

**Information Provenance and  
Semantic Publishing  
A New Vision for Publishing in  
Science**

**Ian Wood**

A subthesis submitted in partial fulfillment of the degree of  
Master of Compting (Honours) at  
The Department of Computer Science  
Australian National University

November 2008

© Ian Wood

Typeset in Palatino by T<sub>E</sub>X and L<sup>A</sup>T<sub>E</sub>X 2<sub>ε</sub>.

Except where otherwise indicated, this thesis is my own original work.

Ian Wood  
5 November 2008



To my mother, who gave to her children her youth and to the significant contributions to communal knowledge that she otherwise would have been able to make.



---

# Acknowledgements

---

I would like to thank my parents for providing a home, Dr. Henry Gardner and Dr. Jay Larson for guiding me patiently through my return to academic life, Dr. Peter Baumgartner, Dr. Ian Barnes, Dr. Roger Clarke, Dr. Scott Sanner, Dr. Catherine Legg, Dr. Jason Grossman and Dr Tom Worthington for their expert advice, Wang Xiaobin for development and support of WebScope4 and to the many others who helped me along the way.



---

# Abstract

---

Science is facing many organisational challenges. Online technologies are opening new ways for scientists to communicate and exchange knowledge. Technological advances in scientific instrumentation are allowing scientists to gather large quantities of data that place increasing demands on storage, processing and data management. New knowledge, data and compute resource technologies are being developed to cope with these demands. These trends have led to unprecedented capacity for the creation of scientific knowledge.

I argue, however, that the area of scientific publishing, though it has seen significant improvements, has not fully capitalised on the available knowledge management and networking technologies. I present a general strategy and some initial steps toward a new *semantic publishing* paradigm that draws together technologies and ideas from knowledge representation, grid technologies, web 2.0 and the Semantic Web.

At the core of semantic publishing is detailed semantic representation of scientific ideas and arguments. With such representation, the concept of *theory provenance*—a record of the arguments and dependencies of published ideas—becomes feasible. When theory provenance is combined with data provenance (the history of acquisition and processing of data) published science can achieve full transparency, allowing effective verification and review.

As an initial foray into semantic representation of scientific arguments, I present a framework for representing arguments based on evidence. Such arguments are fundamental to science, and are essential to detailed semantic representation of scientific arguments.

I argue that a semantically published body of knowledge can be exposed to automatic and semi-automatic analysis such as accurate citation and impact analysis, adapted social bookmarking as a quality and interest measure and provision of different presentations or views of scientific knowledge for diverse audiences. Advances in natural language processing also suggest the future possibility of automatic semantic translation of existing published works. With improved computing power and artificial intelligence technologies, automated reasoning could conceivably be applied to a semantically published body of knowledge as a tool for consistency checking and an aid to discovering new scientific theories.

Finally, I summarise the social changes, technological gaps and missing knowledge that would need to be filled to achieve this vision, and argue that these challenges are not insurmountable.



---

# Contents

---

<b>Acknowledgements</b>	<b>vii</b>
<b>Abstract</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 eScience and The Scientific Method . . . . .	2
1.2 A New Vision for Scientific Publication . . . . .	3
1.3 Evidential Reasoning and Schools of Thought - Small Pieces of the Semantic Publishing Puzzle . . . . .	4
1.4 Roadmap for Future Research . . . . .	4
<b>2 Knowledge Representation, Science and the Web</b>	<b>5</b>
2.1 Knowledge Representation and Ontologies . . . . .	5
2.1.1 Overview . . . . .	6
2.1.2 Ontology . . . . .	6
2.1.3 Description Logic . . . . .	7
2.1.4 The Semantic Web . . . . .	8
2.2 Knowledge Representation in Science . . . . .	8
2.2.1 Domain Markup Languages . . . . .	9
2.2.2 Reasoning with Complex Knowledge . . . . .	9
2.2.3 Ontologies in Science . . . . .	10
2.3 Grid Technologies . . . . .	11
2.3.1 What is a Grid? . . . . .	11
2.3.2 Grids in eScience . . . . .	12
2.3.2.1 Workflows . . . . .	13
2.3.2.2 Data Provenance . . . . .	14
2.3.2.3 Collaborative Tools . . . . .	14
2.3.3 Semantic Grids . . . . .	14
2.4 Publishing in Science . . . . .	15
2.5 Semantic Publishing Tools . . . . .	16
2.6 New Web Technologies and Social Phenomena . . . . .	17
<b>3 A New Vision for Semantic Scientific Publishing</b>	<b>21</b>
3.1 The Publishing Needs of Scientists . . . . .	22
3.2 Semantic Publishing . . . . .	23
3.2.1 Representing Scientific Knowledge . . . . .	23
3.2.2 Semantic Citation . . . . .	23

---

3.2.3	Theory Provenance . . . . .	25
3.2.4	Knowledge Bases and Webs of Knowledge . . . . .	27
3.2.5	Publishing Tools . . . . .	28
3.3	Integration with Existing Infrastructure . . . . .	29
3.3.1	Linking to Existing Knowledge . . . . .	29
3.3.2	Science Publishing Infrastructure . . . . .	29
3.3.3	Semantic Publishing Into and From the Grid . . . . .	30
3.4	Potential Benefits - Answering Scientists Needs . . . . .	31
3.4.1	Implicit and Explicit Verification . . . . .	32
3.4.2	Propagation of Refutation . . . . .	32
3.4.3	Exposure of Knowledge to Semantic Search . . . . .	32
3.4.4	Layers of Information . . . . .	33
3.4.5	Views and User Interfaces . . . . .	35
3.4.6	Trust/Verity Management . . . . .	35
3.4.7	Automated Reasoning and Scientific Knowledge . . . . .	37
3.5	Other Areas of Application . . . . .	38
3.6	Potential Barriers . . . . .	38
3.6.1	Cultural Momentum: Why Change the Way we Publish? . . . . .	39
3.6.2	Cultural Adaptation: Learning new Tools . . . . .	39
3.6.3	Semantic Markup is Hard Work . . . . .	40
3.6.4	Early Adopters . . . . .	41
3.6.5	Technical Barriers . . . . .	42
<b>4</b>	<b>A Framework for Evidential Reasoning</b>	<b>43</b>
4.1	Fundamental Concepts and Definitions . . . . .	43
4.2	Illustrative Examples . . . . .	46
4.2.1	Newton's Second Law . . . . .	46
4.2.2	Statistical Tests . . . . .	48
4.2.3	Climate Modelling . . . . .	49
4.2.4	A Plasma Physics Example . . . . .	52
4.2.5	Kon Tiki . . . . .	55
4.2.6	Classification . . . . .	56
4.3	Summary . . . . .	57
<b>5</b>	<b>Representing Schools of Thought</b>	<b>59</b>
<b>6</b>	<b>Future Research and Development Directions</b>	<b>61</b>
6.1	Social Research . . . . .	61
6.1.1	Study the Needs of Scientists . . . . .	61
6.1.2	Social Impacts . . . . .	62
6.1.3	Changing Roles of Publishers . . . . .	62
6.1.4	Changing Attitudes to Web Technologies . . . . .	62
6.1.5	Identifying Early Adopters . . . . .	62
6.2	Knowledge Representation . . . . .	62

---

6.2.1	Fundamentals of Science Modelling . . . . .	63
6.2.2	Representing Science . . . . .	63
6.2.3	Natural Language Processing . . . . .	63
6.3	Knowledge Management . . . . .	64
6.3.1	Fundamentals . . . . .	64
6.3.2	Semantic Citation Flavours . . . . .	64
6.3.3	Propagation of Refutation . . . . .	64
6.3.4	Folksonomies of Science . . . . .	64
6.3.5	Knowledge Exchange and Trust . . . . .	65
6.3.6	Mining Knowledge Bases . . . . .	65
6.3.7	Adaptation of KBs for Different Reasoning Tools . . . . .	65
6.4	Knowledge Interfaces . . . . .	65
6.4.1	Layers of Information Wiki . . . . .	65
6.4.2	Collaborative Islands . . . . .	66
6.5	Linking to Math and other services . . . . .	66
<b>7</b>	<b>Conclusion</b>	<b>67</b>
<b>A</b>	<b>WebScope: A Data Grid for Fusion Research</b>	<b>69</b>
	<b>Bibliography</b>	<b>71</b>



---

# Introduction

---

In recent years, we have seen substantial changes in the ways we communicate and share knowledge. The internet has become the first port of call when new information is sought, email is often our first choice when we want to contact someone, and new web phenomena are changing the way the coming generation socialises. Along side the Internet phenomena has been an exponential increase in our capacity to gather and store data [153; 142]. Though the processing power of our computers is also increasing exponentially [101; 44] much of our data is still inaccessible due to its volume. It has often been stated that we are “data rich and information poor” (eg: [145]). New information technologies, standards and data management practices are beginning to solve these problems.

These changes are having a significant impact on science. International collaboration has become the norm, made possible by new communication and data sharing technologies. The dissemination of published work has become vastly more effective with digital repositories and services like CiteSeer[24] and Google Scholar [60]. Workflow systems, grid computing and ever faster supercomputers are helping scientists manage complex manipulations of large data sets and data collection systems [44].

These tools and techniques still leave much room for improvement. Our capacity to generate experimental data is fast outstripping our capacity to organise and utilise it. Scientific work is being published at an ever increasing rate, making it more and more difficult to manage our vast store of knowledge. Our scientific publishing techniques are still essentially ‘on paper’ using a publishing model was first used in 1665 when the first editions of “Journal des sçavans” and “Philosophical Transactions of the Royal Society” appeared.

Current work in eScience is tackling many of these issues. Data grids are providing a framework for organisation and access to the vast data stores [44]. Semantic grids promise to streamline resource and data sharing and provide clearer provenance tracking [127]. Automated workflow tools are allowing more efficient utilisation of data and other resources such as supercomputers [44; 61]. Collaborative tools tailored to science are being developed [131]. Knowledge representation and management tools are being developed and applied in many areas of science to add utility to large and cumbersome collections of knowledge [139; 14]. Ideas from the semantic web initiative are starting to be used to track provenance of data and other resources and to manage and enable efficient scientific workflows [96]. These and other developments

in eScience are surveyed in Chapter 2.

Microsoft research sums up these concepts in their report “Towards 2020 Science” [39]:

Our findings have significant implications for scientific publishing, where we believe that even near-term developments in the computing infrastructure for science which links data, knowledge and scientists will lead to a transformation of the scientific communication paradigm.

## 1.1 eScience and The Scientific Method

The scientific method has been long studied. There are many famous publications (eg: [120; 65; 97]), and the basic elements have long been agreed upon. Hilf, Kohlhase, and Stamerjohanns provided a schematic overview of this consensus [68] (see Figure 1.1).

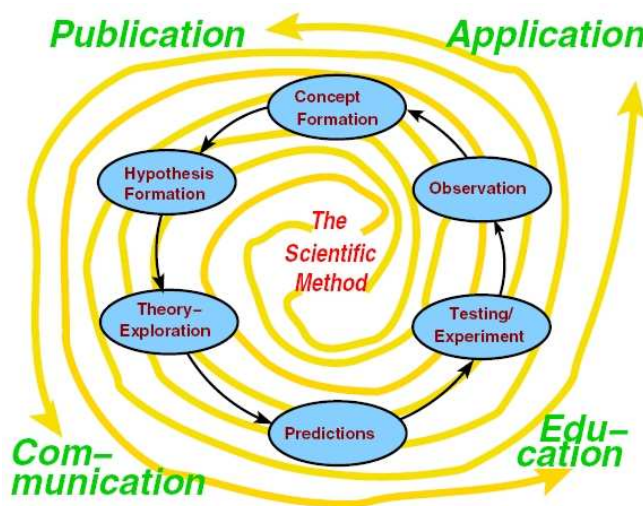


Figure 1.1: The Scientific Method—image from [68]

I do not wish to go into too this process in detail except to note that *communication* is an essential and integrated part of the process. It is through effective communication of our ideas that we act as a community of scientists, and not just a collection of individual researchers. The dissemination of scientific findings is essential to the continued development of our understanding of the world. It allows us to “stand on the shoulders of giants”<sup>1</sup>. Hilf et al describe this need in point DP6 [68]:

...the actual spreading of the information on the findings to other laboratories in the world is part of the operational procedure to gain physics insight.

<sup>1</sup>This metaphor was first recorded in the twelfth century and attributed to Bernard of Chartres - Wikipedia



Figure 1.2: The process of science today—image from [58]

We will see in Chapter 2 that efforts in eScience are helping many of the steps outlined in Figure 1.1. The importance of efficient and effective publication technologies has been recognised in the eScience research community (see Figure 1.2). Efforts to improve our scientific publications, however, have concentrated on improving access to the current ‘on paper’ format.

## 1.2 A New Vision for Scientific Publication

My key proposal in this thesis is a vision of scientific publishing in which science is published in a direct conceptual form. By this I mean that the core published material is in the form of represented knowledge—a form in which the published concepts and the links between them are explicitly represented in a format similar to the knowledge representation formats of today. This vision is not intended to be something that can be attained in the short term, more as a possible future to which to aspire, and which may be closer than we think.

In Chapter 3 I introduce this vision, present some possible advantages and discuss technical and other difficulties that may obstruct the realisation of the vision. New concepts of *semantic citation*, a citation *between concepts* which carries an indication of the nature of the citation (depends on, refutes, is distinct from etc..) and *theory provenance*, a record of the concepts that lead to and support a theory, are proposed.

When theory provenance and data provenance (a record of all the steps taken to obtain data) are combined, the full history of scientific findings is revealed. This has useful implications for effective knowledge management, transparency of scientific endeavour, attribution of merit, recognition of impact etc...

There are many other possible advantages of such a system. For example modified collaborative indexing techniques could be employed to measure importance and validity of published elements, replacing and/or enhancing traditional peer-review and

citation tracking techniques. There could be personalised digests of published results tailored to specific interest areas, but including important results that are more distant. No doubt more advantages will become apparent should such a system come into existence.

### **1.3 Evidential Reasoning and Schools of Thought - Small Pieces of the Semantic Publishing Puzzle**

Fundamental to the scientific method is the development of a theory about a phenomena from which predictions can be made, and subsequent development of a repeatable experiment whose results match the predictions of the theory. Scientific publications often contain these elements. In Chapter 4 I develop a framework for representing arguments based on experimental evidence and present case studies illustrating its application.

A common situation in science is the existence of differing theories to describe a phenomenon or *schools of thought*. In order to allow parallel development of these different strands of science, a knowledge representation system needs to be able to identify and manage (possibly contradictory) groups of ideas. In Chapter 5 I describe one approach to this problem.

### **1.4 Roadmap for Future Research**

The vision of semantic publishing I have presented is a long term goal. Though much of what I describe and develop in Chapter 3 can be done with existing technologies, I have identified many research avenues. Chapter 6 presents a summary of these directions.

---

# Knowledge Representation, Science and the Web

---

Information technology has seen a rapid evolution, both in technological advances, and in the expansion and availability of computer and networking infrastructure. This has had many impacts on the way science is done [58]. In this chapter, I review these changes as they relate to the publishing vision presented in Chapter 3.

Knowledge representation (KR) is the starting point for the publishing vision I will forward in this thesis. In Section 2.1 I present an overview of knowledge representation techniques and tools and their application in the Semantic Web, then go on in Section 2.2 to investigate knowledge representation initiatives in science.

Grid technologies are increasingly important in the data intensive, highly collaborative atmosphere of science today. These are important to semantic science publishing as they are an environment from which science is published (and will, perhaps, become a dominant environment), they provide data provenance and they use represented scientific knowledge as an organisational principle for the resources they manage. In Section 2.3 I summarise current developments in grid technologies for eScience.

The final sections of this chapter contain investigations of other relevant technologies and social phenomena: Section 2.4 summarises recent developments in science publishing such as open access and concerns about peer review. Section 2.5 looks into currently available semantic publishing tools. Section 2.6 investigates new web technologies and social phenomena such as wikis, open source development and Web 2.0.

## 2.1 Knowledge Representation and Ontologies

With KR, represented knowledge can be exposed to automated processing, enabling the services outlined in Chapter 3 and, no doubt, many other innovative services yet to be devised.

### 2.1.1 Overview

Representing knowledge in a formalised way has been a human activity since the dawn of civilisation. In spoken language words and sentences characterise cultural knowledge of the world. More recently, writing has given us a more permanent medium for representing knowledge and mathematics was developed to help us to formalise our knowledge of things in the world.

In computer science, conceptual modelling—formalising and representing some aspect of our knowledge—can be thought of as a part of any programming activity. As our computer systems, the data they operate on and the tasks we ask of them have increased in complexity, conceptual modelling has become an important area of computer science research [17].

The field of Knowledge Representation (KR) originated as a branch of artificial intelligence (AI). The conceptual modelling and representation practises used in other fields have now converged with that of AI and KR is now understood in a broader context. John Sowa describes it succinctly:

“Knowledge representation is a multidisciplinary subject that applies theories and techniques from three other fields:

1. *Logic* provides formal structure and rules of inference
2. *Ontology* defines the kinds of things that exist in the application domain.
3. *Computation* supports the applications that distinguish knowledge representation from pure philosophy.”

(from [137])

A similar but perhaps more restrictive definition can be found in The Description Logic Handbook [6]:

Knowledge Representation is the field of Artificial Intelligence that focuses on the design of formalisms that are both epistemologically and computationally adequate for expressing knowledge about a particular domain.

The ideas presented in this thesis are more in line with Sowa’s definition. I will be considering formalised representations that some would argue to be computationally inadequate, in the sense that automated reasoning with them is not reliable, but which might play a key role in supporting the publishing vision presented in Chapter 3.

A *knowledge base* is a collection of represented knowledge from some domain designed for some specific purpose or purposes. Example might be the human genome, the words of the English language or varieties of fruit.

### 2.1.2 Ontology

The term *ontology* originated (and is still used) as a branch of philosophy which studies the “nature and existence of things in the world” and “aims at the objective description of any domain of objects” [163]. Philosophical ontology often deals with

---

vocabularies of terms describing a particular domain and, in information systems literature, the term ontology has been used to describe collections of knowledge ranging from vocabularies and catalogues to sets of general logical constraints [134]. There are two widely cited definitions of information systems ontology which align better with more recent usage:

“An ontology is an explicit specification of a conceptualisation” [62]

“An ontology is a logical theory accounting for the *intended meaning* of a formal vocabulary”[63]

The difference between these two definitions is subtle. The emphasis on “intended meaning” in the second definition brings it closer to represented knowledge, whereas the first is more pragmatic and perhaps closer to the way in which an ontology would be constructed. Reference [163] covers this distinction in more detail and attempts to unify the two.

### 2.1.3 Description Logic

The study of knowledge representation for artificial intelligence led to questions about the tractability and computational complexity of different systems. Theoretical attempts by logicians to answer these questions led to a deeper understanding of the tractability of automated reasoning and the development of a family of languages for representing knowledge. These languages are called *description logics* (DLs).

Seminal work the computational complexity analysis of DLs was done by Hector J. Levesque and Ronald J. Brachman [89]. They recognised that there is a trade-off between the expressive power of a language for knowledge representation and the difficulty of reasoning with the resultant knowledge bases. A fundamental result they point out is that languages entailing first order logic result in potentially unbounded reasoning operations. In a sufficiently expressive system, there will *always* be questions for which an automatic reasoner will *never* find an answer. Mathematics with the real numbers is such a system.

Levesque and Brachman’s approach was to investigate which features of first order logic could be removed to obtain a decidable logic—that is, a logical system for which all questions could be answered in finite time. Since then, many results on the complexity and decidability of DLs have been found [162].

Two other knowledge representation approaches, frame-based systems and semantic networks, have been shown to be formally equivalent to certain description logics<sup>1</sup> and the complexity and decidability results can also be applied to them [6].

Description logics form the basis of many automated reasoning applications and systems today. In particular Web Ontology Language (OWL; [151]), the W3C standard ontology language for the Semantic Web, is based on a description logic. Some significant areas of application have been in software engineering (managing large software systems), configuration (piecing together a set of components to match a problem

---

<sup>1</sup>If non-monotonic features such as default values are excluded

specification), medical informatics, natural language processing and data management [6].

#### 2.1.4 The Semantic Web

The *Semantic Web* is the brainchild of Sir Tim Berners-Lee. Knowledge representation and automated reasoning technologies had been successfully used in information systems to make efficient use of complex and large data sets. Berners-Lee recognised the potential of these technologies in the context of the World Wide Web. In a famous 2001 article in *Scientific American*, he described his vision of the semantic web:

The Semantic Web is "...an extension of the current Web in which information is given well defined meaning, better enabling computers and people to work in cooperation." [13]

A Web where content and services have machine-readable semantic annotations would offer many advantages. Information could be retrieved accurately and efficiently using search results which match the precise meanings of search terms. Automated *agents* could scour the Web for information and services to fulfil sparsely specified requests such as "find me a dentist I can trust near the airport in London and make a booking for Tuesday morning".

Berners-Lee's fame and the obvious potential of these ideas stimulated much excitement in some areas of the computer science research community. There was an expectation that the Semantic Web would flourish with a similar rapidity to the advent of the World Wide Web, however this has not happened. The Web is still largely semantically un-annotated and society has been slow to develop the ontologies that would be needed to mark up content [66]. But considerable progress has been made, and some would argue that the Semantic Web still appears to be an achievable if not inevitable future [129].

## 2.2 Knowledge Representation in Science

As the quantity and complexity of scientific data and knowledge has increased, new technologies have been developed to organise and effectively utilise it. In many areas of science, knowledge bases have been created or are in the process of creation. These knowledge bases are primarily in the form of DL ontologies. Their application has led to sophisticated data retrieval and resource management systems (see 2.2.3).

Much scientific knowledge cannot be represented using DL ontologies due to their limited expressiveness. More expressive higher order logic formalisms have been developed for disciplines such as mathematics and physics. Other sciences such as chemistry and earth sciences have developed markup languages for data exchange and interoperability.

The importance of these knowledge representation efforts to the thesis presented here is that they provide a format for the representation and communication of the

---

concepts and relationships that are the subject of scientific investigation. Such representational formats are the foundation of the semantic publishing vision presented in Chapter 3.

