Are bioinformaticians doing e-Business?

M. Greenwood, C. Wroe, R. Stevens, C. Goble
Department of Computer Science, University of Manchester, Manchester, UK

M. Addis
IT Innovation Centre, University or Southampton, Southampton, UK

Abstract

We have models of commerce in a Web setting: business to business (B2B) and business to consumer (B2C). Now scientists commonly use Web based services to perform in-silico experiments. Thus we are prompted to ask the question "Are e-Scientists doing e-Business?". Do the infra-structure and models offered by e-Commerce support the activities e-Scientists need to perform? In this position paper we compare e-Science and e-Business using the discipline of bioinformatics. Such a comparison should inform the reuse of existing e-Business technologies in e-Science projects. We argue that the individual e-Scientist is now demanding more than the simple web interfaces prevalent in consumer e-commerce. Individual e-Scientists need to interact in a manner more akin to the B2B model than the B2C style previously used. We examine how the infrastructure prevalent in the B2B arena of e-commerce can be reused and extended to support the needs of today's e-Scientists. We illustrate this argument with reference to the myGrid e-Science middleware project.

1. INTRODUCTION

In its simplest sense, e-business is the use of Internet technologies to improve and transform key business processes. Businesses web-enable core processes to strengthen customer service operations, streamline supply chains and reach existing and new customers. This has led to the requirement for a massively scalable, reliable and secure electronic foundation which includes reliable and available servers, robust software and middleware [1]. It is appealing to think that this infrastructure could be directly used to support the emerging field of e-Science. However this approach will only succeed if the e-Business and e-Science are sufficiently similar. In this position paper we make a first step to compare current e-Business with e-Science, and use the e-Science pilot project ^{my}Grid to illustrate the consequences of the similarities and differences.

E-Science covers a broad range of disciplines but for this paper we will concentrate on the biological domain. Biology is a multi-faceted and increasingly multi-disciplinary science. It has already moved to large interdisciplinary teams distributed throughout the world working together on specific problems. Much biology is based on comparative and speculative reasoning; predicting what might happen based on its relationship to "similar" things seen and studied previously. There is substantial use of electronic resources to compare new results with the existing body of biological knowledge. Discovery is done by combining and collating results obtained from a number of analysis and data resources. These *in-silico* experiments complement experiments *in vitro* by synthesising new information from available data and generating hypotheses for lab-based confirmation.

From the early 1990s web technology has been enthusiastically adopted by the biological community to disseminate data and analysis methods. Bioinformaticians can carry out simple, "low volume" *in-silico* experiments by cutting and pasting data between web pages. However, the complexity of potential *in-silico* experiments together with the volumes of data produced by high throughput experimentation is now threatening to overwhelm standard web technology. The data are often complex, variable in quality, frequently changing and mostly comprised of incomplete data sets. Analysis methods are constantly and rapidly evolving. There is a positive feedback cycle. As more resources become available, more *in-silico* experiments can be done, these generate more resources and knowledge for designing further experiments. The key requirements for bioinformaticians are: discovering web-based resources and gaining access to them, orchestrating their use of resources in *in silico* experiments, and maintaining provenance records of what has been done.

The ^{my}Grid project (http://www.myGrid.org.uk) aims to provide a personalisable collaborative problem-solving platform for an individual e-Scientist working in a distributed environment. The application focus is on data-intensive bioinformatics. The idea is that a scientist can construct long-lived *in-silico* experiments represented as workflows, find and adapt others and publish their own. They can have private personal data collections and also pool them, have their own view on public repositories and be able to easily interoperate these with their own data or tools. By being better informed as to the provenance and the currency of the tools and data directly relevant to them the hope is to improve both the quality of information in repositories and the way repositories or tools are used. The ^{my}Grid architecture is based on services, initially implemented by Web services but is intended to be delivered as Grid Services [7].

2. MODELS OF INTERACTION

E-business interactions can be divided into two broad areas business to business (B2B) and business to consumer (B2C). In the B2B case the driving force is on businesses cooperating within a value chain each business concentrates on its core competencies and outsources other tasks to its partners. Partnerships involve complex negotiation, and coordination of business processes. In the B2C case the driving force is the individual consumer who accesses one business' services through a simple static interface such as a web browser. The most widely used example is ordering a book on from amazon.com. Typically consumers are interested in results; they expect providers or brokers to orchestrate services and present them with a straight-forward interface. If a service does not satisfy their requirements they will seek another service provider.

