Supporting Collaborative Grid Application Development within the e-Science Community

Cornelia Boldyreff, Phyo Kyaw, David Nutter, and Stephen Rank
Department of Computing and Informatics
University of Lincoln
Brayford Pool, Lincoln, LN6 7TS, UK
Email: {cboldyreff,pkyw,dnutter,srank}@lincoln.ac.uk

Abstract: Collaboration by use of common artifacts is at the core of e-science. A recent enabling technology is the Grid, which ties together heterogeneous computation and data resources through the use of middleware, linking the techniques and resources to infer higher-level knowledge. This article presents results from research and development of Grid technology for semantic interoperability between scientific artifacts on the web. The research employs the 'industry-as-laboratory' approach to software development. This means development of theory and models through successive implementations, their deployment in pilot studies and subsequent evaluation studies. The research is exemplified through the c ase of the OSCAR project, which is directed to the domain of bioinformatics.

Key words: Web collaboratory, grid technology, e-science

Reviewed and accepted 14 Jan 2004

1. Background

The systemic representation and organisation of software artefacts, e.g. specifications, designs, interfaces, and implementations, resulting from the development of large distributed systems from software components have been addressed by our research within the Practitioner and AMES projects [1,2,3,4]. Without appropriate representations and organisations, large collections of existing software are not amenable to the activities of software reuse and software maintenance, as these activities are likely to be severely hindered by the difficulties of understanding the software applications and their associated components. In both of these projects, static analysis of source code and other development artefacts, where available, and subsequent application of reverse engineering techniques were successfully used to develop a more comprehensive understanding of the software applications under study [5,6]. Later research addressed the maintenance of a component library in the context of component-based software product line development and maintenance [7]. The classic software decompositions, horizontal and vertical, proposed by *Goguen* [8] influenced all of this research. While they are adequate for static composition, they fail to address the dynamic aspects of composing large distributed software applications from components especially where these include software services. The separation of component co-ordination concerns from component functionality proposed in [9] offers a partial solution.

Recent research within the CoDEEDS project has made some progress towards the determination of design spaces to support both the static and dynamic system composition as well as the determination of the physical deployment and long-term operation of large distributed system composed from heterogeneous components [10]. Our current prototype implementation of collaborative support for the determination, elaboration, and evolution of design spaces, based on the CoDEEDS framework [11], employs at its base another development of our recent research within the GENESIS project, the Open Source Component Artefact Repository, OSCAR [12,13,14,15,16].

The GENESIS project's goal is to develop a generalised environment for process management in cooperative software engineering. A key component of this environment is an underlying distributed repository, OSCAR holding both metadata, describing software artefacts, and their contents. Artefacts within OSCAR range from software products, software processes, software tools to profiles of software developers, thus enabling rich relationships to be recorded between products and the tools used in their production, the processes that were associated with their production, and the developers who carried out these processes. XML DTDs are employed to describe the software artefacts within OSCAR. Castor[1] has been used to allow more flexible processing of the software artefacts as Java objects as well as XML documents. The recording within the metadata and subsequent processing of events related to the evolution of artefacts by awareness filters enables artefacts to *actively* inform interested parties of their evolution and deployment [17]. OSCAR is currently in the process of being released as a web-based service to support distributed developers in both industrial and Open Source Software development projects [18] and feedback from these trial users will drive its future development.

OSCAR's initial population for demonstration purposes has been derived from the Debian Open Source project and from the Java source code modules that comprise the OSCAR system itself. The latter artefacts have been derived using a Java source code import tool developed at Durham. The total population consists of just over 1500 software artefacts. This population with its extracted metadata has been employed in some experimental studies at Durham to gauge the effectiveness of using Self-Organising Maps (SOMs) to organise large collections of software artefacts in the GeniSOM project [19]. In GeniSOM, we have replicated Kohonen's original WebSOM[2] and extended it to the domain of software artefact collections. SOMs have been chosen to solve the problem of organising the incrementally expanding content of such a distributed repository, which we anticipate will result from the usage of OSCAR by a growing number of software development projects. More speculative use of SOMs applied to software artefact's relationship networks may be investigated in future to achieve optimal distribution of artefacts on the underlying OSCAR servers with respect to supporting the virtual user communities more effectively.

