis.

Experiment type: A drop down box enables the ELN user to select the type of *in silico* experiment they are conducting from a defined list of alternatives. For experiment types not included in the defined list, the user is able to define their own custom experiment type.

Species of interest: A text box is provided to record the chemical species the *in silico* experiment focuses on. If the ELN user uses MCM names (the custom names for species used in the MCM) identify the species in the provenance documentation.

Tags / Keywords: A text field enables keywords associated with the experiment to be recorded by the researcher. This allows for a web 2.0 type approach to be adopted whereby tags can be shared, allowing the researcher to define the terms of their scientific discourse in a flexible manner, with minimal overheads.

Experiment owner: The owner of the experiment, can be recorded in a the text box.

Associated researchers: Any other researchers associated (e.g. supervisors of the experiment owner) with the *in silico* experiment can be entered in these text boxes.

Experimental method interface

The experimental method interface (see Figure 8.9) enables the ELN user to record a high level description of the method they will adopt when performing an *in silico* experiment. When conducting pre-hoc annotation, the experimental method can be used to record a high-level plan of action. The interface consists of two elements, each described below.

Model used: This text box allows the ELN user to record the core model they used during his or her *in silico* experiment. In Figure 8.9 a URL for a specific version of the OSBM (pointing to an online svn repository) has been entered into the text box.

Experimental Method: This text field allows the ELN user to record the high level activities executed during an *in silico* experiment. In Figure 8.9 an outline plan is presented for the SOAPEX model development discussed in Chapter 4 of this thesis.

Comments: Again a minimal approach has been adopted to structuring the provenance capture interface of the ELN, with two fields provided to the researcher. Adding more structure within the experimental method interface would be beneficial from the perspective of increasing the informational content of the provenance and facilitating machine processing of the provenance. Potential additional structure includes: separating the experimental method into individual steps; typing each of these steps in the experimental method (i.e. recording whether a given step is a mechanism development, a data processing activity, etc.); the functionality to link each of these steps to components of the experimental process level provenance.

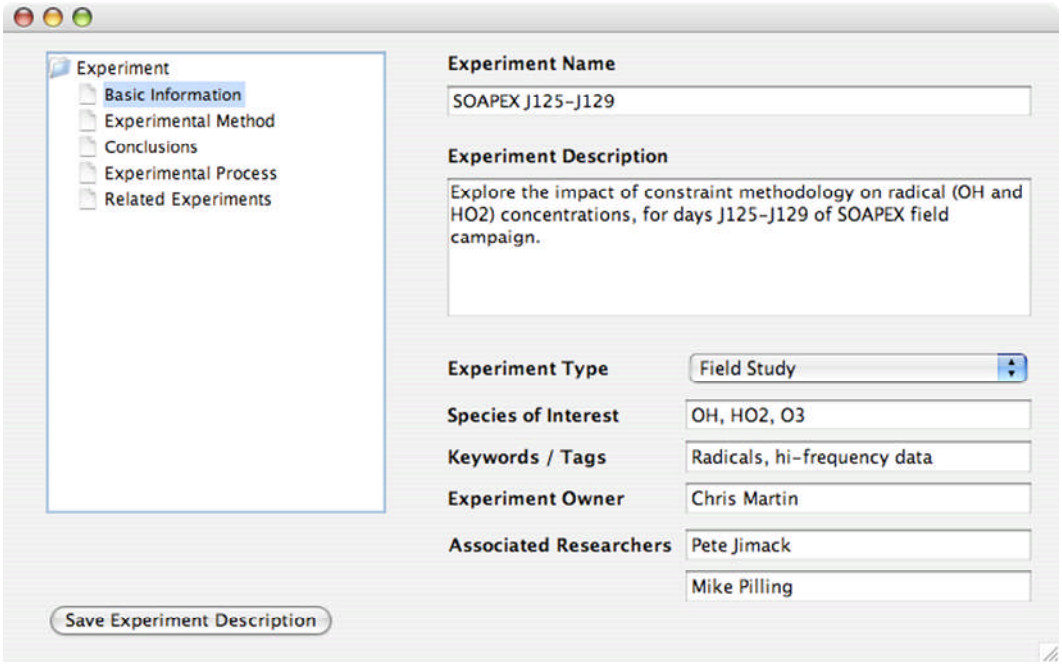


Figure 8.8: The ELN interface for the capture of basic information about an *in silico* experiment. This interface can be used before an experiment takes place (to enable a high level plan to be recorded) or at the later stages of model development (to enable the emerging nature of the experiment to be recorded). This interface enables provenance to be recorded including: the experiment name; a text description of the experiment; and the type of experiment.

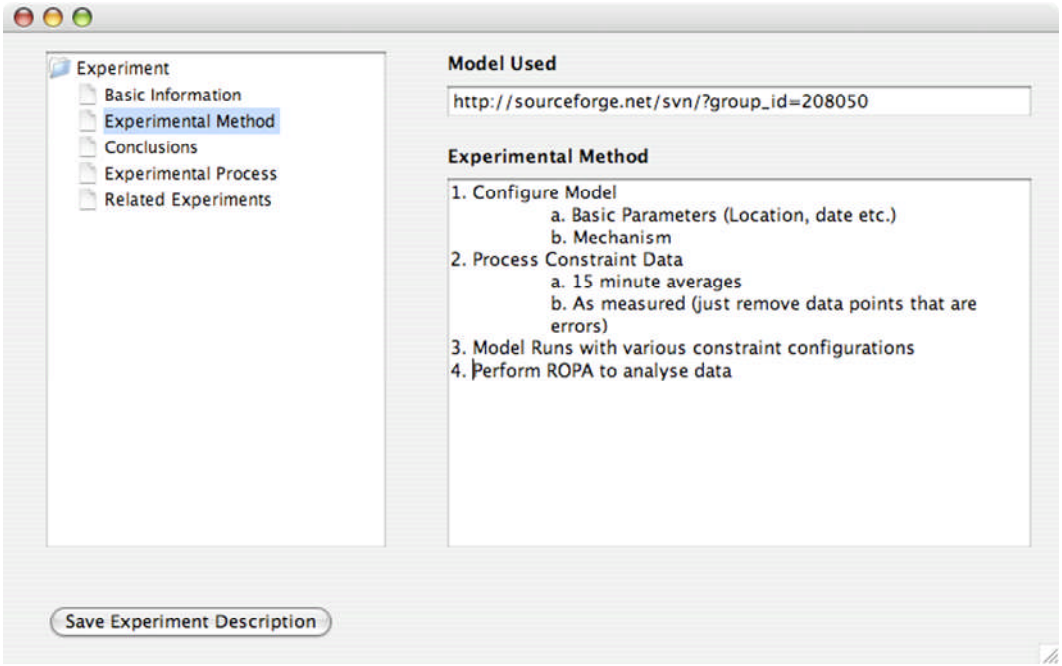


Figure 8.9: The ELN interface for the capture of the experimental method for a given *in silico* experiment. This interface allows the modeller to record the model they used (in the figure populated with a link to an online svn repository). An unrestricted text field is also provided to enable the modeller to record the experimental method they plan to use or have already executed.

Post-hoc provenance capture with respect to an *in silico* experiment

The preceding discussion considered the capture of provenance, prior to commencing an *in silico* experiment. The discussion now progresses to address how an ELN user would capture provenance about an *in silico* experiment once a scientific process has been executed. It is important to note that the *in silico* experiment does not have to be complete for post-hoc annotation to take place, just some element of the scientific process must have been executed (i.e. the *in silico* experiment is in progress).

It is anticipated that the ELN interfaces considered with respect to *pre hoc* annotation would also be used for *post hoc* annotation. For example half way through an experiment the experimental goals may become clear enough to record a high-level description of the experiment (using the basic information interface, see Figure 8.8). For the sake of brevity the ELN interfaces already described above will not be revisited. Two interfaces are particularly likely to be used for post-hoc annotation: first, the conclusion interface (see

Figure 8.10); secondly, the related experiments interface (see Figure 8.11), each of these interfaces is described below.

Conclusion interface

The conclusion interface (see Figure 8.10) enables the ELN user to record the insight they have generated over the course of an *in silico* experiment. The conclusion interface consists of two elements:

- *Conclusions:* In this text field the ELN user can record, in free format text, the conclusions of their *in silico* experiment.
- *Future Plans:* The ELN user can also record their plans for future research separately from their conclusions. The reason for drawing this distinction is to prompt the researcher to consider their future research plans and to make interpretation of the provenance easier (by adding structure).

Related experiments interface

The related experiment interface (see Figure 8.11) allows the ELN user to link their current *in silico* experiment to other experiments and to describe the relationship (if they wish).

- *Related research list:* The related research list presents research projects linked to the current *in silico* experiment. Items can be added to this list by click the 'add' button, the user can then browse their own ELN archive (and potentially the archives of other researchers) to select related *in silico* experiments.
- *Relationship:* For the currently selected experiment (in the related research list) a text field is provided to describe the relationship between the current *in silico* experiment and the related research. In the case shown in Figure 8.11 the related research is a model of an earlier part of the field campaign.

This section has described the design of the interactions between the ELN and the ELN user, for the capture of provenance with respect to both the scientific process and the *in silico* experiment. The decisions made when designing the interaction patterns, outlined above, were informed by the design approach outlined in Section 8.1.2.

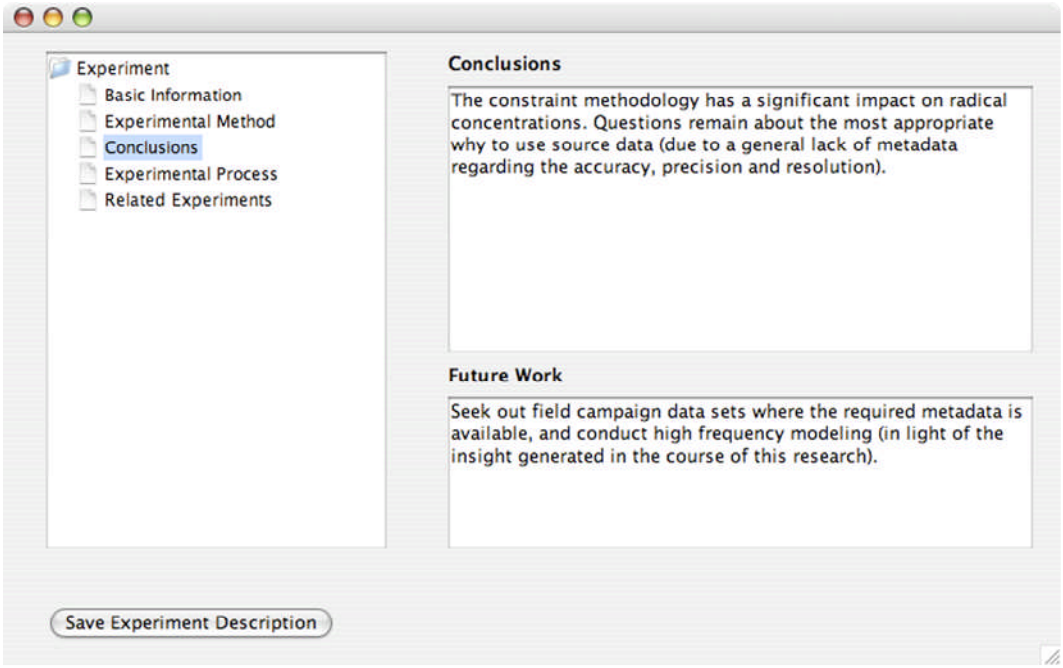


Figure 8.10: ELN conclusions interface. This interface allows the modeller to record their conclusions, following completion of an experiment. Two separate input fields are provides, for conclusions and future plans.

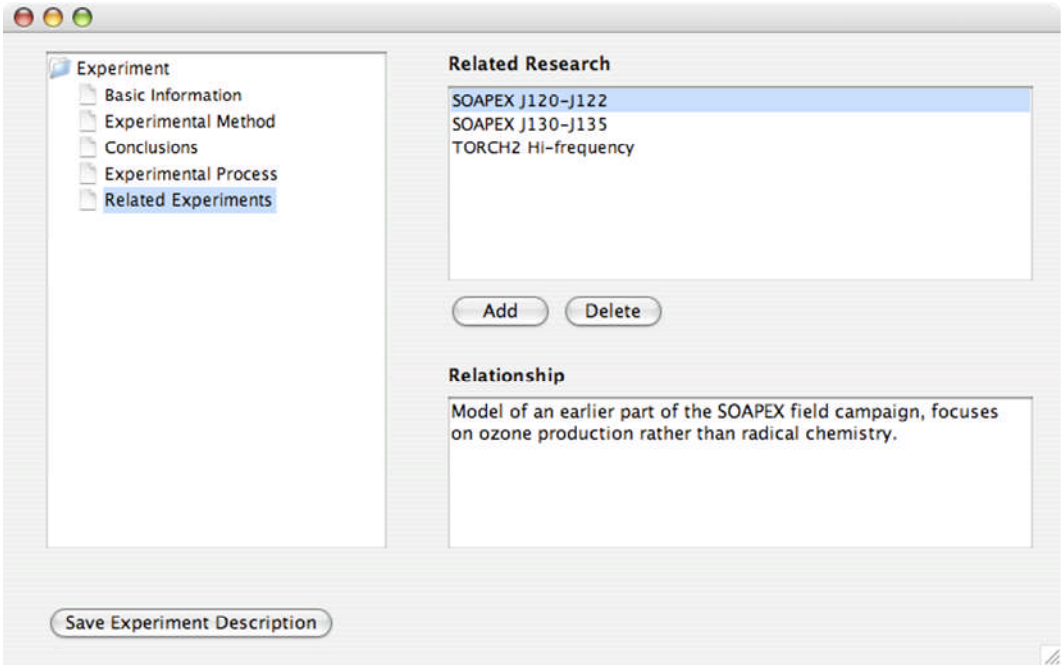


Figure 8.11: ELN related experiment interface. This interface allows the modeller to link their current experiment to other related experiments. These links are created by click the 'add' button (a an ELN archive browser is then presented) and selecting an experiment. The modeller can also annotate the relationship between two experiments with a text description.

8.5 Information Design

In this section the information design for the ELN is described, addressing the representation of the provenance captured by the ELN. The provenance captured by the ELN is represented using a defined vocabulary of terms, i.e. an ontology, that is referred to throughout. This section consists of three sub-sections: first, a conceptual model of the computational modelling process is presented, a distillation of the understanding developed from the analysis of current working practices; secondly, the representation of the provenance with respect to the scientific processes is addressed; and thirdly, the representation of the provenance with respect to *in silico* experiments is addressed. Prior to these sub-sections important elements of the terminology used in this section are described.

Terminology

Model: In this section the terms "model" and "modelling" are used in two contexts: first, as in a conceptual model, an abstract representation of the workings of a system; and secondly, as in a computational model, a computational realisation of a mathematical description of a scientific system. In order to avoid any confusion of terminology, in this section, the former will be referred to as conceptual modelling and the later will be referred to as computational modelling.

Ontology: An ontology is a defined vocabulary of terminology, composed of concepts and the relationships between these concepts. Discussion in this section describes elements of an ontology designed to structure the provenance captured by the ELN.

8.5.1 A Conceptual Model of the Computational Modelling Process

The purpose the conceptual model of the computational modelling process is to provide the core structure for the provenance captured by the ELN. This conceptual model consists of two components: first, a worldview, that describes at the highest level the world in which computational modelling takes place; and secondly, a three layer conceptual model, that describes the computational modelling process.

8.5.1.1 Worldview

When developing the ontology used to structure provenance captured by the ELN, I adopted the view that the restricted domain of computational model development using the MCM could be described in terms of three core concepts: materials, processes and people. These core concepts also form the basis of the CombeChem ontology [1] for describing *in vitro* organic chemistry experiments. Each of the core concepts is discussed below.

- **People:** the researchers involved in *in silico* experiments.
- **Materials:** The materials (either physical or conceptual) that are involved throughout an *in silico* experiment. For example the *in silico* experiment itself, the chemical mechanism, and a reaction within a chemical mechanism are all considered conceptual materials (although they are not tangible materials, in the sense of a given sample of a species).

- **Processes:** The processes that consume materials, transforming them into a different state. For example the mechanism development process transforms a chemical mechanism from one state to another.

These three core concepts are sub-classes of the top-level class. All other elements of the ELN ontology are sub-classes of the three core concepts. A notation (as used in the Comb*e*Chem project [2]) is adopted in the representation of *in silico* experiments, in this chapter, where processes are represented using grey circles and materials are represented with white circles.

8.5.1.2 A Three Layer Conceptual Model

The 3-layer conceptual model (see Figure 8.12) presents a hierarchical decomposition of the computational modelling process adopted by researchers using the MCM. Each layer of the conceptual model is described in detail below.

Experimental Layer

At the highest level model development is viewed as an *in silico* experiment. In the top layer of the 3-layer conceptual model, see Figure 8.12, the experiment can be seen to take a high-level modelling plan as an input and produce a conclusion as an output.

Iteration Layer

At a less abstract level model development is viewed as a network of modelling iterations. An iteration of the modelling process can be considered to take a plan, such as to test the effect of editing a reaction within the mechanism to update the rate coefficient to the latest literature value; and produce a conclusion/plan, such as editing the reaction had no significant effect on model output, now proceed to update the next reaction. So it can be seen that the output of an iteration, the conclusion/plan, is able to form the input to another iteration. The iteration layer shown in Figure 8.12 shows a linear series of three such modelling iterations linked by shared conclusions/plans.

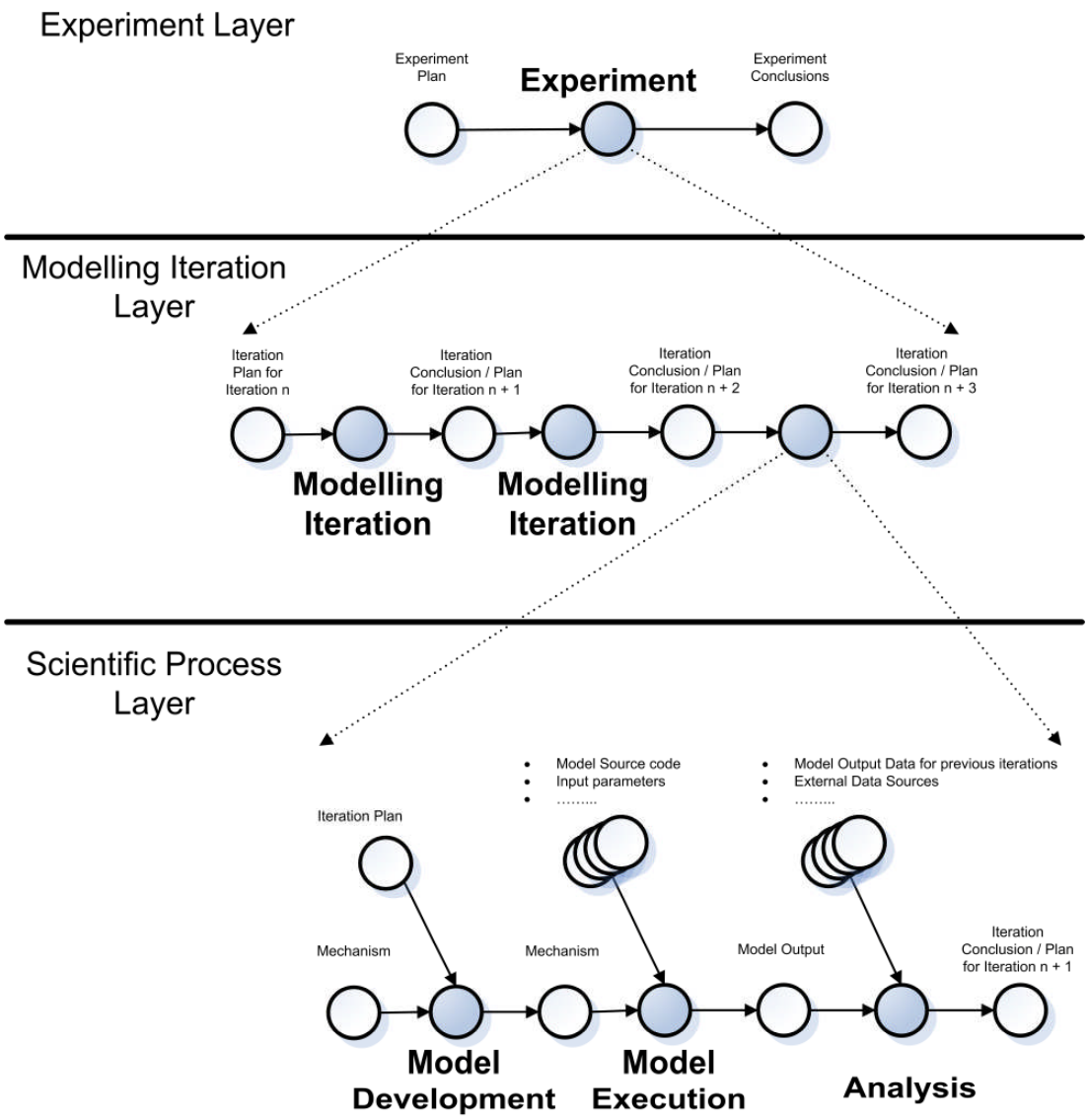


Figure 8.12: A three layer conceptual model of the computational modelling process. This figure shows the conceptual model used to structure the provenance captured using the ELN. The conceptual model consist of three layers: at the highest level the scientific process of a modeller is viewed as an experiment; at a more concrete level, the scientific process is viewed as a series of modelling iterations; and at the lowest level the scientific process is viewed in terms of the modellers actions.

