



eResearch Open Forum eResearch – Col-laboratories and Curation

John Taylor

Presentation to DEST researchers
Canberra, Australia
20 September 2005

eResearch – Col-laboratories and Curation

Outline

This talk will be mainly about *what* and *why*, rather than *how*
- there are lots of experts on the how ...

- John Taylor - background
- e-Science, e-Research
- Grids and information utilities
- UK e-Science program
- Col-laboratories & Curation
- e-Research programs in Australia
 - objectives, needs, policy framework, questions

John Taylor - Background

- Director, Hewlett Packard Laboratories, Bristol, UK, 1984-1998
- Director General of Research Councils,
UK Office of Science and Technology, 1999-2003
- Chairman, Roke Manor Research, 2004 -
- Non Exec Director, Rolls Royce, 2004 -

Director General of Research Councils, UK Office of S&T
-responsible for UK Science Budget & UK Research Councils

Medical - MRC

Biotech & Biological Scs- BBSRC

Natural Environment – NERC

Engineering and Physical Scs– EPSRC

Particle Physics & Astronomy – PPARC

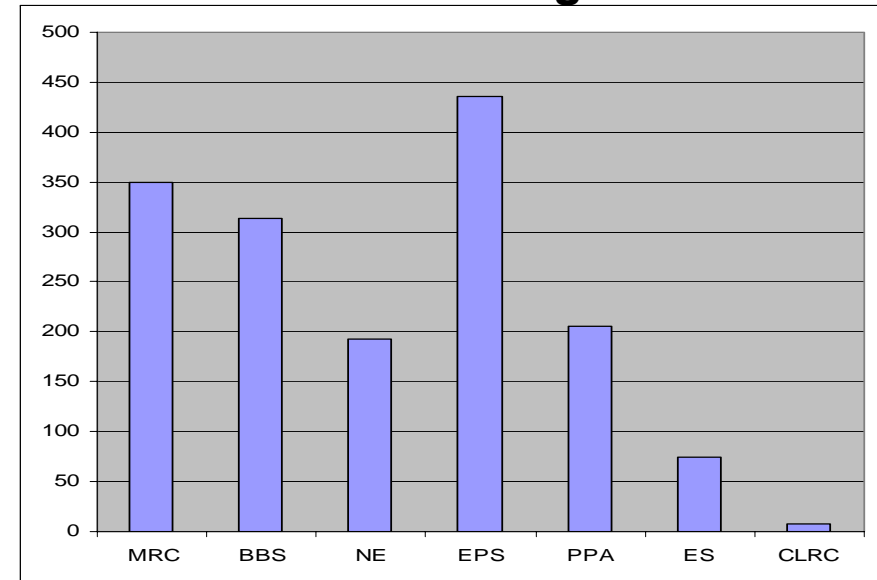
Economic & Social - ESRC

Central Laboratory for RCs – CCLRC

Arts & Humanities - AHRC

Total Science Budget 2005/6: £2.96 billion

Research Council Budgets: 2001-2



- strategic partnership between the 8 UK Research Councils
- promotes cross-council collaboration on research, training and knowledge transfer



John Taylor - background

DGRC, 1999-2003: Key Programs

Policy and Funding

- Investing in Innovation, 2002; UK 10 year Science Strategy, 2004
- UK Science Budget funding doubled, 1998-2005, to £3B
- Research Councils UK (RCUK) & AHRC formed
- Metrics, Selectivity, RAE; SRIF, Full Economic Costs
- Knowledge Transfer – HEIF

