CombeChem: An *e*-Science Research Project

Jeremy G. Frey
Department of Chemistry, University of Southampton
Southampton , S017 1BJ, UK

Email: j.g.frey@soton.ac.uk Tel: 023 8059 3209 www.combechem.org

Introduction: The e-Science programme.
In November 2000 the Director General of Research Councils, Dr John Taylor, announced £98M funding for a new UK e-Science programme. *e*-Science refers to the large scale science that will increasingly be carried out through distributed global collaborations enabled by the Internet. While the World Wide Web gave us access to information on Web pages Internet a much more powerful infrastructure is needed to support e-Science. This infrastructure is the GRID. Besides information stored in Web pages, the Grid will provide scientists with easy access to expensive remote facilities, to computing resources, distributed databases and enable collaboration without regard to geography (well within the limits of the speed of light).

All the UK research councils were allocated funding under this programme and research relevant to Chemistry and the general pharmaceuticals area is being funded by the EPSRC, BBRC. The environmental concerns of the NERC look interesting in this context as well. The funding strengthens and significantly extends the work being undertaken by the UK Interdisciplinary Research Centres (IRC) such as the "Advanced Knowledge Technologies (AKT)" project the EQUATOR (project and the Medical Imaging and Signals project which enables "distributed medicine". The *e*-Science Core Programme set up a national infrastructure to support e-Science/Grid activities including a National e-Science Centre at Edinburgh (www.nesc.ac.uk), Regional Centres and Grid Support and Grid Network teams based round the CCLRC. There are strong links to EU projects in the *e*-Science and Grid Area. An aim of the UK e-Science research is as a bridge between the US viewpoint and the European view.

The UK Computer Science research base brings significant abilities and understanding in the arena of "Knowledge Engineering" to the vision of readily available computational resources. This plays the UK strengths formed via JISC supported activities for library resources such as UKOLN and DNER. The "Knowledge Engineering" aspects of our projects is particularly important as the availability of huge computational and network resources is not sufficient for achieve research - it is not always what resources you have but how you use them that is important.

The Comb-*e*-Chem Project
Chemistry has always made extensive use of the developing computing and information technologies and been an avid consumer of available computing power. Chemical uses of the technology include activities such as modelling, simulation and chemical structure interpretational; activities conveniently summarised as computational chemistry. New procedures in chemical synthesis and characterisation, particularly in the arena of parallel and combinatorial methodologies, have generated ever-increasing demands on both Computational Chemistry and Computer Technology. However, significantly the way in which networked services are being conceived to assist collaborative research pushes well beyond the traditional computational chemistry programmes, towards the basic issue of handling Chemical information and knowledge. The rate at which new chemical data can now be generated in Combinatorial and Parallel synthesis and screening processes means that the data can only realistically be handled efficiently by increased automation of the data analysis as well as the experimentation and collection.

XML Exchanges

In starting to set up the Comb-*e*-Chem project we realised that it is essential to develop mechanisms for exchanging information. This is of course a common feature of all the *e*-science projects, but the visual aspects of chemistry do lead to some extra difficulties. Many chemists have been attracted to the graphical interfaces available on computers. The drag-and-drop, point-and-shoot techniques are easy and intuitive to use but present much more of a problem to automate than the simple command line program interface. Fortunately these two streams of ideas are not impossible to integrate, but it does require a fundamental rethink on how to implement the distributed systems while still retaining (or perhaps providing) the usability required by a bench chemist. One way in which we will ensure that the output of one machine or program can be fed in to the next program in the sequence is to ensure that all the output is wrapped with appropriate XML. In this we have some advantages

as Chemists, Chemical Mark-up Language (cML) was one of the first XML systems to be developed by Peter Murray-Rust[1].

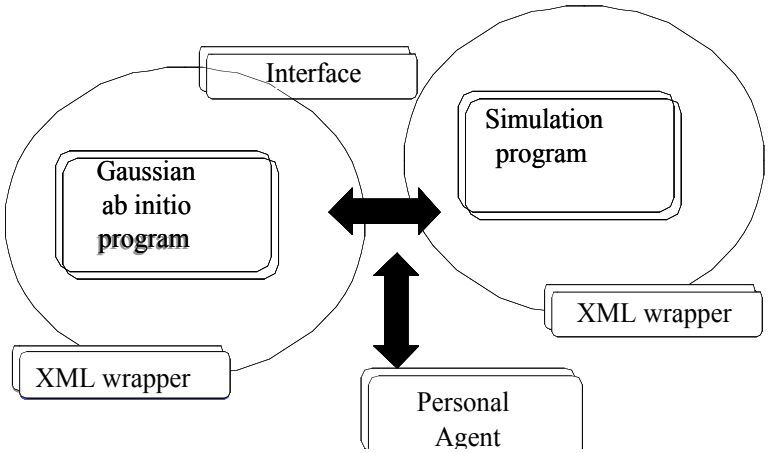

Figure 1:
Showing the use of XML wrappers to facilitate the interaction between two typical chemistry calculation programs. . The XML (perhaps cML and other schema's) ensures that a transparent, flexible, reusable and flexible workflow can be implemented. The resulting workflow system can then be applied to all the elements of a combinatorial library automatically.