Scientific process layer

At a concrete level model development can be viewed as a network of modelling processes (‘model development’, ‘model execution’, ‘data analysis’). In Figure 8.12 the

simplest case is presented: the model parameters are changed ('model development'); the model is run ('model execution'); and the model output is analysed to determine the impact of the parameter change and the fit with experimental data ('data analysis'). The Model Development processes takes an iteration plan (as discussed above) and some set of model parameters as an input, and produces a revised set of model parameters as an output. The Model Execution process takes the revised set of model parameters and the model source code[16] as inputs and produces a set of model outputs. The analysis process takes model output and other data sources (i.e. data from previous model runs or other external data repositories) as an input and produces an iteration conclusion/plan, as an output. There is clearly scope for more complicated networks of modelling processes, for example multiple analysis processes following a model execution. These more complicated networks of modelling process are addressed in Section 8.5.2.3.

Having described the core elements of the ontology for representing process provenance captured by the ELN, in the form of the three-layer conceptual model (presented above), this section continues to discuss in detail two layers of this conceptual model. The two layers discussed further are the scientific process layer and the *in silico* experiment layer. The iteration layer is not discussed further as the provenance associated with this layer, is simply a sub-set of the scientific process layer.

8.5.2 Representation of the Scientific Process

This sub-section describes the representation of provenance at the scientific process level, i.e. process provenance plus scientific reasoning in the form of annotations. There are five components of this section: first, a description of the use of terminology from the atmospheric chemistry domain; secondly, a description of the core of the experiment representation, the process-material spine; thirdly, a description of the scientific processes that could compose a model development iteration; fourthly, the way in which annotations are attached to the scientific process; finally, the representation of the SOAPEX case study provenance is presented to provide a concrete example.

[16] I have assumed that versioning of model source code is managed separately by software version control software.

8.5.2.1 Using Domain Specific Terminology

The three high-level processes shown in Figure 8.12 can be seen as system-orientated concepts for capturing a computational modelling workflow using domain independent concepts. I developed the ontology further through a lower conceptual level to incorporate scientific terminology from the atmospheric chemistry domain. Taking the decomposition of the "model development" process as an example, the modeller can perform a wide variety of operations on the mechanism, see Figure 8.13, including adding, deleting and editing reactions. The ontology also includes the decomposition of the 'edit reaction' process ('edit reactants', 'edit products', 'edit rate coefficient').

8.5.2.2 The Process-Material Experiment Spine

At the core of the provenance captured by the ELN is the spine of the experimental process, this is composed of material-process pairs [2]. In Figure 8.14 a simple experiment is shown where a modeller: adds a reaction (to a mechanism); runs the model; and compares the model output with other data from other sources.

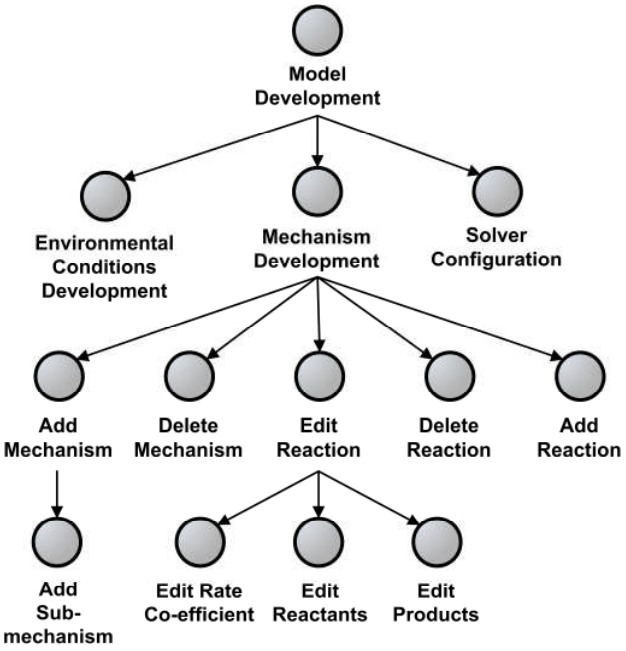


Figure 8.13: Domain-specific terminology for the "model development" process, an example from provenance captured by the prototype ELN. The figure provides a hierarchical decomposition of the model development process, considering developing the

chemical mechanism and editing a reaction within the chemical mechanism as exemplar processes.

The first process in the experiment is “add reaction”, which takes three inputs: a chemical mechanism, a new reaction and some conceptual plan that guides the modelling process. The output of the “add reaction” process is an updated mechanism, which in turn is an input for the second process “model execution”. The model execution process has other inputs, including various other input parameters and the model source code. The “model execution” process outputs a model output dataset, which in turn is analysed in the “compare data sources” process. The “compare data sources” process takes other inputs, such as datasets from other sources (including appropriate *in-silico* and *in-vitro* experiments) and outputs some conclusion about how adding the reaction affected the model’s behaviour. So it can be seen that a spine of alternating materials and processes, (mechanism, “add reaction”, mechanism, “model execution”, model output dataset, “compare data sources”) exists at the core of the provenance representation.

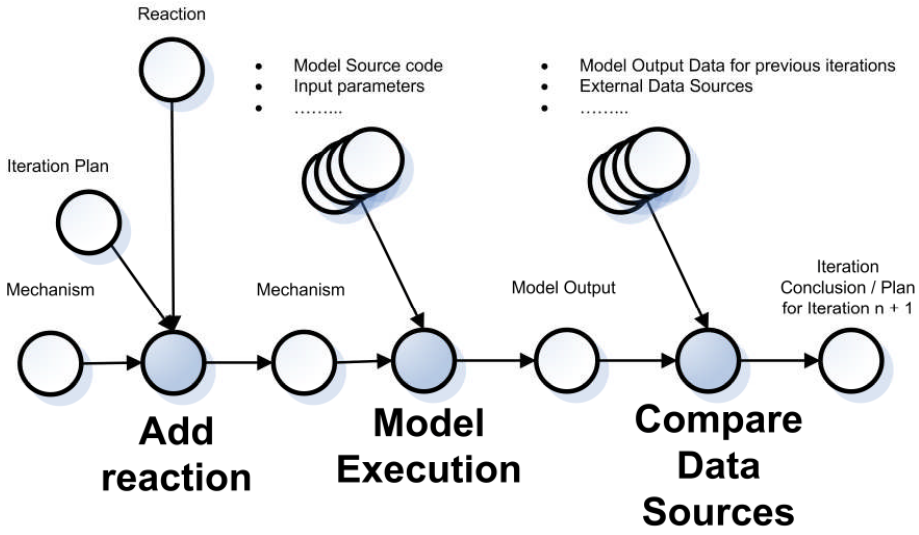


Figure 8.14: The material-process spine. This figure shows the alternating pairs of materials and processes, which form the core of the representation of the scientific process. The simplest case is presented where a modeller: changes the mechanism (add reaction); runs the model (model execution); and then performs some data analysis (compare data sources).

8.5.2.3 Possible Iteration Decompositions

So far only a very simple form of model development iteration has been considered. This form is the ideal case, where one model development process is followed by one model execution process, followed by one data analysis process. Clearly there are number of alternative forms, dependent on the actions of the user, two of these forms are discussed below.

- **Multiple data analysis processes.** A modelling iteration including multiple analysis processes is shown in Figure 8.15. A scenario that could give rise to this workflow, is the user performing two separate analysis processes, say a comparison of concentrations and a rate of production and loss analysis.
- **Multiple runs of the model.** A modelling iteration including multiple mechanism development and model execution processes is shown in Figure 8.16. An example of a scenario that would lead to this modelling iteration is a box model failing due to numerical instabilities in the ODE system. The numerical instabilities are caused by an error, introduced during the first mechanism development process; the user fixes this error and successfully re-runs the model.

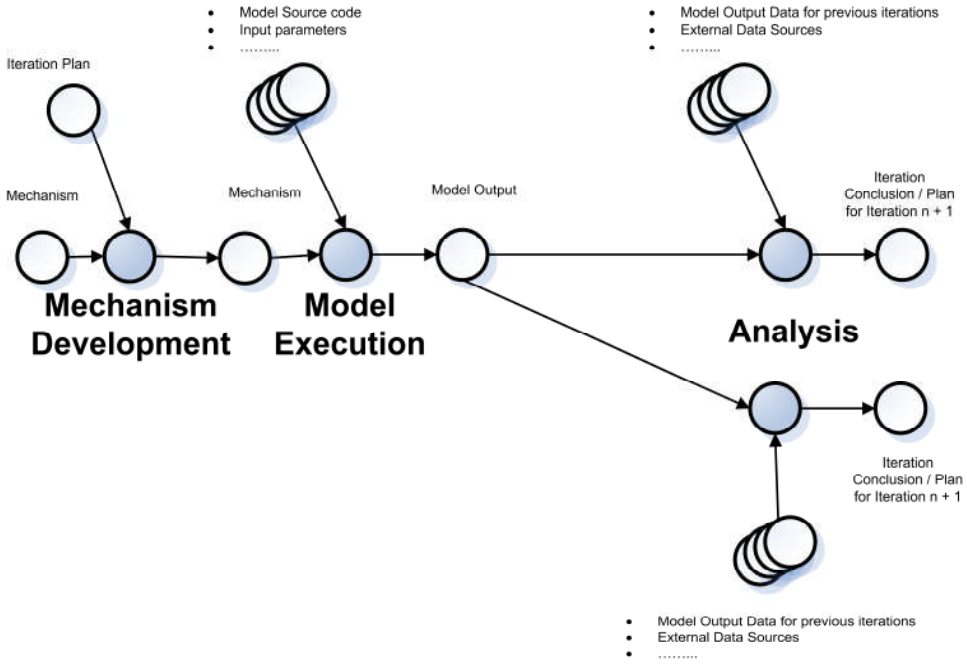


Figure 8.15: A modelling iteration including two data analysis processes. This figure presents a modelling process where a modeller analyses the model output data in two different ways.

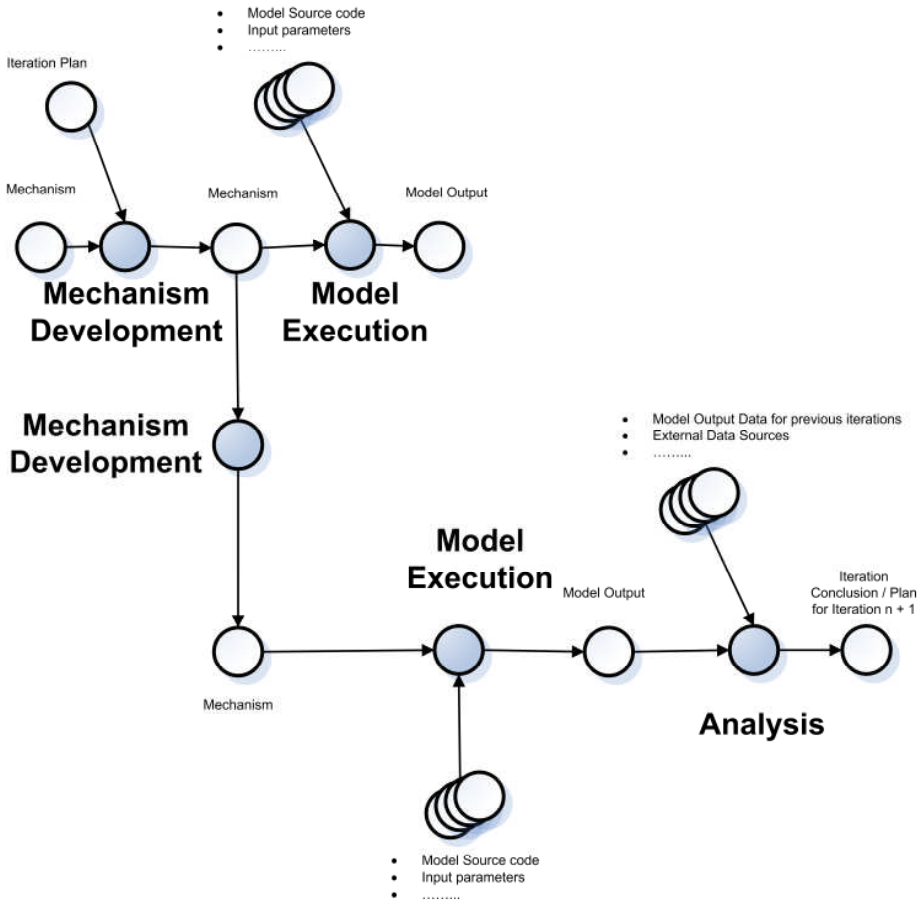


Figure 8.16: Modelling iteration including two mechanism development and two model execution processes. This figure presents a modelling process where a modeller edits the chemical mechanism and runs the model; the model run fails due to an error in the chemical mechanism (introduced by the latest edit). The modeller returns to correct the error and then successfully runs the model. The modelling process concludes with analysis of the model output data (generated by the 2nd model run).

8.5.2.4 Linking Annotations to the Experimental Process

This sub-section describes how annotations are linked to the material-process spine of the scientific process. Figure 8.17 shows the ontology used for representing annotations made by the scientist to record their scientific reasoning. The ontology used has many similarities with the Annotea [3] ontology, the W3C OWL ontology for annotation. The annotation has a property 'annotates' which links the annotation to its subject, in our case a material or process within the experimental workflow. The 'has body text' property

captures the text comments made by the scientist in the form of a simple xml string. The 'has-related-resources' allows the richer annotations to be attached to an object, for example image, audio or video files related to the experiment. Each of these richer annotations themselves can be the subject of further annotations, if they require additional text explanations.

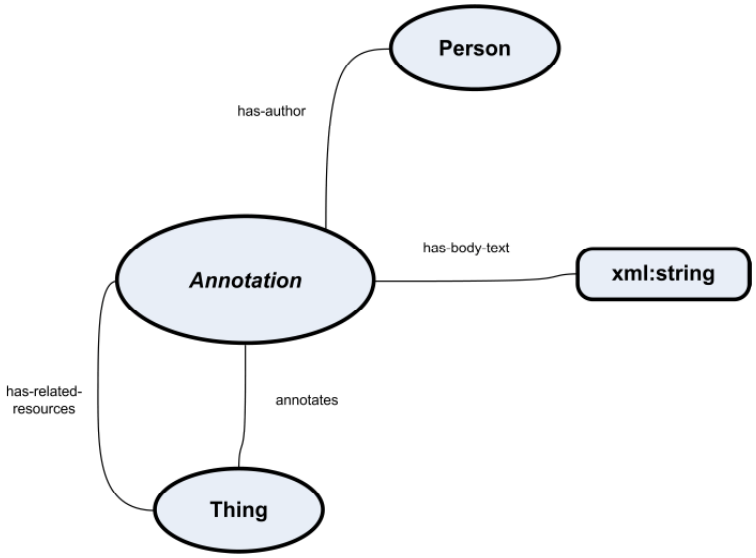


Figure 8.17: Annotation ontology. This figure presents the ontology used to represent annotations. Thing is the parent of all other concepts in the ELN ontology, so anything with the ontology can be annotated (material, process or person).

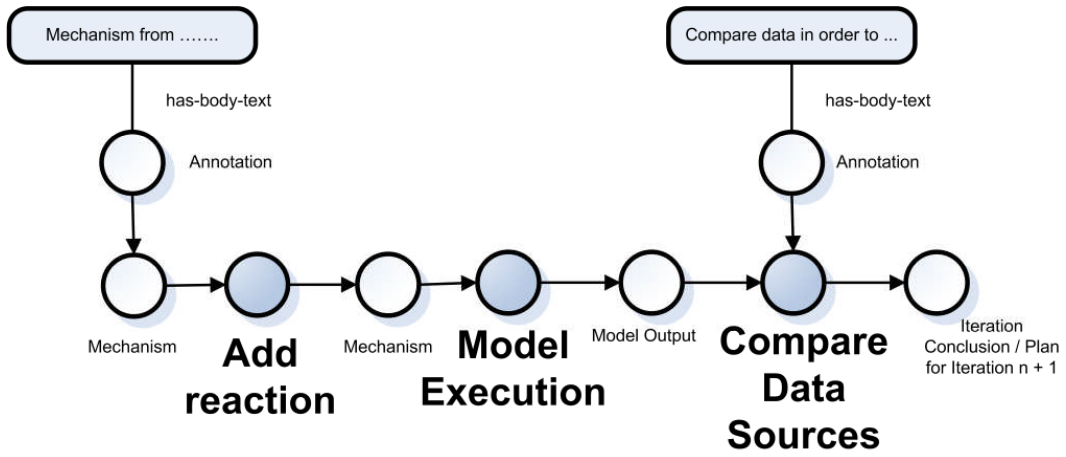


Figure 8.18: Attaching annotations to the scientific process. This figure show the annotates made to materials and processes, within a sample scientific process.

In Figure 8.18 annotations are made to the spine for a simple experimental process. Annotations are made in the same way to both materials, i.e. the initial mechanism, and processes, e.g. the ‘compare data sources process’. In the cases considered within my research only text annotation functionality is implemented, leaving richer annotation as a subject of further work.

8.5.2.5 Representing the SOAPEX Case Study

This sub-section presents an overview of how provenance captured by the ELN for the SOAPEX case study model development process is represented. The provenance representation is depicted in Figure 8.19 and Figure 8.20, and discussed in detail in the later stages of this sub-section.

Representing the SOAPEX case study

In Figure 8.19 (steps 1-6) and Figure 8.20 (steps 7-9) the SOAPEX case study model development process is represented in terms of the ELN ontology. The experimental process is represented in two dimensions: the vertical axis shows the progression of a given model development iteration (from mechanism development, through model execution, to data analysis); the horizontal axis represents progression of model development from one iteration to the next. Annotations in the diagram are presented in a simplified form, attached directly to processes or materials (rather than including the full annotation linkage, as show in Figure 8.18), purely to maintain the clarity of the diagram.

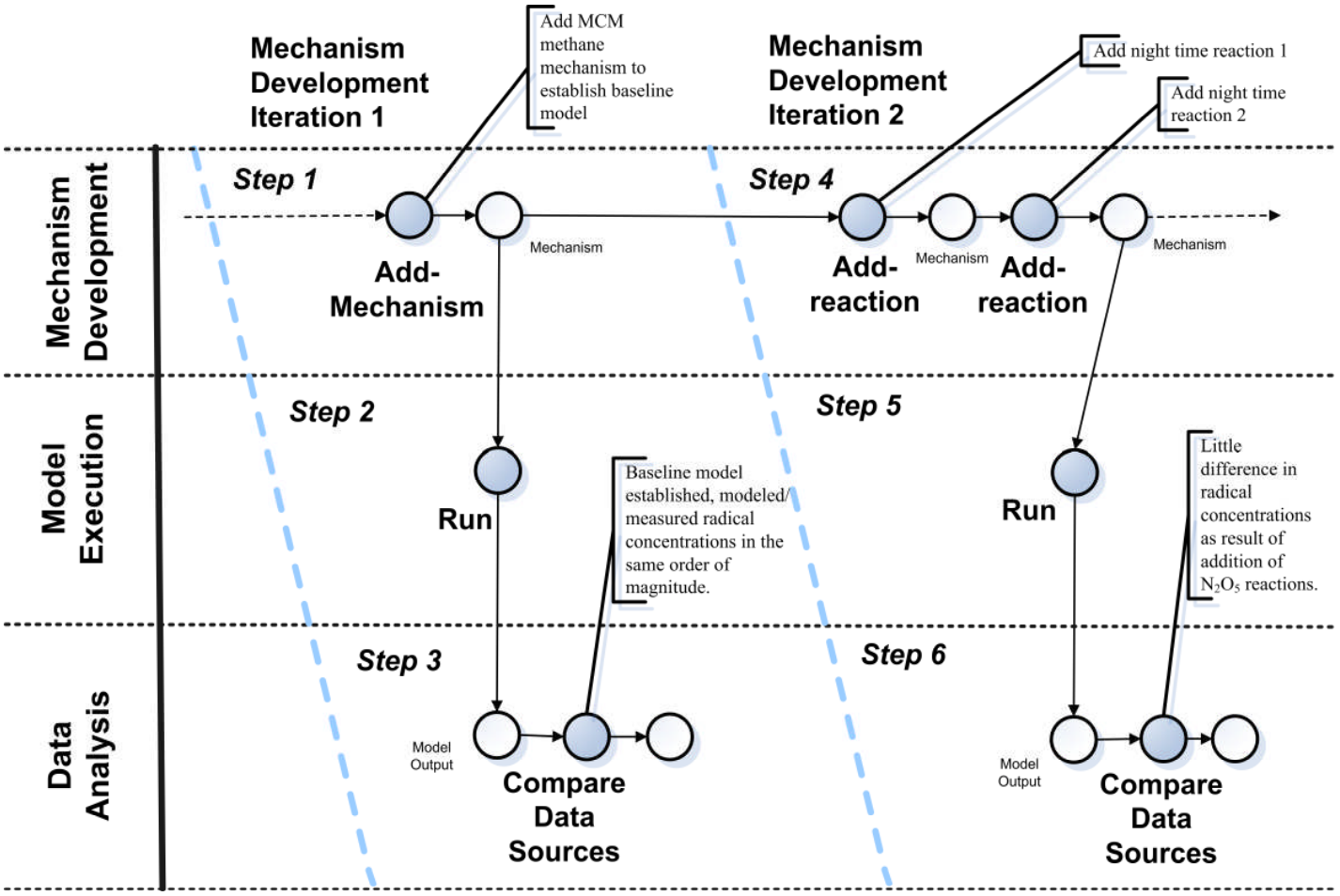


Figure 8.19: Representation of the SOAPEX case study scientific process. This figure show the structure of the provenance captured for the SOAPEX case study (Steps 1 - 6). Figure 8.20, see directly below, shows steps 7 - 9, to complete the representation of the SOAPEX cases study scientific process.

