

Integrating Pathology and Radiology Imaging Data A demonstrator project for the NCRI

Professor Sir Michael Brady, Professor Anthony Finkelstein, Professor David Gavaghan,
Professor Phil Quirke, MERCURY Group, Dr Fiona Reddington

1 Introduction

It is proposed that members of the NCRI Task Force, in collaboration with the NCRI Informatics Unit, and additional scientists both in the UK and elsewhere, develop a demonstrator project that showcases the potential of the NCRI and which results in a prototype of a Grid-enabled service-oriented informatics system that will be of immense value to scientists and clinicians from the many disciplines in which cancer is studied. The NCRI Informatics Task Force aims to maximize the impact of cancer research by application of informatics. Fundamentally this requires compatible systems for handling and sharing data of each type and for integrating different types of data. The aim is to engage the research community in developing the systems, which can then be widely used, creating an ever growing mass of shared, accessible and re-usable data.

2 The Imaging Demonstrator

This demonstrator comes from radiological, microscopy, clinical trials, computer science and integrative biology communities. It develops a framework which re-uses and adapts systems developed for imaging of breast cancer and applies them in rectal cancer, integrating magnetic resonance imaging (MRI) information with macroscopic, microscopy and data from a clinical trial. After testing and validation in this application, the framework will be made available in an open source format to the research community so that it can be used with comparable datasets or adapted for use with other imaging modalities or types of clinical data. In this way it will be a critical tool to enable implementation of the NCRI partners commitment to data sharing, re-use and integration.

2.1 The Subject of the Demonstrator

The subject of the demonstrator is the integration of imaging data from pathology and radiology in rectal cancer. This will add value to the existing tissue and image collections associated with the NCRN-accredited MERCURY trial by archiving the data and ensuring its preservation. It will also provide archived digital images which may be used for quality assurance purposes and teaching. The demonstrator will enable studies which will examine how prognostic and diagnostic features which are apparent in histopathological sections and clinical scans can be related and meaningfully linked into a diagnostic or predictive profile using informatics. This initial study will also begin to deliver the standardisation of data elements, vocabularies, ontologies and exchange formats within and between the Pathology and Radiology communities.

Achieving the effective exchange and integration of data from these diverse disciplines will be an initial step in enabling the comparison of clinical data across spatial and temporal scales in a way that is not currently possible. By taking a relatively narrow, but important slice across the space represented by the NCRI planning matrix (figure 1) this demonstrator project will 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.

	DNA	Functional Genomics	Cytogenetics	Proteomics	Pathophysiology & Visualisation Techniques	Therapeutics	Animal Models	Clinical Trials & Longitudinal Studies	Epidemiology & Population Studies
Data Elements	Yellow	Green	Yellow	Yellow	Green	Yellow	Yellow	Green	Yellow
Controlled Vocabularies & Ontologies	Yellow	Yellow	Red	Yellow	Green	Yellow	Yellow	Green	Red
Data Exchange Formats	Yellow	Yellow	Yellow	Yellow	Green	Yellow	Yellow	Green	Yellow
Protocol Standardisation	Yellow	Yellow	Yellow	Yellow	Green	Yellow	Yellow	Green	Red
Implementation	Yellow	Yellow	Yellow	Yellow	Green	Yellow	Yellow	Green	Yellow
Data Mining	Yellow	Yellow	Yellow	Yellow	Green	Yellow	Red	Green	Yellow
Privacy Enhancing Technologies / Security	Yellow	Yellow	Yellow	Yellow	Green	Yellow	Yellow	Green	Yellow
Knowledge Management	Yellow	Yellow	Yellow	Yellow	Green	Yellow	Yellow	Green	Yellow

Figure 1. Areas of matrix demonstration project would integrate

2.2 Design

The objective of this project is to demonstrate the utility of the multi-scale, multi-disciplinary approach to enhancing the information that can be derived from data collected within a clinical trial, not least by facilitating its linkage to data of the same type from previous studies and to data of different types. A fundamental requirement of this type of work is a domain object model for the area represented by this demonstrator. This model can be linked to the overall platform reference model that delineates the scope of the informatics platform. Extensive work has already been undertaken, by the Task Force and Coordination Unit, to produce a draft platform reference model that can serve as a starting point for this project (Appendix 2).

