Joining up Health and BioInformatics: e-Science meets e-Health

Alan Rector[a], Adel Taweel[a], Jeremy Rogers[a], David Ingram[b], Dipak Kalra[b], Robert Gaizauskas[c], Mark Hepple[c] Jo Milan[d], Richard Powers[e], Donia Scott[e], Peter Singleton[b]

[a]BioHealth Informatics Forum, Department of Computer Science, University of Manchester
[b] Centre for Health Informatics & Multiprofessional Education, University College London,
[c]Department of Computer Science, University of Sheffield
[d] Royal Marsden Hospital Trust, London
[e]Institute for IT Research, University of Brighton

rector@cs.man.ac.uk www.clinical-escience.org

Abstract

CLEF (Co-operative Clinical e-Science Framework) is an MRC sponsored project in the e-Science programme that aims to establish methodologies and a technical infrastructure for the next generation of integrated clinical and bioscience research. It is developing methods for managing and using pseudonymised repositories of the long-term patient histories which can be linked to genetic, genomic information or used to support patient care. CLEF concentrates on removing key barriers to managing such repositories – ethical issues, information capture, integration of disparate sources into coherent "chronicles" of events, user-oriented mechanisms for querying and displaying the information, and compiling the required knowledge resources. This paper describes the overall information flow and technical approach designed to meet these aims within a Grid framework.

Introduction

Our rapidly increasing ability to gather information at the molecular level has not been matched by improvements in our ability to gather information at the patient level. There is a strong convergence of need between current trends towards safer more evidence based patient care and current trends in post-genomic research which seek to link molecular level processes to the progress of disease and the outcome of treatment. Both need to be able to answer the questions:

What happened and why?
What was done and why?

Simple those these questions may seem, they remain difficult to answer without recourse to manual examination of patients' notes – a time consuming process whether the notes are electronic or paper. Yet without answers to these questions, it is difficult either to measure the quality of care or to investigate the factors affecting onset and recurrence of disease.

Barriers & Requirements

Barriers

CLEF categorizes the key barriers and requirements as:

- *Privacy, consent, and security* – at all levels: policy, organisational structure, and technical implementation.
- *Information capture* – extracted from text as well as collected from structured records, reports, and results.
- *Information integration and 'chronicalisation'* – to infer a coherent history of events from the hundreds of diverse documents that make up the raw material of the patient record.
- *Information analysis, presentation and summarisation* – to make the information on individuals and populations easily accessible to both practising clinicians and biomedical researchers with minimal specialist training.
- *Knowledge resources* – to recognise the significance and interrelationships of events.

- *Standards for both data and metadata* – to permit effective information sharing and re-use.

Requirements & information flow

The requirements and technologies are best understood in the context of the CLEF information flow that has emerged from the design process and is shown in Figure 1.

Starting with the "Patient care and dictated text" at the left side of the diagram, the flow is:

- *Capture* of the information. Some information comes from dictated and transcribed text. Other information comes directly from hospital information systems – *e.g.* laboratory results, prescriptions, etc.
- *Pseudonymisation* of all information at the originating hospital by removal of overt identifying items – name, date of birth, etc - and by providing a CLEF Entry identifier that can only be reversed by the provider (or their nominated trusted third party)
- *Depersonalisation* of the texts to remove any residual information that might risk identification – *e.g.* names of relatives, nick names, place names, unusual occupations, etc. Hence a requirement for reliable scalable techniques.
- *Information extraction* of key information from the texts into predefined "templates", possibly with the help of context provided by preceding texts and by structured information already in the repository
- *Integration into the health record repository* of all information including laboratories, radiology, and genomic analyses
- *Constructing the chronicle* to infer a coherent view of the patient's history. Typically the same information occurs in many different documents with different levels of granularity, clarity and sometimes conflicts must be reconciled.

From this point the information can go in two directions.

- *Use for patient care* - back to the clinicians in the form of summaries for patient care. Providing a concise up-to-date summary of a patients' condition is a prime request of clinicians for improving patient care. Because it requires re-identification of patients, this step can only occur at the hospital and after security controls have been stringently tested and agreed to be adequate.
- *Use for clinical e-science research* – on to the repository under the overall control of the Ethical Oversight Committee.
- *Enrichment for e-Science* – with the results of researchers' queries, their workflows, interpretations, curation and links to external information added to the repository so that it becomes the basis for virtual communities of researchers.