Research

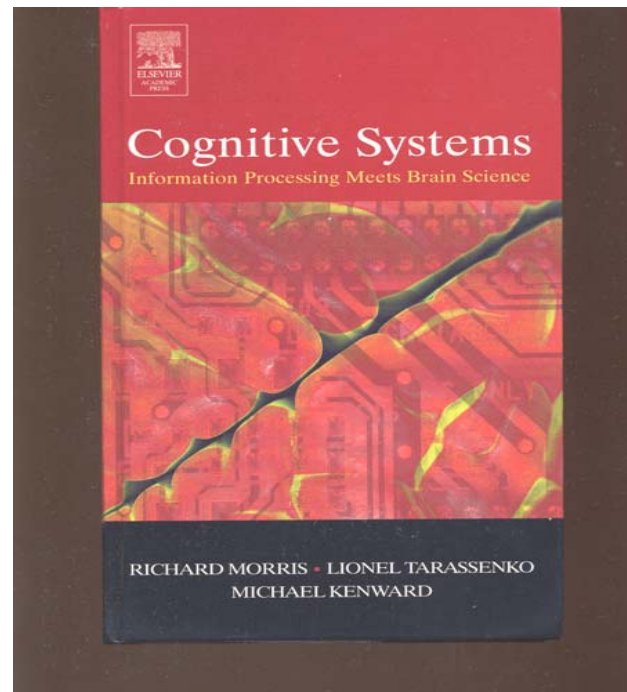
- eScience, GridPP.....
- Large Facilities Roadmap, Diamond Synchrotron
- CERN, ESO
- Energy Research
- Basic Technology Research Prog
- Foresight, Cognitive Systems project
 - good example of “suspend disbelief”
- European Research Council

UK OST Foresight Program

Cognitive Systems Project: 2002-2004

*Richard Morris, Lionel Tarrasenko,
& Michael Kenward (Elsevier, 2005)*

Director: John Taylor



A Definition of e-Science

‘e-Science is about global collaboration in key areas of science, and the next generation of infrastructure that will enable it.’

John Taylor

Director General of Research Councils
Office of Science and Technology

- Purpose of the UK e-Science initiative is to allow scientists to do ‘faster, better or different’ research
- e-Science will change the dynamic of how research is undertaken

e-Science

- Research increasingly done through distributed global collaborations enabled by the Internet (e.g. human genome program, LHC @CERN)
- Uses very large data collections, terascale computing resources, high performance visualisation and col-laboratories – support for trusted teams
- E-science will change the dynamic of the way science is undertaken

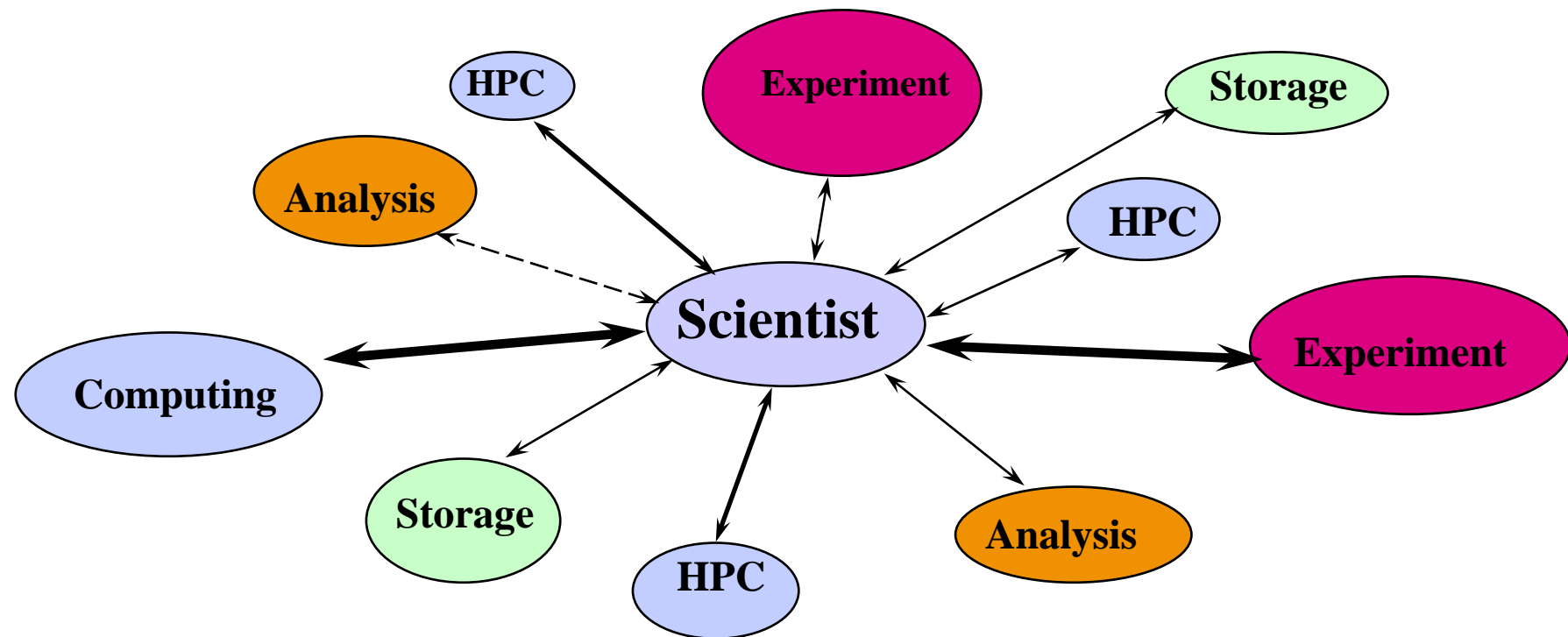
The Grid Vision of Foster, Kesselman and Tuecke

