

NCRI Informatics Task Force

Project Proposal: Construction of a Platform Reference Model

Summary

This proposal outlines a short project necessary to support the work of the Informatics Task Force. The proposal summarises the steps necessary to construct a *Platform Reference Model* and to validate this reference model with respect to a set of Use Cases. The Platform Reference Model will provide a shared basis for understanding the key components of the information sharing and services platform envisaged in the NCRI informatics strategic framework. It will also provide a coherent basis for bringing together existing data sharing schemes. The Use Cases will describe how the platform is to be used and how it will deliver value to researchers and clinicians. The objectives, approach, work plan and team are outlined below. An estimate of the resources required to support the project is given, as are the proposed management arrangements for the project.

Objectives

The aims of the NCRI Informatics Task Force are to enable the development of an “informatics platform in the UK that facilitates access to and movement of, data generated from research funded by NCRI Partner organisations, across the spectrum from genomics to clinical trials”.

The first steps in achieving this aim must be:

- to understand in a *systematic manner* what goals such a platform is required to serve;
- to understand what the key components of the platform are;
- to understand, in terms of these components, what resources are available that can contribute to the construction of the platform.

The objectives of this project are to achieve the first two of these steps and to substantially enable the third.

Approach

Taking the ‘NCRI Strategic Framework for the Development of Cancer Research Informatics in the UK’ as its starting point the project will identify a set of key Use Cases for the projected platform. These Use Cases will be anchored in scenarios that are developed with, and validated by, cancer researchers and clinicians. The Use Cases will support the development of a Platform Reference Model. This Platform Reference Model will be composed of two parts:

A coarse grain ‘domain model’ covering the area of cancer research

informatics.

An 'infrastructure model' that identifies the key components of projected platform and their relationship to each other.

Both parts will be carefully aligned with existing models for data sharing and GRID/escience reference models. The key application of the Platform Reference Model will be to provide a common language and focus to the efforts of the NCRI Informatics Task Force and projects that spin-off from it. For this reason the model is required promptly and at the early stages of our collective work. The model will serve as a superstructure on which to develop a full account of the requirements for the platform. The model, once complete and validated, has two additional applications:

As a model for a meta-data repository, complementing the planning matrix, for use by the NCRI Informatics Unit and by the community.

The planning matrix constructed by the NCRI Informatics Unit presents a two dimensional view of the current state-of-the-art in cancer informatics in the UK. On the horizontal axis are the key communities of interest, loosely ranged left to right by dominant scale of concern. On the vertical axis are broad areas in which these communities have made progress on matters related to information access and sharing. The degree of progress is colour coded. The matrix provides a powerful quick glance summary, valuable for planning, but has limited application beyond this bundling together as it does standards, projects, specific databases etc. A simple metadata repository would be constructed within the framework of this project that would implement the Platform Reference Model (both the domain and infrastructure models). This repository could then be populated and used alongside the matrix to track ongoing work. As work on the platform progresses the repository could serve as the forerunner of service and component broker.

As the basis for a meta-data scheme that can be used to annotate and orchestrate cancer informatics resources;

We anticipate that a meta-data scheme that implements, in significant part, the Platform Reference Model (and is consistent with the repository) would be used actively by those who have implemented services, information resources and similar in the 'cancer research space'. Components would be annotated using the meta-data scheme provided as an XML language. Such a language would be developed and trialled on existing resources within the project and rolled out to new projects such as CancerGrid and the NCRN electronic remote data capture project. It should be stressed that a Platform Reference Model is an *essential* element in the orderly and well-managed conduct of a large system integration and development initiative of the type envisaged by the NCRI. Good systems engineering practice, and experience from the current e-science projects, clearly and unequivocally demands that at the earliest stages of a project requirements

are gathered and managed alongside a model of both domain and architecture that is owned by the clinicians and cancer researchers.

Progress

To demonstrate both feasibility and immediate value a simple infrastructure model has been prepared by the proposal team. It has been, loosely, tested against items in the planning matrix. This model has been favourably received by the Task Force and constitutes a good working basis for this project. It is included below as an Appendix. In short, we have a running start. The resources required to complete this project are however, beyond the scope of the volunteer effort deployed so far. Necessarily, good modelling demands substantial validation effort. A full and wellfounded modelling method must be deployed, at minimum an industry-standard approach such as UML. The implementation of both the repository and annotation scheme, though not inherently difficult once the Platform Reference Model has been agreed, are somewhat time consuming.