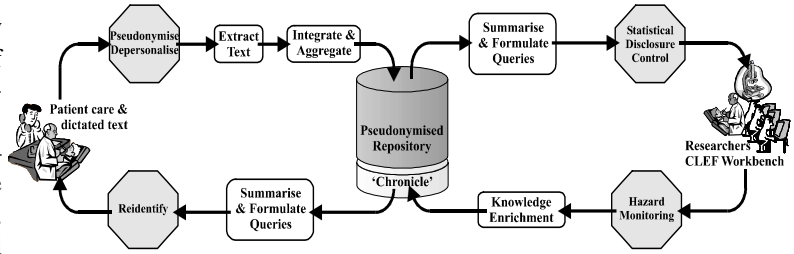

Figure 1 – Basic CLEF Information Flow

At the heart of CLEF is the compilation of a single coherent "chronicle" for each patient from distributed heterogeneous information that makes up the medical record. At one level, the chronicle provides a clear presentation to clinicians and researchers of the course of one patient's illness as shown in Figure 2. At another they are data structures which can be easily aligned on "index events" – diagnosis, first treatment, relapse, etc.- and aggregated for statistical analysis to answer questions such as "Of patients with breast cancer with a particular genetic profile, what is the comparison of the time to first recurrence for those treated with Tamoxifen as against those treated with a new proposed drug regimen". "How many dropped out of each treatment and why?" "How many required supplementary therapy for the side effects of treatment and why?"

Technologies required

Technologies and requirements

CLEF is focusing on the specific technologies which are currently barriers to obtaining and integrating clinical information:

- Privacy , confidentiality, consent, and security
- Information extraction from texts to acquire the information
- Integration of clinical information and the development of "chronicles"
- User oriented querhy formulation and a "What you see is what you meant interfac"

- Knowledge resources and metadata
- e-Science Infrasttructure and Links to the new NHS National Program for IT (NPfIT)
- Links to the new NHS Infrastructure

Privacy, confidentiality, consent and security

As is clear from the "stop signs" in Figure 1, much of the CLEF infrastructure involves privacy and security. The overarching requirement is a policy and oversight framework for privacy and consent. No technical solution can be perfect, so confidence in the organizational measures is the most critical single criterion for success.

Furthermore, no technical solution can succeed without vigilance. A key part of the CLEF policy is the obligation of care for all researchers to report potential hazards to privacy as part of the routine use of the CLEF repository coupled with technical measures to make it easy to do so.

However, technical measures are required and the requirements potentially conflict. Pseudonymous identifiers must be secure but must also support a) linking from multiple sources, b) re-identification with consent by the healthcare provider c) withdrawal or modification of consent by the patient. Both initial pseudonymisation and re-identification must be done solely within the hospital providing the information. Therefore, at least three stages of pseudonymisation are envisaged, one for entry from the hospital level, a second for linkage and use within the repository itself, and a third for any datasets authorised for release to users. Combinations of trusted third parties and techniques from e-Commerce (*e.g.* [22]) and current Grid research are under investigation, but the final choice must be deferred until the new NHS Infrastructure for the Integrated Care Record Service[1] has reached a stable state. To cope with this forced deferral, the current stage of the project deals with records from deceased patients only and uses a simplified scheme but is designing the architecture and user interface insofar as possible to accommodate the leading candidates.

The use of text extraction requires that special attention be paid to removing identifiers from text using language technology – a process we term "depersonalisation" which uses well established techniques from "named entity extraction" [10] and related techniques [21]. The effectiveness of the depersonalisation mechanisms will be rigorously checked using the corpus of records from deceased patients as a condition for use of the system with records of live patients.

The other side of the issue is the employment of *statistical disclosure control* technology to monitor and blur the output of queries to reduce the risk of deliberate or accidental re-identification through queries of the pseudonymised repository. No matter how well pseudonymised, de-identified and depersonalised, there is always a risk that personal data can be re-identified through sophisticated cross referencing, statistical or data mining techniques. This risk of such re-identification is well established and techniques to combat it are developing rapidly [6, 13, 15, 20]. One notable technique is referred to as *statistical disclosure control*. It focuses heavily on the assessment of risk in single, static and cross-sectional datasets [4, 5]. A systematic risk assessment disclosure control methodology [14] for the additional risks posed by *multiple* table releases will be employed to further to reduce the risk of re-identification.

