Scientific workflow design: theoretical and practical issues
Terpstra, F.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 2

Methodology

2.1 Introduction

The research described in this thesis concerns experiments within the context of a Virtual Laboratory also known as e-Science. Scientific methodology is fundamental to conducting experiments. In this chapter we describe how the scientific methodology can be applied in a Virtual Laboratory, and what aspects in particular are affected by the Virtual Laboratory environment. The Virtual Laboratory environment is, for the scientific user, dominated by the workflow environment making up the frontend. This leads to workflow playing a pivotal role in e-Science methodology. We will compare the method of science in general to that of e-Science and discuss if the new assumptions and methods e-Science bring constitute a paradigm shift. We will start by looking at the methodology for science in general and what exactly a paradigm shift is. We continue by looking at what exactly what different people understand e-Science to be and what its defining features are. A short overview is given of the different roles that exist for people that together conduct e-Science in a virtual laboratory. Then we show how the scientific method in e-Science differs from that of science in general and discuss whether this constitutes a paradigm shift. Finally we give a scenario of what the work of a scientist working in a virtual laboratory will look like if all the aims of e-Science have been realized. In this way this chapter sets up a reference against which we can compare whether systems and techniques presented in the rest of this thesis can realize the full potential of e-Science.

2.2 Methodology of Science

To give a better understanding of what aspects of scientific methodology are important in e-Science we will first take a look at scientific methodology in general. This will by no means be an all inclusive overview but it will touch upon the essential problems within the philosophy of science which
are relevant to e-Science.

The foundations for our current scientific methodology were laid in the renaissance when experimental science started. This methodology for experimental science is dominated by the empirical cycle. It is often depicted in the form used here in Figure 2.1 for instance by Adriaans and Zantinge in [26], and based upon the work by scientific philosophers such as Popper, Lakatos and Stegmuller [110, 89, 115]. It is noteworthy though, that there does not seem to be one firm definition of the empirical cycle, upon which everyone in the scientific community agrees. The empirical cycle describes an incremental process for increasing knowledge. In different incarnations it contains the following steps: Define research question, characterization, hypothesis, theory, model, prediction, experiment, observation, analysis, publish results, dissemination, reproduce result, verification. The empirical cycle will be described here in terms of four main elements:

- Theory
- Prediction
- Observation
- Analysis

We start with theory in which all relevant assumptions are defined, these include the research question and methods of observation and measurements as well as existing theories which are relevant for the experiment(s) that are about to be performed. A model would also fall under the theory step. The Prediction step is a logical deduction from the theory defined in the first step, one could also call a prediction a hypothesis. It is defined in such a way that it can be directly compared to observations in the next step. The analysis step performs this comparison, an explanation is sought which explains the differences or similarity between observation and prediction. This can be
done by the scientists themselves or by their peers through publication in the scientific community. Based upon the analysis step multiple actions are possible all of which let the cycle continue for another iteration. The reproducibility can be tested by performing the whole cycle again. This can be done by scientists themselves or by their peers who are reviewing the work based upon a publication. Assumptions in the theory step can be adjusted to find extra evidence either verifying or falsifying the theory. Finally the theory itself can be adjusted based upon the analysis. The meaning of experiment is strongly associated with the actual measurements performed. We take a wider view of what an experiment is and view it as preforming one or more iterations of the empirical cycle.

To illustrate the empirical cycle we will look at an historical example. According to Aristotle heavier objects fell more quickly than lighter objects. The underlying theory being that of the five elements, earth, water, air, fire and aether. The properties of each object are determined by its likeness to four of the elements, earth, water, air and fire, while aether acts as a conduit attracting each object to its likeness. According to this theory a heavy object like a stone had most in common with earth and wanted to move there more quickly than a lighter object. A feather on the contrary had something in common with air and would move to the ground less quickly. Based on this theory Aristotle could make the prediction that a stone would fall more quickly than a feather. Indeed performing this experiment the feather could be observed falling much more slowly than the stone. Other observations could be made as well, the feather does not fall in a straight line. In the analysis this could be explained as the feather showing properties of air, thereby confirming the theory. As we now know Aristotle was wrong. We will return to this example later to show how he was wrong and more importantly illustrate other issues in the philosophy of science.

