Research in a Connected World

Collection Editors:
Alex Voss
Elizabeth Vander Meer
David Fergusson

Research in a Connected World

Collection Editors:

Alex Voss
Elizabeth Vander Meer
David Fergusson

Authors:

Malcolm Atkinson
Tobias Blanke
Ana Lucia DA COSTA
David De Roure
Stuart Dunn
Donal Fellows
David Fergusson
Paul Fisher
Jeremy Frey
Carole Goble
Sarah Harris
Mark Hedges
Andy Kerr
Erwin Laure
Steven Newhouse
Gergely Sipos
Martin Turner
Elizabeth Vander Meer
Alex Voss
Katy Wolstencroft
richard sinnott

Online:
< http://cnx.org/content/col10677/1.12/ >

C O N N E X I O N S

Rice University, Houston, Texas

Collection structure revised: November 22, 2009
PDF generated: February 5, 2011

Table of Contents

Welcome[1]

This book is a very timely contribution. It will help researchers in every discipline grasp the opportunities brought about by the digital revolution. Never before has society undergone such rapid change in the ways in which it communicates and collaborates. This brings untold and, as yet, unimagined new avenues of research and new methods for pursuing research. It demands new thinking and changes in the ways we undertake research.

The digital revolution is probably the most dramatic revolution that humankind has ever experienced. Its foundation is the pervasive growth of digital communication and digital devices. The global reach of digital communication — the Internet, mobile phones and media distribution — means that it is arriving in every nation and reaching most parts of society simultaneously. By contrast, the industrialisation of economies continues to spread after 300 years and the invention and diffusion of printing, telephony and broadcasting is being absorbed and overtaken by the digital revolution.

The rapid changes in personal communication — mobile phones, texting, instant messaging, email, social networking, blogging and twittering — accelerate the propagation of ideas. Governments are opening their data for public scrutiny (President Obama's and Prime Minister Brown's declarations in 2009). Commerce is transformed with Web2.0 models of business: RFID tags in stores, eBay and Amazon trading and marketing wholly online, transactions for tax, payments, books and planning via web pages, and Google's advertising market. In healthcare digital scanning is becoming routine, and will soon be followed by treatment tailored for your genomic variations. Digital cameras, digital video, digital TV and computer games are an intensely competitive global market.

Such a welter of activity is transforming the context of research. Researchers can now find, and be expected to find, an immense number of documents and many sources of data. They create huge volumes of data with faster and higher resolution instruments and generate more precise and larger-scale experiments with laboratory automation. They can use computational notations to describe and refine models precisely. They can access digital replicas of artefacts from museums and libraries — far more than they could ever visit in a lifetime of industrious research — creating virtual assemblies of rarities that could not be contemplated with the original objects. Sensors can be widely deployed to study the atmosphere, oceans and land; they can be worn to study physiology, predation, migration, mating and recreation. Volunteer researchers can gather data worldwide and respond automatically to opportunities and incidents. Global consortia can curate data and make it available for researchers, allowing thousands to collaborate in assembling knowledge and developing models to guide future decisions.

These changes pose new opportunities, raise new questions about research mores and may suggest revision of ethical and legal frameworks. Consequently, researchers need information to exploit the new opportunities, to engage in searches never before possible and to join in worldwide collaborations enabled by the new forms of communication. They will need to be agile and adventurous to thrive in the global competition. Creativity and insight will be amplified by the new methods. Leadership and charisma will assemble complementary skills from around the world to tackle the immense intellectual and practical challenges that face humankind today. Researchers will access the products and by-products of the global revolution as sources of evidence about human behaviour, sampling on scales never previously imagined. Those who engage will shape the way in which future research is done.

[1] This content is available online at <http://cnx.org/content/m32854/1.1/>.

To do this, researchers must gain new skills in computational thinking and data-intensive research. This will be a dynamic process evolving as the pace of the digital revolution throws up new questions and delivers new capabilities. This book is an excellent snapshot; a launching pad from which to get started. Its readers will find key insights and authoritative references, but they must expect to move on to rapidly develop and shape the ideas needed for research in a connected world. They will be on the lookout for claims that appear to break the fundamental principles of distributed systems, but they will also enjoy the rewards of being at the forefront as new methods and technologies make significant advances in research possible.

The most important factor in the success of *Homo sapiens* is their ability to communicate and collaborate. The connected world enables this as never before, as both the speed and scale of collaboration have experienced a step change. Those with the knowledge, enthusiasm and agility to exploit this transformation will pioneer new forms of global behaviour. It is vital that researchers draw on this new resource for combining human talent to address the world's most pressing challenges before it is too late.

When Sir John Taylor launched e-Science, he said, "**e-Science** is about global collaboration in key areas of **science** and the next generation of infrastructure that will enable it". This book shows that Taylor's assertion was a serious understatement. It shows that the new capabilities delivered by the connected world empower new kinds of human collaboration *for all forms of thinking and doing*. Research has a two-fold role: to pioneer these new ways of thinking and doing wherever it will achieve intellectual and practical advances, and to reflect on the deep changes that are underway in global society by recording the massive changes of the digital revolution and better understanding how they shape, and are shaped by, society. This book provides a window into research transformed by the digital revolution, revealing its benefits across disciplines and the added responsibilities that come with these new methods of working. It calls on researchers to observe, record and analyse the digital revolution. It is a valuable resource for researchers as they seize the opportunities brought by the digital age.

Malcolm Atkinson *UK e-Science Envoy* and Director of the e-Science Institute
David De Roure *National Strategic Director for e-Social Science*

Editor's Introduction to Research in a Connected World[2]

The massive availability of networked information and communications technologies today allows us to change the ways we go about our daily working lives as well as the way we spend out leisure time. New ways of shopping, of staying in touch with colleagues and friends, of learning or of navigating places have emerged that are enabled by the ubiquitous electronic devices and networked services that have become available over the past few years. Similarly, as researchers we are today utilising computers in many ways, be it through the use of basic services such as email or the utilisation of the most advanced digital technologies enabling new research methods. No matter what discipline we work in, there are legitimate questions about what potential use we might make of these technologies and what the implications of such use might be.

Over the past decade, funding organisations such as the UK's research councils have funded efforts to make the most advanced information and communications technologies available to researchers and investments are made to develop persistent and sustainable infrastructures to underpin a widespread uptake of digital methods – the development of e-Research. What has been lacking, however, is the development of appropriate learning material such as textbooks that would teach the basics of advanced information systems and digital methods in a way that is accessible to researchers from a wide range of disciplines. This book is an attempt to fill this gap. Its aim is to fill the gap between the initial interest generated by presentations of the potential of e-Research and the various training courses that convey the skills necessary to use specific technologies.

Chapter Outline

The book is divided into four main sections. The first two chapters provide a general introduction to the principles behing e-Research and introduce distributed systems, showing how they differ from single-user desktop systems. The second section discusses a number of different examples of e-Research from a range of disciplines, demonstrating how research can benefit from and be driven forward by the use of advanced information and communications technologies. The third section outlines a number of infrastructures for research that are available to researchers today and discusses the strategies behind the development of European grid initiatives that aim to provide a sustainable environment for the development of e-Research practices. Next, we discuss the role of data and its management over the research lifecycle as well as a number of relevant technologies. The fifth section discusses different ways that researchers can access infrastructure services and the ways they can be factored into actual everyday research practices. Finally, we conclude the book with a collection of resources that we hope will help the reader explore the field of e-Research further and make informed choices about the adoption of the technologies and methods described in this book.

[2]This content is available online at <http://cnx.org/content/m32855/1.2/>.

Acknowledgements

First of all, we would like to thank our colleagues who have contributed chapters to this collection. They have given generously of their time and the essential input of expertise without which this book could not have come into existence. We would also like to thank the organisations that have provided support in cash or in kind:

The UK's JISC has provided financial support through the funding for the e-Infrastructure Use cases and service usage models project.

The Scottish Informatics and Computer Science Alliance has supported the editing process by funding contributions made by Alex Voss.

The National e-Science Centre has supported the editing process by funding the contributions made by David Fergusson and Elizabeth Vander Meer.

The Manchester e-Research Centre has supported the editing process by funding contributions made by Alex Voss and by administering the production process of the first edition of the book.

Research in a Connected World[3]

Key Concepts

- data-rich science
- e-Research

Introduction

Research today is often critically dependent on computation and data handling. The practice has become known under various terms such as e-Science, e-Research, and cyberscience. We would like to avoid using these terms, but when it is unavoidable, in the interests of brevity, we use the term e-Research in a broad sense to include all information processing support for research. Irrespective of the name, many researchers acknowledge that the use of computational methods and data handling is central to their work.

There is no question that scientific research over the past twenty years has undergone a transformation. This transformation has occurred as a result of various factors. New technologies, leading to new methods of working, have accelerated the pace of discovery and knowledge accumulation not only in the natural sciences but also in the social sciences and arts and humanities. Advances in scientific and other knowledge have generated vast amounts of data which need to be managed well so that they can be analysed, stored and preserved for future re-use. Larger scale science enabled by the Internet, and other information and communication technologies (ICTs), scientific instrumentation and automation of research processes has resulted in the emergence of new research paradigms that are often summarised as **'data-rich science'**. A feature of this new kind of research is an unprecedented increase in complexity, in terms of the sophistication of research methods used, in terms of the scale of phenomena considered as well as the granularity of investigation.

e-Research involves the use of computer-enabled methods to achieve new, better, faster or more efficient research and innovation in any discipline. It draws on developments in computing science, computation, automation and digital communications. Such computer-enabled methods are invaluable within this context of rapid change, accumulation of knowledge and increased collaboration. They can be used by the researcher throughout the research cycle, from research design, data collection, and analysis to the dissemination of results. This is unlike other technological "equipment" which often only proves useful at certain stages of research. Researchers from all disciplines can benefit from the use of e-Research approaches, from the physical sciences to arts and humanities and the social sciences.

The following sections in this introduction will elaborate on these transformations in research and the role played by ICT, describing research collaborations, “big research” in a globalised world and participation in research.

Research Collaborations

Today’s research into social and scientific issues and problems often involves increased sharing of resources – because individual research institutions cannot afford having these resources or because they are inherently

[3]This content is available online at <http://cnx.org/content/m20834/1.3/>.

distributed (for example in the case of linked radio telescopes). The research community has changed, so that more work is done in international collaborations and these collaborations have become increasingly multi- or interdisciplinary.

Tackling the grand challenges of many disciplines today requires the coordinated effort of groups of researchers working on different aspects of a problem. Also, individual researchers can more rapidly increase their knowledge in a particular field if they are able to become part of an international and interdisciplinary collaborative network. Instead of working on their own or only with colleagues within their own institutions, researchers now often work in collaborations with colleagues in other institutions, who can provide specialist knowledge, skills or access to resources.

e-Research provides researchers with an environment for sharing resources and facilitates collaborations by making large, distributed data sets accessible, through enabling synchronous or asynchronous collaboration across geographical distances and providing access to resources regardless of location. This opening up of research means that researchers need not be held back by their own resource constraints and can more freely participate in cutting-edge projects.

e-Research Technologies Supporting Collaboration

e-Research technologies support the research collaborations described above by introducing a model for resource sharing based on the notions of "resources" that are accessed through "services". Resources can be computational resources such as high-performance computers, storage resources such as storage resource brokers or repositories, datasets held by data archives or even remote instruments such as radio telescopes. In order to make resources available to collaborating researchers, their owners provide services that provide a well-described interface specifying the operations that can be performed on or with a resource, e.g., submitting a compute job or accessing a set of data.

This simple underlying model of collaboration is complemented by additional functionality such as authentication and authorisation to regulate access to a resource or management functions such as resource reservation. It is important to note that the underlying model is kept simple and that any additional functionality layered on top of it is also formulated in terms of resources and services wherever possible. Using these general principles, it is possible to build a vast range of tools and applications that support collaborative research.

Computer-enabled methods of collaboration for research take many forms, including use of video conferencing, wikis, social networking websites and distributed computing itself. For example, researchers might use Access Grid[4] for video conferencing to hold virtual meetings to discuss their projects. Access Grid and virtual research environments provide simultaneous viewing of participating groups as well as software to allow participants to interact with data on-screen. Wikis have also become a valuable collaborative tool. This is perhaps best demonstrated by the OpenWetWare[5] website, which promotes the sharing of information between researchers working in biology, biomedical research and bioengineering using the concept of a virtual Lab Notebook. This allows researchers to publish research protocols and document experiments. It also provides information about laboratories and research groups around the world as well as courses and events of interest to the community.

Social networking sites have been used or created for research purposes. The myExperiment[6] social website is becoming an indispensible collaboration tool for sharing scientific workflows and building communities. Such sharing cuts down on the repetition of research work, saving time and effort and leading to advances and innovation more rapidly than if researchers were on their own, without access to similar work (for comparison to their own). Other social networking sites such as Facebook have been adopted by researchers and extensions have been built to allow them to be used as portal to access research information. For example, content in the ICEAGE Digital Library[7] can be accessed within Facebook.

[4] http://www.ja.net/services/video/agsc/AGSCHome/whatisaccessgrid.html
[5] http://openwetware.org/
[6] http://www.myexperiment.org/
[7] http://library.iceage-eu.org/

Systems Research in a Globalised World

Many researchers now devote a significant amount of their attention to global issues, which previously could not be addressed due to technological and informational limitations. These global issues include, for instance, climate change, pandemics, rainforest destruction and biodiversity. Such "big research" problems fall under wider contemporary concerns about living sustainably and understanding human biology and health (including the aetiology of diseases and the search for cures).

This ubiquitous global perspective has in large part emerged because of a worldwide exchange of information and the availability of data resulting from use of ICT, coupled with the use of ICT to organise that data. For example, the earth is seen as a system or as systems within systems, which necessitates the need for cross-scale research. Earth system science in geosciences provides a useful example of this change to "systems research". ICT is used to model and simulate integrations of geology, oceanography and environmental sciences, generating a more complex, holistic view than was possible prior to the increased use of computer enabled methods. There has also been a recent concerted development of systems biology, which involves integration of mathematics, engineering and computer science to manage the data deluge in biology in order to answer big questions concerning sustainable living and human health on a global level.

A significant number of researchers in the social sciences and arts and humanities have also taken up this global view. For the social sciences, this perspective is clear, for instance, in the idea of "global knowledge" and attempts to solve social issues relating to sustainable living through large-scale data gathering and analysis. In the arts and humanities, a global perspective is evident in the development of the Global Performing Arts Consortium[8] , an international database of performing arts resources, and in global cultural and international studies research which often relies on/requires access to large amounts of cross-culturally derived data to adequately substantiate conclusions.

Participation in Research – Democratising "Big Science"

e-Research not only enables scientists to tackle "big" questions, but it has also allowed for wider participation in research. Volunteer computing allows members of the public to support and take part in research conducted by teams of professional researchers by providing compute resources or by performing specific tasks that are part of the research process. For example, the SETI@home project[9] makes use of volunteers' desktop computers to search for extraterrestrial life while Folding@home[10] uses the compute power provided by volunteers to study protein folding. In the case of climateprediction.net[11] , any member of the public with appropriate computer equipment can contribute to the study of climate change. In each of these cases, tasks and data are shared across a network of dispersed computers, thus increasing the compute power and storage capacity available far beyond the capabilities of a single computer. Several of the examples of inspiring e-Research projects we will introduce here have been successful as a result of using volunteer computing.

Open Source Science is not just about direct public participation. It is also about transparency, so that the public has access to and can observe the research process. Open Notebook Science enables better collaboration among researchers at the same time that it makes research project records available online for perusal by the lay public. In this way, "big science" is democratised, no longer purely the product and tool of a cloistered research elite but an activity within a wider societal context that society members can take part in.

[8] http://www.glopac.org/
[9] http://setiathome.ssl.berkeley.edu/
[10] http://folding.stanford.edu/
[11] http://www.climateprediction.net/

Research in a Connected World - Fundamental Concepts and Inspiring Examples

Preceding sections in this introduction have presented a strong argument for the uptake of e-Research methods by illustrating their importance in a multitude of research endeavors. The Research in a Connected World brochure serves as an introduction to e-Research for those unfamiliar with such methods, revealing it's potential and promise for all disciplines. The brochure consists of individual modules that give researchers a grounding in fundamental concepts and a taste of what is possible when using computer-enabled methods.

We provide an introduction to distributed systems, contrasting them to desktop PCs, and then move on to detailed discussion of inspiring examples of e-Research, looking at projects in many different fields. These examples are followed by examples that show the wider impact of e-Research and explore the unique collaborations that have developed not only among other academic researchers but also between researchers and the wider public. The subsequent section of the brochure describes elements of and issues relating to distributed systems, beginning with a short history of distributed computing and including modules on the taxonomy of research computation problems, distributed computing architectures, issues concerning managing complex data, visualisation, use of portals and virtual research environments. A final module contains a list of relevant services and contacts.

We hope this resource will not only inform you but also inspire you to begin to use computer-enabled methods to further your research. If you already consider yourself an e-Researcher, we hope to have introduced you to new tools that you can begin to apply in your own work.

What is a Distributed System?[12]

Key Concepts

- distributed systems

Introduction

Over the past decades, as we have begun to explain, we have moved from processing the data that we can hold in a lab notebook to working with many thousands of terabytes of information. (For reference, a terabyte is a million megabytes, and a megabyte is a million letters. A plain textbook might be a few megabytes in size, as might a high-quality photograph — a terabyte is like a huge library.) And yet we keep striving to work with ever more: more genomic data; more high-energy physics data; ever more detailed astronomical photographs; ever richer seismographic measurements; ever more layers of interpretation of artistic details; ever greater volumes of financial data; ever more complex and realistic simulations. We drill down ever deeper into the details. How are we coping with this?

We are in the middle of a huge revolution in information processing, driven by the fact that our tool of choice for working with information — the computer — has been getting exponentially better ever since their invention during the Second World War. We live in the middle of an age of wonder. And yet, despite now being able to hold immense quantities of computation and storage in our hands, our desire to work with ever more has grown even faster.

Thankfully we have been living through another revolution at the same time; the telecommunications revolution. The telecommunications revolution started with the invention of the telegraph, but accelerated with the convergence of computers and telecoms to create the Internet. This not only allows people to share information, but also computers, and it has transformed the world. The first indication of just how amazing this would be came with the WorldWideWeb (WWW), the first internet system to really reflect everything that people do throughout society [link/reference here to history chapter]. But it will not be the last; the ripples from the second wave are now being felt, and it is the global research community that are in the lead. This second wave is Distributed Computing.

The Way Distributed Computing Works

Simply put, distributed computing is allowing computers to work together in groups to solve a single problem too large for any one of them to perform on its own. However, to claim that this is all there is to it massively misses the point.

Distributed computing is not a simple matter of just sticking the computers together, throwing the data at them and then saying "Get on with it"! For a distributed computation to work effectively, those systems must cooperate, and must do so without lots of manual intervention by people. This is usually done by splitting problems into smaller pieces, each of which can be tackled more simply than the whole problem. The results of doing each piece are then reassembled into the full solution.

[12]This content is available online at <http://cnx.org/content/m31661/1.1/>.

The power of distributed computing can clearly be seen in some of the most ubiquitous of modern applications: the Internet search engines. These use massive amounts of distributed computing to discover and index as much of the Web as possible. Then when they receive your query, they split it up into fast searches for each of the words in the query. The results of the search are then combined in the twinkling of an eye into your results. What about locating computers on which to execute the web search? That is itself a distributed computing problem, both in the process of looking up computer addresses and also in finding an actual computer to respond to the message sent on that address.

Early distributed systems worked over short distances, perhaps only within a single room, and all they could really do was to share a very few values at set points of the computation. Since then, things have evolved: networks have got faster, numbers of computers have got larger and the distances between the systems have got larger too.

The speeding up of the networks (from the telecommunications revolution) has been extremely beneficial as it has allowed many more values to be shared effectively, and more often. The larger number of computers has only partially helped; while it has meant that it is possible to use more total computation and to split the problems into smaller pieces (allowing a larger overall problem), it has also increased the amount of time and effort that needs to be spent on communication between the computers, since the number of ways to communicate can increase (see Figure 1).

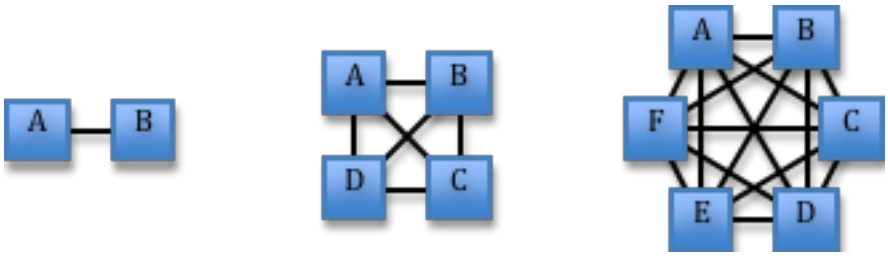


Figure 1: The growth in the number of links as the number of computers goes up.

There are, of course, ways to improve communication efficiency, for instance by having a few computers specialize in handling the communications (like a post office) and letting all others focus on the work, but this does not always succeed when the overall task requires much communication.

The distance between computers has increased for different reasons. Computers consume power and produce heat. A single PC normally only consumes a small amount of power and produces a tiny amount of heat; it is typically doing nearly nothing, waiting for its users to tell it to take an action. With a computational task, it would be far busier and will be consuming electrical energy in the process; the busier it is, the more it consumes and produces heat. Ten busy PCs in a room can produce as much heat as a powerful domestic electric heater. With thousands in one place, very powerful cooling is required to prevent the systems from literally going up in smoke. Distributing the power consumption and heat production reduces that problem dramatically, but at a cost of more communications delay due to the greater distances that the data must travel.

There are many ways that a distributed system can be built. You can do it by federating traditional supercomputers (themselves the heirs to the original distributed computing experiments) to produce systems that are expensive but able to communicate within themselves very rapidly; this remains favoured for dealing with problems where the degree of internal communication is very high, such as weather modelling or fluid flow simulations. You can also make custom clusters of more traditional PCs that are still dedicated to being high-capability computers; these have slower internal communications but are cheaper, and are suited for many "somewhat-parallel" problems, such as statistical analysis or searching a database for matches (e.g., searching the web). And you can even build them by , in effect, scavenging spare computer cycles from across

a whole organization through a special screen saver (e.g., Condor, BOINC); this is used by many scientific projects to analyse large amounts of data where each piece is fairly small and unrelated to the others (e.g., Folding@Home, SETI@Home, Malaria Control).

Special Challenges for Distributed Computing

We are now moving from having the number of computers working on a problem being small enough to fit in a building or two to having tens of thousands of computers working on single problems. This brings many special challenges that mark distributed computing as being a vastly more complex enterprise than what has gone before.

The first of these special problems is security. The first aspect of security is the security of the computers themselves, since few people feel like giving some wannabe digital mobster a free pass to misuse their computers. The second aspect is the security of the data being processed, much of which may be highly confidential or a trade secret (e.g., individual patients' medical data, or the designs for products under development). The third aspect of security is the security of the messages used to control the other computers, which are often important in themselves and could be used to conduct a wide range of other mischief if intercepted.

The second special problem of distributed computing is due to the use of systems owned by others, either other people or other organizations. The issues here are to do with the fact that people ultimately retain control over their own systems; they do not like to cede it to others. This behaviour could be considered just a matter of human nature, but it does mean that it is extremely difficult to trust others in this space. The key worries relate to either the computer owner lying about what actions were taken on their systems (for pride, for financial gain, for spite, or just out of straight ignorance) or the distributed computer user using the system for purposes other than those that the computer owner wants to permit.

Interoperability presents the third challenge for distributed computing. Because the systems that people use to provide large-scale computing capabilities have grown over many years, the ways in which they are accessed are quite diverse. This does mean that a lot of effort has been put into finding out access methods that balance the need for efficiency with those of security and flexibility, but it also means that frequently it is extremely difficult to make these systems work together as one larger system. Past attempts by hardware and software vendors to lock people in to specific solutions have not been helpful here either; researchers and practitioners want to solve a far more diverse collection of challenges than the vendors have imagined there to be. After all, the number of things that people wish to do is limited only by the human imagination.

These challenges can be surmounted though, even if the final form of the solutions is not yet clear. We know that the demands of security can be met through a combination of encryption, digital signatures, firewalls, and placing careful constraints on what can be done by any program. The second challenge is being met through the use of techniques from digital commerce like formal contracts, service level agreements, and appropriate audit and provenance trails. The third, which will become ever more important as the size of problems people wish to tackle expands, is primarily dealt with through standardization of both the access mechanisms (whether for computation or for data) and the formal understanding of the systems being accessed by common models, lexicons and ontologies.

The final major challenge of distributed computing is managing the fact that neither the data nor the computations are open to relocation without bounds. Many datasets are highly restricted in where they can be placed, whether this is through legal constraints (such as on patient data) or because of the sheer size of the data; moving a terabyte of data across the world can take a long time, and in no time the most efficient technique becomes sending disks by courier, despite the large quantity of very high capacity networks that exist out there.

This would seem to indicate that it makes sense to move the computations to the location of the data, but that is not wholly practical either. Many applications are not easy to relocate: they require particular system environments (such as specialized hardware) or direct access to other data artefacts (specialized databases are a classic example of this) or are dependent on highly restricted software licenses (e.g., Matlab, Fluent, Mathematica, SAS, etc.; the list is enormous). This problem does not go away even when the

users themselves develop the software — it is all too easy for them to include details of their development environment in the program so that it only works on their own computer or in their own institution — writing truly portable software is a special and rare talent.

Solving Problems using Distributed Computing

Because of these fundamental restrictions that will not go away any time soon, it is important for someone tackling a problem (often an otherwise intractable problem) with distributed computing to take them into account when working out what they wish to do. In particular, they need to bear in mind the restrictions on where their data can be, where their applications can be, and what sort of computational patterns they are using in their overall workflow.

As a case in point, when performing drug discovery in relation to a disease, the first stage is to discover a set of potential candidate receptors for the drug to bind to in or on the cell (typically through a massive database search of the public literature, plus potentially relevant patient data, which is much akin to searching the web). This is then followed by a search for candidate substances that might bind to the receptor in a useful way, first coarsely (using a massive cycle scavenging pool) and then in depth (by computing binding energies using detailed quantum chemistry simulations on supercomputers). Once these candidates have been identified, they then have to be screened to see if there are warning areas associated with them that might make their use inadvisable (another database search, though this time probably with more ontology support so that related substances such as metabolites are also checked; this step will probably involve real patient data). If we look at the data-flow between these steps, we see that the amount of data actually moved around is kept relatively small; the databases being searched are mostly not relocated, despite their massive size. However, once these steps are completed, the scientist can have a much higher level of confidence that their *in silico* experiments will mean that follow-up clinical trials of the winning candidate will succeed, and in many cases it may be possible to skip some parts of the trials (for example, it might be the case that the literature search uncovers the fact that a toxicity trial has already been performed).

This use-case has other interesting aspects in terms of distributed computing, in that it involves the blending of both public and private information to produce a result. The initial searches for binding receptors relating to a particular disease will often involve mainly public data — archives of scientific papers — and much of the coarse fit checking that identifies potential small molecules for analysis will benefit from being farmed out across such large numbers of computers that the use of public cycle scavengers makes sense; it would be difficult to backtrack from the pair of molecules being matched to exactly what was being searched for. On the other hand, there are strong reasons for being very careful with the later stages of the discovery process; scientists are looking at that stage in great depth at a small number of molecules, making it relatively easy for a competitor to act pre-emptively. Moreover, the use of detailed patient data means that care has to be taken to avoid breaches of privacy. In addition, the applications for the detailed analysis steps are often costly commercial products. This means that the overall workflow has both public and private parts, both in the data and computational domains, and so there are inherent complexities. On the other hand, this is also an application area that was impossible to tackle until very recently, and distributed computing has opened it up.

It should also be noted that distributed computing has many benefits at the smaller scale. For example, it is a key component of providing acceleration for many more commonplace problems, such as recalculating a complex spreadsheet or compiling a complex program. These tasks also use distributed computing, though only within the scope of a workgroup or enterprise. And yet there is truly a continuum between them and the very large research Grids and commercial Clouds, and that continuum is founded upon the fact that bringing more computational power together with larger amounts of data allows the discovery of finer levels of detail about the area being studied. The major differences involve how they respond to the problems of security, complex ownership of the parts, and interoperability. This is because those smaller scale solutions can avoid most of the security complexity, only needing at most SSL-encrypted communications, and they work within a single organization (which in turn allows imposition of a single-vendor solution, thus avoiding the interoperability problems). Of course, as time goes by this gap may be closed from both sides; from

the lower end as the needs for more computation combine with the availability of virtual-machines-for-hire (through Cloud computing) and from the upper end as the benefits of simplified security and widespread standardized software stacks make adoption of scalable solutions easier.

Chapter 1

Examples of e-Research

1.1 Archaeology[1]

Key Concepts

- The early emergence of computing in archaeology
- Data storage and integration – the Archaeology Data Service and Archaeotools
- The role of ICT in excavation – Virtual Research Environment for Archaeology and geospatial archaeology
- The critical importance of integrating any new ICT-based tool or procedure with existing research practices

1.1.1 Introduction

"[A]rchaeologists should look beyond the short term when planning how to use a computer. The world of archaeology is likely to be considerably different in twenty years from now (2009), so archaeologists need to plan with future change in mind."

J. Moffett, in *Computers for Archaeologists*, Ross et. al. (eds.), 1991.

The growth in the last ten years or so of e-Research methods across the academic spectrum, and their impact on archaeology, could not have proved Moffett more correct in his prediction. In many ways, archaeology differs significantly from other arts and humanities disciplines in its uptake, theory and application of computational methods. For one thing, computing has played a central role in the development of archaeology's intellectual traditions for decades; and a coherent community of archaeological computing professionals is now well established. The thirty six year-old *Computer Applications and Quantitative Methods in Archaeology* conference provides a trusted international forum for this community to network and disseminate its outcomes at the cutting edge of technology, and for those experts to undertake critical assessment of that technology (see, e.g., Clark 2007: 11, and http://www.leidenuniv.nl/caa/). More broadly, the history of archaeology as a discipline can be broadly characterized by a progression from 'antiquarian' interest in aesthetically pleasing artefacts to the development of the principles of typology and the evolution of material culture in the late nineteenth century, to the present-day emphasis on systematic and consistent record-keeping (for an overview, see Lock 2003: 1-13). This progression may be seen as an ongoing, iterative transformation in archaeologists' approach to, and relationship with, *information*.