Given this model and the use of existing and adapted vocabularies, our primary aim is to federate histopathology, radiology, photographic, and clinical trial data (from the phase III MERCURY trial www.pelicancancer.org/researchprojects/mercury.html) from several disparate geographical sources into a single, virtual database by leveraging state-of-the-art techniques developed within the e-DiaMoND e-Science project (Gilbert *et al.*, 2004; Berman, Fox and Hey, 2003). The principal aim of the e-DiaMoND project is to develop a federated database of mammograms and associated applications that are sympathetic to the work practices and needs of the NHS Breast Screening Programme. The e-DiaMoND prototype system leverages commercial off-the-shelf products, open source software and bespoke code and has been successfully deployed at four sites. The underlying architecture has been designed to be generic and extensible to other modalities and clinical domains. In this NCRI demonstrator project we will adapt the existing e-DiaMoND architecture to provide appropriate data aggregation and federation services. We will also develop underpinning imaging techniques, through mathematical analyses, to relate diagnostic parameters and features within the various images and data sets to transform that *data* into clinically relevant, patient-specific *information* of direct benefit in diagnosis and treatment.

3 Project Plan

Stage 1

Build on the existing work already undertaken by the Task Force and Unit to refine and validate core Use Cases for the project via continued consultation with user communities



Stage 2

Develop a domain Object Model for pathology and radiology based upon the MERCURY trial. This will utilize existing ontologies and standard vocabularies and will build upon the work undertaken in Stage 1. A series of workshops and meetings will be organised aimed at delivering community-wide acceptance of the model. Given the scope of the demonstrator this model will be necessarily limited and will be developed in such a way as to be consistent with the Platform Reference Model being developed by the Task Force and Coordination Unit



Stage 3

Storage of images

This will require agreement on the meta-data (information about the images) that should be stored and database schemas and services will need to be designed and agreed.

The hardware to host such a database will need to be identified. Non-functional constraints such as security, interoperability, scalability, manageability, and dependability, will also be addressed.



Stage 4

Electronic/GRID enabled transfer of images and associated meta-data

This will require the use of agreed data exchange formats to allow researchers to deposit their images in the database and to view images already stored in the database.

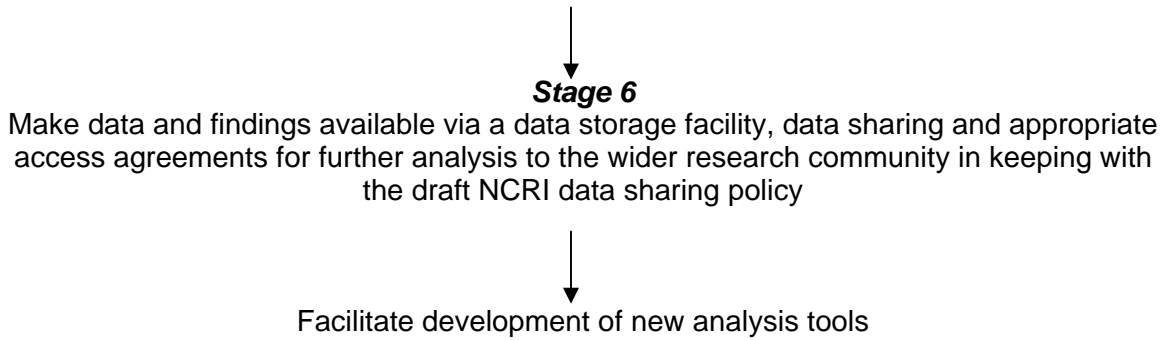
Appropriate security mechanisms will also be identified. The project will extend the approach to DICOM used in the e-DiaMoND project to images of different types.