2.2.1 Philosophical issues

The validity of the scientific method is the core discussion within the philosophy of science. One problem first raised by Hume\cite{78} is the question whether induction, abstracting general knowledge from specific examples is valid. This is very relevant to experimental science as it is this reasoning method that is used to derive knowledge from experimental results. Predictions made, based on a theory that has often been verified, can still not be proven to be the absolute truth. This “problem of induction” is also referred to as Hume’s puzzle, as although induction can not be proven to be valid most of our knowledge is based upon it. Another important problem in the scientific method is the aim to discover objective truths. The question is whether it is possible for humans to perform unbiased objective measurements or to postulate objective hypotheses. Popper argues that all observation is theory laden \cite{110}. Theory guides us to determine which
observations are significant, and in what way they are significant. To put it another way, the scope of our search space for observations is directed by our existing theories. This does not mean we try to find the answers we were expecting and are not open to unexpected, strange, surprising or serendipitous results. In an infinite universe we can only make a finite number of observations, thus we have to restrict our search space based upon our existing theories. According to both Feyerabend and Popper\cite{63, 110} descriptions of observation are necessarily theory laden. We cannot claim that the data we use is purely objective, there is always some theory used to translate our experience, or those of measuring instruments into data.

An often heard term in the philosophy of science is the “paradigm shift”. It was introduced by Thomas Kuhn\cite{88} to counter the idea of gradual and cumulative growth of scientific knowledge, a position taken among others by Popper in his essay “the aim of science” \cite{109}. The basic premise is that instead of gradual evolution there are periodic revolutions in scientific understanding when the scientific community adopts a new theory in favor of an older one. Kuhn argues that to adopt a new theory one also has to change many of the basic assumptions on which a theory is based. Thus it is not a cumulative growth of adding the new theory to existing understanding, but rather a shift of theory and basic assumptions in the area the new theory covers: a paradigm shift. Multiple competing paradigms can exist at the same time. During a shift dominance shifts from one to the other. This, according to Kuhn, is accompanied by fierce arguments of proponents of the different paradigms arguing the correctness of their theories, until results clearly point the community in the direction of the new paradigm.

Now we return to our example: the theory of Aristotle. It was disproved by Galileo. He showed that two balls of different weight, a cannon-ball and a musket ball fell at the same speed despite their weight difference. Galileo came to this experiment because his underlying theory of motion had changed. According to Galileo the distance covered by an accelerating object (for instance through falling) from rest is proportional to the time squared. There is therefore no relation of weight to the speed with which an object falls. Furthermore an object maintains its velocity unless a force acts upon it. The feathers slower fall is due to the force of air resistance acting upon it. This example is a good demonstration of a paradigm shift. The literal meaning of paradigm is example. In this case Galileo’s experiment is an example of a shift in paradigm in the most literal sense. Galileo was not the first to perform this experiment or disprove Aristotle. In fact there is doubt whether he actually ever performed it as is commonly told on the tower of Pisa. Yet this example stuck, because Galileo backed it up with a new theory. The example by which the theory of motion was illustrated shifted to Galileo’s experiment. More interesting though is the underlying shift in theory and assumptions which is most often intended when the term paradigm shift is used.
2.3 Methodology of e-Science

After a brief introduction into scientific methodology in the previous section the focus is now on what differentiates e-Science from science in general and what is important in a methodology for e-Science. To answer this a definition of e-Science is needed first.

2.3.1 Definition of e-Science

There is no single clear definition for the term e-Science. However looking at the collection of definitions presented below a common picture emerges.