In the past, many archaeological approaches have assumed that information handling, processing and visualization is seen as at best ancillary to, and at worst disconnected from, from the interpretive process of understanding the past. This is reflected in early treatments of the subject: in 1985 for example, Martin Carver wrote that '[i]n spite of, or perhaps because of, a great deal of breathless proselytizing, it is the

[1] This content is available online at <http://cnx.org/content/m31033/1.1/>.

computer's relevance to creative archaeology that is still doubted, and it is the wisdom of investing precious thinking-time in such a potential wild-goose chase that must be weighed' (Carver 1985: 47). The misperceptions Richards and Ryan identified in the same year, that computers are 'black boxes producing magic answers', operated by 'practitioners of some mystical black art' (Richards and Ryan 1985: 11) are likely to be less widespread today, due not least to the exposure of archaeologists to ubiquitous internet, email, and other digital technologies, both at home and at work. However, such critical caution is no more misplaced now than it was back then. Indeed the very existenceof the archaeological community's self-awareness (or perhaps outright scepticism) of the approach to computing and Information and Communications Technology provides an excellent background in which to consider the exponential growth and potential of e-Research in recent years. The following examples demonstrate how, in 2009, e-Research tools and methods, can contribute to the critical interpretive process of archaeology. What 2029 will bring is, of course, anyone's guess.

1.1.2 Data: integration and understanding

Archaeology produces a vast amount of data of a massive range of types. These include artefact descriptions, measurements, site plans, context plans, photographs, and cartographic and spatial data. Every excavation has particular challenges for data gathering and recording, and the possible responses of excavators to these challenges are constrained by scale, resources, the type of material, and so on. But there is no question that e-Research methods offer enormous potential for supporting such processes; and few archaeologists would doubt the desirability of integrating data from different sites. Eiteljorg (2004) writes of 'the hope that data storehouses could be used by scholars to retrieve and analyze information from related excavations, thus permitting broader syntheses' (Eiteljorg 2004: 22): broader synthesis is at the core of academic archaeology, and is vital for any interpretation that seeks to embrace any combination of site, inter-site, or regional scale. However, there is an obvious tension between the structures and standards any database must impose in order to be useful, and the unordered (and incomplete) nature of the archaeological record (see Lock 2003: 85-98). e-Research technologies can support researchers faced with such problems in a number of ways. One approach is the construction of domain-specific ontologies and controlled vocabularies, which can describe and link concepts, and map between different groups of concepts. Thus if artefact of type A is found at site X, then a linked ontological system should be able to identify further examples of type A at site Y, even if the artefacts have been otherwise recorded or described differently. This approach has limitations – those producing data still have to describe and/or annotate the information in a way that conforms with, or can be adapted to, the ontology. This will impose extra costs on already-overburdened resources. On the other hand, standardized metadata and data storage systems can be immensely useful and easy to implement, if supported by centralized support services and repositories like the Archaeology Data Service (http://ads.ahds.ac.uk[2]).

Other approaches seek to apply Natural Language Processing (NLP) technologies to primary archaeological material and secondary archives. One such example is the Archaeotoolsproject, conducted as part of the AHRC-JISC-EPSRC Arts and Humanities e-Science Initiative by the universities of York and Sheffield (http://ads.ahds.ac.uk/project/archaeotools/[3]). Archaeotools identifies and extracts references to 'what', 'when' and 'where' entities in so-called 'grey literature'. Grey literature refers to reports of (usually small-scale) archaeological investigations that have been produced and archived, often never to be seen again. The NLP process allows information to be tagged in a systematic way according to 'what', 'where' and 'when' and structured into *facets* for facetted browsing. It should therefore be possible, for example, to search across a range of disparate archaeological reports for references to data concerning Early Medieval coins from North Eastern England [*when, what, where*], even if the information has not been tagged or described in such terms and the point of being recorded. In another important development, Archaeotools uses NLP-generated entities to search for the information according to the terms in existing controlled vocabularies such as Sites and Monuments Records (SMRs): as will be seen below, integrating e-Research methods within existing practices is essential for archaeology, so allowing researchers to search using the terms and conventions they

[2] http://ads.ahds.ac.uk/

[3] http://ads.ahds.ac.uk/project/archaeotools/

are already familiar with is critical.

1.1.3 Computers and excavation

As indicated above, the data recovered from excavations are often hugely complex, but excavation itself is also a very complex task. When situated within the nexus of data gathering and the realities of excavation practice, e-Research presents both significant challenges and great opportunities. Digital methods have been integrated with excavation practice at a low level for many years. For example, it is common practice for excavators to take the points of particular positions from a Total Station Theodolite (TST), place these in a local data store, and download them to a computer for processing back at base. However, the ubiquity of networked systems, along with the availability of (often proprietary) software such as ArchaeoData, has meant that e-Research technologies are now being more widely applied in field archaeology. In many cases, this has 'only' meant speeding up and/or facilitating existing work; allowing for the documentation of objects and their contexts and transferral of this information to the excavation's database faster and more efficiently. In essence, many of the software packages are database-oriented, aiming to support excavation directors and post-excavation researchers in organizing and structuring the site's data according to existing organizing principles and structures.

Some projects however have considered in greater depth the intellectual and interpretive implications of using such technology, thereby addressing Carver's 'relevance to creative archaeology' critique. Ian Hodder for example has reflected on the implications of separating observation from interpretation, and noted that '[i]nterpretation occurs at the trowel's edge. And yet, perhaps because of the technologies available to deal with very large sets of data, we have as archaeologists separated excavation methods out and seen them as prior to interpretation. Modern data-management systems perhaps allow some resolution of the contradiction. At any rate, it is time it was faced and dealt with' (Hodder 1997: 693). Hodder's own response to this problem, the online site database of the Çatalhöyük project (http://www.catalhoyuk.com/database/catal/[4]), seeks to present fully and simply all the data about the site, including the free text interpretations of the recorders.

An archaeological project frequently referenced in the literature is the Roman urban excavation of Silchester in Hampshire, which has trialled the use of e-Research technologies in the Virtual Research Environment for Archaeology (VERA: see http://vera.rdg.ac.uk[5] , also case study at http://engage.ac.uk[6]) project in conjunction with its existing Integrated Archaeological Database (http://www.iadb.co.uk/). VERA, funded under JISC's VRE programme, has tested use of a broadband network at the site and various onsite digital capture methods. Those used earlier in the project, such as PDAs and tablet PCs for recording information about artefacts and plans of trenches and features, proved less successful for a variety of reasons (a major one being that liquid screens perform badly in bright sunlight). Currently however, the project is trialling the use of digital pens for recording information. This follows exactly the procedure for recording information using 'normal' pens, with the exception that users can 'dock' the digital variety at the end of the working day, downloading handwriting and converting it to ASCII text using automated handwriting recognition. The VERA project has noted that integrating such technologies with existing onsite workflows is critical if they are to stand any chance of wider adoption (see Warwick at. al. 2009 for full discussion). This greatly speeds up and facilitates the process of entering the data into the IADB; and it may well be that, as the method is further refined and deployed in the field, it will provide some hitherto unforeseen contribution to understanding the data as well.

e-Research methods and technologies have also played a significant role in the development of geospatial archaeology. Geographic Information Systems (GIS) have long been at the forefront of computational archaeology: the large quantities of data from large-scale surveys and site-wide analysis, and the need to reference it within a broader spatial framework such as a global coordinate system, has ensured this. However, the emergence of the so-called 'Geospatial Web' in recent years (for a recent review see Scharl and Tochtermann 2007) has led to new ways of linking, sharing and understanding geospatial information online.

[4] http://www.catalhoyuk.com/database/catal/
[5] http://vera.rdg.ac.uk/
[6] http://engage.ac.uk/

The availability of high quality satellite imagery from services such as Google Earth (GE) has generated a good deal of recent interest in the archaeological community (see Ullman and Gorokhovich 2006), as have the means of marking up and describing data in such environments. In GE's case this is Keyhole Markup Language (KML), which allows a dataset to be created in a GE view and then shared, updated and added to by another user. Although its impact on field archaeology is not likely to be great in the near future, GE and other 'virtual earth' platforms are undoubtedly of interest to scholars wishing to link and contextualize archaeological data online (e.g. Elliott and Gillies 2009).

1.1.4 Summary: Improving Archaeological Research

Archaeology has always thrived on technological innovation. The increasingly information-rich ways of working into which the UK's academic milieu is moving forms a backdrop for the ever-convoluted relationships between archaeologists and their data. Current e-Research technologies will not provide any panaceas: these equate with what Hodder describes (above) as 'modern data-management systems'. They may have yet to prove that they can transfer very 'fuzzy' data from the ground into the highly structured and quality assured forms that appear in archaeological publications; but there seems little doubt that tools and methods such as relational databases, Natural Language Processing, cultural heritage ontologies, quantitative profiling, geospatial computing, and field-based digital data capture, form a 'methodological commons'. Whether taken together for the discipline as a whole, or separately in individual projects or research exercises, this collective set of e-Research tools and methods can provide a type of 'enabling support' that is simply unprecedented for archaeologists so that they may undertake the research process in better, faster and – possibly – completely new ways.

1.1.5 References / Further Reading

Carver, M. 1985: The friendly user. In Cooper, M. A. and Richards, J. D. (eds.), Current issues in archaeological computing. British Archaeological Reports International Series 271: 47-61.

Clark, J. T. 2007: An introduction to digital discovery: Exploring new frontiers in human heritage. In Clark, J. T. and Hagemesiter, E. M. (eds.), Digital Discovery: Exploring new frontiers in human heritage. Computer Applications and Quantitative Methods in Archaeology: Proceedings of the 34th conference, Fargo, ND, April 2006: 11-14.

Eiteljorg, H. 2004: Computing for Archaeologists. In Schreibman, S., Siemens, R. and Unsworth, J. 2004: *A Companion to Digital Humanities*. Blackwell, London: 20-30.

Elliott, T. and Gilles, S. 2009: Digital geography and classics. In *Changing the Center of Gravity: Transforming Classical Studies Through Cyberinfrastructure*. Special issue of *Digital Humanities Quarterly* (Winter 2009: v3 n1), Gregory Crane and Melissa Terras (eds.): http://digitalhumanities.org/dhq/vol/3/1/000031.html[7] .

Hodder, I. 1997: 'Always momentary, fluid and flexible': towards a reflexive excavation methodology. Antiquity 71: 691-700.

Lock, G. 2003: Using computers in archaeology: towards virtual pasts. Routledge, Taylor and Francis, London.

Moffett, J. 1991: Computers in archaeology: approaches and applications past and present. in Ross, S., Moffett, J. and Henderson, J. (eds.), Computing for archaeologists. Oxford University Committee for Archaeology Monograph No. 18. Oxford: 13-39.

Richards, J. D. and Ryan, N. 1985: Data processing in archaeology. Cambridge Manuals in Archaeology.

Scharl, A. and Tochtermann, K. (eds.), *The Geospatial Web* (Springer 2007).

Ullmann, L. and Gorokhovich, Y., 2006: 'Google Earth and some practical applications for the field of archaeology', CSA Newsletter Vol. XVIII, No. 3 (2006), published online: http://csanet.org/newsletter/winter06/nlw0604.html[8]

[7] http://digitalhumanities.org/dhq/vol/3/1/000031.html
[8] http://csanet.org/newsletter/winter06/nlw0604.html

Warwick, C., Baker, M., Clarke, A., Fulford, M., Grove, M., O'Riordan, E. and Rains, M. 2009: iTrench: A study of user reactions to the use of information technology in field archaeology. Literary and Linguistic Computing 24 (2).

1.2 Text Analysis in the Arts and Humanities[9]

Key Concepts

- Digital scholarship
- Data-driven research
- TextGrid and collaborative working
- HiTHeR (High ThroughPut Computing in Humanities e-Research) and use of e-Infrastructure

1.2.1 Introduction

According to UNESCO reports, Britain tops the European lists of research publications per year in philology, literature and other text-based studies such as philosophy. Worldwide, Britain overtook the U.S. in 2006 in terms of book publications per year. In the list of countries for number of book publications in the latest available year, Britain is no. 1. These figures emphasize the urgent need for the British textual studies communities to explore new ways of dealing with this deluge of research data.

Based on these quoted figures, collaboration becomes fundamental to digital scholarship in textual studies. No researcher alone will be able to cope with the plethora of new daily published material. Furthermore, text analysis in the humanities can be a tedious and time-consuming task. But advanced computer-enabled methods make the process easier for digital or digitised works. Researchers can search large texts rapidly, conduct complex searches and have the results presented in context. The ease brought to the analysis process allows the researcher to engage with texts more thoroughly and can then lead to the development of insightful, well-crafted interpretations of texts.

Various projects have emerged internationally in recent years that allow for a new scale of textual studies research, in keeping with the idea of new data-driven research. Software developed by the US MONK (Metadata Offer New Knowledge) project helps humanities scholars discover and analyze patterns in texts,[10] while its sister project SEASR (Software Environment for the Advancement of Scholarly Research) enables digital humanities developers to design, build, and share software applications that support research and collaboration in textual studies.[11] Aus-e-Lit is an Australian project aimed at Australian literary scholars to allow them to seamlessly search across relevant databases and archives to retrieve reliable information on a particular author, topic or publication.[12] These are just three projects quite closely linked to e-Research initiatives, but there are many more. For over fifty years, there has been a worldwide academic movement to work on Digital Humanities, resulting in many achievements, especially in the field of textual studies. It is impossible in the space of this chapter to list all of these projects (for a history of Digital Humanities and some of the involved textual scholarship see Schreibman, Siemens et al. 2004). Instead, we shall concentrate on two projects linked to both Digital Humanities and e-Research, which exemplify in very particular ways two major new developments in textual studies research that are directly linked to the shift in methodologies based on data-driven research: the German TextGrid project illustrates the value of new collaborative research in textual studies, while the UK project HiTHeR (High ThroughPut Computing in Humanities e-Research) demonstrates the effective use of e-Infrastructure to support everyday research in the Digital Humanities.

[9] This content is available online at <http://cnx.org/content/m31502/1.2/>.
[10] http://monkproject.org/ (<http://monkproject.org/>)
[11] http://seasr.org/
[12] http://www.itee.uq.edu.au/~eresearch/projects/aus-e-lit/

1.2.2 Collaboration in textual studies - TextGrid

TextGrid[13] is primarily concerned with historical-critical editions for modern cross-language researchers. Such historical-critical editions often form the basis for more light-weight editions for study and reading. Such editions can be very large and very detailed. They cannot be the result of the work of one individual researcher alone, but have to be the result of a collaborative effort. It is TextGrid's key innovation to facilitate such (virtual) collaboration across language and national barriers.

In its first phase of funding, TextGrid delivered a modular platform for collaborative textual editing, mainly based on the community standard of the Text Encoding Initiative (TEI).[14] As a community grid for textual studies, TextGrid forms a cornerstone in the emerging German e-Humanities agenda. Its success has also been noted in the UK, where the arts and humanities e-Science initiative allowed researchers to experiment with new technologies to cope with the research data deluge in textual studies. The UK e-Science Scoping Study for textual studies, written by Professor Peter Robinson from Birmingham University, quotes TextGrid as a prime example of how to advance literary and textual studies with new digital services, because it addresses the need for collaborative resource creation, comparison (that is, collation and alignment), analysis and annotation.

TextGrid focuses on advancing digital scholarship for a particular community: TEI-based textual studies research. At the centre of its technology innovation is the deployment of an integrated development environment for the creation of critical editions called TextGridLab. Based on the Eclipse platform, TextGridLab uses Grid technologies for storage and retrieval of textual studies resources. It supports all activities, stakeholders and challenges in the textual studies research lifecycle. Resource discovery, via the web interface or TextGridLab modules, is aided by searching across the entire TextGrid data pool – either full text or metadata-restricted.

Decentralized and collaborative work is always sensible when primary sources grow very large and need to be made available and linked to each other in complex metadata schemes. This is due to the quality of these resources, which demand an integration of different viewpoints. Additionally, new mass quantities of resources need the support of high-performance technology in new investigations of ways that advanced text mining solutions can add to the linking and discovery of textual studies resources. The UK JISC Engage funded HiTHeR project has taken on this challenge.

1.2.3 Use of e-Infrastucture in textual studies – HiTHeR

In the Digital Humanities, many text-based collections are exposed via searchable websites. One of these resources is the Nineteenth Century Serials Edition (NCSE) in the UK.[15] The NCSE, a free online scholarly edition of nineteenth-century periodicals and newspapers, has been created as a collaborative project between Birkbeck, University of London, King's College London, the British Library, and Olive Software. The UK Arts and Humanities Research Council funded the project from January 2005 to December 2007. The NCSE corpus contains circa 430,000 articles that originally appeared in roughly 3,500 issues of six 19th Century periodicals. Published over a span of 84 years, materials within the corpus exist in numbered editions and include supplements, wrapper materials and visual elements. A key challenge in creating a digital system for managing such a corpus is to develop appropriate and innovative tools that will assist scholars in finding materials that support their research, while at the same time stimulating and enabling innovative approaches to the material. One goal would be to create a 'semantic view' that would allow users of the resource to find information more intuitively. Such a semantic view can be created by offering users articles with common content through a browsing interface. This is a typical classification task known from many information retrieval and text mining applications. (Nentwich 2003)

According to Toms and O'Brien (2008), the work of humanities researchers using digital resources is concerned with access to sources, the presentation of texts and the ability to analyse texts using a well-defined set of analysis tools. HiTHeR promises direct retrieval of relevant primary sources for research on

[13] http://www.textgrid.de
[14] http://www.tei-c.org/index.xml
[15] http://www.ncse.ac.uk

the NCSE collections. It provides an automatically generated browsing interface, which allows for the crucial Humanities 'chain of readings' activities that define most Humanities researchers' work. In Humanities research processes, new relevant resources are based on the initial discovery of other relevant resources. HiTHeR offers an interface to primary resources by automatically generating a chain of related documents for reading.

However, the advanced automated methods that could help to create such a browsing view using text mining to aid the information retrieval task by users require greater processing power than is available in standard desktop environments. Prior to the current case study, we experimented with a simple document similarity index to allow journals of similar contents to be represented next to each other. Initial benchmarks on a stand-alone server allowed us to conclude that (assuming the test set was representative) a complete set of comparisons for the corpus would take more than 1,000 years!

Governments, private enterprise and funding bodies are investing heavily in digitization of cultural heritage and humanities research resources. With advances in the availability of parallel computing resources and the simultaneous need to process large and complicated historical collections, it seems logical to turn attention towards the best parallel computing infrastructures to support work as envisioned in the HiTHeR project. In HiTHeR we set up an infrastructure based on High Throughput Computing (HTC), which uses many computational resources to accomplish a single computational task.

The HiTHeR project created a prototype infrastructure to demonstrate to textual scholars, and indeed to humanities researchers in general, the utility of HTC methods using Condor. It uses Condor to set up a Campus Grid. In our case, we have built a Campus Grid using underutilized computers from two institutions, which share a building at King's College London: the Centre for Computing in the Humanities (CCH) and the Centre for e-Research (CeRch). We use two types of computer systems: underutilized normal desktops and dedicated servers. Both, CCH and CeRch, have a large number of desktop machines and servers, used to present their vast archives and online publications. While the servers contain several Terabytes of data, they have underused processing capabilities which can be made available for advanced processing. Additionally, the Condor Toolkit can use the national research infrastructure in the UK, the National Grid Service (NGS), which is a free service to UK researchers and provides dedicated advanced computing facilities.

The evaluation showed that the time used for calculating document similarity could be reduced significantly by using the HTC resource. However, it also showed that more work is needed to exactly determine how text mining for humanities can best be served by UK research infrastructures. More research is also needed to determine when HTC can serve the needs and when dedicated hardware is required.

1.2.4 Summary: The Potential of e-Research Technologies in Textual Resource Analysis

There is great unplugged potential for using e-Research technologies in textual resource analysis. Computation of textual resources is quite well researched and there are by now many well performing algorithms and data structures to serve the needs not only of the general user, but also the specific needs of researchers. But less work has been done to consider infrastructural needs for the future of research based on these methodologies. More user studies are required to analyse existing work in Digital Humanities involving textual resources. We need to better understand how new methods such as text mining could be used, or how the discipline of textual studies and humanities in general is transformed by the ability to do more data-driven empirical research. The field of humanities has the opportunity to move towards a new more empirical way of working in which more and more resources, increasing not only in number but in size, become easily available. Interest in such new working practices already exists as repeatedly shown in research reports for conferences. This chapter has presented just a few of the many projects working on this agenda. TextGrid looks at how collaboration can enable new research in the textual studies, while HiTHeR looks at enhancing online editions in Digital Humanities using text mining approaches. As the need for research using large digital corpora increases, other projects will emerge that will further advance computational text analysis in arts and humanities research.

1.2.5 References

Brockman, W. S., L. Newmann, et al., Eds. (2001). Scholarly Work in the Humanities and the Evolving Information Environment. Washington DC, Digital Library Federation. Council on Library and Information Resources.

Gietz, P., A. Aschenbrenner, et al. (2006). TextGrid and eHumanities. Proceedings of the Second IEEE International Conference on e-Science and Grid Computing, IEEE Computer Society.

Nentwich, M. (2003). Cyberscience. Research in the Age of the Internet. Vienna, Austrian Academy of Science Press.

Schreibman, S., R. Siemens, et al., Eds. (2004). A Companion to Digital Humanities. Oxford, Blackwell Publishing.

Toms, E. and H. L. O'Brien (2008). " Understanding the information and communication technology needs of the e-humanist." Journal of Documentation**64**.

1.3 Climate Prediction[16]

Key Concepts

- Climate prediction
- Climate modeling

1.3.1 Introduction

Over 100 years ago Svante Arrhenius, who would go on to win the Nobel Prize for Chemistry, postulated that changes in levels of carbon dioxide in the atmosphere could affect global temperatures. We now know that a number of natural and industrial chemicals, including water vapour and carbon dioxide, affect the properties of the atmosphere. Levels of carbon dioxide, methane and nitrous oxide have increased markedly as a result of fossil fuel use, agriculture and land use change since the start of the industrial revolution. This past and projected future rise in emissions, coupled with the observed rise in global mean temperatures over the past three decades, has led to considerable concern about future climate change. For geoscientists seeking to understand how the global climate system operates, the challenge has been how to represent a system that is not fully accessible, because of the time and space constraints of experiments conducted on, and observations of, environmental systems. The solution has involved the development of numerical models to represent physical processes.

Climate is the statistical average of the weather over long (30 year) periods of time. An old saying goes: "Climate is what you expect; weather is what you get." Climate models have evolved over four decades from simple energy balance models to the massively complex global system models of today, which are largely extensions of models used for weather forecasting. This evolution has been enabled by the extraordinary development of computational power, largely in supercomputers but increasingly through dispersed applications, and the management of the correspondingly massive data sets. Not only have these developments provided far greater scope for more complex numerical models, the technology has fundamentally changed the scientific questions that can be posed.

At the heart of such climate models are the mathematical equations representing geophysical processes. Such relationships are non-linear, necessitating a range of numerical methods to provide approximate iterative solutions to the equations. There is no one"best" climate model. Model components and subcomponents are combined to answer specific questions and at a time and space scale of interest to the user. For example, a global system model might include ocean, atmosphere, ecosystem and ice sheet components at coarse resolutions. A regional model might use finer numerical grids to resolve small-scale meteorological phenomena, but will need to use the outputs of the global model as a boundary condition to its more detailed study.

[16]This content is available online at <http://cnx.org/content/m31704/1.1/>.

The trade-off between model components and scale has typically reflected the computational efficiency of the model and its ability to include as much of the detailed physical processes as possible. However, even for models which incorporate detailed physical processes, the non-linear nature of the problem means that equations can only be solved approximately. There is inherent loss of information at scales below the averaging (grid) scales of models and through the process of parameterizing physical relationships within the model. As a result, confirmation that a complex climate model actually represents the underlying physical processes of the global climate is rather challenging; instead, the onus is on the modeller to establish a sufficient degree of confidence in the model through its ability to recreate observed data to a reasonable accuracy. This chapter first introduces the different approaches used by modellers for climate prediction, detailing the complexities of this endeavour. It concludes by considering the importance of collaborative working in development of predictive models and the future challenges facing climate science.

1.3.2 Predicting the future? Model Ensembles

Attempting to predict the future has profound implications for model development and application. Until very recently, information from climate models about possible future climates has been presented as a scenario or projection, without specified probabilities. This has reflected the difficulty of managing the core uncertainties associated with climate modelling:

- Changing boundary conditions: these are the factors affecting the climate system that are treated as separate from, or outside, the climate system. These include changes in solar output, volcanic eruptions and human factors, such as levels of emissions of greenhouse gases
- The natural internal variability of the global climate system: the global climate system is chaotic, which means that very small changes in one location and at one point in time can lead to large differences in other locations at a future point in time
- The extent to which the models accurately represent (parameterize) the physical processes of the climate system; in other words our understanding of the component parts of the global climate system and how they respond to change

With increasing demands from the public and private sectors for information to manage future changes in climate, and with enhanced computational power, climate modellers can now begin to explore this range of uncertainty. Different approaches exist for developing probabilistic climate predictions. One relies on brute force, based on large *ensembles* of simulations from computationally efficient models. This approach carries out large numbers of model runs in which model parameters are varied within their current range of uncertainty. Model parameterizations which fail to replicate existing climate observations are rejected, with the remainder used to explore future climate scenarios. This approach is complemented by continuous improvement in model representations of physical processes and higher resolution data, which improves the parameterizations – the model representation of physical processes. The second approach for developing probabilistic predictions relies on "expert judgement", drawn from small *ensembles* of state-of-the-art models.

An *ensemble* consists of many simulations run with a specific climate model, each one slightly different from the rest. The uncertainty associated with natural climate variability is studied using "initial condition" *ensembles*, which vary the distribution of temperature, wind, humidity and other factors at the beginning of the simulation. The uncertainty associated with the model boundary conditions is studied using *ensembles* with different scenarios for human-induced or natural greenhouse gas emissions. These seek to examine the full range of possible boundary conditions of, for example, future global greenhouse gas emissions from society under different economic futures. The final source of uncertainty reflects the quality of the model representation of the climate; this is studied by using ensembles of different climate models. This approach assumes that the available models from climate modelling centres capture the full range of plausible behaviour, though this is unlikely to be the case. This source of uncertainty remains least studied, and potentially most important.

These uncertainties, associated with natural variability, boundary conditions and model representation, are not independent of each other. To explore the full "parameter" space of the models requires the integration of these different model runs into a grand ensemble of ensembles!

Obviously, these climate model ensembles are extraordinarily computationally-intense. A global climate model might typically need to solve its equations for each grid point on a simulated 30 minute time-step. While weather forecasting might require simulated time to be of the order of a few days, climate forecasting requires simulated time to be of the order of decades. So each simulated time-step in the model must be repeated tens of thousands of times. As a result, accurately modelling atmospheric processes alone can require high performance computer runs of several days. Because of these constraints, until recently complex climate models have only been run on supercomputers. With the need for grand ensembles of model runs to explore climate uncertainty more fully, the growth of distributed computing approaches is appealing.

A public example of a grand ensemble has been the work of *climateprediction.net*, which used distributed computing resources of the general public to run ensembles of a version of the UK Meteorological Office Unified Climate Model to examine the implications of doubling levels of carbon dioxide in the atmosphere. Members of the public donate spare computing capacity on their personal computers to do one or more of the simulations. Public interest has been staggering. Over 100,000 people from 150 countries have taken part to date. More than 70,000 simulations of the climate model have been completed. In contrast, at the time climateprediction.net started, the largest model ensemble reported in the literature was 53.

Outputs from these model ensembles provide a snapshot of the range of uncertainty associated with future climate – whether internal model uncertainty or external uncertainty associated with future global greenhouse gas emissions. Although these outputs have been termed probabilistic predictions, they are not objective probabilities. Instead they are subjective probabilities, based on the quality and availability of information at the present time. They cannot capture all uncertainties about future climates: different climate models produce different results based on the same forcing scenarios. For example, one model might suggest that rainfall increases across the Amazon; another might suggest that rainfall will decrease. Different models are better at representing different elements of the climate system. When many model outputs converge, we might have more confidence that we are seeing a robust result from the many possible representations of the climate system.

In this way, "probabilistic" predictions do not reduce uncertainty in future climates, but they do make the range of possible futures more transparent. It should, in theory, make decision making more transparent. One recent example of such probabilistic predictions is the publication of the UK Climate Impacts Programme (UKCIP) climate scenarios, which provide probabilistic projections for seven overlapping 30 year time slices through until 2099 for 25km x 25km land grid squares in the UK.

1.3.3 Summary: The Importance of Collaborations and Future Challenges for Climate Research

Twenty years ago, it was not uncommon for a doctoral graduate student to be expected to produce a numerical model of some element of the climate system, for example an ice sheet or an ecosystem model, track down the necessary data to test the model, and use the model to examine a science question. Typically, the model details (parameter values; computer code) would not be published, and only sparse elements of the model outputs would be made available through published papers. To all intents and purposes, these model results cannot be replicated, since the model is likely to be discarded or developed through time, with poor notification of model versions and parameter values.

In time, modellers recognised the importance of benchmarking models against other models, to examine the strengths and weaknesses of their particular representation of the climate system. Funding was made available to bring together modellers and to ensure effective data and model management. More recently, community-wide models of different elements of the global climate system have been developed. For example, in the UK, the GLIMMER ice sheet model is available as a resource to the academic community. This model captures the learning developed over many years by different modelling groups across the UK. A new doctoral graduate student who is exploring science questions about ice sheets is likely to use this as the starting point, perhaps with the aim of improving physical processes within the model, or using the model to answer new science questions. These community model developments, enabled by distributed access to computer models and more effective model and data management, have changed the way modellers work together. Instead of

a largely individual approach, collegiate approaches are now the norm.

e-Research methods have had a fundamental impact on the way in which climate science is undertaken. These methods have changed how individuals and modelling groups work together. They have changed the very science questions that can be posed. However, the very success of high performance or distributed computing to produce colourful ensemble model outputs has also disguised critical questions about what models can usefully offer and how the outputs are used by decision makers and politicians. To those outside the modelling community, "probabilistic predictions" might well be assumed to be objective probabilities of future events, rather than subjective assessments based on incomplete information. Such a perception will affect the decisions that are taken about managing future climate impacts. Yet, climate models are not truth machines; they are inherently partial. In practice, there is an asymmetry between explanation and prediction of complex systems. Satisfactory explanation of the future is possible even when absolute prediction is impossible.

Separately, extensive work by behavioural economists has shown that humans are inherently poor at calculating and managing probabilities when making monetary decisions. Yet the output of these ensemble model runs shows a wide range of probabilistic outcomes, from futures with little change to futures with catastrophic change. Taking the next step and enabling more effective decision making on the basis of these model outputs remains challenging.

While computer-enabled methods of research may not be able to address these problems of human decision making, they have enhanced climate science through the development of models that expand our knowledge of a range of possible future climates that could occur based on different variables.

1.4 e-Malaria[17]

Key Concepts

- Computational drug design – docking programs
- Drug target
- eMalaria project as outreach – linking research and teaching
- 3D Stereochemistry
- SMILES (Simplified Molecular Input Line Entry System)

1.4.1 Introduction: Motivation and Drug Design

The original motivation for the eMalaria project[18] was to bring together school students with university researchers in the hunt for a new anti-malarial drug. The challenge was offered, via the web, to school students to design molecules and test them using a computational drug design approach to see if the molecule might be suitable for further research as an anti-malarial drug. The participants were not merely going to be passive suppliers of computational resources in the model of the very successful cycle stealing grid drug screening systems as pioneered by Graham Richard's group[19] and followed most recently by the World Community Grid[20] . Instead, they were involved with the design and selection of potential drug molecules so that they could learn more about modern approaches to drug design and development. This necessitated a significant amount of background and tutorial material to support the investigations.