2. Proposed Programme of Research and Methodology

The principal aim of the research proposed is to develop new models to enable more effective collaborative design via shared active grid artefacts. More explicitly the research aims to achieve the following:

Semantic interoperability between scientific artefacts, including artefacts describing grid resources and computational services, across distributed computing platforms employed by the e-Science community such as various computational grids and the WWW.

Development of semantically-enabled systems and services facilitating scientific collaboration over the Web and Grid to support the design of Grid based scientific research: grid-based software applications, consolidated data sets, and experiments employing these applications and data sets.

Development of semantically based and context-aware systems to acquire, categorise, organise, process, share and reuse the knowledge embedded in the designs of scientific experiments and associated scientific artefacts.

To ensure accountability within and the replication of complete experimental design lifecycles within e-Science developments and

To maintain records of provenance with respect to associated scientific artefacts.

The research resulting in the OSCAR development has achieved many of these aims with respect to software artefacts in the context of collaborative software engineering. However, extending this work to more generally apply to scientific artefacts in the context of Grid based scientific research collaboration will constitute a major challenge. Currently, there is existing research within the myGrid[3] project directed at developing such support in the domain of Bioinformatics.

3. Objectives

Collaboration is a core activity at the heart of large-scale co-operative scientific experimentation. However, with the advent of large scale, distributed computational resources and the new affordances to scientific collaboration available in this context, specialised e-Science collaborative support is needed. In order to develop such support, it is necessary to develop new models of collaboration in this context and of the artefacts resulting from and deployed within such collaborations and resultant co-operative experiments. As these are all likely to evolve as the grid technology is more widely adopted, it is necessary to ensure that the modeling itself is done in an evolutionary manner. The modeling must take into consider the extensive legacy of the scientific community, their current software applications and data sets, recorded experimental designs, and current forms of collaboration. These models will be the basis of the initial prototype systems developed by the project. Deployment of these systems within pilot studies in the e-Science community will enable their refinement based on feedback from scientists and thus, further evolution of the models and systems.

Semantic interoperability between scientific artefacts will be achieved by developing an extensible form of artefact type descriptions building on our established research on describing software artefacts within OSCAR. Existing interchange standards and conventions employed by the e-Science community will be investigated. Ontology developments relevant to e-Science in the context of the Semantic Grid will be investigated and applied where relevant as the basis for semantic interoperability. Semantically enabled systems and services facilitating scientific collaboration over the Web and Grid will be developed to support the design of Grid based scientific research: grid-based software applications, data sets, and experiments employing these applications and data sets. These will be based on the initial modeling research and identified semantics relevant to describing the scientific artefacts in the e-Science context.

Key services to be developed will be ones which will support scientists in acquiring, organising, processing, sharing, and reusing the knowledge embedded in the designs of scientific experiments and associated scientific artefacts. It is hypothesized that the concepts

of design spaces and active artefacts coupled with the application of Self Organising Maps to organise the contents of the distributed artefact repositories can be developed to achieve the core functionalities required here. Accountability and replication of complete experimental design lifecycles within e-Science and establishment of records of provenance with respect to associated scientific artefacts will be ensured by maintenance of historical records. This will be achieved through instrumentation of the systems developed and by recording change and usage data within scientific artefacts in line with identified conventions within the scientific communities of practice studied by the project.

4. Methods

This research will employ the approach to software engineering research known as "industry-as-laboratory" with the e-Science community taking the role of industry, progressively refining its solutions based on feedback from the e-Science community. Initially we will collaborate with colleagues in the UK e-Science Centres and, as the project work progresses, with the wider e-Science community worldwide through the releases of our systems as open source.