- ‘The Grid is a software infrastructure that enables flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions and resources’
 - Includes computational systems and data storage resources and specialized facilities
- Long term goal for Grid middleware infrastructure is to allow scientists to build transient ‘Virtual Organisations’ routinely

The Grid

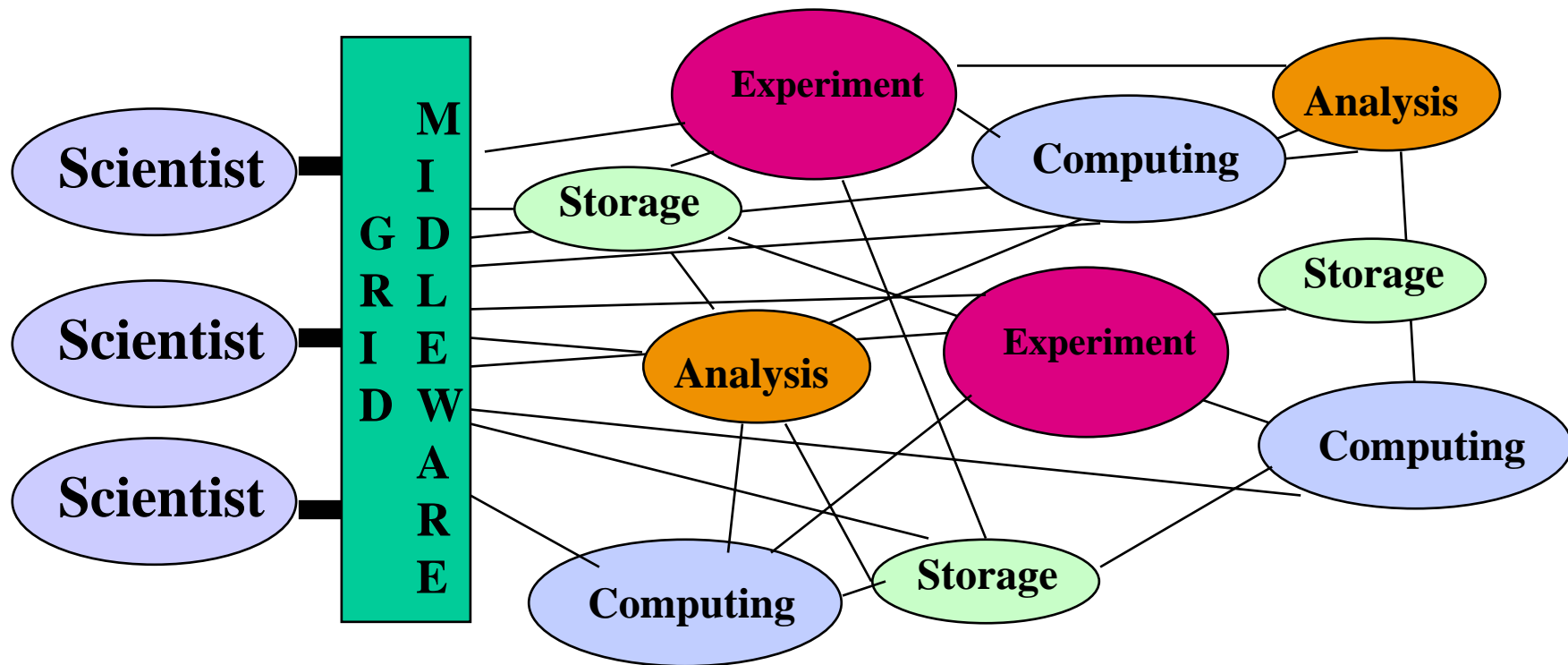
- First stages of the *information utility* – successor to client-server Internet & WWW architectures
- Solution to fulfilling the data-intensive problems and compute-intensive needs of many scientific communities and industries over the next decade (e.g. particle physics, astronomy, protein folding, earth sciences, engineering design...)
- Middleware, software and hardware to access, process, communicate and store huge quantities of data
- Grid is infrastructure enabler for e-Science
- e-Science/e-Research is main driver for Grid development