At its core the eMalaria system uses the Cambridge Crystallographic Data Centre[21] and the Gold docking engine Software[22] . The docking program needs to investigate the whole conformational space of the potential drug and its fit into the enzyme active pocket. Gold evaluates the quality of the fit with an energy scoring

[17] This content is available online at <http://cnx.org/content/m31767/1.1/>.
[18] http://emalaria.soton.ac.uk
[19] http://www.chem.ox.ac.uk/curecancer.html
[20] http://www.worldcommunitygrid.org/
[21] http://www.ccdc.cam.ac.uk/
[22] http://www.ccdc.cam.ac.uk/products/life_sciences/gold/

function. This function takes into account the main contributions to the forces acting between the atoms in the enzyme and the potential drug; the intermolecular forces as well as any strain imposed on the drug molecule.

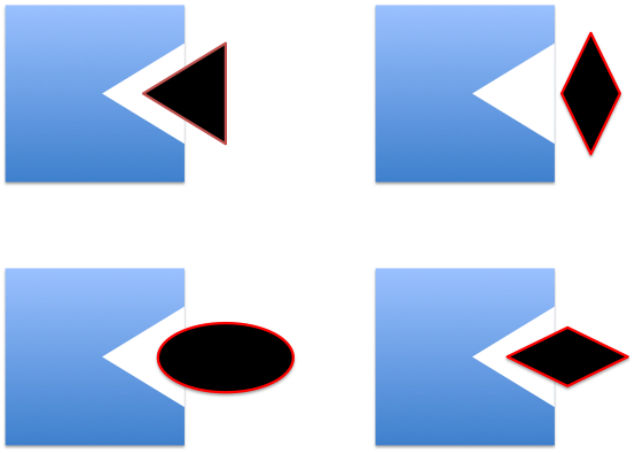
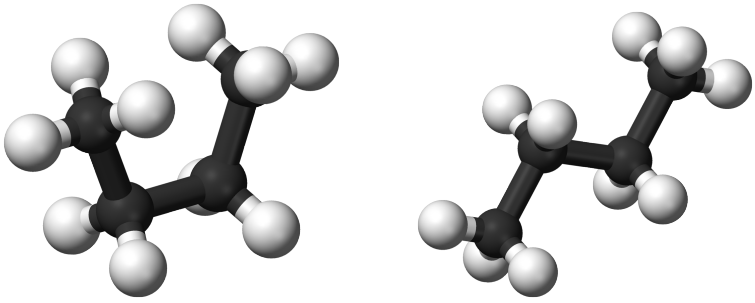


Figure 1.1: showing how the shape of the potential drug molecule influences the ability to bind into the pocket of the enzyme active site

With even moderate sized molecules the size of the conformational space is very large. That is, there are very many ways in which the molecule can be twisted without breaking any bonds. Each different shape can produce different interactions between the potential drug and the enzyme pocket. It is a very significant calculation to evaluate these different possible ways in which the drug could bind into the pocket and locate the best fit.



(a) (b)

Figure 1.2: Two possible conformations of the hydrocarbon butane demonstrating how rotation about carbon-carbon single bonds gives rise to different molecular shapes. For larger molecules with many possible 'rotatable' bonds a very large number of conformations are possible. (Images from Wikipedia)

To support the anticipated computational load and provide a model that would grow in capacity in parallel with the increase in users, we chose to use a cycle steeling grid. Several machines within the University of Southampton initially provided the computational resource. The expectation was that as schools joined the

project they would provide additional computational cycles. This was made possible while protecting the commercial code by using the United Devices (UD)[23] software, which ensured that the core Gold code was secure.

To supply the docking engine we originally built software to enable molecules to be drawn and converted from a 2D sketch to a 3D model with a realistic molecular conformation using empirical rules and molecular mechanics and semi-empirical quantum codes. The drug target for the initial studies was the DHFR protein, chosen for the different way the DHFR protein is regulated in mosquitoes and humans, which makes it very suitable as a target to block. The structure of the DHFR protein was obtained from the protein data bank (PDB)[24] and a scoop suitable for docking produced.

The scripts developed set up the computational job, bringing together the molecule submitted by the user with the drug target, and then submitted the job to the UD system. Web interfaces were provided to keep track of the individual's runs and allowed the 2D, 3D molecular and docked structures to use the very versatile Java based Jmol program[25] . Keeping track of structures and results for each user required some system to identify the users. We were careful to collect no personal information as we had to be sure that we were clear of any data protection requirements, especially in regard to those users who might be under 18 years of age. In subsequent use for undergraduate teaching we linked the eMalaria system to the University (LDAP) authentication system, allowing access to designated students.

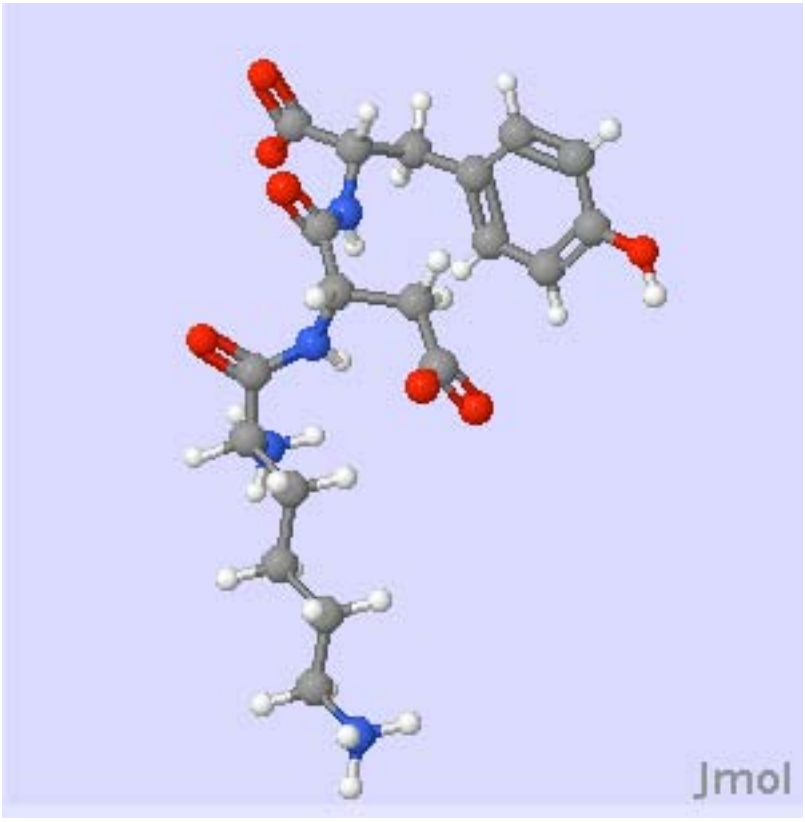
The site was designed to be as accessible as possible for students with special educational needs, particularly those with dyslexia. Information is presented in manageable amounts and boxed away from the navigation tools to avoid confusion. The website has been designed to use cascading style sheets (CSS) to allow students to pick the text font, colour and background that makes it easiest for them to read the website. At all times the user's built in browser settings are able to override the standard website style sheet. This means that if a user has set their computer up to give the best font and colour options for them, our site will follow these instructions. A text only version is available to make the site accessible to students using a screen reader, and the downloadable documents are available as word documents so that font colour and size can be changed before printing if necessary. The site and associated materials have been designed in accordance with the British Dyslexia Association[26] 's guidance, which turned out to provide a good visual feel for all students.

1.4.2 Details of the Project

1.4.2.1 Initial Outreach & Schools Project

The eMalaria system was used in a number of schools in the region and supported by visits by the team. The project was also supported by workshops and talks delivered by project team members at the request of participating schools. The general format of these consisted of a 45-minute talk covering drug design methodology and how the project follows these principles, followed by a 45 minute session where students were given a compound to use as a lead for modification. Students then tried to find modifications to this molecule that would lead to better docking scores. The visualization of molecular shape was highly productive in supporting chemical education and the drug challenge taken up actively by many of the students; some interesting molecules resulted. We deployed the system for a number of University Open Days with perhaps the youngest participant being about 6 years old; this group designed molecules that came close to breaking the system!

[23] http://www.univaud.com/
[24] http://www.pdb.org/pdb/
[25] http://jmol.sourceforge.net/
[26] http://www.bdadyslexia.org.uk/



(a)

We found that the display of the highest docking score achieved by a School or Open Day group was a valuable incentive to try and develop a molecule with a higher score. Tools to manage a group collectively were very useful in motivating and providing feedback to a class.

1.4.2.2 Evolution of the project

New chemical services became available as the project moved forward and in its new incarnation it makes use of one of the available 2D molecular sketch packages (Marvin[27] from ChemAxon[28]) and the 3D optimization within this package. This reduced the number of bespoke services which we needed to provide, but removed some of the filtering for sensible molecules that we had previously put in place. We supplied additional interfaces to be able to automate the molecular design, and the SMILES[29] input has proved very useful for more advanced teaching applications with undergraduates. It is now easier to bring many structures from a paper, for example, into the eMalaria system and to manipulate them as we did in automatically constructing the SMILE sequences for all the possible tri-peptides from the 20 most common naturally occurring amino-acids.

As well as serving as a general reminder of 3D stereochemistry and examples of using SMILE strings, we made extensive use of the eMalaria system in the final year undergraduate course on Chemical informatics for which the students, as part of their project work, were required to build a QSAR (Quantitative Structure Activity Relationship) model to model docking score based on simple descriptors that were calculated using web based freely accessible programs (e.g. molecular mass, volume, LogP).More recently the task set the undergraduate class was to model real experimental data on inhibitors of Plasmemsin taken from the literature (K. Ersmark etal, J. Med Chem. 2005, **48**, 6090-6106), using the eMalaria docking scores as one of the possible descriptors in the QSAR model of inhibitor activity.

1.4.2.3 Summary and Future Plans

The eMalaria approach to outreach and teaching was formally evaluated as part of the eBank project with a review of the material and interviews with students who had used eMalaria as part of their course. Typical teaching scenarios for using eBank and eMalaria included getting students to retrieve small molecules from the eBank[30] and eCrystals[31] systems and then putting them into the eMalaria site in order to undertake a range of manipulations and investigations. Such a linkage between eBank, eCrystals and eMalaria provides a demonstration of a means of explicitly integrating research activities with learning and teaching practice. The course was designed to progressively build up complexity and use of real and authentic data, clearly linked and of relevance to work-based learning, which was timely, as these students had just completed their six-month work placement. Very positive outcomes were noted and the lesson of linking research and teaching is exploited in the outreach context as well as in university teaching.

Returning to our original aims of outreach to the school community, eMalaria now has a presence on the University of Southampton's island in Second Life. A 'fruit machine' allows a visitor to the island to select three amino acids to make a tri-peptide. The molecule is rendered as a floating 3D image and the docking score against the DHFR enzyme target is shown. In the future a visitor will be able to design more general molecules and obtain a docking score by launching an eMalaria calculation against a selected target.

The eMalaria system is actively being used in undergraduate teaching but currently less so for outreach. We are planning to increase the support for the system and to make it available again regionally and nationally with the addition of more computational power. We are exploring distributed computing solutions with other systems such as BOINC to see if they can provide a more appropriate solution for some of the longer calculations that could be useful in advanced teaching. This project has shown the wider impact of e-Research through its involvement of young students in the process of drug design. While learning how to

[27] http://www.chemaxon.com/marvin/
[28] http://www.chemaxon.com/
[29] http://www.daylight.com/smiles/
[30] http://www.ukoln.ac.uk/projects/ebank-uk/
[31] http://ecrystals.chem.soton.ac.uk/

design molecules using computer-enabled methods, students also provided research results which have fed into university researchers' work on finding anti-malarial drugs. Projects such as this one can not only lead to new discoveries but also inspire students to become enthusiastic about science early on.

1.4.3 Acknowledgments

Thanks to EPSRC and JISC for the financial contributions, CCDC for making the Gold software available, and Robert Gledhill, Sarah Kent, Andrew Milsted, Brian Hudson, Steve Wilson, Simon Coles and Jon Essex for their contributions to the project.

1.4.4 Bibliography

The eMalaria website http://eMalaria.soton.ac.uk[32]

http://www.rsc.org/chemistryworld/News/2005/July/07070501.asp

Frey, Jeremy G., Gledhill, Robert J., Milsted, Andrew, Kent, Sarah, Essex, Jon W. and Richards, G.W. (2006) A computer-aided drug discovery system for chemistry teaching. In, American Chemical Society 232 National Meeting, San Francisco, USA, 10-14 Sep 2006. USA, American Chemical Society.

Frey, Jeremy G., Gledhill, Robert, Kent, Sarah, Hudson, Brian and Essex, Jon (2006) Schools Malaria Project. In, Proceedings of the UK e-Science All Hands Meeting 2005. UK, Engineering and Physical Science Research Council. (UK e-Science All Hands Meeting 2005).

Geldhill, Robert, Kent, Sarah, Milsted, Andrew, Chapman, Richard, Essex, Jonathan. W. and Frey, Jeremy. G. (2008) e-Malaria: the schools Malaria project. Concurrency and Computation: Practice & Experience, 20, (3), 225-238. (doi:10.1002/cpe.1193)

Gledhill, Robert, Kent, Sarah, Hudson, Brian, Richards, W. Graham, Essex, Jonathan W. and Frey, Jeremy G. (2006) A computer-aided drug discovery system for chemistry teaching. Journal of Chemical Information and Modeling, 46, (3), 960-970. (doi:10.1021/ci050383q)

Woodgate, Dawn and Fraser DanaÎ S., eScience and Education 2005: A Review

http://www.jisc.ac.uk/media/documents/programmes/eresearch/escienceineducation_study_final.pdf

1.5 nanoCMOS Device, Circuit and System Simulations[33]

Co-Authors:
G. Stewart (*National e-Science Centre, University of Glasgow*)
A. Asenov, C. Millar, D. Reid, G. Roy, S. Roy (*Dept of Electronics and Electrical Engineering, University of Glasgow*)
C. Davenhall (*National e-Science Centre, University of Edinburgh*)
B. Harbulot, M. Jones (*e-Science North West, University of Manchester*)

Key concepts

- Use of 3D simulations to understand statistical fluctuations across devices
- Infrastructure to address security issues
- Results and progress in the management of device variability

NOTE: Please note that there is a related video (p. 90).

[32] http://eMalaria.soton.ac.uk/

[33] This content is available online at <http://cnx.org/content/m32874/1.1/>.

1.5.1 Introduction

The increasing variability present in modern CMOS devices demands revolutionary changes in the design methodology, algorithms and tools used to produce integrated circuits and systems. Progressive scaling of CMOS transistors has driven the success of the global semiconductor industry, as captured by the reknowned Moore's Law which stipulates an exponential increase of the number of transistors in chips, with some modern chips now comprising several billion transistors. Until recently the transistors in silicon chips were assumed to be of uniform design, having identical physical properties and characteristics. However as illustrated in Figure 1, as transistor dimensions approach the nanometer scale this assumption no longer holds true with microscopic differences in atomic structure, doping configuration and material granularity producing differences in the macroscopic behaviour of individual devices.

As a result Moore's Law as tracked by the International Technology Roadmap for Semiconductors[34] (ITRS) is now reaching the physical, atomistic limits of silicon and radical new approaches are needed that encompass this atomistic variability. To address this problem, it is now widely recognized that a paradigm shift must happen in circuit and system design. Strong links have to be established between system, circuit and fundamental device technology research in order to allow modern integrated circuits to cope with the statistically varying behaviour of individual transistors on a chip. Design methods must evolve to accommodate the increasing statistical variability of transistors and absorb the impact that this can have on circuit and system performance. The nanoCMOS project confronts these engineering challenges. This chapter introduces the project in detail and concludes with a brief look at future work in this area.

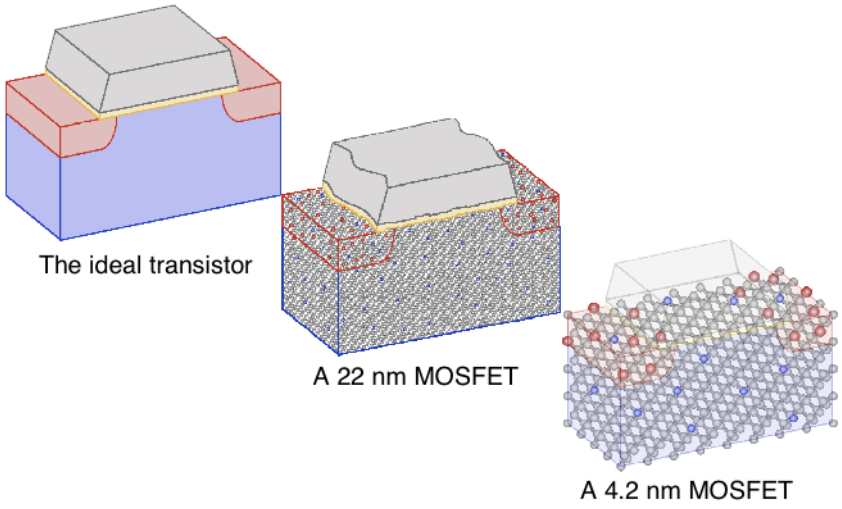


Figure 1.4: Origin of atomic scale variability in nano CMOS transistors

1.5.2 The nanoCMOS Project: Innovative Engineering

Changing design rules for new device architectures and device variability adds significant complexity to the design process, requiring the orchestration of a broad spectrum of tools by geographically distributed teams of device experts, circuit and system designers. In the nanoCMOS project the challenges that this working method presents are being addressed by embedding e-Science technology and know-how in the device

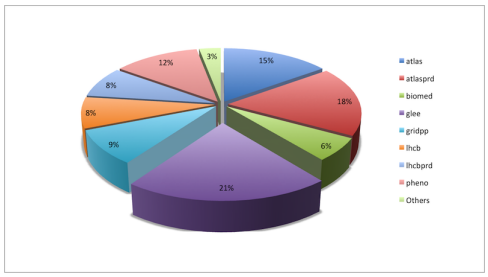
[34] http://www.itrs.net

modeling and design groups and changing the ways in which these disparate groups currently work. The nanoCMOS project[35] was funded for 4 years and started in October 2006. It involves collaboration between world leading device modelling and circuit and system design research groups at the universities of Glasgow, Edinburgh, Manchester, Southampton and York. This academic grouping is enhanced by strong links and collaboration with industrial partners including leading semiconductor, EDA tool vendors and design companies such as Freescale, Fujitsu, National Semiconductor, Synopsis, ARM, Wolfson Microelectronics amongst others. The project will provide valuable insights for industry on the challenges faced in the nanoCMOS domain and how the global semiconductor industry can address them.

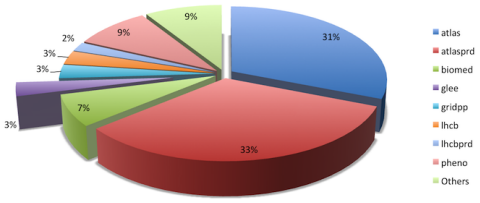
In order to study the statistical fluctuations introduced by the discreteness of charge and matter it is necessary to perform 3D simulations of very large ensembles of hundreds of thousands of devices, rather than a single representative device. Given the increasing number of transistors in modern chips, simulation of very large statistical samples of devices is required to allow statistically rare devices with potentially fatal effects on circuit performance and functionality to be examined. This requires access to significant distributed high performance computing resources, including the UK e-Science National Grid Service[36] , ScotGrid[37] and a wide variety of other resources including Condor pools and campus clusters across partner sites. However, this is not simply another large scale simulation problem, since the commercially sensitive nature of the information and stringent IP protection requirements necessitate fine grained security on access to, and usage of, licensed software; protection of the intellectual property associated with circuit and device designs, data and simulations belonging to industrial partners and key stakeholders.

To this end, the project has developed an infrastructure capable of providing comprehensive security. This includes exploitation of Kerberos for secure global file based access through the Andrews File System; authorization technologies such as PERMIS for definition and enforcement of access policies using centralized attribute authorities such as the Virtual Organisation Membership Service (VOMS), and simple user-oriented access to a project portal through the Internet2 Shibboleth technology using the UK Access Management Federation. Furthermore, the project has identified that a key challenge is in data annotation and management. The simulations that are undertaken can generate large quantities of data and meta-data and the electronics domain unlike other domains does not have agreed standards on data format, rather, the data formats tend to be driven by commercial tool providers.

[35] http://www.nanocmos.ac.uk
[36] http://www.ngs.ac.uk
[37] http://www.scotgrid.ac.uk



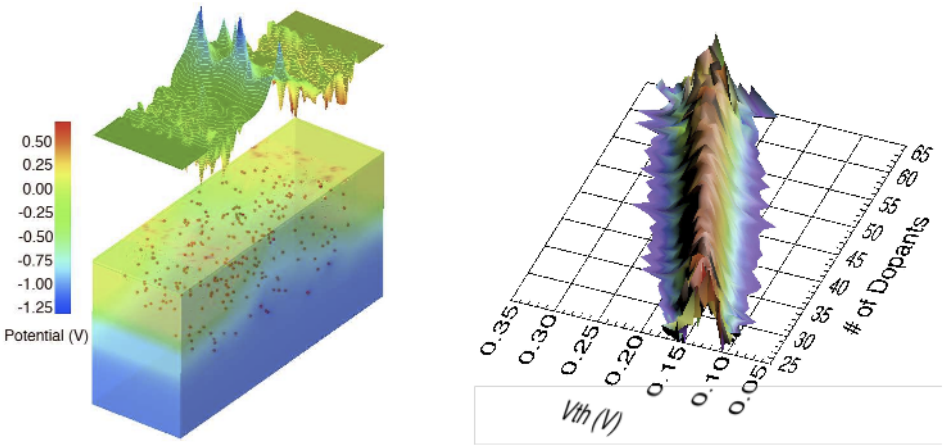
(a)



(b)

Figure 1.5: nanoCMOS Resource Usage on ScotGrid (a) Total CPU time per VO between 16.11.2006 and 19.3.2009 (b) Total number of submitted jobs per VO between 16.11.2006 and 19.3.2009

Traditionally, due to the computational complexity of 3D device simulation, studies of variability have been based on small ensembles of devices typically up to 200 devices and simulations on much larger scale have hitherto never been undertaken. Simulations of ensembles of up to 100,000+ devices enabled by the grid technology are shedding new light on the impact of atomic structure variation on the behaviour of devices, especially at the extreme limits of device variability. Furthermore, based on these simulations, we have been able to examine the effect of device variability at a simple circuit level and have simulated over 1 million CMOS inverters using random configurations of devices. Figure 3(a) shows the potential and dopant position of a statistically rare device. Figure 3(b) shows the threshold voltage variation as a function of the number of dopants.



(a) (b)

Figure 1.6: (a) Potential/Dopant Distribution for Statistically Rare Device (b) Threshold Voltage Variation as Function of the Number of Dopant atoms in the transistor

1.5.3 Summary and Further Workd

The nanoCMOS project has shown the value of e-Research methods in the field of microprocessor circuit design. Use of distributed high performance computing and other dispersed computing resources has allowed for large scale simulations that assist the CMOS design community in managing device variability. The work on nanoCMOS is still progressing and higher level circuit and system design tools are being incorporated into the e-Infrastructure. The systems are being extended in numerous other ways including seamless access to multiple-HPC facilities depending upon user privileges. Optimisation of job submission and management based upon data distribution and security constraints is another area that is currently being investigated. More information on the nanoCMOS project is available at www.nanocmos.ac.uk[38] , or through contacting Prof. Asenov (a.asenov@elec.gla.ac.uk - science related questions) or Prof. Sinnott (r.sinnott@nesc.gla.ac.uk - e-Infrastructure related questions).

1.5.4 Acknowledgements

This work was funded by a grant from the UK Engineering and Physical Sciences Research Council. We gratefully acknowledge their support.

1.5.5 References

Sinnott, R.O. *et al.* (2006). Meeting the Design Challenges of nanoCMOS Electronics: An Introduction to an EPSRC Pilot Project. *UK e-Science All Hands Meeting*, Nottingham UK, September.

Reid, D. *et al.* (2008). Prediction of Random Dopant Induced Threshold Voltage Fluctuations in NanoCMOS Transistors *International Conference on Simulation of Semiconductor Processes and Devices*, Sept.

[38] http://www.nanocmos.ac.uk

1.6 Computational Chemistry[39]

Key Concepts

- Simulation of Biomolecules

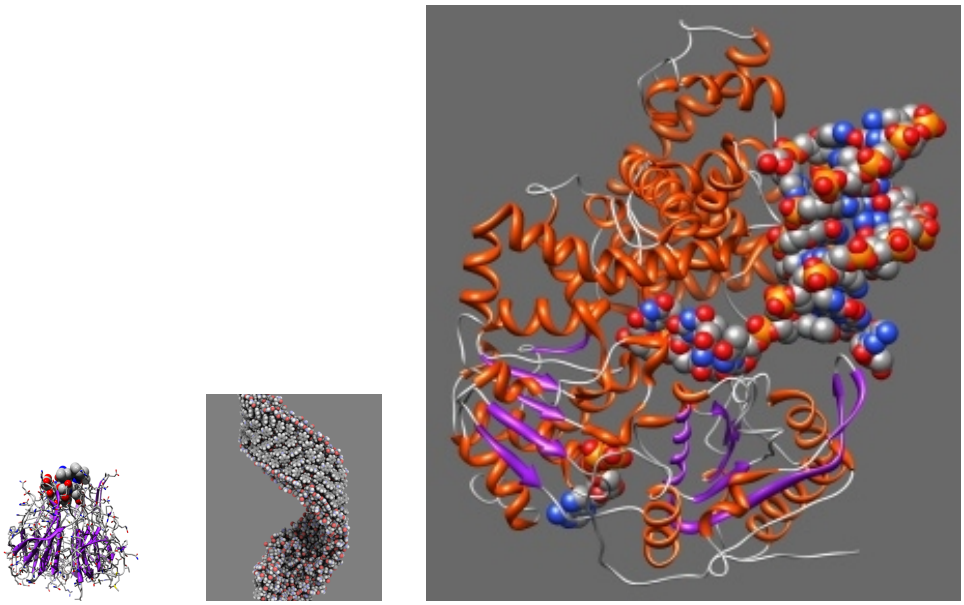
1.6.1 Introduction

In many of the physical sciences, our theoretical understanding has developed alongside experimental discoveries, for example in the fields of electromagnetism, optics and semiconductor physics. Theory has provided the design principles which have then enabled engineers to maximise the potential applications of these new technologies. However, there are many instances in which simple phenomenological models cannot capture the complexity of the systems in question: notable examples are the chemistry of the atmosphere, which has implications for prediction of weather patterns and climate change, or the properties of materials at the nanoscale, such as chemically functionalised carbon nanotubes, which will become increasingly important in nanoengineering.

Arguably, the most complex materials of all are biological macromolecules; namely proteins, DNA, lipids, sugars and their interactions. Biological macromolecules routinely perform extraordinary functions such as biomolecular recognition (Figure 1a), enzyme catalysis, self-assembly (Figure 1b) and self-organisation. Moreover, there are many examples of molecular motors within the cell (Figure 1c). These are nanoscale machines capable of burning chemical energy to perform work. The theoretical challenge of understanding these systems is more than offset by the potential benefits. For example, our current understanding of molecular recognition has already enabled us to rationally design new drugs *in silico* to inhibit or promote a given biomolecular interaction; in the future nano-computer aided design may be used to design our own molecular devices of equivalent complexity to biological molecular motors, but which perform a bespoke function.

If we had an equivalent theoretical understanding of biological systems as we have of semiconductors, then whole new regimes of bio-inspired engineering at the nanoscale would become possible. To achieve this, we need to combine our existing physical understanding of mechanics and thermodynamics with a theoretical technique that is capable of including chemical complexity. The only suitable methodology is High Performance Supercomputing (HPC).

[39] This content is available online at <http://cnx.org/content/m32928/1.1/>.



(a) (b) (c)

Figure 1.7: Atomistic Molecular Dynamics Simulation: Keap1 Protein

The most successful biomolecular simulation methods to date use Newtonian mechanics in conjunction with an empirical force-field to produce a mathematical model of the interactions between every single atom in the macromolecule with chemical accuracy; the calculation results in a series of molecular conformations (or a "movie") that illustrates the changing shape of the biomolecule due to thermal fluctuations.

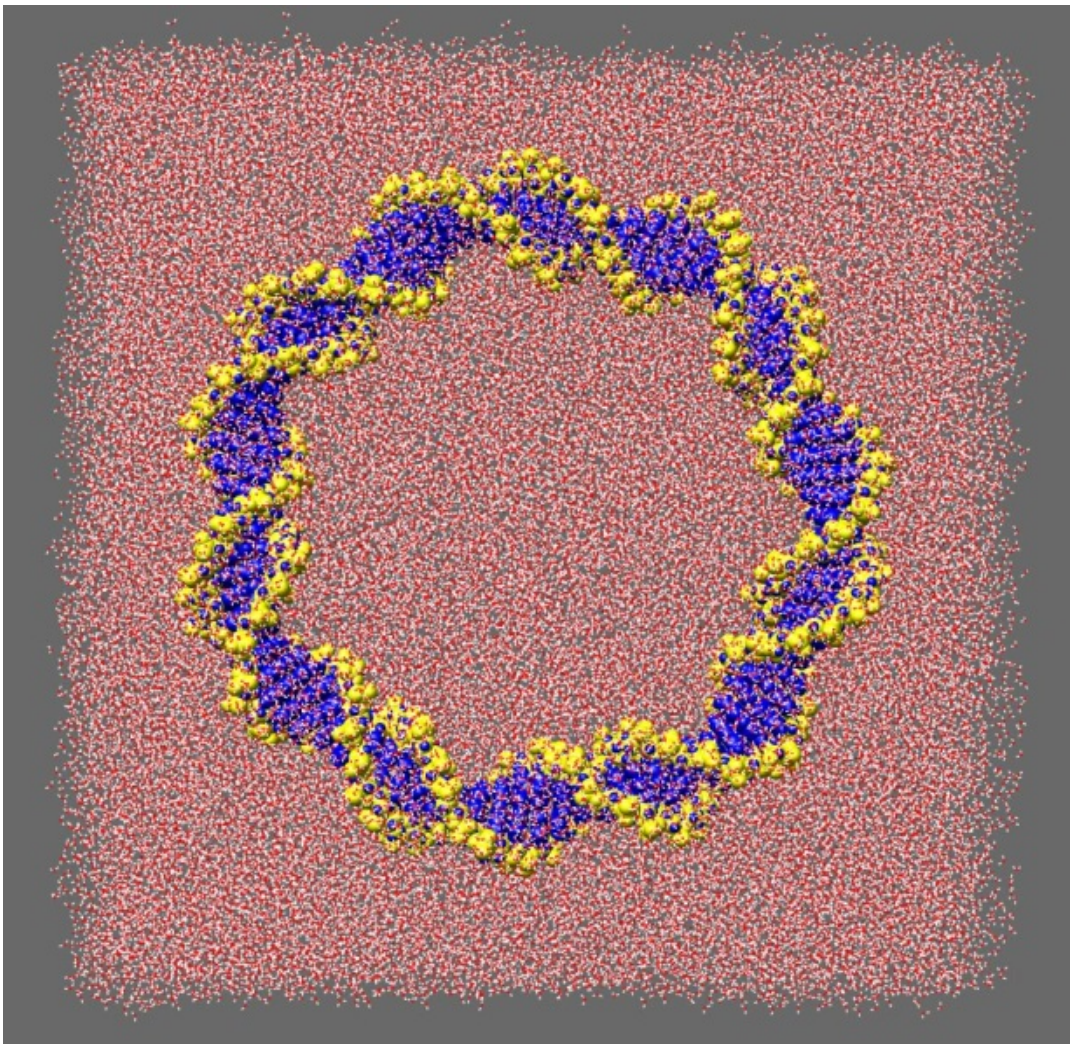


Figure 1.8: An atomistic model of a 90 base pair DNA nano-circle, showing the presence of explicit water molecules [1].