### 2.2.1 Domain Markup Languages

Specialised scientific markup languages have been driven by the need to extend HTML to perform typesetting of technical information such as mathematical formulae, by the need for standard information and data exchange formats, and the need for standard formats for automated processing. Numerous markup languages supporting science and eResearch exist or are under development[9]. A substantial list of markup languages can be found on-line[143]. This list is likely out of date and may contain obsolete standards, however it is a good indication of the substantial development efforts in this area. The following are a few notable examples from science:

**MathML, OpenMath and OMDoc** Mathematics markup languages - MathML is intended for presentation of mathematical formulae [93]. OpenMath captures the meaning or semantics of mathematical formulae and can be used to complement MathML. OMDoc incorporates both MathML<sup>2</sup> and OpenMath and OMDoc allows the definition of entities (such as definitions and theories) and relations between them [111].

**PhysML** Physics markup language is an extension of OMDoc intended to represent results and theories from physics [68].

**CML** Chemistry markup language describes molecules. An open standard to enable interoperability of analytical tools. It covers disciplines from macromolecular sequences to inorganic molecules and quantum chemistry [25].

**ESML** Earth science markup language [41]. Similar in intent to CML, but with geophysics data formats.

**QCdml** An XML schema for marking up gauge configurations in lattice quantum chromodynamics [94].

### 2.2.2 Reasoning with Complex Knowledge

In many areas of science, core results are expressed in mathematics. In the words of Galileo:

*Mathematics is the key and door to the sciences.*

We can devise formalised representation systems for such knowledge, however as we have seen in Section 2.1.3, automated reasoning based on logic is unreliable with such highly expressive systems. That does not mean, however, that automated reasoning is not useful for scientific applications.

---

<sup>2</sup>OMDoc captures *Content MathML* - that is, the non-presentation aspects of MathML

In mathematics, automated theorem proving systems have successfully aided researchers to find mathematical proofs that had not previously been known [95]. These systems often require (sometimes substantial) human intervention. Due to the undecidability of the underlying system, they may, at times, effectively get stuck, and in order to move forward, they need a hint.

This may still be helpful—the points at which the reasoning system gets stuck may be informative, and the system will likely be able to check consistency without help in many cases. On the other hand, the system may need a huge number of hints before it can arrive at an answer. Also, the algorithms used in these systems are highly complex. Even if answers can be found without help, the execution times to find them may be prohibitively long. One anecdotal estimate of the level of complexity this could entail is:

It's like being asked to carry out an exhaustive case analysis of chess, starting from scratch, but much much harder [10].

Though it may be worth experimenting with current technologies to gauge their potential and move toward a possible future with more powerful automated reasoning systems, we are unlikely to see fully-automated reasoning providing substantial support for mathematical representations of scientific knowledge in the near future.

One such experiment would be to build a fully-automated system that attempts to check consistency of a knowledge base containing complex mathematical relations. Such a system would not always be able to complete the check, however it would still be useful when it succeeds. To the best of my knowledge, this has not been attempted on general mathematical theories, however theorem provers have been successfully used to check consistency of theories in first order logic [22].

### 2.2.3 Ontologies in Science

The potential for ontology based knowledge management was recognised early, and already in 2000 there were several ontologies in use in scientific research [139]. Ontologies are now widely used in bioinformatics, with substantial well developed reference ontologies [38; 69] and ontologies developed by the MyGrid project [140]. Another area that has seen significant work in ontology development has been in association with the construction of semantic grids (see Section 2.3.3). Numerous platforms and methodologies for ontology construction and maintenance have been developed [140; 8; 90; 156].

One application of ontologies in science is data integration. Well developed ontologies, made in collaboration with relevant expert communities, serve as a standard form of annotation and allow diverse data formats to be utilised interoperably. There are several projects working on these issues [47; 158].

---

## 2.3 Grid Technologies

Loosely speaking, a Grid is a digital federation of geographically dispersed resources. These resources could be data, computers, grid users or scientific instruments. The Grid provides a uniform platform to view, communicate with or otherwise utilise these resources. Grid implementations for science provide data provenance, process automation, access and security control, collaborative tools and simplified data access. We shall see that Grid infrastructures are effective (perhaps essential) tools for eScience in today's highly-collaborative and computationally-intensive scientific milieu. Grid technologies are important in the discussion here for several reasons. Firstly, in the proposed semantic publishing vision, publishing is an integral part of the research process and is derived from and feeds back into the Grid. Secondly, Grid infrastructures would enable effective management of the collected semantically published knowledge.

Another area of interest is the development of provenance tracking collaborative tools. There is some potential to incorporate represented knowledge into these systems, providing a less work-intensive semantic authoring tool. More on this in Section 3.6.3.

### 2.3.1 What is a Grid?

The initial motivation for the invention of Grids was to fully utilise and manage heterogeneous, dispersed super-computer facilities and very large data stores—the vision was to create a large and powerful virtual supercomputer. The term *Grid* was inspired by the electrical power grid: in the power grid, you receive electricity and do not know where it comes from, similarly, in a computational grid, you do not know which physical computer is processing your data. This allows the compute resources to be efficiently and automatically utilised. As the technology evolved, its denizens realised that the management and coordination of the resources that the Grid offers could have a far reaching impact on the nature and abilities of our interconnected computer infrastructure including the internet and the WWW [44].

Ian Foster, a prominent figure in the grid research community, has attempted to provide a definition of what a '*Grid*' is:

#### A Grid Checklist

I suggest that ... a Grid is a system that:

1. *Coordinates resources that are not subject to centralised control ...*
2. *... using standard, open, general-purpose protocols and interfaces ...*
3. *... to deliver nontrivial qualities of service.* [43]

Foster is not specific about the nature of the resources—they could be data, computational power, other grid users or anything else appropriate to the application. Nor is he specific about the nature of the coordination—later we will see that what is known

as a *Semantic Grid* is a Grid equipped with Semantic Web technologies and annotations.

With the lack of centralised control, a Grid is scalable and robust to computer failures. Foster's second point ensures that there will be many applications and other resources that can effectively utilise a Grid. Thirdly, delivering nontrivial qualities of service means that a Grid becomes an effective and useful resource to its users.

A Grid that has resources requiring controlled access must implement non-centralised mechanisms for authentication and rights management, and this is a common feature in grid systems. This feature is important if scientific knowledge is to be digitally represented as scientists may want to, at times, allow access only to those in close collaboration lest someone else publish the ideas before them. We shall see in Section 3.4.6 that a semantic publishing system could reduce this need.

An important concept in Grid systems is *virtualisation*. This principally concerns the provision of views and organisations that can be designed independent of the underlying network structures and physical locations of resources [44].

An nice example of virtualisation is the Earth System Grid (Figure 2.3.1), whose data catalogue(s) look like a single filesystem to the user. In fact, with OPeNDAP-G technology, not only does one not need to know where the files are, one doesn't even have to know what or how many files are used to build a particular (space-time) data set [40]. Figure 2.3.1 is a schematic of the Earth System Grid indicating its primary services and data storage facilities. On the schematic NCAR MSS (NCAR mass storage system) and ORNL HPSS (Oak Ridge National Laboratory High Performance Storage System) are deep, high-capacity robot tape storage systems. The Storage Resource Manager (SRM) middleware, used to retrieve data from the deep archives and the Replica Location Service (RLS) support the data virtualisation described above.

Grid technology is an evolving and dynamic area. In order to foster the interoperability between diverse grid applications and implementations, an open community focused on the development of Grid standards has been formed. The Open Grid Forum [110] accomplishes its work through open forums that build the community, explore trends, share best practises and consolidate these best practises into standards.

Foster has also put forward a vision for *The GRID*—a future in which the internet is full of grids, all communicating and inter-operating, effectively one large grid similar in its ubiquity to the World Wide Web. The Grid would be one enormous marketplace of computing services, data and consumers, heralding unprecedented efficiencies and possibilities [44].

### 2.3.2 Grids in eScience

Grids have been widely used in areas of science that involve large quantities of data and which are analysed and processed by geographically distributed (often international) collaborations. Data grids are often combined with mechanisms for sharing and utilising computing resources such as supercomputers. Some prominent examples are the Earth System Grid [40] for climate and environmental data, the EU Data-Grid Project which will manage data from the Large Hadron Collider [27], the Open

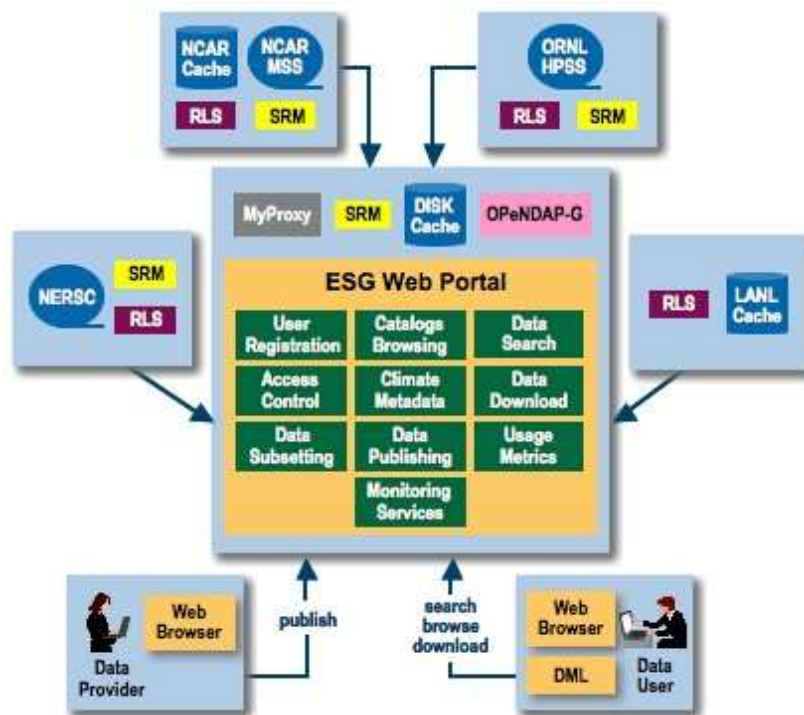


Figure 2.1: Earth System Grid Architecture [40]

Science Grid in the USA [115], the International Lattice Data Grid [18; 71] for quantum chromodynamics data and the National FusionGrid of the USA [49]. A data grid technology for plasma physics has been developed at the ANU [51]. The list of implemented Grids is long and growing [44].

### 2.3.2.1 Workflows

A *workflow* is a series of tasks describing a method for achieving something. The term is used in business and government contexts to describe work processes [42].

In the context of Grids and eScience, a workflow consists of a high level description of a task and tools and data formats which can be interpreted by Grid middleware to perform the task in an efficient and distributed way. Typically these workflows are so large and the data sources and participants so widely distributed that they cannot be executed within one system or institution [44]. Ludascher et al state effective description of workflows in the science context [91]:

*Scientific workflows* ... are networks of analytical steps that may involve database access and querying steps, data analysis and mining steps, and many other steps including computationally intensive jobs on high performance cluster computers.

Workflow tools such as Pegasus [31], Swift [159] and Kepler [91] act as meta-programming languages to automate Grid services and are being used in many data

and computationally intensive areas of science such as climate science, particle physics, biology and medicine [44].

### 2.3.2.2 Data Provenance

*Provenance* is a term that has long been used to describe the history of ownership and modification of works of art in order to track their value and authenticity. This principle can be applied to data to provide transparency in its acquisition and processing, allowing those who use the data to determine its validity and to verify its accuracy [102].

The term *data provenance* has sometimes been used in a relatively restricted technical sense: a record of the computational steps that transformed raw experimental data into that which is published [44]. I will be using a somewhat broader sense, including details of how the experiment was conducted. The intention is for data provenance to provide sufficient information for another researcher to *reproduce* the result.

There have been concerns expressed in the scientific community about the lack of provenance in published work. For example, new algorithms published in computational science often lack sufficient detail to reproduce the published results [122]. There have been some attempts at publishing formats with embedded source code to deal with these problems [123; 88].

In another example of the need for provenance, complex derivation of data from scientific workflows can obscure the data's significance and meaning not only from those reading published work, but also from those who created the data [99]. Many Grid workflow implementations, therefore, have implemented effective provenance gathering mechanisms. Several surveys of data provenance practises in eScience have been compiled [133; 16]. They report that, though provenance issues are being addressed, there is still much work to be done, in particular on standards to allow the portability of provenance metadata.

### 2.3.2.3 Collaborative Tools

State of the art collaborative tools are being incorporated into Grid middleware [131; 103; 104]. These tools incorporate knowledge management services for annotating and organising records of meetings and other collaborative events. In a sense, these tools are tracking the provenance of ideas. This leads to an interesting possibility for an easy adoption path to semantic publishing—see Section 3.6.3 for more on this.

## 2.3.3 Semantic Grids

The term *Semantic Grid* was coined by De Roure, Jennings and Shadbolt to describe “the application of Semantic Web technologies both on and in the Grid” [126]. The basic idea is to use Semantic Web technologies such as web services and autonomous agents to manage resource discovery/allocation of the data/services that the Grid is serving (on the Grid) and to manage grid middleware that realises virtualisation, security and other Grid services (in the Grid). The effectiveness and efficiency of Grid

---

services is substantially enhanced by this approach, particularly when the Grid contains large and complex resources [127; 57]. This can also be seen in research on workflow automation [132] and resource discovery [136; 3].

There are many semantic grid implementations in science today [98; 23; 56; 44; 140]. Virtual observatories [14], though they do not claim to be semantic grids, satisfy Foster's grid criteria and apply Semantic Web technologies. Virtual observatory middleware has been implemented [46] and successfully deployed to serve data from fields spanning upper atmospheric terrestrial physics to solar physics [96]. Work on grid development [12] indicates that this is also a rich and growing field.

## 2.4 Publishing in Science

As with many areas of human endeavour, the advent of the Internet and the World Wide Web has stimulated changes in the science publishing industry. The majority of scientific journals and conferences (possibly all) now publish their material in digital form. Indeed there are now journals that publish *only* in digital form (e.g. [118; 79; 112]). Repositories of scientific articles often have associated knowledge management services such as keyword searches and subject categorisation (eg: Springer, IEEE and ACM) and second party search and categorisation services such as CiteSeer [24], Google Scholar [60] and the ISI Web of Knowledge [76] provide one-stop portals to much of the worlds published science.

Despite these advances, our scientific publishing techniques are still essentially 'on paper', albeit digital paper, largely failing to utilise many of the powerful knowledge management techniques that are available [29]. This publishing model was first used in 1665 when the first editions of "Journal des sçavans" and "Philosophical Transactions of the Royal Society" appeared and has not changed significantly in the ensuing 350 years.

Science publishing today is in flux. The stayed core of reputable journals has not changed significantly, however controversy and new ideas abound concerning problems of quality and access in the current system. The effectiveness of peer review has been brought into question [78; 21; 32] and is the topic of open discussions (e.g. [106]). Un-refereed pre-print archives such as arXiv [5] have been very successful (arXiv passed 1 million articles this year, many of which have been published in refereed journals) and open access journals have proliferated [35] with some high impact journals among them [119]. The debate about open access to published science involves major players in the scientific community—recently, US government funding required public access to the results of funded research [7]. Peter Suber has compiled a Timeline of the Open Access Movement [141] that gives some indication of the accelerating uptake of an open access publishing model for science.

One journal worth noting here is PLoS One [118; 54]. This innovative on-line journal published by the Public Library of Science has an unusual publishing philosophy. Firstly, it does not restrict itself to any particular field. The motivation for this is to provide a publishing space for interdisciplinary articles whose content spans many

disciplines but does not fit the narrow subject area of any discipline-specific journals. Secondly, it publishes any paper that meets their technical standards, not attempting to assess the importance of the content—this is achieved by comments and critiques left by readers. Thirdly, it operates with a publisher pays system, allowing them to provide free public access to published material.

The general science journal Nature [106], apart from support for community discussions of issues such as peer review, has been experimenting with Web 2.0 technologies, for example reference [138] has an associated wiki where the public can edit and update the content as well as facilities for leaving comments and discussion.

Another interesting example is the UN-supported Intergovernmental Panel on Climate Change (IPCC) that periodically publishes reports summarising the scientific understanding and knowledge of the Earth's climate. These publications have significant impact on government and business decisions around the world, and indeed on our societies future. They have a detailed and highly rigorous process with which they build these reports [72].

## 2.5 Semantic Publishing Tools

Many tools and platforms have been developed for publishing text based and multimedia material with semantic annotations [149; 82; 147]. Many tools also support ontology additions and maintenance and are oriented toward realisation of the vision of the Semantic Web. The tools provide a possible starting point for a semantic publishing approach that I propose in Section 3.2.5.

A recent knowledge management initiative in the business sector, commonly referred to as *single sourcing*, builds on the idea that a single piece of information should be represented only once within a collection of knowledge. Many business documents contain the same piece of information (for example, the name of the CEO). If the piece of information changes, that information must be found and changed in all the documents the business uses. This approach is especially useful for technical documentation and help systems, the same instruction may appear in several contexts. It is much easier to maintain the consistency of such documentation if that instruction needs updating in only one place. DITA [33] is a widely used standard for single sourcing, with many XML editors now supporting it [34].

The single-sourcing process is very similar to semantic publishing in that they both build structured documents that refer to elements of other documents. In single-sourcing, this structure relates to layout and the elements are generally blocks of text. In semantic publishing, the structure represents arguments and the elements represent concepts. Semantic publishing can be thought of as single sourcing with a higher level of abstraction.

Single sourcing is important here for several reasons. The existence of and experiences from content creation and editing tools is relevant to semantic publishing tools. Secondly, knowledge management culture evolving around single sourcing is positive: a significant community of knowledge users is learning to think in ways similar

to knowledge representation.

Another class of applications that incorporate machine readable semantics are *semantic desktops* [128; 77]. These are semantically aware knowledge management systems for personal computers. Files and programs are have semantic tags, either given automatically or by the user, or already present. As with other applications mentioned here, these tools are essentially doing semantic publishing to some degree, and again, are important both for design experience and tool availability and for the user base that is gaining awareness of semantically represented knowledge.

Another relevant semantic tool is the “living book” [11; 15]. Living books adapt their content to the (known) prior knowledge of a user, constructing their content from the extra information the user requires to achieve a learning goal and were used to create course notes for mathematics course at universities. This technology (or one like it) could be used to build layers of content relevant to different audiences (see Section 3.4.4 for more information).

## 2.6 New Web Technologies and Social Phenomena

The internet and the world wide web provide unprecedented possibilities for information transfer and communication. In recent years, people have been exploring these possibilities in new and innovative ways. Much of this research has happened on a community level, outside of traditional research institutions, and these institutions have been slow to adopt the new technologies [146]. As young scientists who spent their youth in an internet enabled world enter the research community, these technologies are gaining more interest within the scientific community.

The term *Web 2.0* has been used to describe a variety of modern internet applications. The term has had many differing uses, however in general it talks of *web platforms* with enhanced dynamic utility and often with mechanisms for users to contribute to the content [113]. Some common types Web 2.0 platforms are wikis, folksonomies, mashups, social networking, blogs and platforms used for open source software development. These are described below, along with some notes on their relevance to the semantic publishing vision I present in Chapter 3.

Wikis are on-line knowledge repositories that allow open access and editing to a community of users. Many, such as Wikipedia, are open to anybody. Others, such as business intranet wikis, may have limited access. Wikis can be powerful because they draw on the knowledge and expertise of all the users within their community and help those users to organise content in sensible ways. They rely on the inclination of users to provide information that is missing and to correct or update information that they find to be wrong. Users can also “watch” pages, receiving notification of changes. The phenomenal success of Wikipedia is a tribute to their potential—as of writing, the English language version of wikipedia has 2,604,656 content pages [155], Alexa [1] reports that Wikipedia ranks 8<sup>th</sup> in the world for web traffic, up from 500<sup>th</sup> in October 2004. The accuracy of Wikipedia has be favourably compared to Encyclopaedia Britannica on scientific subjects [53]. In Section 2.4 we described experiments with

wiki technologies in science publishing.

Open source software development is another web phenomenon that has achieved remarkable successes. High profile projects such as the Apache web server (50% of web sites are hosted on Apache<sup>3</sup>), Linux operating system (12.7% of web servers<sup>4</sup>) and Mozilla Firefox browser (19% of wikipedia visitors<sup>5</sup>) have seen incredible successes. Open source development projects generally use forums and issue tracking systems such as Trac<sup>6</sup> for managing the development process and community discussion. Open source version control systems such as Subversion<sup>7</sup>, Concurrent Versions System<sup>8</sup> (CVS) and Mercurial<sup>9</sup> are also used to enable multiple active development streams and for coordination of stable software versions. Version control systems such as these and the experiences of open source projects that have used them will likely be important in the management of knowledge bases produced by semantic publishing.

The term *blogs* is a shortened form of *web log*. There are many types of blogs [154], but they all have an author or group of authors (who adds primary content to the blog on a regular basis if the blog is active) and most have a facility for the public to attach comments to blog entries. The blog search engine Technorati<sup>10</sup> claimed, at the time of writing, to track 112.8 million blogs, over 250 million pieces of tagged social media and over 1.6 million posts per day, or over 18 updates a second. Blogs are impacting science both directly (for example, the RealClimate blog [124]) and through specialised notebook and journal applications (such as the ORNL Electronic Notebook Project [52]).

The principles of open discussion and community contribution and review, and the experiences of applying those principles in wikis and open source projects and blogs can be used to help guide scientific knowledge management. I will discuss this further in Section 3.4.6.

Another notable class of Web 2.0 application are *folksonomies*. Wikipedia gives a concise definition:

**Folksonomy** (also known as **collaborative tagging**, **social classification**, **social indexing**, and **social tagging**) is the practise and method of collaboratively creating and managing tags to annotate and categorise content<sup>11</sup>.

Some examples of Web folksonomies are social (Web) bookmarking sites del.ic.us<sup>12</sup>

---

<sup>3</sup>[http://news.netcraft.com/archives/web\\_server\\_survey.html](http://news.netcraft.com/archives/web_server_survey.html)

<sup>4</sup><http://www.linux-watch.com/news/NS5369154346.html>

<sup>5</sup>See the Wikipedia page "Usage share of web browsers". Other sites report 43% (w3schools.com), 32% (w3counter.com), 17% (thecounter.com) etc...

<sup>6</sup><http://trac.edgewall.org/>

<sup>7</sup><http://subversion.tigris.org>

<sup>8</sup><http://www.nongnu.org/cvs/>

<sup>9</sup><http://www.selenic.com/mercurial/wiki/>

<sup>10</sup><http://technoratimedia.com/about/>

<sup>11</sup>Taken from Wikipedia 29/10/08.

<sup>12</sup><http://del.ic.us>

---

and rabble<sup>13</sup>, product review systems epinions<sup>14</sup> and amazon reviews<sup>15</sup>, and image indexing in flickr<sup>16</sup>. Folksonomies, like wikis, utilise the collected knowledge and opinions of a community of individuals.