Figure 1 illustrates this for a common situation where information needs to be passed between a Quantum mechanical calculation that has evaluated molecular properties[2] and a simulation programme to calculate the properties of a bulk system or interface. It equally applies to the exchange between equipment and analysis. A typical chemical application would involve for example a search of structure databases for the details of small molecules, followed by a simulation of the molecular properties of this molecule, then matching these results by further calculations against a protein binding target selected from the protein database and finally visualisation of the resulting matches. Currently the transfer of data between the programs is accomplished by a combination of macros and Perl scripts each crafted for an individual case with little opportunity for intelligent reuse of scripts. Proper analysis of this process and the implementation of a workflow will enable much better automation of the whole research process[3].

Figure 1 illustrates another issue; more information may be required by the second program than is available as output from the first. Extra knowledge (often experience) needs to be added. The Quantum program provides for example a molecular structure but the simulation program requires a force field (describing the interactions between molecules). This could be simply a choice of one of the standard force fields available in the packages (but a choice never-the-less that must be made) or may be derived from additional calculations from the QM results. This is where the interaction between the "Agent" and the workflow appears[4,5].

Virtual Data

A Combinatorial Library could be thought of as a "Library Collection" with the material itself and all the information on that material all ideally cross-referenced. If information is requested from a library then it can be provided if it is held in the collection, if not a search can be made to locate and deliver it from elsewhere. For the chemical information searches spilling out to a from a wide variety of databases are common but the power of the Grid based approach to the handling of the combinatorial data is that we can go further than this "static" approach.

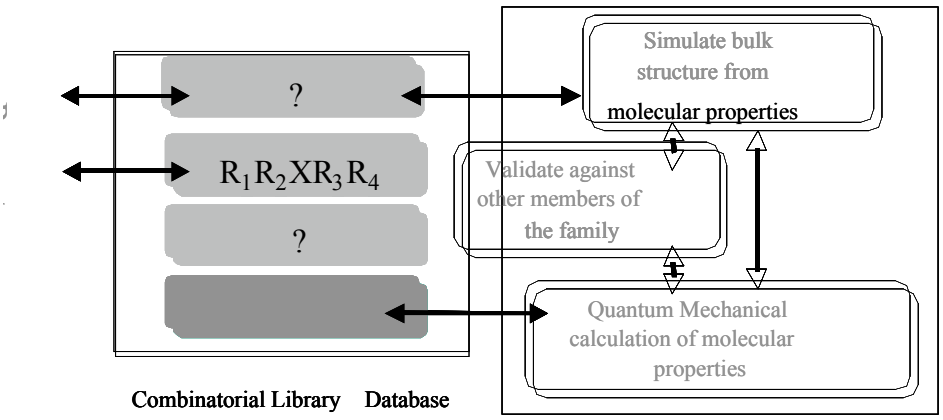

Figure 2 Virtual Data

An example of virtual data interactions in the context of a combinatorial family. The calculations of an unknown material can be calibrated against known members of the library.

The combination of the high throughput laboratory equipment and the resulting information, together with the calculation resources of the Grid allows for a much more interesting "library network" to be created. As shown in Figure 2 an appropriate model can calculate the requested data. These models are themselves validated by comparison with the measured properties of the actual physical members of the library. Depending on time or resource constraints different types of model, or different levels of implementation of a model can be chosen, ranging from resource hungry high level quantum mechanical calculation, through extensive simulations, to an empirically based approached; we thus have in effect a virtual entry in the database. In the limit this process has a close connection with the ideas of virtual screening of combinatorial libraries.

As in our model the Grid extends down in to the laboratory this virtual data idea can be extended to not only calculations but also to new experimental data acquisition or even automated synthesis. That is the direction of the synthesis or analysis in the automated laboratory would be controlled via a database request. The delegated activation of computational and physical resources has many implications for security, accountability and accounting.

Statistical Models

The presence of a large amount of related data such as that obtained from the analysis of a combinatorial library suggests that it would be productive to build simplified statistical models to predict complex properties rapidly. A few extensive detailed calculations on some members of the library will be used to define the statistical approach, building models using, for example, appropriate regression algorithms or neural nets or genetic algorithms, that can then be applied rapidly to the very large datasets.

The rapid acquisition of new information means that the very processing of data via these statistical models and the resulting comparisons that will be made generate new information that can be used to modify the statistical models. The speed with which can envisage the design – experiment-analysis-model building-modify-design cycles means that it is more appropriate to consider the new data flowing through the workflow as potentially modifying that workflow. The automated analysis becomes an example of a self modifying network bring us very close to the frontiers of Computer Science Research.