"Science increasingly done through distributed global collaborations enabled by the Internet, using very large data collections, tera-scale computing resources and high performance visualization."\(^1\)

"e-Science is about global collaboration in key areas of science and the next generation of infrastructure that will enable it."\(^2\)

"e-Science refers to science that is enabled by the routine use of distributed computing resources by end-user scientists. e-Science is often most effective when the resources are accessed transparently and pervasively, and when the underlying infrastructure is resilient and capable of providing a high quality of service. Thus, grid computing and autonomic computing can serve as a good basis for e-Science. e-Science often involves collaboration between scientists who share resources, and thus act as a virtual organization."\(^3\)

"The next generation of scientific research and experiments will be carried out by communities of researchers from organizations that span national boundaries. These activities will involve geographically distributed and heterogeneous resources such as computational systems, scientific instruments, databases, sensors, software components, networks, and people. Such large-scale and enhanced scientific endeavors, popularly termed as e-Science, are carried out via collaborations on a global scale."\(^4\)

"The e-Science vision is a future research environment based on virtual organizations of people and agents highly dynamic and with large-scale computation, data, and collaboration."\(^5\)

\(^1\) U.K. Government
\(^2\) Dr John Taylor, Director General of Research Councils, 2000 http://www.rcuk.ac.uk/escience/
\(^3\) e-Science Gap Analysis June 2003 UK e-Science Grid
\(^4\) e-Science 2005 conference Australia
\(^5\) e-Science: The Grid and the Semantic Web IEEE Intelligent Systems vol 19 no1, January/February 2004
Important factors in e-Science can be brought together in a few distinct clusters

- massive computing: grid computing, distributed computing resources, large scale computation
- massive data handling: data management, very large data collections, large scale data, distributed databases
- virtual organizations: large scale/global collaboration
- Resource sharing: computational, storage, software, data, measuring equipment, human experts

The development of e-Science environments is being driven by applications such as ATLAS[37] from the field of high energy physics, which place a big demand on all four of the previously mentioned factors. E-Science environments aim for the sharing of as many resources as possible of all types of resources. This allows other scientific applications, which are not the primary driving force behind e-Science development, to benefit from the available shared resources. Within an e-Science environment scientists will have access to resources which are outside of their own expertise. These can be resources within a large interdisciplinary project or resources shared with other projects in a common e-Science environment. Within e-Science there are different roles that perform the various tasks in a functioning e-Science environment. In his thesis[82] Ersin Kaletas identifies four types of users which are briefly explained here:

- **Scientist**, uses resources in an e-Science environment, but does not necessarily have specialist knowledge about these resources.

- **Domain Expert**, someone who has specialized knowledge about one particular domain and the resources involved. Tries to make resources available in a generic and easily reusable form.

- **Tool Developer**, a (scientific) programmer who develops the core functionality of resources. Access to scientific instruments data processing software tools etc.

- **Administrator**, performs user- infrastructure and resource management. Keeps a Virtual laboratory in proper working order.

Scientific methodology should only be relevant for the scientist and domain expert, whose responsibility it is to ensure the scientific validity of experiments. There is a big burden on the domain expert in developing a shared resource. Not only should he himself be able to use it in a scientifically sound way, others - e.g. non-experts - should be able to do this
as well without his direct supervision. The scientist using shared resources should also be very much aware that he is relying on the developers of the shared resource for the scientific soundness of his experiment. Due to the importance of sharing and reuse within e-Science, performing an experiment should be interpreted in the wider sense as defined in the previous section. An e-Science experiment includes the definition of all assumptions as well as dissemination of results, be they resources which can be reused by others or just data. e-Science environments are setup in such a way that they allow for easy reproducibility. Repeating experiments can be made easy through the use of a workflow environment. In such an environment all steps of an experiment are defined and can be easily repeated many times without much extra administrative burden. Recreating an experiment performed by another scientist is also relatively easy through the use of shared resources when the original experiment and the reproduction are both performed in similar e-Science environments. This does bring with it the danger that errors inherent in either the e-Science environment in general or in the shared resources are repeated unnoticed in the reproduction of an experiment.