We shall see in Chapter 3 that semantically published science would be well positioned to benefit from new web technologies, both those existent today and new technologies that we have not yet seen.

*Mashups* are web pages whose content is drawn from diverse sources and services. Common examples are news digest sites such as GoogleNews and embedded maps such as can be found on WhatIf.com.

Online social communities such as myspace.com, facebook.com and QQ.com (in china) have flourished in recent years, with millions of active users. Technologies for analysing network structures to determine things like trust [59], likely friendships and common interest groups have developed in association with this phenomena. These technologies have interesting implications for science publishing. In Section 3.4 I will discuss some of their possible applications. There are open source social networking applications that could be adapted to the needs of science [92].

There are some interesting discussions [45; 117] and application (eg: [105]) of Web 2.0 technologies in an eScience context.

---

<sup>13</sup><http://rabble.ca>

<sup>14</sup><http://epinions.com>

<sup>15</sup><http://amazon.com>

<sup>16</sup><http://flickr.com>



---

# A New Vision for Semantic Scientific Publishing

---

The basic idea I am proposing is to publish science in a machine readable semantic form. That is, for the primary published product to be a representation of the knowledge that is being published. A textual representation (as is the current form of academic publications) would be secondary, and possibly derived from the semantic representation. This could potentially be transparent to the author—publishing tools could, for example, present a textual representation to an author<sup>1</sup>.

This would be a substantial change to the way science is published today, and many things will have to happen before it can be realised. Such a change cannot be expected to happen quickly and there are many questions that we must answer before we embark on that road: What are the potential benefits of semantic publishing? What are the potential drawbacks? What are the needs of scientists from a publishing perspective? What are the technical and social barriers that would have to be overcome?

In this chapter I begin to answer these questions, investigating ways in which semantic publishing could be implemented, proposing the ideas of *semantic citations* (citations that accurately track the flow of knowledge, Section 3.2.2), *theory provenance* (a record of the development of ideas, Section 3.2.3) and *conceptual workflows* (compiled scientific arguments and evidential reasoning tools, Section 3.2.3). I then describe in more detail how the idea of semantic publishing could relate to the current state of affairs. I map out key areas of research, development and change that would need to be addressed for this vision to become a reality.

Though there has been mention of publishing semantics along with the traditional paper format [30; 29; 32], there has been surprisingly little discussion on this topic in the scientific literature.

---

<sup>1</sup>With current technologies, the author would still need to link the text to digitally represented knowledge and may need to edit represented knowledge to create new concepts to represent his or her ideas.

### 3.1 The Publishing Needs of Scientists

As noted in Section 1.1, publishing is an important part of the continuing evolution of science. It allows other scientists to access new ideas to gain new inspirations and to incorporate these ideas into their research. It provides a forum for scientists to debate the validity of ideas and to come to eventual consensus. These aspects are the primary goals of a scientific publishing media.

But what about the needs of the scientists themselves? Anecdotal evidence suggests the following needs:

1. A medium in which to communicate their findings to their peers.
2. To be informed about new research findings, and for this information to be filtered for relevance to their own specific concerns.
3. A body of knowledge from which they can easily extract the leading edge of research in any given field.
4. An objective judgement of the quality and impact of their work and that of others.
5. To be able to maintain some level of understanding of advances in areas that may be distant or remotely connected to their own speciality (this need is not well met in the current system).
6. To be able to learn about an area different to their own speciality for interdisciplinary research.

In Section 2.4 we saw that the current publishing systems satisfy most of these needs, however some are only partially or unreliably achieved (for example quality and impact). In the following sections we see that my proposal for a more granular semantic publishing mechanism has the potential to fulfil these needs far better.

The publishing ethos varies in different fields of science. This could be due to several factors, some social, some political and some pragmatic. A publishing system would have to be careful to identify and cater to these differences. In many of the ideas presented in this chapter, care would need to be taken to respect this. Adjustments and extensions to interfaces and overall system functionality may be needed.

There is another important role for scientific publishing: to inform people outside of science about relevant scientific advances. Governments need to be informed to make good policy decisions (for example [73]). Businesses need to understand how scientific advances can be used in their industry, and how they may effect their business in other ways. Educational institutions need to be informed about changes in our understanding of things, so they can adapt their materials. Another group whose importance has recently been gaining recognition is the general public. Effective science communication is essential for the continued improvement of society [109].

---

## 3.2 Semantic Publishing

In this section I describe the basic elements of my proposal for *semantic publishing*.

### 3.2.1 Representing Scientific Knowledge

At the core of semantic publishing is a machine readable representation of scientific knowledge. It is with such represented knowledge that we can hope for some of the benefits outlined below. The difficulties with achieving and then managing represented scientific knowledge drive the key research directions outlined later in this chapter.

The representational needs of different scientific fields vary significantly. In some areas of science, such as psychology, precise measurements are impossible and it is questionable whether the many concepts under consideration can be given precise meanings (that is not to say they have no value). In other areas such as genetics and biochemistry and geology, there is substantial work in identifying and classifying the elements of complex structures. In many areas such as paleontology and biochemistry, the development of new investigative and experimental techniques is especially important. And so on.

The languages and mechanisms for representing science need to respect the needs of the areas of science in which they are used. They need to be flexible and extensible in order to track our changing understanding of the world. And yet they need enough consistency to allow for continued intelligibility and generic knowledge management tools. These are non-trivial problems and much work needs to be done. In Section 3.2.4 we will see the beginnings of some answers to these problems.

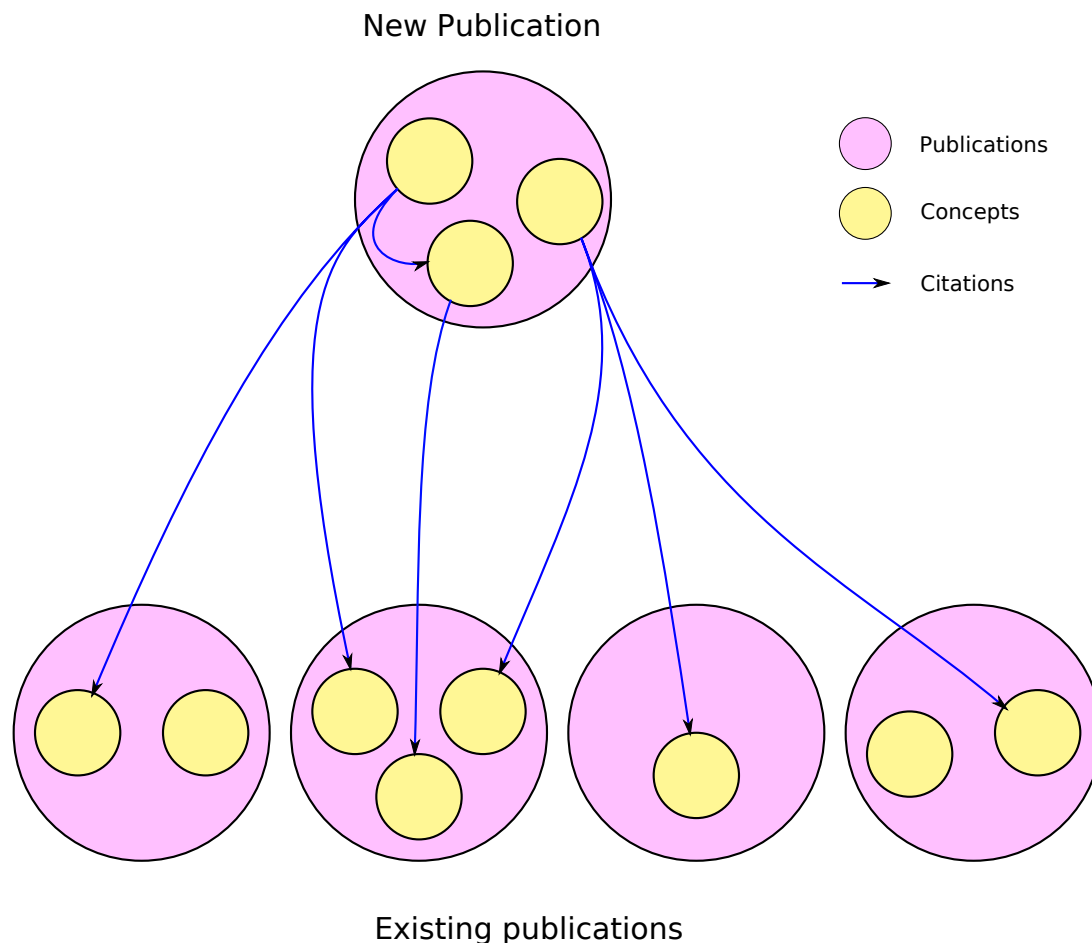
We saw in Section 2.2 that there are many initiatives to create standard annotations and formats for scientific data. These attempts have been *ad-hoc* and application specific, responding to particular needs for interoperability and knowledge management in the fields for which they were developed. But they all embody key concepts and physical entities from their fields and so represent a beginning of the road to full knowledge representation structures for science. In general however, they are far from representing scientific knowledge to a depth that would be required for semantic scientific publishing as presented in this thesis. We shall see in the following sections what extra representational structures would be needed.

### 3.2.2 Semantic Citation

The current citation techniques are coarse. Without reading the text (a task difficult for machines and onerous for humans), a citation tells you nothing about which specific concepts or results from the cited paper are relevant nor which concepts or results they relate to in the paper you are looking at, and nothing about the relationship between those concepts or results. Currently, this information can only be obtained by reading the paper and considering the context in which the citations appear.

With semantically represented scientific knowledge, it would be possible to provide all that information. A citation could link individual elements of a represented

publication with individual elements of the cited publication. Also, this citation could contain information about the nature of the relationship between those elements. For example, it could indicate that the new element assumes the validity or truth of the first or conversely that the new element contradicts the first, or it could simply indicate that the new concept is distinct from the first or is a refinement or sub-concept of the first. One important semantic role is an indication that the cited concept is the same as the first. An author may not always succeed in locating similar concepts to those he or she is presenting, and thus would create a new concept entity. Later, someone may realise the connection and add an appropriate citation. Issues of rights to edit or annotate would need to be addressed. I discuss this and similar issues in Section 3.4.6. I will refer to citations between individual concepts with semantic information about their relationship as *semantic citations* (see Figure 3.1).



**Figure 3.1:** Semantic citations

With such a network of linked publications, the ramifications of a new result that contradicts some established ideas could be easily tracked. It can take some time with the current system for new results to filter into our education systems and government or business decision makers. A semantically linked network of publications would

---

help to pass new knowledge on to where it is needed in a timely manner.

Imagine a scientist who has a profound idea, but is unable to convince others of its significance. Later, another scientist with a louder voice in the scientific community takes this idea and publishes some startling results with it. Likely it is this second publication that will attract many citations. If semantic citations were employed, the origin of the idea could be easily identified, and the original inventor of the idea could be given due praise.

In practical terms, given a format for representing scientific knowledge, semantic citations are relatively trivial to implement. Analogous to URI's (Universal Resource Indicators [150]) and DOI's (Digital Object Identifiers [36]), elements of represented knowledge (data, theories, entities etc..) could be given unique identifiers which could be quoted in the citing document. It would not be difficult to implement a system of back-referencing also.

A related idea was presented by Carr et. al. in [20]. They present a service that semantically links documents that contain similar concepts, utilising existing document metadata. In essence, they are creating something similar to semantic citations between existing documents on the web. The approach here is to take existing forms information and attempt to enrich its semantics. My proposal works in the other direction: it explores how to utilise and organise semantically rich information.

In Section 3.4 we will see some other potential benefits that semantic publications could bring.

### 3.2.3 Theory Provenance

A key concept in my proposal is *theory provenance*. Analogous to data provenance as developed in many Grid implementations (see Section 2.3.2), theory provenance deals with the history of the development of a theory or other published result. Figure 3.2.3 depicts the flow of ideas and data as new theories are formulated and tested. As with data provenance (which exposes experimental and computational techniques), theory provenance exposes in a declarative way the logic, mathematics, statistics, models, experimental results etc.. that were used to obtain and justify a result. This entails three aspects:

**theory dependencies:** The previously published results on which a new result depends. Dependencies could have different forms. For example, the new result could depend on a previous one, a previous result could be refuted or simply be indicated to be different to the new result. Ideally, there would be both standardisation of dependency types and flexibility for the evolution of new dependencies. These are a subset of the semantic citations.

**supporting arguments:** A representation of the logical steps that led to the new result. These steps could reference already published results (both experimental and theoretical) and link them with new experiments and/or theoretical constructs to arrive at new conclusions. The theory dependencies could be discovered by analysing the supporting arguments.

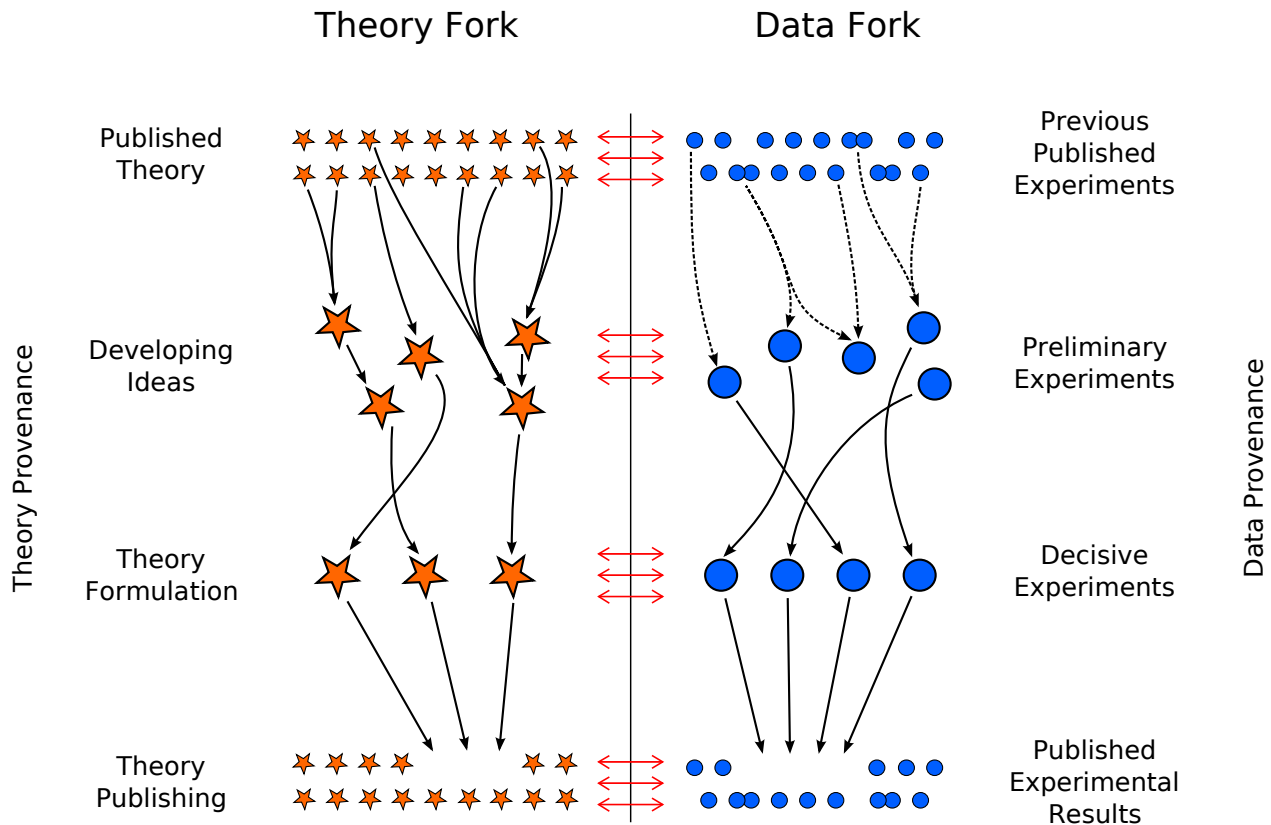


Figure 3.2: Science Lifecycle

Scientists use a rich set of logical arguments. As discussed later in Chapter 4, a thorough investigation of the types of arguments would be appropriate before committing to any particular representational framework. It would also be well for that framework to be designed with inherent flexibility for future extension.

**attribution:** Who presented the result and the institutions and funding bodies that supported the research.

Analogous to Grid workflows, *conceptual workflows* could be compiled from established forms of scientific argument. They could then be referenced when presenting the supporting arguments for a new result. This would become non-trivial, for example when advanced statistical tests are applied to complex data. The established techniques would require data and theory to have specific forms, which could be checked automatically<sup>2</sup>, and indicate the underlying assumptions about the ways in which the data was collected<sup>3</sup>. Conceptual workflows could themselves be published, reviewed

<sup>2</sup>The principle of *garbage in, garbage out* would still apply of course. However, with careful provenance initialisation and tracking, conceptual workflows could help prevent the accidental application of inappropriate techniques.

<sup>3</sup>The assumptions could, for example, be provided in the form of a checklist for experimental design that must be completed before the test can be applied.

---

and accepted (or not) in a similar way to other published science. As described here, conceptual workflows are merely packaged scientific arguments with well defined inputs and outputs. At some future time, these may indeed become workflows for automated reasoning services. I discuss the possible futures of automated reasoning in science in Section 3.4.7.

Of the representation languages mentioned in this thesis, perhaps the nearest to the needs of representing theory provenance is PhysML [68]. PhysML has constructs for representing observables, experiments and instruments and inherits theorem constructs and abstract types from the mathematical representation language OMDoc [85]. The theorem constructs represent mathematical proofs with references to other theorems (similar to semantic citations as described in the next section). Other formalised languages for mathematics also have this feature (such as [100]).

It is worth noting the compiled libraries of formally represented mathematics and their attendant theorem provers/checkers such as MIZAR [100] and IsarMathLib [75]. These libraries faithfully represent theory dependencies and supporting arguments, though to my knowledge they do not record *attribution*. As such they provide a substantial step toward theory provenance for science.

### 3.2.4 Knowledge Bases and Webs of Knowledge

The formalised mathematics libraries mentioned in the previous section are built “from the ground up”—they start from fundamental axioms of mathematics, and build everything from there. For scientific publishing, this level of detail is not necessary in order to build a useful representation. Accepted fundamental results from science and mathematics could be treated as axioms. Indeed, we could start with cited results as our effective axioms for a given publication.

To make sense of this, it is useful to think of the knowledge structures implicit in these knowledge representation approaches. A publication and its citations is similar to a theorem in a mathematics library and its dependencies. There are, however, a few instructive differences. Firstly, an interconnected collection of publications offers no guarantee of consistency. Secondly, the collection of publications is an inherently dynamic system and new publications are continually added. Lastly, the collection of publications could be distributed across the internet, whereas a mathematics library is usually held in one place, likely within one computer system (though there could be other copies).

What I am trying to point out here is that a natural medium for scientific publishing is a distributed *web* of interconnecting publications, similar in structure to the World Wide Web. On the other hand, for standardisation and (potentially) automated reasoning<sup>4</sup> a single, consistent knowledge base is needed.

I envisage a combination of the two. The collection of all semantically published science would have the form of a dynamic and continually growing web of knowledge with semantic citations as the links between knowledge nodes (publications). This web could be automatically or semi-automatically analysed to produce domain

---

<sup>4</sup>See Section 3.4.7

knowledge bases that represent the current thinking and terminologies of a particular area<sup>5</sup>. This task would be substantially simplified with semantic publications. Such knowledge bases would have a role not unlike review and summary articles, technical books and textbooks, and would be regularly maintained and updated. As our knowledge deepens, they would also have a role as accessible repositories of large and complex collections of factual information. This process is already underway with the construction of knowledge bases representing the current state of knowledge in a particular area, particularly in the biological sciences [38; 69; 19].

The provenance of the theories in such knowledge bases would be maintained. Newly published knowledge could reference such knowledge bases in preference to original publications (knowledge provenance ensuring that the original sources of ideas are not forgotten). Further structures and refinements to this framework would likely be made—for example, layered knowledge bases with differing levels of abstraction and detail (similar to undergraduate texts vs. technical publications—see Section 3.4.4).

How to build such knowledge bases and other questions surrounding the management of evolving knowledge are open research questions. Research around the concept of *knowledge grids* is looking into these and related problems [160].

### 3.2.5 Publishing Tools

In Section 2.5 we discussed the emergence of semantic annotations in common document editing tools as well as sophisticated DITA XML editing tools that enable links to repositories of reusable XML segments. We also saw a “living book” [11] that was able to adapt its content to the needs of a particular user. These all represent aspects of the sort of editing tools semantic publishing would require.

Tools for semantic annotation could generate semantic representations of documents. Many also support ontology additions and maintenance: this would be necessary for scientific publishing, as scientific publications often contain new concepts (not always—for example, research summaries or new measurements of known quantities). These tools in their current form are not ideal for semantic publishing however; the focus of semantic publishing is on the *semantic content*, text or graphical representations would be secondary. These tools work in the opposite direction: Semantic markup is added to text, and does not need to reflect the full semantic content. As a bridging mechanism, they are, however, valuable.

These tools are in their infancy. Their interfaces are cumbersome and people are not accustomed to the idea of authoring from or with semantically represented content. These factors represent a barrier to the semantic publishing vision. I will discuss them in greater detail in Section 3.6.2.

As with any publishing medium, you would expect to find many publishing tools appearing. It would be wise to cater for diverse publishing tools tailored to different personalities and cognitive approaches and personal histories [50]. For example, some people may find a spacial/graphical approach more intuitive and others more

---

<sup>5</sup>Measures of verity and consensus would be needed for this—see section 3.4.6

---

comfortable with words. Younger generations may prefer radically new techniques, while those already used to an older system may see no reason to change. One advantage to a semantic publishing approach is that it would be agnostic to the preferred tool of a given user, so long as it conforms to an underlying knowledge representation standard.

An interesting project launched in September this year intends to build a mathematical wiki that uses theorem provers to vet the correctness of submitted proofs [48]. Another proposed mathematical wiki is presented in [86]. These wikis will be semantic publishing platforms for mathematics.

### 3.3 Integration with Existing Infrastructure

In this section, I discuss different elements of the current eScience and scientific publishing infrastructure and how semantic publishing could be integrated.

#### 3.3.1 Linking to Existing Knowledge

Currently, conceptual scientific knowledge is contained in a collection of text books, technical references and a substantial corpora of published scientific articles. For effective theory provenance, semantically published results would need to be able to cite 'on-paper' published results. There are two possibilities for implementing these citations.