This technique is known as atomistic molecular dynamics (MD) simulation. Biomolecules are naturally highly responsive materials, as is required by their function. Consequently, the most accurate simulations of biological macromolecules must also include a description of the solvent environment (see Figure 2), which usually consists of water and counterions. Typically, such a calculation will contain ∼150,000 atoms, and will require over 750 CPU hrs to obtain a 1ns MD trajectory using the AMBER suite of MD programs.

An example of the use of atomistic simulation to develop new medicines is shown in Figure 1a. The protein Keap1 is responsible for inactivating protective anti-cancer genes, which the cell only uses under conditions of environmental or chemical stress [2]. It achieves this function by recognising components of the cellular machinery that ultimately detect and repair genetic damage, and tagging them for destruction. By designing synthetic molecules that block the action of Keap1, the aim is to enhance the natural ability of the cell to protect itself from cancerous agents, which could be beneficial to those facing a particularly

high risk of developing cancer. Atomistic MD simulations can provide detailed chemical information about the specific inter-atomic interactions that drive molecular recognition between the Keap1 protein and its natural target. Understanding which of these interactions is most important will lead to the virtual design of synthetic molecules which mimic the natural substrate and block it binding.

All molecular recognition is ultimately driven by thermodynamics; the complex is formed because this lowers the overall free energy of the system. The protein recognises its target through shape specific favourable energetic interactions, such as van der Waals, electrostatic and hydrogen bonding interactions, all of which are encoded in the computer simulation of the complex. However, since the protein is a soft nanoscale object, the complex is also highly flexible, and it is important to include dynamics (specifically the entropy) to calculate the correct free energy change. The interaction energy between the protein and its complex is relatively straightforward to calculate using computational models. However, methods for calculating the entropic contribution are still under development [3]. The computational expense of the calculations limits MD timescales to ∼100ns for a small protein. Since a molecule of this size executes very slow conformational changes that take place on a longer timescale than we are able to simulate, the full contribution from the entropy can be difficult to calculate accurately. As supercomputer resources continue to expand, and as simulation codes become more efficient, our ability to quantify dynamic changes will improve the virtual design of synthetic molecules which intervene in biological processes.

1.6.2 Amyloid Fibril Simulation

In addition to molecular recognition, many biological molecules are also able to self-assemble in a remarkably specific manner. Microtubules, which are large cellular scaffolds formed from the polymerisation of many tubulin protein monomers, generate forces which drive cell motility by assembling and dissembling in a switchable manner. Amyloid fibrils are long fibrous structures that form when many copies of the same protein self-assemble through hydrogen bonding interactions along the peptide backbone. Although amyloid has been shown to have advantageous functions within cells, most remarkably amyloid has been shown to play a role in inheritance in yeast [4], amyloid fibrils are best known for their involvement in human diseases such as Alzheimer's and Parkinson's, and the transmissible (prion) diseases such as Creutzfeldt-Jakob disease [5]. Synthetic self-assembling peptides (see Figure 1b) have enormous potential as new nanomaterials, and have been tested for applications in molecular electronics, drug delivery and tissue engineering [6]. However, these systems are generally so structurally irregular that traditional experimental methods for investigating protein structure and function, such as X-ray crystallography, cannot be used to study them.

Computer simulations of amyloid-like fibrils can be used to generate model structures of small peptide aggregates at the atomic level, and are able to provide information about the interaction energies that drive the self-assembly. As well as providing new insight into the thermodynamics of disease, these methods can also be used in the design of synthetic self-assembling systems with engineered material properties. Nevertheless, there are many aspects of fibril growth that cannot currently be probed by atomistic simulation due to the length and timescales involved. Amyloid formation is a nucleated process, which implies that there is a "lag phase" due to a kinetic barrier to aggregation. It is therefore a slow process, and it is not computationally possible to study the kinetics of fibril formation in full atomic detail. Since the likelihood of developing Alzheimer's, for example, may well depend on fibril growth rates, it is hoped that in the future there will be sufficient computational resources to investigate the sensitivity of amyloid formation to external factors such as cell pH, or the presence of metal ions such as copper and aluminium.

1.6.3 Future Directions for HPC in Computational Chemistry

The ability of biological macromolecules to function as nanoscale machines is particularly remarkable. DNA helicases can be some of the smallest molecular motors (see Figure 1c). They perform the essential function of separating double stranded DNA to provide access to the genetic code by burning the chemical energy provided by ATP hydrolysis. The catalysis step occurs over femtosecond timescales, and since it involves the breaking and formation of chemical bonds it requires quantum mechanical (QM) simulation methods, which are immensely computationally expensive. This chemical reaction is coupled to the large conformational

changes that enable the motor to function, which take place over millisecond timescales, resulting in a machine that moves along DNA at a rate of around 50 DNA bases per second. Currently, only multi-scale methods which integrate models operating at different length and time-scales (in this case QM/ atomistic MD, and stochastic modelling) can be used to study even the smallest molecular motors. An exciting future prospect for high performance supercomputing is the expectation that it will be possible to study the entire thermodynamic cycle of such a nanoscale machine with sufficient accuracy to fully understand its mechanism.

Acknowledgements: Many thanks to Geoff Wells and Binbin Liu for providing Figures 1a and 1b respectively.

References

[1] Harris S. A., Laughton C. A. & Liverpool T. B. "Mapping the phase diagram of the writhe of DNA nanocircles using atomistic molecular dynamics simulations" (2008) *Nucleic. Acids. Res.* 36, 21-29.

[2] Lee J. & Surh Y. "Nrf2 as a novel molecular target for chemoprevention" (2005) *Cancer Lett.* 171-184.

[3] Harris S. A. & Laughton C. A. "A simple physical description of DNA dynamics: quasi-harmonic analysis as a route to the configurational entropy" (2007) *J. Phys.:Condens. Matter.* 19, Art. No. 076103.

[4] Fowler D. M., Koulov A. V., Balch W. E. & Kelly J. W. "Functional amyloid – from bacterial to humans" (2007) *Trends in Biochem. Sci.* 32, 217-223.

[5] Chiti F. & Dobson C. M. "Protein misfolding, functional amyloid and human disease" (2006) *Annu. Rev. Biochem.* 75, 333-366.

[6] Cherny I. & Gazit E. "Amyoids: Not only pathological agents but also ordered nanomaterials" (2008) *Angew. Chem. Int. Ed.*, 47, 4062-4069.

1.7 Biomedical Research[40]

Key Concepts

- biomedical research
- drug discovery
- docking
- telemedicine
- radiotherapy
- epidemiology

1.7.1 Introduction

Biosciences research provides an exemplar of the dramatic transformation occurring within the context of data rich science. Increased emphasis on biology as an "informational science" focused on genomics has resulted in the production of huge data sets that can only be adequately managed and decoded by using advanced information technologies (ICTs). Researchers studying biology at higher levels of organisation than the genome also rely on ICTs to develop models of cells, tissues, organisms and ecologies, in order to come to terms with complexity. Biomedical science aided by ICTs has made significant advances in areas such as the understanding of disease processes (for instance in heart or cancer modelling) and drug discovery.

Biomedicine consists of various aspects which can benefit from a grid-based approach including the search for new drug targets into the genome and the proteome, identification of single nucleotide polymorphisms (SNPs) relating to drug sensitivity, drug resistance mechanism elucidations as well as epidemiological monitoring of disease outbreaks. Well-identified areas of relevance of the grid paradigm are epidemiology and computer-intensive analysis of geographically distributed medical images. Grids are defined as fully distributed, dynamically reconfigurable, scalable and autonomous infrastructures to provide location independent, pervasive, reliable, secure and efficient access to a coordinated set of services encapsulating and virtualising resource. Their relevance for biomedical research has been investigated within the framework of

[40] This content is available online at <http://cnx.org/content/m32938/1.1/>.

the HealthGrid initiative (Breton et al. 2005, SHARE Project 2008). Here we focus on the use of e-Research methods in the study of infectious diseases such as flu viruses and malaria and Medical Data Management.

1.7.2 Grid as a surveillance tool for diseases

Epidemiology focused on population-level research requires access to distributed, critically sensitive and heterogeneous data, resulting in overall costly computing processes. The study of flu viruses and their treatment is one notable example of e-Research in biomedical sciences. Recent years have seen the emergence of diseases which have spread very quickly around the world, either through human travel, like SARS and SIV (H1N1), or animal migration, like avian flu (H5N1). Swine Flu has been in the headlines in 2009, officially classified as a "pandemic" by the World Health Organization in response to the virus's worldwide geographic spread (Neumann, Noda and Kawaoka 2009). International collaboration has involved use of grid computing to model potential circumstances surrounding such extreme outbreaks.

Among the biggest challenges from emerging infectious diseases is the relation between early detection and surveillance of the diseases, as new cases can appear anywhere. This results from the globalization of exchanges and the circulation of people and animals around the world as recently demonstrated by the avian flu epidemics. An international collaboration of research teams in Europe and Asia has been exploring some innovative in silico approaches to better tackle avian flu and more recently swine flu, taking advantage of the very large computing resources available on international grid infrastructures (Brenton *et al.* 2008). Existing data sources have been integrated to form a global surveillance network for molecular epidemiology, based on Service Oriented Architecture (SOA) and grid technologies. The idea is to dynamically analyze the molecular biology data, made available on public databases using computing, storage and automatic updating services offered by grid technology. Bioinformatics methods of sequence alignment can highlight mutations on a virus's genome that could impact transmission mechanisms, pathogenicity or drug sensibility. In addition, phylogenetic analyses help to characterize evolutionary history, a key point in understanding the geographic and molecular source of this outbreak, when the virus seems to be a reassortant from avian and human forms. If another pandemic strikes, bioinformatics is expected to have an impact by adding a new weapon to researchers' arsenal: the grid.

1.7.3 Grid as a discovery tool for new drugs - WISDOM

Another challenge for infectious disease research is the constant mutation of the viruses. The mutation of these viruses makes them perpetual moving targets for drug and vaccine discovery. In this context, the WISDOM (World-wide In Silico Docking On Malaria) collaboration, comprised of experts on all continents, was launched in 2005 to exploit the resources of grid infrastructures for in silico drug discovery (Chien, Foster and Goddette 2002). Virtual screening is a computational technique used in drug discovery research. It involves the rapid in silico assessment of large libraries of chemical structures in order to identify those structures most likely to bind to a drug target, typically a protein receptor or enzyme. Although virtual high throughput screening (HTS) is mainly achieved through clusters of computers physically connected to one another that can screen compound sets against the target, recent advances in grid technology are allowing powerful grid-computing strategies to be applied to HTS to enlist a larger number of compute resources. Discovering hits with the potential to become usable drugs is a critical first step to ensure a sustainable global pipeline for finding innovative products to treat neglected and emerging disease (Richards 2002).

The primary goal of the WISDOM initiative was to support research on diseases that do not receive sufficient attention from the research community and pharmaceutical laboratories for the development of new drugs and vaccines, despite critical situations and the efforts of international agencies and foundations. Among neglected diseases, malaria causes more than one million deaths every year mostly in tropic and sub-tropic regions. Dengue fever which is also a viral disease transmitted by mosquitoes share common geographic areas with malaria with additional prevalence in urban areas of the tropics. In the meanwhile, tuberculosis has reemerged as a major threat to international public health in recent years due to its correlation with AIDS. Four years after its launch, WISDOM has been able to successfully screen a handful of biological targets involved in major societal threats like avian flu, diabetes and malaria, using very large databases

of drug-like compounds. A few hundred compounds selected in silico were tested in vitro and about 20% have demonstrated significant inhibition activity against the target of interest, showing the relevance of grid technology to address drug discovery issues.

1.7.4 Grid Technology for Distributed Medical Data Management

Providing patients with "google-like" secure access to their medical records requires the information to be available for querying and retrieval. Google is able to query and search for any data published on the Internet. However, it will be absolutely necessary to ensure the security of this Internet environment before storing any medical data on it. An alternative is provided by grid technology which allows distributed data to be queried securely according to personal access rights. Some platforms in medical data management (Erberich *et al.* 2007) of paediatric data (Freud *et al.* 2007) or medical radiography data (Warren *et al.* 2007) already benefit from grid technologies to manage medical data securely thanks to dedicated grid middleware services such as MDM8 or Globus Medicus. The use of grids overcomes the difficulties inherent in a centralized storage system, especially high cost and complexity. Grids also make it possible to store data where or very close to where they are produced. Through grid authentication, authorization and accounting, only duly authorized persons can gain access to data which are encrypted and made anonymous when they are transmitted (Mohammed *et al.* 2007).

Early attempts at epidemiological applications of grids (Blanquer and Hernández 2005) have demonstrated their relevance for patient customized research. Users ought to be able to take it for granted that the security mechanisms are sufficient to protect their data; that the results of their research will be private and available to third parties only if designated; that the system will meet the concerns of the ethical and legal committees of their research institutions; that the services are reliable, efficient and permanent; that they do not have to change significantly their current procedures; protocols or workflow, and finally that the data is somehow automatically organised and gathered, and thus available for further exploitation. In the next chapter, we will present an epidemiological application of grids for cancer surveillance which is currently being used in France. Another attractive field of application for grid technology is computer-intensive analysis of distributed medical images. The impact of grid technology comes from the secure management of distributed images together with the capacity to gain access to large computing resources on demand to analyze them. In the field of oncology, the use of Computer-Aided Detection (CAD) for the analysis of mammograms was addressed by the MammoGrid project as early as 20059. Other efforts focus on using grid computing resources to plan radiotherapy treatment (Benkner *et al.* 2001) a case of the use of this technology currently exploited in collaboration with a French Cancer Treatment Centre will be further documented in the case study 2.

1.7.5 Case Study 1 - Cancer surveillance network

Cancer screening programs aim at the early detection of the malignant tumors in order to improve the prognosis. Most EU countries have launched a national program for breast cancer screening (von Karsa et al. 2007). In France, when a woman is positively diagnosed with a risk of tumour, cancer associations are responsible for providing a second diagnosis on the mammograms and have to follow-up the pathology data about the tumour, which are stored by the laboratories. At present, the patient's data are faxed on request or carried physically by the patient to the associations where they are recorded again. This process is costly and error prone as data has to be typed and reinterpreted twice. The cytopathology data are also relevant for epidemiological analysis. The INVS (Sanitary Surveillance Institute), the French equivalent of the (E)CDC in the USA, is in charge of publishing indicators about global health and particularly about cancer. To produce its indicators, the INVS relies on regional cancer registries (CRISAPs) set up to collect relevant information to support statistical and epidemiological studies about cancer incidence, mortality, prevalence or screening. CRISAPs (Centre de Regroupement Informatique et Statistique en Anatomie et cytologie Pathologiques) are like regional data warehouses collecting anonymous data from pathology laboratories or from healthcare establishments involved in cancer treatment.

Healthcare professionals in laboratories are reluctant to release data because of cost and also because they lose some control over the data they have produced. An alternative is for clients to query databases of the pathology laboratories. A grid, federating the laboratories, would provide a secure framework enabling the screening associations to query databases and fill their local patient files (De Vlieger *et al.* 2009). No action is required by physicians to put their data on the network. Thanks to the grid security architecture, the cytopathologists are able to define and modify the access rights of the users querying their data.

Several projects in Europe have studied or are currently exploring the advantages of grid technology with regard to breast cancer, particularly computer-aided diagnosis of mammograms, most notably the e-Diamond (Brady *et al. 2003*) and MammoGrid (Warren *et al.* 2007) projects. If a sentinel network is able to federate pathology databases, it can be used by the epidemiological services of the National Institute for Health Surveillance (Institut National de Veille Sanitaire) and the regional epidemiological observatory. In the present case, it means that women could consult their own data in the pathology laboratories as well as see mammographic images stored in the radiology services through the proposed network. A cancer surveillance network is presently being implemented in the Auvergne region in France within the framework of the AuverGrid regional grid initiative (http://www.auvergrid.fr). It uses grid technology developed by EGEE, such as the AMGA metadata catalogue (Koblitz, Santos and Pose 2008) and the MDM Medical Data Manager (Montagnat *et al.* 2006), as well as by the Health-e-Child project, for example, the Pandora Gateway (http://www.health-e-child.org).

1.7.6 Case Study 2 - Application in radiotherapy

Radiotherapy is one of the three major treatments for cancer. It has demonstrated its efficacy in curing cancer and is also the most cost effective strategy. From a technology point of view, radiotherapy is a highly complex procedure, involving many computational operations for data gathering, processing and control. The treatment process requires large amounts of data from different sources that vary in nature (physics, mathematics, biostatistics, biology and medicine), which makes it an ideal candidate for healthgrid applications. Nowadays, in radiotherapy and brachytherapy, commercial treatment planning systems (TPS) use an analytical calculation to determine dose distributions near the tumor and organs at risk. Such codes are very fast (execution time below one minute to give the dose distribution of a treatment), which makes them suitable for use in medical centres.

For some specific treatments using very thin pencil beams (IMRT) and/or in the presence of heterogeneous tissues such as the air-tissue, lung-tissue and bonetissue interfaces, it appears that Monte Carlo simulations are the best way to compute complex cancer treatment by keeping errors in the dose calculation below 2%. The accuracy of Monte Carlo (MC) dose computation is excellent, provided that the computing power is sufficient to allow for extreme reduction of statistical noise. In order to finish MC computations within an acceptable time period for interactive use, parallel computing over very many CPUs has to be available. In this way, MC dose computations could become standard for radiotherapy quality assurance, planning and plan optimisation years before individual departments could afford local investment that is able to support MC. With the objective of making Monte Carlo dose computations the standard method for radiotherapy quality assurance, planning and plan optimisation, we are participating in the development of a Monte Carlo platform dedicated to SPECT, TEP, radiotherapy and brachytherapy simulations together with 21 other research laboratories which are involved in the international collaboration OpenGATE (http://www.opengatecollaboration.org, Jan *et al.* 2004). This GATE software with its accuracy and flexibility was made available to the public in 2004 and now has a community of over 1000 users worldwide.

The limiting issue of GATE right now is its time consuming simulations for modeling realistic scans or treatment planning. A secured web platform enabling medical physicists and physicians to use grid technology to compute treatment planning using GATE Monte Carlo simulations and share medical data has been developed. This platform, the Hospital Platform for E-health (HOPE, Diarena *et al.* 2008) provides quick, secure and easy to use tools to physicians or medical physicists to perform treatment planning on the Grid infrastructure. When the user is logged in, he/she has the possibility to upload or access medical data located on the hospital's PACS (Picture Archiving and Communication System) server In the case of medical imaging for radiotherapy, the metadata server (AMGA) services located at the hospital collect metadata as

attributes like the name of the patient, the characteristics of the disease, etc. SSL (Secure Socket Layer) connections in addition to encryption systems are used for the transfer of data. Authentication using ACLs (Access Control Lists) are used for the access to metadata in the database. The metadata server provides a replication layer which makes databases locally available to user jobs and replicates the changes between the different participating databases. Information contained in electronic patient sheets is also registered as parameters in the metadata server. The anonymized medical images are registered on the grid. GridFTP (File Transfer Protocol) is used to enable advanced security transfers. Medical images are associated with patient sheets and the user can automatically visualize them.

By visualizing the tumour, the physician can choose what kind of device is the most appropriate to treat the patient using ionizing particles (field size, type of particle, energy, brachytherapy sources, ...). The treatment plans can be directly visualized from the HOPE portal and downloaded onto the personal computer of the user. The web portal offers to the user a transparent and secure way to create, submit and manage GATE simulations using realistic scans in a grid environment. The conviviality of the web portal and the Grid performances could make it possible, in the near future, to use Monte Carlo simulations from clinical centres or hospitals to treat patients in routine clinical practice for specific radiotherapy treatments. In addition, the web platform functionalities enable direct access to medical data (patient sheets, images...) and secure sharing between two users located in different hospitals.

1.7.7 References

Benkner S., Berti G., Engelbrecht G. *et al.* (2001). GEMSS: Grid-infrastructure for Medical Service Provision. *Methods of Information in Medicine*, 44(2). pp. 177-181.

Blanquer, I. and Hernández, V. (2005). The Grid as a Healthcare Provision Tool. *Methods of Information in Medicine* vol. 44. pp. 144–148

Brady, M. *et al.* (2003) eDiamond: a grid-enabled federated database of annotated mammograms. Berman, F., Fox, G. and Hey, T. (eds.) *Grid Computing: Making the Global Infrastructure a Reality*, Wiley.

Breton, V., Da Costa, A.L., De Vlieger, P., Maigne, L., Sarramia, D., Kim, Y-M, Kim, D.,Nguyen, H. Q., Solomonides, T. and Wu Y-T. (2009). Innovative In Silico Approaches to Address Avian Flu Using Grid Technology. *Infectious Disorders - Drug Targets* 9(3), June. pp. 358-365.

Breton V, Dean K, Solomonides T, editors on behalf of the Healthgrid White Paper collaboration (2005). The Healthgrid White Paper. *Proceedings of the Healthgrid conference*. Studies in Health Technology and Informatics, IOS Press, vol. 112. pp. 249–321.

Chien, A., Foster, I., Goddette, D. (2002). Grid technologies empowering drug discovery. *Drug Discovery Today*, 7 Suppl 20. pp. 176-180

Diarena, M. *et al.* (2008). HOPE, an open platform for medical data management on the grid. *Proceedings of HealthGrid*. Studies in Health Technology and Informatics, vol. 138. pp. 34–48.

Erberich, S.G., Silverstein, J.C., Chervenak, A., Schuler, R., Nelson, M.D. and Kesselman, C. (2007). Globus MEDICUS: federation of DICOM medical imaging devices into healthcare Grids. *Proceedings of HealthGrid*. Studies in Health Technology and Informatics, vol 126. pp. 269–278.

Freund J, Comaniciu D, Ioannis Y, *et al.* (2007). Health-e-child: an integrated biomedical platform for grid-based paediatrics. *Proceedings of Healthgrid*. Studies in Health Technology and Informatics, IOS Press, vol. 120. pp. 259–70.

Jan, S. *et al.* (2004). GATE: a simulation toolkit for PET and SPECT. *Physics in Medicine and Biology*, 49(19). pp. 4543–61.

von Karsa, L. *et al.* (2007) Cancer Screening in the European Union; Report on the implementation of the Council Recommendation on cancer screening.

Koblitz B., Santos, N. and Pose, V. (2008). The AMGA Metadata Service. *Journal of Grid Computing*. vol. 6. pp.61–76.

Mohammed, Y., Sax, U., Viezens, F. and Rienhoff, O. (2007). Shortcomings of current grid middlewares regarding privacy in HealthGrids. *Proceedings of HealthGrid*. Studies in Health Technology and Informatics. vol. 126. pp. 322–329.

Montagnat, J., Jouvenot, D., Pera, C., Frohner, A., Kunszt, P., Koblitz, B., Santos, N. and Loomis, C. (2006). Bridging clinical information systems and grid middleware: a Medical Data Manager. *Proceedings of HealthGrid*. Studies in Health Technologies and Informatics, vol. 120. pp. 14-24.

Neumann, G., Noda, T. and Kawaoka, Y. (2009). Emergence and pandemic potential of swine-origin H1N1 influenza virus, *Nature* 459, 931-939, June

Richards, W.G. (2002). Virtual screening using grid computing: the screensaver project. *Nature Reviews Drug Discovery*, vol. 1, pp. 551-555

SHARE Project (2008). *SHARE, the journey: a European HealthGrid roadmap*, printed by the European Commission, Information Society and Media DG, ISBN n° 9789279096686

De Vlieger, P., Boire, J.Y. and Breton, V. *et al.* (2009). Grid-enabled Sentinel Network for Cancer Surveillance. *Proceedings of HealthGrid*. Studies in Health Technologies and Informatics. vol 147

Warren, R., Solomonides, T. and del Frate, C. *et al.* (2007). Mammogrid: A Prototype Distributed Mammographic Database for Europe. *Clinical Radiology*, June, 62(11). pp. 1044–51

Chapter 2

Distributed Systems

2.1 The European e-Infrastructure Ecosystem[1]

Key Concepts

- Managed e-Infrastructures
- Different needs for general-purpose and specialised e-Infrastructures
- Layers of e-Infrastructures and the role of standards
- Academic and commercial e-Infrastructures
- Convergence of e-Infrastructures

2.1.1 Introduction

e-Research requires seamless access to computational, storage, and network resources, which can be provided by a variety of different means ranging from volunteer systems, to community based infrastructures, to general-purpose infrastructures federating resources across different institutions. These resources are made available to different scientific communities via well-defined protocols and interfaces exposed by a software layer (**Grid middleware**). Such federated infrastructures are referred to as **e-Infrastructures** and provide a number of advantages to researchers and service providers alike.

Of these different approaches, well managed e-Infrastructures are of particular importance. Apart from enabling seamless access to heterogeneous, independently managed resources, well managed e-Infrastructures also provide their users with common operational procedures such as accounting, support and support systems, and usage policies (etc). Moreover, different service levels can be negotiated, allowing the user to establish service level agreements with such an e-Infrastructure. As a consequence, researchers experience the usage of an e-Infrastructure in the same way as using a single system managed by a local resource provider. It is important to realise that in order to achieve this deployment of standardised services is needed as well as a harmonization of operational and security procedures across the different independent resource providers. Multi-purpose e-Infrastructures are also desirable from a resource provider point of view, as a single infrastructure can serve several communities and thus reduce the need for dedicated community services that require additional operational effort.

However, it is unlikely that a single common infrastructure will eventually be able to serve **all** needs as different legislative regulations, usage models, and other regional or thematic peculiarities demand the creation of separate e-Infrastructures. As a consequence, national and regional e-Infrastructures as well as thematic ones like e-Infrastructures focusing on the federation of supercomputing resources have emerged. Europe, through ambitious national research and infrastructure programs and dedicated European Commission programs, is playing a leading role in building multi-national, multi-disciplinary e-Infrastructures

[1] This content is available online at <http://cnx.org/content/m31215/1.1/>.

and has devised a roadmap for a pan-European e-Infrastructure. This road-map acknowledges the need for different infrastructures but also envisages these infrastructures embedded in an ecosystem that allows users to easily access resources managed by different infrastructures.

2.1.2 Two Ecosystem Paths – the EGI and PRACE Infrastructures

The establishment of a European e-Infrastructure ecosystem is currently progressing along two distinct paths: the EGI and PRACE. The European Grid Initiative (EGI) intends to federate national and regional e-Infrastructures, managed locally by National Grid Initiatives (NGIs) into a pan-European, general-purpose e-Infrastructure as pioneered by the EGEE (Enabling Grids for E-sciencE) project that unites thematic, national and regional Grid initiatives. EGI is a direct result of the European e-Infrastructure Reflection Group (e-IRG) recommendation to develop a sustainable base for European e-Infrastructures. Most importantly, funding schemes are being changed from short term project funding (like 2 years funding periods in the case of EGEE) to sustained funding on a national basis. This provides researchers with the long-term perspective needed for multi-year research engagements. All European countries support the EGI vision and at the time of writing the organisational and legal details are being defined with the aim of starting EGI in 2010.

Unlike EGEE, which has strong central control, EGI will consist of largely autonomous NGIs with a lightweight coordination entity on the European level. This setup asks for an increased usage of standardised services and operational procedures to enable a smooth integration of different NGIs exposing a common layer to the user while preserving their own autonomy. In this context the work of the Open Grid Forum (OGF) is of particular importance. OGF aims at standardising services for e-Infrastructures and increasingly tackles operational issues.

At the same time, the e-IRG has recognised the need to provide European researchers with access to petaflop-range supercomputers in addition to high-throughput resources that are prevailing in EGI. The Partnership for Advanced Computing in Europe (PRACE) aims to define the legal and organisational structures for a pan-European High Performance Computing (HPC) service in the petaflops range. These petaflop range systems are supposed to complement the existing European HPC e-Infrastructure as pioneered by the Distributed European Infrastructure for Supercomputing Applications (DEISA) project. DEISA is federating major European HPC centres in a common e-Infrastructure providing seamless access to supercomputing resources and, thanks to a global shared file system, data stored at the various centres. This leads to a three tier structure, with the European petaflops systems at tier 0 being supported by leading national systems at tier 1. Regional and midrange systems complement this HPC pyramid at tier-2 as depicted in Figure 1 below.

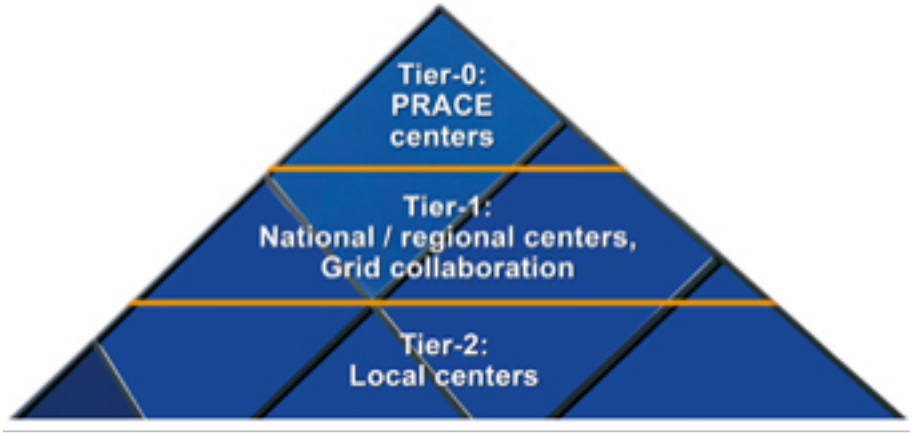


Figure 2.1: The PRACE HPC Ecosystem

Although the goals of PRACE/DEISA are similar to the ones of EGI/EGEE, the different usage and organisational requirements demand a different approach and hence the establishment of two independent, yet related infrastructures. For researchers it is important, however, to have access to all infrastructures in a seamless manner, hence a convergence of the services and operational models in a similar way as discussed in the EGI/NGI case above will be needed.