2.3.2 Empirical Cycle for e-Science

Now that the definition of e-Science is understood and we know the roles of the different people involved a clear methodology is needed. This is needed in particular for both the domain expert making his work available as a shared resource and for the non-expert scientist using the e-Science environment. First we define what we consider to be the empirical cycle for e-Science (see figure 2.2) and compare each step of this cycle to the one for science in general that was defined earlier in this chapter. Then we discuss how these differences differentiate e-Science from science in general and if these differences can be considered a paradigm shift.

Theory: e-Science differentiates itself from science in general in the way experiments are performed. Thus the theory step of the cycle remains more or less unchanged for e-Science.

Design Workflow: The workflow design process is analogous to formulating a Hypothesis. In its most abstract view a workflow consists of available input and desired output. This is similar to the research question which is used as the starting point for formulating a hypothesis. To move from the most abstract view of a workflow, the input and output requirements, to a concrete workflow is a process of refinement into multiple workflow steps. The assumptions made in this refinement process are clearly defined until an executable workflow is produced. This is similar to arriving at a testable hypothesis based upon the original research question. Defining workflow elements and resources as well as defining data are all part of explicitly defining assumptions in e-Science. This can be limited to data types (i.e. integer) and resource location (i.e. ip-nr 145.50.10). It can go further by providing meta
data such as temperature measured in Celsius with -273.15 lower boundary or MySQL database located at Foo maintained by John Doe. The advantage of using meta data, based on an ontology, is that it makes the theory laden nature of data more explicit: the assumptions made for representing data are clearly defined. This can potentially increase standardization and reusability of results. While not every workflow is a hypothesis in e-Science, a hypothesis can be expressed in a workflow. In fact a workflow can be as valid a method of expressing a hypothesis as a mathematical function is. This does require that both the workflow language as well as the processes making up each workflow step are formally described. In this way a workflow effectively expresses the hypothesis in a formal language. Insight into how this can be achieved will be provided in later chapters.

**Perform Workflow:** Actually performing a workflow is similar to doing observations in science in general. Not only are observations done within the workflow, the fact whether the workflow itself is actually executable and delivers the expected type of data is also verified. To ensure reproducibility, but also to aid the analysis of the workflow execution, intermediate data for each of the workflow steps is stored. A record is kept of the provenance of data, the steps that led to their creation. Similarly the provenance of the workflow steps can be kept: in which workflows these components were employed.

**Analysis / Share Resources:** Furthermore workflow patterns, common topologies of resources in a workflow, can be stored as abstract workflows. All of this helps the scientist trying to reproduce his own work or that of his peers. Very much in the same way that a detailed written description and laboratory logbook can help in traditional science. It also helps the scientist to analyze the data produced by the workflow, possibly leading to a change in the workflow, or even in the underlying theory. Just like in science in general, multiple experiments are usually performed before the underlying theory is changed or considered to be confirmed, in e-Science multiple executions of different workflows are performed.

**Roles in Empirical Cycle for e-Science**

The scientific process in e-Science is divided over multiple roles, the scientist and the domain expert. The Scientist is mainly concerned with the empirical cycle, whilst the domain expert is involved in the dissemination of resources and associated knowledge, as well as ensuring reproducible behavior for the resources he makes available. The scientist who performs experiments uses the e-Science specific version of the empirical cycle. Within this cycle, as presented in figure 2.2, the domain expert can also have a small task in defining workflow resources. This occurs when the scientist uses shared resources. Furthermore when previously defined abstract workflows are used to compose the workflow for an experiment. The domain expert, which pro-
2.3. METHODOLOGY OF E-SCIENCE

Figure 2.2: The empirical cycle within an e-Science context, as defined in this thesis.

vided these abstract workflows, has taken part in the workflow definition phase as well. The main methodology for the domain expert concerns itself with dissemination and to a lesser extent reproducibility. In practice this means creating shared resources and associated methodologies. Methodologies can consist of both workflow patterns and documentation on how to use resources. In many ways this is analogous to publishing results in scientific publications, because by making resources available you invite other scientists to test their validity as well as reproducing results from previous experiments. It can also augment scientific publications by offering a workflow interface to the scientists work which a reviewer can access and review remotely. Clearly validation of the shared resources is a very important part in making shared resources available in a scientific context.