Relationship with Other Projects

This project is strongly related to, and complementary with, the proposed Integrating Pathology and Radiology Imaging Data demonstrator. The demonstrator project will, by taking a relatively narrow, but important slice across the space represented by the NCRI planning matrix provide an important test of the feasibility of the integration that is envisaged by the NCRI informatics Initiative. It will in addition provide us with an improved understanding of the issues that are to be faced by the Informatics Task Force. The project outlined in this proposal will provide the context for the demonstrator and the means by which it can be integrated within the larger platform.

Workplan

Because of the importance of ensuring the Platform Reference Model is in place at a relatively early stage in the work of the NCRI Informatics Task Force we are proposing to divide the work in four Packages summarised below. Two Research Fellows will be required, each for a period of 6 months with the appointment of the second Research Fellow (RF2) lagging the first (RF1) by 3 months. Note that these 3 months coincide with the initial 3 months of the demonstrator for which resources have been requested within that project. The overall project will thus take 9 months with key milestones at 3, 6 and 9 months. This structure should ensure that the work is distributed and is used as soon as possible. Packages run in parallel as set out in the table with distinct deliverables for each Package. RF1 picks up the modelling work while RF2 is more concerned with the implementation.

Package 1

Task 1.1: Gather scenarios

Task 1.2: Build Use Cases

Task 1.3: Develop 'quick and dirty' Use Case simulations

Package 2

Task 2.1: Review GRID/e-science reference models and test against Use Cases

Task 2.2: Validate Use Cases

Task 2.3: Test domain and infrastructure models against Use Cases

Task 2.4: Instantiate models using data from matrix and other resources

Task 2.5: Populate repository with skeleton data

Task 2.6: Annotate exemplar components using meta-data scheme

Package 3

Task 3.1: Build domain model

Task 3.2: Build infrastructure model

Package 4

Task 4.1: Construct repository

Task 4.2: Construct meta-data scheme

Package 5

Task 5.1: Website (at <http://www.cancerinformatics.org.uk>) and dissemination

Who

This work will be conducted under the supervision of Prof Anthony Finkelstein (UCL) jointly with Prof Jeff Kramer (Imperial College), Dr Helen Parkinson (EBI) and Dr Fiona Reddington who will in addition provide liaison with the Informatics Coordination Unit. Profs Finkelstein and Kramer work together as part of London Software Systems a newly established joint Institute of UCL and Imperial College in the area of software engineering. The project will report progress to a Project Board constituted as a sub-group of the NCRI Informatics Task Force chaired by Prof Richard Begent. Other members of the Project Board would include Prof David Gavaghan, Prof David Ingram, Prof John Fox and Dr Alvis Brazma.

Prof Anthony Finkelstein

Anthony Finkelstein is Professor of Software Systems Engineering and Head of the Department of Computer Science at UCL. He is an international research leader in the broad field of software systems engineering. His research has largely been in the area of software development methods and tool support. His current research includes work on the construction of large-scale data grids and on the development of scalable modelling technologies to support systems biology. He has published more than 150 papers and held research grants totalling in excess of £10m. A list of publications is available at <http://www.cs.ucl.ac.uk/staff/A.Finkelstein>. He is a Fellow of both the IEE and BCS. In 2003 he published in both ACM Transactions on Software Engineering and Methods (TOSEM) and at the International Conference on Software Engineering, the leading venues in the field. In 2003 he was a joint winner of the prestigious ICSE 'most influential paper' prize for work on 'viewpoints'. In 2004 he was joint winner of the first RE 'most influential paper' prize for work on requirements traceability. He has served on numerous editorial

boards including that of ACM TOSEM and was founder editor of Automated Software Engineering. He is currently a member of the editorial board of IEEE Transactions on Software Engineering. The 'state-of-the art' review he edited remains the publication with the highest impact factor in software engineering. He has chaired numerous international meetings and was General Chair of the International Conference on Software Engineering 2004. He has also been an invited speaker at many meetings. Most recently he was keynote speaker at Automated Software Engineering 2003 in Montreal, Canada and at SBES 2004. He is currently Chair of IFIP WG 2.9 (Software Requirements Engineering). He established a leading research group in software systems engineering at UCL and played a key role in the foundation of London Software Systems a joint Institute of UCL and Imperial College in the area of software engineering. Anthony Finkelstein is a founder of Systemwire, a UCL spinout company.