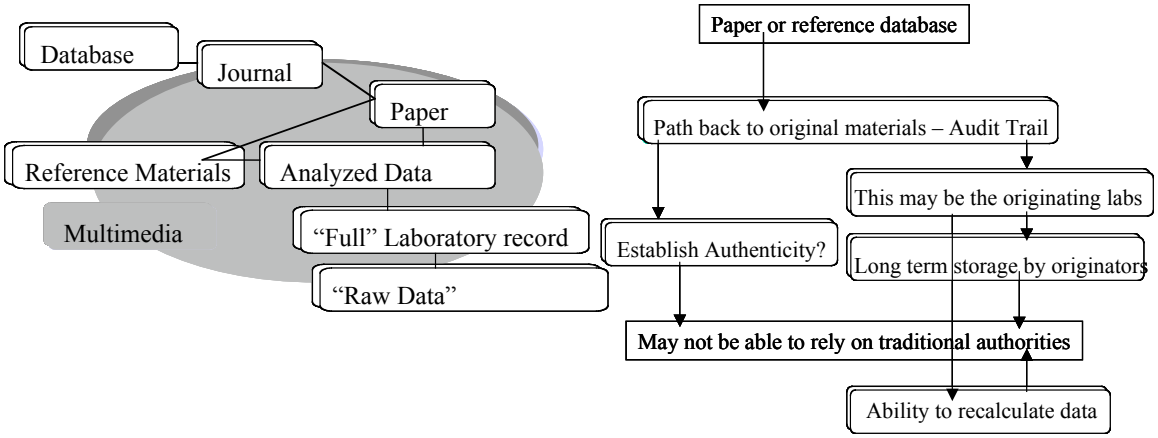

Figure 3 The e-Publication chain for multimedia Chemistry Information

Publication@Source

Chemistry is a multimedia subject – 3D structures are key to our understanding of the way in which molecules interact with each other. The historic presentation of results originally as text and then on a flat sheet of paper is to limiting for current research. 3D projectors are now available, dynamic images and movies are now required to portray adequately the chemist's view of the molecular world. This dramatically changes expectations of what a journal will provide and what is meant by "publication. Chemistry is becoming an information science[6],, but exactly what information should be published? And by whom? The traditional summary of the research with all the important details will continue to provide a productive means of dissemination of the chemical ideas. The databases and journal papers link to reference data provided by the authors and probably held at the journal site or a subject specific authority. Further links back to the original data take you to the author's laboratory records. The extent type of access available to such data will be dependent on the authors as will be the responsibility of archiving these data. There is thus inevitably a growing partnership between the traditional authorities in publication and the people behind the source of the published information, in the actual publication process.

We seek to formalise this process by extending the nature of publication to include links back to information held in the originating laboratories. In principle this should lead right back to the original records (spectra, laboratory notebooks) which will give much greater use and re-use of the original data. The consequent checking of the data and application of different approaches to the analysis will also be beneficial.

Other *e*-Science Projects

CombeChem has strong links with two of the other e-Science projects Reality Grid and MyGrid. RealityGrid is a collaboration whose aim is to grid-enable modelling and simulation of condensed matter structures at the meso- and nanoscale levels, and the discovery of novel materials. MyGrid aims to design, develop and demonstrate higher level functionalities over an existing Grid infrastructure that support scientists in making use of complex distributed resources producing a virtual laboratory workbench that will serve the life sciences community and bioinformatics community.

Conclusions

The Grid infrastructure when fully developed will enable the Chemist to sit at the centre of a virtual world with simple, rapid access to a wide range of physical, computational and informatics resources. The implementation of automatic knowledge handling right from the inception of an experiment, through all stages of analysis and use of the information generated, will enable the single human mind to work collaboratively with others to keep pace with the exponentially growing quantities of chemical information generated by combinatorial techniques taking place in "dark" high throughput laboratories. Without such techniques we run the very real risk of generating even more information that is effectively hidden from the very people who should be using it.

Acknowledgements

In addition to myself the main investigators of the Comb-*e*-Chem project are M.B. Hursthouse, D.C. De Roure, J.W.Essex, S.M. Lewis, A.H. Welsh, M. Surridge. Together with the Universities of Southampton and Bristol several companies and organisations are involved in the project in particular IBM and CCDC have made significant contributions to the early stage of the project.

References

[1] Chemical Markup Language, P. Murray-Rust World Wide Web Journal, 1997, pp135-147

[2] The Past present and Future of Quantum Chemistry, T. D. Crawford, S. S. Wesolowski, E.F. Valeev, R.A. King, M.L. Leininger and H.F. Schaefer III, Chapter 13 (pages 219 – 246) in Chemistry for the 21st Century, E. Keinan and I. Schechter Eds. Wiley-VCH, 2001, Weinheim, ISBN 2-527-30235-2.

[3] Leymann, F. and D. Roller. "Workflow-based Applications", IBM Systems Journal, 36 (1997) 1, pp. 102-123.

[4] M. Wooldridge and N.R. Jennings: Intelligent Agents: Theory and Practice, The Knowledge Engineering Review, 10 (2), pp. 115-152, 1995.

[5] N. R. Jennings (2001) "An agent-based approach for building complex software systems" Comms. of the ACM, 44 (4) 35-41.

[6] "Some Reflections on Chemistry", J.M. Lehn, Chapter 1 (page 1-7), in Chemistry for the 21st Century, E. Keinan and I. Schechter Eds. Wiley-VCH, 2001, Weinheim, ISBN 2-527-30235-2.