Privacy is relative to risk and consent. All records in the repository contain detailed metadata on the level of consent granted for their use by patients. One of CLEF's major activities is to seek agreed standards for metadata on consent within the community.

Information Extraction & Language Technology

Doctors dictate. Much of the key information in clinical records continues, and will continue for the foreseeable future, to be contained in unstructured or at best minimally structured texts. Hence a major part of CLEF is devoted to adapting and evaluating mechanisms for information extraction from text [8, 11]. Four features of the cancer domain make information extraction feasible a) the very limited sublanguage, even more so than for medicine as a whole [7];b) much of the specialised information is in common with molecular biology which is a major target for current text extraction efforts *e.g.* [9, 19]; c) the well defined list of index events and signs that allows the template for extraction to be well defined; d) the existence of multiple reports for most events.

The existence of multiple reports is particularly important and has not been widely noted elsewhere to the best of our knowledge. Cancer patients are seen over a long period of time and their records summarized repeatedly so that there are many parallel or near parallel texts – often 150 or more text documents per patient. What may be unclear or ambiguous in one text can be

[1] http://www.doh.gov.uk/ipu/whatnew/specs_12d.htm

refined from others. This is particularly important when dealing with records from a referral hospital where the system usually will start in the "middle of the story". For example, first document might simply mention breast cancer in the past, concentrating on the current recurrence. A summary later might give a date for a mastectomy but no details of the tumour type. Eventually, perhaps after information from the referring hospital was received, a definitive statement of the time, tumour, spread, and treatment might be found. Subsequent notes might again refer to the initial cancer vaguely while concentrating on current concerns. By cross checking information, the picture of the overall "chronicle" gradually comes into focus, although still with varying degrees of certainty.

What this means for the architecture is that extracting information from one document may involve reference to the repository as a whole – hence the extra loop back from information extraction in Figure 1.

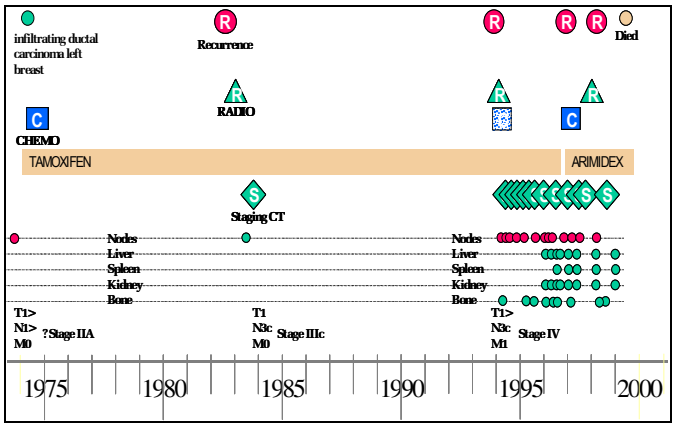

Figure 2: An individual patient chronicle in graphical form

"Chronicalisation" and Integration

The classic problem for electronic health records is to maintain a faithful, secure, non-repudiatable record of what healthcare workers have heard, seen thought and done [17]. The CLEF EHR repositories follow standards designed to achieve these aims – *e.g.* OpenEHR[2] [12], CEN standard 13606[3], and associated development of "archetypes"[1]

However, the central issue for CLEF is different – to infer a single coherent view of each patients' history from the myriad documents and data in the EHR including and to align them with other similar patients in aggregates for querying and research.

Furthermore, CLEF is interested not only in the literal information in the documents but in their clinical significance – not only what was done but also *why*. It is not enough to know that the report of a bone scan claimed "only osteoporotic changes". It is necessary to recognise that this indicates that there are "no bony metastases found". It is not enough to know that the patient was taken off chemotherapy, it is important to know what side effect or concurrent illness intervened.

Assembling the chronicle is therefore a knowledge intensive task that relies on inferences. The reliability of these inferences may vary, and it is essential to record not only the inferences but also the evidence on which they were based and their reliability. A graphical presentation of a chronicle developed manually as part of the requirements exercise for CLEF is shown in Figure 2. A human observer can quickly infer many of the reasons from the juxtaposition of events; an effective computer based "chronicle" must capture those inferences.