2.1.3 Layers of the Ecosystem

This convergence and the addition of other tools (like sensors, for instance) will eventually build the computing and data layer of the e-Infrastructure ecosystem. Leveraging the physically wide area connectivity provided by the network infrastructure (operated by GÈANT and the National Research and Education Networks in Europe) this computing and data layer facilitates the construction of domain specific knowledge layers that provide user communities with higher level abstractions, allowing them to focus on their science rather than the computing technicalities. The resulting three-layered ecosystem is depicted in Figure 2 below.

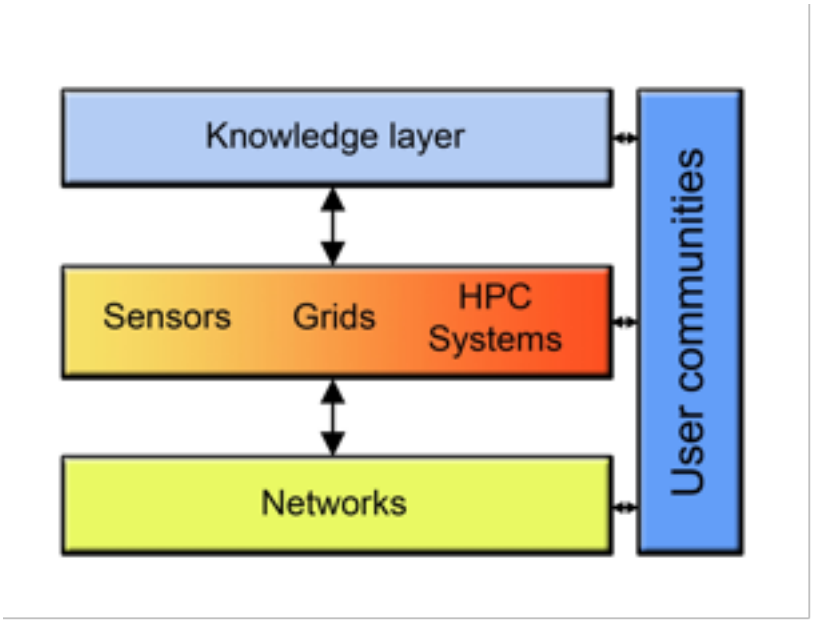


Figure 2.2: e-Infrastructure Ecosystem

Europe is also active in technology transfer to foster the usage of e-Infrastructures in other areas such as South America, India, China, Asia-Pacific and the Mediterranean. These efforts, together with similar efforts in the US, Japan, and Australia, ensure that large parts of the world are covered by e-Infrastructures as shown in Figure 3 below.

Figure 2.3: Worldwide e-Infrastructure Coverage

2.1.4 Clouds in the Ecosystem

In the commercial sector, dynamic resource and service provisioning as well as "pay per use" concepts are being pushed to the next level with the introduction of **"cloud computing"**, successfully pioneered by Amazon with their Elastic Compute Cloud (EC2) and Simple Storage Service (S3) offerings. Many other major IT businesses offer cloud services today, including Google, IBM, and Microsoft. Using virtualisation techniques, these infrastructures allow dynamic service provisioning and give the user the illusion of having access to virtually unlimited resources on demand. This computing model is particularly interesting for start-up ventures with limited IT-resources as well as for the dynamic provisioning of additional resources to cope with peak demands, rather than over-provisioning ones own infrastructure. As of today the usability of these commercial offerings for research remains yet to be shown although a few promising experiments have already been performed. In principle, clouds could be considered yet another resource provider in e-Infrastructures, however, it is currently not trivial to bridge the different interfaces and operational procedures to provide the researcher with a seamless infrastructure. More work on interface standardisation and on how commercial offerings can be made part of the operations of academic research infrastructures is needed.

2.1.5 Summary: The Need for Convergence

In summary, a variety of different e-Infrastructures are available today to support e-Research. Convergence of these infrastructures in terms of interfaces and policies is needed to provide researchers with seamless access to the resources required for her research, independently of how the resource provisioning is actually managed. Eventually, a multi-layer ecosystem will greatly reduce the need for scientists to manage their computing and data infrastructure, with a knowledge layer eventually providing high-level abstractions according to the

needs of different disciplines. Initial elements of such an e-Infrastructure ecosystem already exist and Europe is actively striving for sustainability to ensure that it continues to build a reliable basis for e-Research.

2.2 The EGEE Distributed Computing Infrastructure[2]

Key Concepts

- gLite middleware, providing access to shared resources
- Security and access to resources
- Data storage
- Compute resources
- Deployment of the infrastructure
- Application development

2.2.1 Introduction

The Enabling Grids for E-sciencE (EGEE)[3] project provides an e-Research platform to the European research community and their international collaborators for high throughput data analysis for over 17,000 users across 160 projects. With a heritage stretching back over nearly a decade, EGEE-III (and its proceeding projects EGEE-II, EGEE-I and the European Data Grid) is a 32M € project funded by the European Commission to implement and deploy a distributed computing infrastructure to support researchers in many scientific domains, such as astrophysics, biomedicine, computational chemistry, earth sciences, high energy physics, finance, fusion, geophysics and multimedia. In addition, there are several applications from business sectors running on the EGEE Grid, such as applications from geophysics and the plastics industry. This chapter introduces the EGEE project in detail, considering middleware, security issues, access to information, data, compute resources, deployment and application development. It concludes with a look at sustainability.

[2] This content is available online at <http://cnx.org/content/m32047/1.1/>.
[3] http://www.eu-egee.org/

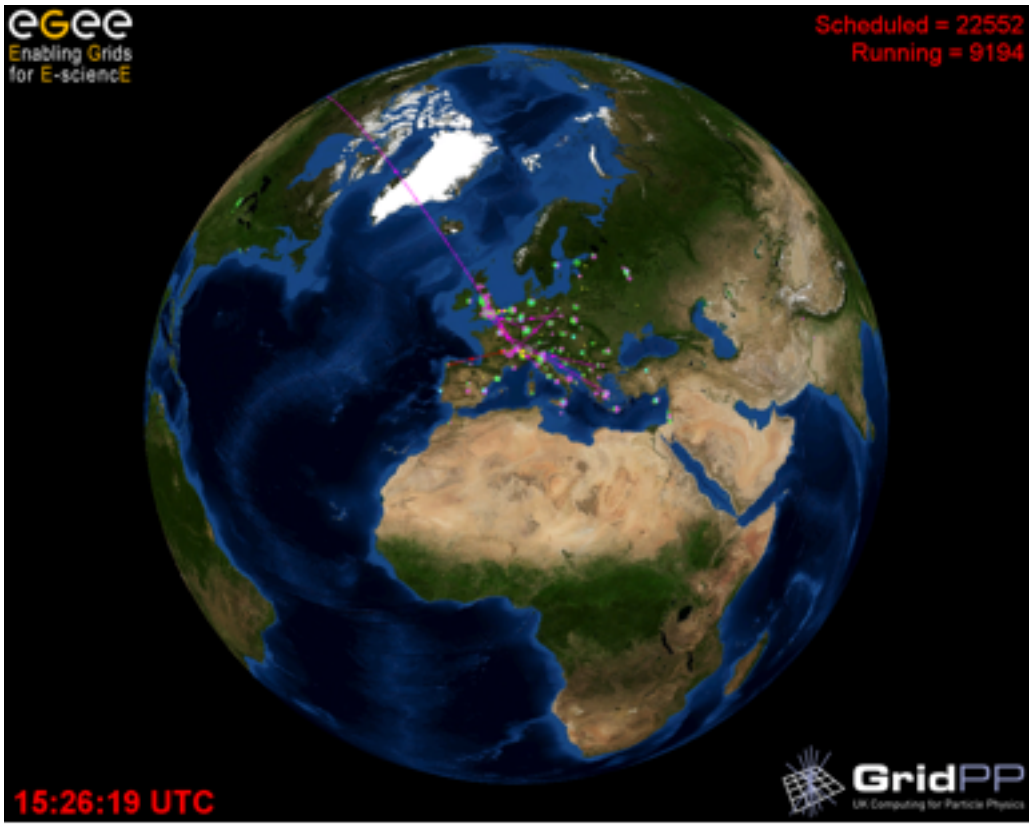


Figure 2.4: EGEE nodes mapped onto the globe

2.2.2 Facilitating Access To Shared Resources – the gLite Middleware

Grids are characterised by decentralised access to shared resources. These resources consist of computers, disks, and the network connections that link them together. Seamless, secure and scalable access to these resources, which may be owned by different organisations and could encompass different operating systems and architectures, is provided through software called **middleware**. Organisations that wish to cooperatively share their resources with their collaborators can do so without central control.

The gLite middleware distribution produced by the EGEE project is composed primarily of open-source software from many sources – some developed within the project and others from external providers. This software is integrated into a single software distribution before being tested and made available to sites for installation.

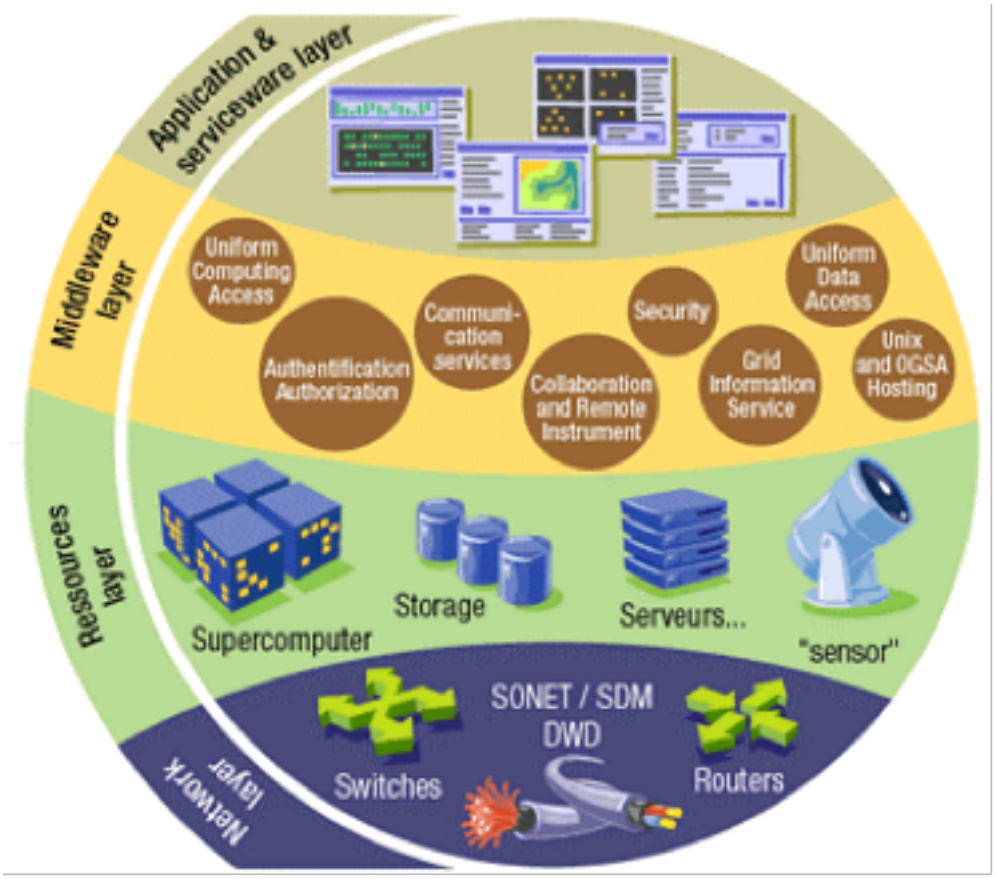


Figure 2.5: Layers of the EGEE e-Infrastructure

The software services used within gLite are there to enable researchers, through their own applications, to access the physical resources (disks, computers, instruments) that are attached to the EGEE infrastructure. These services have defined interfaces which allow developers to build their own applications.

2.2.3 Security

The resources that make up the EGEE infrastructure are extremely valuable and access to these resources needs to be strictly controlled. Access to resources within EGEE is restricted to members of research collaborations (commonly called Virtual Organisations). Some resources may only allow individuals from a single Virtual Organisation to access their resources, while other resource providers will provide a shared resource to multiple Virtual Organisations spanning several research communities.

To join a Virtual Organisation you need to be able to prove who you are electronically. This is similar to the way that a passport is used to prove your identity when you cross international borders. Within the Grid community this is frequently done through the use of a certificate – generally issued by proving your identity to someone at your local institute. Some organisations allow you to generate a certificate through your existing ability to access your organisation's own network.

Once you are able to identify yourself you can apply to join a Virtual Organisation. Different Virtual Organisations exist to cover the needs of different communities. A community may have more than one

Virtual Organisation within it with each one having different entry criteria and possibly providing access to different resources.

2.2.4 Information

With many thousands of services potentially available to a user, discovering which one to use presents many challenges. The infrastructure is continually changing – services are appearing, disappearing or being upgraded as the sites evolve. Being able to discover, in near real-time, the types of services that are available, the Virtual Organisations that are able to access them, and the characteristics of each service (i.e. the data that it stores or the speed of the processors), and the load on the service, are all information points that can drive which service to select.

The information collected by gLite on the resources within the infrastructure can be presented in many ways. The information can be browsed directly through a web portal, searched manually through command line tools, or programmatically from within an application.

2.2.5 Data

Many of the researchers that use EGEE's infrastructure do so in order to analyse data stored in files. Frequently these files are stored at locations different from the currently available computational resources. EGEE provides services that allow users to retrieve a file from off-line tape storage onto disk, and to then move that file to the site where the computational resources for that user is going to be available. (This method of data storage and retrieval is normal in high energy physics experiments and frequently used in other communities dealing with large archived data sets such as climate modelling and satellite observation records.)

How does a user locate the file that they need to use on an infrastructure where the location of files is continually changing? File catalogues run by some communities provide a register where a 'logical' file name can be mapped to a number of physical replicas. Having files stored in multiple places has many benefits - files are still available even if one of the sites storing the files is temporarily disconnected from the network or the service is down. Software can be written to exploit the distributed location of the files so as to run an application on the computing resources located near to their storage location – thereby reducing the time taken to move the files from their storage location to the compute resources.

Many of EGEE's sites are linked together through dedicated network connections. EGEE provides a service that is able to coordinate the bulk transfer of files across the network that allows the connection to be shared between different communities and file transfers to be managed and prioritised.

2.2.6 Computing

Key to nearly all communities supported by EGEE is the ability to analyse file-based data. Generally, applications need to be installed on the compute resources before they can be used to analyse their data. Their availability on a resource is something that can be advertised through the information system and allows the user to select the resources where their applications are already available. These applications are then started through services that encapsulate the compute resource – regardless of the operating system or the internal structure of the compute cluster that will be used to analyse the data.

The user, or their application, uses the EGEE information service to select a compute resource that they have access to and where their application is installed. Any input files needed by the application are transferred from the user's computer, located in storage out on the EGEE infrastructure through the file catalogue, or by knowing its explicit location, to the compute resource that will be used for the analysis. Once the file is in place, the request to start the application and analyse the file, is passed to the compute service. When the analysis is complete any output files will be available on the compute resource. If the user wishes the output files to remain available for future use they will need to be transferred back to the user's desktop or stored elsewhere.

Within EGEE some user communities undertake this process manually by knowing where their files are and which compute resources they wish to use. Other communities have written their own applications that directly mimic the manual processes thereby simplifying the life of the user. EGEE provides a generic resource brokering service that is able to automatically perform these tasks for many of the core scenarios previously done manually by a user.

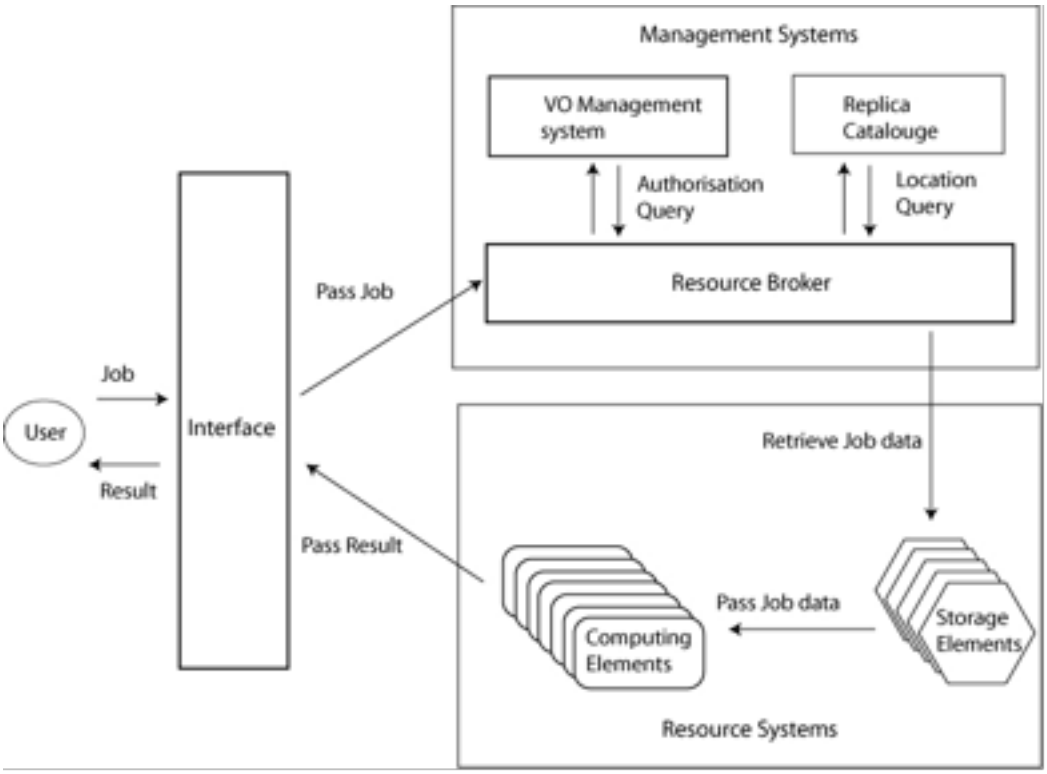


Figure 2.6: Overview of the architecture of EGEE

2.2.7 Accounting

As the compute, storage and network resources are contributed by different organisations for shared use by groups outside their organisations, it is important that this use is accounted for. Many organisations share their resources through 'service level agreements' that specify the proportions of the resource that can be used by different communities. Within EGEE the use of individual computing resources is accounted for and recorded centrally for later analysis and reporting. These usage agreements are validated through these centralised accounting records. Similarly, the volume of data transferred over the dedicated network links between the primary resource centres is also reported. This usage is generally reported for each Virtual Organisation using the infrastructure.

2.2.8 Operations

Vital to EGEE, and for any project that aims to deploy and support an infrastructure, is its operational effectiveness and availability. EGEE's infrastructure is deployed in over 50 counties on over 280 sites and

encompasses over 80,000 processors and 20PB of data enough to store 400 million four-drawer filing cabinets full of text or 50 million CDs – a stack around 50km high. This infrastructure is available continuously and supports over 300,000 jobs a day and the research network connecting these sites and the distributed user community sustains transfer speeds of over 900MB/s each day.

2.2.9 A Platform for Application Development

Over the last decade the software interfaces to EGEE have stabilised, matured and now form a platform that provides a basis for external developers to build their own applications. These developers come from both the research community, that use EGEE for their own work, and the broader software community that provide higher-level tools and services for the research community to use. Some of the latter work is starting to appear in the RESPECT programme (Recommended External Software for EGEE CommuniTies - http://technical.eu-egee.org/index.php?id=290) that aims to publicise software and services that work well in concert with the EGEE gLite software and thereby expand the functionality of the grid infrastructure for users, promote the reuse of existing software to reduce duplicated development, and to provide software more oriented to end users than the core gLite middleware distribution.

This activity has expanded over the last year and now includes various software packages:

- meta-schedulers able to run unattended applications and their related workflows by dynamically selecting resources
- portals that provide access to EGEE's resources through a web interface
- tools that simplify the specification and execution of applications – especially those that involve the execution of the same application over large sets of data
- services that provide access to data stored in files or in relational databases
- tools that help developers to build applications to access grid resources

2.2.10 Summary

As the EGEE-III project enters its final year work continues on improving the effectiveness, usability, availability and reliability of the infrastructure. The user community continues to expand and an increasing number of researchers are coming to depend on this e-Infrastructure as part of their regular daily work. In recognition of the increasing maturity of this e-infrastructure, the community has been studying over the last year how this infrastructure can be made more sustainable. The result, the European Grid Infrastructure (EGI), establishes a small organisation that federates and coordinates the work of independent National Grid Infrastructures (NGI) and is due to start in May 2010.

Chapter 3

Managing Complex Data

3.1 Scholarly Communication and the Web[1]

3.1.1 Introduction

The social process of sharing research results underpins the progress of research. For many decades our research has been published in journal articles, conference proceedings, books, theses and professional magazines. With increasing availability of tools to disseminate knowledge digitally, and with increasing participation in the digital world through widespread access to the Web, we are seeing this scholarly knowledge lifecycle become digital too. Although we have seen some welcome changes, including open access publishing which makes material free for all to read, the shared artefact in this lifecycle is predominantly still the academic paper. We might call this "Science 1.0".

e-Science is taking us into the "Science 2.0" world where we have new mechanisms for sharing (Schneiderman 2008) but also new artefacts to share. The tooling of e-Science produces and consumes data, together with metadata to aid interpretation and reuse, and also the scripts and experiment plans that support automation and the records that make the results interpretable and reusable – our new forms of artefact include data, metadata, scripts, scientific workflows, provenance records and ontologies. Our tools for sharing include the array of collaboration tools from repositories, blogs and wikis to social networking, instant messaging and tweeting that are available on the Web today, though these are not always designed around the new artefacts nor do they always have the particular needs of the researcher in mind.

These are already the familiar tools of the next generation of researchers and their uptake may seem inevitable, though it may take time for them to be appropriated and embedded in research practice. But crucially the other driver for change is the evolution of research practice as more work is conducted *in silico* and as we pursue multidisciplinary endeavours in data-intensive science to tackle some of the biggest problems facing society, from climate change to energy.

In this chapter we look at emerging practice in collaboration and scholarly communication by focusing on a case study which exemplifies a number of the principles in the paradigm shift to Science 2.0 and gives us a glimpse into the future needs of researchers.

3.1.2 myExperiment

myExperiment is an open source repository solution for the born-digital items arising in contemporary research practice, in particular *in silico* workflows (see the contribution by Fisher *et al.* (Section 3.2)) and experiment plans (DeRoure *et al.* 2009). Launched in November 2007, the public repository (myexperiment.org) has established a unique collection of workflows and a diverse international user community. The collection serves both researchers and learners: ranging from self-contained, high value research analysis

[1] This content is available online at <http://cnx.org/content/m32860/1.3/>.

methods referenced by the journal publications that discuss the results of their use, to training workflows that encode routine best practice scientific analyses or illustrate new techniques for new kinds of research data.

myExperiment has focused on support for sharing pieces of research method, such as scientific workflows and experimental plans, in order to address a specific need in the research community in both conducting research and training researchers. Experimental plans, standard operating procedures and laboratory protocols are descriptions of the steps of a research process, commonly undertaken manually. Scientific workflows are one of the most recent forms of scientific digital methods, and one that has gained popularity and adoption in a short time – they represent the methods component of modern *in silico* science and are valuable and important scholarly assets in their own right. Repositories often emphasise curation of data, but in digital research the curation of the process around that data is equally important – methods are crucial intellectual assets of the research life cycle whose stewardship is often neglected (Goble and De Roure 2008), and by focusing on methods, myExperiment provides a mechanism for expert and community curation of process in a rapidly changing landscape.

While it shares many characteristics with other Web 2.0 sites, myExperiment's distinctive features to meet the needs of its research user base include support for credit, attributions and licencing, fine control over privacy, a federation model and the ability to execute workflows. Hence myExperiment has demonstrated the success of blending modern social curation methods (social tagging, crowd sourcing) with the demands of researchers sharing hard-won intellectual assets and research works within the scholarly communication lifecycle.

3.1.3 Research Objects

The Web 2 design patterns (O'Reilly 2005) tell us "Data is the next Intel Inside. Applications are increasingly data-driven. Therefore for competitive advantage, seek to own a unique, hard-to-recreate source of data." Significantly, myExperiment also recognises that a workflow can be enriched as a sharable item by bundling it with some other pieces which make up the "experiment". Hence myExperiment supports aggregations of items stored in the myExperiment repository as well as elsewhere. These are called "packs", and while a pack might aggregate external content stored in multiple specialised repositories for particular content types, the pack itself is a single entity which can be tagged, reviewed, published, shared etc. For example, a pack might correspond to an experiment, containing input and output data, the experimental plan, associated publications and presentations, enabling that experiment to be shared. Another example is a pack containing all the evidence corresponding to a particular decision as part of the record of the research process. Packs are described using the Open Archives Initiative's Object Reuse and Exchange representation which is based on RDF (Definition: "RDF", p. 90) graphs and was specifically designed with this form of aggregation in mind (Van de Sompel 2009).

While some publishers are looking at how to augment papers with supplemental materials, raising concerns about peer-review and about decay, myExperiment is tackling this from first principles by starting with the digital artefacts and asking "what is the research object that researchers will share in the future?" These Research Objects have important properties:

- Replayable – go back and see what happened. Experiments are automated and may occur in milliseconds or in months. Either way, the ability to replay the experiment, and to study parts of it, is essential for human understanding of what happened.
- Repeatable – run the experiment again. There's enough in a Research Object for the original researcher or others to be able to repeat the experiment, perhaps years later, in order to verify the results or validate the experimental environment. This also helps scale to the repetition of processing needed for the scale of data intensive science.
- Reproducible – run a new experiment to reproduce the results. To reproduce (or replicate) a result is for a third party to start with the same materials and methods and see if a prior result can be confirmed.

- Reusable – use as part of new experiments or Research Objects. One experiment may call upon another, and by assembling methods in this way we can conduct research, and ask research questions, at a higher level.
- Repurposeable – reuse the pieces in a new experiment. An experiment which is a black box is only reuseable as a black box. By opening the lid we find parts, and combinations of parts, available for reuse, and the way they are assembled is a clue to how they can be re-used.
- Reliable – robust under automation, which brings systematic and unbiased processing, and also "unattended experiments" without a human in the loop. In data-intensive science, Research Objects promote reliable experiments, but also they must be reliable for automated running.

To achieve these behaviours it is crucial to store provenance records and full contextual metadata in the Research Object, so that results can be properly interpreted and replicated. This complete digital chain from laboratory bench to scholarly output is exemplified by the work on repositories and blogs in laboratories (Coles and Carr 2008), and also in the use of electronic laboratory notebooks.

We believe that in the fullness of time, objects such as these will replace academic papers as the entities that researchers share, because they plug straight into the tooling of e-Research. This means it is Research Objects rather than papers that will be collected in our repositories, and as well as a workflow repository, myExperiment has become a prototypical Research Object repository.

3.1.4 Linked Data

To achieve these properties, a Research Object must be self-contained and self-describing – containing enough metadata to have all the above characteristics and have maximal potential for re-use, whether anticipated or unanticipated. To support this, myExperiment provides a SPARQL endpoint (rdf.myexperiment.org) that makes myExperiment content available according to the myExperiment data model – a modularised ontology (Definition: "Ontology", p. 90) drawing on a set of emerging ontologies and standards in open repositories, scientific discourse, provenance and social networking.

myExperiment also aims to be a source of *Linked Data* so that myExperiment content can be readily integrated with other scientific data. The Linked Data initiative (linkeddata.org) enables people to share structured data on the Web as easily as they can share documents – as with documents, the value and usefulness of data increases the more it is interlinked with other data. To be part of the Linked Data web, data has to be accessible as RDF over the HTTP protocol in line with guidelines. At the time of writing there are 8 billion triples in Linked Data datasets.

With linked data a user can assemble a workflow in minutes to integrate data and call upon a variety of services from search and computation to visualisation. While the linked data movement has persuaded public data providers to deliver RDF, we are now beginning to see assembly of scripts and workflows that consume it – and the sharing of these on myExperiment. We believe this is an important glimpse of future research practice: the ability to assemble with ease experiments that are producing and consuming this form of rich scientific content.

3.1.5 Discussion

There is an open debate about the extent to which open publication should be mandated through the project lifecycle. A common pattern is to share artefacts with friends and colleagues and then make them available more broadly at time of publication. myExperiment supports this model, providing privacy and facilitating openness. In contrast, some sites like openwetware.org oblige their members to make everything public and still enjoy considerable adoption, exemplifying the *open science* approach.

Scholarly communication is evolving (Hey and Hey 2006) but the traditional academic publishing system has reinforced silos and made communication between disciplines more difficult. In contrast, important challenges like climate change research, which cut across different research communities, demand a social infrastructure to support resource sharing in large teams, and new shared artefacts. With this there also needs to be a culture of sharing and of making shared artefacts re-usable. It is clear that the behaviour of researchers

is closely related to incentive models, and these are currently set up around traditional publications. The creation of data and digital methods needs also to be rewarded if they are to flourish as a powerful enabler of new research.

myExperiment shares many characteristics with social networking sites for scientists and also with open repositories and contemporary content management systems, but it also exemplifies some important principles in developing Science 2.0 solutions (DeRoure and Goble 2009). One is the focus on providing a specific solution to meet the immediate requirements of its users and to make it highly configurable to the immediate needs of new communities. Another is that the user can come to myExperiment and find it familiar to use, but equally the myExperiment functionality can be appropriated and integrated in the familiar working environment of the user, be it loosely coupled or tightly integrated. Through linked data, myExperiment realises the network effects of scientific data as well as the network effects of the scientific community. It is an example of the kinds of systems that enable Science 2.0.

3.1.6 References

Coles, S. and Carr, L. (2008). Experiences with Repositories & Blogs in Laboratories[2] . *Third International Conference on Open Repositories*, 1-4 April, Southampton, UK.

De Roure, D., Goble, C., Aleksejevs, S., Bechhofer, S., Bhagat, J., Cruickshank, D., Fisher, P., Hull, D., Michaelides, D., Newman, D., Procter, R., Lin, Y. and Poschen, M. (2009) Towards Open Science: The myExperiment approach[3] . *Concurrency and Computation: Practice and Experience*. (In Press)

De Roure, D. and Goble, C. (2009). Software Design for Empowering Scientists[4] . *IEEE Software*, 26(1). January/February 2009. pp. 88-95.

Goble, C. and De Roure, D. (2008). Curating Scientific Web Services and Workflows[5] . *Educause Review*, 43(5), September/October.

Hey, T. and Hey, J. (2006). e-Science and its implications for the library community[6] . *Library Hi Tech*, 24(4). pp. 515-528

O'Reilly, T. (2005) What Is Web 2.0? Design Patterns and Business Models for the Next Generation of Software[7] , September.

Shneiderman, B. (2008). Science 2.0[8] . *Science*, 319. pp. 1349-1350.