Stage 5

Develop methods for comparison to other data

This will require standard descriptors between data types and we will design a framework in which algorithms for image standardization (e.g. the SMF algorithm for mammography) and/or standard protocols for image formation (e.g. in PET, MRI) can be incrementally incorporated into the framework. For the demonstrator we will import algorithms from the e-DiaMoND project, the Medical Images and Signals to Clinical Information (MIAS Interdisciplinary Research Consortium) and from the National Library of Medicine Insight Segmentation and Registration Toolkit (ITK) website. This will allow for the comparison of digitised pathological and radiological images and for the development of mathematical algorithms to describe specific features within the images (for further information please see Appendix 1)



	0 months	3 months	6 months	9 months	Ongoing
Stage 1					
Stage 2					
Stage 3					
Stage 4					
Stage 5					
Stage 6					

Figure 2. Flow chart of project plan and timeline

4 Deliverables

Responsibility for the following deliverables is as follows:

Oxford, Computing Laboratory: David Gavaghan and Andrew Simpson

Oxford, Engineering: Mike Brady

IBM Hursley: David Watson

Leeds, Pathology: Phil Quirke

UCL, Computer Science: Anthony Finkelstein

1. Validated Use Cases and domain object model (UCL Computer Science)
2. Agreed ontologies, vocabularies and data exchange formats based on the MERCURY trials (Oxford, Computing Laboratory)
3. Database schemas for each imaging modality (Oxford, Computing Laboratory)
4. Adapted e-DiaMoND architecture (Oxford, Computing Laboratory)
5. A secure Grid-enabled, federated database deployed at four sites (Oxford, Computing Laboratory)
6. Grid monitoring and administration software adapted from the e-DiaMoND system to allow on-the-fly grid configuration (e-DiaMoND consortium)
7. MRI, macroscopic (Tiff/JPG) and digital histological slides of rectal cancers (MERCURY trial, Leeds, Marsden)
8. Image analysis algorithms for data aggregation and transformation to useful information (Oxford, Engineering)
9. Wrapping of image analysis algorithms as web services so that they can be invoked in real time within the system (e-DiaMoND consortium)
10. A prototype demonstrator (All project partners).

In day-to-day terms delivery of the above might, for example, enable the following;

A query is made to the system for patients meeting certain criteria. Images are retrieved for the relevant set of patients in real-time from the individual databases (possibly behind local firewalls). A single patient is chosen and all relevant images displayed. Services are invoked to undertake a 3D reconstruction of the macroscopic resected rectum. Initially this will be from the photographs but will be extended to also build in the virtual histology allowing correlation of in vivo pathology with the cellular pathology. A diagnostic feature of tumour close to the margin or suspected high risk features such as venous invasion or peritoneal invasion is highlighted in one image and a further service is invoked to identify the same feature (and highlight) in other images and within the 3D reconstruction. It would be advantageous if pre-treatment scans and post treatment scans could be fused with differences automatically highlighted. The data needs to be seen at Leeds, the Royal Marsden, Basingstoke (Pelican) and Oxford. Patient identifiers will be removed and substituted with trial code number held only by the trials office.

5 Existing Resources

[e-DiaMoND](http://www.ediamond.ox.ac.uk) (www.ediamond.ox.ac.uk)

The underlying architecture the e-DiaMoND project has been designed to be generic and extensible to other modalities and clinical domains. In this NCRI demonstrator project we will adapt the existing e-DiaMoND architecture to provide appropriate data aggregation and federation services.

[CancerGrid](http://www.cancergrid.org) (www.cancergrid.org)

The project will ensure complementarity with the MRC funded CancerGrid project which aims to develop open standards for cancer clinical informatics, work on data integration and use Grid-enabled collaborative tools to manage clinical trials. The project also intends to collaborate with the PathGrid project which is in preparation. The integration of such projects to cover an increasing area of the planning matrix is one of the strategic aims of the Initiative. The investigators of both this demonstrator project and PathGrid have agreed collaboration on the workplan listed here.