In a commercial setting, businesses will typically interact according to a well-defined process for outsourcing, which is predominately applicable to B2B interactions. UN/CEFACT defines the steps of outsourcing to be implementation, discovery, design and runtime [2]. Implementation is the provision and publication of a service's particular e-Business capabilities. Discovery refers to the process by which a potential business partner can discover the service provider and the capabilities of their service. Design is the stage of defining the process by which one or more services are orchestrated to fulfil an overall objective. Finally, runtime refers to the execution of the services. The ebXML community defines a similar set of steps, including process definition; partner discovery; partner sign-up; process execution; and process management [3]. The implementation of the outsourcing process can be readily mapped onto various underlying technologies, for example by using the 'publish', 'find', 'bind' model of Web Services [4] accompanied by a suitable Web Service orchestration language such as BPEL4WS [5] or WSFL [6].

Although currently not well characterised, e-Science interactions can be divided into two distinct types analogous to e-Business: enterprise e-Science and personal e-Science.

Enterprise e-Science occurs between organisations or institutions of scientists and is analogous to B2B. Groups cooperate in order to pool limited resources. As in business, partnerships increasingly involve complex negotiation concerning intellectual property rights. For example, any submission of a nucleotide sequence to either the EMBL, Genbank or DDBJ databanks is transferred between the other resources. The three organisations have negotiated a collaboration to ensure that all three resources contain the same data.

Personal e-Science needs support beyond a simple B2C model. The individual scientist is not a long term partner keen to spend time and effort negotiating contract agreements. However the scientist is interested in maintaining control of the orchestration of distributed services into *in silico* experiments. They can not be considered an outsider with little interest in how the service works as in the classic B2C scenario. They must act as an active if short-lived business partner with rights to orchestrate the surrounding services.

3. DISCOVERY AND NEGOTIATION

One of the characteristics of e-Science is that scientists are very interested in the discovery and selection of services. They need to know the most applicable services so that their results are valid within their community, so their outlook is that of a strong business partner rather than a casual consumer. As more services become available, there must be ways to discover those that are relevant in the context of a specific *in-silico* experiment.

Within B2B, the process of service discovery is in two phases. The first is a rough identification of potential business pertners. This requires service providers to advertise themselves sufficiently as a candidate with the right service capability and distinguish themselves from competitors. The provider needs to advertise just enough to hook a potential partner. The yellow and white pages of UDDI [8], for example, give rather coarse-grained and weak descriptions of service function and quality of service. The second phase is a one-one negotiation of service level agreements. In contrast to the shop-front of the first phase, this is conducted behind closed doors. Detailed service capabilities are requested and traded under non-disclosure contractual obligations.

In personal e-Science an individual has little scope to force the modification of a service in terms of functionality or performance. What you see is what you get. Negotiation is replaced by a much more detailed discovery process in which the exact requirements are certain to be met. Thus, the service description must be rich enough to make a decision about whether it will fulfil the task. This discovery mechanism is not a one-off exercise during the design of an *in-silico* experiment. Often the life of an experimental design may outlive the availability of a service, or a more appropriate service may be published in the interim. It is therefore desirable to describe the functionality required and defer actual selection of a service until the experiment is run.

It is also important that the service discovery mechanisms are open and extendable. New more specialised data sets and more sophisticated computational methods will continue to become available and will need to be described. These will need to be discovered as easily as those available today. The consequences for ^{my}Grid are:

1. To extend the description of a service beyond a brief advertisement. A controlled vocabulary is used to describe the functionality of the service in terms of its inputs, outputs, resources used and overall goal [9]. DAML-S [10] is an ontology based schema with which to describe a web service in the business setting. It has to be extended beyond its business roots to cover the relevant functional attributes on which the scientist could base their choice. For example, the algorithm used, the name of the application used to implement the service, and the generic nature of the task performed.
2. ^{my}Grid's description of services needs to deal with the rapidly evolving nature of bioinformatics services. The ontology of services must be open and extensible. The autonomous service providers must have access to a vocabulary and a formal grammar that enables them to independently construct novel service descriptions.