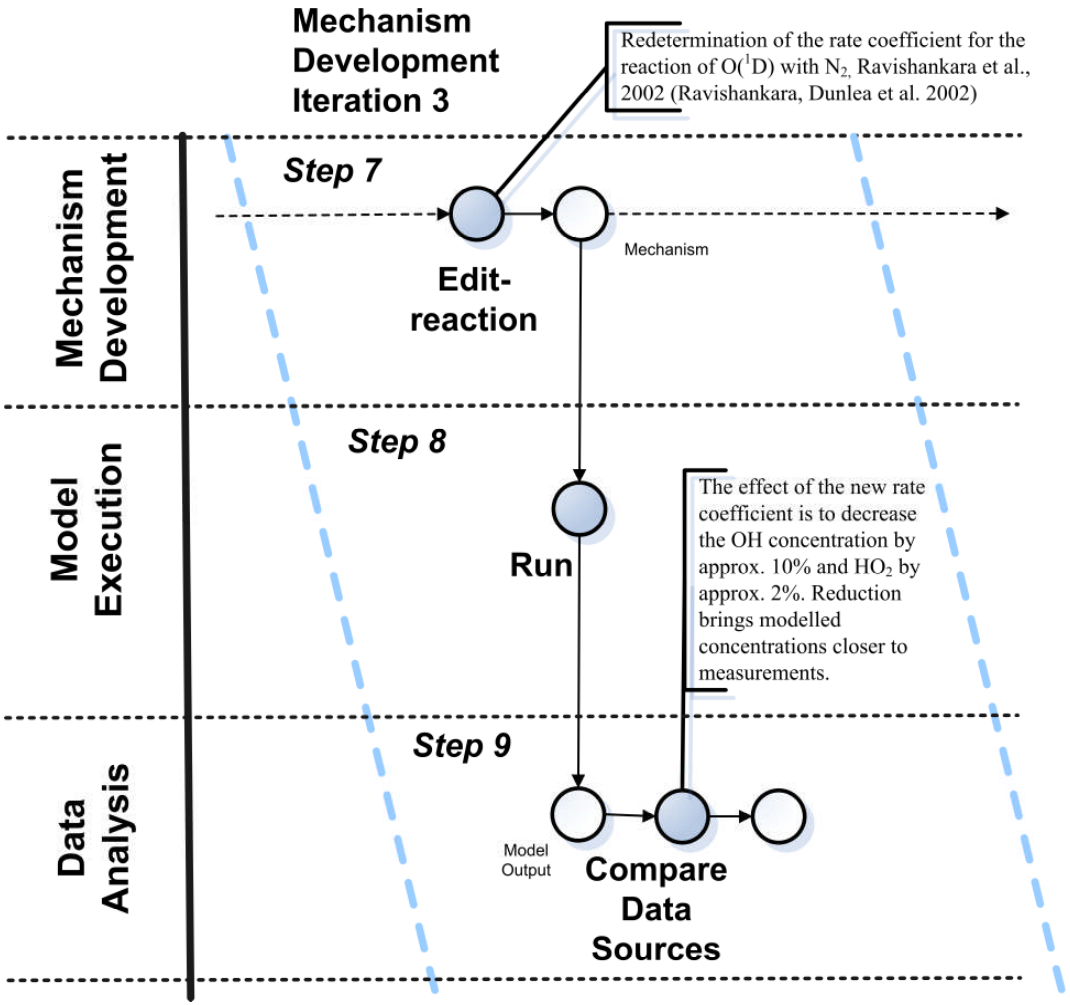


Figure 8.20: Representation of the SOAPEX case study scientific process (part 2). This figure show the structure of the provenance captured for the SOAPEX case study (Steps 7 - 9). This figure follows on from Figure 8.19, to complete the representation of the SOAPEX case study.

Up until this point scientific processes have only been considered for single model development iterations (e.g. add reaction, run model, data analysis), so in order to represent the SOAPEX case (with three model development iterations) the method of linking iterations together must be addressed. When considering mechanism development as the mode of model development, the linkage between model development iterations is made using the chemical mechanism. It is easiest to consider this linkage using a specific case, so in Figure 8.19 the mechanism added (at step 1) is the input for a "model run" process (going down the page) and also an input to the "add reaction" process (part of the second iteration, going right across the page). So the mechanism, which is edited throughout the iterations of mechanism development, can be seen to link together the provenance for modelling iterations.

8.5.3 Representation of the *In Silico* Experiment

The preceding sub-section described the ontology at the level of the scientific process, in this section an ontology for capturing provenance at a higher conceptual level, the *in silico* experiment level, is considered.

8.5.3.1 A High Level Overview of the I*n Silico E*xperiment Ontology

An overview of the *in silico* experiment ontology is provided below; this overview provides a guide to the accompanying ontology diagram (see Figure 8.21). In the centre of the ontology diagram, the *in silico* experiment class is shown, with various properties. It is this set of properties, used to characterise and describe the *in silico* experiment, that are described below, each in turn. These properties are aligned with the fields provided in ELN interface and described in section 8.4.2.

- **'has-owner':** Links the *in silico* experiment to the person responsible for running the experiment, typically a post-doctoral researcher or PhD student. Ontologies exist for describing people including the friend of a friend (FOAF) ontology [4], and will not be addressed further in this section.

- **'has-associated-researchers':** Links the *in silico* experiment to associated researchers. An experiment may be contributed to by a number of researchers, beyond the experiment owner, such as the research group leader or the experimental scientist responsible for the *in-vitro* experiment being modelled.
- **'has-associated-in-situ-experiment':** Typically, when using the MCM to develop a model, an *in situ* experiment, either field or chamber, will be the subject of model. For models of chamber and fields experiments the 'associated-in-vitro-experiment property' could point to the EUROCHAMP or BADC online databases respectively, for the experiment in question.
- **'has-associated-in-silico-experiment':** Links the *in silico* experiment in question, to other related *in silico* experiments.
- **'has-experiment-type':** There are a number of possible experiment types for models development using the MCM, the simplest types are either field model or chamber model.
- **'has-keyword':** This property allows the researcher to tag their experiment with terms from a vocabulary or free-text.
- **'has-associated-documentation':** Where an *in silico* experiment has led to the production of publication, thesis chapter, PhD report, or other unpublished document, a link to this document can be provided using this property.
- **'has-species-of-interest':** Often an experiment can be associated with specific chemical species, this relationship can be used to capture this relationship.
- **'is-executed-by':** Links the *in silico* experimental level provenance to the scientific process level provenance (that describes the modelling process executed).
- **'has-experimental-method':** Links the experiment to a description of the experimental method used.
- **'has-conclusion':** links the experiment to a description of the experiments conclusions.

The has-experimental-method and has-conclusion relationships link to ontology, so are discussed in further detail in the following sub-sections.

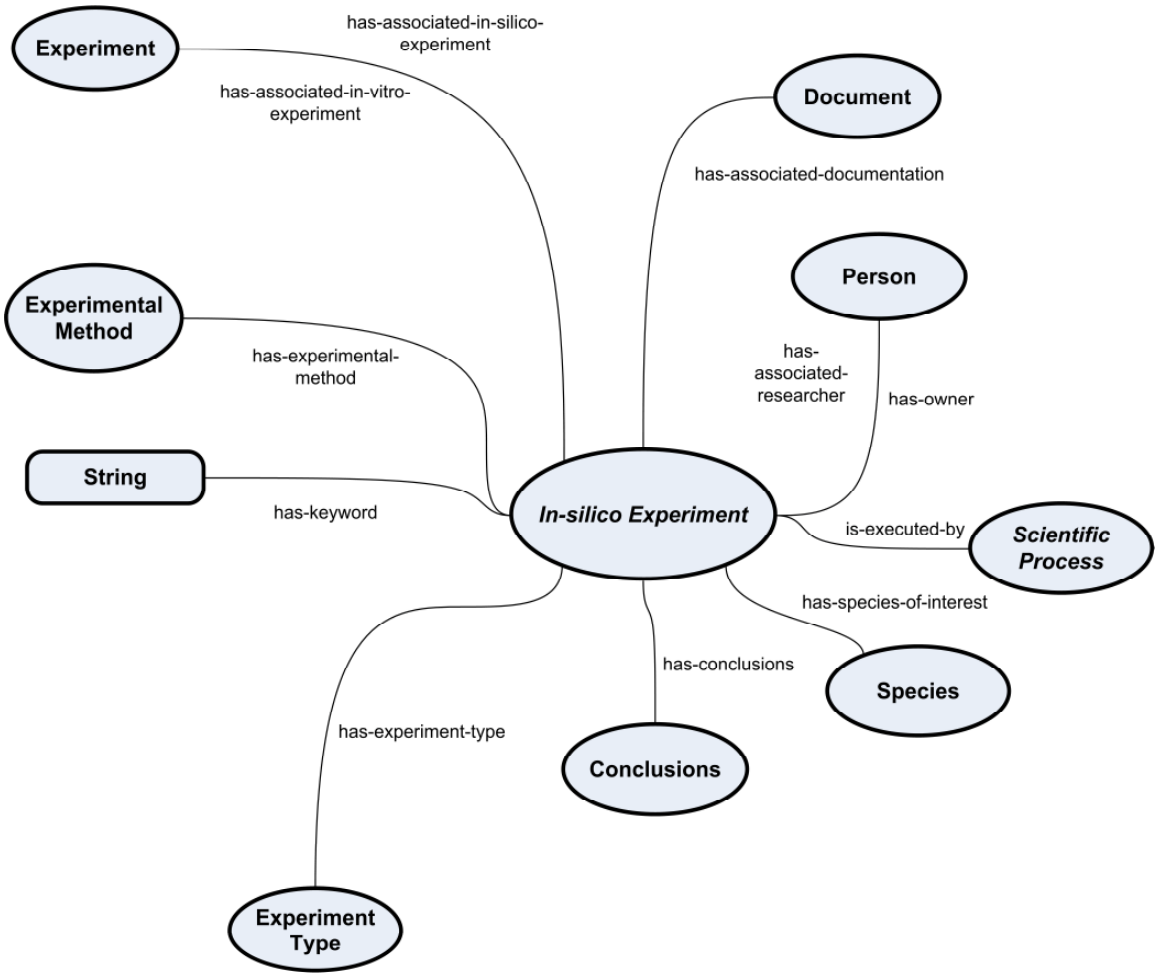


Figure 8.21: The core of the ontology uses to describe *in silico* experiments. This figure provides an overview of the ontology used to describe *in silico* experiments, including properties that describe: the owner of the experiment; the experimental method used to complete the experiment; the conclusions of the experiment; etc.

8.5.3.2 The Experimental Method

The 'has-experimental-method' property allows the researcher to provide a text description of their experimental method, for example for the experiment that underpins the Chapter 4 of this thesis, the experimental method description could be as follows.

1. Configure Model
    i. Basic Parameters (Location, date etc.) and mechanism
2. Process Constraint Data
    i. 15 minute averages
    ii. As measured (just remove data points that are errors)
3. Model Runs with various constraint configurations
4. Perform ROPA to analyse data

As shown in Figure 8.22 the experimental method has the following associated property; 'uses-model' records the models used within an experiment, typically a URL pointing to a model within some source control system. The provenance for model source code is considered to be managed by a version control system, such as svn (http://subversion.tigris.org/).

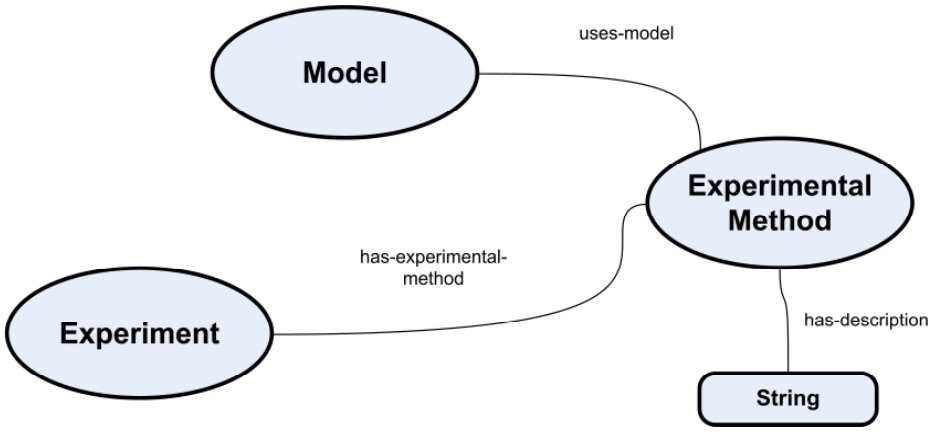


Figure 8.22: Experimental method ontology. This figure shows the ontology used to structure the experimental method. The experimental method has two proprieties: first, a text description; and secondly, a link to the model used.

8.5.3.3 Conclusions Ontology

Figure 8.23 shows the conclusions ontology used by the ELN. The conclusions of a modelling experiment can be described with free form text, using the 'has-description'

relationship, for example considering the *in silico* experiment presented in Chapter 4 the following conclusion may be attached.

> Constraining species and environmental conditions at appropriate frequencies is important to ensure the model realistically maps to the physical system. In the SOAPEX-2 case this did not deliver definitive benefits in terms of improving the model-measurement comparison.

When drawing conclusions about an experiment it is important to also look forward and identify the potential future work, the 'includes' relationship between conclusions and future plans enables a link to be established. Future plans can be linked to experiments by the 'is-executed-by' property, allowing links to proposed, completed or in progress experiments.

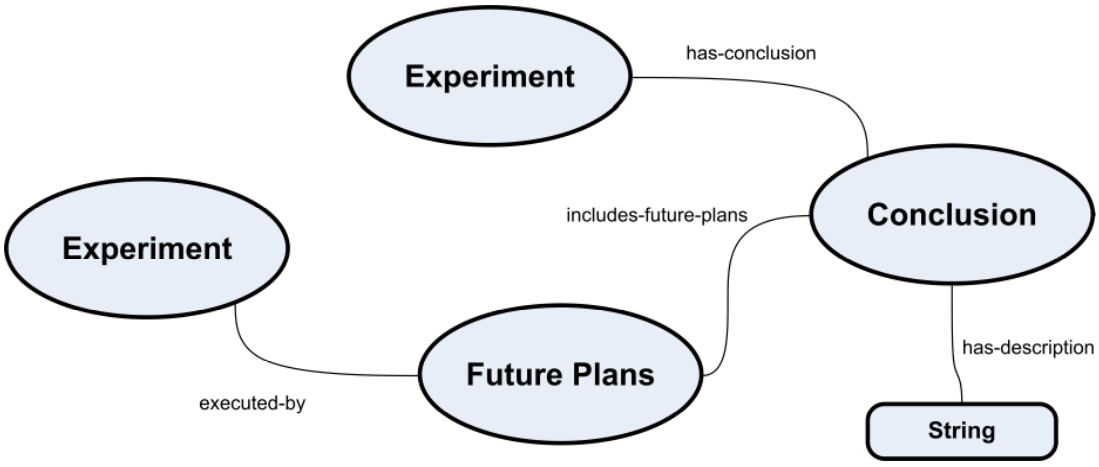


Figure 8.23: Conclusions ontology. This figure shows the conclusion ontology use to structure provenance captured by the ELN. The conclusion consists of a text description and a link to future research plans.

This section has presented the information design implemented during the development of the prototype ELN. The information design has been presented in the form of the ontology used to structure provenance captured by the ELN. The implementation of the information design is addressed in the next chapter.

Chapter Summary

This chapter has presented a description of the design of the ELN, for computational modellers using the MCM. The early sections of the chapter provided a link between the analysis of current working practices and the design of ELN, specifically the characteristics of current working practices (with respect to provenance capture when developing computational models) were analysed to determine a set of implications for the ELN design. The design approach was then outlined, including the goals, scope and principles that guided the development of the ELN. The design scope focused ELN development on the capture and representation of provenance, leaving querying provenance beyond the scope of the content presented in this thesis. The chapter then progressed to consider the envisioned working practices of a researcher developing a model using the MCM and the ELN. These envisioned working practices were presented in the form of an activity design scenario and provided a high-level description of the user experience that the ELN delivers. A high-level architecture for the ELN was then described, to provide an overview of the ELN from a systems perspective. The detail of the ELN design was then presented, in two sections: first, the interaction design (i.e. how the user interacts with the ELN); and secondly, the information design (i.e. the ontology used to structure the provenance captured by the ELN). For both the interaction and information design a distinction was drawn between provenance captured with respect to the scientific process and provenance captured with respect to *in silico* experiments. The implementation of the ELN design will be addressed in the next chapter, covering details including: the technologies and tools used to implement the ELN; and the representation of provenance captured by the ELN.

References

1. Taylor, K.R., et al., *Bringing Chemical Data onto the Semantic Web.* Journal of Chemical Information and Modeling, 2006. **46**(3): p. 939-952.
2. Frey, J., et al., *Less is More: Lightweight Ontologies and User Interfaces for Smart Labs*, in *The UK e-Science All Hands Meeting 2004*. 2004, EPSRC: Nottingham, UK.
3. Jose Kahan and M.-R. Koivunen, *Annotea: an open RDF infrastructure for shared Web annotations*, in *Proceedings of the 10th international conference on World Wide Web*. 2001, ACM: Hong Kong, Hong Kong.
4. 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.

Chapter 9 Implementation of the ELN

This chapter describes the implementation of the ELN (as described in the preceding ELN design chapter) and consists of three sections. First, the ELN architecture (as introduced in the previous chapter) is revisited and the implementation of each architectural component is described. Secondly, the interactions of ELN components for the capture of scientific process provenance are described. Thirdly, the chapter concludes with some examples of how provenance captured by the ELN is represented using semantic web technologies. The ELN was implemented jointly with Dr. Mohammed H. Haji, School of Computing, University of Leeds. The ELN source code can be found on the CD associated with this thesis.

9.1 Implementation of the System Architecture

The implementation of system architecture, introduced in the previous chapter and shown again in Figure 8.1, is presented in this section. Each component of the system architecture is briefly discussed in turn below.

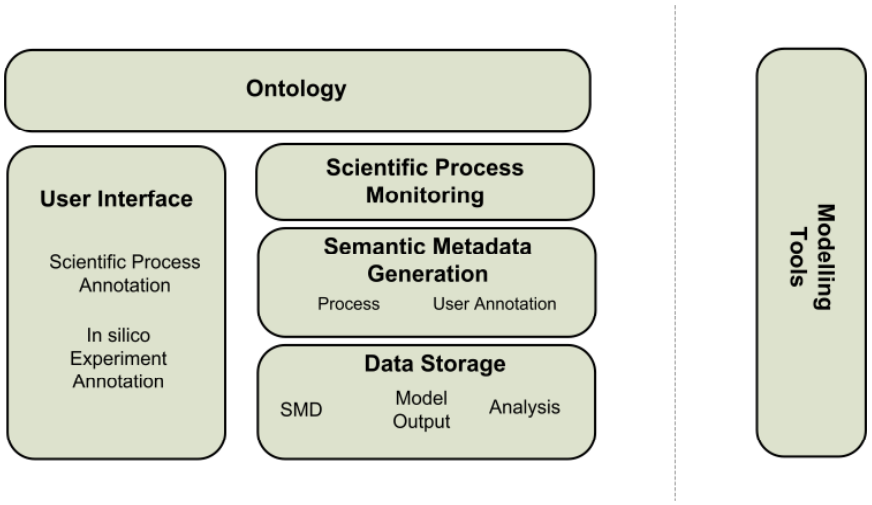


Figure 9.1: The ELN system architecture, as presented and discussed in chapter 8.