Behind The Wall: ad hoc *Client-Server*



Behind The Wall: The Grid Model

-Utilities and col-laboratories



The Information Utility

- Behind the Wall
 - Huge amounts of connectivity, computing power, software and data resources
- In Front of the Wall
 - Billions of smart devices connected (wirelessly) to each other and to the utility
 - from clothing to synchrotrons
 - People!
- Through the Wall
 - Crucial two-way interaction between people and things connected to the utility
 - Col-laboration

Evolution of all three:

=> Global Information Infrastructure

The Information Utility

- e-Science Grids building on the internet and the World Wide Web represent the first generation of global Information Utilities
- Will rapidly develop beyond science for a huge range of other applications

The UK e-Science Paradigm

- The Integrative Biology Project involves seven UK Universities lead by Oxford and the University of Auckland in New Zealand
 - Models of electrical behaviour of heart cells developed by Denis Noble's team in Oxford
 - Mechanical models of beating heart developed by Peter Hunter's group in Auckland
- Researchers need robust middleware services to routinely build secure 'Virtual Organisations to' support an international "collaboratory"
 - Goal is to enable 'faster, better or different' research

Aims of UK program

- to enable research community to recognise that many areas of research will *HAVE TO HAVE* e-science capabilities in order to do their next cycle of work:
 - to work with the best in the world
 - to exploit huge data resources
 - (including huge instrumentation capabilities)
- to provide common, generic technologies and infrastructure for this
 - economies of scale, effort, complexity
 - facilitate multi-disciplinary e-research
 - get critical mass of influence on emergence of next gen standards
- to stimulate next cycle of computer science and communications research - problems of huge scale, security, persistence, curation....
- to train new kinds of researcher and knowledge worker
- to provide a “lab” to explore how new adopters will use the technology to transfer it to business and industry applications