[European Grid of Solar Observation \(EGSO\) project](http://www.egso.org) (www.egso.org)

The demonstrator project will have access to resources from the European Grid Of Solar Observations project (EGSO), a major European data grid project in which UCL is a key partner that also supports federated image databases (albeit in a different domain).

[The MERCURY Trial](http://www.ncrn.org.uk/portfolio/data.asp?ID=1192) (www.ncrn.org.uk/portfolio/data.asp?ID=1192)

MERCURY is an NCRN adopted trial of rectal cancer to test the hypothesis that MRI with a body coil and precise positioning can predict the subsequent pathology, especially the presence of a complete resection. This is defined as the presence of tumour within 1 mm of the pathological circumferential margin. There has been training of the radiologists and histopathologists and all data has been collected by proforma and is available. The trial is in follow up but the early results have been announced and prove that MRI can predict likelihood of clearance.

The trial has collected both MRI scans and photographs, which are available for use in the demonstrator project, and that may be amenable to 3D reconstruction. Large mount sections have been cut as well as standard histology and are also available to the

demonstrator project. The Principal Investigator of the MERCURY trial has given their approval for the data to be used in this demonstrator.

MRI's from the trial would be entered into the proposed database via Oxford, histopathology slides would be scanned in Leeds, photographs are digital and would need to be made available at the Marsden Hospital, Leeds, Oxford and the Pelican centre in Basingstoke (the coordinating centre for the MERCURY trial).

The Integrative Biology Consortium (www.integrativebiology.ox.ac.uk)

The Integrative Biology consortium is developing a service-based infrastructure leveraging middleware developed within the UK e-Science Programme to facilitate the world-wide research efforts in complex systems biology. This involves providing user-friendly access to tools enabling access to: high-performance computing facilities, state-of-the-art collaborative visualization applications, and a comprehensive suite of data management services. The aim is to provide to the experimental scientist the capability of performing in silico experiments as a routine tool to support and enhance laboratory-based experiments. Via an iterative interplay between mathematical modelling, computer simulation and wet-lab experiments, the ultimate aim is to enhance the ability of the researchers to meet the post-genomic goal of biological function determination.

Other projects in the NCRl Planning Matrix

The project will also ensure close links with the other projects listed in the NCRl Planning Matrix (www.cancerinformatics.org.uk/planning_matrix.htm) especially the Clinical e-Science Framework (CLEF) project which is developing security and confidentiality solutions for use in the clinical domain.

This project not only utilizes the strengths of Oxford (software engineering, e-Science, radiology networking, medical imaging), Leeds (pathology scanning and interpretation), Marsden (radiology interpretation), UCL (software engineering and GRID technologies) and the Pelican Centre in Basingstoke but links to a phase III NCRN trial, the national MDT colorectal cancer training programme (NHS), a DH funded scanning project and 3 NTRAC centres (UCL, Oxford and Leeds Bradford).

6 Resources requested

7 Conclusion

The aims of the NCRl 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 NCRl Partner organisations, across the spectrum from genomics to clinical trials". A prototype platform for achieving this has been devised and requires development and investigation with real data. This demonstrator project will use radiological, macroscopic, microscopic and clinical trials data to test and investigate the platform. It feeds on components on which much work has already been done and illustrates how value can be added by appropriate use of informatics. Potential benefits in this application could include improved staging and prediction of prognosis of cancer, characterisation of novel high resolution imaging techniques in relation to microscopy and the application of functional imaging to investigation of the mechanisms of novel

therapeutics. Broader benefits will accrue by taking these first steps towards developing an informatics platform that can be adapted for use with different types of data across the full range of cancer research.