Prof Jeff Kramer

Professor Jeff Kramer is Professor of Distributed Systems and leads the Distributed Software Engineering Group at Imperial College. Until September 2004 he was Head of the Department of Computing. Jeff Kramer is Associate Director of London Software Systems. His research interests include requirements engineering, software architectures and analysis techniques, particularly as applied to concurrent and distributed software. He was a principal investigator in the various research projects which led to the development of the CONIC environment for configuration programming and the Darwin architectural description language which is used in commercialised form by Philips for the software for high end television sets. His current research work is on behaviour analysis, the use of models in requirements elaboration and architectural approaches to self-organising software systems. Jeff Kramer is a Chartered Engineer, Fellow of the IEE and Fellow of the ACM. He was program co-chair of the 21st ICSE (International Conference on Software Engineering) in Los Angeles in 1999, Chair of the Steering Committee for ICSE from 2000 to 2002, associate editor and member of the editorial board of ACM TOSEM from 1995 to 2001 and is currently associate editor and member of the editorial board of IEEE TSE. He was winner of the Most Influential Paper Award at ICSE 2003. He is co-author of a recent book on Concurrency, co-author of a previous book on Distributed Systems and Computer Networks, and the author of over 150 journal and conference publications.

Dr Helen Parkinson

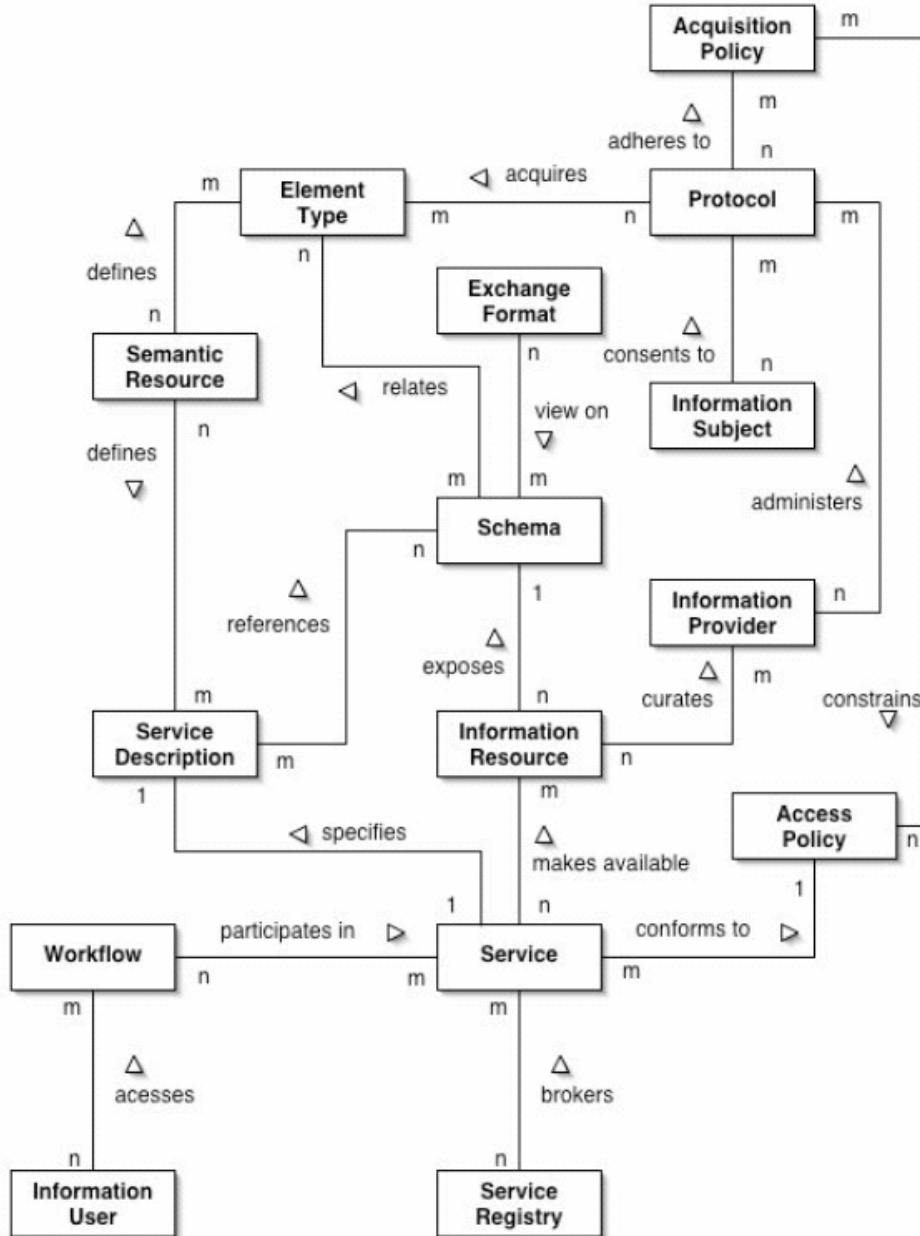
Helen Parkinson is a staff member at the European Bioinformatics Institute. She is the Curation Coordinator for the ArrayExpress database run the by the Microarray Informatics Team. She is also a Scientific Programme Manager for the National Cancer Research Institute (NCRI) Informatics Unit. She has a degree in Biochemistry and Genetics from University of Leeds; a Ph.D. in Genetics from the University of Leicester and prior to moving the EBI performed post-doctoral research in eukaryote genetics. She moved to the EBI in 2000 to