A pragmatic and immediately implementable approach would be for semantic citations to simply indicate the paper or other work which contains the cited ideas. Such citations would be semantic—they would indicate the nature of each citation (extends, contradicts, etc. . .). They could also contain more precise information about where the cited concept can be found in the cited publication.

It would be possible to begin semantically publishing immediately with such citations. There are, however, many concepts in each field of science that are considered fundamental and are not referenced. Also, it would be good to identify unique sources of important concepts. To ensure consistency within each field, knowledge bases reflecting the current consensus on fundamental concepts would need to be built. These would essentially kick-start the web of semantically represented knowledge.

A possible future approach to integrating existing literature with a semantically published body of knowledge may become available if research in natural language processing and/or data mining enables automatic and detailed semantic representation of text. In this case, 'on-paper' works could be automatically represented and made available for true semantic citation.

#### 3.3.2 Science Publishing Infrastructure

The current science publishing network is essentially identical to the web structure described in Section 3.2.4. Citations (though not semantic) link collections of ideas (ie: papers) forming a web or tree-like structure.

In the first instance, this infrastructure may not need to change beyond adopting knowledge representation and management technologies to adopt a semantic publishing system. However, in order for the scientific community to get full benefit from semantic publishing, access to the complete web of semantic information would be needed. Without this, theory provenance would not be practicable.

One possibility would be for the semantic information to be freely accessible, but for the detailed human readable forms to require subscription, as is the case with abstracts and bibliographies vs full text in today's paid subscription journals. The semantic content essentially carries the ideas present in the publication, far more information than the citations and abstracts of current publications—there would be less motivation for users to pay for the better representation. The knowledge representation strategy could be designed to be difficult to interpret without the associated textual forms, however this contradicts the whole purpose of semantic representation—to give computers access to the full semantic content. If computers are able to access that content, they will also be able to translate it into a form people can readily understand. Restricting that access would detrimentally effect the benefits semantic publishing can bring.

In Section 2.4 we saw that the scientific publishing industry is undergoing self-examination, with wide debates about open access and peer review and many new and experimental publishing systems. Perhaps these movements could result in a more open philosophy of scientific knowledge access.

In order to effectively support aggregated knowledge bases as described in Section 3.2.4, new standards for knowledge exchange and trust would need to be devised. The standards under development in the Grid community would likely be good starting places for the development of such standards.

Many of the possible advantages outlined in Section 3.4 would involve more radical changes to the publishing infrastructure. I will describe these in the relevant sections.

### **3.3.3 Semantic Publishing Into and From the Grid**

In Section 2.3 we saw that semantic grid infrastructure is having an increasingly important role in modern science. Grid technologies provide federated resources, sophisticated access control measures, automated data processing workflows and collaborative tools for scientists.

The relationship between semantic publishing and grid technologies will likely be highly integrated. Grids provide data provenance information that would be included as evidence for published ideas and they provide the infrastructure within which much of the publishing will likely take place. Published ideas, on the other hand, provide the grid with extra information about the resources it manages - information linking data to the theory/theories it supports. If the grid were used to store and/or manage the published ideas, grid services such as virtualisation would greatly enhance the utility and accessibility of published knowledge, resulting in a true knowledge grid.

---

As we saw in figure 1.1, scientific knowledge and data are intimately connected throughout the process of science. There is a relationship between the ontologies used to organise data on semantic grid's and the represented knowledge in semantically published results.

Obviously, the data referenced in a publication may reside on a semantic grid. The semantic tags used in the grid to describe that data represent concepts that scientists use to understand what the data is about. These tags could naturally include elements of the data's provenance—which experiments and algorithms were done by who and where.

But there is more: Once the data has been recognised as an example of or evidence for some phenomenon, it could have a tag relating to that phenomenon. Such tags effectively indicate that the data is relevant to some scientific theory.

When new theories/phenomena/entities are discovered by scientists, they would desire tags on the data representing those new concepts. Now we begin to see the link to semantically published results—*the new concepts in a published result need to be attached to the relevant data*. In a semantic grid, it would be natural to do this with semantic tags in the grid's ontology. In other words, the grid's ontology should accept newly published concepts as possible semantic tags. In practise, new concepts will likely be added to the grid ontology first—the research that scientists do before they publish creates the new ideas, and the grid would be the workspace for this research.

To represent science, representation of mathematical relations is important. We saw in Section 2.1.3 that description logic ontologies such as those used in semantic grid's cannot properly represent mathematical relationships. This does not prevent them from representing mathematically related concepts—the offending relationships could simply be left out. Grid ontologies are used to organise grid resources. Though representing mathematical relationships would add to their utility, ontologies built by omitting these relationships would still provide useful organisational principles for the data and resources on the Grid. For example, in order to search the grid for data, services or apparatus that relate in some way to a given theory, the grid middleware does not need to know the mathematical relationships embodied by that theory.

This does, however, raise an interesting point: We may want our represented knowledge to be used in different systems with differing levels of expressiveness. An initial approach is for the more expressive features to simply be invisible to the simpler systems, as described above. Can we do better? Perhaps there are approximations or simplifications of the expressive features that could be used in the simpler system? Such questions form interesting directions for future research.

### 3.4 Potential Benefits - Answering Scientists Needs

In Section 3.2 I talked about some immediate potential benefits of semantic publishing. In this section I elaborate on some of the advantages that could lead from widespread adoption of semantic publishing combined with other technologies.

With the possible exception of automated reasoning on scientific knowledge, all

of the ideas here can be implemented using existing technologies. These ideas are, in fact, adaptations of techniques that are currently in use, often in large scale applications.

### 3.4.1 Implicit and Explicit Verification

The availability of theory provenance and conceptual workflows could have two beneficial effects. Firstly, implicit in the application of sophisticated techniques to analyse data, the referenced template for that technique would remind scientists of the requirements of the technique. In this way, for example, we may see fewer errors due to the incorrect application of statistical tools.

Secondly, the arguments presented would be explicitly stated and easily accessible to someone scrutinising the publication. With standardised formats, flaws in the logic or incorrect application of analytical tools could be easier to recognise.

### 3.4.2 Propagation of Refutation

Sometimes in science, a long established result is later refuted. In such cases, that refutation may take some time to become widely known, especially to people outside the specific area of research where it lies, as existing scientific literature that builds on the refuted result remains indistinguishable from other research that is still considered valid. This can especially be problematic with high level summaries and educational material for schools.

With effective theory provenance, the consequences of changes in our understanding of the world can be readily propagated to *all* digitally published knowledge that relates to the changed knowledge. This connection to the underlying knowledge would also aid in propagating new knowledge to higher level materials, for example, outdated ideas in school teaching materials would be flagged, and could be more readily updated.

### 3.4.3 Exposure of Knowledge to Semantic Search

The current document format for science publishing does not lend itself to effective searches. This is improving as text mining and automated recognition of semantic content improves [69], however it is still primitive.

Semantically published knowledge on the other hand would be searchable by specifying individual concepts or expressions containing concepts. This could be achieved, for example, with current description logic reasoners<sup>6</sup>.

A poignant example of the utility of conceptual searches is a review by Naomi Oreskes of 928 peer reviewed papers to determine the level of scientific consensus on the issue of anthropogenic climate change [114]. Oreskes read the abstracts of all the papers, a time consuming, laborious and error prone process. If these papers had been semantically published, the task would have been almost trivial. I should note,

---

<sup>6</sup>Assuming the knowledge can be assimilated and translated into an appropriate DL knowledge base.

however, that the efficacy of semantic publishing relies on the accuracy of the semantic representations.

#### 3.4.4 Layers of Information

To accommodate different information needs of scientists (such as points 1, 5 and 6 in Section 3.1) it would be advantageous to enable views of published work with differing levels of detail and with differing assumed knowledge—*layers of information* (Figure 3.3). Higher level synopsis views could be used by non-scientists in government, business or just through general interest.

Of course, these views would need to be authored by someone. If the knowledge base into which ideas are published were a kind of wiki system, with (appropriately managed) scope for viewers to leave comments or edit different aspects of published material, interested or professional viewers could create extra knowledge layers. Wikipedia is a fine example of this process with open editing rights—with semantic publishing, the full provenance (theory and data) could be referenced in the articles, thus enriching the content. To improve fidelity. As with Wikipedia, there is a danger that the information added may not faithfully represent the underlying knowledge. The presence of full provenance would be an aid to ensuring the correctness of information. Trust and verity systems as discussed in Section 3.4.6 could also be used to manage information reliability and/or editing rights.

The higher level summaries would be overviews with varying authorship and presenting different levels of detail for those less expert or less interested in the details. Review articles and wikipedia pages, the IPCC [73] and government or industry reports are current examples.

Each layer is linked to the layer above it, thus exposing the full provenance (data and theory) of the ideas presented in each layer. For example, if I were reading a summary of an area of research, and want to know what argument was presented for a given theory when it was published, and a history of debate and new results that affect it, I can find them.

The layers described below and in Figure 3.3 are divided into two groups. The knowledge representation (KR) layers contain represented, machine readable knowledge. Evidence has been included in this list. It would typically consist of experimental data and other observations, appropriately tagged. It is not, strictly speaking, machine readable represented knowledge, however it fits in with this group as it is generally not human readable.

##### **Knowledge representation layers** machine readable knowledge

- *Concept* layer - representation of domain concepts
- *Theory*.layer - representation of theories as relationships between concepts
- *Evidence* layers - observational data and its provenance
- *Justification* layer - representations of structured *arguments* using explicitly stated argumentative steps

**Textual layers** human readable descriptions

- *Publication layer* - a textual view of the justification layer. This is a *presentation for experts*, and would have a similar appearance to publications of today.
- *Summary layers* - summaries of key concepts, starting with *abstracts*, then *summaries* of small branches of research, then *higher level summaries* and so on.

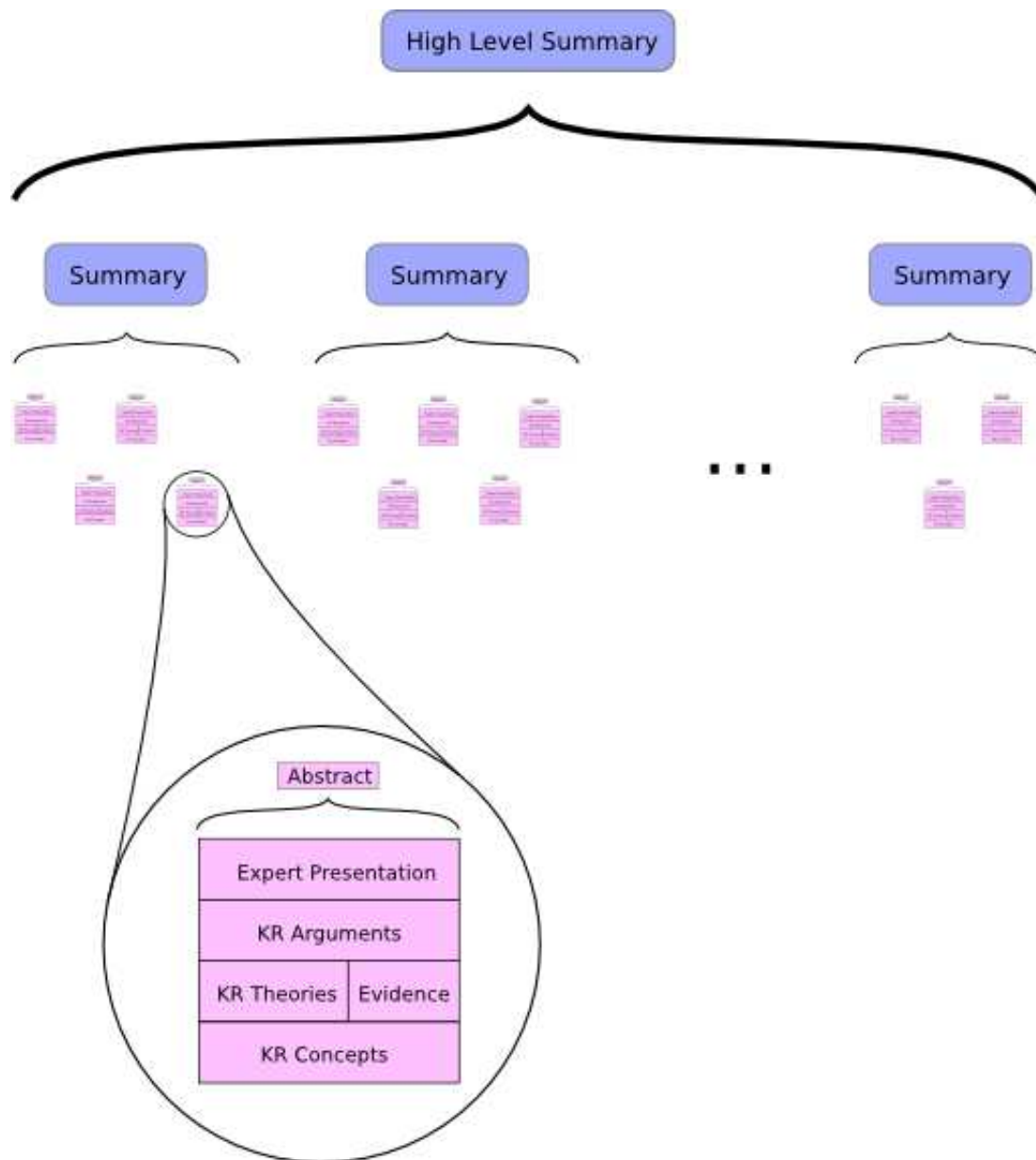


Figure 3.3: Layers of Information

### 3.4.5 Views and User Interfaces

In Section 3.2.5 we noted that a variety of publishing tools should be available that cater to the individual needs of users. This is also true of user interfaces in general. Different user interfaces could be created for viewing/editing the information. Some possible examples using current technologies could be:

- A graphical view of the represented knowledge and relationships, perhaps similar to the ontology editor Protege [121].
- Text-focused views, similar to pdf or html pages.
- Hybrid or XML views similar to semantic annotation tools presented in Section 2.5.
- Views for differing levels of familiarity with the area of research.
- RSS or email alerts - individuals could choose the what areas of research interest them, setting thresholds for verity, significance and relevance and trust and the level of detail retrieved from each area. Trust and significance measures are important here! (see Section 3.4.6).

In Section 3.2.5 we noted that as semantic knowledge management technologies mature and gain acceptance there will likely be new and innovative ways of interacting with semantically represented information. Semantically published knowledge is (at least in principle) agnostic to its presentation, allowing such new ideas to be easily integrated with existing systems and bodies of knowledge. This would also lend to the potential for automated translation into and from many languages.

Baumgartner et. al. have implemented a system for personalised views of mathematics textbooks that adapt to knowledge you are presumed to already know, presenting only things that are new to you in a given course [11]. This idea applied to views of published science could supply richly dynamic interfaces tailored to a users perspective and preferences. This would be like a scientific version of Google News, but much more powerful.

### 3.4.6 Trust/Verity Management

Measures of *trust* (the credulity of a source) and *verity* (the degree to which we should believe a piece of information) are important in scientific publishing. Measures of journal impact factors and editorial and peer review provide this information in the current publishing system. In Section 2.4 we saw that there is concern in the scientific community about the efficacy of the current system. In Section 3.4.5 we saw that an important application of measures of trust and verity is filtering the available information.

There have been some interesting suggestions for improving the current peer and editorial review system. Rodriguez et al utilised social networking technologies to choose relevant reviewers [125]. De la Rosa and Szymansk suggested a system of

citation dollars that are earned by people citing your publications, and spent when you publish. High impact journals would cost more, and there would be a mechanism for ‘lending’ your citation dollars [28].

An alternative mechanism to peer and editorial review for achieving measures of trust and verity is through community or public discussion (and dispute!) of published claims and through community ranking (folksonomy). In Section 2.6 we saw successful social phenomena like Wikipedia and online folksonomies. Their experiences could help guide effective implementation of similar techniques in science publishing. For example, Wikipedia has implemented a system for automatic detection of conflict [84] utilising machine learning techniques. A related example is the introduction of an endorsement system by the pre-print archive ArXiv [5]—articles can only be submitted with endorsement by an author who already has an article in the archive<sup>7</sup>.

The ability for users to edit the semantic content, with appropriate version tracking and constraints, may also be of value. For example, a missing citation may need to be added, or concepts merged (after consensus between authors). Trust management approaches using social networks could help to measure the quality of these comments and rankings [59] and enable the calculation of trust and verity measures for published materials.

A folksonomy approach would allow for sliding scales of verity and trust. This would have many advantages. It would be possible, for example, for junior scientists to publish unusual ideas in an unobtrusive way (as low verity scores would hide them from most scientists) whilst allowing interested individuals to assess their value and quality<sup>8</sup>.

A possible enhancement to such a system that is worth considering is a system of weighted voting. Based on publishing and other history, the system could make an assessment of a scientist’s level of expert authority in a particular area, as well as a measure of trust, and use this to weight that scientist’s ranking of a published element. This would, of course, be highly controversial, and great care would need to be taken to find a balance between authoritarianism and effectiveness, perhaps leaning toward less authority!

In a semantic publishing paradigm, current techniques used for analysis of citation graphs, with some adaptation to interpret semantic citations, could also be applied to citation graphs from semantically published knowledge. The resultant significance measures would, however, have a deeper meaning, and would in a sense be more accurate. The original inventor of an important idea would be more likely to receive due recognition<sup>9</sup>.

A semantic publishing system would allow scientists to publish much smaller pieces of knowledge, or publish their larger findings progressively. “Salami publish-

---

<sup>7</sup>ArXiv does no peer review

<sup>8</sup>this is similar to the pre-print archives of today, however, published results would automatically become visible to a wider audience if they attract good verity, significance and interest scores—see Section 3.4.6.

<sup>9</sup>There is still a chance that a connection to the first appearance of an idea may be missed, however, the author would have the opportunity to remedy the situation.

---

ing”, publishing findings in many papers, each with little new content, is a technique that has received criticism as an attempt to maximise publication count and improve job success in the current paradigm. In a semantic publishing paradigm, this effect could be mitigated with sophisticated measures of significance. Scientists would have a motivation to publish their ideas quickly in order to publish before others who may have the same idea <sup>10</sup>. This would improve the speed with which ideas are circulated and have a beneficial effect on the advancement of science. On the other hand, with the ability to publish ideas quickly and before substantial review, they may be less inclined to hold their ideas back until they are highly developed <sup>11</sup>.

In the current system, peer review, citation count and the impact of the journal in which an article is published are the available measures of verity and trust. In a semantic publishing paradigm with folksonomy like assessment of published work, these factors would be greatly improved and extra measures would be available. Citation analysis would have greater accuracy due to the greater accuracy of semantic citations and the possibility of tracking ideas through the network of published material <sup>12</sup>. Journals could, for example, take on the role of filtering published science, essentially providing a trust, verity and relevance service. With improved citation analysis, measurements of impact factor would also be improved. The concept of a *deconstructed journal* presented by Smith [135] is similar to this idea.

Great care needs to be taken in designing systems for verity and trust management. The measures obtained can have a significant effect on the lives of scientists and funding for scientific institutions and projects. Ineffective or biased measures curb the effectiveness of scientific research, rewarding some who contribute little while punishing others who would contribute greatly, given the opportunity. In depth studies of social structures, similar experiences perhaps from open source, trial and error experience from existing experimental publishing efforts are appropriate.

### 3.4.7 Automated Reasoning and Scientific Knowledge

In Section 2.2.2 we saw that mathematics is fundamental to science and that current automated reasoning systems are weak when it comes to reasoning on mathematically formulated knowledge. There will always be queries in such a system for which the reasoner will keep looking for an answer forever in vain.

If a reasoning system were available that could reason on mathematical relationships, it could provide a number of useful services. Checking the consistency of the knowledge base is perhaps the most fundamental - users could be alerted when a newly found result contradicts some existing theory or combination of theories. For example, we would like the system to ensure that the force on an object is equal to its

---

<sup>10</sup>This is, of course, true in the current system as well, however with the possibility of smaller published elements its effect is multiplied.

<sup>11</sup>This is also true currently, with pre-print archives, however, again, the smaller published elements would magnify the effect.

<sup>12</sup>Currently, if you publish a great idea, then someone else publishes a famous paper based on your idea, you get no credit for citations of the famous paper. With semantic citations, the origin of the idea is not lost.

mass multiplied by its acceleration (newtons second law) and inform us if some data was added for which this wasn't true.

Another potential service a reasoning system could provide is discovering new theories that are implied by the existing body of scientific knowledge, but as yet not recognised. There is current research into automated theory formation in mathematics that shows promise for the development of such a system [26].

One service that can be automated, however, is checking the correctness of proofs. If a theoretical derivation of a new theory exists, that derivation can be automatically checked for correctness. Assuming the consistency of the theorems that are called upon in the derivation, the a new theory with a valid derivation will not break the consistency of the body of knowledge. There are several mathematical representation languages and attendant theorem provers that would be up to this task, for example MIZAR [100] and Isabelle [74]. TPS [2], a semi-automatic proof writing assistant is also interesting in this context.

### 3.5 Other Areas of Application

Though the focus of these discussions has been on scientific publishing, many of the conclusions and considerations would apply equally well in other publishing settings where a collection of knowledge is used and maintained by a community.

Some obvious examples where semantic publishing could be beneficial are:

- business and administrative processes
- engineering processes
- community interest groups
- educational tools - semantic wikis and KB interfaces for learning
- maintenance of ontologies in semantic grids
- evidence-based public policy formation

There are no doubt many other application areas and will likely be many more in the future - The flow of knowledge is an important part of a functioning society [160].

### 3.6 Potential Barriers

There are many barriers to the potential widespread adoption of semantic publishing in science. It is likely that we may see several generations pass before it happens, and also possible that other technologies and ideas appear that will make the vision of semantic publishing obsolete. For example, natural language processing and knowledge tracking techniques may improve to the extent that semantic publishing is not necessary. None the less, our need to better organise our knowledge is great, and we

---

will very likely see substantial changes and advances in the way we do that, in science and in general.

In this section, I describe some of the barriers that may block the birth or spread of semantic publishing. I do not claim to have predicted all the factors that may act against semantic publishing, but I have identified some significant issues.

Arguably, it is the social barriers that will be the most difficult to overcome.