There is a speculative element to the proposed research in terms of the innovative elements such as determination of new collaborative affordances and their implementation and the semantic modeling of the wide range of scientific artefacts likely to developed and deployed in the context of the grid. By adopting an incremental approach to the modeling and grounding it in progressive studies with feedback from the e-Science community, our approach will minimize risk and accommodate evolution of the models and associated systems development. The research proposed here follows from earlier work supporting roll designers in the steel industry and recent research to support collaborative software development by industry. The research activities will in part be forming abstractions over this earlier research in order to develop a customisable framework to support scientific collaboration in the context of e-Science. Our current research on software artefacts will be generalised and extended to accomplish the modeling of scientific artefacts. Our system implementations will follow from this framework and build on existing implementations carried out in our earlier research projects cited above.

5. Delivery to e-Science Community and e-Science programme

Within the University of Lincoln, the Faculties of Health & Life Sciences and Applied Computer Sciences provide an initial ideal basis for piloting the project's prototype systems as they are developed. As we plan to work with researchers from the Lincoln Science Faculties' departments in our initial studies, we will already have in place the groundwork for the initial delivery and trial usage of the prototypes. The feedback obtained through early deliveries will contribute to the evolution of the models and systems during the course of the research. Once the prototype systems have reached an appropriate stage of development, we will release them as open source and will carry out their further development as within the framework of an open source project. We will promote the project research and resulting software through the All Hands e-Science meetings regionally in the UK and through participation via national and international e-Science community web sites and events, such as the Grid Forum, as well as Software Engineering conferences and workshops.

6. Proposed Evaluation

The underlying theory and models developed throughout this research will be evaluated through the implementation of the framework described and deployed in pilot studies within the e-Science community. The initial prototypical implementation will demonstrate the feasibility of applying the concept of design space to the design of scientific experiments. An evolutionary approach will be employed with respect to the development of the proposed framework as feedback from initial pilot studies is obtained. Development of the theory and models will progress through successive implementations, their deployment in pilot studies and subsequent evaluation studies. The evaluation will employ an evaluation framework initially proposed by *Ramage* in the context of CSCW Evaluation studies which has been used to evaluate groupware for software engineering students at *Durham* [20] and extended by *Boldyreff* to be more generally applicable in the context of evaluating web-based collaboration [21]. Two key aspects of these extensions are the links made between evaluating usability and maintainability and the recognition that the context of usage must be conditioned by the users' goals in their primary collaborative work.

References

Boldyreff, Cornelia (2002) Determination and Evaluation of Web Accessibility, Proceedings of IEEE 11th Intl. Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), IEEE Computer Press.

Boldyreff, C, Burd, E.L, Hather, R.M, Mortimer, R.E, Munro, M, Younger, E.J (1995). The AMES Approach to Application Understanding: a case study, Proceedings of the International Conference on Software Maintenance, IEEE Computer Press.

Boldyreff,C, Burd, E.L, Hather, R.M, Munro, M,. Younger, E.J. Greater Understanding Through Maintainer Driven Traceability, Proceedings o Press, 1996.

Boldyreff, Cornelia (1992). A Design Framework for Software Concepts in the Domain of Steel Production, Proceedings of the Third International Conference on Information System Developers Workbench, Gdansk, 22-24 September.

Boldyreff, C. Elzer, P. Hall, P, Kaaber, U. Keilmann J, Witt, J (1990). PRACTITIONER: Pragmatic Support for the Reuse of Concepts in Existing Software, Proceedings of Software Engineering (SE90), U.K, Brighton, Cambridge University Press

Boldyreff, C., Kyaw, P., Nutter, D., and Rank, f the 4th Workshop on Program Comprehension, April 1996, pp. 100-106, IEEE Computer S.(2003). Architectural Framework For a Collaborative Design Environment, Proceedings of Second ASERC Workshop on Software Architecture (Banff, Canada, 2003).