work as a data curator for the EMBL database. She is currently Head of Production for the ArrayExpress database at the EBI and manages a team of data curators. She is a member of the Microarray Gene Expression Data Society (MGED) board and is a member of the MGED Ontology Working Group that developed the MGED Core Ontology for microarray experiments. She teaches regularly on EMBO courses. She is also a co-organizer of the Standards and Ontologies for Functional Genomics Conference (SOFG) <http://www.sofg.org>

Dr Fiona Reddington

Fiona Reddington is a Scientific Programme Manager with the NCRI Cancer Informatics Initiative. She has a degree in Pharmacology from University College Dublin and a Ph.D in Neurophysiology from Guys, Kings and St Thomas (GKT) Medical School. After GKT, she worked as a Project Manager for Cancer for the UCL Clinical Research Network before becoming Centre Manager of the UCL NTRAC Centre. She was also seconded to the NTRAC Coordinating Centre in Oxford where she was responsible for Bioinformatics and involved in the development of the information system to support the National Cancer Tissue Resource (NCTR). She has previously developed ontologies to describe clinical networks and is involved in the development of a cancer therapy ontology at the Royal Free Hospital.

Appendix: Draft Infrastructure Model



Model Key

Information User — The agent that requires access to information or analytical support.

Workflow — A connected set of tasks that must be performed in order to meet the needs of an information user.

Service — A loosely coupled software component that provides a set of contractually defined behaviours accessible from a published interface..

Information Resource — A repository of managed information such as a relational database or collection of flat files.

Service Description — A description of a service that formally specifies behaviour of that service.

Service Registry — A meta-service that publishes and locates services.

Semantic Resource — A resource that defines the meaning of data element types and service descriptions (typically an ontology or controlled vocabulary).

Element Type — A class of 'things' relevant to the domain of discourse.

Information Provider — An agent that makes information available.

Information Subject — An agent (generally a patient or animal) about whom data is generated.

Schema — An abstract description of the structure and organisation of an information resource.

Exchange Format — A format in which data is encoded, for the purposes of exchanging data between organisations.

Protocol — Definition of how an experiment, part of experiment or other means of data acquisition should be performed so as to ensure the data is valid.

Acquisition Policy — Policy that prescribes how, and by whom, data may be properly acquired, covering such matters as consent.

Access Policy — Policy that prescribes how, and by whom, information may be properly accessed.

Draft Model with Examples