Ontology

An ontology can be defined as "a formal explicit specification of a shared conceptualisation", where a conceptualisation is an abstract model of the world or some phenomenon within it [1]. A more straight-forward definition is "computer ontologies are

structures or models of known knowledge" [2]. In the specific case of the ELN ontology, the purpose of the ontology is to provide a structure for the provenance captured by the ELN. The core components of an Ontology are:

- Concepts; for example *Person*
- Abstract Concept; a concept that cannot be instantiated, similar to an abstract class in the object-orientated programming paradigm, used as an organising structure. For example *vertebrate*.
- Properties; of the concepts. For example *name.*
- Relationships; between concepts. For example *parentOf.*

The ontology developed for the ELN was implemented using OWL (Web Ontology Language) [3], the W3C standard ontology language for the Semantic Web. Adhering to the Semantic Web architecture, enables applications to process and derive value from the provenance records generated by the ELN (such as the data and provenance aggregation application to support the development of the MCM as discussed in Chapter 5). Protégé, an OWL ontology editor, was used throughout the implementation of the ontology described in the design chapter of my thesis. Protégé enables the ontology developer to develop classes and their properties, reasoning over the resulting ontology and OWL individuals to infer new knowledge.

User interface

The user interface, as presented in Chapter 8, was implemented using Java Swing.

Computational modelling tools

The modelling tools used in conjunction with the ELN, exist outside the ELN architecture, and consist of the OSBM and a diverse set of analysis tools (including Microsoft Excel and customised python scripts).

Scientific process monitoring

This component monitors the activity of the modeller and captures process provenance. For example changes to the chemical mechanism are detected, by the scientific process monitoring component, when the user runs the model; this component compares the latest mechanism to the previous mechanism (stored locally by the ELN), to determine any changes, making use of the UNIX diff utility.

Semantic metadata generation

This component of the architecture processes provenance captured by other components of the ELN, to create a semantic metadata (SMD) representation of the provenance, that adheres to the ELN ontology. The provenance generated by the ELN is represented and stored in RDF [4] (The Resource Description Framework), adhering to the ontology (described above). RDF was originally designed as a metadata language for XML, however it is now a widely used for knowledge representation within and beyond the Semantic Web [5]. Extensive use was made of Jena [6] [7], an open source framework, that has emerged from the research work of Hewlett Packard Semantic Web Research Program (http://www.hpl.hp.com/semweb/). The Jena functionality used in the development of the ELN included the RDF API; which provides functionality to read and write RDF, in RDF/XML, N3 and N-Triples formats. Functionality is also provided to create programmatic RDF models within an application.

Data storage

Model output data and data analysis documents are stored in a MySQL database (http://www.mysql.com/). Future work will look at the additional use of a triplestore, for storage of the SMD produced by the ELN; research that has focused on the scalable storage of rdf includes CombeChem[8].

Programming language

The core programming language used during the development of the ELN prototype was Java. Java was selected for three main reasons: first, the team involved in development of the ELN had previous experience developing Java applications (negating the learning curve associated with developing applications using an unfamiliar language); secondly, Java applications benefit from the inherent portability of the Java language (a guiding principle of the language is “write applications once and run anywhere”, thanks to the Java Virtual Machine); thirdly, a variety of Semantic Web tools and libraries providing Java interfaces exist within the public domain, facilitating the development of the ELN as an application capable of producing Semantic Web content.

9.2 Capturing Provenance with Respect to the Scientific Process

In this section the way in which the components of the ELN interact, during the capture of provenance with respect to the scientific process, is discussed. If the simplest model development iteration (the chemical mechanism is edited, the model is run, and the model output data is analysed) is considered, then the interactions are as follows.

Mechanism development: When starting a new model development project a unique global URI is automatically assigned to the experiment. The modeller can then proceed to develop the chemical mechanism e.g. adding an MCM mechanism, editing existing reactions or inserting new reactions. The modeller commits to the changes in the mechanism by calling the model execution command. The scientific process monitoring layer then places a lock on the model (temporarily preventing the model running), to enable provenance to be captured before the model runs. The scientific process monitoring layer then identifies and captures any such changes to the chemical mechanism and drives the annotation interface to prompt the user for scientific reasoning. Once user annotation has been completed, the provenance from the user interface and the scientific process monitoring layers is combined and the SMD generation layer produces an rdf representation of the provenance captured.

Model execution: Prior to the model running the scientific process monitoring layer takes the model input files and passes them to the data storage layer for archival, and then releases its lock on the model. The model then compiles and runs, and the scientific process monitoring layer takes the model output files and again passes them to the data storage layer for archival, driving the user interface to generate a prompt for annotation. The user completes the prompt and the provenance from the user interface and the scientific process monitoring layers is combined by the semantic metadata generation layer.

Model input and output files are inserted into the ELN database using JDBC-ODBC (Java Database Connectivity - Open Database Connectivity). As each input or output file is added to the database it is allocated a resolvable URI that is referenced from the SMD. The archival of these files enables the experiment results to be quickly accessed for future analysis.

Data analysis: The analysis interface is presented, giving the user the opportunity to record: the data sources they have used; the type of analysis conducted; their conclusions; and their plans for the next modelling iteration. A lock is placed on the model, preventing users from editing or running the model until the analysis interface has been completed. Once the user has performed the analysis of the model output they complete analysis interface. The SMD generation layer then generates a rdf representation of the data analysis provenance and releases the lock on editing or running the model. The SMD for the model development iteration is then aggregated and submitted to the data storage layer for archival. The next iteration of model development can then commence.

9.3 Provenance Representation

This section presents example rdf representations of provenance captured by the ELN, for elements of the SOAPEX model development case study. The rdf samples presented were generated by the ELN. These samples were then edited by hand to improve readability and simplify the content of the sample. The rdf samples consider the provenance captured by the ELN for step 4 -6 of the SOAPEX case study, a single model development iteration consisting of: step 4, mechanism development, two reactions are added to the mechanism; step 5, model execution, the model is run; step 6, data analysis, the model output data is compared with a number of other data sources. In this section each step in the case study is considered in turn, with the provenance captured by the ELN presented in both a graphical and rdf forms.

9.3.1 Mechanism Development

Process Description: The modeller adds two reactions to the mechanism, to characterise elements of chemistry taking place during the night, and provides an annotation when prompted by the ELN.

Reaction added: %2.50d-22*H2O : N_2O_5 = HNO_3 + HNO_3**;**
Reaction added: %1.80d-39*H2O*H2O: N_2O_5 = HNO_3 + HNO_3**;**

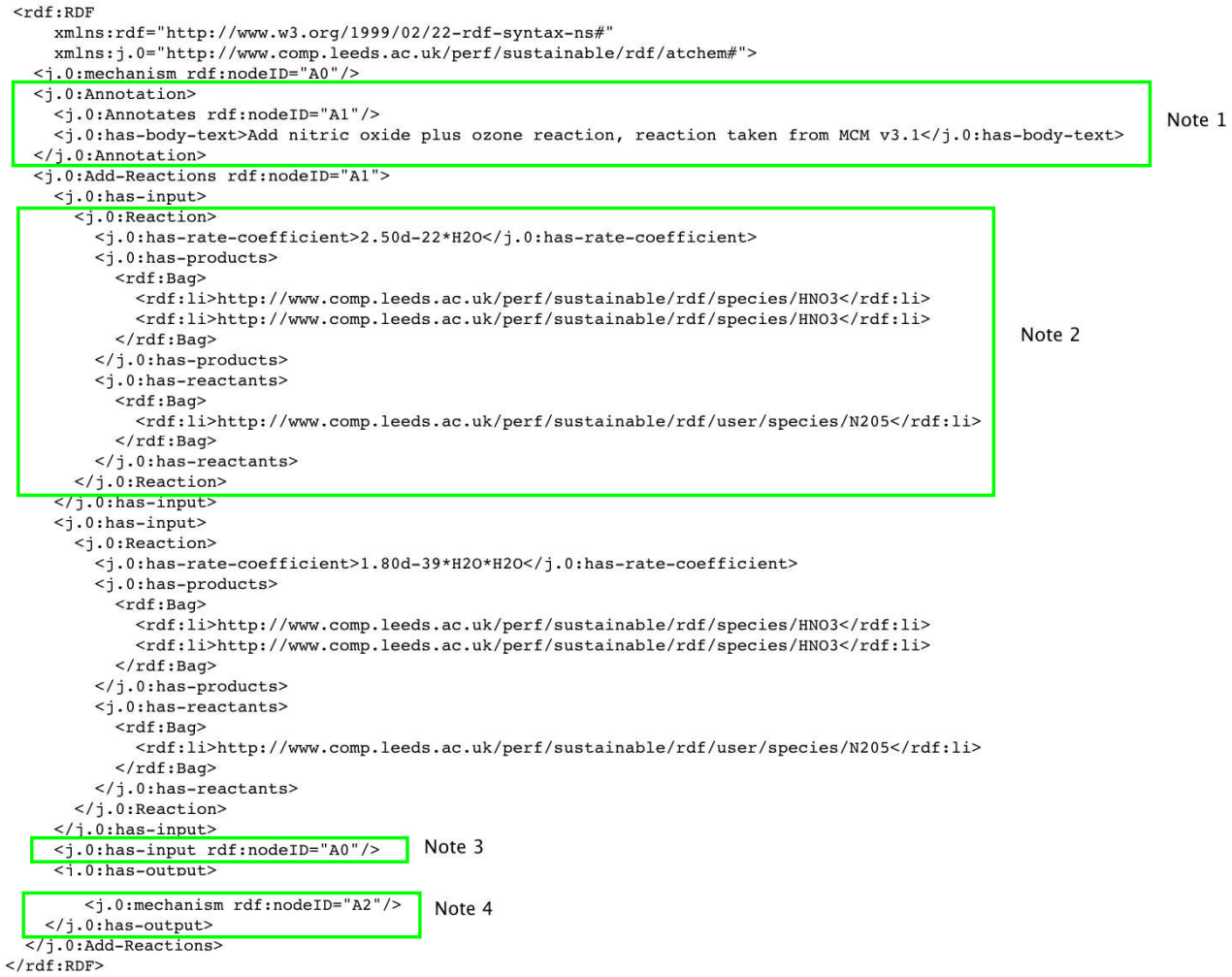
Figure 9.2 shows the rdf representation of this process while Figure 9.3 provides a graphical representation (using a notation familiar from the design chapter) of this process, these figures present the same information and are described in conjunction in the following text. Figure 9.2 and Figure 9.3 are both annotated, with sections highlighted and labelled, each of these annotations is discussed below.

Note 1. Highlights the annotation of the ‘add reaction’ process. Here the subject of the annotation is defined as rdf:nodeID=“A1”, where this identifier has been automatically assigned to the ‘add reaction’ process. The user’s annotation is captured by the ‘has-body-text relationship’.

Note 2. Highlights the representation of one of the reactions added to the mechanism. The ELN parses the reaction and splits it in it components: reactants; products; and a rate co-efficient. The ‘has-reactant’, ‘has-product’ and ‘has-rate-coefficient’ relationships capture these components of a reaction.

Note 3. The chemical mechanism, that forms a key input to the ‘add reaction’ process, is highlighted. This mechanism has been given the identifier A0.

Note 4. Highlights the chemical mechanism produced, as an output of the mechanism development process, i.e. the original mechanism plus the two reactions added. No details about the output mechanism are stored because they can deduced from the input mechanism and the details of the reactions added. The output mechanism is identified in order to link together the mechanism development and model execution activities.




```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:j.0="http://www.comp.leeds.ac.uk/perf/sustainable/rdf/atchem#">
  <j.0:mechanism rdf:nodeID="A0"/>
  <j.0:Annotation>
    <j.0:Annotates rdf:nodeID="A1"/>
    <j.0:has-body-text>Add nitric oxide plus ozone reaction, reaction taken from MCM v3.1</j.0:has-body-text>
  </j.0:Annotation>
  <j.0:Add-Reactions rdf:nodeID="A1">
    <j.0:has-input>
      <j.0:Reaction>
        <j.0:has-rate-coefficient>2.50d-22*H2O</j.0:has-rate-coefficient>
        <j.0:has-products>
          <rdf:Bag>
            <rdf:li>http://www.comp.leeds.ac.uk/perf/sustainable/rdf/species/HNO3</rdf:li>
            <rdf:li>http://www.comp.leeds.ac.uk/perf/sustainable/rdf/species/HNO3</rdf:li>
          </rdf:Bag>
        </j.0:has-products>
        <j.0:has-reactants>
          <rdf:Bag>
            <rdf:li>http://www.comp.leeds.ac.uk/perf/sustainable/rdf/user/species/N205</rdf:li>
          </rdf:Bag>
        </j.0:has-reactants>
      </j.0:Reaction>
    </j.0:has-input>
    <j.0:has-input>
      <j.0:Reaction>
        <j.0:has-rate-coefficient>1.80d-39*H2O*H2O</j.0:has-rate-coefficient>
        <j.0:has-products>
          <rdf:Bag>
            <rdf:li>http://www.comp.leeds.ac.uk/perf/sustainable/rdf/species/HNO3</rdf:li>
            <rdf:li>http://www.comp.leeds.ac.uk/perf/sustainable/rdf/species/HNO3</rdf:li>
          </rdf:Bag>
        </j.0:has-products>
        <j.0:has-reactants>
          <rdf:Bag>
            <rdf:li>http://www.comp.leeds.ac.uk/perf/sustainable/rdf/user/species/N205</rdf:li>
          </rdf:Bag>
        </j.0:has-reactants>
      </j.0:Reaction>
    </j.0:has-input>
    <j.0:has-input rdf:nodeID="A0"/>
    <j.0:has-output>
      <j.0:mechanism rdf:nodeID="A2"/>
    </j.0:has-output>
  </j.0:Add-Reactions>
</rdf:RDF>
```

Figure 9.2: RDF representation of provenance captured by the ELN, for step 4 of the SOAPEX model development case study. The provenance represents the process of adding two reactions to an existing mechanism.

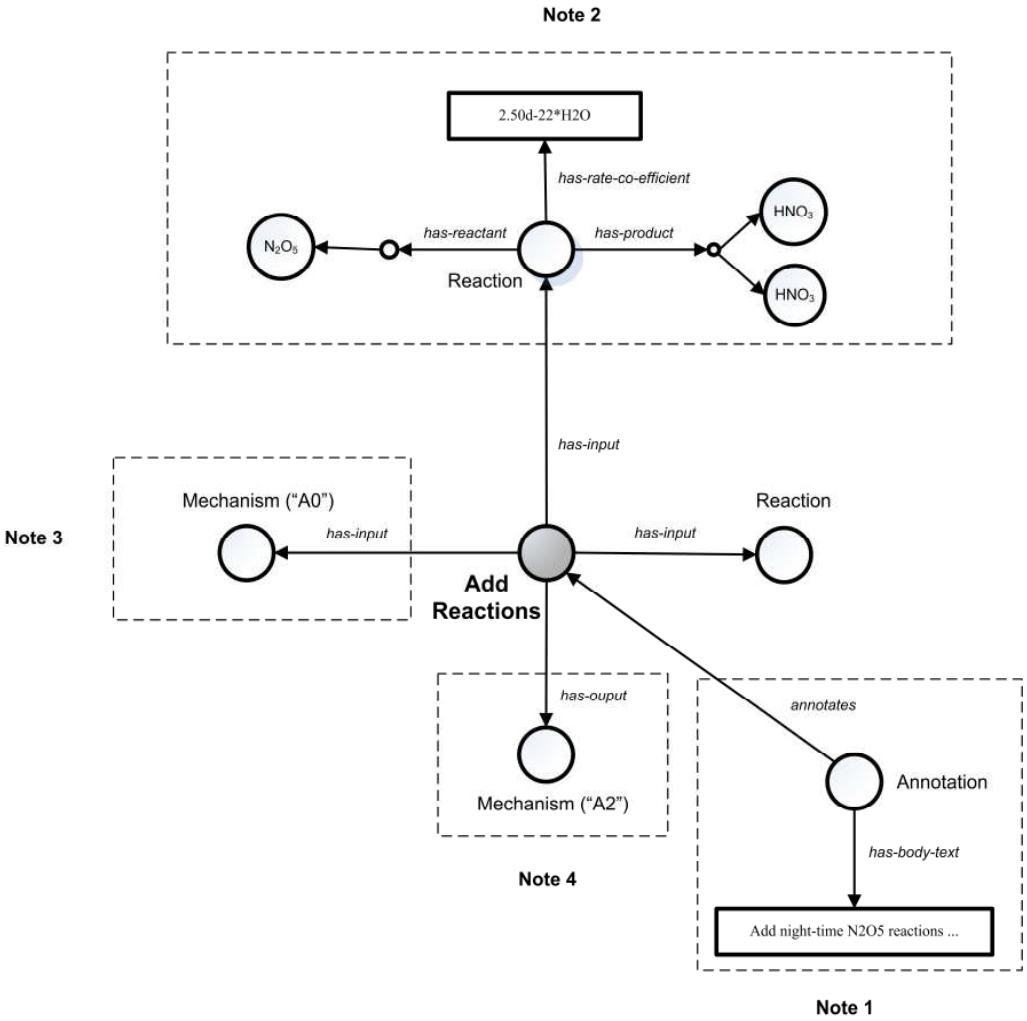


Figure 9.3: Graphical representation of provenance captured by the ELN, for step 4 of the SOAPEX model development case study. The provenance represents the process of adding two reactions to an existing mechanism.

9.3.2 Model Execution

This sub-section considered the rdf representation of provenance for a model execution process. A specific example is examined, step 5 from the SOAPEX model development case study, described below.

Process Description: having completed the ELN prompt (driven by adding two reactions to the mechanism), the model runs.

Again two representations of the provenance captured by the ELN are presented and annotated, a RDF representation (see Figure 9.4) and a graphical representation (see Figure 9.5). The provenance presented for the model execution process is simplified in order to provide a concise overview of the rdf structure, simplifications include: showing the provenance for a sub-set of model input and output files; omitting basic metadata about the model execution such as execution location. Two aspects of the rdf representation, as annotated on Figure 9.4 and Figure 9.5, are described below.

Note 1. Highlights the model input and output files. The URIs for each of these input and output files, were generated when they were submitted to the ELN database, and are resolvable to retrieve the files from the database.

Note 2. Highlights the linking of input and output files to the model execution process, using the *has-input* and *has-output* relationships respectively.

```
<j.0:Input-File rdf:about="http://www.comp.leeds.ac.uk/Demo/2008-04-1_2:50:46/input/mechanism_species/"/>
<j.0:Input-File rdf:about="http://www.comp.leeds.ac.uk/Demo/2008-04-1_2:50:46/input/mechanism_reac/"/>
<j.0:Input-File rdf:about="http://www.comp.leeds.ac.uk/Demo/2008-04-1_2:50:46/input/mechanism_prod/"/>
<j.0:Output-File rdf:about="http://www.comp.leeds.ac.uk/Demo/2008-04-1_2:50:46/output/concentration_output/"/>
<j.0:Model-Execution rdf:nodeID="A3">
  <j.0:has-input rdf:resource="http://www.comp.leeds.ac.uk/Demo/2008-04-1_2:50:46/input/mechanism_species/"/>
  <j.0:has-input rdf:resource="http://www.comp.leeds.ac.uk/Demo/2008-04-1_2:50:46/input/mechanism_prod/"/>
  <j.0:has-input rdf:resource="http://www.comp.leeds.ac.uk/Demo/2008-04-1_2:50:46/input/mechanism_reac/"/>
  <j.0:has-input rdf:nodeID="A2"/>
  <j.0:has-output rdf:resource="http://www.comp.leeds.ac.uk/Demo/2008-04-1_2:50:46/output/concentration_output/"/>
</j.0:Model-Execution>
<j.0:Annotation>
  <j.0:Annotates rdf:nodeID="A3"/>
  <j.0:has-body-text>Normal model run</j.0:has-body-text>
</j.0:Annotation>
```

Note 1

Note 2

Figure 9.4: RDF representation of provenance captured by the ELN, for step 5 of the SOAPEX model development case study. The provenance represents the process of running the computational model

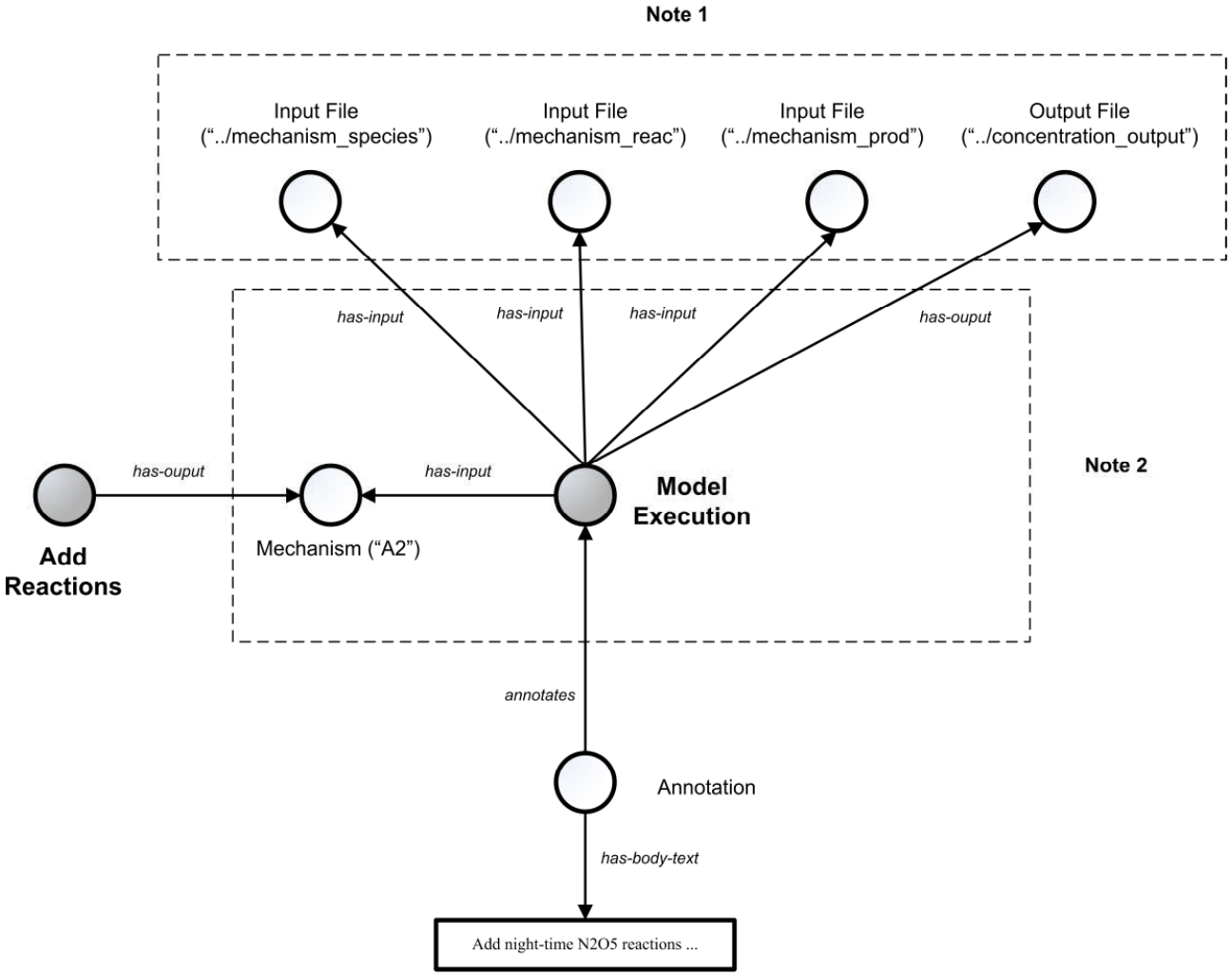


Figure 9.5: Graphical representation of provenance captured by the ELN, for step 5 of the SOAPEX model development case study. The provenance represents the process of running the computational model.