Although it is not possible for the scientist to modify a service, general services such as those that provide access to databases, do allow complex configuration to perform a specific task. Translating from the goals of the scientist to the appropriate services is not just about selecting between equivalent services based on cost or reliability. Geodise [11] (http://www.geodise.org) is another e-Science pilot project, which as part of its remit, is examining the explicit knowledge required to configure complex computational services.

4. ORCHESTRATION USING WORKFLOW

Capturing in-silico experiments as a reusable process that can be defined, published and repeatedly executed is essential when sharing scientific best practice. In bioinformatics, processes are reusable at many levels.

At a basic level, a process might capture how to use a programming API of a remote application, for example proving input data, executing and then retrieving the results for a protein similarity search. This process is analogous to simple B2C or B2B processes such as use of an on-line flight reservation service.

Basic bioinformatics processes can be combined in order to orchestrate a set of applications to perform a higher-level function, for example the combination of protein similarity searching with a protein structure database to provide protein structure prediction. These processes are analogous to aggregate or brokered B2C processes such as coordination of flight, hotel and excursion booking to form a holiday package. However in e-Science the consumer wants control over the coordination.

Finally, in bioinformatics the semantics of service orchestration can be abstracted away from the syntactic details of data exchange, data transformation and specific application invocation, to create reusable processes that can be customised to individual scientists and service providers. For example, general bioinformatics processes include structure prediction, literature knowledge extraction, and annotation curation in databases. These are examples of e-business patterns, as identified by IBM [12] in this case 'extended enterprise', 'information aggregation', and 'collaboration' respectively.

^{my}Grid aims to support personal interaction with workflow at these multiple levels. Users must be able to create novel workflows and specialise workflow templates held in a library. Workflows must be described as resources in their own right to enable their discovery.

The scientist must be able to control the enactment or delivery of a set of services. The speculative nature of the task, requires the scientist to monitor the state of the process, and possibly steer it in the light of intermediate results. An enactment engine has been built by the ^{my}Grid partner IT Innovation (http://www.it-innovation.soton.ac.uk), which interprets workflows specified in WSFL, and allows interaction with the user during the enactment process. In current ^{my}Grid workflows this allows certain steps to be under-specified; for example denoting the functional class of service to be used rather than a specific instance. The user can then be asked to choose which of the services available at that time should be used.

5. PROVENANCE

Although the physical confederation of services may be transient, the results are not. In science as in many fields the mechanism by which the results are obtained can be of more importance than the results themselves. There needs to be an explicit record of the origin of results which resources were used, when, what parameters were supplied etc. This is essential evidence for assessing whether an *in silico* experiment needs to be re- run in the light of new evidence, such as updated resources. A provenance record is also essential if for legal reasons an e-scientist needs to prove that a given result was obtained on a specific date, using a particular set of resources available at that time.

Provenance of data is analogous to traceability of raw materials in business. "...The capability to follow the path of a specified unit of a product through the supply chain as it moves between organisations. Products are traced for purposes such as product recall and investigating complaints.." [13]. A key mechanism for achieving traceability is a unique identifier for all items in the supply chain and the European Article Numbering (EAN) provides just such a mechanism. In ^{my}Grid we aim to identify all input data, intermediate and final results together with the process used to create the results. ^{my}Grid is monitoring the development of the Life Sciences Identifier (LSID). A LSID is a standard mechanism in development by I3C (http://www.i3c.org) for identifying biologically significant resources and may become the bioinformatics equivalent of the EAN.

It is equally important to record the process by which results are obtained. Business processes often include secure and verifiable business transaction semantics [14]. These are directly applicable to scientific processes since they provide a mechanism for provenance to be established. Security techniques allow a verifiable audit trail to be created to establish what was done at the time of an *in-silico* experiment in terms of who was involved, what applications were executed, what data was used, and most importantly what process was followed. Whilst there is the possibility to use this for legal purposes, one of the key values is to enable *in-silico* experiments to be re-run if new data or services become available and then results meaningfully compared with historical analysis.

6. DISCUSSION

We have asked ourselves the question "Do bioinformaticians perform e-Business?". Simply, the answer is "yes". E-Scientists use Web based services in a range of activities, from a single step analysis akin to the B2C model, up to the orchestration of multiple services like that of the B2B model. ^{my}Grid has thus been able to reuse e-Business technologies such as UDDI registries, workflow languages and enactment systems to support personal e-Science.