### **3.6.1 Cultural Momentum: Why Change the Way we Publish?**

The world is full of great ideas that promised better results than some existing system, but which never came to fruition. A good idea is not the only requirement for cultural change (indeed, it is not even a necessary requirement!). People who have an effective (if not efficient) way of fulfilling a need are wary of changing it. People under time pressures are reluctant to invest time in new ways of doing things, even if there is a promise of more efficient use of time as a result. It is likely that scientists will not be interested in quickly changing the way in which they publish their work. Their focus is on the research, and learning a new way of publishing will generally not be a high priority.

To overcome this resistance, systems should be designed to be easily operable, ideally with some degree of backward compatibility so that known publishing skills can still be applied, as discussed in Section 3.2.5. Semantic publishing tools are needed that are intuitive for scientists to use and which make it easy to locate and insert the appropriate citations.

Though it may be difficult to ask the current generation of scientists to change their ways, perhaps future generations will find such publishing techniques more familiar. The youth of today (at least in the developed countries) have been described as “digital natives”. Many young people in China are active internet and social networking users [70]. These young people are already creating digital content in new and innovative ways—sms messages, social networking, folksonomies and web 2.0 mashups. They are inventing new ways to interact with and over the Web. I feel that it is likely that they will embrace and extend knowledge technologies, perceiving and pursuing the advantages of greater knowledge availability as has happened with the advent of the World Wide Web. Research into this aspect of the evolution of our cultures and into the tools that are generated by it would be relevant to semantic publishing.

### **3.6.2 Cultural Adaptation: Learning new Tools**

In the previous section I talked about resistance to change. Another related barrier to the adoption of a new system is the investment required for a community of users to adapt to it. Here I mean adaptation in a pragmatic sense: individuals learning new cognitive models and system interfaces, and communities developing new social norms. This is related to resistance to change, systems that are difficult to adapt to will be resisted more, but not the same. Resistance may be overcome, but the learning curve for a new system will remain!

In the case of semantic publishing, we might expect substantial difficulty for individuals to learn new conceptualisations of the authoring process and new interfaces. As noted above, carefully and intelligently designed interfaces and processes could reduce the difficulty.

Social changes related to semantic publishing are perhaps harder to predict. We might expect, however, that these changes would not be trivial. Publishing is an integral part of the scientific community, driving social institutions such as standing, prestige and wages. Some investigation of the social evolution of science and the potential impacts of semantic publishing could aid us in understanding the social barriers to its adoption and guide us in designing semantic publishing systems.

### **3.6.3 Semantic Markup is Hard Work**

We saw in Section 2.3.2 that alongside the development of Semantic Grids for eScience, new web or Grid based collaborative tools for science are being created. Many of these tools collect and organise provenance data from the collaborative process. This data represents the development of ideas by the team and is organised according to conceptual aspects of the project. It would be a small step for these annotations to be represented in semantic publishing form, and as such they would become the semantic pieces of the final published result. In this way, publishing would be a process of gathering and organising and connecting the already developed pieces, thus much reducing the task of semantic markup that plagues Semantic Web initiatives.

A recent paper in Nature talked about the need for biological scientists to learn to use modern knowledge management tools:

In order to become active and effective contributors to the cyberinfrastructure, biological researchers will need to become familiar with the basics of computer science, learn to use ontologies to describe their data and protocols unambiguously, and have the skills to put this information in a form that can be readily adapted and reused by others in the community. [138]

This would hold true for any branch of science in which Grid technologies are becoming widely used, a list that is and will likely continue to grow into the future.

Other interesting avenues for the creation of represented knowledge are data mining and natural language processing. Recent research in this area is promising. For example Shaparenko and Joachims [130] have been able to track the flow and of ideas in collections of documents containing text. Encouragingly, their work was able to reassemble citation graphs in corpora of scientific publications with high accuracy. Another notable example is the work of Hu et.al. [69], who have developed an effective tool for extracting knowledge from large corpora of scientific publications. Other natural language processing initiatives in science with differing degrees of success include [116; 107; 144].

Though currently unable to provide rich semantic interpretation of written documents, there is hope that in time these and related fields will be able to provide tools that will 'understand' natural language texts (in the sense of automatically generating

---

a faithful semantic representation). This would allow scientists used to current authoring media to publish semantically with little extra effort. In the words of Carole Golbe: “Machines will be reading the library” [58]

### 3.6.4 Early Adopters

It is worth thinking about adoption paths - who will be inclined to adopt a new technology before it is widely used? Can the technology be built in stages, each of which may be more easily adopted and lead on to the next?

We have seen that there will be some areas of science that are already embracing sophisticated knowledge management technologies and/or published data provenance. Scientists in these areas will have less work to adopt a semantic publishing approach, and will be more familiar with knowledge management tools in general.

A few potential early adopters could be:

- Users of bioinformatics technologies. Tools based on knowledge representation and automated reasoning are in use in bioinformatics. Familiarity with the use of semantic tools may make these researchers amenable to semantic publishing as an effective means of maintaining the currency of the tools.
- Data intensive research areas such as plasma and particle physics, quantum chromodynamics and climate science. Semantic grid technologies are being deployed in many such areas. In these areas, many of the published results are found by post-processing experimental or simulation data [51]. In a semantic grid, it is advantageous for these results to be incorporated into the grid ontologies. This would provide a base infrastructure for semantic publishing and provide a sensible mechanism for maintaining the grid’s semantic structures.
- There are some recently initiated projects to create “mathematical wikis” drawing upon knowledge representation standards and initiatives in mathematics [86; 48]. These embody many of the potential benefits of semantic publishing described above. If such efforts succeed to represent mathematics on a large scale, they could become a platform for semantic publishing in mathematics and lead to similar platforms in other sciences. This is, however, unlikely in the short term.
- There has been some work in education to provide relevant syllabuses and content to students given their learning goals and prior knowledge [15; 11; 37]. Such systems could be relatively easily extended to include a semantic publishing element for course maintenance and as a collaborative educational tool for students. This second use is important, as students represent the next generation of scientists - experience with new and effective technologies will make them more inclined to support their deployment later in their professional lives.

A more thorough investigation of incremental development paths and possible early adopters would be beneficial.

### **3.6.5 Technical Barriers**

Though much of the technology needed to implement semantic publishing exists, there are many research questions that will need to be answered before semantic publishing can become a reality. I will summarise the research directions identified in this thesis in detail in Chapter 6—among these are the technical barriers to realisation of my semantic publishing vision. Suffice to say that the main areas in which technical development is needed before semantic publishing can become a reality concern knowledge representation schemas and systems for science, semantic publishing tools and initial reference knowledge bases.

---

# A Framework for Evidential Reasoning

---

In this chapter I investigate some initial formalisms that would be fundamental to representing scientific reasoning and knowledge. Please note that these are but small steps toward a large goal. A proper investigation should delve deeply into the nature of scientific reasoning and into techniques and formalisms for knowledge representation before presenting a framework for representing science. Such a study, even of a very preliminary nature, would require significantly more effort than the scope of this thesis permits. It is my hope, however, that the ideas presented here could form a part of such a study in the future.

I present a framework for evidential reasoning. It would be well to consider frameworks for representing scientific knowledge in general, however that also falls beyond the scope of the current investigation. Mention of some current attempts has been made in Section 2.2, and Section 3.2.1 has a short discussion of some issues facing their development.

Fundamental to the scientific method is the development of theories about phenomena from which predictions can be made, and subsequent development of repeatable experiments whose results match those predictions. Scientific publications often contain these elements. I present here an abstract framework that captures this process while making few assumptions about how that knowledge is represented.

## 4.1 Fundamental Concepts and Definitions

Science is, at its deepest level, an attempt to apprehend the nature of the world in which we live. Acknowledging that individual perception is imperfect, science attempts to achieve objectivity through repeatable *experiments* and *observations*. In apprehending the nature of the world, science attempts to understand the *processes* in the world. It attempts to *categorise* the things in the world and determine *relationships* between them <sup>1</sup>.

---

<sup>1</sup>*Ontology* (on the nature of things) and *epistemology* (on the nature of knowledge) examine these concepts in greater detail. I will not delve more deeply into these topics here.

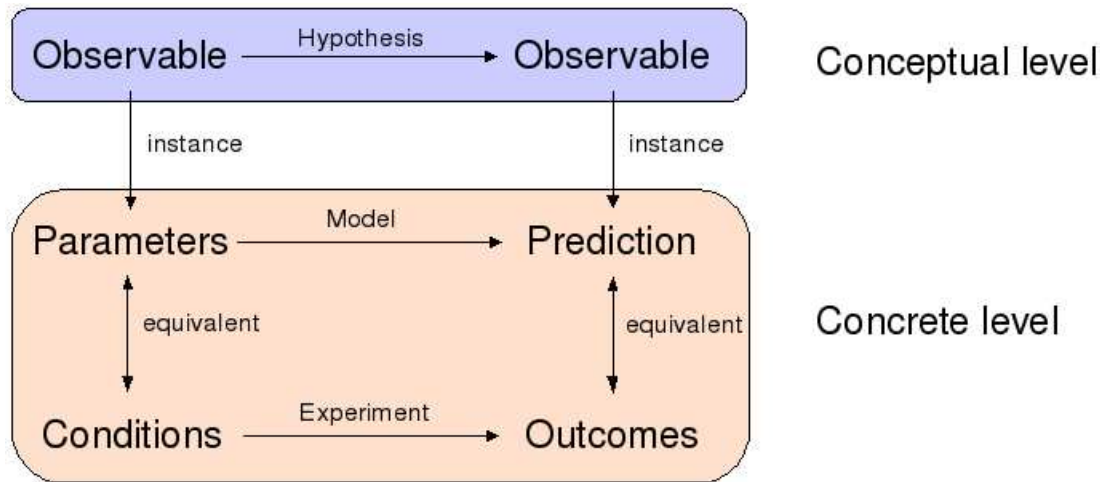


Figure 4.1: Evidential reasoning

From this perception of the nature of science, I propose the following basic conceptual framework for representing scientific arguments based on evidence. I do not claim that this framework is exhaustive—that it is effective for representing any argument presented by science. For example, the classification of rocks in geology could conceivably be moulded to this framework, but it would not be a particularly natural fit. Also, theoretical arguments are not covered here.

**A Hypothesis** is a rule connecting and constraining measurable *quantities* or observable *qualities* that can be seen in the world. A hypothesis operates on a conceptual level. It attempts to capture processes and relationships between the things we can observe. A theory will generally start life as a hypothesis. By repeated application of evidential arguments, it will attain greater certainty and we may start to refer to it as a theory (or it may be disproved!).

**Observations** are quantities or qualities that we can measure or observe in the world. They are concrete conceptualisations of things we believe to exist or have existed in the world. They are the results of *experiments*. They may be divided into *initial conditions* and *outcomes* (see Figure 4.1), though this will not always be the case. Observations will generally carry some level of uncertainty.

**An Experiment** is a recipe for measuring or observing the aspects of the world. An experiment will often involve an apparatus (with appropriate instructions on how to build it). It should be consistently repeatable, producing the same observations each time<sup>2</sup>. It should be designed to contradict the hypothesis as much as possible if the argument is to be strong.

I am using experiment here in a fairly broad sense, including mundane activities such as taking measurements or reporting the colour of a rock. The essence

<sup>2</sup>We will see in Section 4.2.3 that repeatability is not always possible.

---

here is the capture the connection between the objective world and our understanding of it, the process through which our understanding is molded by the world.

**A Model** of a hypothesis makes *predictions* about the outcomes of an experiment. It is through models that the other half of the dialogue with the world takes place—we say to the world “ah! I think I understand! Is it like this?” and through our experiments, we receive a reply. A model will often use sophisticated (perhaps numerical) techniques to make predictions—these techniques should be well tested and thoroughly understood.

**A Prediction** is the product of a model. It will generally carry some sense of imprecision, either through imprecise numerical techniques and/or imprecisely measured initial conditions from which the prediction was made.

There are situations in which predictions may not be made per se. For example,

**An Equivalence Relation** is a process for determining whether a prediction from a model is acceptably similar to an observation from an experiment. The notion of acceptability is highly subjective—the intention here is to capture the accepted standards within a field. It is envisaged that equivalence relations will be publishable entities in their own right. Since we believe our observations are never perfect, an equivalence relation will always be imprecise. For example, statistical tests may give us a p-value (a probability that equivalence is not met) or measurements may be equivalent within some margin of error.

**An Evidential Argument** links these parts together. It is a hypothesis, a model of the hypothesis and matching experiment, the predictions of the model and the observations of the experiment, and the assessment of an equivalence relation about the similarity between the predictions and observations.

You may notice that this framework does not include the things in the world that we think have some objective existence. Things like trees, clouds and atoms. These things are fundamental to our scientific knowledge, however I claim that they are not necessarily fundamental to evidential reasoning. It is in our formulation of hypotheses, by theoretical derivation or otherwise, and in experimental design that we think about the things in the world. Evidential reasoning as presented here comes after that.

This framework attempts to capture essential conceptual components of the way we link our conceptual understanding of the world (our hypotheses and theories) with our observations, and the techniques we use to make that connection. The terms I have used to label those components have many uses in and out of science. In some areas of discussion, the terms may have uses that are quite different to the way they are used here. We will see such differences in some of the examples described in the following sections.

Each of these features would be represented as concepts or their instances. The representations would have attendant provenance information. There would also be

other attached information such as semantic citations and textual descriptions (perhaps in several languages).

In the following sections, I will use examples to illustrate the way this framework could be applied and begin to examine in more depth the sorts of information that would be stored in their representations.

## 4.2 Illustrative Examples

This section contains illustrative examples of how the proposed framework for evidential reasoning could be applied. These examples show the flexibility and broad applicability of the proposed framework. They also show that modelling scientific arguments is complex and requires a flexible approach to its implementation.

### 4.2.1 Newton's Second Law

To illustrate the conceptual framework above, let us consider a hypothetical experiment to test Newton's second law of classical mechanics. Our experimental apparatus consists of a wheeled trolley on a flat table, a pulley on the edge of the table, a light and inelastic thread tied to the trolley and running over the pulley, and three weights that can be either tied to the other end of the thread or placed on the trolley. There is a mechanism to hold the trolley in one place, and release it as required. We also have an accurate acceleration measuring device. In the experiment, we will try two combinations: weights 1 and 2 on the trolley and weight 3 the end of the string and then move weight 1 from the trolley to the end of the string. For each combination, we will measure the trolley's acceleration when we release it. All these components are tested to ascertain if factors such as friction in wheels and the pulley, vibrations caused by the release mechanism, the flatness of the table etc... and we are convinced that there is not observable bias. Our accelerometer is from a trusted manufacturer of scientific equipment, and has their stamp indicating recent calibration.

In this simple experiment, the elements of the evidential reasoning framework are easy to identify:

**Hypothesis:** Newton's second law  $F = ma$  will govern the behaviour of the weights and trolley. Its representation would have a formal representation of  $F = ma$  linked to the concepts of *force*, *mass* and *acceleration*. It would reference the evidential argument as evidence. Since this is a fundamental law, its *theory provenance* could be quite simple: Issac Newton would be indicated as the author. Perhaps a derivation from general relativity may have been added, given appropriate assumptions about speeds, masses and distances given which Newton's second law is a valid approximation.

**Model:** We expect that the acceleration will vary proportionally to the weight on the thread. The representation of the model in this case could be a short program in python that accepts two numbers tagged as a masses 1 and 3, then outputs

---

a number tagged “ratio of accelerations”. All these numbers would have error margins. This number would be calculated as  $m_3/(m_1 + m_3)$ . We would calculate this using a (trusted) package that propagates error margins in simple arithmetic calculations.

**Experiment:** The representation of the experiment would contain a detailed description of the experimental setup and steps taken to ensure there is no bias. Since this is a common high school experiment, perhaps many of the details may already have been partially encoded in an online physics text. We could refer to this and fill in any placeholders and with details specific to our setup.

**Observations:** We observe four numbers, each with an indication of margins of error: the mass of weights 1 and 3, and two acceleration readings. The representation of the observation would contain the masses of weights 1 and 3 and the ratio of the two acceleration readings. These values would be tagged as masses 1 and 3, and “ratio of accelerations” respectively, and would be associated with appropriate margin of error values (also represented). The provenance information would contain our names (or perhaps open Id’s), the date we did the experiment, a reference to the representation of the experiment, an indication of the technique used to calculate the ratio and its margin of error.

**Prediction:** This would be a number generated by the model when fed the representation of the observation—ie: the result of  $m_3/(m_1 + m_3)$  for these weights. The model would ignore the acceleration measurements, since it does not recognise their tags. It would have provenance information referencing the observation and the model (and, indirectly via the observation, the experiment).

**Equivalence Relation:** The equivalence we are investigating with the setup thus far is a simple comparison of two numbers with error margins. It would be a program that accepts two numbers with error margins and output its assessment as “equivalent within margin of error” or “different”. Its representation may reference some statistical theories that justify its application (I plead ignorance here, but I think such a comparison has no derivation but an appeal to common sense!—in that case, the appeal would be indicated).

**Evidential Argument:** This is represented as a container object that associates the other elements of this argument. An automated consistency checking service could examine the tags of its components, checking that they match, then execute the model on the observations, check the result against the prediction, apply the equivalence relation and add a tag indicating “valid” and “does not refute”, “valid” and “refutes” or “invalid”.

There are a couple of things to note here. We could have done away with the prediction: the model could contain only formalised mathematics and the equivalence relation could be made to implement mathematical formula applying error propagation. Also, the hypothesis could have been posed in a more specific way, perhaps as

the mathematical relation used in the model. In this case, the model becomes trivial. We can see that there is some flexibility in how to implement this framework. Poppers principles [120] could be called upon to insist that the hypothesis should be as unspecific as possible and thus choose the less specific formalism, however there will often be arbitrary choices about how to represent the parts of an evidential argument. In the following examples, that will become more clear.

#### 4.2.2 Statistical Tests

Statistical tests are often used to determine if a change in conditions effects some phenomenon in a significant way. As an example, lets examine an experiment to test whether a new document scrolling technique improves user search times over a standard scrollbar. A group of people are chosen and given search tasks with different scrolling techniques and the time it takes them to perform the tasks is recorded.

**Hypothesis** In this example, the proposed *hypothesis* states that the new scrolling technique will result in faster search times than a scroll bar. This would be represented as a logical assertion. It would also indicate its author(s) and list its supporting evidential argument(s)—if the current experiment is not the first that tests this interface, it may have more than one supporting argument. It may cite other theories concerning, for example, document navigation and/or other scrolling techniques. It could cite cognitive and behavioural studies that help support its validity on a theoretical level.

**Experiment** The *experiment* consists of a description of how the participants were chosen, tasks performed etc... Details of steps taken to design the experiment to minimise bias in the results would be included. Some features of the experimental design may draw on established results, for example standard design techniques or general psychological or ergonomic results about computer interaction. Perhaps these standards could be combined as a single user interface experimental design standard. The representation of the experiment would contain references to these and indicate how this experiment complies with the standard.

**Observations** The *observations* are the collected search speed measurements and information about the participants. A reference to or details of the experiment would be part of the provenance of the observation data.

**Model and Prediction** In this case, it is hard to distinguish the *hypothesis*, the *model* and the *prediction*. We could say that the hypothesis asserts faster search times in general whereas the model or prediction asserts faster search times in this experiment. Doing this does not, however, enrich the representation of the argument. It would make more sense to dispense with the model and prediction when representing this type of argument.

---

**Equivalence Relation** The *equivalence relation* here would be the statistical test. Its representation would first indicate the data structures the test requires and provide an indication of how to apply the test. This would likely be a concept in a statistics ontology that statistical packages could interpret along with the data to perform the test. It could equally be a mathematical formula or the URI of a web service that can perform the test. The representation would also contain information about the requirements of the test. For example, a strong T-test could insist on a minimum sample size and some measure of the ‘normality’ of the data (a weaker T-test might relax these constraints to some degree). Standards for experimental design that help to ensure that the test is valid would also be referenced.

**Evidential Argument** The *evidential argument* ties all these elements together. It asserts that the conditions of the hypothesis match those of the experiment. It asserts that the experiment satisfies the requirements of the equivalence relation. It references the data and contain the result from application of the equivalence relation to the data. That result would have provenance information indicating when and where it was calculated.

It is the evidential argument that the scientist would publish.

### 4.2.3 Climate Modelling

Climate is commonly defined as a set of weather statistics computed from instantaneous data over a long period—the standard definition is thirty years—of time[55]. Modelling the climate is a daunting challenge; equations of evolution for climate statistics do not exist per se, and one must instead model the instantaneous behaviour of the system, log its history, and then compute climate statistics using this history.

Climate models are large and complex computer programs, comprising sub-models for atmosphere, ocean, sea-ice, and land-surface (vegetation and soil), with all of these sub-models evolving in mutual interaction. Atmosphere general circulation models (GCMs) comprise fluid flow based on Navier-Stokes, radiative transfer, and also transport of water vapour. Ocean GCMs model ocean momentum flow based on Navier-Stokes, but also must track salt transport and its effect on ocean thermodynamics. Sea ice models capture the thermodynamics of sea ice formation and melting and the dynamics of large masses of sea ice, including ice pack motion as a plastic flow and deformation properties—that is, its rheology. The inherent complexity of these models is recognised in the common technical term used to describe them: *Coupled Climate Models*.

Coupled climate models are used in *numerical experiments*. That is, the model is used as the apparatus, and run with carefully chosen data (e.g., initial or boundary conditions or modified parameters), or modified in some way to explore model behaviour. The four main types of numerical experiments performed in climatology are: 1) sensitivity experiments; 2) internal variability studies; 3) process studies; and 4) studies of past climates (a.k.a. paleoclimate). Model sensitivity experiments are

performed to gauge response of the model to some changed condition. Perhaps the most famous model sensitivity experiments are those cited in the Intergovernmental Panel on Climate Change (IPCC; <http://www.ipcc.ch>) reports such as last year's Fourth Assessment Report [73], which measure climate sensitivity to increased atmospheric carbon dioxide concentration.

Process studies are performed by modifying some low-level (but important!) process model in the system—for example parameterisation of convection—and performing simulations to determine whether the new representation of the physical process yields overall a better climate.

Internal variability studies are simply very long runs (many decades to many centuries in duration), and studying patterns of interannual, interdecadal, and centurial variability, comparing where possible with available climate data.

Paleoclimate studies are climate model integrations run with conditions corresponding to past geological eras. Continental masses may be shifted to replicate land-mass distribution during past eras such as the cretaceous period, or ice masses such as those found in the Late Pleistocene.

In all of these types of numerical experiment, a *control simulation* may be run along with other modified simulations to assess impacts of model or forcing data changes. In the case of the IPCC climate sensitivity experiments, the control simulations are simulations of twentieth-century climate, using historically accurate atmospheric carbon dioxide concentrations, and temporally correct forcings due to variability in solar input (i.e., sun spots) and aerosols due to volcanic activity.