9.3.3 Data Analysis

This sub-section considers the rdf representation of provenance for the data analysis process. A specific example is examined, step 6 from the SOAPEX model development case study, described below.

Process Description: the modeller takes the model output and analyses it, by plotting a series of graphs (using Microsoft Excel). During the data analysis the modeller makes use of field experiment data, taken from the BADC (the British Atmospheric Data Centre). The modeller submits provenance describing the data used and the conclusions of the analysis process, using the ELN interface.

Again two representations of the provenance captured by the ELN are presented and annotated, an RDF representation (see Figure 9.6) and a graphical representation (see Figure 9.7). Four aspects of the provenance representation, as annotated on Figure 9.6 and Figure 9.7, are described below.

Note 1. The external data sources used in the data analysis process are identified. The data used is taken from the BADC (an online database), so a resolvable URL is used as an identifier.

Note 2. The model output, identified by the URI allocated when it was submitted to the ELN database, is an input to the data analysis process. This provides the linkage between the model execution process and the data analysis process.

Note 3. The annotation of the data analysis process is shown, capturing the conclusions of the modeller; "Adding N_2O_5 night-time reactions has little impact on the radical concentrations.".

Note 4. The annotation of a material, one of the external data sources taken from the BADC, is shown. The annotation gives a description of the data, as provided by the modeler via the ELN interface; "Experimental OH data".

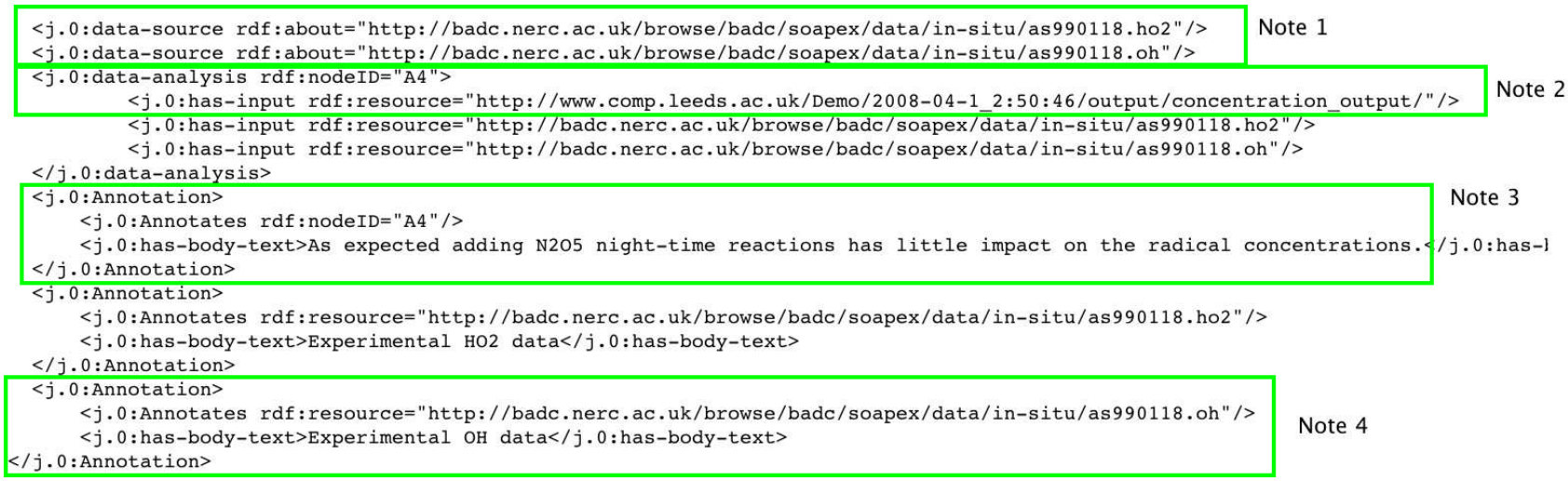


Figure 9.6: RDF representation of provenance captured by the ELN, for step 6 of the SOAPEX model development case study. The provenance represents the process of analysing model output data.

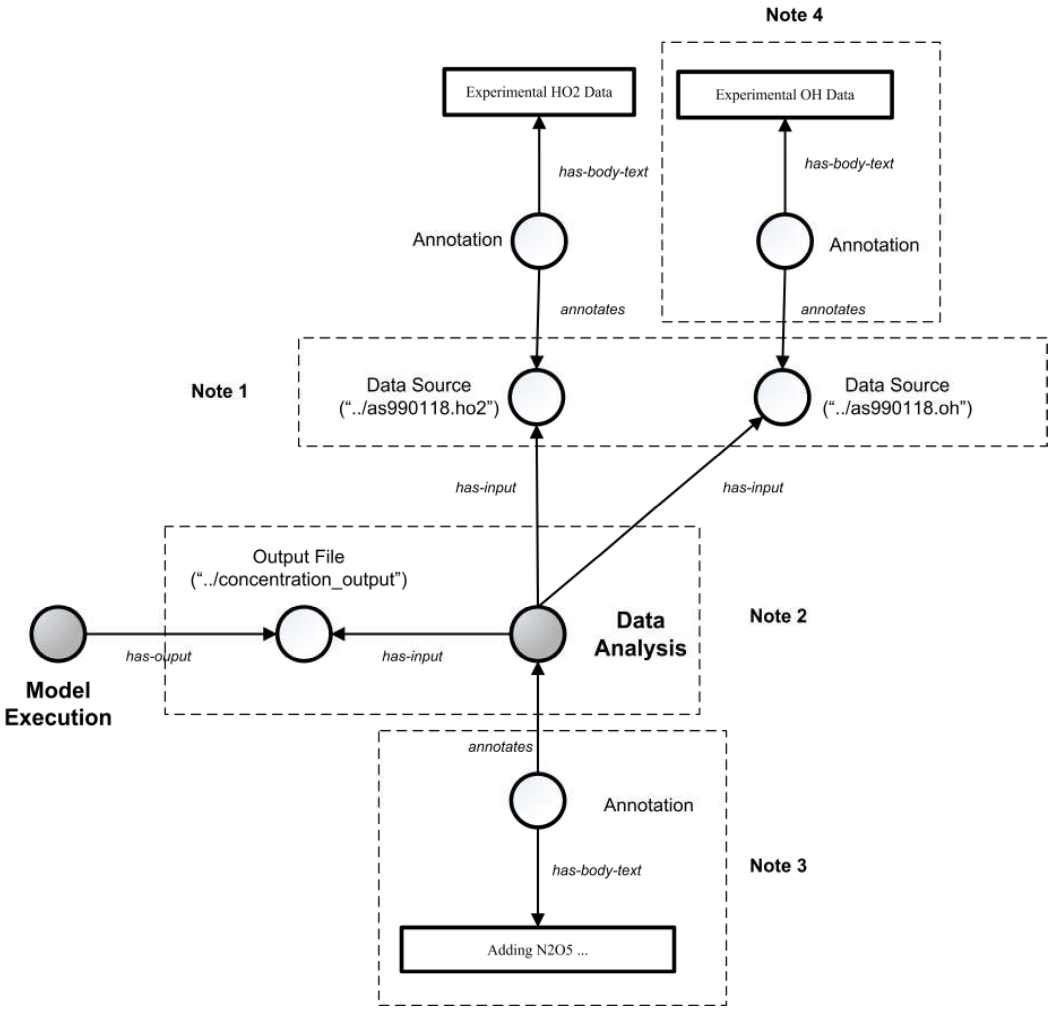


Figure 9.7: Graphical representation of provenance captured by the ELN, for step 6 of the SOAPEX model development case study The provenance represents the process of analysing model output data.

Chapter Summary

This chapter has provided an overview of the implementation of the ELN, including a description of the tools and technologies involved, and detail of how the components of the ELN interact with each other to capture and store provenance. The representation of provenance using a Semantic Web technology (rdf) has also been discussed, with the aid of three annotated examples taken from the SOAPEX case study. The next chapter proceeds to discuss the evaluation of the ELN, addressing the evaluation approach and results.

References

1. Gruber, T.R., *A translation approach to portable ontology specifications.* Knowl. Acquis., 1993. **5**(2): p. 199-220.
2. Leuf, B., *The Semantic Web: Crafting Infrastructure for Agency*. 2005, Chichester: Wiley
3. Bechhofer, S., et al. *OWL Web ontology language reference*. 2004 [cited 5th March 2009]; Available from: http://www.w3.org/TR/owl-ref/.
4. *Resource Description Framework (RDF) Model and Syntax Specification*. 1999 [cited 29th January 2009]; Available from: http://www.w3.org/TR/REC-rdf-syntax/.
5. Decker, S., et al., *The Semantic Web: The Roles of XML and RDF*. IEEE Internet Computing, 2000. **4**(5): p. 63-74.
6. McBride, B., *Jena: a semantic Web toolkit.* Internet Computing, IEEE, 2002. **6**(6): p. 55-59.
7. Carroll, J.J., et al., *Jena: implementing the semantic web recommendations*, in *Proceedings of the 13th international World Wide Web conference on Alternate track papers \& posters*. 2004, ACM: New York, NY, USA.
8. Hughes, G., et al., *The semantic smart laboratory: a system for supporting the chemical eScientist.* Organic and Biomolecular Chemistry, 2004. **2**: p. 1-10.

Chapter 10 Evaluation of the ELN

This chapter presents the results of the evaluation of the ELN. The goal of the evaluation was to elicit responses from potential ELN users that will inform the design of a production quality ELN for use by the wider community. The evaluation consisted of an in-depth, qualitative, semi-structured interview; two members of the MCM-user community were selected as evaluators[17]. This chapter consists of three sections: the first section presents an overview of the evaluation itself; the second section presents the evaluation results, with the implications of these results highlighted; and in the final section, the implications of the evaluation results are aggregated and discussed in greater detail.

10.1 Evaluation Overview

The mode of evaluation was very much formative [1], 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 provenance capture 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 ensured important topics are addressed.

The scope of the evaluation was limited to considering the capture of provenance with respect to the scientific process. The ELN interface prototypes for the capture of provenance for *in silico* experiments (as described in Chapter 8) were not evaluated in order to restrict the evaluation activity to a manageable domain. The ELN functionality for capturing provenance with respect to the scientific process was selected for evaluation as it is firmly grounded in analysis of current working practices, so the evaluators could easily relate to and understand the ELN functionality they were presented with. Additional detail, describing the nature of the evaluation is presented below, in two parts: first, the

[17] Both of the evaluators regularly perform *in silico* experiments that make use of the MCM.

structure of the evaluation is described; and secondly, the methodology used to analyse the evaluation results is discussed.

10.1.1 Evaluation Structure

The evaluation was structured around a set of scenarios developed during the analysis of current practice and ELN design phases of prototype development. The evaluation itself addressed four topics, each discussed below.

Current practice: The evaluation opened with a semi-structured interview and discussion of current practice for provenance capture. This discussion was prompted by a problem scenario; a high level description of the way in which an individual might use their laboratory notebook to capture provenance.

Envisioned practice (with an ELN): Following on from the topic of current practice, the evaluators were presented with an activity design scenario; a high level description of the way in which an individual may use his or her ELN to capture provenance. Again the format for this section of the interview was semi-structured interview, prompted by the activity design scenario and number of associated design documents.

Demonstration of the ELN: The functionality of the ELN was then demonstrated to the evaluator, with explanation where required. The demonstration followed a predefined modelling process, based on modelling work conducted for the SOAPEX field campaign [2]. The modelling was conducted using the OSBM with which the ELN is loosely coupled. The evaluators were familiar with OSBM and other modelling tools used, enabling them to focus on the evaluation of the ELN.

User testing: The evaluators were then invited to test the ELN prototype; they were provided with the option of starting a new piece of modelling or continuing from where the demonstration had left off. The evaluators were encouraged to verbalise their thought processes, ask questions and suggest improvements throughout their time testing the ELN prototype. Following the user testing the evaluator was asked to provide some comments about their general impressions of the ELN.

10.1.2 Evaluation Methodology

Having discussed the structure of the evaluation, this sub-section describes the evaluation methodology adopted. This description addresses two key topics: first, the scope of the evaluation; and secondly, how the qualitative data generated by the evaluation was analysed.

Evaluation scope

Two researchers, with substantial experience of developing atmospheric chemistry models using the MCM, evaluated the ELN. The evaluators 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 use of a small number of potential users, with close links to the software development team, to evaluate scientific software has also been applied successfully in the large e-Science projects such as MyGrid [3, 4] and MyExperiment [5-7]. Goble and De Roure [8] recommend that when developing software for scientists, one should "act local, think global". By acting to meet the requirements of a small number of local, well known scientists (who acts as pioneers); whilst thinking about the requirements of the wider user community, a widely adopted software application can be developed.

Analysing the qualitative data produced by the evaluation

Audio recordings of both the evaluations were transcribed, to form a qualitative data set. This dataset was then analysed using techniques from grounded theory [9] [10] [11]. Grounded theory is a systematic qualitative research methodology, used across the social science research disciplines [10]. The defining characteristic of grounded theory is that qualitative data are analysed to generate theory or a hypothesis (rather than generating a theory and seeking to capture qualitative data that supports the theory, as in other qualitative research methodologies) [9].

Within grounded theory the process of analysing qualitative data is referred to as coding. Coding involves reading (and re-reading), the qualitative data source (in this case the transcripts of the evaluation interviews) to identify concepts and interrelationships occurring with the data [9]. It is from these concepts and interrelationships, that the individual performing the coding can generate theory about the world that the qualitative data describes.

The coding of the qualitative data (produced by the ELN evaluations) took place iteratively, working through two different types of coding: open coding [11] and axial coding [10]. Both of these types of coding are explained below, with reference to a fragment of problem scenario 1 (introduced in chapter 7), which serves as an exemplar piece of qualitative data.

> Helen is developing models of a set of chamber experiments, the model is developed iteratively. A modelling iteration typically involves model development, running the model, and analysing the model output to identify the appropriate model development for the next iteration. The goal of her piece of modelling research is to obtain a good agreement between model output and experimental measurements, deriving some insight into the chemical mechanism in the process. …

Figure 10.1: Qualitative Data Sample: Taken from problem scenario 1; describing current working practices for the capture of provenance and data.

Open coding

During open coding[18] [11] the analyst seeks to identify and label concepts within the text. So when conducting open coding of the data sample above many concepts can be identified including: model output and experimental measurements; these concepts could be labelled "model output dataset" and "*in situ* experimental dataset" respectively. Creating labels allows multiple references to the same concept to be collected together. Having identified a set of concepts, categories (that group together a set of concepts) can be identified [11]; in the example both "model output dataset" and "*in silico* experimental data" are members of the category "dataset".

During open coding the analyst also seeks to identify the properties of categories and concepts. So in the example above, one concept is a "dataset comparison":

> "The goal of her piece of modelling research is to obtain a good agreement between model output and experimental measurements"

[18] The open in open coding refers to the qualitative analyst approaching the coding process with an open mind, free of preconceptions.

A dataset comparison has a property that defines the level of agreement between the datasets, “a good agreement” in the text above.

Axial Coding

During axial coding [10] the analyst seeks to link concepts and categories identified during open coding. By linking together categories and concepts, the analyst gains an understanding of the domain described by the qualitative data. In order to generate a manageable number of these links, the analyst will typically focus on a number of specific relationships, such as: consequences (*A* happened as consequence of *B*); casual conditions (*A* caused *B*); and context (i.e. background information). So in the example above “deriving some insight into the chemical mechanism” can be seen to be a consequence of conducting “model research”. This raises questions in itself that cannot be answered by the text, such as under what conditions is insight derived from modelling research.

10.2 Evaluation Results

Having presented an overview of the evaluation goals, structure and methodology in the preceding section, this chapter now progresses to present the evaluation results. These results are presented in five sub-sections: first, the evaluator’s general perceptions of, and attitudes towards, provenance are explored; secondly, the evaluators’ perceptions of current provenance capture practices are discussed; thirdly, the evaluators’ perceptions of an ELN, as a concept, are discussed; next, the evaluators’ responses to the ELN prototype are discussed; and finally, specific improvements to the ELN prototype, as suggested by the evaluators, are presented. Prior to these sub-sections a brief summary of the evaluation results is presented, outlining the key results that are expanded upon during the more detailed discussions that follow.

Terminology: Throughout this section reference is made to adopting a holistic approach to provenance, as an implication of the evaluation results. In this context I define a holistic approach to provenance as considering within the design scope of the ELN development: the degree to which current processes and tools facilitate provenance capture; and, the full set of model development processes. So adopting a holistic approach opens up the possibility of reengineering existing model development processes and tools to facilitate

provenance capture by the ELN. This holistic approach is in contrast to the more conservative approach adopted during the design of the ELN, where: the ELN was designed to fit in with existing tools and processes; and scope of the provenance captured by the ELN was restricted.

10.2.1 Summary of Evaluation Results

The overall response to the ELN and the approach to provenance capture adopted was positive, with the evaluators able to see significant value in capturing provenance using the ELN and drawbacks to current provenance capture practices. The evaluators highlighted a number of benefits of adopting the ELN, including: the well structured nature of the provenance captured; prompts for annotations encouraging good practice in provenance capture and model development; and the automation of provenance capture. The evaluators also highlighted a number of limitations of the ELN prototype and the ELN design approach, including: the limited scope of prototype development; and the inflexible nature of the ELN user interface. The following sub-sections pick up these themes and examine them in greater detail, with the evaluation results presented in the form of a commentary developed following the coding of the evaluation transcripts. The commentary provides supporting evidence in the form of quotations taken from the evaluation transcripts.

10.2.2 General Perceptions of Provenance

This sub-section presents a set of evaluation results, which outline the general perceptions of the evaluators to provenance. The general perception was that provenance capture is a secondary consideration, and that getting on with the research at hand is the primary consideration of the researcher. This perception is discussed in three parts: first, evidence supporting this perception is presented; secondly the factors that lead to this perception are identified and discussed; and thirdly, the implications of this perception are discussed.

10.2.2.1 Provenance Capture as a Secondary Consideration

The general attitude of evaluators was that recording provenance plays a secondary role to executing the scientific process. This perception has been supported by the analysis of the initial evaluations:

> "you are looking at the science and not the way you are doing it"
>
> "[Provenance capture is] not absolutely necessary but beneficial."

In some cases provenance capture can even be seen as burden, something to be avoided:

> "I have got away without doing it completely for a long time."

10.2.2.2 Causes of Provenance Capture being Viewed as a Secondary Consideration

From the analysis of the evaluation results three factors can be seen to contribute to the perception of provenance as a secondary consideration. These factors are: the task focus of researchers; time constraints; and the fact that the value of provenance can only be realised after its capture. Each of these factors is described in detail below.