Van de Sompel, H., Lagoze, C., Nelson, ML., Warner, S., Sanderson, R. and Johnston, P. (2009). Adding eScience Assets to the Data Web[9] . CoRR abs/0906.2135

3.2 Scientific Workflows[10]

Key Concepts:

- Scientific workflows
- Data-intensive research

3.2.1 Introduction

The use of data processing workflows within the business sector has been commonplace for many years. Their use within the scientific community, however, has only just begun. With the uptake of workflows

[2] http://pubs.or08.ecs.soton.ac.uk/22/
[3] http://eprints.ecs.soton.ac.uk/17270/
[4] http://dx.doi.org/10.1109/MS.2009.22
[5] http://www.educause.edu/library/erm0857
[6] http://dx.doi.org/10.1108/07378830610715383
[7] http://oreilly.com/web2/archive/what-is-web-20.html
[8] http://dx.doi.org/10.1126/science.1153539
[9] http://arxiv.org/abs/0906.2135
[10] This content is available online at <http://cnx.org/content/m32861/1.3/>.

within scientific research, an unprecedented level of data analyses is now at the fingertips of individual researchers, leading to a change in the way research is carried out. This chapter describes the advantages of using workflows in modern biological research; demonstrating research from the field where the application of workflow technologies was vital for understanding the processes involved in resistance and susceptibility of infection by a parasite. Specific attention is drawn to the Taverna Workflow Workbench (Hull *et al.* 2006), a workflow management system that provides a suite of tools to support the design, execution, and management of complex analyses in the data intensive research, for example, in the Life Sciences.

3.2.2 Data-Intensive Research in the Life Sciences

In the last decade the field of informatics has moved from the fringes of biological and biomedical sciences to being an essential part of research. From the early days of gene and protein sequence analysis, to the high-throughput sequencing of whole genomes, informatics is integral in the analysis, interpretation, and understanding of biological data. The post-genomic era has been witness to an exponential rise in the generation of biological data; the majority of which is freely available in the public domain, and accessible over the Internet.

New techniques and technologies are continuously emerging to increase the speed of data production. As a result, the generation of novel biological hypotheses has shifted from the task of data generation to that of data analysis. The results of such high-throughput investigations, and the way it is published and shared, is initially for the benefit of the research groups generating the data; yet it is fundamental to many other investigations and research institutes. The public availability means that it can then be reused in the day to day work of many other scientists. This is true for most bioinformatics resources. The overall effect, however, is the accumulation of useful biological resources over time.

In the 2009 *Databases* special issue of *Nucleic Acids Research*, over 1000 different biological databases were available to the scientific community. Many of these data resources have associated analysis tools and search algorithms, increasing the number of possible tools and resources to several thousand. These resources have been developed over time by different institutions. Consequently, they are distributed and highly heterogeneous with few standards for data representation or data access. Therefore, despite the availability of these resources, integration and interoperability present significant challenges to researchers.

In bioinformatics, many of the major service providers are providing Web Service (Definition: "Web Service", p. 91) interfaces to their resources, including the NCBI, EBI, and DDBJ; many more are embracing this technology each year. This widespread adoption of Web Services has enabled workflows to be more commonly used within scientific research. Data held in the NCBI can now be analysed with tools available at the EBI, within analysis pipeline.

3.2.3 In Silico Workflows

One possible solution to the problem of integrating heterogeneous resources is the use of *in silico workflows*. The use of workflows in science has only emerged over the last few years and addresses different concerns to workflows used within the business sector. Rather than co-ordinating the management and transactions between corporate resources, scientific workflows are used to automate the analysis of data through multiple, distributed data resources in order to execute complex *in silico* experiments.

Workflows provide a mechanism for accessing remote third-party services and components. This in turn reduces the overheads of downloading, installing, and maintaining resources locally whilst ensuring access to the latest versions of data and tools. Additionally, much of the computation happens remotely (on dedicated servers). This allows complex and computationally intensive workflows to be executed from basic desktop or laptop computers. As a result, the researchers are not held back by a lack of computational resources or access to data.

A workflow provides an abstracted view over the experiment being performed. It describes *what* analyses will be executed, not the low-level details of *how* they will be executed; the user does not need to understand the underlying code, but only the scientific protocol. This protocol can be easily understood by others, so can be reused or even altered and repurposed. Workflows are a suitable technology in any case where

scientists need to automate data processing through a series of analysis steps. Such mechanisms have the potential to increase the rate of data analysis, from a cottage-scale to industrial scale operation.

There are many workflow management systems available in the scientific domain, including: Taverna (Hull *et al.* 2006), Kepler (Altintas *et al.* 2004) and Triana (Taylor *et al.* 2003). Taverna, developed by the the myGrid consortium (http://www.mygrid.org.uk/), is a workflow system that was built with the Life Sciences in mind but it has since been used in other fields as well, including Physics, Astronomy and Chemistry. Like many others, the Taverna Workbench provides:

- an environment for designing workflows;
- an enactment engine to execute workflow locally or remotely;
- support for workflow design in the form of service and workflow discovery;
- and provenance services to manage the results and events of workflow invocations.

3.2.4 Understanding Disease Resistance in Model Organisms

Taverna workflows are used in many areas of Life Science research, notably for research into genotype-phenotype correlations, proteomics, genome annotation, and Systems Biology. The following case study demonstrates the use of Taverna workflows in the Life Sciences domain for genotype-phenotype studies (Stevens *et al.* 2008).

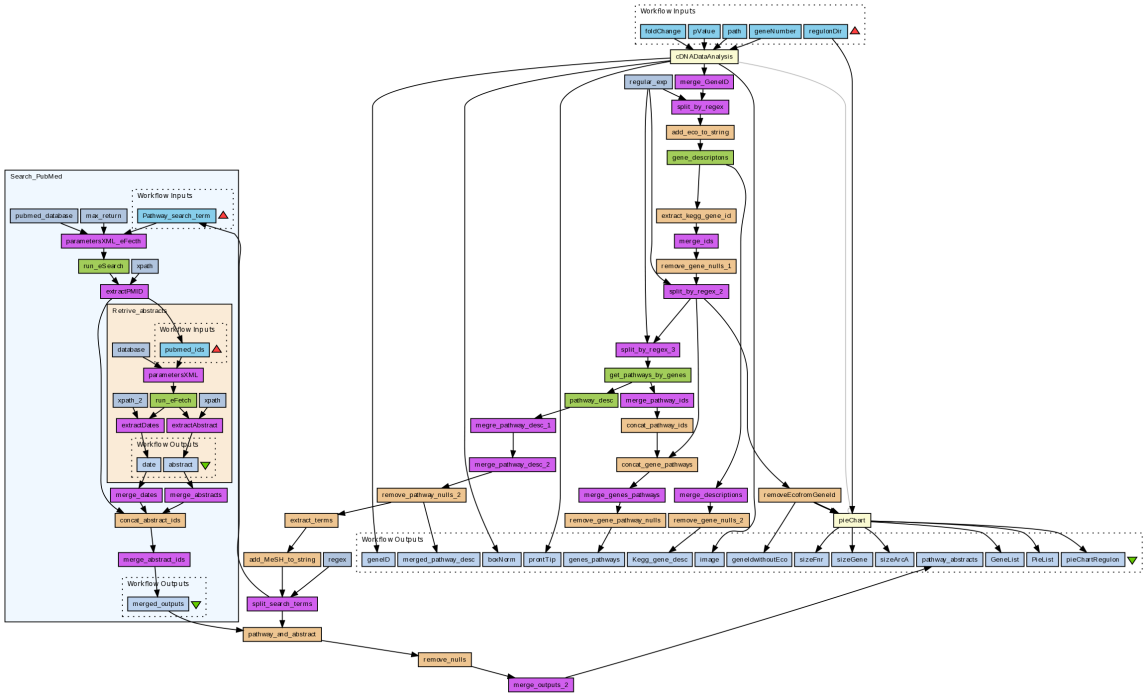


Figure 3.1: This figure shows the conversion of a microarray CEL image file to a list of candidate genes, pathways, and pathway publications. The workflow makes use of a local statistical processor, services from the National Centre for Biotechnological Innovation (NCBI) and the Kyoto Ecyclopedia of Genes and Genomes (KEGG)

Sleeping sickness (or African trypanosomiasis) is an endemic disease throughout the sub-Saharan region of Africa. It is the result of infection from the trypanosome parasite, affecting a host of organisms. The inability of the agriculturally productive Boran cattle species to resist trypanosome infection is a major restriction within this region. The N'Dama species of cattle, however, has shown tolerance to infection and its subsequent disease. The low milk yields and lack of physical strength of this breed, unfortunately, limit their use in farming or meat production. A better understanding of the processes that govern the characteristics of resistance or susceptibility in different breeds of cattle will potentially lead to the development of novel therapeutic drugs or the construction of informed selective breeding programs for enhancing agricultural production.

Research conducted by the Wellcome Trust Host-Pathogen [11] project is currently investigating the mechanisms of resistance to this parasitic infection, utilising Taverna workflows for a large-scale analysis of complex biological data (Fisher *et al.* 2007). The workflows in this study combine two approaches to identify candidate genes and their subsequent biological pathways: classic genetic mapping can identify chromosomal regions that contain genes involved in the expression of a trait (Quantitative Trait Loci or QTL) while transcriptomics can reveal differential gene expression levels in susceptible and resistant species.

Previous studies using the mouse as model organism identified 3 chromosomal regions statistically linked to resistance to trypanosome infection. One of these regions, the Tir1 QTL, showed the largest effect on survival. Previous investigations using this QTL identified a region shared between the mouse and cow genomes. As the scale of the data analysis task is large, researchers performing such an analysis manually would often triage their data and in this case have tended to focus on this shared region in their search for candidate genes contributing to the susceptibility to trypanosome infection. While this approach may be scientifically valid, there is a danger that candidate genes may be missed where additional biological factors may contribute to the expression of the phenotype. With a workflow, this triage of data is no longer necessary. *All* data can be analysed systematically, reducing the risk of missing vital information.

Researchers on the Wellcome Trust Pathogen-Host project conducted a wider analysis of the entire QTL region using a set of workflows to identify pathways whose genes lie within the chosen QTL region, and contain genes whose expression level changes. As a result of this research, a key pathway was identified whose component genes showed differential expression following infection from the trypanosome parasite. Further analysis showed that, within this pathway, the Daxx gene is located within the Tir1 QTL region and showed the strongest change in expression level. Subsequent investigations using the scientific literature highlighted the potential role of Daxx in contributing to the susceptibility to trypanosome infection. This prompted the re-sequencing of Daxx within the laboratory, leading to the identification of mutations of the gene within the susceptible mouse strains. Previous studies had failed to identify this as a candidate gene due to the premature triage of the QTL down to the syntenous region.

This example shows that conducting this kind of data-driven approach to analysing complex biological data at the level of biological pathways can provide detailed information of the molecular processes contributing to the expression of these traits. The success of this work was primarily in data integration and the ability of the workflow to process large amounts of data in a consistent and automated fashion.

3.2.5 Workflow Reuse

Workflows not only provide a description of the analysis being performed, but also serve as a permanent record of the experiment when coupled with the results and provenance of workflow runs. Researchers can verify past results by re-running the workflow or by exploring the intermediate results from past invocations. The same workflow can also be used with new data or modified and reused for further investigations.

The ability to reuse workflows and to automatically record provenance of workflow runs gives workflow management systems a large advantage over manual analysis methods and scripting. Manual analysis techniques are inherently difficult to replicate and are compounded by poor documentation. An example is the wide-spread use of 'link integration' in bioinformatics (Stein 2003). This process, of hyper-linking through

[11] http://www.genomics.liv.ac.uk/tryps/index.php

any number of data resources, further exacerbates the problem of capturing the methods used for obtaining *in silico* results where it is often difficult to identify the essential data in the chain of hyper-linked resources.

Workflow reuse is also an important area within the sciences, and provides a mechanism for sharing methodologies and analysis protocols. As a result, repositories for finding and sharing workflows are emerging. One such resource, myExperiment[12] , developed in collaboration between the Universities of Manchester and Southampton, provides a workflow repository and a collaborative social networking environment to support the *in silico* experimental process, and to enable scientists to connect with those with similar interests. The workflows dicussed in the trypanosomiasis use-case study are available on myExperiment[13] , as part of a workflow pack. Many of these have already been reused in other studies. One such example includes the re-purposing of the microarray gene expression workflow to analyse gene expression data from *E. Coli*. This workflow[14] appends a further workflow to include a means of information retrieval for future text mining applications (shown in Figure 1).

3.2.6 Discussion

Manually processing and examining results in biology is no longer feasible for many scientists. Data is dynamic, distributed, and often very large. This will not change in the near future.

The integration and interoperation of data between different and distributed resources is a vital part of almost all experiments. With the exception of a few supercomputing centres, most institutions do not have the storage, computational, or curation facilities to consider integrating resources locally. The ability to access and utilise many different resources from all over the world is consequently a large advantage of workflow technologies. It allows scientists to access computing resources far beyond the power available through their own desktop machines.

Building workflows is a practical solution to problems involving access to data and applications, but care still needs to be taken to exploit these advantages. Interoperation without integration may lead to unmanageable results which are difficult to analyse. In this event, the problem has not been solved, but simply transferred further downstream. Considering *how* results will be used and *who* will be analysing them is important. For example, designing workflows to populate a data model, or to feed into external visualization software, could reduce these problems. The provenance traces of the workflow runs can also help scientists to explore their results.

Designing these 'advanced' workflows requires a significant amount of informatics knowledge that many laboratory researchers cannot be expected to have. They do, however, need to use tools and software to analyse their data. The introduction of workflow repositories, like myExperiment, provides the wider research communities with access to pre-configured, complex workflows. Researchers can re-use established analysis protocols by downloading and running them with their own data. In some circumstances, they can even run Taverna workflows through the myExperiment interface.

Increasingly, workflows are becoming applications that are hidden behind web pages like myExperiment, or other domain specific portals. Instead of stand-alone tools, they are becoming integral parts of virtual research environments, or e-Laboratories. Users may not necessarily know they are invoking workflows.

The use of workflows in research can reduce many problems associated with data distribution and size. In the post-genomic era of biology, for example, this is extremely important. Biomedical science is a multi-disciplinary activity which can benefit from advances e-Science in equal measure to advances in laboratory techniques. Sharing workflows and *in silico* analysis methods, with tools like Taverna and myExperiment, can lead to significant contributions to research in this and other disciplines.

3.2.7 References

Altintas, I. *et al.* (2004). Kepler: an extensible system for design and execution of scientific workflows. *Proceedings of the 16th International Conference on Scientific and Statistical Database Management*

[12] http://myexperiment.org
[13] http://www.myexperiment.org/packs/83
[14] http://www.myexperiment.org/workflows/187

Fisher, P., Hedeler, C., Wolstencroft, K., Hulme, H., Noyes, H., Kemp, S., Stevens, R. and Brass, A. (2007). A systematic strategy for large-scale analysis of genotype phenotype correlations: identification of candidate genes involved in African trypanosomiasis. *Nucleic Acids Resesearch*, 35(16). pp. 5625-5633.

Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M., Li, P. and Oinn, T. (2006). Taverna: a tool for building and running workflows of services. *Nucleic Acids Research*, vol. 34, Web Server issue, W729-W732.

Stein, L. (2003). Integrating biological databases. *Nat Rev Genet*, 4(5). pp. 337-345.

Stevens, R. *et al.* (2004). Exploring Williams-Beuren syndrome using myGrid. *Bioinformatics*, 20 Suppl 1

Stevens, R. et al. (2008). Traversing the bioinformatics landscape. W. Dubitzky (ed.) *Data Mining Techniques in Grid Computing Environments*. John Wiley and Sons. pp. 141-164.

Taylor, I. *et al.* (2003). Triana Applications within Grid Computing and Peer to Peer Environments. *Journal of Grid Computing*, 1(2). pp. 199-217.

3.3 Repositories[15]

- Digital libraries
- Dataset repositories
- Repositories for and within workflows
- Repositories as Virtual Research Environments

3.3.1 Introduction

The digital material generated from and used by academic and other research is to an increasing extent being held in formally managed digital repositories. Digital repository systems arose in the self-archiving community – for example, arXiv[16] and Cogprints[17] , the latter of which gave rise to the EPrints [18] repository software – and in their earlier incarnations they were used to manage relatively simple content, primarily pre-prints and post-prints, sometimes less formal material such as presentations or lecture notes. A major motivation in setting up and populating such repositories was (and continues to be) to make the results of research available to a wider audience, by encouraging or mandating deposit and open access principles. In any case, from the point of view of the system these were individual objects, unrelated except via having metadata fields in common.

However, digital repositories have been changing, both in the type of content that they hold, and in the ways in which they are used; indeed, these two things are connected. Repository software has become more sophisticated, allowing complex digital content to be stored in such a way that its internal structure and external context can be explicitly represented, managed and exposed. Institutions are beginning to use them to manage research data in a variety of disciplines, including physical sciences, social sciences, and the arts and humanities, in part as a result of various programmes funded by the Joint Information Systems Committee (JISC[19]) in the UK.

Such systems allow us to move on from the model of a stand-alone repository, where objects are simply deposited for subsequent access and download. Instead, researchers are developing more sophisticated models in which repositories are integrated components of larger infrastructures, incorporating advanced tools and workflows. They are being used to model complex webs of information and capture scholarly or scientific processes in their entirety, from raw data through to final publications.

Within e-Science communities, much of the focus regarding data management has been on techniques for the efficient organisation of and access to large and distributed data sets, an issue that has been well

[15] This content is available online at <http://cnx.org/content/m31391/1.1/>.
[16] http://arxiv.org/
[17] http://cogprints.org/
[18] http://www.eprints.org/
[19] http://www.jisc.ac.uk/

addressed by various flavours of grid middleware. The particular challenge raised here, however, is not just size, but rather the very nature of the data, which can be highly diverse, complex, fuzzy and context-dependent, as well as the highly interpretative character of research in many disciplines, for example the humanities.

Another issue to be addressed is the silo mentality. Even if data is held in formally managed digital repositories, these are often managed on an institutional basis, resulting in information that is widely dispersed and not easy for researchers to locate and access. Although the repository content is in principle accessible via the internet, it is often held at a "deep" level that is not amenable to traditional discovery techniques. If, as we expect, digital repositories take on a central and pivotal role in the research lifecycle, then there is a clear strategic need to develop methods and tools to enable collaborative research through the coordination and federation of such complex and dispersed resources. This chapter will present case studies of repositories to show the range of ways in which they are used.

Rather than attempt an exhaustive analysis, it seems best to look at a number of categories of repository application: digital libraries, repositories of data sets, repositories of workflows and within workflows and repositories as Virtual Research Environments. Examination of each application will include concrete examples.

3.3.2 Repositories as Digital Libraries

While realising that the term is fuzzy, by digital library we understand here a system for managing, curating and delivering digital content that is primarily focused on (web) delivery to a human user, who is able to search, browse and access the material therein. While this also covers basic publication repositories, digital content may also be rich, varied and complex in structure, and digital libraries typically provide specialised access mechanisms and functionality for particular content types.

While some may dispute that the term should be so restricted – see for example the classic article 'What Is a Digital Library Anymore, Anyway?[20] ' – systems that provide functionality for manipulating, annotating etc. material may be better thought of as virtual research environments.

Examples of such repositories are many. Early examples are the University of Virginia Library[21] or The Encyclopedia of Chicago[22] , which use repositories to manage extensive collections of diverse material, including books, documents, images, maps and pre-existing websites. Many of these objects are themselves compound, and possess contextual relationships with other objects, all of which need to be represented in the data model. Different types of object are provided with particular access mechanisms that are required to be consistent across these object types.

More recently, there has been an emphasis on lightweight mechanisms for producing this sort of digital library. An interesting example is Active Fedora[23] , which exploits RESTful web services (for an explanation, please see http://en.wikipedia.org/wiki/Representational_State_Transfer?) and rapid web development technologies to provide a low cost and low effort means for producing such applications, for example the Jewish Women's Archive[24] .

While a user may see only a website, there are important differences between using a repository and just using, say, a Content Management System (CMS); the **structure** of the information is entirely managed within the repository, and the **services** that support the behaviours and functionality of the information are associated with the repository objects. Both are decoupled from the online delivery mechanism, thus "future proofing" the content. Moreover, the repository provides a generic way of representing content, in contrast to specialised (often commercial) systems for delivering particular types of content, such as book digitisations, which provide high levels of specific functionality but less flexibility.

[20] http://www.dlib.org/dlib/november05/lagoze/11lagoze.html
[21] http://www.lib.virginia.edu/digital/
[22] http://encyclopedia.chicagohistory.org/
[23] http://projects.mediashelf.us/projects/show/active-fedora
[24] http://jwa.org/

3.3.3 Repositories of Datasets

Take up of repositories for managing more complex material occurred first in the arts, humanities and cultural heritage fields. In the sciences, the focus for data management was more on the need to manage, transfer and access large data sets distributed over multiple, collaborating research centres. The eCrystals[25] archive, developed by the JISC-funded eBank[26] project, was a pioneer in the use of digital repository technology – in this case EPrints[27] – to archive, curate and disseminate research data, specifically chemistry datasets in the particular field of small-molecule crystallography. The repository here is integrated within the researcher's work activities – datasets are deposited once the experiment is complete – and it exploits the publication mechanisms established for more traditional institutional repositories.

More recently, this team has also been in the forefront of federating repositories. While a solid foundation within an institution may be essential for the longevity and sustainability of the material, from the point of view of a researcher the location of a dataset is irrelevant. A researcher, or indeed any user, wants to be able to locate, access, reuse and combine datasets independently of where they happen to be managed. The eCrystals Federation [28] is establishing an international partnership of crystallography data repositories implemented using heterogeneous technologies, with the aim of integrating them into a broader framework for the curation and dissemination of crystallographic data.

There is also a recognised need for linking research publications in institutional repositories with the primary scientific datasets on which the research is based, which will in general be held in separate repositories such as the ones described here. A number of projects – for example StORe[29] and CLADDIER[30] – have been investigating the potential for such linkage.

3.3.4 Repositories of Workflows

There is a need to curate scientific processes as well as the data. myExperiment[31] uses digital repository systems to curate and share scientific workflows as part of a social networking environment in which researchers can annotate, discuss and execute these workflows. Initially developed as a community resource for Taverna[32] users, the scope of myExperiment[33] has expanded to incorporate users of other workflow systems.

myExperiment[34] is more than a website for sharing these objects – it addresses directly many of the issues around publication in more traditional digital repositories, including object versioning, and – importantly for a repository of published artefacts – author attribution and credit, making use of the trust models implicit in the Web 2.0 world. Workflows are just as much an intellectual creation of someone's research as a traditional publication. myExperiment[35] is currently being enhanced as a result of further JISC[36] funding, to support additional content types, and provide a richer range of functionality, such as user-defined vocabularies for more effective tagging and discovery, and a mechanism for better expressing relationships between items in the repository, which can be used for (e.g.) provenance capture.

3.3.5 Repositories Within Workflows

The RepoMMan[37] project developed an environment that allowed users to interact with a repository as part of their natural workflow. The repository is not just viewed as a stand-alone "silo" of information into which

[25] http://ecrystals.chem.soton.ac.uk/
[26] http://www.ukoln.ac.uk/projects/ebank-uk/
[27] http://www.eprints.org/
[28] http://wiki.ecrystals.chem.soton.ac.uk/index.php/Main_Page
[29] http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2005/store.aspx
[30] http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2005/claddier.aspx
[31] http://www.myexperiment.org/
[32] http://taverna.sourceforge.net/
[33] http://www.myexperiment.org/
[34] http://www.myexperiment.org/
[35] http://www.myexperiment.org/
[36] http://www.jisc.ac.uk/
[37] http://hull.ac.uk/esig/repomman/

researchers explicitly deposit completed research outputs (and then perhaps forget about them), but is rather a tool that supports researchers by providing services for managing digital content throughout their research activities, from initial conception and experimental work through to archiving and publication. This was implemented by using a workflow engine to orchestrate web services supporting common repository-related tasks.

Although RepoMMan[38] was developed in the context of creating and managing documents rather than complex material and research data, this restriction is not essential. Indeed, this pioneering work is being continued in other current projects, notably Hydra[39] , which is developing a Scholar's Workbench[40] with a repository at its core, and CLIF[41] , which is investigating the repository as an embedded part of an environment supporting the lifecycle of digital content in various forms.

3.3.6 Repositories as Virtual Research Environments

The term Virtual Research Environment (VRE) covers a range of different systems, and this is not the place for a definitive definition. However, as virtually all research deals with data in some form, a repository is likely to be an important component of any such environment.

An example is provided by the eSciDoc[42] system, one of the aims of which was to provide a digital library/archive ensuring permanent access to the variety of research outputs of the Max Planck Society (MPG)[43] . For the present purposes, however, more interesting is its use as a generic framework for building virtual research environments for specific research communities within the MPG[44] .

eSciDoc[45] provides a generic infrastructure[46] and a rich set of services[47] , built around a digital repository, to support the entire lifecycle of a research project, including visualisation, manipulation, processing and publication of the various information objects used by researchers. This framework is used to develop "solutions[48] ", which are researcher-centric applications for supporting particular research communities; a number of these have been developed by the library service of the MPG[49] .

eSciDoc[50] is a system with very rich functionality and potential; in consequence, developing these solutions for specific communities requires a significant amount of programming effort. At the other end of the spectrum, the Islandora[51] project is developing a VRE framework based around the Fedora[52] repository system, which like eSciDoc[53] allows distinct environments to be created for individual research groups around a common repository infrastructure. It is however a much more lightweight solution, developed in the form of plug-ins for an open source content management system, and allows new environments[54] to be created in rapid and agile fashion.

A final example illustrates the potential synergies between repositories and wider e-infrastructure technologies, specifically the gLite[55] grid middleware. The gCube[56] system allows the ad hoc creation of *ad hoc* data-centric virtual research environments that are tailored to the needs of specific resource groups,

[38] http://hull.ac.uk/esig/repomman/
[39] http://www.dlib.org/dlib/may09/green/05green.html
[40] http://www.fedora-commons.org/confluence/display/FCCWG/Scholars+Workbench
[41] http://www.jisc.ac.uk/whatwedo/programmes/inf11/clif.aspx
[42] http://www.escidoc.org/
[43] http://www.mpg.de/english/portal/index.html
[44] http://www.mpg.de/english/portal/index.html
[45] http://www.escidoc.org/
[46] http://www.escidoc.org/JSPWiki/en/Infrastructure
[47] http://www.escidoc.org/JSPWiki/en/Services
[48] http://www.escidoc.org/JSPWiki/en/Solutions
[49] http://www.mpg.de/english/portal/index.html
[50] http://www.escidoc.org/
[51] http://vre.upei.ca/dev/islandora
[52] http://www.fedora-commons.org/
[53] http://www.escidoc.org/
[54] http://vre.upei.ca/dev/VREsites
[55] http://glite.web.cern.ch/glite/
[56] http://www.gcube-system.org/

and built on top of gLite[57] -based grid infrastructures such as EGEE[58] [will add link to Steve Newhouse's chapter here]. These environments provide virtual repositories that allow pre-existing data resources, diverse in terms of formats and metadata standards, to be combined, manipulated and annotated.

3.3.7 Summary: The Value of Repositories in Research

We have seen how repositories have evolved greatly from their beginnings as a "home" for research papers, both in terms of the material that they hold and the uses to which they are put, and our examples are moreover far from exhaustive. Crucially, repositories are being integrated with broader processes and infrastructures, not only within research groups and institutions, but globally. In particular, the architecture of the web is being exploited to facilitate the exposure and re-use of digital material in repositories. This movement towards a more joined up world may be expected to continue, making use of approaches such as Linked Data[59] , which aims to use the web, and in particular Semantic Web[60] , technologies, to forge connections between related data, leading to a vision of a digital ecosystem in which repositories form a key component.

3.4 Resource Sharing: Trust and Security[61]

Key Concepts

- Single sign-on access to distributed resources
- Certification Authority (CA) and its problems
- Shibboleth technologies
- Portlets for finer grained security of portals – SCAMP, CCP and SPAM-GP ACP

3.4.1 Introduction

Many researchers require environments providing seamless access to and usage of a heterogeneous variety of distributed resources: on-line journals, data repositories and archives, software, large scale high-performance computing facilities (HPC) or indeed support for collaborations between distributed research teams themselves. The internet-age is truly upon us and there are few disciplines where radical IT-driven change in the way research is undertaken has not been felt. The vision of e-Science and the Grid, as part of e-Research, has been to support seamless and transparent access to such heterogeneous resources. Solutions within the e-Science model should support user/research-oriented environments offering seamless single sign-on to a range of research-specific distributed resources. For many disciplines however, trust and security are paramount and many existing models of single-sign on security are inadequate. Instead controlled trust-driven environments are required where sites can remain autonomous and in strict control of their resources through their own discretionary local access and usage policies. In this paper we outline how the UK Access Management Federation[62] , augmented with advanced authorization solutions, supports this model. This UK example can serve as a more general exemplar for other national contexts.

3.4.2 Single Sign-on and a Centralized Certification Authority

It is a fact that security is essential for much, if not all, inter-organizational collaborative research. Many disciplines place a higher emphasis on security of resources, e.g. the clinical health domain, but even those

[57] http://glite.web.cern.ch/glite/
[58] http://www.eu-egee.org/
[59] http://www.w3.org/DesignIssues/LinkedData.html
[60] http://www.w3.org/2001/sw/
[61] This content is available online at <http://cnx.org/content/m32872/1.1/>.
[62] http://cnx.org/content/m32872/latest/www.ukfederation.org.uk

disciplines where security is not a primary focus, e.g. the particle physics domain, would be seriously affected by downtime or compromise of HPC facilities that they use.

From a security perspective, the vision of e-Science and the Grid has been to provide *single sign-on* access to distributed resources, i.e. where a user is able to access multiple resources without the need for multiple, individual authentications (username/password challenges for example). This has been largely tackled in the UK through establishment of a centralized Certification Authority (CA – www.grid-[63] support.ac.uk/ca[64]). Through recognizing and trusting a CA in associating the identity of a researcher with a particular digital certificate (typically through a local institutional Registration Authority charged with ensuring that the user presents in person their passport or matriculation card as evidence of their identity), single sign-on authentication can be supported. Thus researchers use their X509 certificate (or more often a proxy credential created from that X509 certificate) with a common username given by the distinguished name (DN) associated with that credential and a single (strong) password. Through sites trusting the CA that issued the certificate, the end user is able to access a wide range of resources that recognize that credential without the need for multiple usernames and passwords across those sites. In short, the approach is based upon a model of public key infrastructure (PKI) supporting user authentication.

Relying solely upon X509-based PKI models for Grid security suffers from numerous problems. Firstly, users must acquire and manage their own X509 digital certificates often having to convert them to different Grid-oriented formats using specialized software, often far removed from their own domain of expertise. Secondly, X509-based PKI security as used to access resources such as the UK e-Science National Grid Service (NGS – www.ngs.ac.uk) is based on associating a local account on an NGS cluster for a user identified by their DN, to "do stuff" without limiting (authorizing) what this stuff actually is! For security considerations, this is obviously a major issue for many domains. Thirdly, a centralized CA model has issues with supporting longer term identity management. Thus if an individual leaves the organization they are associated with, they probably will still be in possession of their X509 digital certificate. A better model is to support decentralized authentication. In the UK this has been supported through adoption and roll-out of the Internet2 Shibboleth technologies to support devolved (federated) access management (www.ukfederation.org.uk).