Query formulation, WYSIWYM and Information Generation

For the data in the CLEF Repository to be useful, it must be easily accessible to scientists and clinicians. CLEF is experimenting with a variety of textual and graphical query interfaces to the repository. However, the prime interface for researchers is being designed around techniques from language generation known as WYSIWYM –"What you see is what you meant"[2, 16]. An example is given in Figure 3.

The WYSIWYM interface allows users to expand a natural language like query progressively to produce queries of arbitrary complexity and then summarises the results, again in generated natural language.

[2] http://www.openehr.org/

[3] http://www.chime.ucl.ac.uk/~rmhidsl/EHRcomMaterials.htm

Query
Treatment profiles: Patients who received **[this type of treatment],** compared with patients who did not. *Outcome measure*: Percentage of patients alive after **[this interval of time]**.
Relevant subjects: Patients with **[this type of cancer]**
Answer|
It was found that out of 1790 patients diagnosed with cancer of the pancreas, 1300 had a pancreaticoduodenectomy and 490 didn't. Out of the 1300 patients who had a pancreaticoduodenectomy, 890 (68.46%) were alive after 5 years. Out of the 490 patients who did not have a pancreaticoduodenectomy, 87 (17.75%) were alive after 5 years.

Figure 3: Example of WYSIWYM query formulation and natural language response

Knowledge resources and Metadata

All the key technologies in CLEF are knowledge intensive. The overall approach in CLEF is based on "ontology anchored knowledge bases" – knowledge bases anchored in common conceptual models but conveying additional domain knowledge about the concepts represented. Examples include which drugs are used for which purposes, the significance of different results from different studies, the fact that a seemingly positive finding such as "evidence only of degenerative changes" may in practice convey the negative information that "no metastases were found". Some of this information exists in established resources such as the UMLS[4]. However, much of it needs to be compiled. CLEF works with both myGrid[5] and the new CO-ODE[6] project to developer-usable knowledge resources and tools.

The CLEF repository is intended to be more than simply a data collection. Rather it is intended, in the spirit of "collections based research and e-Science" to be a repository of both data and what the interpretations of that data by various researchers, their conclusions, and the methods they have used to achieve them. In this, it requires intensive metadata of at least five types:

- *Resource discovery information:* what is in the repository and what services does it provide
- *Provenance information:* where information came from, the evidence for any inferences, and the uncertainty of the information.
- *Usage and workflow information:* how the information has been used, including information allowing monitoring potential compromises of privacy
- *Consent and sensitivity information* about what information may be included in queries for different purposes.
- *Clinical significance and consequences:* why things were done and what they are believed to mean, always annotated by provenance metadata

The first three appear generic and analogous to metadata within other projects in e-Science and the semantic web, such as myGrid. CLEF also shares much in common with clinical trials, and some of the metadata schemas must take into account the emerging standards for clinical trial metadata[7]. The fourth and fifth types of metadata are more specific to CLEF's biomedical and care focus. CLEF will be actively promoting interchange standards in these areas.

e-Science Infrastructure

CLEF is building on and/or extending technologies and middleware components developed in myGrid[8] pilot and other e-Science projects in several areas. One area is related to the Grid based security framework. The role of privacy and security in handling clinical and other person-based information has become even more critical and central since CLEF was formulated because of external pressure in society. For CLEF to be usable and integrate-able within the e-Science infrastructure/Grid whilst meeting the clinical domain privacy and security stringent requirements, it is basing its technical solutions on the underlying Grid authentication, authorisation and access control services being developed in FAME-PERMIS[9], PERMIS [3] or other e-Science related projects. FAME-PERMIS is an authentication strength linked authorisation system. It has a multi-factor authentication model supporting a wide range of authentication methods including IP addresses, passwords, certificate-based soft tokens, and Java cards. The use of different authentication tokens imply different *authentication strength*, or the *Level of Assurance* (LoA), that enable multi-level authorisation and access control to the underlying information.