Task focus

The first factor to be considered is what I have described as task focus, i.e. the researcher is focussed on the task of developing models (not the task of capturing provenance). The evaluators referred to the task in question as 'the science'.

> "you are looking at the science and not the way you are doing it."
>
> "you … concentrate on the scientific process"

The use of the term 'the science' is interesting in itself, although the actual meaning of 'the science' was not probed during the evaluation; here 'the science' is taken to mean research work that aims to produce publishable (or interesting) scientific findings. For a fixed-term researcher or PhD Student, conducting model development, 'the science' is their key motivation and means of gaining recognition within their field. 'The science' leads to publications, which play a critical role in the development and progression of an

academic career. So motivation to pursue 'the science' is clear; it is interesting, leads to recognition and career development; whereas the motivation to capture provenance is not so clear, i.e. the information captured may or may not be of use at some unknown point in the future.

Time constraints

The second factor, referred to during the initial evaluations, contributing to a lack of focus on provenance capture was time constraints. In this context time constraints have been defined as having insufficient time to complete all the work (that is desirable to complete) leading to prioritisation of individual tasks. This prioritisation is to the detriment of accurate and complete provenance capture, as from the discussion above it can be seen that the task of conducting research is the primary goal of a researcher, and as such has higher priority than provenance capture.

> "[Limited, ad-hoc provenance capture is] Less time consuming than having to organise … [provenance] … in a logical way [which] will take time …[away from] focussing on the … science."

> "if you are under time restrictions, which you are to a certain extent, to get the data out … [Limited, ad-hoc provenance capture] … is the way you would do it (provenance capture) although it's not the best way to do it."

I would suggest that the greater the time constraints, i.e. the greater the pressure to deliver some research, the more likely provenance captured is of a low quality (incomplete and unstructured). The relationship between time constraints and quality of provenance captured, is a potential subject for future research.

The value of provenance is realised in the future

The third factor that contributes to provenance capture being a secondary consideration is that the value of provenance is realised (at some point) in the future. So at any given time, when provenance is being captured, the potential benefits of the provenance always lie in the future, maybe even after the immediate value of the associated data has been realised (i.e. after initial publication of the data).

> "I can definitely see benefits after the event of capturing provenance"

> "[Provenance is] certainly useful when you go back to something a few months or few years later"

It is also entirely possible that for any given item of provenance, no benefits will ever be realised as no situation where someone becomes interested in the item of provenance in question will ever occur. At the time of provenance capture it is impossible to predict if the provenance being captured will be of use or not, although the likelihood of the provenance being reused can be estimated (using some mix of experience and intuition). So when considering provenance capture, a researcher will weigh up the uncertain value at some time in the future of the provenance against the time taken to record it. Current methods of provenance capture, e.g. the laboratory notebook, require significant effort to record extensive provenance and minimal effort to record minimal provenance. This situation tips the balance in favour getting on with research and recording the minimal provenance set (i.e. the minimum a researcher believes they can get away with).

10.2.2.3 Implications of Provenance Capture as a Secondary Consideration

This sub-section considers the implications of the provenance being perceived as a secondary consideration. Two key implications are addressed in turn below.

Adopting a holistic approach to provenance capture

As discussed above at the time of capture provenance has little value to the researcher, but at some point in the future a given item of provenance may have significant value to either the researcher themselves or to an interested third party. Attempting to determine which items of provenance will be valuable seems to be a very tricky proposition, it is also the approach that researchers seem to take in current practice, where they record a minimal sub-set of provenance (presumably) based on what they believe is likely to valuable in the future. This approach does not seem to succeed, see the drawbacks researchers experience with current practice (Section 10.2.3.1). These findings imply that the ELN design must adopt a holistic approach to provenance capture (i.e. capturing as much provenance as possible), rather than focussing on a limited subset of activities (i.e. provenance for mechanism development) as in ELN design to date.

Design approach

The perception that provenance capture is a secondary consideration, places a premium on capturing provenance without requiring any input for the ELN user. So this perception validates two key elements of the design approach.

- First, the high level goal of the ELN development to: *Enable the capture of provenance for the process of developing models using the MCM, whilst minimising the burden on ELN users.*
- Secondly, one of the design principles: *to automate provenance capture where possible.*

10.2.3 Reflections on Current Provenance Capture Practices

This sub-section reviews the evaluation results that relate to the evaluators' experience of current provenance capture practices. The evaluators recognised the deficiencies of the laboratory notebook, as presented in the problem scenario for provenance capture, and their dialogue centred on the consequences of these deficiencies. So whilst the scenario talks about the provenance captured being incomplete, structured in an *ad hoc* manner and time consuming to record, the evaluators talked about difficulties interpreting laboratory notebook provenance (i.e. a consequence of incompleteness and the *ad hoc* structure) and how provenance degrades over time (i.e. a consequence of incompleteness and the associated reliance on the tacit knowledge of the researcher).

This divergence in the dialogue was not anticipated but can be understood in light of the discussion above. An individual item of provenance has no value at the time of capture and uncertain value in the future, so the deficiencies of the laboratory notebook at the time of provenance capture are not issues for the user. Issues only become apparent to the user at some later date, when they experience the consequences of the deficiencies (e.g. difficulties interpreting in an incomplete provenance record).

10.2.3.1 Drawbacks of Current Practice

The evaluators identified three main drawbacks of current provenance capture practices. These drawbacks are presented below, followed by a discussion of the implications, for the design of the ELN, of this set of drawbacks.

Difficulty interpreting provenance records

The main drawback raised by evaluators was the difficulties experienced interpreting laboratory notebook provenance, when returning to a piece of work. This issue was referred to using emotive language (such as "trawl through" and "nightmare"), showing that there is a real burden associated with returning to a piece of work.

> "if she hadn't written [extensive, logical provenance records] … we would have [just had] several models to trawl through which would be a nightmare"

The evaluators identified *ad hoc* provenance structure, as a cause of the difficulties experienced in interpreting provenance records.

> "[when] you haven't recorded … [provenance] … in a logical way it's difficult to find exactly what you were doing"

Reliance on tacit knowledge

The reliance of the laboratory notebook, as a means of provenance capture and storage, on the tacit knowledge of the laboratory notebook owner was also stated as an issue. With the difficulty in recalling the tacit knowledge, at some later date, required to supplement the laboratory notebook contents adding to the challenges of interpreting the provenance.

> "You will forget … [the details required to supplement laboratory notebook provenance] … and this is a problem I am having today"

In the case where someone other than the laboratory notebook owner attempts to interpret provenance contained with the laboratory notebook the issue of a reliance on tacit knowledge is even more significant. The reader of the laboratory notebook lacks the contextual information (held as tacit knowledge by the laboratory notebook owner), required to make full sense of the provenance record in question.

Time consuming

The evaluators found current provenance capture practices to be time consuming. The quotation below refers to a former member of the modelling group, who rather than use her laboratory notebook used word processor documents to record the model development provenance for the EXACT chamber modelling campaign (as described in Chapter 7). The benefit of using word processor documents as a means of provenance recording is that

subsets of model parameters could be copied, pasted and annotated (without the need to hand write large subsets of model parameters). Despite this it still took up time, which could have otherwise been spent focusing on conducting research.

> "Claire [, a former member of the Leeds modelling group,] has written all her changes in a word document, which takes time"

Implications

The drawbacks, of current working practices, identified above have two key implications for the design of the ELN, discussed below.

Avoid reliance on tacit knowledge

Researchers have a tendency to rely on the capture of some provenance as tacit knowledge, this has the benefit of having no cost at the time of provenance capture. The clear drawback of tacit knowledge as a means of provenance storage is that people forget. This implies that, where possible, the ELN should capture provenance, currently stored as tacit knowledge, automatically (as again this has no cost to the researcher at the time of provenance capture, without the drawbacks of the current practice). Where automatic provenance capture is not possible, researchers can be encouraged to avoid a reliance on tacit knowledge by the ELN prompts for inline annotation (as described in Chapter 8). The issue of what researchers typically record using physical artefacts and what they rely on tacit knowledge for will require further investigation.

Structure provenance to facilitate interpretation

Current practice for provenance capture relies on *ad hoc*, unstructured provenance formats, this causes issues with regards to returning to provenance records to interpret them. This supports the use of the ELN ontology to structure the provenance captured by the ELN, but also brings up the question of how the ontology should be maintained and developed once the ELN is used within the community.

10.2.3.2 Benefits of Current Working Practices

Having considered the drawbacks of current provenance capture practices, their benefits are now addressed. The key benefit of current provenance capture practices, particularly using a laboratory notebook, is the flexibility allowed in terms of what provenance to

capture, and, how and when to do so. This benefit and its implications, for ELN design, are discussed in further detail below.

Flexibility

In contrast to the drawbacks, of current practices for capturing provenance, the benefits were attributed at the time of provenance capture. The main benefit discussed was the flexibility of the laboratory notebook as a provenance capture medium:

> "in your lab book you can write it straight down"
>
> "in the lab book you can write what you like"

In something of paradox, the flexibility quoted here as a benefit can be related to the drawbacks of current provenance capture practices (i.e. incomplete provenance records, *ad hoc* provenance structure). It can be seen that the flexibility of the laboratory notebook, enables the researcher to apply minimal effort to provenance capture.

> "[You can] just get on with your modelling without writing things down."

Again this can be seen as a result of provenance capture's secondary status (to 'the Science') and the idea that provenance capture can even be seen as a burden (rather than an integral part of conducting 'the Science').

Implications

The positive response to the flexibility of the laboratory notebook has two implications for the design of the ELN, are discussed below.

The ELN design must encourage users to spend time on provenance capture

The benefits of the laboratory notebook, as a means of provenance capture, centre on its ability to enable the researcher to capture a minimal set of provenance, so minimise the time spent on provenance rather than conducting research. This has a number of implications, including:

- The ELN should automate provenance capture where possible;
- The ELN interface should be optimised for speed of entry of provenance;
- The ELN interface should be optimised to minimise the learning curve;

- This issue of how to shift provenance capture from a secondary consideration to an integral part of the research process, needs to be addressed.

The ELN design must balance requirements for flexibility and structure

The dialogue of the researchers highlights the importance of flexibility in a provenance capture medium. The implications for how requirements for flexibility and structure should be balanced remains unclear, questions arising include:

- Should the user be able to turn off provenance capture, when they don't want to use it?
- Should the user be able to turn off the annotation prompts, but leave on the automated provenance capture functionality (running in the background, then they can get on with conducting research without interruption)?
- Is richer annotation (beyond text) required?

Providing this flexibility could be detrimental, in some cases, to the quality of provenance captured by the ELN and is likely to reinforce the message that provenance is a secondary consideration (rather than an integral part of the research process).

10.2.4 Response to Envisioned Provenance Capture Practices

The next stage of the evaluation introduced the concept of an ELN, using a high-level descriptive scenario (the activity design scenario introduced in chapter 8), to gauge the response of the evaluator to the concept on an ELN. This introduction set the scene for the later stages of the evaluation where the ELN prototype was demonstrated and subject to hands-on user tests. No significant alterations to the scenario were suggested during the evaluations, so this section proceeds to discuss the perceived drawbacks and benefits of the ELN described in the activity design scenario. Again the evaluators' dialogue centred on the consequences of features of the provenance captured.

10.2.4.1 Drawbacks of Envisioned Working Practices

The main drawback of the ELN, as described in the activity design scenario, was identified by the evaluators. This drawback was that making annotations, recording scientific reasoning, would require too much effort. This drawback and its implications are addressed in detail below.

Making annotations will require too much effort

The drawback raised by the evaluators was that the effort required to make annotation may deter them from using an ELN. This drawback can be seen to be a related to the general issues of provenance being a secondary consideration and the impact of time constraints on provenance capture (as discussed in Section 10.2.2).

> "Again, it will be the time issue. In your lab book you can write wherever you want. … [With the ELN] she is prompted to annotate the process, in your lab book you can write it straight down. This is going to take time to go through the different protocol steps."

The quote above indicates a concern about a lack of flexibility of the ELN, in comparison to the laboratory notebook, that there may be insufficient scope to record or omit items of provenance. Concern was also expressed with regard to the necessity or desirability of capturing all the provenance. For example, in response to the question concerning what cases would it be preferable not record provenance:

> "If you are doing something straight forward where you don't change a lot."

Implications

There are two main implications of the evaluators' perception that making annotation will require too much effort; each of these implications are described below.

The ELN should be designed for speed of provenance input

The concerns raised with regard to the time taken to use the ELN can be mitigated to some extend by optimising the interface design for speed of data entry, but writing something down (in an ELN prompt or in a laboratory notebook) is always going to take more time than not writing it down.

The ELN must balance requirements for flexibility and structure

The evaluators' concerns regarding a lack of the flexibility of the ELN should be addressed in the redesign of the ELN. Striking a balance between the flexibility and structure of provenance capture will be critical, as they are competing factors both identified as desirable by the evaluators.

10.2.4.2 Benefits of Envisioned Working Practices

The evaluators perceived five main benefits of adopting the ELN as a means of capturing and managing provenance, based upon the activity design scenario's description of envisioned provenance capture practices using the ELN. Each of these benefits is discussed in detail below, followed by a discussion of the implications, for ELN design, of these benefits.

Provenance captured by the ELN will be easily interpretable

The ELN description was perceived to capture provenance that was easily interpretable, in contrast to the provenance captured using current practices:

> "You can see exactly what you have done, whereas before you had to rifle through various lab books to find out exactly what you had done"

The evaluators were able to envision cases where other researchers would want to review their data and provenance, and cases where they would want to review the data and provenance of other researchers.

> "if somebody else wants to look at … [your work] … then they know exactly what you have done and exactly where you have been and where to go next"

Provenance captured by the ELN will be stable over time

Again in contrast to a drawback of the current provenance capture practice, where provenance was perceived to degrade over time, a benefit of the ELN was perceived to be the provision of provenance that is stable over time. This stability is a result of using a fixed, well defined structure when capturing and representing provenance.

> "If you are doing a PHD you could … [capture provenance for your model development research] on a field work campaign and come back to it [when you write your thesis, back in the office]."

Provenance captured by the ELN will be complete

The activity design scenario suggested that the ELN would capture complete provenance, but the scenario was deliberately vague in its definition of completeness:

> *The ELN ensures that the provenance recorded is complete; providing sufficient detail to fully recreate a given experiment.*

This enabled the evaluator to create their own definition of completeness, in the context of a provenance use scenario from their own experience. So it is no surprise that the evaluators identified the completeness of the provenance captured by the ELN as a benefit.

> "I think not only does it (the ELN) capture everything, if …"

The risk with allowing the user to use a personal definition of completeness, it that what an individual believes is sufficient detail to recreate a given experiment may vary greatly, due to the provenance use scenario they envision, their personal experiences and views etc.

Provenance captured by the ELN will be structured

The evaluators responded very positively to the use of a fixed structure for capturing and representing provenance. The evaluators viewed this structure as making provenance records easier to interpret as the reader can become familiar with the format and the semantics of the provenance presented.

> "I think it [, ELN captured provenance,] will be much better, more ordered."

Using the ELN will add structure to the modelling process itself

An unanticipated benefit of the ELN identified by the evaluators was that encouraging provenance capture would help researchers to structure their modelling process. By prompting users to provide annotations including their scientific reasoning, literature references and justifications for a given course of action it encourages the user to take a more structured, logical approach to their modelling.

> "[the ELN] sets your mind into a certain way of processing data – it can focus your mind more on what you are doing by providing more of a framework. It's an interesting way at looking at it."

> "[the ELN] will prompt you to change it [, the chemistry in the model,] in an iterative, more logical order therefore making your brain think in a more scientific way as well. Therefore speeding up the process."

Implications

The evaluators' perceptions of the benefits of using the ELN (as described in the activity design scenario) have three main implications for the ELN design; described below.

Adopt a holistic approach to provenance capture

One evaluator referred to the possibility of the provenance capture discipline enforced by the ELN leading to a more structured, logical modelling process. Being prompted by the ELN to provide justifications and annotations to record their scientific reasoning; the researcher is also prompted to think a little more rigorously about the process they are constructing and executing. This implies that there is an opportunity, whilst developing the ELN, to re-engineer the modelling processing and modelling tools, to produce more rigorous, structured working practices. Again this supports adoption of a holistic approach to provenance capture, encompassing reengineering the modelling process in conjunction with the development of the ELN.

Design the ELN for provenance use by multiple stakeholders

One of the benefits of the ELN referred to above is that provenance records could be used by individuals other than the data/provenance owner. One example of this use of provenance by third parties, examined in detail in chapter 6, is where the MCM developers aggregate and reviewed provenance from across the MCM user community, in order to inform the MCM development process. Supporting the use of provenance by third parties has a number of implications for the presentation of provenance, including two listed below.

- Users should be able to customise their view of the provenance records according to their individual requirements.
- Users should be able to access to ELN archives remotely. This raises questions including; what are the interface, security and infrastructure implications for remote access?

The ontology requires direct evaluation

The structure applied to provenance captured by the ELN was regarded as one of the key benefits of the ELN. This structure, in the form of the ontology, has not been validated directly or explored by users (to date). Given the importance of the ontology for the success or otherwise of the ELN there is a need to evaluate the ontology with members of the community. The evaluation of the ELN ontology will be considered in greater detail in the final section of this chapter.

10.2.5 Response to ELN Prototype

Having introduced the evaluators to the concept of the ELN with the activity design scenario, earlier in the evaluation, the evaluation proceeded to provide a demonstration of the prototype ELN, followed by hands-on user tests. This gave the opportunity to test how well the prototype fulfilled the high-level requirements stated in the scenario and to assess if the prototype implementation was able to address any of the concerns raised when discussing the scenario. This sub-section describes the response of the evaluators to the ELN prototype in terms of the perceived drawbacks and benefits of using the ELN.

10.2.5.1 Drawbacks of the ELN Prototype

The evaluators identified a number of drawbacks, based on their experience: watching the ELN demonstration; and using the ELN first-hand. Some of these drawbacks has been discussed in the evaluation results above, but are discussed again in this sub-section as the evaluators' comments enable further understanding of the issues to be developed. Five drawbacks are identified and discussed below, followed by a discussion of the implications of these drawbacks.

The ELN lacks flexibility

One of the main drawbacks raised during the discussion of the ELN scenario, was again raised in discussion of the prototype. The issue that re-emerged was the ELN being insufficiently flexible to capture the diverse annotations the researcher wants to capture.

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

Having developed the annotation interface based on the requirements of a small group, this quote suggests that effort needs to be made to understand the diverse requirements of community members. Considering the issue of the flexibility of the annotation interface also raises the question of the degree to which the functionality of the ELN should be developed to enable user customisation/personalisation of the ELN.

The ELN captures too much provenance

The second issue, that re-emerged during the prototype testing, was the idea of capturing too much provenance.

> "If the model falls over, do you end up with lots of annotations for updating the model? … That would probably [be] the less useful aspect of … [the ELN], if it records every time and you end up with lots of stuff that you don't necessarily need a record of."

The concept of capturing too much provenance is tricky to address, as it is difficult for the ELN to determine if an individual item of provenance should not be recorded (or not drive a prompt for annotation). The case discussed in the quote related to a model failing to execute successfully (due to some error in the parameter configuration) and the subsequent actions of the modeller seeking to correct the error in the parameter configuration. The evaluator perceived provenance on this workflow to be of no value, that is not say that some other individual would view this as potentially valuable information worth archiving.

The ELN learning curve is too steep

Concerns were also expressed about the learning curve that an ELN user would experience when first using the ELN within their day-to-day work.

> "Initially, getting to know the system [would be a challenge] because in the lab book you can write what you like"

In the quotation above the flexibility of the lab-book is reiterated, suggesting the learning curve relates to the process of structured, coherent provenance capture (rather than the tool interface). It is this change of working practices from ad-hoc, unstructured provenance capture to a structured, coherent provenance capture that represents the main challenge to users. The issues of usability and the ease of learning system functionality of course remain, and need, to be addressed in the form of efforts to improve the ELN user interface.

Low quality provenance of the analysis processes

The provenance captured for the analysis phase of the modelling process is of a lower quality than the provenance capture for the model development and model execution phases. This is a result of the mechanism development and model execution phases being well understood (and reliant on a limited number of operators), whereas the analysis phase is much more flexible and tied to the working practices of a given researcher. This led to the development of a generic provenance capture interface for the analysis process,

providing functionality to capture user annotations, the data sources used, and the locations of relevant analysis documents. The evaluators identified the main drawback, with this generic interface for the analysis activity provenance capture, as the interface not being tailored to the specific analysis process performed by the user.

> "It would be nice to be prompted when you are doing analysis; I think that would be difficult or impossible."

Getting full value from ELN requires community engagement

The other main drawback highlighted by the evaluation was that for many of the benefits of the ELN to be realised, the community as a whole needs to engage with, and use, the ELN.

> "You need to get everyone using it at the same time to get the most use out of it"

> "I think the … [as the ELN is adopted across the community] more people will use it. [Being able] to access other people's records would encourage you to do it (ELN provenance capture) yourself."