3.4.3 Shibboleth: Decentralized Authentication

The core of Shibboleth is a trust relationship between institutions within a federation, where each institute in the federation is expected (trusted) to authenticate their users properly. The architecture of Shibboleth defines several entities which are necessary to achieve this seamless integration of separate collaborating institutional authentication systems. The main components of Shibboleth consist of Identity Providers (IdPs); a Where-Are-You-From (WAYF) service, and one or more Service Providers (SP). The IdP is typically the users' home institution and is responsible for authenticating the end users at their institution. Each institution will have their own local systems for authenticating their users, e.g. LDAP or other mechanisms. The WAYF service is generally run by the federation that the institutions are subscribed to. It typically presents a dropdown list to the user that contains all the participating institutions (or projects) that are subscribed to within the federation. Users choose their home institution from this list and are then redirected to the home institution (IdP). The SP provides services or resources for the federation that the end user wishes to access.

A typical scenario of this process is shown in Figure 1, where a user types in the URL of the service or Grid portal (SP) they wish to access.

[63] http://www.grid-support.ac.uk/ca
[64] http://www.grid-support.ac.uk/ca

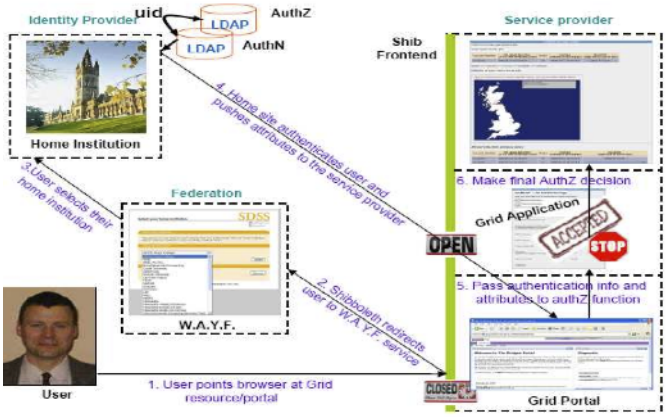


Figure 3.2: Shibboleth-oriented Federated Access

In this model, if the SP is protected by Shibboleth, the user will be redirected to the WAYF service where they select their home institution. Once redirected to their IdP they will provide the username and password they would normally use for authentication at their home institution. Once successfully authenticated, the user will be automatically redirected to the SP they are trying to access. At the same time, the security attributes (privileges) of this user will also be passed to the SP in a secure manner for further authorization from either the IdP or one or more known attribute authorities (AA). What attributes will be released by an institutional IdP or AA and what attributes will be accepted by a given SP needs to be configurable however and targeted towards the needs of particular virtual organizations. It is important that all of this is transparent to the end users (who simply log-in to their home site).

Key to this model is trust and security. Sites need to be sure that collaborating sites have adopted appropriate security policies for authentication for example and that they have appropriate rigor in management of strength of user passwords and ideally, support a unified institutional account management. It is also the case that for some virtual organizations, some sites may be recognized (trusted) whilst others may not. A given SP might only be available to members of a particular virtual organization for example, and only researchers at the sites involved in the collaboration should be able to access and use the SP resources. Similarly, given institutions need to be able to define which attributes they wish to send to which SPs. A naïve model would be that all attributes for a given user from a given site would be sent through to a given SP. However an improved model would support a targeted (reduced) subset of attributes that are released based upon what that service requires to make an access control decision. The Open Middleware Infrastructure Initiative[65] (OMII-UK) funded the Security Portlets simplifying Access and Management of Grid Portals[66] project (SPAM-GP), which developed tools that support precisely such requirements.

3.4.4 Scoping of Trust and Protecting of Resources: Finer Grained Security of Portals

Figure 1 illustrates how a given Grid portal can be accessed through Shibboleth. However, it is essential for many domains that access to such resources is restricted further. Simply knowing that someone has authenticated at the University of Glasgow will typically not be sufficient for access control to clinical data sets that might be available through the portal for example. To support this finer-grained scoping of trust and associated authorisation, the SPAM-GP project implemented a variety of JSR168 compliant portlets

[65] http://www.omii.ac.uk
[66] http://www.nesc.gla.ac.uk/projects/omii-sp

that a portal administrator could apply to support finer-grained security of their portals and the portal content, e.g. portlets that give access to remote VO-specific resources, in a Shibboleth-environment.

The first such portlet developed was the SCAMP (Scoped Attribute Management Portlet). This portlet allows restricted and syntactically correct manipulation of the Attribute Acceptance Policy (AAP) of a Shibboleth SP to streamline the subset of IdPs from whom a portal will accept user attributes. Thus if a collaboration only involves a subset of organizations in the UK federation, then SCAMP allows to limit access to the portal to that subset of sites. To achieve this, the portlet parses the federation metadata for the list of all the IdPs within the federation, and stores the values of the 'scope' entry for each IdP.

When the SP is provided with a scoped attribute, the suffix will by definition be one of these scoped values. The list of IdP scopes in the federation is provided to the user/portal administrator in the form of a drop down list, one per user attribute, where the institutions from whom attributes are to be recognized/accepted from may be selected. The first time the portlet runs, the policy will set all attributes to 'scoped' but with no scope defined, so the default behaviour will be to accept attributes from no institutions – a default common with most security infrastructures, i.e. deny all. Subsequently collaborating sites can be iteratively added to build a VO at the attribute level by the portal manager. Once defined, these changes can then be added to the AAP file. This policy information will then subsequently be available for the next browser session referencing that resource, i.e. only allowing access to the resources from known and trusted sites with expected attributes.

A Content Configuration Portal (CCP) was developed to allow the dynamic configuration of contents of a portal based upon presented user privileges that are presented from Shibboleth. Intuitively, CCP allows users to access and use client portlets that are associated with their given privilege sets as defined by a portal administrator. An example of application of the CCP in the neurological domain is shown in Figure 2 where the portlet on the right is for researchers with a *brainIT-investigator* role and the portlet on the left for researchers with a *brainIT-nurse* role. The definition of these roles and their association with specific portlets (giving access to remote services/data), and finally their dynamic configuration through information presented by Shibboleth is at the heart of the CCP.

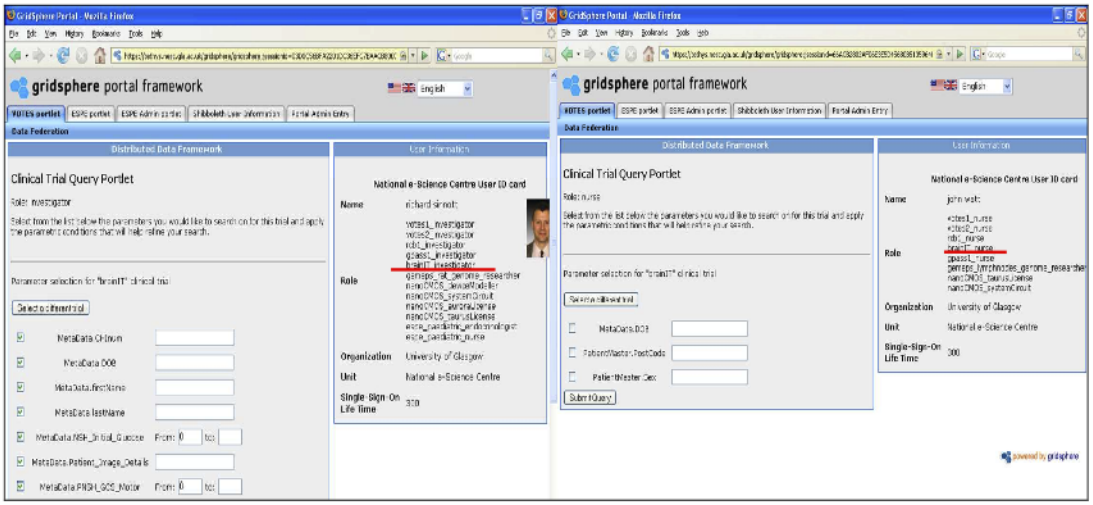


Figure 3.3: Example of CCP Application in the Neurological domain

The SPAM-GP Attribute Certificate Portlet (ACP) is a JSR-168 compliant portlet that supports sites wishing to define and enforce their own remote authorization. Thus it is unlikely that a resource provider will simply delegate their security and access control to a potentially remotely managed portal. ACP allows the

security information (attributes) to be signed and stored in a local attribute store for use by remote providers in enforcing their own local authorization policies. Currently the ACP has been applied successfully to a range of distributed resources and resource providers including those in the clinical domain through projects such as the Medical Research Council (MRC) funded Virtual Organisations for Trials and Epidemiological Studies[67] project (VOTES), the social science domain through projects such as the Economic and Social Science Research Council (ESRC) funded Data Management through e-Social Science[68] project (DAMES) and the geospatial domain through projects such as the Joint Information Systems Committee (JISC) funded Secure Access to Geospatial Services[69] project (SeeGEO).

3.4.5 Summary

Employment of an adequate trust and security model is crucial to e-Research. It must provide researchers with easy access to shared heterogeneous, distributed resources while ensuring adequate protection of those resources. Some disciplines place more emphasis on security, for instance the medical field where patient confidentiality is paramount, but all disciplines undertaking e-Research involving distributed systems require a model that takes into account issues of single sign-on authentication and authorisation which should apply to all computational and data resources. In this chapter, we have presented examples of models that work well to address these issues of resource sharing. These models are now being used in a variety of diverse large scale projects including major European clinical trials; nanoCMOS electronics; social sciences and the arts and humanities amongst others. More information on these solutions and the work of NeSC at Glasgow is available at www.nesc.ac.uk/hub[70] .

3.4.6 Acknowledgements

This work was funded by a variety of grants from the EPSRC, ESRC, JISC and the MRC. We gratefully acknowledge their support. More information on NeSC at Glasgow and the security-oriented projects that they are involved in is available at www.nesc.gla.ac.uk/projects[71] .

3.4.7 References/Further Reading

Watt, J., Sinnott, R.O., Ajayi, O., Jiang, J. and Koetsier, J. (2006). A Shibboleth-Protected Privilege Management Infrastructure for e-Science Education[72] . *6th IEEE International Symposium on Cluster Computing and the Grid*, CCGrid2006, May, Singapore.

Sinnott, R.O., Watt, J., Chadwick, D.W., Koetsier, J., Otenko, O., and Nguyen, T.A. (2006). Supporting Decentralized, Security focused Dynamic Virtual Organizations across the Grid[73] . *2nd IEEE International Conference on e-Science and Grid Computing*, Amsterdam, December.

Sinnott, R.O., Watt, J., Ajayi, O. and Jiang, J. (2006). Shibboleth-based Access to and Usage of Grid Resources[74] . *IEEE International Conference on Grid Computing*, Barcelona, Spain, September.

Sinnott, R.O. (2008). Grid Security. In Wang, L, Jie, W. and Chen, J. (eds.) *Grid Computing: Technology, Service and Application*[75] , CRC Press, May.

Sinnott, R.O., Chadwick, D., Doherty, T., Martin, D., Stell, A., Stewart, G., Su, L. and Watt, J. (2008). Advanced Security for Virtual Organizations: Exploring the Pros and Cons of Centralized vs Decentralized

[67] http://www.nesc.gla.ac.uk/projects/votes
[68] http://www.dames.org.uk
[69] http://edina.ac.uk/projects/seesaw/seegeo/
[70] http://www.nesc.ac.uk/hub
[71] http://www.nesc.gla.ac.uk/projects
[72] http://dx.doi.org/10.1109/CCGRID.2006.11
[73] http://dx.doi.org/10.1109/E-SCIENCE.2006.261106
[74] http://dx.doi.org/10.1109/ICGRID.2006.311008
[75] http://www.crcpress.com/product/isbn/9781420067668

Security Models[76] . *8th IEEE International Symposium on Cluster Computing and the Grid* (CCGrid 2008), May, Lyon, France.

Wei, J., Arshad, J. and Sinnott, R.O. (2010). A Review of Grid Authentication and Authorization Technologies and Support for Federated Access Control, to appear in ACM Computing Surveys, January.

[76] http://dx.doi.org/10.1109/CCGRID.2008.67

Chapter 4

Using Distributed Systems in Research

4.1 Portals[1]

4.1.1 Portals

4.1.1.1 Introduction

A portal (or Web portal) presents information from diverse sources in a unified way. A Web site that offers a broad array of resources and services, such as e-mail, forums, search engines, online shopping are referred to as portal. The first portals grow out of online services, such as AOL, and provided access to the Web, but by now most of the traditional search engines have transformed themselves into Web portals to attract and keep a larger audience. Apart from the basic search engine feature, these portals often offer services such as e-mail, news, stock prices, information, and entertainment.

Portals provide a way for enterprises, research and other communities to generate a consistent look and feel with access control and procedures for multiple applications, which otherwise would have been different entities altogether. In a research environment a portal integrates online scientific services into single Web environment which can be accessed and managed from a standard Web browser. The most remarkable benefit of portals is that they simplify the interaction of users with distributed systems and with each other, because a single tool – the browser – and a standard and widely accepted network protocol – HTTP – can be used through all communications.

After the proliferation of Web browsers in the mid-1990s many companies tried to build or acquire a portal, to have a piece of the Internet market. The Web portal gained special attention because it was, for many users, the starting point of their Web browser. Similarly, but a bit later, research communities recognized the value of Web portals in integrating various services into coherent, customizable environments. Research collaborations began developing portals in the late 1990. These environments can be broadly categorized as horizontal portals, which cover many areas, and vertical portals, which are focused on one functional area. Horizontal research portals often provide services that are independent from any scientific discipline and represent generic functionalities that are common across disciplines. Vertical portals target specific group of researchers that are involved in the same experiment or work within the same scientific field.

4.1.1.2 Portals for distributed science

In the simplest form a scientific portal is a collection of links to external Web pages and Web services that are scattered on the Internet and aimed to serve scientists with similar interest. These portals often include search engines that are customized for the interest of the user community, e.g. for publications, job positions, news from a scientific domain. Various research communities, ranging from mathematics to art and

[1] This content is available online at <http://cnx.org/content/m24551/1.1/>.

humanities, all have their own portals, sometimes even multiple portals localized for different geographical regions and languages.

For researchers who perform massive simulations on distributed computing systems, portals mean Web environments that make computing resources and services accessible via Web browsers. Such portals are typically referred to as Grid portals, or science gateways. Grid portals are Web interfaces that are integrated with PC, cluster or supercomputer based computing resources. These environments very often include high level services that are not included in the underlying infrastructure, they are implemented on the portal server instead. Such services can be brokers, load balancers, data replica services, data mirroring and indexing components. All these services together with the front-end portal provide an integrated solution that enables e-scientists to deal with various aspects of data-intensive research. Usage scenarios in such portals often involve the generation and storing of research data, analysis of massive datasets, drawing scientific conclusions and sharing all these entities with colleagues or with the general public. Scientists of these portals tend to organize their frequently used steps of simulations into reusable components, called workflows. Workflows provide repeatable, traceable experiments and improve both the productivity and quality of research.

Hardware resources, software services, applications and workflows that are made accessible through Grid portals are typically provided by multiple independent organizations and are managed by different administrators. Portals connect to these distributed entities in a service oriented fashion, namely through well defined interfaces that expose the functionality of each shared component. For the sake of scalability and fault tolerance these services are accessed by the portal through some catalogues. A typical difficulty of portal design is how to integrate the content from the dynamic services of the catalog into a user friendly view that is ergonomic, provides coherent information, and at the same time flexible and easily customizable for different user preferences.

In the early years of Grid computing the portal systems were implemented using various different programming approaches and languages. Despite Grid solutions have much simplified since those early years, developers of science gateways must still know relatively high number of technologies. Portal operators must closely follow the evolution of Grid middleware services, because updates on the portal are required when the Grid middleware is changed. While enterprise portals also demand regular updates due to changes in the back-end systems, the technological evolution in Grid computing is more rapid than in business environments. As grid portals shield users away from middleware evolution, they are very often the only way for scientists to stay connected with Grids over longer period of time.

There are a few other technical difficulties that are specific to Grid portals. The issue of how to map the security model of Grid systems to the security model of the "Web" is one of these. While Web portals identify and authenticate users with account-password pairs, Grids and other sorts of distributed computing environments use certificates. Certificates enable users to authenticate only once in a distributed system and perform complex operations without being asked for account details over and over again (e.g. in every stage of the orchestration of a workflow). By certificates the users can delegate their access rights to workflow managers, brokers, catalogs and other services that perform activities on their behalf. Grid portals typically translate username-password pairs to certificates by either through certificate repositories (such as MyProxy), or by importing certificates from Web browsers.

From the technological perspective a portal is a dynamic Web page that consists of pluggable modules, called portlets. These portlets run in a portlet container. The container performs basic functionalities such as management of system resources and authenticating users, while portlets generate the actual Web interface. Portal developers realized that reusability of portlets is a key to the customizability of portals, and the interoperability of portlets across different container platforms in an important step towards this. Around 2001 the Java Specification Request (JSR) 168 emerged as a standard that allows portal developers, administrators and consumers to integrate standards-based portals and portlets across a variety of portal containers. Most of current science gateways are built from JSR-168 compliant portlets and provide easily customizable and reusable solutions for various purposes.

4.1.1.3 Grid portal examples

The usefulness of Grid portal technologies for computational science has been established by the number of portals being developed in Europe, the United States and Asia. In Europe the most relevant portal developer consortiums have gathered around the Enabling Grids for E-sciencE project (EGEE) and its national Grid counterparts. Some of these portals provide tools that are independent from scientific disciplines, others emphasize solutions that specific communities are familiar with and can utilize efficiently.

P-GRADE Portal provides facilities to create and execute computational simulations on cluster based Grids. Various user communities of the EGEE Grid and several European national Grids apply P-GRADE Portal as a graphical front-end to manage workflow applications on their infrastructures. While P-GRADE Portal is primarily a generic environment, it can be customized to any scientific domain by generating application specific portals from it that grants access only to pre-defined, domain specific workflows and simulations.

The NGS Applications Repository is an open access portal used to describe and list applications and their associated artefacts that are available on the National Grid Service (NGS) of the UK. Applications hosted by the repository are described using middleware agnostic documents, which can be searched for by categories of interest. The repository currently holds over 50 applications from various fields such as bioinformatics, engineering, chemistry, astrophysics or image analysis.

4.1.1.4 Outlook

The concept of content aggregation seems to still gain momentum and portal solution will likely continue to evolve significantly over the next few years. The Gartner Group recently predicted to expand on the Business Mashups concept of delivering a variety of information, tools, applications and access points through a single mechanism. Mashups are Web applications that combine data or functionality from two or more sources into a single integrated application. The term mashup implies easy, fast integration, frequently done by access to open programming interfaces and data sources to produce results that were not the original reason for producing the raw source data. An example of a mashup is the use of cartographic data from Google Maps to add location information to real estate data, thereby creating a new and distinct Web service that was not originally provided by either source.

Programmers of Grid and high performance computing portals still often find it hard to bridge between user friendly Web interfaces and low-level services of Grid middleware. Errors and faults sent back from the Grid are often difficult to interpret and deal with automatically, meanwhile it is inevitable that easy to use and autonomous portals are important tools to attract larger user communities to Grids. Grid portals will definitely improve in the near future in this respect.

4.1.1.5 References

- M. Thomas, J Burruss, L Cinquini, G Fox, D. Gannon, I. Glilbert, G. von Laszewski, K. Jackson, D. Middleton, R. Moore, M. Pierce, B. Plale, A. Rajasekar, R. Regno, E. Roberts, D. Schissel, A. Seth, and W. Schroeder. Grid Portal Architectures for Scientific Applications. Journal of Physics, 16, pp 596-600. 2005.
- "Web portal" entry in Wikipedia: http://en.wikipedia.org/wiki/Web_portal[2] , last accessed 06/06/2009
- Enabling Grids for E-sciencE project (EGEE): http://www.eu-egee.org/[3]
- P. Kacsuk and G. Sipos: Multi-Grid, Multi-User Workflows in the P-GRADE Portal Journal of Grid Computing, Vol. 3, No. 3-4, Springer Publishers, pp. 221-238, 2005.
- NGS Job Submission Portal: https://portal.ngs.ac.uk/[4] , last accessed: 06/06/2009

[2] http://en.wikipedia.org/wiki/Web_portal
[3] http://www.eu-egee.org/
[4] https://portal.ngs.ac.uk/

- "Mashup" entry in Wikipedia: http://en.wikipedia.org/wiki/Mashup_(web_application_hybrid)[5] , last accessed 06/06/2009

4.2 Visualization Matters[6]

Key Concepts

- Themes in the science of visualization
- Simulation models
- Visualization tools – graphs created using Excel and MATLAB
- Distributed visualization
- Metadata and Paradata for scientific visualization

4.2.1 Introduction

"We don't see with our eyes. We see with our brains", Paul Bach-y-Rita.

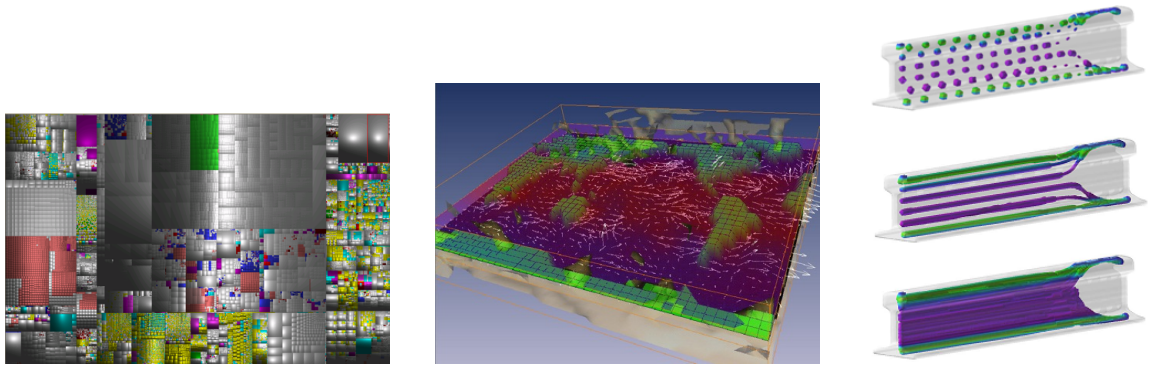
In the last thirty years computer-based visualization has moved from an informal *ad hoc* tool designed to create particular results, to becoming a proper science in its own right. Universal generalisations and specifications as well as best practice guidelines are now available. Visualization methods are now being studied as an individual topic within various courses and modules; at all levels from undergraduate to postgraduate. Visualization is now the basis of numerous PhD titles and further research projects and programmes, funded across all the research councils and the infrastructure HE/FE funding agencies. This research and development has created a large toolkit for general use as well as individual methodologies for specialist user data sets, and has helped in understanding the barriers between the computer display and the human visual system. Visualization, it should be emphasised, is as much about gaining investigative insight as it is about enhancing presentations to tell a clearly specified story.

The science of visualization has been split into three themes; information visualization that studies methods for the representation of large-scale collections of often non-numerical information as well as the recommendations for use of graphical techniques to aid in the analysis of data. Scientific visualization, the second theme, was developed from previous often natural and experimental methods of displaying data, which has seen an explosion of users due to the deluge of in-silico experimental data (e.g. supercomputing and high throughput computing results) as well as real experimental capture equipment (e.g. 3D medical scanners, climate sensor data and astrophysical telescopes). Results often mimic reality, for example creating virtual wind-tunnel visualizations, but can be abstract, for example visualizing 6-dimensional tensor components using different geometric shapes (as in Figure 1). Visual analytics is the third theme. This merges both of these fields to focus on the user's analytical reasoning, which often involves interactive visual interfaces and commonly employs various data-mining techniques as well as combining data across different databases.

This chapter introduces examples within these visualization themes, first providing an overview of simulation models and then specific examples from the creation of graphs using popular tools such as Excel and MATLAB. It then moves on to present the complexities of distributed visualization, as well as the role of adding metadata and paradata.

[5] http://en.wikipedia.org/wiki/Web_portal

[6] This content is available online at <http://cnx.org/content/m31926/1.1/>.



(a) (b) (c)

Figure 4.1: Visualization examples: information visualization example showing the content of the $\frac{1}{2}$ million files on my hard disc (Sequoiaview[7]); and two scientific visualizations, the first showing climate modelling using various animated glyphs to show flow strength (Avizo[8]); and the second a selection of interactive superquadric glyphs selecting various forms from the six dimensions available within tensor stress components (AVS/Express[9]).

4.2.2 The Human Visual System: The User is Key.

Will Schroeder et al. in The Visualisation Toolkit (Schroder et al. 1998) stated *"informally visualisation is the transformation of data or information into pictures. Visualisation engages the primal human sensory apparatus, vision, as well as the processing power of the human mind. The result is a simple and effective medium for communicating complex and/or voluminous information."* Based upon using the massive amount of brain power within the human visual system that constitutes about 1/3 of the total brain size, visualizations have been shown to be one of the best and sometimes the only way of conveying a huge amount of data in a short period of time. One of the key reasons for visualization as a specific field to study was the rapid increase in quantity of data being produced by simulations on supercomputers of physical, natural and theoretical problems. This has been termed as the data-deluge problem and frequently has been so large that graphical representations offer the only viable way to assimilate the data.

The simulation models themselves have also been increasing in complexity, involving large numbers of independent and dependent variables whose relationships need to be understood. For example, in climate modelling, we may wish to explore how temperatures, water vapour content, pressure, wind directions and velocities vary within a 3D region, over time and all at once. The process of visualization is therefore concerned with ways to represent the data as well as defining tools for interactively exploring the multidimensional and multi-variant models. One of the early active research areas was to find ways to link this visualization process with interactive control of the simulations themselves, opening up completely new possibilities for interactive exploration and understanding of complex phenomena. Over the years a number of visualization systems have emerged, which provide a framework for this kind of model exploration.

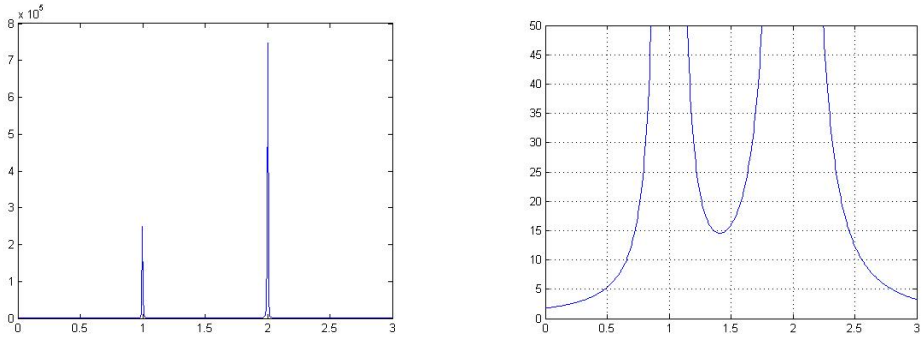
4.2.3 Visualization Tools: Evaluating a Graph

Plenty of literature and course notes are now available but as a simple example, a few rules are presented next on how to create an effective graph. A graph should present a reasonable amount of data, say something

[7] http://www.win.tue.nl/sequoiaview/
[8] http://www.vsg3d.com/
[9] http://www.avs.com/

about the behaviour of that data and it should avoid giving a false impression of the data. In other words, the graph must communicate something. Tukey (1977) pp 128,157 said *"there cannot be too much emphasis on our need to see behaviour. Graphs force us to note the unexpected; nothing could be more important"*.

Excel and MATLAB are two of the most popular visualization tools currently used, even though users may not consider them as such. They produce numerous 2D and 3D graphs of different sizes and dimensions but the visualization choices are rarely thought about. Figure 2 shows two views using MATLAB of a simple formulae ($y = (x\text{-}1)^{-2} + 3(x\text{-}2)^{-2}$). Both show the same numerical sampled data but the second, by cropping the y-axis to a limited range [0:50], could be said to present a large amount of extra information highlighting an important area. This process has been termed focus and context zoom interaction.

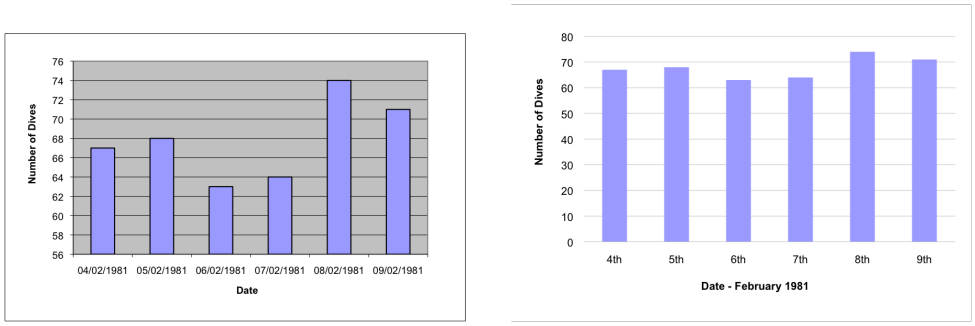


(a) (b)

Figure 4.2: The default and a cropped version of the numerical evaluation of the MATLAB plot command for the formulae ($y = (x\text{-}1)^{-2} + 3(x\text{-}2)^{-2}$).

Users make choices about the data to be used and its visualization, and these affect both the quality and the quantity of the information presented. Good visualization requires graphical integrity and there are many standard techniques to help quantify and qualify between different versions. As a short exercise three well used simple objective tests (adapted from Tufte (2001)) are presented here as applied to the data shown in Figure 3.

- *Objective Test 1:* The **Lie Factor** emphasises the variation in the data which can cause misleading interpretations. The variation in height of the smallest and largest bars in the graph on the left is (74 - 56) / (63 - 56) = 2.57; however, the variation in the data is 74 / 63 = 1.17. These two numbers being different indicate how visually the variations appear more extreme on the left-hand graph.
- *Objective Test 2:* The **Data Ink** represents the non-erasable items of a graphic and often represents the non-redundant ink. The horizontal grid lines, tick marks and the frame around the graph are all erasable – and can be, within reason, as they may distract more than guide the observer.
- *Objective Test 3:* The **Data Density** represents the ratio defined as the number of entries in the data matrix divided by the area of the data graphic. In this case by removing frames the data density of the right-hand graph slightly increases.



(a) (b)

Figure 4.3: An Excel bar chart example showing the statistics of the number of dives for a Female Elephant Seal in early February 1991 (numbers and example adapted from Tufte 1997). The graph on the left is the default format.

It is possible and recommended to try these kinds of tests and many others on any images including those found in national newspapers.

4.2.4 Distributed Visualization: Massive Datasets

The visualization of large datasets has become a new key bottleneck in applications where validation of results and data acquisition from scientific equipment is required at an early stage. Such validation would allow correctness of methods (such as the set up of a physical experiment) to be determined prior to further spending of computational or imaging machine resources. Datasets can far exceed the capabilities of modern graphics hardware (GPUs) and so visualization systems are turning to parallel compute facilities to render them.