References

Fiona Gilbert, Sharon Lloyd, Marina Jirotko, David Gavaghan, Andrew Simpson, Ralph Highnam, Tom Bowles, David Schottlander, David McCabe, David Watson, Brian Collins, John Williams, Alan Knox, Manfred Oevers, Michael Brady, Paul Taylor

EDiaMoND: the UK's Digital Mammography National Database

7th International Workshop of Digital Mammography, 18-21 June 2004, Chapel Hill, NC, USA

Michael Brady, David Gavaghan, Andrew Simpson, Miguel Mulet Parada and Ralph Highnam

EDiamond: a Grid-enabled federated database of mammograms, in Fran Berman, Geoffrey Fox and Tony Hey (eds.), Grid Computing: Making the Global Infrastructure a reality

Wiley, 923-944, 2003

Appendix 1 Complex Systems Biology: An integrative approach

The biological sciences have undergone a revolution over the last decade. Advances in biotechnology, underpinned by the massive leap in computational resources, have provided a wealth of biological data at all levels of biological organisation. At the molecular and cellular levels the various genome and proteome projects, coupled with advances and innovations in microscopy and biological imaging, have provided unprecedentedly detailed descriptions of the constituent parts and basic structures of living organisms.

The ultimate goal of this research and data collection programme is the determination of *biological function*. The bulk of this data, particularly at the molecular level, is obtained from in vitro, and usually static, laboratory experiments: where dynamic changes are considered, experimental complexity usually restricts observations to single, or very restricted, spatial and/or temporal scales. However, biological systems consist of myriad complex interactions between non-linear processes occurring on widely differing *spatio-temporal scales*, so that biological and physiological function emerges only through the *dynamic* interplay between these processes at all levels of organisation from molecular through to whole organism and environmental. A full understanding of biological function can therefore be gained only if we are able to integrate all relevant information at multiple levels of organisation to recreate these dynamic interactions. This, in general, cannot be done purely by experimental observation\footnote {It is now thought that the Human Genome contains approximately 40,000 genes coding for approximately 100,000 proteins. If we were to consider experimentally only pairwise interactions between these proteins there would be of the order of 5 times 10⁹ possibilities.}.

Whilst the molecular biologist can, after detailed experiment, gain some insight into the genetic pathways involved in such complex processes, and the biochemist might determine the relevant signaling pathway and encompass it within static diagrams, these descriptions will not yield a full understanding of biological function. Reducing a complex system down to its constituent parts yields only partial information, and the only feasible approach to recreating the dynamic interactions from which function emerges is to develop mathematical and computational models. An iterative process between experiment and modelling then provides descriptions of biological processes with the dynamic complexity to give the required biological function. Such models also have predictive power - giving virtual cells, tissues, organs and systems - which can be used in the development of novel drugs and treatments, and ultimately for patient-specific care regimes.

A key issue in furthering this approach is how such models can be validated. This is done by comparison with experimental data (both in vivo and in vitro) but again this is often static. A promising alternative is to make use of imaging technologies (particularly MRI, ultrasound, and, of increasing importance, PET), to provide dynamic temporal data across a range of spatial scales. Within Oxford, much work is being done in this area, and strong collaborations exist between the groups of Professors Brady and Gavaghan, with joint projects aimed at modelling both the underlying physiological processes (for example, acid production by tumours, or the elastic properties of tumours), and matching these to the physics of the imaging process (in these cases PET and Ultrasound, respectively). This approach potentially gives the means of quantifying physiological

changes both spatially and temporally, greatly increasing the diagnostic information available to the clinician. Ultimately, our aim to be able to compose these various models of differing imaging modalities, again advancing diagnostic tools and bringing a multidimensional and multidisciplinary approach to the assessment of disease.

Appendix 2 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/e-science 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 meta-data 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 trialed 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 well-founded 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

Month	1	2	3	4	5	6	7	8	9
P 1	1.1	1.2		1.3					
P 2			2.1	2.2	2.3	2.3	2.4	2.5	2.6
P 3				3.1	3.2				
P 4						4.1	4.1	4.2	
P 5	5.1	5.1	5.1	5.1	5.1	5.1	5.1	5.1	5.1
Notes	<i>RF1 starts</i>			<i>RF2 starts</i>		<i>RF1 stops</i>			<i>RF 2 stops</i>

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 scaleable 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