It is not clear what proportion of the value of the ELN (to a given individual) is related to the adoption of the ELN across the wider community, because the quotes above were captured in general discussion about the prototype and not probed further during the interview. So whilst these comments above can be seen as a drawback in terms of the difficulty of realising the full value of the ELN and understanding how best to tailor the ELN to encourage adoption across the community, they also show that provenance can be of value across the community. The sharing and use of model development provenance across the community represents an opportunity to enhance existing working practices and enable new processes such as aggregating provenance for *in silico* experiments to support development of the MCM (as described in Chapter 6).

Implications

In the discussion above the drawbacks of the ELN prototype are outlined, these drawbacks have a number of implications for the design of the ELN. These implications have been grouped together and are discussed below.

The ELN design needs to balance requirements for structure and flexibility

In order to provide the flexibility (i.e. the ability for an individual to write what they want), that the evaluators feel is missing from the ELN prototype, the ELN needs to be developed to:

- Allow freeform annotations and other digital objects (graphs etc.) to be attached to the process provenance (in addition to process specific prompts);
- Provide annotation prompt interfaces that suit the preferences of the individual user, i.e. allow the user to customise their prompt interface to the level of detail they require.

The need to prompt the ELN user is context dependent

The evaluator perception that the ELN captures too much provenance maybe related to the user being prompted to record annotations that they see no need to make, thus interrupting the modelling process. This is a significant issue as, if the ELN is seen to get in the way of conducting research, by prompting at inappropriate times, then this would discourage potential users from adopting the ELN. This concern about too much provenance could also be related to difficulty in navigating large provenance reports and the representation of "less useful" provenance. Again this is an interesting issue to explore, in terms of what provenance is useful for inclusion in summary-level provenance reports and what provenance should be available only by drilling-down through summary reports.

The quality of provenance for data analysis processes should be improved

The weakness of provenance captured by the ELN for data analysis processes can be seen as a result of not having adopted a holistic approach to provenance capture, in the prototype development to date. Adopting a holistic approach to provenance capture would improve the quality of provenance for data analysis processes. This will require automatic capture of this provenance, which in turn requires the automation of a number of manual data analysis tasks.

The processes for encouraging the adoption of the ELN must be considered

As the value of the ELN to a given individual is partially dependent on other researchers across the community using the ELN and making their ELN archives available to view, the initial ELN implementations must be carefully selected to ensure that this value can

come to the fore. Also individuals across the community must be sought, at operational and senior management levels, to champion ELN adoption.

10.2.5.2 Benefits of the ELN Prototype

This section discusses the benefits of ELN identified during the prototype demonstration and user testing. The evaluators found many of the benefits of ELN prototype were the same benefits they identified when considering the activity design scenario. This provided reassurance that the scenario and the prototype are well aligned. Given that some benefits had been discussed in detail previously, much of the discussion in this section of the evaluation was curtailed to avoid repetition. Three of the key benefits are outlined below, followed by a discussion of the implications of these benefits.

Prompts encourage good practice

During the design of the ELN the decision to drive annotation capture by prompting the user had caused some concerns including: users finding prompts annoying or an unwelcome interruption to getting on with conducting research, and that the prompts may not be sufficiently context sensitive to be useful. 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 analysis."

This supports the idea that prompting users will encourage them to adopt good practice in their provenance capture, being driven by the ELN to record their annotations more frequently and in a more structured manner than with a traditional laboratory notebook. It also offers some hope that by embedding the prompts within the modelling workflow, annotation and provenance capture can become part of business as usual operation (rather than a secondary concern).

The provenance captured by the ELN is well structured

The structured nature of the provenance captured by the ELN was also highlighted as a benefit of the prototype system. When compared to provenance capture in laboratory notebook (with an *ad hoc* structure) the ELN provides:

> "A clearer record of what you have done, what you have changed, what input you have used and what output belongs [with what model version]."

Provenance captured by the ELN will be usable by third parties

The evaluators also noted that the provenance captured by the ELN would be useful not only to themselves, but also to other researchers seeking to interpret their data:

"[The ELN provides] general clarity for looking back or for someone trying to figure out somebody else's work."

Again it is interesting that the evaluator has referred to use of provenance by another member of their community, but without providing any reason for this community interaction. This could merely be because, at the time of interview, this point was not probed further, or potentially it could be due the evaluators' lack of clarity about the community interaction (they just have a feeling it would be useful).

Implications

The benefits of ELN use, outlined immediately above, have three keys implications; outlined below.

Adopt a holistic approach to provenance capture

The positive response to being prompted to annotate the scientific process as it takes place and the way in which this could encourage good modelling practice, suggest it would be worth exploring the use of prompts for annotation across the modelling process. This course of action would be in line with adopting a holistic approach to development of the ELN.

A direct evaluation of the ELN ontology is required

The ontology used by the ELN to structure provenance should be directly evaluated, as it is a critical component of the system design. This implication has arisen previously within this section, so is not expanded upon any further here.

The presentation of archived provenance requires further consideration

The use of provenance by individuals other than the data/provenance owner ELN functionality to compose customised provenance reports to meet the requirements of the provenance viewer.

10.2.6 Potential Improvements to Prototype ELN

Having discussed the evaluators' responses to the ELN prototype above, this sub-section discusses potential developments that would enhance the usability of the ELN prototype. These potential developments were identified and raised for discussion by the evaluators during the prototype demonstration and user testing. The improvements fall into three categories: first, improving the annotation functionality for the mechanism development process; secondly, improving the provenance capture for model constraint data; thirdly, improving the quality of provenance captured for data analysis processes.

10.2.6.1 Mechanism Development

The prototype ELN provides a single text field (i.e. minimal structuring) to enable annotation of changes to the chemical mechanism. This minimal structuring of the annotation prompts was used due to concerns about the burden placed on the user by the ELN prompts. Presenting just a single text field to the user was intended to provide a flexible means of annotation, that mimicked the traditional laboratory notebook. The feedback during the evaluation suggested that this minimal structuring of the annotation is not in line with the requirements of users. Two suggestions for providing more structure to the ELN annotations are discussed below.

Separate annotation fields for the scientific reasoning for changing a given reaction and an associated literature reference:

> "[It would be useful to have] Two text boxes, one says provide justification and one saying reference."

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

> "I think the reference …[,annotation field,]… would have to be optional because you don't always have a reference."

Currently, where a reaction has been edited in multiple ways, e.g. the products and the rate co-efficient have been changed, a single prompt is presented to the user. This prompt states the nature of the multiple changes that have taken place, but only allows the one annotation. In the evaluation it was suggested that functionality is required to allow separate annotations of the each of the multiple changes, so the change to the reaction's products can have a separate justification and reference to the change of the rate co-efficient.

> "it might have been useful if when you change the rate and the products and the reactants, you could perhaps put a separate annotation on each bit or the option to. Just in case you had different references."

Implications: Rather than adopt a minimal approach to the structuring of annotations, as to the ELN development to date (based on the assumption researchers would not like the prompts as they are a distraction from conducting research), more structure can be added to the annotation prompts to enable a finer grain of information.

10.2.6.2 Model Constraint Data Provenance

The prototype ELN was developed to capture provenance for a mechanism development based modelling process (i.e. changing the chemistry in the model). When discussing the prototype with a researcher, whose experience primarily lies in the area of field modelling, it quickly became clear that the configuration of model constraints was of comparable importance to mechanism development. Typically for field modelling a limited amount of time will be spent developing the mechanism (selecting a mechanism from the MCM, adding/editing/deleting reactions or sets of reactions) then time will be spent developing the data constraint set (obtaining data, adding constraints, updating data when the source data changes). So given the importance of the constraint data in field modelling the ELN needs to be developed to support provenance capture for developing constraint datasets.

> "the thing that would be most useful is … [annotation of the constraints for example the] source where the … [constraint] had come from and on what date"

An important point to address for the provenance of the constraint datasets, is that the constraint data can change due to re-interpretation of raw experimental data.

> "In an ideal world it wouldn't happen but these experimentalists will recalibrate and change everything [(i.e. the constraint dataset changes, so the model must be rerun with the latest data)]. Also, with the input you could, say you have the ozone data and there are three sets, it's all come from the BADC, and you might want to look up which one you have been using."

Implication: By adopting a holistic approach to provenance capture the acquisition and processing of model input data will come within scope for process re-engineering and provenance capture.

10.2.6.3 Data Analysis Provenance

As discussed above the provenance of the analysis phase of the model process is weaker than desirable. This issue of weak references from the provenance to an object within a researchers file system (e.g. analysis spreadsheets) could be partially addressed by allowing researchers to submit objects to the database, providing a fixed reference that can be used in the provenance, annotating the object and associating it with a given piece of scientific workflow.

> "you could [put] the spreadsheet in the database to associate it with a set of [model] runs and … [the ELN prompts] you comment. That would be the best way. I think the prompting is good."

The alternative approach to improving the provenance of the analysis phase of the modelling process is to consider the development of a set of provenance aware analysis tools. In the quote below the evaluator identifies the automatic plotting of some graphs, a simple form of analysis, as a way of speeding up the modelling process (by automating a manual task). This kind of automation of analysis processes would also enable automatic capture of process provenance.

> "Probably beyond this study, but if you just plotted … [the data sets] up and added upon the same time grid. That would be really useful to speed up the process."

FACSIMILE (discussed in Chapter 2), a modelling system often used in conjunction with the MCM, offers some automated data analysis functionality. This functionality is not

typically used by researchers, and the limitations of this functionality could be examined in order to gain an understanding of the type of data analysis tools that are required.

Implication: By adopting a holistic approach to provenance capture data analysis will processes come within scope for process re-engineering and automatic provenance capture.

10.2.6.4 Improving Provenance Reports

The provenance report functionality developed for the ELN prototype was intended to provide a user-accessible representation of the provenance captured during the demonstration and to show the provenance at varying levels of abstraction. The evaluation of these reports was fairly light touch, but one major potential improvement was identified in the form of a complementary mechanism provenance report.

> "What would be useful as well at the end is if you had the mechanism that you ran … [including] … all the … [reactions] you have taken out or changed."

The evaluator went on to describe a potential colour coding scheme for the mechanism provenance report. The purpose of the mechanism provenance report was established, through discussion with the evaluators, to be providing a complete record of a mechanism and its evolution within a single, easily interpretable report. Developing such a report requires further requirements capture and presents a number of challenges including: representing the history of a given reaction and showing the chronology of development.

Implications

A wider variety of provenance reports are required (beyond the simple chronological reports currently available). Possible reports including: reports for a set of processes (e.g. a report detailing all the data analysis that took place); and customised, user-composed provenance reports.

10.3 Implications for Prototype Design

The section gathers together the implications for the ELN design, identified during the discussion of the evaluation results, and suggests actions to be taken as a result of the ELN evaluation. Related implications have been gathered together and addressed under a single

heading to simplify the presentation of this content. These implications need to be addressed to ensure the successful transition of the ELN from a prototype, research project output, to a production quality tool for use by the MCM-user community. The identification and discussion of these implications is the topic of the remainder of this chapter. Addressing these implications and implementing a production quality ELN are beyond the scope of my thesis (and remain a topic for future work). This section begins with a discussion of the implications of adopting a holistic approach to provenance capture; followed by a discussion of the need to balance the flexibility of the ELN against the structure that the ELN adds to the modelling process; the need to directly evaluate the ELN ontology is then discussed; and the section concludes with a discussion of the issues associated with the adoption of the ELN across the MCM user community.

10.3.1 Adopting a Holistic Approach to Design of the ELN

During the discussion of the evaluation results the most frequently cited implication, for the ELN design, was that a holistic approach to the design of the ELN should be adopted.

10.3.1.1 Revisiting the Design Approach

In the light of the evaluation results presented above and the need to adopt a more holistic approach to the design of the ELN, this sub-section revisits and revises the ELN design approach (as introduced in chapter 8). The design approach consists of three components: the design goal; the design scope; and the design principles; each of these components are revisited and revised in turn below.

Design goal

The design goal is now revisited: first, the original design goal is presented, and secondly, the implications of the evaluation results are discussed.

Original design goal

> *Enable the capture of provenance for the process of developing models using the MCM, whilst minimising the burden on ELN users. The provenance capture should be, where possible, complete. Complete provenance is defined as the*

provenance required to re-implement a given model, recreate a given dataset and understand why the model was implemented in a given way.

Implications of evaluation results

The evaluation results support the design goal in its original form, so there is no need to revise the design goal. Relevant evaluation results are discussed below.

Enable the capture of provenance for the process of developing models using the MCM.

The evaluators saw significant value in capturing provenance in general and using the ELN to capture a higher quality of provenance than possible using current working practices.

Whilst minimising the burden on ELN users.

The evaluators saw provenance capture very much as a secondary consideration, with conducting research their primary concern. So it follows the ELN should seek to minimise the effort that the modeller must invest in provenance capture, particularly for process provenance (which can generally be captured automatically). When capturing scientific reasoning, in the form of annotations, it is essential to engage the modeller in the process, as scientific reasoning cannot be captured automatically.

The provenance captured should be, where possible, complete.

The evaluators responded positively to the idea of seeking to capture complete provenance, but suggested that this would not be possible for all elements of their modelling process (e.g. data analysis due to the use of *ad hoc* processes and proprietary software).

Design scope

The scope of the ELN design is now revisited in light of the evaluations results: first, each of the original scoping statements is discussed; and secondly a revised set of scoping statements are presented.

Original scoping statements

- *Retain existing model development processes and tools:* Adopting a holistic approach to the design of the ELN renders this scoping statement obsolete. Rather than retaining existing tools and process, the scope of the ELN design should be

extended to allow ELN users to benefit from the re-engineering of existing tools and processes to facilitate provenance capture and improve the user experience.

- *Address the capture of provenance for model development:* Again the scope of the ELN design needs extending, in this case to include the querying and presentation of provenance (rather than just provenance capture and representation as considered for the ELN prototype).
- *Address the most frequently occurring modes of core activities:* Having explored the capture of provenance for mechanism development, model execution and data comparison; adopting a holistic approach to provenance requires the scope of the ELN design to be extended. Provenance capture for other commonly occurring modelling activities (such as editing model constraints, or rate of production and loss and analysis) must also be considered, along with methods of recording provenance for *ad hoc* modelling activities.

Revised Scoping Statements

Having considered the original scoping statements above, a revised set of scoping statements are presented immediately below.

- *Seek opportunities to reengineer existing processes and tools:* During the design of the ELN, opportunities should be sought to reengineer and enhance existing modelling tools and processes to facilitate provenance capture and improve the user experience.
- *Consider the full provenance lifecycle:* The ELN design should encompass functionality to enable both provenance capture and subsequent provenance use. Provenance use should be considered from the perspectives of multiple stakeholders across the MCM-user community.
- *Enable provenance capture for standard and non-standard modelling activities:* The ELN should capture provenance for standard (frequently occurring) modelling activities. Functionality should also be provided to capture provenance for one off, *ad hoc* activities.

By adopting a holistic approach to developing the ELN the scope of the problem space has expanded considerably. Figure 10.2 shows the scope for the initial prototype of the ELN, focussing on the core modelling workflow (model development, model execution, data

analysis) and only considers one mode of model development (developing the chemical mechanism) and one mode of data analysis (comparing two data sources by plotting graphs using Microsoft Excel). Figure 10.3 shows the design scope for a production quality ELN. Taking a holistic approach expands the problem space to include the acquisition, processing and local storage of data consumed by the core modelling workflow, and multiple modes of model development and analysis. Capturing provenance across this expanded problem space requires an understanding of all of these processes, and where possible manual processes to be automated (to facilitate provenance capture and improve the user experience).

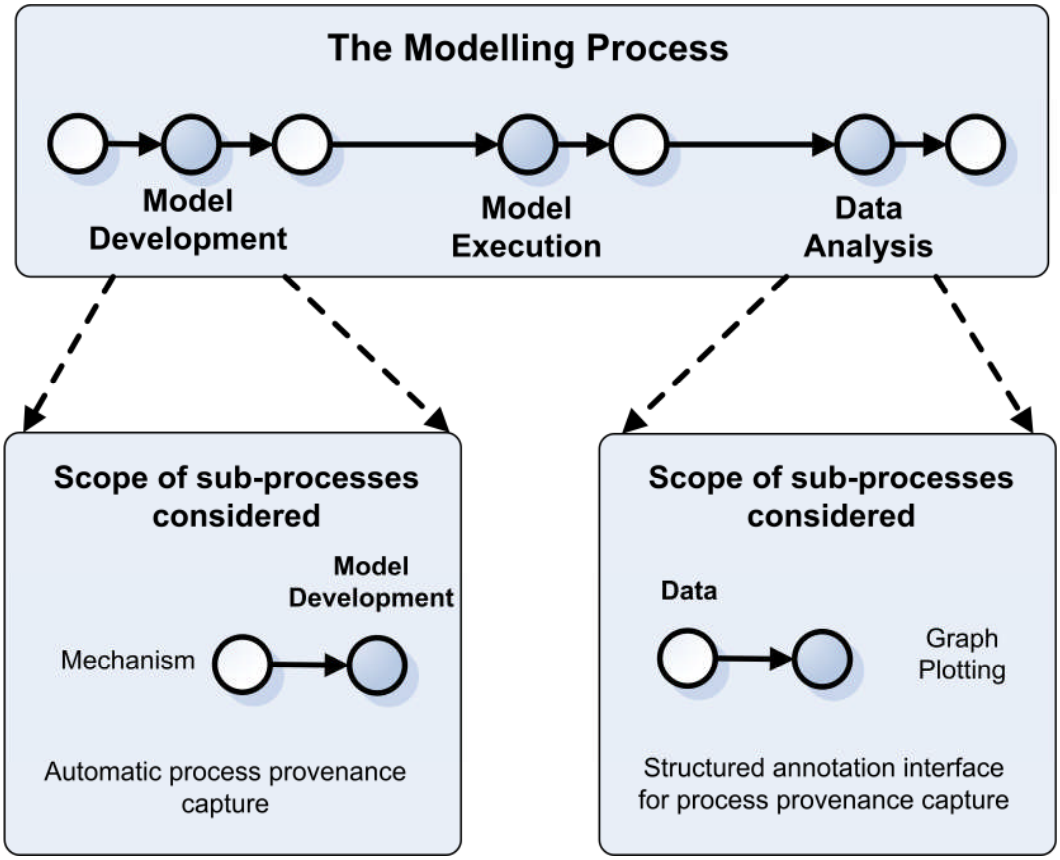


Figure 10.2: This figure shows the design scope adopted for the development of the prototype ELN. The core modelling activities (model development, model execution, and data analysis) are shown. The modes of the core activities considered are shown below, with mechanism development and graph plotting to compare data sources being the only modes considered.

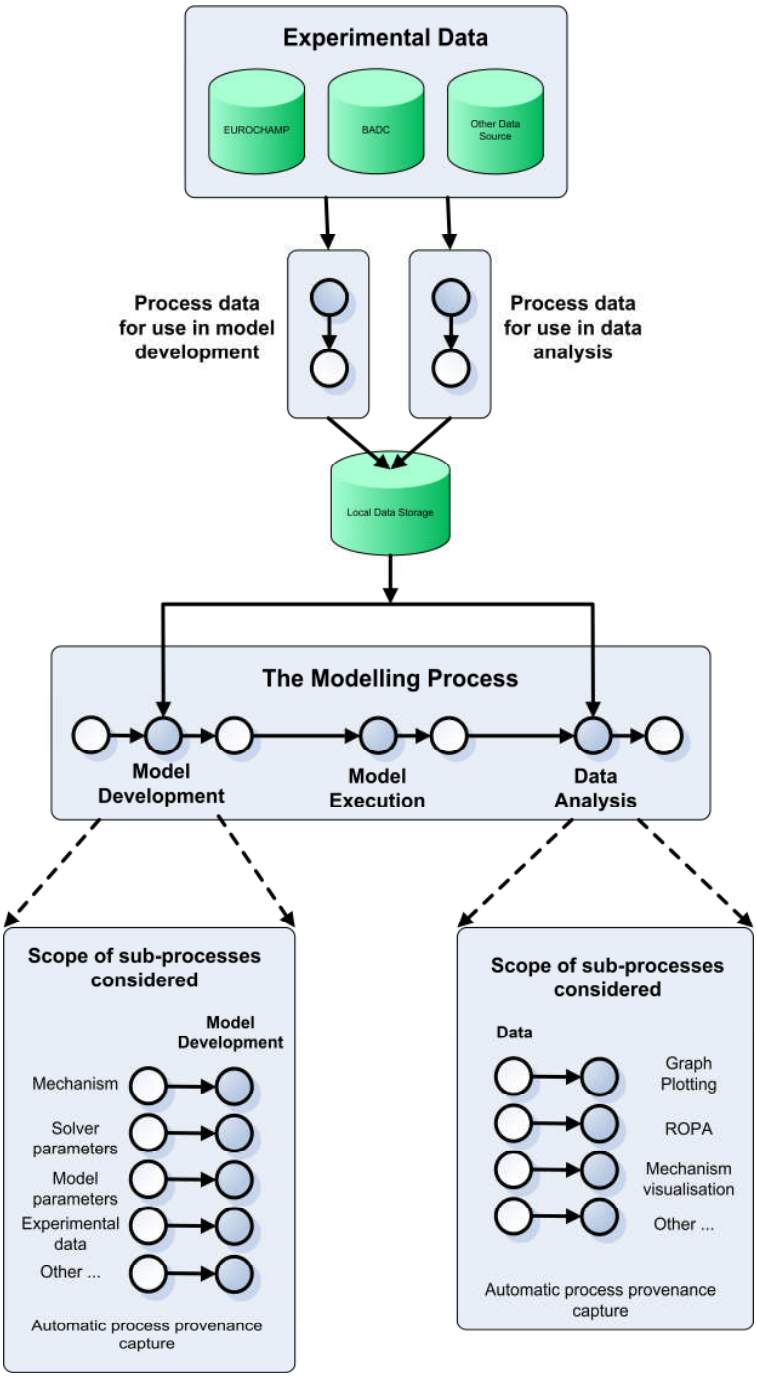


Figure 10.3: This figure shows the design scope revised in the light of the evaluation results. The core modelling activities (model development, model execution, and data analysis) are shown, with experimental data feeding into these core activities (an addition to the original design scope). The modes of the core activities considered are shown at the bottom of the diagram; a more extensive set of modes is presented (as a result of the findings of the evaluation).