Sharing resources in e-Science

The methodology as presented in figure 2.3 deals with sharing resources. It is a simple abstraction from the concrete executable resource. The creation of a shared resource from an existing dedicated resource, whether this is a small software component or something fundamental like grid access, is guided by four main criteria that have to be satisfied:

- **Consistency** A workflow component must always be able to reach a final finished state and the content of its output must always remain consistent with the associated data definitions.

- **Generality** A workflow component should minimize its dependence on: specific resources on which to run as well as specific types of data required for either input or output.

- **Simplicity** The number of inputs/outputs parameters needed for the correct functioning of the component needs to be limited. The com-
CHAPTER 2. METHODOLOGY

Figure 2.3: Methodology for creating shared resources

- Communication and computation performed should happen as efficiently as possible with a minimum of overhead.

- **Useability** The design time of a workflow employing a shared resource should be minimized, a shared resource should be quick and easy to implement.

It is obvious that there will be a trade-off to some extent between simplicity and generality on the one hand and useability on the other. Below we describe a three step process to derive a shared resource from an existing dedicated resource.

- First all experiment specific parts of a resource have to be reviewed. A decision needs to be made on how far to abstract. The higher the abstraction the broader the applicability of the shared resource. At the same time higher abstraction means a longer route to implementation when a resource is used in a different context. With higher abstraction usability goes down. Abstraction can thus be as simple as removing instantiations of resource parameters, usually though it will be a more complex task. Every aspect of a resource that in some way interacts with other resources will have to be reviewed to see if it is general enough. This includes: parameter defaults and boundaries, types of data input which are accepted as well as the form of data output.

- Knowledge associated with using a resource has to be made explicit i.e. documentation help, links to proper successful uses of the resource, links to abstract workflows and ontologies of associated expert knowledge.
Finally a validation has to take place in which the generalized component is tested by scientists outside of the domain of the domain expert who is generalizing his resource.

In 2.2.1 the theory laden nature of both observation and hypothesis was touched upon. In e-Science these issues manifest themselves among other things as a bias in the representation of data and the implementation of resources. It should be the aim to minimize these biases when creating a shared resource.

The validation part in practice never stops, a workflow should be seen as a form of hypothesis which is falsifiable. If any of the assumptions used in defining a scientific theory turns out to be false, the theory itself can be falsified. Similarly if any of the assumptions made in the construction of a workflow is incorrect the whole workflow can be incorrect. Assumptions here include assuming the work of others is correct. Using any shared resource in a workflow should be done with appropriate skepticism to its correctness. Anyone offering a resource or using it should be aware to what extent it has been validated before giving too much value to results attained using this resource.

The scientist constructing an experiment using shared resources should carefully consider for each resource whether its use is appropriate and after running the experiment evaluate if the shared resource performed its task properly. When this is not the case he should consider a different shared resource, or build his own.

### 2.4 Differences Science and e-Science

Do the features of e-Science that distinguish it from regular science like massive data, massive computing power, virtual organizations and sharing of resources truly enhance science as we know it? What improvements do they bring to classical scientific problems? First the ability of e-Science to handle massive amounts of data and have massive computing power allows the scientist a bigger search space in his quest for knowledge. As the amount of observations and associated hypotheses grows larger, the scientist can put less constraints based on prior assumptions on his search for knowledge, thereby potentially increasing the objectivity of his work. Similarly virtual organizations and the sharing of resources allow more people to work on one potentially larger problem. Projects such as the Large Hadron Collider[37] would not be possible without some form of e-Science infrastructure. Sharing of resources also allows scientists to build on each others work in a more efficient way than before. There are also potential problems in e-Science. Working towards sharing resources encourages standardization especially in data representation. The prospect of interoperability can encourage the use of a standard representation which is less suitable to the
problem but works well with other parts of the workflow. Interoperability therefore discourages scientists to experiment with different representations based on alternative assumptions, thus potentially decreasing the objectivity of data and hypotheses.