Comparing control simulation results from climate models with “reality” is a complex problem. Models typically present their data on regular grids (e.g., logically rectangular latitude-longitude grids), whereas meteorological observations from surface stations, sounders, and balloons tend to occur where people have settled (along rivers and coastlines) in a spatially non-uniform fashion!. Satellite data is more regularly distributed, but only along the path of a satellite's orbit. Aircraft and ship observations are available along commercial aviation and sea routes. Taken together, these instruments form the world's weather observing system.

The data stream from this network can be combined with weather model calculations and statistical analysis to produce results whose spatial distribution is more uniform and readily comparable with a model spatial mesh. This process is called *data assimilation*[80]. In addition to instantaneous data assimilation products, there exist data products detailing estimates of the meteorological state at gridded locations and regular time intervals (typically six-hourly) for long periods stretching from the mid-twentieth century to the present. These data sets are called *retrospective analyses* or more simply *reanalyses*. Major reanalysis efforts in the United States and Europe have produced publicly available reanalyses that can be used for climate model evaluation [81; 83; 148].

**Hypothesis** The hypothesis is that a numerical climate model constructed from first principles, and run using present-day values for relevant parameters (in particular atmospheric carbon dioxide concentration) is capable of reproducing climate

---

statistics at the global scale (e.g., global means) and in terms of geographic distributions (e.g., correct placement of maxima and minima such as high and low pressure centers). The hypothesis typically would state that the model reproduces observed average statistics (the mean) and variability (the variance) for important meteorological variables such as surface temperature and precipitation, mean sea-level pressure, et cetera.

The theory provenance of this hypothesis would include the collection of all of the theoretical concepts implemented by the climate model, including (but not limited to) the primitive equations for atmospheric and oceanic flow, physical laws governing radiative transfer, physical laws governing phase transitions for water, bulk energy, moisture, and momentum formulae found in a land-surface model, etc... Also included in the body of theory are the approximations made in mathematical arguments to derive these formulae.

**Model** The model in this case is the climate model, but also includes post-processing of daily model history output to create climatologies. Typical post-processing initially entails computation of a series of monthly averages of model output, followed by subsequent processing to create long-term mean monthly and seasonal averages and variances. Note that these data are computed on the model's spatial grid.

**Prediction** The prediction in this case is the raw model output, along with any post-processed data to construct climatologies.

**Observations** For a control experiment, reanalysis data sets mentioned above are the data used as the "ground truth" against which model results are compared, and are presented as long-term monthly and/or seasonal averages and variances. As such, it is these data sets that fill the role of *observations*.

It is important to note that these statistics are computed on the spatial grid used by the data assimilation system used to generate the reanalyses. That spatial grid will, in general, be different to the grid used by the model.

**Equivalence Relation** The equivalence relations used can vary in levels of complexity. A widely used and simple technique is to plot fields of data and analyse them visually by comparing model alongside observations. Based on visual analysis, one can claim something is "good enough". A more rigorous approach is interpolation of the model data onto the observations' grid, or vice-versa, and computation of differences. The resulting difference fields can be analysed either visually or numerically and criteria for equivalence assigned accordingly.

**Evidential Argument** The evidential argument combines the above elements, using the Equivalence Relation to test the Hypothesis with the Observations and Prediction.

It is clear that the data we are calling observations has undergone a great deal of analysis. Climate modelling is perhaps an extreme case, however it is usual that observational data will receive some level of treatment before being presented as evidence. One other example is statistical treatment of data before comparison with a prediction. It is important that the analytical treatments applied to observational data are trusted, and that the data's full provenance is available.

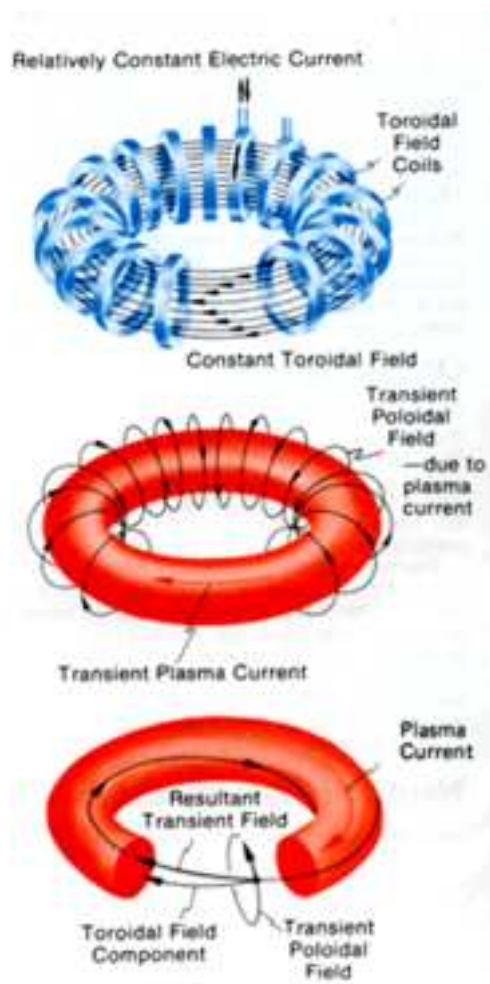


Figure 4.2: Depiction of plasma in a tokamak taken from [154].

#### 4.2.4 A Plasma Physics Example

In a prize winning plasma physics paper, Angioni et al [4] intuited a subtle, but relatively simple phenomena in tokamak plasmas. Their paper gives a sophisticated theoretical derivation and a survey of plasma 'shots'<sup>3</sup> from the ASDEX Upgrade divertor tokamak (Axially Symmetric Divertor EXperiment). Plots showing the distri-

<sup>3</sup>A shot is one plasma experiment. Gasses are added to the tokamak chamber and heated. Various measurements of the plasma are made until the plasma breaks down.

bution of shots according to selected properties are presented. The reader is invited to note visual features of the plots as evidence supporting their ideas. Their result is an important observation of plasma behaviour, and may help in the design of fusion reactors to generate electricity at some point in the future.

In highly condensed form, the core contribution of this paper is a proposed mechanism involving electron thermodiffusion to account for observed anomalous ‘flattening’ of plasma density, or *pump out*<sup>4</sup>, in response to central electron cyclotron heating (central ECH). The theoretical treatment suggests that when *trapped electron mode* (TEM) instability is dominant over *ion temperature gradient* (ITG) instability in the plasma, electron thermodiffusion will cause pump out.

Their theoretical derivation involves techniques for approximating the full mathematical treatment of the equations governing the behaviour of plasma as well as numerous numerical simulations. As with the central evidential argument of the paper, plots of these simulations are presented, and the reader is invited to recognise features that support their derivation. I will not go into the representation of this argument—suffice to say that it would have a similar form to the main evidential argument. It is interesting to note that this paper contains many linked arguments and conclusions. A full representation of the paper would contain each and indicate the connections as semantic citations. These citations would form part of the theory provenance of each conclusion.

Angioni et al presented a number of finer points of physical evidence and logic, however it is their central argument that I will describe here—it is summed up in the following statements from the papers conclusion:

... from a theoretical viewpoint, it has been shown that in the framework of the theory of ITG and TEM instabilities, outward electron thermodiffusion is predicted when the most unstable mode is a TEM, and the mode propagates in the electron drift direction. On the contrary, when the dominant instability is an ITG and the mode propagates in the ion drift direction the electron thermodiffusive flux is directed inwards and is smaller. A mechanism has been presented which leads to density flattening with increasing central electron heat power.

...

In support, it has been shown that plasmas exhibiting density flattening in response to central ECH are found to be in the domain of a dominant TEM instability. On the contrary, plasmas which do not show significant profile modifications during central ECH, are found in the ITG instability domain, while in almost the totality of these plasmas TEMs are stable.

**Hypothesis** The *hypothesis* of this evidential argument could be presented in first order logic. Let *ITG* and *TEM* assert that ITG (respectively TEM) is the dominant

<sup>4</sup>A plasma in a tokamak has a toroidal or doughnut shape (see Figure 4.2), typically with greater density in the middle of the doughnut ring. When *pump out* occurs, material shifts away from the middle of the ring towards the ‘skin’ of the toroid. See figure 4.2.

instability mode, *ECH* assert that a plasma was given ECH and *pump out* assert that density flattening occurs. The hypothesis would then read <sup>5</sup>:

$$\text{central ECH} \wedge \text{TEM} \Rightarrow \text{pump out}$$

$$\text{central ECH} \wedge \text{ITG} \Rightarrow \neg \text{pump out}$$

The representation of the hypothesis would also include citations (with 'depends on' flavour) of the theory of ITG and TEM instabilities.

**Experiment** The *experiment* in this case (in the sense of the conceptual framework) is most naturally viewed as a collection of plasma shots (with central ECH) that were examined to compile the presented figures and the (presumably trusted) analyses used to determine which shots exhibited ITG or TEM and which showed anomalous pump out. In terms of the framework, each shot has an experiment entity and an observations entity of its own. The representation of the experiment entity would indicate the tokamak control parameters and (probably in the form of links) the design and specifications of the ASDEX Upgrade divertor tokamak. The representation of the observations entity would have the data gathered during the shot and provenance information such as a link to the experiment entity, the scientists involved in the shot, any preprocessing done to the data etc..

**Observations** The *observations* referenced by this evidential argument are the presented plots. It is these plots that will be compared to the predicted relationship between central ECH and ITG and TEM instabilities. There are several types of plots presented as experimental evidence, but they all show data from one or more shots and indicate the presence or lack of pump out in response to central ECH, and an indication of the dominant instability mode. The representation of these observations would include provenance data for each shot presented in the plots. It is worth noting that Angioni et al referenced shot numbers for all of their plots, thus providing provenance information.

**Equivalence Relation** The *equivalence relation* is an appeal to the visual judgements of the reader. This is a common technique used across many areas of science. Though this method does not lend itself per se to automated verification (since it explicitly requires human perception), it can still be represented. Someone browsing the represented knowledge would be presented with the plots and an indication of what is meant to be seen in them. Data mining, machine learning or statistical techniques could also be applied to the data as a means of automated verification and to provide an objective measure of the strength of the reported features.

<sup>5</sup>The notation  $A \wedge B$  indicates that assertion  $A$  and assertion  $B$  both hold.  $\neg A$  indicates that  $A$  does not hold.  $A \Rightarrow B$  indicates that  $A$  implies  $B$ .

---

**Model** In the text of the paper, Angioni et al describe the features of the presented plots and explain the ways in which they support the core hypothesis. These descriptions embody the *model* of this evidential argument as they connect the hypothesis to the presented evidence. In fact, Angioni et al present several sub-arguments, one for each of a number of plasma types—effectively several models. What constitutes the *predictions* in this argument is difficult to discern. It would probably be fairest to say that they are embedded in the models. Another approach would be to say that the predictions are left to the reader examining the plots.

This paper is interesting in that it presents a very sophisticated theoretical derivation of what is, in essence, a very simple hypothesis. An attempt to represent the theoretical derivation would likely be a complex and difficult task, as many sophisticated mathematical and numerical techniques are applied.

Another interesting aspect of this paper is that it does not claim to be conclusive. Angioni et al are careful to point out the limitations of their analysis. For example they write:

The flattening predicted by quasi-linear theory appears too small also with respect to the experiments. A conclusive quantitative comparison between the theoretically predicted magnitude of the thermodiffusive  $u_x$  produced by ITG and TEM instabilities and the measured modifications of the density profile in response to central electron heating requires full non-linear simulations, taking into account collisionality effects.

It will often be the case that presented scientific work is inconclusive, either without evidence (ie: a conjecture) or with limited evidence (but enough to suggest further study). The need for grades and shades of verity in a knowledge representation system for science, as discussed in Section 3.4.6, is apparent here.

#### 4.2.5 Kon Tiki

In 1947, Thor Heyerdahl sailed a raft made of from balsa wood and other native materials from Peru, attempting to reconstruct ancient designs, and sailed it across the Pacific Ocean to Polynesia. The raft was named Kon Tiki [67].

Heyerdahl had a theory that a race of fair skinned people, referred to in Incan legends, may have colonised Polynesia around A.D. 500. He had several reasons to think this a possibility: reports of fair skinned people in Polynesia often with fair hair and hooked noses, unlike most Polynesians; South American sweet potatoes that were a staple in Polynesia and more. In order to add credulity to his theory, he built Kon Tiki and sailed it across the Pacific, thus showing the plausibility of such a voyage in prehistoric times. According to Wikipedia, most archaeologists still believe, based on genetic, linguistic and physical studies, that Polynesia was settled from Asia. However, evidence such as the sweet potato suggests that there may have been some South American contact.

The evidential argument he presented with his voyage is relatively trivial to represent with the proposed evidential reasoning framework. The *hypothesis* would read something like “the legendary people of *Kon-Tiki* could have sailed to Polynesia”. It would be represented as a logical proposition linking the concepts “Polynesia” and “the legendary people of *Kon-Tiki*” with the qualified action “could have sailed to”. Each of these concepts would have its own provenance and descriptive information, for example the people of *Kon-Tiki* would be linked to the Incan legends and who reported them, archaeological evidence from Peru, the fact that they were purported to live in Peru etc.. If Heyerdahls theory of colonisation were represented elsewhere, the evidential argument from the voyage would be linked to it as supporting evidence.

The *model* would read something like “I can sail to Polynesia from Peru in a boat similar to what the people of *Kon-Tiki* might have had”. The *prediction* would read something like “If I build a boat similar to what the people of *Kon-Tiki* might have had, I will be able to sail across the Ocean”. The *experiment* would be a description of all the steps taken to research and build the boat, and then to sail it across the Ocean. The *evidence* would be “we built *Kon Tiki* and sailed it across the Ocean according to the specifications in the experiment”. Attached to each concept, such as “we” and “*Kon Tiki*”, there would be appropriate extra information and provenance.

Though the formalisation of this argument may seem trivial, it shows the flexibility of the approach and demonstrates its ability to capture the link between our interactions in the world and our conceptualisations of how it may be, even in our simplest arguments.

#### 4.2.6 Classification

Many areas of science are concerned in part with classifying objects in the world, in fact it could be argued that this is an element of *all* areas of science. Geology, botany and biochemistry are areas where classification is of particular interest. They seek to classify things in the world according to sets of criteria, and justify those classifications according to observed differences between things in the world. At times, there will be exceptions to the classification rules: objects that fall neatly between classes or share features of several classes, or continuum’s that connect classes.

Evidential arguments around classification are perhaps the hardest to fit into this framework, and probably would be best treated with a different framework. Lets have a look at how the evidential reasoning elements described above fit into a classification setting. A theory of classification has:

- A collection of classification criteria. These are like hypotheses postulating that things will be clearly divided by the criteria.
- Theoretical derivations. Perhaps there is reason for us to expect the classification (for example evolution creates classes of organisms—except viruses are now known to move genes “horisontally”).
- Possibly some exceptions (not unlike some physical theories, for example fluid dynamics begins to fail when particle sizes or densities become too large).

- 
- Evidence: a survey of known things, noting that they classify nicely (bar the exceptions!)

Seen in this way, we appear to have a special collection of related hypotheses and evidential arguments. That means a classification framework can be built around the evidential argument framework, however extra structure would be needed for it to be practicable. Also, a more thorough analysis would be appropriate in order to identify situations or requirements that cannot be fit sensibly into the proposed evidential argument framework.

### 4.3 Summary

I have attempted to unravel the fundamental elements of the ways in which we derive our knowledge from observation of and interaction with the world.

We have seen that the evidential reasoning framework presented here is robust for representing the evidential arguments examined, but requires interpretation in unusual and sometimes counter-intuitive ways, and requires the flexibility to combine the roles in some cases. Also, the terms used for the framework elements have, in some circumstances, common usage quite different from their interpretation according to the framework. This is unacceptable as a working terminology—it is important that the terminology used for representation of knowledge match that used in the area of the knowledge being represented. For these reasons, any implementation of representational frameworks would do well to adapt to the conceptual and terminological needs of scientists in each area of specialisation.

As mentioned at the beginning of this chapter, what I have presented here is a preliminary analysis. Even in the few examples I have considered, issues concerning the nature of the framework and its ability to unambiguously represent evidential arguments were discovered. The need for larger and deeper analysis is clearly indicated.



---

## Representing Schools of Thought

---

Scientific enquiry often results in differing and perhaps contradictory ideas about the processes at work in the world. These different conceptualisations about some aspect of the world are often termed *schools of thought*. Generally, these contradictory stances are resolved through experiments designed to clearly indicate that one or the other is incorrect, however the controversy may last in the literature for some time before this happens.

Another situation in which the body of scientific knowledge may contain apparently contradictory theories can be seen with Newtonian and relativistic mechanics. Newtonian mechanics is an exact theory which works with 'reasonable' accuracy when masses, relative velocities and physical scales are within 'reasonable' limits. Here 'reasonable' could be quantified in theoretical terms and with experiments. Fluid dynamics, quantum mechanics and economic theories also have restricted domains<sup>1</sup>.

In all these situations each theory is correct (read experimentally verifiable) given certain assumptions. Though 'pure' science is interested primarily in theories that work in all situations, engineering and applied sciences are more pragmatic—if a theory works in a real application, it is worth investigation.

A widely useful semantic publishing system should, therefore, accommodate apparently contradictory theories.

In a web of semantically published science, representing different schools of thought does not present any difficulties—there is no assumption about consistency between published concepts. Inconsistencies would, however, be undesirable within knowledge bases constructed from such a web if automated reasoning were to be applied or if they were intended as a reference on scientific consensus.

One approach to this problem would be to have a collection of knowledge bases, one for each school of thought. This is a cumbersome solution however, as most knowledge would be common to many schools of thought, and combinations of schools of thought may be required—you could quickly end up with a very large number of knowledge bases.

A better approach is to switch groups of theories on and off as required. A simple trick is commonly used for this purpose. In propositional logic form, given a theory

---

<sup>1</sup>Quantum mechanical theories operating on large scales have been notoriously hard to formulate [87]

---

$T$  (where  $T$  is a conjunction of closed axioms  $A_i : i = 1..n$  - the set of rules that the theory asserts), we define a simple boolean proposition  $t$  and an axiom  $t \Rightarrow T$ . If we assert the sentence  $t$ , then  $T$  holds, and conversely, if we do not assert  $t$  or assert  $\neg t$ , then  $T$  may not hold. I will call  $t$  a *switch* for  $T$ .

A school of thought  $S$ , consisting of a collection of related theories with switches  $t_i : i = 1..m$ , could have a switch:

$$s \equiv t_1 \wedge t_2 \wedge \dots \wedge t_m$$

Asserting the sentence  $s$  would turn on the school of thought, making all the theories in  $S$  hold. Switches could be combined in any combination required. This technique could be used to effectively manage diverse and interrelated schools of thought.

We have seen in Section 3.4.7 that automated reasoning on large mathematically rich knowledge bases is currently infeasible. To implement the approach to schools of thought presented here however, some reasoning would need to be done (to calculate the consequences of asserting that  $t$ ). However, we saw in Section 3.3.3 that simplified 'snap shots' of rich knowledge bases could be used as grid ontologies, providing organisational principles to the data and other resources on a grid. This technique for managing schools of thought could be applied to such snapshots. It would be useful as an organisational principle on a grid and could also be used to filter the more expressive knowledge base, essentially making the technique feasible for mathematically rich knowledge bases as well.

---

# Future Research and Development Directions

---

In this chapter, I summarise the research areas identified in the earlier chapters. There are many research and development directions in the list, of varying scope and importance. From this list, it is evident, as we might expect, that this research has vast scope, and can only feasibly be approached by a community of researchers and over an extended period of time. Many of the research directions listed here are interdependent. Those directions that represent fundamental studies and standards will need to be completed (at least partially) before more specific areas are approached.

## 6.1 Social Research

In section 3.6 we saw that the social barriers to a semantic publishing paradigm are arguably the most significant. The research topics outlined in this section attempt to identify the nature of those barriers and find ways to ease the adaptation process were semantic publishing to become widespread.

### 6.1.1 Study the Needs of Scientists

Any technological system has as an eventual goal the service of the needs of people. This realisation has become fundamental in technology design strategies [108].

I am proposing here possible new technologies for scientific publishing. A thorough and thoughtful analysis of the publishing needs of scientists and the scientific community and the role of publishing within the overall process of science is therefore appropriate. Apart from surveys of the habits and opinions of scientists, such a study should entail a review of relevant literature, for example from the philosophy of science and studies and discussions concerning scientific publishing. The scope of such a study will likely be significant.

Identifying the needs of scientists has been identified as an important step in system (and in particular, ontological) development, and methodologies incorporating such steps have been developed and used in eScience [12]. Some results from these efforts may also be applicable in the area of scientific publishing.

### 6.1.2 Social Impacts

In Section 3.6.2 we noted that publishing is an integral part of the scientific community, driving social institutions such as standing, prestige and wages. Some investigation of the social evolution of science and the potential impacts of semantic publishing could aid us in understanding the social barriers to its adoption and guide us in designing semantic publishing systems.

A possible area where impacts of social impacts the ideas for semantic publishing presented in this thesis may be significant is the application of trust and scientific authority measures. In Section 3.4.6 we saw that such measures would be highly controversial, and great care would need to be taken to find a balance between authoritarianism and effectiveness, perhaps leaning toward less authority! For example, fine grained measures of scientific authority may be inappropriate, as scientific authority is a highly subjective consideration. Coarse measures, for example using two grades of authority—“is expert” and “is not expert” are more likely to be faithful to the opinions of most people.

### 6.1.3 Changing Roles of Publishers

In Section 3.4.6 I suggested that in a semantic publishing paradigm, journals could, for example, take on the role of filtering published science, essentially providing a trust, verity and relevance service, similar to the deconstructed journal concept presented by Smith [135].

### 6.1.4 Changing Attitudes to Web Technologies

In Section 3.6.1 I stated that I feel that it is likely that the youth of today will embrace and extend knowledge technologies, perceiving and pursuing the advantages of greater knowledge availability as has happened with the advent of the World Wide Web. Research into this phenomenon is relevant to semantic publishing.

### 6.1.5 Identifying Early Adopters

In Section 3.6.4, I identified a number of potential areas of science in which scientists may be inclined to adopt semantic publishing technologies through need or familiarity with knowledge representation and management technologies. A more thorough investigation into possible early adopters would be appropriate.