Boldyreff, C., Kyaw, P., Nutter, D., and Rank, S.(2003). Rationale for a Metalevel Collaborative Design Environment, Proceedings of the Workshop on Advanced Collaborative Environments (WACE), Washington, USA, 2003.

Boldyreff, Cornelia, Lavery, Janet, Nutter, David, Rank, Stephen.(2003) Open-Source Development Processes and Tools, Proceedings of Taking Stock of the Bazaar: 3rd Workshop on Open Source Software Engineering, Portland, Oregon.

Boldyreff, Cornelia, Nutter, David, Rank, Stephen(2002). Architectural Requirements for an Open Source Component and Artefact Repository system within GENESIS, Proceedings of the Open Source Software Development Workshop, Newcastle upon Tyne, U.K. 25-26th February 2002, pp 176-196.

Boldyreff, Cornelia, Nutter, David, Rank, Stephen.(2002).Open-Source Artefact Management for Distributed Software Engineering", Proceedings of the 2nd Workshop on Open-Source Software Engineering at The 24th International Conference on Software Engineering in Orlando.

Boldyreff, Cornelia, Nutter, David , Rank, Stephen (2002).Active Artefact Management for Distributed Software Engineering", Proceedings of the Workshop on Cooperative Supports for Distributed Software Engineering Processes, in the Proceedings of the 26th IEEE Annual International Computer Software and Application Conference.

Boldyreff, C, Nutter, D, Rank, S, Smith, M Wilcox, P Dewar, R Weiss, D. Ritrovato, P (2003) Environments to Support Collaborative Software Engineering, Proceedings of 2nd Workshop on Cooperative Supports for Distributed Software Engineering Processes, CSSE 2003, Benevento, Italy

Boldyreff, C, Nutter, D, Rank, S, Smith, M Wilcox, P Dewar, R Weiss, D. Ritrovato, P (2003) Environments to Support Collaborative Software Engineering, Proceedings of 2nd Workshop on Cooperative Supports for Distributed Software Engineering Processes, CSSE 2003, Benevento, Italy.

Brittle, James Boldyreff, Cornelia (2003). Self-Organising Maps Applied in Visualising Large Software Collections, Proceedings IEEE VISSOFT.

Drummond, Sarah Boldyreff, Cornelia Ramage, Magnus (2001). Evaluating Groupware for Software Engineering Students, *Journal of Computer Science Education* 11,(1) 33-54.

Goguen, J.A, (1986). Reusing and Interconnecting Software Components, IEEE Computer, 16-28.

Joanna, Fyson, Boldyreff, M (1998). Cornelia. Using Application Understanding to support Impact Analysis, Journal of Software Maintenance: Research and Practice, 10, pp. 93-110, Wiley, 1998.

Kwon, O.C, Boldyreff, C, Munro, M (1997). An Integrated Process Model of Software Configuration Management for Reusable Components", Proceedings of the Ninth International Conference on Software Engineering & Knowledge Engineering (SEKE'97), June 18-20, Madrid, SPAIN.

Kyaw, Phyo, Boldyreff, Cornelia, Xu, Jie .(2002). Co-ordination Adaptors: The Evolution of Component-Based Distributed Systems, chapter in Systems Engineering for Business Process Change: New Directions (Volume 2), edited by Peter Henderson and published by Springer-Verlag, pp 298-308, 2002.

Nutter, D, Boldyreff, C,Rank, S (2003). An Artefact Repository to Support Distributed Software Engineering, Proceedings of 2nd Workshop on Cooperative Supports for Distributed Software Engineering Processes, CSSE 2003, Benevento, Italy.

Nutter, David, Boldyreff, Cornelia. (2003). Architectures for Awareness Support in Collaborative Software Engineering, IEEE Proceedings of WETICE.

Zhang, Jian Boldyreff, Cornelia (1990). Towards Knowledge-Based Reverse Engineering, Proceedings of the Fifth Annual Knowledge-Based Software Assistant Conference, Syracuse, NY, 24-28.