The second area is based around the use of technologies developed to support the e-Science lifecycle, notably technologies for workflows and provenance. Using these technologies, clinical scientists can make the most effective use of the

[4] http://umlsks5.nlm.nih.gov
[5] mygrid.semanticweb.org
[6] www.co-ode.org

[7] *e.g.* see http://www.cdisc.org/; http://ncicb.nci.nih.gov/core
[8] http://www.mygrid.org.uk
[9] Led by D. Chadwick, and N. Zhang, and funded by the Joint Information Systems Committee (JISC)

clinical resources being developed. Experience in myGrid has shown that adoption of automated workflows within a bioinformatics setting can drastically reduce the time taken to perform complex analysis and also aid the sharing and reuse of e-Science practice between scientific groups [18]. CLEF is currently addressing the challenge of adapting metadata, provenance and workflow technologies and methods from the relatively precise delimited world of molecular biology to the imprecise and much wider world of clinical practice. myGrid has a sophisticated model of provenance – who, what, where, why, when, how - metadata associated with every experimental entity represented by the Information Model[10], including components that can generate, store and visualise provenance represented in RDF [23]. Although properties are already in place in the model to hold the appropriate disclosure information, extensions are required to cope with the enhanced levels of privacy required by CLEF and to make use of that metadata when browsing the metadata and data in the myGrid Information Repository.

Another important area is to use Semantic Web and relevant Grid tools and technologies to support the CLEF chronicle described above. myGrid is already developing middleware components with which the scientist can directly interact. Domain specific ontologies written in the Ontology Web Language (OWL) allow resources to be described in terms that are familiar to the scientist. These ontologies and associated semantic web components pervade the middleware and applications, being used in repositories, registries and workflow environments.

Discussion

CLEF is aiming to contribute to a UK national infrastructure for advanced clinical trials and longitudinal studies using the emerging e-Science/Grid technology for distributed collaborative research and information sharing. Technologies developed in CLEF will enable a broad integration of clinical information from multiple sources and eventually aiming to joining up patient care with biomedical research. It builds on the basis of e-Science projects, such as myGrid and others, to bring their insights to the clinical domain. It is complementary with projects such as the National Cancer Tissue Resource and National Translational Cancer Network that focus more on actual specimens and genomic information *per se*. It seeks to provide clinical and knowledge resources that will be re-usable, for example within the broad framework being overseen by the National Cancer Research Institute (NCRI) and to lessen the barriers to using clinical information in collaborative research.

CLEF is playing a central role in addressing privacy and security issues with a group of cooperating projects within the e-Science initiative. Other projects, using clinical or other person-based information, include Integrative Biology (EPSRC e-Science pilot), proposed MRC e-Science projects PsyGrid, CancerGrid, and GEOGRAPHICAL would benefit from CLEF privacy and confidentiality policy framework, data and metadata management. Although current technical approaches for addressing privacy and security solutions are promising, developing a solution that enable distributed secure Grid-based computing and sharing of information remains a challenge for the whole GRID and e-Science community.

Current work CLEF is centred at the Royal Marsden Hospital Trust, one of the UK's premier cancer research centres. The next step is to broaden CLEF technology adaptability and scalability to include other trusts. Also planned evaluations include practical clinical trials at the London Institute of Genetic Medicine. Once CLEF developed privacy policies and security solution are evaluated, CLEF will seek ethical approval from the MREC to use live patients data.

Acknowledgements

CLEF is supported in part by grant G0100852 from the MRC under the e-Science Initiative. Special thanks to its clinical collaborators at the Royal Marsden and Royal Free hospitals, to colleagues at the National Cancer Research Institute (NCRI) and NTRAC and to its industrial collaborators – see www.clinicalescience.org.

References

1. Beale, T., Archetypes: Constraint-based domain models for future-proof information systems. in *OOPSLA-2002 Workshop on behavioural semantics*, (available from http://www.oceaninformatics.biz/publications/archetypes_new.pdf, 2002).
2. Bouayad-Agha, N., Scott, D. and Power, R. Integrating content and style in documents: a case study of patient information leaflets. *Information Design Journal*, *9* (2-3). 161-176.