## 6.2 Knowledge Representation

Fundamental to the semantic publishing vision is effective knowledge representation formats and techniques. This is currently an active area of research, and many new results will have an impact on the feasibility and efficacy of semantic publishing. Below I list some specific topics that are particularly relevant to semantic publishing.

---

### 6.2.1 Fundamentals of Science Modelling

We saw in Section 3.2.1 that the concepts of science are many and varied, and in a continual state of flux (at least on the frontiers of research). Any system hoping to effectively capture the richness of scientific concepts needs to take special care to allow sufficient flexibility for new ideas to be faithfully represented. On the other hand, there must be sufficient structure to allow effective organisation of ideas and the application of generic knowledge management technologies.

To achieve this, a thorough study of the conceptual practices of science and the history of scientific thought would provide an understanding of the conceptual issues that representation schemes would need to face. Philosophical studies in this vein are not uncommon, for example [120; 97; 152; 161]. These studies could provide a solid basis to a representational framework, however a systematic survey of published science would also be appropriate.

Such a study would result in two products: a representational language for science and its underlying logic, and upper ontologies expressed in this language. In Section 3.2.3 we saw that current mathematical representation languages and their extensions are good initial candidates for a representational language for science. For further implementation, means of interpreting existing represented scientific knowledge in this framework would also be required, perhaps in the form of plugins for knowledge management tools or translators.

Further investigation of the framework for representing evidential reasoning as described in chapter 4 and subsequent development of a prototype system would form a part of this research direction.

### 6.2.2 Representing Science

In Sections 2.2.1 and 2.2.3 we saw that representational languages and knowledge bases of scientific knowledge are currently undergoing rapid expansion. As noted in Section 3.3.1 continued work in this area is necessary to “kick start” semantic publishing.

In Section 3.2.1 we saw that these attempts have been ad-hock and application specific, responding to particular needs for interoperability and knowledge management in the fields for which they were developed. For semantic publishing, they may need to be translated and further annotated in order to follow science modelling standards once such standards are developed (as described in Section 6.2.1).

### 6.2.3 Natural Language Processing

In Sections 3.3.1, 3.6 and 3.6.3 I have mentioned that natural language processing (NLP) could play a significant role in enabling semantic scientific processing. NLP research oriented to extracting semantics specific to science could provide important support to the feasibility of semantic science publishing.

## 6.3 Knowledge Management

If semantic publishing becomes widely used, tools and techniques for managing and organising that knowledge are needed. We have seen that many existing technologies provide powerful techniques for knowledge management, however work is needed to adapt these to a semantically published scientific knowledge context.

### 6.3.1 Fundamentals

In Section 3.2.4 we talked about two possible structures for semantically published knowledge—a web structure and summary knowledge bases. Management of knowledge access (such as search techniques and alerts) and storage structures presents many challenges and opportunities.

Hai Zhuge's book *"The Knowledge Grid"* [160] outlines some current understandings and research directions in this area.

### 6.3.2 Semantic Citation Flavours

In our discussion of semantic citation in Section 3.2.2, we saw that there are many types of citation. Modelling these could involve choices that are not immediately obvious, and would need careful consideration. Such a study would form part of a more fundamental study of modelling scientific knowledge proposed in Section 6.2.1.

The discussion in Section 3.2.2 also touched on the need to adapt current technologies for analysing citation graphs to utilise semantic citations. Initial adaptations could be trivial (for example, simply ignore citations other than 'depends on'), however richer analysis could be achieved through utilising the full semantics of citations.

### 6.3.3 Propagation of Refutation

In Section 3.4.2 we discussed the possibility that refuted results could be propagated through a web of published knowledge. Mechanisms for achieving this would need to be integrated into knowledge management tools. Such mechanisms would operate similarly to mechanisms for calculating and storing measures of trust and verity from annotations and other measures applied to individual published elements (Section 3.4.6 discusses trust and verity management).

### 6.3.4 Folksonomies of Science

In Section 3.4.6 we considered an alternative mechanism to peer and editorial review for achieving measures of trust and verity through community or public discussion (and dispute!) of published claims and through community ranking (folksonomy). As noted above, research into social impacts of such measures is appropriate. There is also much work to be done to adapt current folksonomy tools to the semantic science publishing context.

In Section 2.6 we saw that there are open source social networking applications that could be adapted to the needs of science [92].

### 6.3.5 Knowledge Exchange and Trust

In Section 3.3.2 we noted that new standards for knowledge exchange and trust would need to be devised. The standards under development in the Grid community would likely be good starting places for the development of such standards.

### 6.3.6 Mining Knowledge Bases

Processes and techniques for automatically generating knowledge bases from webs of published knowledge based on verity, trust and significance measures are needed. This is not a trivial task, as differing schools of thought would need to be identified and managed. In Chapter 5 a possible technique for managing schools of thought was discussed.

### 6.3.7 Adaptation of KBs for Different Reasoning Tools

In Section 3.3.3 we noted that we may want our represented knowledge to be used in different systems with differing levels of expressiveness. For example, we may want to use knowledge in a highly expressive science representation for as an organisational principle in a description logic driven semantic grid. An initial approach is for the more expressive features to simply be invisible to the simpler systems, as described above. Can we do better? Perhaps there are approximations or simplifications of the expressive features that could be used in the simpler system? Such questions form interesting directions for future research.

## 6.4 Knowledge Interfaces

In Section 3.2.5 we talked about the need for flexible semantic publishing tools adapted to the needs of different scientific disciplines and individual scientists. In Section 3.6.1 we noted that semantic publishing systems should be designed to be easily operable, ideally with some degree of backward compatibility so that known publishing skills can still be applied. Semantic publishing tools are needed that are intuitive for scientists to use, and which make it easy to locate and insert the appropriate citations.

Designing and implementing intelligent interfaces for publishing and accessing semantically represented knowledge is a broad and challenging area of research.

### 6.4.1 Layers of Information Wiki

In Section 3.4.4 we discussed the utility of different interconnected layers of information and presentation for semantically published scientific knowledge. An interesting research topic would be the development of a wiki system that implements this idea.

Integrating grid authentication services connected to measures of trust and expert authority could be a useful feature of such a system.

#### 6.4.2 Collaborative Islands

To allow for individual scientists or collaborative groups to independently develop new concepts and theories, there would be a need for managing access and exposure of subsets of science knowledge [157]. A group of scientists could, for example, first develop an idea in their private collaboration space, or *collaborative island*, then expose it for comment (much like pre-prints are used today).

Grid services for authentication and rights management (see Section 2.3.1) would also make a logical starting point for implementing collaborative islands. These services would need to be integrated into knowledge management tools for semantically published material.

### 6.5 Linking to Math and other services

In order to take advantage of the benefits that automation could provide to verification and other activities around semantic publishing, knowledge management and authoring tools would need to use mathematical, reasoning and other services. These could be incorporated into the tools, however the power of grid and semantic web technologies for distributed services could add significant improvements in performance and quality.

Such technologies are well into development, and some are quite mature. Utilising them, however, requires ontologies of function and implementations of their application programming interfaces (API's). Development of open standards for the services required by a scientific semantic publishing paradigm is an important research direction.

---

## Conclusion

---

The dissemination of new findings is an important and integral part of the scientific process. In this thesis I present a new vision for science publishing—*semantic publishing* that I argue has the potential to greatly improve the efficacy of dissemination and quality control of published science.

With the advent of the World Wide Web and new search technologies, scientific publishing has, and is continuing to experience what some would call a revolution. I have argued that what we see today is only the beginning of what could be significant change in the way we publish and manage our scientific knowledge.

After a broad survey of current practices and trends in scientific publishing, eScience and Web technologies, I conclude that there are significant opportunities to apply knowledge representation, knowledge management and web technologies to science publishing.

Semantic publishing is a new publishing paradigm in which the conceptual content of scientific work is published directly in machine readable form. This opens up many possibilities for automated management and analysis of published science.

I have introduced two key concepts as a structural basis for a body of semantically published science. *Semantic citation* involves citations between individual concepts which indicate the nature of the relationship between the concepts. *Theory provenance* records the conceptual history of published ideas. I have argued that these concepts, combined with existing Web and knowledge management technologies could afford far greater access to and more accurate assessment of our knowledge than the current publishing system and outlined several practical possibilities for services and systems built around semantically published scientific knowledge.

Evidential reasoning, drawing tenuous conclusions about the world from our observations and experimentation, is fundamental to scientific research. As an initial step toward the semantic publishing vision, I have developed a conceptual framework for representing evidential arguments and made an initial assessment of its utility through examples from the application of statistical tests, climate modelling and plasma physics.

Finally, I have presented a roadmap for future research that would be required in order to realise the semantic publishing vision.



---

# WebScope: A Data Grid for Fusion Research

---

WebScope is a data grid connecting distributed MdsPlus data bases of experimental results from high temperature fusion, developed at the ANU [51; 157].

I had intended to implement a description logic representation system for published fusion research which would link experimental data served by WebScope to a semantic representation of the researches conceptual content. The description logic reasoning and representation system Racer [64] was to be used to formally represent the semantic content of research papers and provide reasoning services such as consistency checking and semantic searches.

It became evident late in the year that Racer's representation language was not equal to the task of usefully representing research findings and arguments in fusion research because of its inability to properly represent complex mathematics (see Section 2.2.2). This discovery and the fact that I had conducted a substantial literature review led to a change in focus for this thesis to a more theoretical presentation.

We did however succeed in porting WebScope to a linux system and I implemented a simple API in WebScope for a knowledge representation system.



---

# Bibliography

---

- [1] ALEXA. Web traffic statistics from alexa, the web information company. data drawn from alexa toolbar users and other sources. [http://www.alexa.com/data/details/traffic\\_details/en.wikipedia.org/wiki/Main\\_Page](http://www.alexa.com/data/details/traffic_details/en.wikipedia.org/wiki/Main_Page). (p.17)
- [2] ANDREWS, P., AND BROWN, C. TPS: A hybrid automatic-interactive system for developing proofs. *Journal of Applied Logic* 4, 4 (2006), 367–395. (p.38)
- [3] ANDRONICO, G., BARBERA, R., AND FALZONE, A. Grid portal based data management for lattice qcd. In *Enabling Technologies: Infrastructure for Collaborative Enterprises, 2004. WET ICE 2004. 13th IEEE International Workshops on* (June 2004), IEEE, pp. 347 – 351. (p.15)
- [4] ANGIONI, C., PEETERS, A., GARBET, X., MANINI, A., RYTER, F., AND TEAM, A. U. Density response to central electron heating: theoretical investigations and experimental observations in asdex upgrade. *Nuclear Fusion* 44, 8 (2004), 827–845. (p.52)
- [5] ARXIV. Eprint archive for physics, mathematics, computer science, quantitative biology and statistics. <http://arxiv.org/>. (pp.15,36)
- [6] BAADER, F., NARDI, D., BRACHMAN, R. J., NUTT, W., DONINI, F. M., SATTLER, U., CALVANESE, D., MOLITOR, R., CALVANESE, D., KUSTERS, G. D. G. R., WOLTER, F., MCGUINNESS, D. L., PATEL-SCHNEIDER, P. F., MOLLER, R., HAARSLEV, V., HORROCKS, I., BORGIDA, A., BRACHMAN, R. J., WELTY, C. A., MCGUINNESS, D. L., FRANCONI, A. R. E., LENZERINI, M., AND ROSATI, R. *The description logic handbook: theory, implementation, and applications*. Cambridge University Press New York, 2003. (pp.6,7,8)
- [7] BACHRACH, S. Who should 'own' scientific papers? *Science* 281 (1998), 1459–60. (p.15)
- [8] BAO, J., HU, Z., CARAGEA, D., REECY, J., AND HONAVAR, V. A tool for collaborative construction of large biological ontologies. *Database and Expert Systems Applications, 2006. DEXA '06. 17th International Conference on* (Sept. 2006), 191–195. (p.10)
- [9] BARTOLO, L. M., COLE, T. W., GIERSCH, S., AND WRIGHT, M. NSF/NSDL Workshop on Scientific Markup Languages. *D-Lib Magazine* 1, 11 (2005). (p.9)
- [10] BAUMGARTNER, P. Private communication with Peter Baumgartner., October 2008. (p.10)

- 
- [11] BAUMGARTNER, P., FURBACH, U., GROSS-HARDT, M., AND SINNER, A. Living Book–Deduction, Slicing, and Interaction. *Journal of Automated Reasoning* 32, 3 (2004), 259–286. (pp. 17, 28, 35, 41)
- [12] BENEDICT, J., MCGUINNESS, D., AND FOX, P. A semantic web-based methodology for building conceptual models of scientific information. In *AGU 2007 Fall Meeting. San Francisco, California, December 10-14, 2007* (2007). (pp. 15, 61)
- [13] BERNERS-LEE, T., HENDLER, J., LASSILA, O., ET AL. The Semantic Web. *Scientific American* 284, 5 (2001), 28–37. (p. 8)
- [14] BERRIMAN, B., KIRKPATRICK, D., HANISCH, R., SZALAY, A., AND WILLIAMS, R. Large Telescopes and Virtual Observatory: Visions for the Future. In *25th meeting of the IAU, Joint Discussion* (2003), vol. 8, p. 17. (pp. 1, 15)
- [15] BILLINGSLEY, W., AND ROBINSON, P. An Interface for Student Proof Exercises Using MathsTiles and Isabelle/HOL in an Intelligent Book. (pp. 17, 41)
- [16] BOSE, R., AND FREW, J. Lineage retrieval for scientific data processing: a survey. *ACM Comput. Surv.* 37, 1 (2005), 1–28. (p. 14)
- [17] BRODIE, M., MYLOPOULOS, J., AND SCHMIDT, J. *On Conceptual Modelling: Perspectives from Artificial Intelligence, Databases, and Programming Languages*. Springer Verlag, 1984. (p. 6)
- [18] BÜCHNER, O., ERNST, M., JANSEN, K., LIPPERT, T., MELKUMYAN, D., ORTH, B., PLEITER, D., STÜBEN, H., WEGNER, P., AND WOLLNY, S. Datagrids for lattice qcd. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 559, 1 (Apr. 2006), 57–61. (p. 13)
- [19] CAMON, E., MAGRANE, M., BARRELL, D., LEE, V., DIMMER, E., MASLEN, J., BINNS, D., HARTE, N., LOPEZ, R., AND APWEILER, R. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucl. Acids Res.* 32, suppl 1 (2004), D262–266. (p. 28)
- [20] CARR, L., HALL, W., BECHHOFFER, S., AND GOBLE, C. Conceptual linking: ontology-based open hypermedia. In *WWW '01: Proceedings of the 10th international conference on World Wide Web* (New York, NY, USA, 2001), ACM, pp. 334–342. (p. 25)
- [21] CASATI, F., GIUNCHIGLIA, F., AND MARCHESE, M. Publish and perish: why the current publication and review model is killing research and wasting your money. *Ubiquity* 8, 3 (2007), 1–1. (p. 15)
- [22] CHENG, J., GRUNINGER, M., SRIRAM, R., AND LAW, K. Process Specification Language for project scheduling information exchange. *INTERNATIONAL JOURNAL OF IT IN ARCHITECTURE ENGINEERING AND CONSTRUCTION* 1 (2003), 307–328. (p. 10)
- [23] CHOI, Y., JEONG, K., KIM, D., LEE, J., LIM, S. B., JUNG, S., HEO, D., HWANG, S., AND HWAN BYEON, O. Glyco-mgrid: A collaborative molecular simulation

- 
- grid for e-glycomics. *e-Science and Grid Computing, IEEE International Conference on* (Dec. 2007), 213–220. (p. 15)
- [24] CITESEER. <http://citeseer.ist.psu.edu/> or <http://citeseersx.ist.psu.edu/>. (pp. 1, 15)
- [25] CML. Chemistry markup language. <http://cml.sourceforge.net/>. (p. 9)
- [26] COLTON, S. Computational discovery in pure mathematics. In *Computational Discovery of Scientific Knowledge*, S. Džeroski and L. Todorovski, Eds. Springer, 2007, pp. 175–201. (p. 38)
- [27] DATAGRID. The (EU) DataGrid Project. <http://eu-datagrid.web.cern.ch/eu-datagrid/>. (p. 12)
- [28] DE LA ROSA, J., AND SZYMANSKI, B. Selecting scientific papers for publication via citation auctions. *Intelligent Systems, IEEE 22*, 6 (Nov.-Dec. 2007), 16–20. (p. 36)
- [29] DE WAARD, A. Science publishing and the semantic web, or: Why are you reading this on paper. In *European Conference on the Semantic Web* (2005). (pp. 15, 21)
- [30] DE WAARD, A., AND KIRCZ, J. Modeling scientific research articles shifting perspectives and persistent issues. In *Proceedings ELPUB2008 Conference on Electronic Publishing Toronto, Canada* (June 2008). (p. 21)
- [31] DEELMAN, E., BLYTHE, J., GIL, Y., KESSELMAN, C., MEHTA, G., PATIL, S., SU, M.-H., VAHI, K., AND LIVNY, M. Pegasus: Mapping scientific workflows onto the grid. In *Grid Computing* (2004), no. 3165/2004 in Lecture Notes in Computer Science, Springer, pp. 11–20. (p. 13)
- [32] DI DONATO, F. Designing a semantic web path to e-science. In *Proceedings of SWAP 2005, 2nd Italian Semantic Web Workshop* (December 14–16 2005), CEUR Workshop Proceedings. ISSN 1613-0073. (pp. 15, 21)
- [33] DITA. Introduction to the darwin information typing architecture. <http://www-128.ibm.com/developerworks/xml/library/x-dita1/>. (p. 16)
- [34] DITA EDITORS. Xml editors that support dita: XMetal Author [www.xmetal.com](http://www.xmetal.com/); XML Pro [www.vervet.com](http://www.vervet.com/); XML Spy/Authentic [www.altova.com](http://www.altova.com/); FrameMaker - [www.adobe.com](http://www.adobe.com/); many others... (p. 16)
- [35] DOAJ. Directory of open access journals. <http://www.doaj.org/>. (p. 15)
- [36] DOI. The digital object identifier system. <http://www.doi.org/>. (p. 25)
- [37] EDUTECH. Edutech wiki - page on darwin information typing architecture (dita) and it's uses in education. <http://edutechwiki.unige.ch/en/DITA>. (p. 41)
- [38] EMBL-EBI. Biological ontology databases. European Bioinformatics Institute, an Outstation of the European Molecular Biology Laboratory. <http://www.ebi.ac.uk/Databases/ontology.html>. (pp. 10, 28)
- [39] EMMOTT, S., AND RISON, S. Towards 2020 Science. Tech. rep., Microsoft Research, Cambridge, UK, 2006. (p. 2)

- 
- [40] ESG. Earth system grid. <http://www.earthsystemgrid.org/>. See also the ESG Services page: <http://www.earthsystemgrid.org/about/explainPage.do>. (pp.12, 13)
- [41] ESML. Earth science markup language. <http://esml.itsc.uah.edu/>. (p.9)
- [42] FISCHER, L. *2008 BPM and Workflow Handbook: methods, concepts, case studies and standards in business process management and workflow*. Future Strategies Inc., Lighthouse Point, Florida, 2008. (p.13)
- [43] FOSTER, I. What is the Grid? A Three Point Checklist. *Grid Today* 1, 6 (2002), 22–25. (p.11)
- [44] FOSTER, I., AND KESSELMAN, C., Eds. *The Grid 2: Blueprint for a New Computing Infrastructure*. The Morgan Kaufmann Series in Computer Architecture and Design. Morgan Kaufmann, November 2003. (pp.1, 11, 12, 13, 14, 15)
- [45] FOX, G., GUHA, R., MCMULLEN, D., MUSTACOGLU, A., PIERCE, M., TOPCU, A., AND WILD, D. Web 2.0 for Grids and e-Science. In *INGRID - Instrumenting the Grid 2nd International Workshop on Distributed Cooperative Laboratories S.Margherita Ligure Portofino, ITALY, April 18 2007* (2007), INGRID. (p.19)
- [46] FOX, P., MCGUINNESS, D., MIDDLETON, D., CINQUINI, L., DARNELL, J., GARCIA, J., WEST, P., BENEDICT, J., AND SOLOMON, S. Semantically-Enabled Large-Scale Science Data Repositories. *LECTURE NOTES IN COMPUTER SCIENCE* 4273 (2006), 792. (p.15)
- [47] FOX, P., MCGUINNESS, D., RASKIN, R., AND SINHA, A. Semantically-Enabled Scientific Data Integration. *US Geological Survey Scientific Investigations Report 5201* (2006). See also <http://sesdi.hao.ucar.edu/>. (p.10)
- [48] FREER, C. vdash.org: a formalized math wiki. <http://www.vdash.org/e-club.pdf>. (pp.29, 41)
- [49] FUSIONGRID. National fusiongrid (of the usa). <http://www.fusiongrid.org/>. (p.13)
- [50] GARAVELLI, A. C., GORGOGLIONE, M., AND SCOZZI, B. Managing knowledge transfer by knowledge technologies. *Technovation* 22, 5 (May 2002), 269–279. (p.28)
- [51] GARDENER, H., KARIA, R., AND MANDUCHI, G. A web-based, dynamic metadata interface to mdsplus. *Fusion Engineering and Design* 83 (April 2008), 448–452. (pp.13, 41, 69)
- [52] GEIST, A., SCHWIDDER, J., JUNG, D., AND NACHTIGAL, N. Ornl—electronic notebook project. <http://www.csm.ornl.gov/geist/java/applets/enote/>. (p.18)
- [53] GILES, J. Internet encyclopaedias go head to head. *Nature* 438, 7070 (Dec. 2005), 900–901. (p.17)
- [54] GILES, J. Open-access journal will publish first, judge later. In *Nature* [118], pp. 9–9. (p.15)