Of course, the reality is a little more complex. We have correlates of the B2B and the B2C models, but modern bioinformatics needs an intermediate level of interaction. In this style, an individual e-Scientist needs to transiently behave in a B2B model, orchestrating autonomous, distributed, heterogeneous services to analyse, discover or create new data. Users must have access to detailed service descriptions to inform unilateral partnership formation and orchestration. They must be supported in the interaction with workflow representations and in the recording of experiments performed.

For such an infrastructure, a far more explicit semantic representation of service capabilities and interaction is needed. The encoded knowledge must support a model of interaction by which a middleware broker allows access to the full functionality of a set of dynamically orchestrated services, while shielding the scientist from the full complexity of implementing that orchestration. ^{my}Grid has begun to build the additional middleware components and explicit knowledge necessary to support the personal e-Scientist and will deliver these to the community.

ACKNOWLEDGEMENTS

This work is supported by the ^{my}Grid e-Science pilot (EPSRC GR/R67743), and the Geodise e-Science pilot (EPSRC GR/R67705). The authors would like to acknowledge the rest of the ^{my}Grid team: Nedim Alpdemir, Rich Cawley, Vijay Dialani, David De Roure, Alvaro A. A. Fernandes, Justin Ferris, Rob Gaizauskas, Kevin Glover, Chris Greenhalgh, Ian Horrocks, Peter Li, Xiaojian Liu, Phillip Lord, Darren Marvin, Simon Miles, Luc Moreau, Tom Oinn, Norman Paton, Stephen Pettifer, Milena Radenkovic, Angus Roberts, Alan Robinson, Tom Rodden, Martin Senger, Nick Sharman, Brian Warboys, Paul Watson.

REFERENCES

[1] IBM e-business. Getting the tools you need to take a strategic approach to a smart e-business infrastructure. http://www-3.ibm.com/e-business/doc/content/overview/28212.html
[2] UN/CEFACT Electronic Business Transition ad-hoc Working Group (eBTWG) "eBTWG - Electronic Business Architecture - Revision 0.50" http://www.ebtwg.org/news/pdf/architecture_v0.50.pdf
[3] ebXML Business Process and Business Information Analysis Overview http://www.ebxml.org/specs/bpOVER.pdf
[4] Web Services Conceptual Architecture http://www-3.ibm.com/software/solutions/webservices/pdf/WSCA.pdf
[5] Business Process Execution Language for Web Services http://www-106.ibm.com/developerworks/library/ws-bpel/
[6] F. Leymann. Web Services Flow Language (WSFL 1.0) IBM Software Group. (May 2001) Available: http://www-3.ibm.com/software/solutions/webservices/pdf/WSFL.pdf
[7] I. Foster, C. Kesselman, J. Nick, S. Tuecke. The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. Available: http://www.globus.org/ogsa/ (2002).
[8] UDDI Technical White Paper. (September 2000) Available: http://www.uddi.org
[9] C. Wroe, R. Stevens, C. Goble, A. Roberts, M. Greenwood. A suite of DAML+OIL ontologies to describe bioinformatics web services and data. International Journal of Cooperative Information Systems. In press.
[10] A. Ankolekar, M. Burstein, J. Hobbs, O. Lassila, D. Martin, S. McIlraith, S. Narayanan, M. Paolucci, T. Payne, K. Sycara, H. Zeng . DAML-S Semantic Markup for Web Services. Proceedings of the International Semantic Web Working Symposium (SWWS) (2001).
[11] S.J. Cox et al. Grid Services in action: Grid Enabled Optimisation and Design Search. 11th IEEE International Symposium on High Performance Distributed Computing HPDC-11 2002 (HPDC'02) pp. 413.
[12] R. Bloor, M. Hanrahan. Patterns of Experience. A Review of IBM's Patterns for e-business Initiative http://www-106.ibm.com/developerworks/patterns/guidelines/bloor.pdf
[13] Fresh Produce Traceability Guidelines Available: http://www.ean-int.org/agro-food/Opmaak%20tekst%20Fresh%20Produce%20.pdf
[14] ebXML Business Process Specification Schema http://www.ebxml.org/specs/ebBPSS.pdf