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>

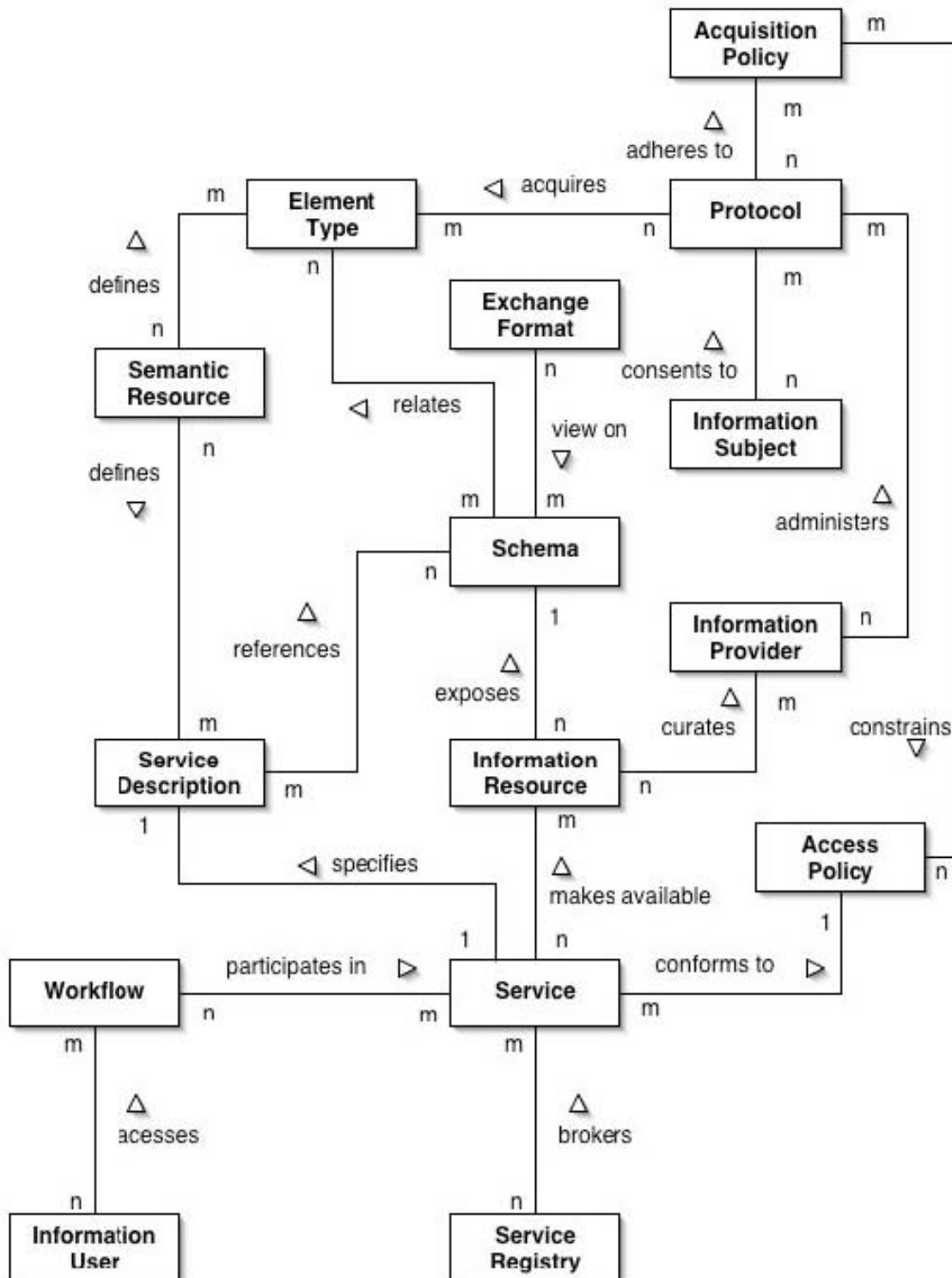
Dr Fiona Reddington

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 asa 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.

Funding

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