Design principles

Having a presented a revised design scope, above, the design principles adopted during the development of ELN are now revisited, in light of the evaluation results: first, the original design principles adopted are re-iterated; followed by a discussion of the implications of the evaluation results.

Original design principles

- Where possible provenance will be captured automatically;
- Provenance captured will be represented, stored and queried using the terminology of the atmospheric chemistry domain;
- Provenance will be captured for both human and computational processes;

- Equal importance will be placed on capturing process provenance (generally automatically) and scientific reasoning (requiring the ELN user to record their scientific reasoning);
- Provenance will be captured with respect to two frames of reference: first, the scientific process; secondly, the *in silico* experiment;
- Provenance capture will be interleaved with the modelling process;
- Minimise the learning curve for the ELN (as a tool) and provenance capture (as a process).

Comments

The design principles were, in general, supported by the evaluation results. Some principles were directly validated by the evaluation results. For example “where possible provenance will be captured automatically”, was validated by results showing provenance is a secondary consideration for modellers. Other principles were not directly addressed by the evaluation but were indirectly validated, as contributing factors to overall favourable response to the ELN. For example the use of scientific terminology in the representation of provenance was not directly evaluated, as the evaluators found it difficult to relate to a concept of generic representations of provenance for their scientific processes, but the evaluators were positive about the provenance reports presented.

This sub-section has revisited and revised the ELN design approach, in light of the evaluation results. The discussion focused on updating the design approach to align with a

holistic approach to provenance capture and use. The following sub-sections continue to address the other key implications of the evaluation results.

10.3.2 Balancing Flexibility and Structure

The second key implication of the evaluation results was that the ELN design has not struck the right balance between flexibility and structure. The evaluation results showed that the flexibility of the laboratory notebook was a key benefit of its use, i.e. the user can write what they think is important. The flexibility of the laboratory notebook was also identified as a drawback, leading to a lack of structure and difficulties in interpreting provenance records. The evaluators appreciated the structure applied, by the ELN, to the provenance capture process in terms of: annotation prompts encourage them to consider and record their scientific reasoning; and the format of provenance captured by the ELN aiding the interpretation of provenance records. The evaluators had concerns about the lack of flexibility of the ELN, i.e. the ELN is not tailored to their individual working practices and preferences.

Given these ambivalent responses to both flexible (i.e. the traditional laboratory notebook) and structured (i.e. the ELN) tools, it becomes difficult to balance requirements for flexibility and structure. The logical approach seems to be to enable the ELN user to select or configure the ELN to provide the best mix of flexibility and structure according to the user's personal context (i.e. the type of research they are conducting, the user's personal preferences, the user's level of experience, etc.). An example of how the ELN could be developed to allow the user to configure their own mix of structure and flexibility is presented below.

Annotation strategies

The ELN could be developed to allow the user to select their own annotation strategy. Here an annotation strategy is defined as the detail and scope of the annotations prompted for by the ELN. Three example annotation strategies are presented below.

- ***Annotation-full:*** The ELN user is prompted for all changes made during model development processes, all model execution processes and all data analysis processes. The prompts are structured to capture scientific reasoning in detail (i.e. prompts present separate fields for each component of their reasoning). For

example after adding a reaction to the mechanism the user is prompted to record why they added the reaction and a reference for the reaction added.

- ***Annotation-standard:*** The ELN user is prompted for all changes made during model development processes, all model execution processes and all data analysis processes. Prompts are minimally structured (i.e. a single field is presented to capture all scientific reasoning). This strategy roughly corresponds to the strategy employed in the design of the ELN prototype.
- ***Annotation-lite:*** As a default the ELN user is prompted for all changes made during model development processes, all model execution processes and all data analysis processes. The ELN user is able to configure which prompts they want to see and how much structure they want the prompts to provide. For example the ELN user could select only to be prompted for changes to the chemical mechanism, with minimal structuring. Alternatively, the ELN user could select not to be prompted at all, in this case only process provenance captured automatically would be recorded.

10.3.3 Direct Evaluation of the Ontology

The third key implication of the evaluation results was that the ontology itself should be directly evaluated. The ontology plays a crucial role in structuring the provenance captured by the ELN; and the structure of the provenance captured by the ELN was identified as one of the key benefits of adopting the ELN. To date the ELN ontology has only been evaluated indirectly, by presenting provenance reports (which use the same terminology as the ontology and roughly the same structure) to the evaluators. The most appropriate means of evaluating the ontology would be through inspection by one or more domain experts (e.g. researchers with extensive experience of developing models using the MCM). In order to overcome the difficulties that a chemist may have in understanding the ontology, a "guided tour" of the ontology would need to precede the evaluation of the ontology.

10.3.4 Adoption

The evaluation results drew attention to the need to consider how adoption of the ELN across the MCM user community can be achieved. The quality and usability of the ELN software will play a significant role in ensuring adoption, but there are a wide variety of

other factors that will play a role in encouraging members of the MCM user community to adopt the ELN. These factors include: the role of senior members of the research community in encouraging ELN use; the role of earlier adopters; and the network effects created by the ability to explore provenance archives across the MCM user-community. Addressing these issues is beyond the scope of the research presented in this thesis, but I offer some thoughts and predictions in the final chapter of this thesis.

Chapter Summary

This chapter has presented an overview of the results the evaluation of the ELN. I chose to perform an in-depth evaluation with a small evaluation panel (of two people), in order to understand how to develop an ELN that meets the needs of a small group of local users. This approach is based upon the belief that from this position it will be a relatively straightforward task to extend the ELN to meet the needs of the wider user community. The evaluation results were presented and analysed to determine their implications for the design of the ELN. The chapter concluded with presentation of a revised ELN design approach, informed by these implications.

References

1. Scriven, M., *Types of Evaluation and Types of Evaluator.* American Journal of Evaluation, 1996. **17**(2): p. 151-161.
2. Sommariva, R., et al., *OH and HO2 chemistry in clean marine air during SOAPEX-2.* Atmos. Chem. Phys., 2004. **4**(3): p. 839-856.
3. Hull, D., et al., *Taverna: a tool for building and running workflows of services.* Nucl. Acids Res., 2006. **34**(suppl_2): p. W729-732.
4. Oinn, T., et al., *Taverna: lessons in creating a workflow environment for the life sciences.* Concurrency and Computation: Practice and Experience, 2006. **18**(10): p. 1067-1100.
5. De Roure, D. and C. Goble, *myExperiment: A Web 2.0 Virtual Research Environment,* in *International Workshop on Virtual Research Environments and Collaborative Work Environments*. 2007: Edinburgh, UK.
6. De Roure, D., C. Goble, and R. Stevens, *The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows.* Future Generation Computer Systems, 2008.
7. Goble, C. and D. De Roure, *myExperiment: social networking for workflow-using e-scientists,* in *Proceedings of the 2nd workshop on Workflows in support of large-scale science*. 2007, ACM: Monterey, California, USA.
8. De Roure, D. and C. Goble, *Six Principles of Software Design to Empower Scientists.* IEEE Software, 2008. **26**(1).

9. Strauss, A.C. and J.M. Corbin, *Basics of Qualitative Research: Second Edition: Techniques and Procedures for Developing Grounded Theory*. 2nd ed. 1998: Sage Publications.
10. Somekh, B. and C. Lewin, *Research Methods in the Social Sciences: A Guide for Students and Researchers*. 2005: Sage Publications.
11. Creswell, J.W., *Qualitative Inquiry and Research Design: Choosing among Five Traditions: Choosing Among 5 Traditions*. 2nd ed. 1997, Thousand Oaks, CA: Sage Publications.

Chapter 11 Conclusions and Future Work

This final chapter draws together conclusions from the research presented for both topics addressed in this thesis: atmospheric chemistry model development; and provenance for *in silico* experiments. Following on from this set of conclusions, areas for future work are outlined.

11.1 Conclusions

The first section of this chapter presents the conclusions of my research, revisiting each of the research objectives (presented in Chapter 1). Prior to considering the research objectives, a general reflection on the research approach I adopted is presented.

11.1.1 Research Approach

The multidisciplinary, ethnographic approach to the research presented in this thesis was described in the opening chapter of this thesis, the successes and limitations of this approach are outlined below.

The key successes of the research approach adopted were that:

- Research contributions were made to both the atmospheric chemistry and e-Science domains.
- A novel, user-orientated approach to provenance, for *in silico* experiments, was explored. Had the research been conducted along the more traditional, disciplinary lines a very different research output may have been delivered;
- Fundamental assumptions about the nature of both the e-Science and the atmospheric chemistry domains were challenged (e.g. the impact of constraint implementation on modelled radical concentrations had been assumed to be unimportant);
- As a researcher embedded within the atmospheric chemistry community (but with a computer science / information systems background) I was able to bridge the gap in understanding and terminology between the computer science and atmospheric chemistry researchers associated with my work.

The limitations experienced with the multidisciplinary, ethnographic approach included the following.

- Efforts to apply the understanding of the impact of constraint implementation on modelled radical concentrations (as described in Chapter 4), to generate insight into chemical processes in the atmosphere were unsuccessful. This was a result of there being insufficient time to gain the level of understanding required to produce a chapter of purely atmospheric chemistry research;
- The insight developed throughout the development of the ELN is coupled to my personal experiences (within the atmospheric chemistry community). Hence it could be argued that there is a lack of objectivity and breadth in the research findings. However, the alternative of adopting the role of a more passive observer of the atmospheric chemistry community could have led to arguments that the resulting research was insufficiently grounded in the practice of individual scientists.

11.1.2 Research Objective 1: OSBM Development

The development of the OSBM played two important roles within the wider research presented in this thesis: first, reviewing the modelling process itself, led to the opportunity to conduct research in to the impact of constraint implementation on modelled radical concentrations; secondly, the challenges of benchmarking the OSBM, led to the opportunity to explore the role of provenance within the atmospheric chemistry community (particularly for *in silico* experiments). So, in this case the development of a tool for use by the wider research community can be seen to have led to two valuable outputs: first, enabling researchers to conduct high quality research, using the MCM, with a freely available, open source tool; and secondly, as a by-product of achieving the original research goal, interesting and productive branches of research have been opened up and partially explored.

11.1.3 Research Objective 2: Constraint Implementations

The investigation presented in Chapter 4 considered the impact of constraint implementation (i.e. a way in which experimental data is used to configure a

computational model) on modelled radical concentrations (i.e. the model output). The constraint implementation had significant impact on model output, but it was not possible to show that any given constraint implementation led to improved model accuracy[19]. Given that it was not possible to demonstrate the superiority of a single constraint implementation, in terms of improving the accuracy of model output, a constraint implementation was recommend based upon the following.

- A set of tests with simplified models, where the impact of constraint implementation could be measured against a known solution.
- The mapping of the constraint implementation to the characteristics of the physical system being modelled; i.e. if the constraint implementation for a variable (e.g. a species concentration being constrained to vary every 15 minutes instantaneously) does not match the characteristics of the variable in the physical system (e.g. a species concentration in the atmosphere varying on a sub one minute timescale continuously) then a good mapping does not exist. Where possible a good mapping is a desirable characteristic to seek when configuring a model.
- The improved computational efficiency associated with the use of a continuous interpolant.

Determining the appropriate constraint implementation to adopt for a given model will depend on the nature of the experimental data being used to constrain the model. In the research in presented in Chapter 4, I took the experimental data at face value, i.e. if the experimentalist had presented their data (for say ozone concentration) with a frequency of 1 minute; I assumed that the data could be used to constrain the model at a 1 minute frequency. This assumption rests on the experimental data owner having processed the raw experimental data in such a way that it is appropriate to use the data for constraints at a 1 minute frequency. The experimental data used in the constraint implementation research was taken from the British Atmospheric Data Centre (BADC), which is, amongst many other duties, responsible for the long-term archival of data from atmospheric chemistry field campaigns (funded by UK research councils). The provenance associated with the experimental data, taken from the BADC, was not sufficient to determine the

[19] As measured by the model's ability to produce results that match the experimental measures for radical concentrations.

manner in which data is could be used (e.g. the question "does the data require further averaging prior to use in a computational model?" can not be answered).

11.1.4 Research Objective 3: Mapping Provenance-related Working Practices

Mapping provenance-related working practices led to an enhanced understanding of the role of Community Evaluation Activities (CEAs) across the atmospheric chemistry community. The current working practices of MCM developers gathering and evaluating feedback from *in silico* experiments[20], to inform the ongoing development of the MCM, was examined as an exemplar CEA. I developed envisioned working practices provenance for this *in-silico* experiment CEA, where an Electronic Laboratory Notebook (ELN), for *in silico* experiments, plays a critical role in capturing the data and provenance that informs the CEA. The design, development and evaluation of the ELN was then considered in detail (Chapter 6-10), focusing on the requirements of individual modellers performing *in silico* experiments. This sub-section now reflects upon how the provenance requirements of individual modellers align with the requirements of the MCM developers (in the context of the *in silico* experiment CEA).

Aligning the provenance requirements of modellers and MCM developers

When considering how to align the provenance requirements of modellers (i.e. the researchers generating data and provenance) and the MCM developers (i.e. the researchers consuming data and provenance), it will be important to consider the following four factors.

[20] e.g. A paper reporting the application of the MCM, to model a particular chamber experiment, will provide a summary of the modelling methodology, results and conclusions. Necessarily a published paper will abstract away details of the modelling methodology and results, in order to present a concise and comprehensible record of the *in-silico* experiment and its findings. Whilst the published paper provides a good introduction to an *in-silico* experiment it provides insufficient detail, in terms of data and information, to enable the MCM developer to confidently evaluate the published findings and update the MCM as necessary.

The relative importance of experimental-level provenance: Experimental level provenance is likely to be more important to MCM developers than modellers; as MCM developers will be seeking to navigate the content generated by the whole community for items of interest. This navigation will require provenance describing the goals and conclusions of experiments. Modellers focused on their own research, and managing their own provenance archive, are likely to require less provenance at the experiment level, since they posses the requisite tacit knowledge to navigate their personal archive.

The MCM developers may need more detailed provenance for the scientific process: MCM developers have not been involved in the execution of the research they are evaluating; hence they will need detailed provenance records to provide the contextual information they require. Again the modeller, considering their own archive, may be able to rely on their tacit knowledge to "fill in the gaps".

Minimal provenance standards: The MCM developers are likely to be interested in playing a role in defining minimum standards for the provenance captured by the ELN. These minimum standards would ensure that the data and provenance captured by researchers performing *in silico* experiments would be fit for purpose (i.e. supporting the ongoing development of the MCM).

The ELN design must, first and foremost, satisfy modellers: Balancing the requirements, where they are in competition, of MCM developers and modellers will prove a challenging task; but a guiding principle of first and foremost satisfying the modellers should be adhered to. The rational for this guiding principle is grounded in common sense; modellers will only adopted the ELN if they can personally benefit from its use; and the MCM developer can only benefit from provenance captured by researchers if ELNs are adopted across the community.

11.1.5 Research Objective 4: Development of an ELN

At the outset of Chapter 5 a definition for a user-orientated approach to provenance for *in silico* experiments was presented. Having designed, developed and evaluated the ELN it is possible to expand the definition as follows. A user-orientated approach to provenance adopts the following principles:

- Take a holistic view of the scientific process and provenance capture practices, reengineering the scientific process, where feasible, to facilitate provenance capture and improve users' experience of the scientific process;
- Capture, where possible, provenance automatically;
- Represent, store and query provenance using the terminology of the scientific domain;
- Capture provenance for both human and computational processes;
- Place equal importance on capturing process provenance (generally automatically) and scientific reasoning (requiring the ELN user to make annotations);
- Capture provenance with respect to two frames of reference: first, the scientific process; secondly, the *in silico* experiment;
- Make use of inline provenance capture, to encourage annotations.

Benefits

Adopting a user-orientated approach to provenance can enable a number of benefits to be realised, including those listed below.

- Engaging users in the provenance system development process from the outset;
- Capturing and representing provenance in terms that scientific users can immediately relate to;
- Taking a holistic view (incorporating the scientific process and provenance capture practices) challenges the assumption that provenance is a secondary priority, rather than an integral component of the scientific process;
- Encouraging users to engage with, and invest in, provenance as a valuable resource for personal use and to be shared across the community;

Drawbacks

Adopting a user-orientated approach to provenance also has a number of drawbacks, including the drawbacks listed below. These drawbacks can be seen to be consequences of realising the benefits of a user-orientated approach, outlined above.

- The level of resources required to develop the domain understanding required to adopt a user-orientated approach;
- The tools and ontologies developed are necessarily domain specific, so transferability of these tools and ontologies is an issue;

11.2 Future Work

This section concludes this thesis, by looking to the future and identifying the key areas where interesting and potentially important research will take place.

11.2.1 Modelling

Specific elements of future work for further development of the OSBM and better understanding the impact of constraint implementations were suggested in Chapters 3 and 4 respectively. The main themes are reiterated below.

Maintenance and ongoing development of the OSBM: The transition of the OSBM from a development project, with small number of developers and users, to a tool used across the atmospheric chemistry will be addressed in the near future. Limited ongoing support for the OSBM will be provided by the MCM development team, with other interested parties able to contribute to the OSBM as part of a collaborative open source development project.

Exploring the impact of constraint implementation: The research in Chapter 4, highlighted the impact of constraint implementation on the modelled radical concentrations. Opening up a number of new research opportunities including: modelling radical concentrations on a timescale comparable with *in situ* experimental measurements of radical concentrations; and exploring the impact of the uncertainty associated with the data used in constrain models.

11.2.2 Developing a Production Quality ELN for Modellers using the MCM

The ELN implemented and evaluated was a prototype system, intended to act as a proof of concept rather than a fully functional, production quality system. In order to realise, and fully evaluate, the benefits of adoption of the ELN across the MCM user community, a production quality version of the ELN will be required. I believe that, prior to the production quality ELN being deployed, it will be necessary to conduct another iteration of prototyping in light of the evaluation results presented in Chapter 10. In the next prototyping iteration key findings from the evaluation will be addressed, including:

adopting a holistic approach and the associated increase in design scope; and offering customisable annotation interfaces to the ELN user. The next prototype will be developed within the context of the EUROCHAMP, described in the next sub-section. In order to encourage the adoption of the production quality ELN across the MCM-user community it will be important to engage with the publishers and researcher funders (the interest of these stakeholders was outlined in Chapter 7). If either of these stakeholders seek impose minimum provenance standards as a condition of: research publication, or research funding, respectively; a significant impetus for ELN adoption would be provided.

11.2.3 EUROCHAMP Project

The EUROCHAMP project [1] consists of a consortium of 12 laboratories throughout Europe. Each laboratory brings 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, this has two benefits: facilitating the analysis of *in situ* experimental data, to produce scientific knowledge; and ensuring that the performance of the MCM is frequently tested. The computational modelling infrastructure will build upon the Open Source Box Model (described in Chapter 3). 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, a provenance and knowledge management architecture will be implemented. I envisage that each researcher using an ELN will be able to make sections of their ELN available to community, the security and sharing models for the ELN remain of topic for research. 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. I envision 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 *in silico* experiments that could inform their current research;
- 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 general in Chapter 5 (as the MCM development CEA) and my associated publications [2] [3].

11.2.4 Transferability

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

References

1. Wiesen, P., *The EUROCHAMP Integrated Infrastructure Initiative Environmental*, in *Environmental Simulation Chambers: Application to Atmospheric Chemical Processes*. 2006. p. 295-299.
2. 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

3. 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.

Appendix I: Semantically-Enhanced Model-Experiment-Evaluation Processes (SeMEEPs) within the Atmospheric Chemistry Community

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

Appendix II: Semantically enhanced provenance capture for chamber model development with a master chemical mechanism

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.

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

In press:

Martin, C.J., et al. *A User-Orientated Approach to Provenance Capture and Representation for in silico Experiments: Explored within the Atmospheric Chemistry Community*. in *Crossing Boundaries: Computational Science, E-Science and Global E-Infrastructure*. 2009: Philosophical Transactions of the Royal Society A.