- 
- [55] GLICKMAN, T. S., AND THE AMERICAN METEOROLOGICAL SOCIETY. *Glossary of Climate*. American Meteorological Society, 1999. (p. 49)
- [56] GLOBUS. The globus alliance is a community of organizations and individuals developing fundamental technologies behind the "grid". <http://www.globus.org/>. (p. 15)
- [57] GOBLE, C. Putting semantics into e-science and grids. *e-Science and Grid Computing, 2005. First International Conference on* (Dec. 2005), 1 pp.–. (p. 15)
- [58] GOBLE, C. The future of research. Presentation to British Library Board, York, September 2008. (pp. 3, 5, 41)
- [59] GOLBECK, J. Computer Science: Weaving a Web of Trust. *Science* 321, 5896 (2008), 1640–1641. (pp. 19, 36)
- [60] GOOGLESCHOLAR. <http://scholar.google.com/>. (pp. 1, 15)
- [61] GRAY, J. Trident: Scientific workflow workbench for oceanography. <http://www.microsoft.com/mscorp/tc/trident.msp>. (p. 1)
- [62] GRUBER, T. A translation approach to portable ontology specifications. *KNOWLEDGE ACQUISITION* 5 (1993), 199–199. (p. 7)
- [63] GUARINO, N. *Introduction - Formal ontology in information systems*. IOS Press, 1998, ch. Introduction, pp. 3–18. (p. 7)
- [64] HAARSLEV, V., AND MÜLLER, R. Racer system description. In *Automated Reasoning, Lecture Notes in Computer Science*. Springer, 2001, pp. 701–705. (p. 69)
- [65] HACKING, I. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge University Press, 1983. (p. 2)
- [66] HEPP, M. Possible ontologies: How reality constrains the development of relevant ontologies. *Internet Computing, IEEE* 11, 1 (Jan.-Feb. 2007), 90–96. (p. 8)
- [67] HEYERDAHL, T. *The Kon-Tiki Expedition: By Raft Across the South Seas*. Allen & Unwin, 1950. (p. 55)
- [68] HILF, E. R., KOHLHASE, M., AND STAMERJOHANN, H. Capturing the content of physics: Systems, observables, and experiments. In *Lecture Notes in Computer Science, Mathematical Knowledge Management* (2006), vol. Volume 4108/2006 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 165–178. (pp. 2, 9, 27)
- [69] HU, X., LIN, T., SONG, I., LIN, X., YOO, I., LECHNER, M., AND SONG, M. Ontology-Based Scalable and Portable Information Extraction System to Extract Biological Knowledge from Huge Collection of Biomedical Web Documents. In *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence* (2004), IEEE Computer Society Washington, DC, USA, pp. 77–83. (pp. 10, 28, 32, 40)
- [70] HUANG, G. T. China special: Beyond the great firewall. *New Scientist*, 2629 (2007). (p. 39)

- 
- [71] ILDG. International lattice data grid. <http://ildg.sasr.edu.au/Plone>. (p. 13)
- [72] INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE. Procedures for the preparation, review, acceptance, adoption, approval and publication of ipcc reports. In *Principles Governing IPCC Work*. 2003. (p. 16)
- [73] INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE. *Climate Change 2007—The Physical Science Basis: Working Group I Contribution to the Fourth Assessment Report of the IPCC*. Cambridge University Press New York, 2007. (pp. 22, 33, 50)
- [74] ISABELLE. Isabelle is a generic proof assistant. <http://isabelle.in.tum.de/>. See also [IsarMathLib]. (p. 38)
- [75] ISARMATHLIB. Library of formalized mathematics for isabelle/isar (zf logic). <http://savannah.nongnu.org/projects/isarmathlib>. See also [IsarMathLib]. (p. 27)
- [76] ISI. Isi web of knowledge. <http://apps.isiknowledge.com/>. (p. 15)
- [77] ITURRIOZ, J., DIAZ, O., AND ANZUOLA, S. Toward the semantic desktop: The semouse approach. *Intelligent Systems, IEEE* 23, 1 (Jan.-Feb. 2008), 24–31. (p. 17)
- [78] JEFFERSON, T., ALDERSON, P., WAGER, E., AND DAVIDOFF, F. Effects of Editorial Peer Review: A Systematic Review. *JAMA* 287, 21 (2002), 2784–2786. (p. 15)
- [79] JEP. The journal of electronic publishing. <http://www.journalofelectronicpublishing.org/>. Sponsored by Elsevier, Lexis-Nexis, O'Reilly Media, NewsBank Readex, Aptara, Wiley-Blackwell. (p. 15)
- [80] KALNAY, E., Ed. *Atmospheric Modeling, Data Assimilation, and Predictability*. Cambridge University Press New York, 2002. (p. 50)
- [81] KALNAY, E., KANAMITSU, M., COLLINS, W., DEAVEN, D., GANDIN, L., IREDELL, M., SAHA, S., WHITE, G., WOOLEN, J., ZHU, Y., CHELLIAH, M., EBISUZAKI, W., HIGGINS, W., JANOWIAK, J., MO, K. C., ROPELEWSKI, C., WANG, J., LEETMA, A., REYNOLDS, R., JENNE, R., AND JOSEPH, D. The NCEP/NCAR 40-Year Reanalysis Project. *Bulletin of the American Meteorological Society* 77 (1996), 437–471. (p. 50)
- [82] KALYANPUR, A., PARSIA, B., SIRIN, E., GRAU, B. C., AND HENDLER, J. Swoop: A web ontology editing browser. *Web Semantics: Science, Services and Agents on the World Wide Web* 4, 2 (June 2006), 144–153. (p. 16)
- [83] KANAMITSU, M., EBISUZAKI, W., WOOLEN, W. J., YANG, S.-K., HNILO, J. J., FIORINO, M., AND POTTER, G. L. NCEP-DOE AMIP-II Reanalysis (R-2). *Bulletin of the American Meteorological Society* 83 (2002), 1631–1643. (p. 50)
- [84] KITTUR, A., SUH, B., PENDLETON, B. A., AND CHI, E. H. He says, she says: conflict and coordination in wikipedia. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems* (New York, NY, USA, 2007), ACM, pp. 453–462. (p. 36)
- [85] KOHLHASE, M. Omdoc: an infrastructure for openmath content dictionary information. *SIGSAM Bull.* 34, 2 (2000), 43–48. (p. 27)

- 
- [86] LANGE, C. Swim a semantic wiki for mathematical knowledge management. In *The Semantic Web: Research and Applications* (2008), vol. Volume 5021/2008, Springer Berlin / Heidelberg, pp. 832–837. (pp.29, 41)
- [87] LAUGHLIN, R. B. *A Different Universe: Reinventing Physics from the Bottom Down*. Basic Books, 2005. (p.59)
- [88] LEISCH, F. *Sweave: Dynamic generation of statistical reports using literate data analysis*. Springer, 2002, pp. 575–580. (p.14)
- [89] LEVESQUE, H., AND BRACHMAN, R. Expressiveness and tractability in knowledge representation and reasoning 1. *Computational Intelligence* 3, 1 (1987), 78–93. (p.7)
- [90] LIN, H.-N., TSENG, S.-S., WENG, J.-F., LIN, H.-Y., AND SU, J.-M. An iterative, collaborative ontology construction scheme. *Innovative Computing, Information and Control, 2007. ICICIC '07. Second International Conference on* (Sept. 2007), 150–150. (p.10)
- [91] LUDASCHER, B., ALTINTAS, I., BERKLEY, C., HIGGINS, D., JAEGER, E., JONES, M., LEE, E., TAO, J., AND ZHAO, Y. Scientific workflow management and the Kepler system. *Concurrency and Computation: Practice & Experience* 18, 10 (2006), 1039–1065. (p.13)
- [92] MASHABLE. Top 10 open source social platforms. <http://mashable.com/2007/07/25/open-source-social-platforms/>. (pp.19, 65)
- [93] MATHML. W3c mathematical markup language. <http://www.w3.org/Math/>. (p.9)
- [94] MAYNARD, C., AND PLEITER, D. Qcdml: First milestones for building an international lattice data grid. *Nuclear Physics B - Proceedings Supplements* 140 (Mar. 2005), 213–221. (<http://www.lqcd.org/ildg>). (p.9)
- [95] MCCUNE, W. Solution of the robbins problem. *Journal of Automated Reasoning* 19, 3 (Dec. 1997), 263–276. (p.10)
- [96] MCGUINNESS, D., FOX, P., CINQUINI, L., WEST, P., GARCIA, J., BENEDICT, J., AND MIDDLETON, D. The Virtual Solar-Terrestrial Observatory: A Deployed Semantic Web Application Case Study for Scientific Research. In *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE* (2007), vol. 22, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, p. 1730. (pp.1, 15)
- [97] MEDAWAR, P. *Induction and Intuition in Scientific Thought*. American Philosophical Society, 1969. (pp.2, 63)
- [98] MGRID. Michigan grid research and infrastructure development. <http://www.mgrid.umich.edu/index.html>. (p.15)
- [99] MILES, S., DEELMAN, E., GROTH, P., VAHI, K., MEHTA, G., AND MOREAU, L. Connecting scientific data to scientific experiments with provenance. *e-*

- 
- Science and Grid Computing, IEEE International Conference on* (Dec. 2007), 179–186. (p. 14)
- [100] MIZAR. The mizar project for formalized representation of mathematics. <http://www.mizar.org/>. (pp. 27, 38)
- [101] MOORE, G. Excerpts from a Conversation with Gordon Moore: Moore’s Law, 2005. Video transcript from Intel. (p. 1)
- [102] MOREAU, L., LUDÄSCHER, B., ALTINTAS, I., BARGA, R. S., BOWERS, S., CALLAHAN, S., JR., G. C., CLIFFORD, B., COHEN, S., COHEN-BOULAKIA, S., DAVIDSON, S., DEELMAN, E., DIGIAMPIETRI, L., FOSTER, I., FREIRE, J., FREW, J., FUTRELLE, J., GIBSON, T., GIL, Y., GOBLE, C., GOLBECK, J., GROTH, P., HOLLAND, D. A., JIANG, S., KIM, J., KOOP, D., KRENEK, A., MCPHILLIPS, T., MEHTA, G., MILES, S., METZGER, D., MUNROE, S., MYERS, J., PLALE, B., PODHORSZKI, N., RATNAKAR, V., SANTOS, E., SCHEIDEGGER, C., SCHUCHARDT, K., SELTZER, M., SIMMHAN, Y. L., SILVA, C., SLAUGHTER, P., STEPHAN, E., STEVENS, R., TURI, D., VO, H., WILDE, M., ZHAO, J., AND ZHAO, Y. Special Issue: The First Provenance Challenge. *Concurrency and Computation: Practice and Experience* 20, 5 (2008), 409–418. (p. 14)
- [103] MYERS, J., MENDOZA, E., AND HOOPES, B. A Collaborative Electronic Notebook. In *Proceedings of the IASTED International Conference on Internet and Multimedia Systems and Applications (IMSA 2001), August (2001)*, ACTA Press, pp. 13–16. (p. 14)
- [104] MYERS, J. D., CHAPPELL, A., ELDER, M., GEIST, A., AND SCHWIDDER, J. Re-integrating the research record. *Computing in Science and Engineering* 5, 3 (2003), 44–50. (p. 14)
- [105] MYEXPERIMENT. Freeware application to “find, use, and share scientific workflows and other files, and to build communities”. developed by a joint team from the universities of southampton and manchester. <http://www.myexperiment.org/>. (p. 19)
- [106] NATURE. Nature’s peer review debate. <http://www.nature.com/nature/peerreview/debate/index.html>. (pp. 15, 16)
- [107] NEUROCOMMONS. The neurocommons. [http://neurocommons.org/page/Main\\_Page](http://neurocommons.org/page/Main_Page). (p. 40)
- [108] NORMAN, D., AND COLLYER, B. *The design of everyday things*. Basic Books New York, 2002. (p. 61)
- [109] NOWOTNY, H., SCOTT, P., AND GIBBONS, M. *Re-Thinking Science: Knowledge and the Public in an Age of Uncertainty*. Polity Press, 2001. (p. 22)
- [110] OGF. Open grid forum. <http://www.ogf.org/>. An open community committed to driving the rapid evolution and adoption of applied distributed computing. (p. 12)
- [111] OMDOC. A markup format and data model for open mathematical documents. <http://www.omdoc.org/>. (p. 9)

- 
- [112] OPEN JOURNAL SYSTEMS. Open source journal system used by over 1400 titles using ojs (as of march 2008) in ten languages. <http://pkp.sfu.ca/?q=ojs>. (p.15)
- [113] OREILLY, T. What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *Communications & Strategies* (First Quarter 2007), 17. (p.17)
- [114] ORESKES, N. BEYOND THE IVORY TOWER: The Scientific Consensus on Climate Change. *Science* 306, 5702 (2004), 1686–. (p.32)
- [115] OSG. Open science grid. <http://www.opensciencegrid.org/>. (p.13)
- [116] OTMI. Open text mining interface. [http://opentextmining.org/wiki/Main\\_Page](http://opentextmining.org/wiki/Main_Page). (p.40)
- [117] PIERCE, M., FOX, G., CHOI, J., GUO, Z., GAO, X., AND MA, Y. Using Web 2.0 for Scientific Applications and Scientific Communities. *Concurrency and Computation: Practice and Experience Special Issue for 3rd International Conference on Semantics, Knowledge and Grid SKG2007* (October 28-30 2007). Xian China. (p.19)
- [118] PLOS. Plos one - public library of science open access journal. <http://www.plosone.org/home.action>. (pp.15,74)
- [119] PLOS-IMPACT. 2007 impact factors for plos journals. <http://www.plos.org/cms/node/366>. (p.15)
- [120] POPPER, K. *The Logic of Scientific Discovery*, 6th impression revised. ed. London : Hutchinson, 1972. (pp.2, 48, 63)
- [121] PROTÉGÉ. Protege ontology editor. <http://protege.stanford.edu/>. (p.35)
- [122] QUIRK, J. Computational science “same old silence, same old mistakes” “something more is needed ...”. In *Adaptive Mesh Refinement - Theory and Applications* (2005), vol. 41 of *Lecture Notes in Computational Science and Engineering*, Springer, pp. 3–28. (p.14)
- [123] RAMSEY, N. Literate programming simplified. *Software, IEEE* 11, 5 (Sep 1994), 97–105. (p.14)
- [124] REAL CLIMATE. Real climate blog—climate science from climate scientists. <http://www.realclimate.org/>. (p.18)
- [125] RODRIGUEZ, M. A., BOLLEN, J., AND VAN DE SOMPEL, H. The convergence of digital libraries and the peer-review process. *Journal of Information Science* 32, 2 (2006), 149–159. (p.35)
- [126] ROURE, D., JENNINGS, N., AND SHADBOLT, N. Research agenda for the semantic grid: a future escience infrastructure. *National e-Science Centre, Edinburgh, UK* 9 (2001). (p.14)
- [127] ROURE, D. D., JENNINGS, N. R., AND SHADBOLT, N. R. The semantic grid: A future e-science infrastructure. In *Grid Computing*, T. H. Fran Berman, Geoffrey Fox, Ed. Wiley, 2003, pp. 437–470. (pp.1, 15)

- [128] SEMANTIC DESKTOP LINKS. Applications and other links related to semantic desktops. [www.semanticdesktop.org](http://www.semanticdesktop.org); [www.gnowsis.org](http://www.gnowsis.org); [www.deepamehta.de](http://www.deepamehta.de); [www.openiris.org](http://www.openiris.org). (p.17)
- [129] SHADBOLT, N., HALL, W., AND BERNERS-LEE, T. The semantic web revisited. *Intelligent Systems, IEEE* 21, 3 (Jan.-Feb. 2006), 96–101. (p.8)
- [130] SHAPARENKO, B., AND JOACHIMS, T. Information genealogy: uncovering the flow of ideas in non-hyperlinked document databases. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (New York, NY, USA, 2007), ACM, pp. 619–628. (p.40)
- [131] SHUM, S., DE ROURE, D., EISENSTADT, M., SHADBOLT, N., AND TATE, A. CoAKTinG: Collaborative Advanced Knowledge Technologies in the Grid. In *2nd Workshop Advanced Collaborative Environments*. See also <http://www.aktors.org/coacting/>. (pp.1, 14)
- [132] SIDDIQUI, M., VILLAZON, A., AND FAHRINGER, T. Semantic-based on-demand synthesis of grid activities for automatic workflow generation. *e-Science and Grid Computing, IEEE International Conference on* (Dec. 2007), 43–50. (p.15)
- [133] SIMMHAN, Y. L., PLALE, B., AND GANNON, D. A survey of data provenance in e-science. *SIGMOD Rec.* 34, 3 (2005), 31–36. (p.14)
- [134] SMITH, B., AND WELTY, C. Fois introduction: Ontology - towards a new synthesis. In *FOIS '01: Proceedings of the international conference on Formal Ontology in Information Systems* (New York, NY, USA, 2001), ACM, pp. .3–.9. Conference Chair-Barry Smith and Conference Chair-Christopher Welty. (p.7)
- [135] SMITH, J. The deconstructed journal-a new model for academic publishing. *Learned Publishing* 12, 2 (1999), 79–91. (pp.37, 62)
- [136] SOMASUNDARAM, T., BALACHANDAR, R., KANDASAMY, V., BUYYA, R., RAMAN, R., MOHANRAM, N., AND VARUN, S. Semantic-based grid resource discovery and its integration with the grid service broker. *Advanced Computing and Communications, 2006. ADCOM 2006. International Conference on* (Dec. 2006), 84–89. (p.15)
- [137] SOWA, J. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Pacific Grove : Brooks/Cole, 2000. ISBN: 0534949657. (p.6)
- [138] STEIN, L. D. Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. *Nat Rev Genet* 9, 9 (Sept. 2008), 678–688. (pp.16, 40)
- [139] STEVENS, R., GOBLE, C., AND BECHHOFFER, S. Ontology-based knowledge representation for bioinformatics. *Briefings in Bioinformatics* 1, 4 (2000), 398–414. (pp.1, 10)
- [140] STEVENS, R. D., ROBINSON, A. J., AND GOBLE, C. A. myGrid: personalised bioinformatics on the information grid. *Bioinformatics* 19, suppl\_1 (2003), i302–304. (pp.10, 15)

- 
- [141] SUBER, P. Timeline of the open access movement. <http://www.earlham.edu/peters/fos/timeline.htm>. (p.15)
- [142] SZALAY, A., AND GRAY, J. 2020 computing: Science in an exponential world. *Nature* 440, 7083 (Mar. 2006), 413–414. (p.1)
- [143] TEAM, P. M. L. D. Standards. <http://web.mit.edu/mecheng/pml/standards.htm>. (p.9)
- [144] TENENBAUM, M., AND WILBANKS, J. Health commons white paper. see also <http://sciencecommons.org/projects/healthcommons/>, May 2007. (p.40)
- [145] TIEN, J., AND BERG, D. A case for service systems engineering. *Journal of Systems Science and Systems Engineering* 12, 1 (2006), 13–38. (p.1)
- [146] TOMLIN, S. Science in the web age: The expanding electronic universe. *Nature* 438, 7068 (Dec. 2005), 547–547. (p.17)
- [147] TOPAZ. Open source software around collaborative creation, management and sharing of information. <http://www.topazproject.org/trac/wiki>. (p.16)
- [148] UPPALA, S. M., KALLBERG, P. W., SIMMONS, A. J., ANDRAE, U., DA COSTA BECHTOLD, V., FIORINO, M., GIBSON, J. K., HASELER, J., HERNANDES, A., KELLI, G. A., LI, X., ONOGI, K., SARRINEN, S., SOKKA, N., ALLAN, R. P., ANDERSSON, E., ARPE, K., BALMASEDA, M. A., BELJAARS, A. C. M., VAN DE BERG, L., BIDLOT, J., BORMANN, N., CAIRES, S., CHEVALLIER, F., DETHOF, A., DRAGOSAVAC, M., FISHER, M., FUENTES, M., HAGERMANN, S., HOLM, E., HOSKINS, B. J., ISAKSEN, L., JANSSEN, P. A. E. M., JENNE, R., MCNALLY, A. P., MAHOUF, J.-F., MORCETTE, J.-J., RAYNER, N. A., SAUNDERS, R. W., SIMON, P., STERL, A., TRENBERTH, K. E., UNTCH, A., VASILJEVIC, D., VITERBO, P., AND WOOLLEN, J. The ERA-40 Re-analysis. *Quarterly Journal of the Royal Meteorological Society* 131 (2005), 2961–3012. (p.50)
- [149] UREN, V., CIMIANO, P., IRIA, J., HANDSCHUH, S., VARGAS-VERA, M., MOTTA, E., AND CIRAVEGNA, F. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: Science, Services and Agents on the World Wide Web* 4, 1 (Jan. 2006), 14–28. (p.16)
- [150] W3C. Naming and addressing: Uris, urls, ... <http://www.w3.org/Addressing/>. (p.25)
- [151] W3C. Web ontology language. <http://www.w3.org/TR/owl-features/>. (p.7)
- [152] WALLACE, D., AND GRUBER, H. *Creative people at work*. Oxford University Press, 1989. (p.63)
- [153] WALTER, C. Kryder’s law. *Scientific American Magazine* (July 2005). (p.1)
- [154] WIKIPEDIA. Wikipedia, the free encyclopeida (english version). <http://en.wikipedia.org/>. (pp.18, 52)
- [155] WIKIPEDIA STATISTICS. Statisics drawn in october 2008 from. <http://en.wikipedia.org/wiki/Special:Statistics>. (p.17)

- [156] XEXEO, G., DE SOUZA, J., VIVACQUA, A., MIRANDA, B., BRAGA, B., ALMENTERO, B., D' ALMEIDA, J.N., J., AND CASTILHO, R. Peer-to-peer collaborative editing of ontologies. *Computer Supported Cooperative Work in Design, 2004. Proceedings. The 8th International Conference on 2* (May 2004), 186–190 Vol.2. (p.10)
- [157] XIAOBIN, W. Web services architecture for a fusion data grid. Master's thesis, Australian National University, 2008. (pp. 66, 69)
- [158] ZHANG, X., HU, C., ZHAO, Q., AND ZHAO, C. Semantic data integration in materials science based on semantic model. *e-Science and Grid Computing, IEEE International Conference on* (Dec. 2007), 320–327. (p.10)
- [159] ZHAO, Y., HATEGAN, M., CLIFFORD, B., FOSTER, I., VON LASZEWSKI, G., NEFEDOVA, V., RAICU, I., STEF-PRAUN, T., AND WILDE, M. Swift: Fast, reliable, loosely coupled parallel computation. *Services, 2007 IEEE Congress on* (July 2007), 199–206. (p.13)
- [160] ZHUGE, H. *The Knowledge Grid*. World Scientific, 2004. (pp. 28, 38, 64)
- [161] ZIMAN, J. *Real Science: What it Is, and what it Means*. Cambridge University Press, 2000. (p.63)
- [162] ZOLIN, E. Description logic complexity calculator. <http://www.cs.man.ac.uk/~ezolin/dl/>. (p.7)
- [163] ZÚÑIGA, G. L. Ontology: its transformation from philosophy to information systems. In *FOIS '01: Proceedings of the international conference on Formal Ontology in Information Systems* (New York, NY, USA, 2001), ACM, pp. 187–197. (pp. 6, 7)