The term paradigm shift is used frequently in connection with e-Science. It can refer to any number of distinct paradigm shifts:

- A paradigm shift for scientific disciplines who have become dependent upon computational resources.
- It can be used to denote a shift from Object Oriented Computing towards Service Oriented Computing.
- The virtualization of computational and data resources.
- A paradigm shift for the methodology of science, from ”conventional science” to ”enhanced science”

Can it be argued that e-Science is a paradigm shift in the specific way that Kuhn defines it, or are we dealing here with a more general use of the term? And is e-Science a paradigm shift for all of science, for computer science concepts or for the scientific disciplines newly introduced to the concept. Kuhn describes a paradigm shift as being about the acceptance of scientific theory and the assumptions which go with this theory. One classic example of a paradigm shift is the proof for the four color theorem in 1977 by Appel and Haken[36]. The four color theorem states that any map or indeed any plane divided into regions can be colored using four colors in such a way that every border between two regions has different colors on each side. The proof by Appel and Haken was unique in that it was generated using a computer. Another example which seems relevant here is the use of the telescope and microscope for observations. Before their introduction direct observation by the human senses was the only method considered in science. The arrival of instruments such as the telescope and microscope brought about a mechanization of scientific observation. In both of these examples the assumptions changed, in case of the four color problem the assumption that only humans can create proofs was changed. It was now clear that computers could generate proofs no human could realistically generate. Similarly the telescope and microscope changed the assumption that observation was done only directly by the naked eye. If e-Science is a paradigm shift, it needs to change some important assumption in the scientific process.

First let us turn our attention to two types of paradigm shift mentioned earlier. For scientific disciplines such as Biology and Biochemistry that have recently started to explore the massive amounts of data contained in the genetic code, e-Science has brought about a new type of experimentation. Where before studies into relevant related work could be done by hand, now
2.4. DIFFERENCES SCIENCE AND E-SCIENCE

databases all over the world need to be consulted in an automated fashion. One can argue that this is a paradigm shift. The assumption of what the maximum size of the search space is and what can be found in this search space has dramatically changed. Similarly in computer science the shift towards service oriented architecture (SOA) which is employed in e-Science can be seen as a paradigm shift. SOA can be seen as shift in paradigm where previously there were only the concepts of client server architecture and object oriented computing. On the other hand one could argue SOA is an additional concept since it has not replaced nor made obsolete concepts such as client server architectures and object oriented computing.

The infrastructure used in e-Science aims to virtualize data and computational resources through grid technology, while human resources are virtualized as virtual organizations. We need to know whether this virtualization is just an addition of another tool that can be employed for scientific experiments or whether it brings with it a change in our understanding of the basic concepts of science. e-Science as a scientific method, one could say is a theory about what the best method for doing certain types of science is.

There are different assumptions of what the basic concepts are in this scientific method as compared to traditional science. For instance an observation is not just a numerical representation as it is often in traditional science. An observation also has meta data detailing its provenance: where was this data generated, where has it been used in the past, what is its relation to other data and so on. This is information which is usually kept implicit or at least separate from observations. Yet all information that can be kept in meta data associated with an observation could have been recorded in a traditional scientific methodology. Virtualization demands that it is made more explicit but the concept itself is not a new one, scientists have been recording provenance in lab logbooks for a very long time. Similarly one can say that in e-Science a hypothesis is an abstract workflow. This is a new formal method of notation, but just as with observations there is no fundamental change in the semantics of the concept. Neither concept of observation nor of hypothesis changes in a way which makes it incompatible with previous scientific methodologies.