Figure 4 shows a use case of a current system being developed. Here multiple render processes are executed to render small sections of a decomposed dataset (right hand side). In this case the GPU output from each render process is visible; although usually these windows are not visible and only the left hand composite image is shown. However, this conveys the idea of distributed rendering with the final composited image, shown on the left, viewable by the user. This final real-time interactive image can be transmitted across the internet at fast rates (experience is about 15 frames per second across countries), to be displayed within an application window (as shown), within a portal, or within a Virtual Learning or Research Environment. There is no current national based visualization service in the UK; but various test services exist within JISC and research council funded projects, including on the National Grid Service (NGS http://www.ngs.ac.uk/[10]) and two initiatives currently running on the UK national supercomputing service (HECToR http://www.hector.ac.uk/[11]) are leading the way.

[10] http://www.ngs.ac.uk/
[11] http://www.hector.ac.uk/

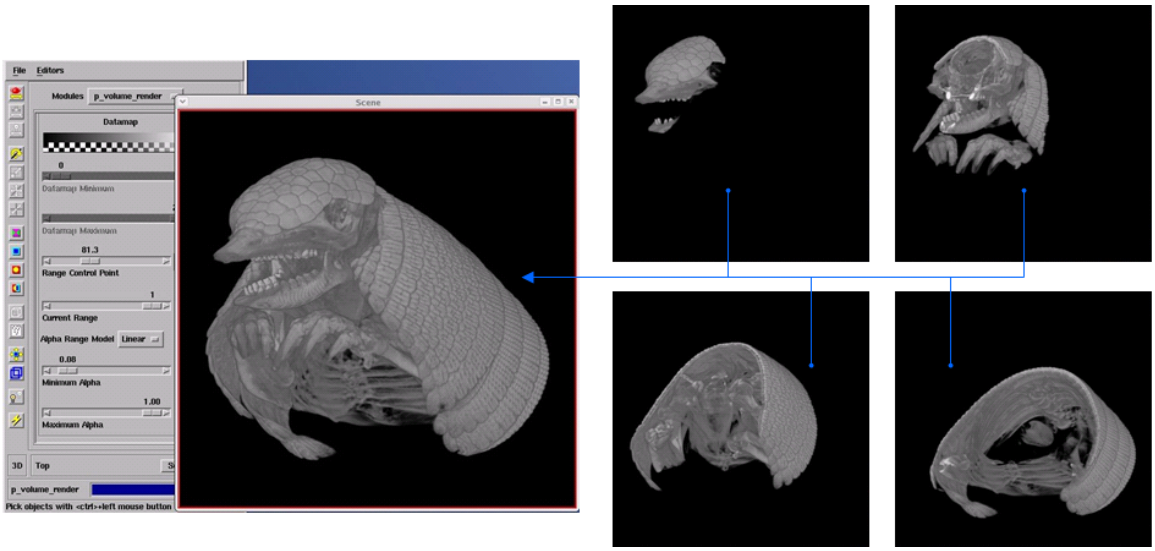


Figure 4.4: End-user volume viewer application (left) displays a composited image from raycasting volume rendering processes running in parallel on the four cluster nodes (right). AVS/Express pre-test version for the MRBV (Massive Remote Batch Visualizer) project running on a CRAY XT4[12].

4.2.5 Making Choices: Metadata and Paradata

Rules can be broken with the addition of appropriate metadata, and this has been known for a long time. The addition of good metadata including all forms of annotations is very important even if it takes time and careful thought. Metadata can include all details describing the source of the data, the methods used to pre-manipulate the data and create the visualization, as well as the contact details of the author, creation date etc. Recently there have been tools developed to help record this process. These include the development of software within e-Science, creating a set of middleware linking computing resources together – adding semantic tags which define meaning to these components – and creating ontologies, which describe how human terms relate to computer terms.

A proposal is to add paradata that extends the concept of metadata to consider issues of choice and alternatives by recording the subjective decisions. For example, Figure 5 shows fourteen different visualization variations for a simple vortex fluid flow data set. Often only one or two images will be used to illustrate a specific scientific phenomenon, but it is very rarely considered in detail what decisions have been make and it is even rarer for these decisions to be written down, as to why a particular version has been chosen. The use of paradata would now allow and even force the authors to describe the reasons for their choices.

It is said that an image is worth a thousand words, but we can rephrase this to say a good visualization may need a thousand words of annotations, in both metadata and paradata, in order to properly describe it.

[12] http://www.hector.ac.uk

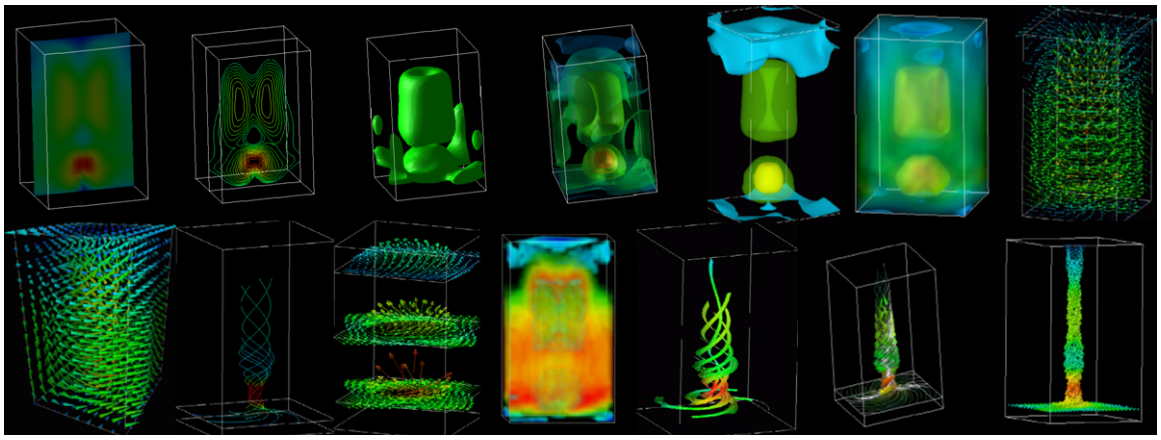


Figure 4.5: Fourteen different versions of scientific visualizations for the same data flow field (McDerby 2007).

A couple of solutions to address visualization are the introduction of recordable and shareable workflows (myExperiment[13]), and the controlled recording of researchers' choices creating a visualization provenance (VisTrails[14]). These and similar tools are going to be more available within VREs (Virtual Research Environments[15]) that are already considering the use of collaborative environments; including an emphasis on the web 2.0 generic principles of being able to store and annotate everything.

4.2.6 Conclusions: "Lying" with Visualizations

They always say you can lie with statistics, but similarly you can lie with visualizations as well. This is especially true as visualizations not only can be selective in choice of data, but as they employ the human visual system they can create visual illusions as well. Often this process is not deliberate but is accidentally misleading, caused by authors who only have space for a few visualizations and make quick, possibly uninformed, decisions.

We have presented a few very simple examples to describe how small changes can improve the presentation of information. Also we have given a warning that without defining and describing the choices made, through metadata and possibly paradata, there can be confusion. Fortunately there are now methods, just starting to be introduced, to help in the process, although more need to be actively used, tested and developed.

4.2.7 Acknowledgement

At the University of Manchester, Research Computing Services, starting with the Computer Graphics Unit, has for over 30 years been considering how to efficiently create and present visual stimuli and is still learning the best way to integrate and transfer information from computer source to the human user. Thanks to all those who indirectly have contributed ideas to this short article from numerous sources (including the MSc module taught at Manchester[16]). It is recommended that readers explore the topic further as this barely

[13] http://www.myexperiment.org/
[14] http://www.vistrails.org
[15] http://www.jisc.ac.uk/whatwedo/programmes/vre1.aspx
[16] http://wiki.rcs.manchester.ac.uk/community/Viz_for_HPC

covered the topic. Dr Martin J. Turner, the University of Manchester, and part of the JISC funded national vizNET[17] support network, Martin.Turner@manchester.ac.uk.

4.2.7.1 References

Schroeder, W., Martin, K. and Lorensen, B. (1998) The Visualization Toolkit Prentice Hall 1998 2nd Edition

Tukey, J. W. (1977) Exploratory Data Analysis. Addison-Wesley, Reading, MA

Tufte, E.R. (2001) The Visual Display of Quantitative Information Graphics Press, Cheshire, Connecticut 2nd Edition

Tufte, E.R. (1997) Visual Explanations: Images and Quantities, Evidence and Narrative Graphics Press Cheshire, Connecticut

4.3 Virtual Research Environments[18]

4.3.1 Introduction

e-Research is, by definition, a collaborative activity that combines the abilities of distributed groups of researchers in order to achieve research goals that individual researchers or local groups could not hope to accomplish. Very often, e-Research is also multidisciplinary, spanning not only geographical and organisational boundaries but also disciplinary ones. There is hardly any kind of research that does not make use of electronic resources of one kind or other and in some disciplines ICTs play such a central role that without them, the advancement of research would not be possible.

The notion of a virtual research environment has gained prominence in the e[U+2011]Research community (Fraser 2005, Borda *et al.* 2006). For all practical puposes, the term is synonymous with other concepts such as collaboratories, cyberenvironments or science gateways that are used in the US and elsewhere (*cf.* Olson, Zimmerman and Bos 2008, Wilkins-Diehr 2007). The aim of a VRE is to provide an *integrated* environment that supports the work of a community of collaborating researchers. That is, a VRE brings together previously separate tools needed for conducting the research and for *collaboration*, support for which is increasingly recognised as an integral aspect of researchers' work rather than something that can be added on as an afterthought.

4.3.2 Providing Rich Functionality

Behind the scenes, a VRE makes use of a set of services proving secure access to various kinds of resources such as datasets, large-scale storage facilities and computational facilities for execution of scientific codes. The resources used are distributed, they are provided by different organisations and under different policies governing their usage. Therefore, the infrastructure needs to support their management by providing, for example, appropriate authentication and authorisation mechanisms to ensure that only authorised individuals access files and that computational resources are accessed with the correct credentials. During the execution of a scientific application, intermediate data and runtime information is created that may be retained to provide a provenance record and simulation outputs are stored in a storage system. For example, researchers might want to:

- *authenticate* using an authentication service,
- *communicate* and *collaborate* with colleagues,
- *transfer* data,
- *configure* a resource,
- *invoke* a computation,
- *re-use* data and *give credit* to the original producer,
- *archive* output data and runtime data,

[17] http://www.viznet.ac.uk/

[18] This content is available online at <http://cnx.org/content/m32637/1.1/>.

- *publish* outputs, both informally through blogs or wikis and formally through conference or journal papers,

- *discover* what resources are available,
- *monitor* the state of a resource or process,
- *maintain awareness* of who is currently doing what,
- find out where particular data has come from and how it was processed (*provenance*),
- find out who has access to a resource and what they can do with it (*authentication* and *authorisation*).

The list above is not meant to be a comprehensive list but it gives an indication of the type of functionality many VREs will contain and what sorts of interfaces they have with other systems and services. The scope of virtual research environments tends to be defined to encompass activities such as project management and research administration (cf., e.g., Borda et al. 2006, p. 3) and their relevance is therefore not limited to research active academics but extends to other professionals in the research context such as administrative or library staff. It is this breadth of vision that sets VREs apart from earlier concepts that were much more focused on solving specific problems arising from particular scientific endeavours.

A range of VREs have been funded by the UK's Joint Information Systems Committee (JISC), which has funded a series of research programmes in this area (http://www.jisc.ac.uk/whatwedo/programmes/vre2.aspx). The VREs funded range from exemplars and demonstrators to institutional implementations and generic tools as well as supporting projects. Further examples of e-Research and VREs can be found in the literature, for example, in Foster and Kesselman (2004), Berman, Fox and Hey (2003) and Olson, Zimmerman and Bos (2008). Examples of VREs are mentioned in other chapters of the book and you can find videos in the Resources section of the online version.

4.3.3 Infrastructure and Communities

As the technologies mature, the focus shifts from the technical problems of distributed computation to the embedding of these technologies into organisational settings, into arrangements within research communities and into the wider societal context. It has been pointed out that infrastructures for research need to be seen as socio-technical arrangements (e.g., Edwards *et al.* 2007) and that the use of advanced ICTs is not just limited to "big science" endeavours in a small number of disciplines but, rather, is starting to affect research activities across the board.

Most research data today exists in digital form, either because it is created digitally ('born digital') or through digitisation programmes (*ibid.*). The concept of 'content as infrastructure' emphasises the increasing importance of collections of research data as a re-usable infrastructure that builds on top of the physical research computing infrastructure and traditional infrastructures such as scientific instruments or libraries.

As the name virtual research *environment* implies, the aim is not to build single, monolithic systems but rather socio-technical configurations of different tools that can be assembled to suit the researchers' needs without much effort, working within organisational, community and wider societal contexts. The concept of a VRE suggests the seamless integration of resources needed by researchers throughout the lifecycle of a research undertaking. While current VRE implementations are difficult to customise by the individual researcher to meet their specific research needs, there is a trend to provide environments that allow for the dynamic configuration and assemblage of research tools.

In any VRE, there are some components that are generic and can potentially be used by researchers in many disciplines. A wide range of commoditised components and systems are available and efforts are underway to develop interoperability frameworks to foster flexible integration to form seamless collaborative work environments (Voss *et al.* 2007). This offers the opportunity for reuse on a large scale and thereby to avoid duplication of effort where supporting tools already exist. For example, synchronous and asynchronous collaboration support can be provided through integration of tools like instant messaging, Access Grid (www.accessgrid.org), EVO (evo.caltech.edu), wikis, blogs, feeds etc. Likewise, generic tools for the

management of job submissions to computational grids or for the management of data in storage resource brokers exist and are quite mature and stable. Their general applicability leads to wide support for their development and, consequently, it does not make sense to re-invent them.

In addition, re-use of components also facilitates the re-use of skills on the side of technology and service providers as well as on the side of the end user. If every environment came with its own authentication system, for example, this would hinder uptake significantly, so re-use of common solutions such as Shibboleth supported by a large access management federation is clearly important. In effect, the tools and services making up a VRE should become part of the seen-but-unnoticed e-Infrastructure that enables researchers to collaborate with their peers easily and without having to pay much attention to the technology.

More specific support, however, for the management and conduct of specific research tasks will require configuration and adaptation of tools as well as the development of new ones. In order to maximise reuse in the light of heterogeneous requirements, a modular approach is needed that leaves the user in charge of managing their research environment and provides support for this through automated processes such as service and tool discovery. Inevitably, there will be a point when technical support staff will need to intervene. These interventions should be supported by the system in a way that enables users to learn and become more independent in the future.

The nature of research means that VREs will require constant adaptation to fit the specific research projects being undertaken. As research thrives on a constant modification of its practices, it is likely that new functionality will be required that is not already available. At the same time, economic pressures and the fact that some aspects of research are routine mean that existing functionality needs to be made use of and adapted wherever possible. What is required is a close collaboration between researchers on the one hand and technology and service providers on the other to establish configurations of technologies and social arrangements that allow the researchers to focus on the innovative aspects without having to concern themselves with technical details.

Creating an integrated e-Research experience fundamentally relies on the creation of communities of service providers, tool builders and researchers working together to develop specific support for research tasks as well as the creation of a technical and organisational platform for integrating these tools into an overall research process.

4.3.4 Conclusions

The vision of e-Research argues persuasively for the need to generate, keep and re-use an expanding range and volume of research resources. VREs are crucial to e-Research because they are the sites where these resources will be both consumed and created. VREs are also the sites of experiments in scholarly communications, which have the potential to transform their conduct.

4.3.5 References

Berman, F., Fox, G. and Hey, T. (eds.) (2003), *Grid Computing: Making the Global Infrastructure a Reality*. John Wiley and Sons.

Borda, A. *et al.* (2006), "Report of the Working Group on Virtual Research Communities for the OST e-Infrastructure Steering Group", London, UK, Office of Science and Technology, available at http://www.nesc.ac.uk/documents/OSI/vrc.pdf[19] (accessed 12.02.2009)

Brown, S. and Swan, A. (2007), "*Researchers' Use of Academic Libraries and their Services*". A report commissioned by the Research Information Network and the Consortium of Research Libraries. Available at: http://www.rin.ac.uk/researchers-use-libraries[20]

Edwards, P.N., Jackson, S.J., Bowker, G.C. and Knobel, C.P. (2007), *Understanding Infrastructures: Dynamics, Tensions, and Design*. Report of a Workshop on "History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures", National Science Foundation.

[19] http://www.nesc.ac.uk/documents/OSI/vrc.pdf
[20] http://www.rin.ac.uk/researchers-use-libraries

Foster, I. and Kesselman, C. (eds.) (2004). *The Grid: Blueprint for a New Computing Infrastructure*, 2nd edition, Morgan Kaufman Publishers.

Fraser, M.A. (2005), Virtual Research Environments: Overview and Activity, *Ariadne*, No. 44, July 2005.

Olson, G.M., Zimmerman, A. and Bos, N. (2008), *Scientific Collaboration on the Interent*, MIT Press.

RIN (2007), *Research and the Scholarly Communications Process: Towards Strategic Goals for Public Policy*. Available at http://www.rin.ac.uk/sc-statement

Voss, A., Procter, R., Budweg, S. and Prinz, W. (2007), "Collaborations in and for e[U+2011]Research: making the 'O' in virtual organisation work", Proceedings of the German e-Science Conference, Baden Baden, May 2007.

Wilkins-Diehr, N. (2007), Special Issue: Science Gateways – Common Community Interfaces to Grid Resources, *Concurrency and Computation: Practice and Experience*, Vol. 19, pp.743-749.

Chapter 5

Resources

5.1 Examples of e-Research Videos - from the eIUS project[1]

NOTE: The videos in this section have been produced by the e-Infrastructure Use Cases and Service Usage Models project[2] and are licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 2.0 UK: England & Wales License[3] .
External Image
Please see:
http://i.creativecommons.org/l/by-nc-nd/2.0/uk/88x31.png
[4]

5.1.1 Bioinformatics

This media object is a Flash object. Please view or download it at
<http://www.youtube.com/v/Y6_Kz5L010g&hl=en&fs=1&rel=0>

5.1.2 Astronomy

This media object is a Flash object. Please view or download it at
<http://www.youtube.com/v/UDPy7-UHCOY&hl=en&fs=1&rel=0>

5.1.3 Archaeology

This media object is a Flash object. Please view or download it at
<http://www.youtube.com/v/LxZcI0ikKV0&hl=en&fs=1&rel=0>

5.1.4 Earth Sciences

This media object is a Flash object. Please view or download it at
<http://www.youtube.com/v/HWjADafkji8&hl=en&fs=1&rel=0>

[1] This content is available online at <http://cnx.org/content/m32619/1.2/>.
[2] http://www.eius.ac.uk
[3] http://creativecommons.org/licenses/by-nc-nd/2.0/uk/
[4] http://creativecommons.org/licenses/by-nc-nd/2.0/uk/

5.1.5 nano-Electronics

This media object is a Flash object. Please view or download it at
<http://www.youtube.com/v/W70bz2JDJzw&hl=en&fs=1&rel=0>

5.1.6 Computational Chemistry

This media object is a Flash object. Please view or download it at
<http://www.youtube.com/v/bkbRwOWmiwo&hl=en&fs=1&rel=0>

5.2 Virtual Research Environments - Videos[5]

The JISC VRE programme has produced a number of videos that are available on the JISCMedia channel on Youtube[6] . We will seek to embed these videos directly into this resource. In the meantime, please access them through the links below.

- Virtual Environments for Research in Archaeology [7]
- Study of Documents and Manuscripts[8]
- myExperiment[9]
- Collaborative Research Events on the Web[10]

5.3 e-Research Glossary[11]

In this section we define a number of terms used in the Research in a Connected World book. We have aimed to keep these definitions simple to aid understanding of the core concepts rather than trying to come up with definitive definitions of what are often very complex and ill-defined concepts.

Definition 5.1: Cloud
An infrastructure that makes use of virtualisation technologies to provide on- demand services to users. Clouds provide access to dynamically scalable resources reserved for individual users within a shared infrastructure. Like grids, clouds can be private (in-house), community-specific or public. Access to clouds is often provided through web services (Definition: "Web Service", p. 91).
See Also: Grid (Definition: "Grid", p. 90), Web Service (Definition: "Web Service", p. 91)

Definition 5.2: Grid
A type of distributed computing infrastructure. Grids can exist within organisations (private grids, campus grids) but often span organisational boundaries. The The EGEE grid infrastructure (Section 2.2) is an example of an international grid infrastructure.
See Also: Cloud (Definition: "Cloud", p. 90)

Definition 5.3: Ontology
A description of a set of concepts and their relationships within a given domain. Ontologies can be used as a basis for automatic reasoning.

[5] This content is available online at <http://cnx.org/content/m32636/1.1/>.
[6] http://www.youtube.com/user/JISCmedia
[7] http://www.youtube.com/user/JISCmedia#p/u/6/8s8T7nFSMWw
[8] http://www.youtube.com/user/JISCmedia#p/u/7/Ey7Zp4V69RA
[9] http://www.youtube.com/user/JISCmedia#p/u/8/x83pzMMw7lk
[10] http://www.youtube.com/user/JISCmedia#p/u/9/ceFaFA4WUJc
[11] This content is available online at <http://cnx.org/content/m32631/1.3/>.

Definition 5.4: RDF
Resource Description Framework: a set of specifications for data representation that facilitate the conceptual representation of data by forming subject- predicate-object style expressions. RDF is a core component of the Semantic Web (Definition: "Semantic Web", p. 91)
See Also: Ontology (Definition: "Ontology", p. 90), Semantic Web (Definition: "Semantic Web", p. 91)

Definition 5.5: REST(-ful) Service
A web service (Definition: "Web Service", p. 91) that follows specific architectural principles. More often than not, REST-ful services are implementing the standard HTTP protocol.
See Also: Web Service (Definition: "Web Service", p. 91), HTTP Protocol

Definition 5.6: Semantic Web
The plan to 'semantically annotate' data that is available on the web, thus enabling it to be processed more automatically. An example is that a postal address would not be just written like an address (making it recognisable by a human reader as an address) but also be 'tagged' so that a computer can relate the data to the concept of a postal address.
See Also: RDF (Definition: "RDF", p. 90)

Definition 5.7: Web Service
A service that can be invoked across a network using an machine-to-machine interface that ensures interoperability across system boundaries.
See Also: REST(-ful) service (Definition: "REST(-ful) Service", p. 91), WSDL (Definition: "WSDL", p. 91)

Definition 5.8: WSDL
Web Services Description Language: a language used to describe web service (Definition: "Web Service", p. 91) interfaces, specifying the methods available and their parameters.
See Also: Web Service (Definition: "Web Service", p. 91)

Glossary

C Cloud

An infrastructure that makes use of virtualisation technologies to provide on-demand services to users. Clouds provide access to dynamically scalable resources reserved for individual users within a shared infrastructure. Like grids, clouds can be private (in-house), community-specific or public. Access to clouds is often provided through web services[12] .

G Grid

A type of distributed computing infrastructure. Grids can exist within organisations (private grids, campus grids) but often span organisational boundaries. The The EGEE grid infrastructure[13] is an example of an international grid infrastructure.

O Ontology

A description of a set of concepts and their relationships within a given domain. Ontologies can be used as a basis for automatic reasoning.

R RDF

Resource Description Framework: a set of specifications for data representation that facilitate the conceptual representation of data by forming subject- predicate-object style expressions. RDF is a core component of the Semantic Web[14]

REST(-ful) Service

A web service[15] that follows specific architectural principles. More often than not, REST-ful services are implementing the standard HTTP protocol[16] .

S Semantic Web

The plan to 'semantically annotate' data that is available on the web, thus enabling it to be processed more automatically. An example is that a postal address would not be just written like an address (making it recognisable by a human reader as an address) but also be 'tagged' so that a computer can relate the data to the concept of a postal address.

W Web Service

A service that can be invoked across a network using an machine-to-machine interface that ensures interoperability across system boundaries.

WSDL

Web Services Description Language: a language used to describe web service[17] interfaces, specifying the methods available and their parameters.

[12] http://cnx.org/content/m32631/latest/
[13] http://cnx.org/content/m32631/latest/
[14] http://cnx.org/content/m32631/latest/
[15] http://cnx.org/content/m32631/latest/
[16] http://cnx.org/content/m32631/latest/
[17] http://cnx.org/content/m32631/latest/

Index of Keywords and Terms

Keywords are listed by the section with that keyword (page numbers are in parentheses). Keywords do not necessarily appear in the text of the page. They are merely associated with that section. *Ex.* apples, § 1.1 (1) **Terms** are referenced by the page they appear on. *Ex.* apples, 1

Attributions

Collection: *Research in a Connected World*
Edited by: Alex Voss, Elizabeth Vander Meer, David Fergusson
URL: http://cnx.org/content/col10677/1.12/
License: http://creativecommons.org/licenses/by/3.0/

Module: "Welcome"
By: Malcolm Atkinson, David De Roure
URL: http://cnx.org/content/m32854/1.1/
Pages: 1-2
Copyright: Malcolm Atkinson, David De Roure
License: http://creativecommons.org/licenses/by/3.0/

Module: "Editor's Introduction to Research in a Connected World"
By: Alex Voss, Elizabeth Vander Meer, David Fergusson
URL: http://cnx.org/content/m32855/1.2/
Pages: 3-4
Copyright: Alex Voss, Elizabeth Vander Meer, David Fergusson
License: http://creativecommons.org/licenses/by/3.0/

Module: "Research in a Connected World"
By: Alex Voss, Elizabeth Vander Meer
URL: http://cnx.org/content/m20834/1.3/
Pages: 5-8
Copyright: Alex Voss, Elizabeth Vander Meer
License: http://creativecommons.org/licenses/by/3.0/

Module: "What is a Distributed System?"
By: Donal Fellows
URL: http://cnx.org/content/m31661/1.1/
Pages: 9-13
Copyright: Donal Fellows
License: http://creativecommons.org/licenses/by/3.0/

Module: "Archaeology"
By: Stuart Dunn
URL: http://cnx.org/content/m31033/1.1/
Pages: 15-19
Copyright: Stuart Dunn
License: http://creativecommons.org/licenses/by/3.0/

Module: "Text Analysis in the Arts and Humanities"
By: Tobias Blanke
URL: http://cnx.org/content/m31502/1.2/
Pages: 19-22
Copyright: Tobias Blanke
License: http://creativecommons.org/licenses/by/3.0/

Module: "Climate Prediction"
By: Andy Kerr
URL: http://cnx.org/content/m31704/1.1/
Pages: 22-25
Copyright: Andy Kerr
License: http://creativecommons.org/licenses/by/3.0/

Module: "e-Malaria"
By: Jeremy Frey
URL: http://cnx.org/content/m31767/1.1/
Pages: 25-30
Copyright: Jeremy Frey
License: http://creativecommons.org/licenses/by/3.0/

Module: "nanoCMOS Device, Circuit and System Simulations"
By: richard sinnott
URL: http://cnx.org/content/m32874/1.1/
Pages: 30-34
Copyright: richard sinnott
License: http://creativecommons.org/licenses/by/3.0/

Module: "Computational Chemistry"
By: Sarah Harris
URL: http://cnx.org/content/m32928/1.1/
Pages: 35-39
Copyright: Sarah Harris
License: http://creativecommons.org/licenses/by/3.0/

Module: "Biomedical Research"
By: Ana Lucia DA COSTA
URL: http://cnx.org/content/m32938/1.1/
Pages: 39-44
Copyright: Ana Lucia DA COSTA
License: http://creativecommons.org/licenses/by/3.0/

Module: "The European e-Infrastructure Ecosystem"
By: Erwin Laure
URL: http://cnx.org/content/m31215/1.1/
Pages: 45-50
Copyright: Erwin Laure
License: http://creativecommons.org/licenses/by/3.0/

Module: "The EGEE Distributed Computing Infrastructure"
By: Steven Newhouse
URL: http://cnx.org/content/m32047/1.1/
Pages: 50-55
Copyright: Steven Newhouse
License: http://creativecommons.org/licenses/by/3.0/

Module: "Scholarly Communication and the Web"
By: David De Roure, Carole Goble
URL: http://cnx.org/content/m32860/1.3/
Pages: 57-60
Copyright: David De Roure, Carole Goble
License: http://creativecommons.org/licenses/by/3.0/

Module: "Scientific Workflows"
By: Katy Wolstencroft, Paul Fisher, David De Roure, Carole Goble
URL: http://cnx.org/content/m32861/1.3/
Pages: 60-65
Copyright: Katy Wolstencroft, Paul Fisher, David De Roure, Carole Goble
License: http://creativecommons.org/licenses/by/3.0/

Module: "Repositories"
By: Mark Hedges
URL: http://cnx.org/content/m31391/1.1/
Pages: 65-69
Copyright: Mark Hedges
License: http://creativecommons.org/licenses/by/3.0/

Module: "Resource Sharing: Trust and Security"
By: richard sinnott
URL: http://cnx.org/content/m32872/1.1/
Pages: 69-74
Copyright: richard sinnott
License: http://creativecommons.org/licenses/by/3.0/

Module: "Portals"
By: Gergely Sipos
URL: http://cnx.org/content/m24551/1.1/
Pages: 75-78
Copyright: Gergely Sipos
License: http://creativecommons.org/licenses/by/3.0/

Module: "Visualization Matters"
By: Martin Turner
URL: http://cnx.org/content/m31926/1.1/
Pages: 78-84
Copyright: Martin Turner
License: http://creativecommons.org/licenses/by/3.0/

Module: "Virtual Research Environments"
By: Alex Voss
URL: http://cnx.org/content/m32637/1.1/
Pages: 84-87
Copyright: Alex Voss
License: http://creativecommons.org/licenses/by/3.0/

Module: "Examples of e-Research Videos - from the eIUS project"
By: Alex Voss
URL: http://cnx.org/content/m32619/1.2/
Pages: 89-90
Copyright: Alex Voss
License: http://creativecommons.org/licenses/by/3.0/

Module: "Virtual Research Environments - Videos"
By: Alex Voss
URL: http://cnx.org/content/m32636/1.1/
Page: 90
Copyright: Alex Voss
License: http://creativecommons.org/licenses/by/3.0/

Research in a Connected World
The Research in a Connected World collection provides an overview of distributed computing technologies and their use in research.

About Connexions
Since 1999, Connexions has been pioneering a global system where anyone can create course materials and make them fully accessible and easily reusable free of charge. We are a Web-based authoring, teaching and learning environment open to anyone interested in education, including students, teachers, professors and lifelong learners. We connect ideas and facilitate educational communities.

Connexions's modular, interactive courses are in use worldwide by universities, community colleges, K-12 schools, distance learners, and lifelong learners. Connexions materials are in many languages, including English, Spanish, Chinese, Japanese, Italian, Vietnamese, French, Portuguese, and Thai. Connexions is part of an exciting new information distribution system that allows for **Print on Demand Books**. Connexions has partnered with innovative on-demand publisher QOOP to accelerate the delivery of printed course materials and textbooks into classrooms worldwide at lower prices than traditional academic publishers.