[10] http://twiki.mygrid.org.uk/twiki/bin/view/Mygrid/InformationModel

3. Chadwick, D.W., A. Otenko, E.Ball. "Implementing Role Based Access Controls Using X.509 Attribute Certificates", IEEE Internet Computing, March-April 2003, pp. 62-69.
4. Cox LH. (2001) Disclosure Risk for Tabular Economic Data. In Confidentiality, Disclosure and Data Access (P. Doyle, J. Lane, J. Theeuwes and L Zayatz, eds) pp 167-183. Elseiver, Amsterdam.
5. Domingo-Ferrer J, Torra V. (2001) A quantitative comparison of Disclosure Control methods for Microdata. In Confidentiality, Disclosure and Data Access (P. Doyle, J. Lane, J. Theeuwes and L Zayatz, eds) pp 111-133. Elseiver, Amsterdam.
6. Elliot, M.J., Manning, A.M. and Ford, R.W. A computational algorithm for handling the special uniqueness problem. *Int Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, *10* (5). 493-511.
7. Friedman, C., Kra, P. and Rzhetsky, A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, *35* (3). 222-235.
8. Gaizauskas, R., Cunningham, H., Wilks, Y., Rogers, P. and Humphreys, K., GATE: An environment to support research and development in natural language engineering. in *Proceedings of the 8th IEEE International Conference on Tools with Artificial Intelligence*, (Toulouse, France, 1996), 58-66.
9. Gaizauskas, R., Demetriou, G., Artymiuk, P. and Willett, P. Protein structures nd information extraction from biological texts: The PASTA system. *Journal of Bioinformatics*, *19* (1). 135-143.
10. Gaizauskas, R., Hepple, M., Davis, N., Guo, Y., Harkema, H., Roberts, A. and Roberts, I., AMBIT: Acquiring Medical and Biological Information from Text. in *Second UK e-Science "All Hands Meeting"*, (Nottingham, 2003), (in press).
11. Humphreys, K., Demetriou, G. and Gaizauskas, R. Bioinformatics applications of information extraction from journal articles. *Journal of Information Science*, *26* (2). 75-85.
12. Ingram, D. GEHR: The Good European Health Record. in Laires, M., Ladeira, M. and Christensen, J. eds. *Health in the New Communications Age*, IOS Press, Amsterdam, 1995, 66-74.
13. Lin, Z., Hewett, M. and Altman., R.B., Using Binning to Maintain Confidentiality of Medical Data. in *AMIA Fall Symposium 2002*, (Austin Texas, 2002), Henry Belfus, 454-458.
14. Manning A. and Elliot MJ. (2003). Applying disclosure control to temporal data. To appear in Proceedings of Federal Committee on Statistical Methodology (FCSM) Research Conference. Washington DC. November 2003.
15. Murphy, S.N. and Chueh, H.C., A security architecture for querytools used to access large biomedical databases. in *AMIA Fall Symposium 2002*, (Austin Texas, 2002), Henry Belfus, 452-456.
16. Power, R., Scot, D. and Evans, R., What you see is what you meant: direct knowledge editing with natural language feedback. in *Proceedings of the 13th Biennial European Conference on Artificial Intelligence (ECAI-98)*, (1998), Spring-Verlag, 677-681.
17. Rector, A., Nowlan, W. and Kay, S. Foundations for an Electronic Medical Record. *Methods of Information in Medicine*, *30*. 179-186.
18. Stevens, R., Tipney, H. J. , Wroe, C., Oinn, T. , Senger, M., Lord, P., Goble, C., Brass, A., Tassabehji, M.. Exploring Williams-Beuren Syndrome Using myGrid. Accepted for Twelfth International Conference on Intelligent Systems for Molecular Biology (ISMB), Glasgow. July 2004.
19. Swanson, D.R. and Smalheiser, N.R. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, *91*. 183-203.
20. Sweeney, L. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, *10* (5). 557-570.
21. Taira, R.K., Bui, A.A.T. and Kangarloo, H., Identification of patient name references within medical documents using semantic selectionalrestrictions. in *Amia Fall Symposium*, (Austin Texas, 2002), Henry Belfus, 757-761.
22. Zhang, N., Shi, Q. and Merabti, M. Anonymous public-key certificates for anonymous and fair document exchange. *IEE Proceedings-Communications*, *147* (6). 345-350.
23. Zhao, J. Wroe, C., Goble, C., Stevens, R., Quan, D. and Greenwood, M. Using Semantic Web Technologies for Representing e-Science Provenance. 3rd International Semantic Web Conference (ISWC2004) Japan, 2004 (submitted)