What does change - as mentioned at the beginning of the discussion - is the size of the search space and with that the assumption on what can potentially be found in scientific experiments. Just as the search space increases for bio-informatics, it can increase for many other scientific disciplines. It is important to note that not all scientific disciplines are as data dependent as bio-informatics. Thus the sharing of other resources, particularly scientific software, will need to increase dramatically in order for the same type of shift to occur for other disciplines. This is one of the aims of e-Science but in case of software it is inherently more complex than sharing data. The sharing of software on the scale that bio-informatics share data has not occurred yet.
Another property of a paradigm shift according to Kuhn is that it is accompanied by vigorous arguments of the proponents and detractors of the new theory. Thus for e-Science to be a paradigm shift according to Kuhn we should see heated debate, and scientific papers both arguing for, but also against e-Science and perhaps in favor of a different methodology. The discussion that perhaps needs to occur most for e-Science is whether the increase in search space is worth the loss of control and understanding of the resources used; whether shared resources under the control of others can be trusted enough to produce accurate results.

In its current state e-Science is clearly not a paradigm shift for science as a whole. There has not been the widespread reuse of resources within in all disciplines that employ e-Science, that would constitute a dramatic change in the way science is performed. Nor has there been a a heated debate or even a vigorous discussion on the merits of sharing resources. For the moment it can only be a contributing factor in the paradigm shifts for a number of specific disciplines such as bio-informatics. In the following section we take a look at what it would take for e-Science to cause a paradigm shift for science in general.

2.5 Future Scenario e-Science

To get a better picture of what e-Science looks like when it does constitute a paradigm shift for science in general we present a future scenario for e-Science. As described in the previous section such a paradigm shift would be achieved if many resources can be shared. By many we mean not only the storage and computational resources that can currently be shared through grid technology. We also mean data resources such as the databases of gene information as used by bio-informatics or scientific publications accessible through webservices such as the medline database. These are all currently available for sharing. Most work is needed in facilitating the sharing of software resources used for performing experiments, scientific instruments as well as expert knowledge.

Problems that need to be addressed by a future e-Scientist when building an experiment with a large (and diverse) number of shared resources at his disposal are:

- Connectivity, do different software resources connect and do they have the same semantic understanding of the data they communicate.

- What workflow model should be employed for each stage of the experiment, and if multiple models are to be employed how do they relate to each other.
  - Hierarchically where a workflow using one model has to execute inside a workflow using another model.
2.6. CONCLUSION

- Sequentially where a workflow using one model is executed after the other has finished.

- Ensure different workflow models can inter-operate with each other, and do not violate each others rules for proper operation.

- What level of abstraction is needed when composing the workflow to enable either hierarchical or sequential relationships.

- What level of abstraction is needed to best represent the experiment for dissemination (both reuse and cooperation).

It is clear that workflow plays a central role in these problems. The solutions are therefore also associated with workflows. A future e-Scientist will have the following workflow related solutions at his disposal when building and performing an experiment:

- A workflow system suitable to his domain which can interact with workflow systems of other domains.

- Software resources specific to his domain and from other domains which apply to his research. As well as a way to easily discover these resources.

- Methodologies associated with these software resources which make explicit all relevant knowledge to properly integrate and use them in an experiment.

- The validity of each software component itself is proven, the main task for the scientist is to ensure the combination of different software resources possibly running in different environments is valid.

- A set of tools which can help the scientist to determine the correctness of his workflow. Whether it will run, produce the desired output and whether it provides an answer to his research question.

2.6 Conclusion

In this chapter a very high level methodology for e-Science was derived from general scientific methodology. Furthermore the merits of e-Science as a paradigm shift were discussed. It is clear from the methodology that workflows and workflow components play a central role in e-Science. The next chapters will go into more detail concerning the formal grounding for workflow design. We look at the current state of the art in many of the areas mentioned in the future scenario and what is needed to get closer to this future scenario.