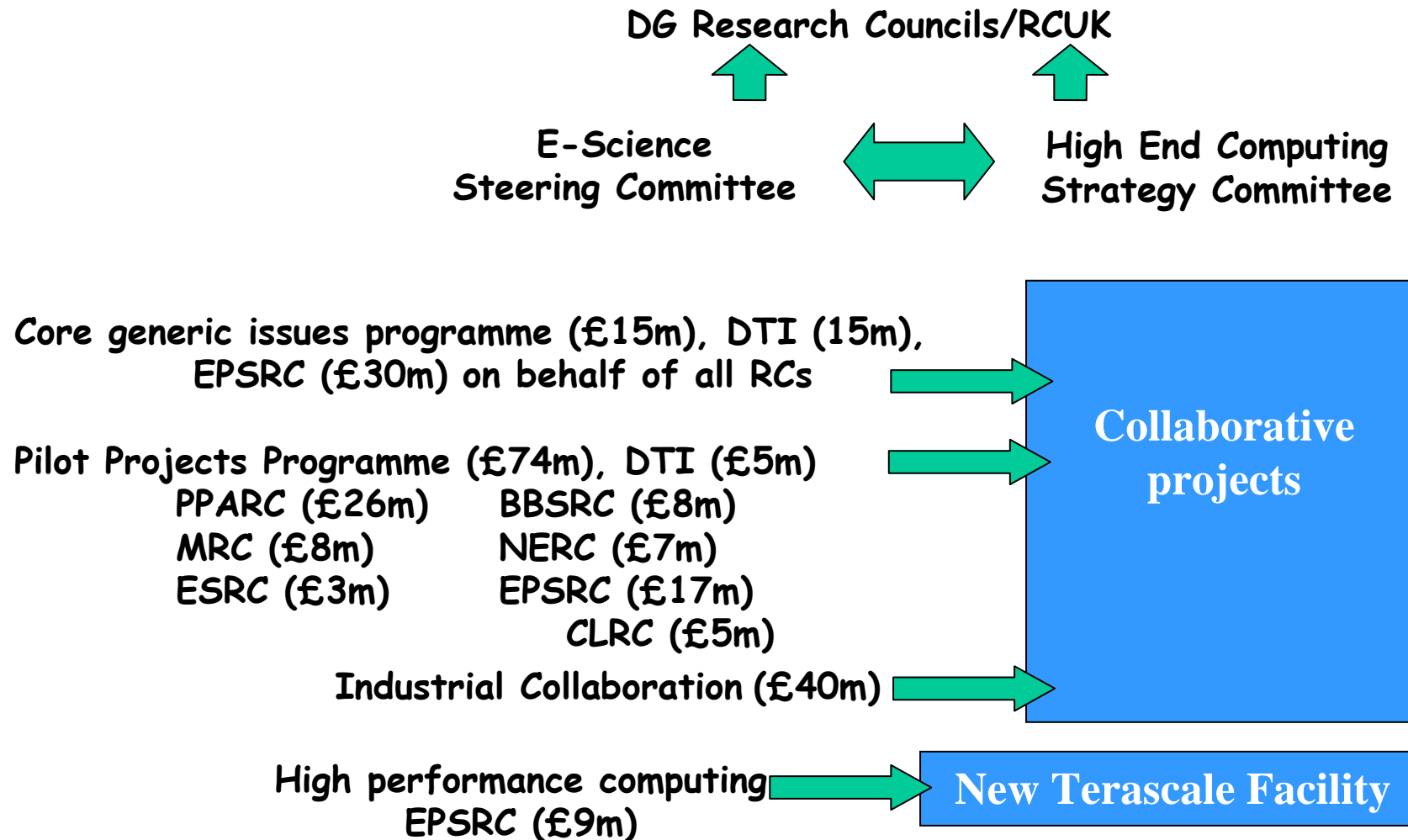
Why we needed a UK e-Science program

To provide the UK-wide national infrastructure and facilities needed for the UK's participation in worldwide science research across all disciplines

- One national information utility infrastructure not one per discipline
- in silico experimentation,
- huge data collections,
- global col-laboratories, not individual client-server
- integrated campus infrastructure for all disciplines

UK e-Science Program

Total: £118m + £40m



RCUK e-Science Funding Matrix

First Phase: 2001 –2004

Total: £109m

- Application Projects
 - £74M
 - All areas of science and engineering
- Core Programme
 - £15M Research infrastructure
 - £20M Collaborative industrial projects

Second Phase: 2003 –2006

Total: £123m

- Application Projects
 - £96M
 - All areas of science and engineering
- Core Programme
 - £16M Research Infrastructure
 - £11M DTI Technology Fund

- An exciting portfolio of Research Council e-Science projects
 - Beginning to see e-Science infrastructure deliver some early 'wins' in several areas eg
 - DiscoveryNet success at SC02
 - TeraGyroid success at SC03: 'heroic' achievement
- The UK is unique in having a strong collaborative industrial component
 - Nearly 80 UK companies contributing over £30M
 - Engineering, Pharmaceutical, Petrochemical, IT companies, Commerce, Media, ...
- Evaluation report due autumn 2005

- Particle Physics
 - global sharing of data and computation
- Astronomy
 - 'Virtual Observatory' for multi-wavelength astrophysics
- Chemistry
 - remote control of equipment and electronic logbooks
- Engineering
 - industrial healthcare and virtual organisations
- Bioinformatics
 - data integration, knowledge discovery and workflow
- Healthcare
 - sharing normalized mammograms
- Environment
 - Ocean, weather, climate modelling, sensor networks

AHM₂₀₀₅*Innovating through e-Science*

Particle Physics: CERN, Global PP Grid for LHC

Astronomy: Astrogrid, working Virtual Observatory,
International VO Alliance

Drug discovery, genetics, structure prediction

DiscoveryNet, myGrid, CombeChem, e-materials

Environment:

Climate prediction.Net, SARIS weather forecasting,
DEWS - Delivery via WebServices,

Chemistry:

Teragyroid, RealityGrid, – steering distributed
simulations & visualisation on US Teragrid & UK
HPC/CSAR supercomputers

AHM₂₀₀₅
Innovating through e-Science



Medical:

e-Diamond - mammography

Clinical e-Science Framework (CSEF) – secure querying
of patient records – 22k cancer cases

Engineering:

DAME – aeroengine data

Geodise – collab eng design – aircraft wings

Social Sciences

Archaeology

AHM₂₀₀₅

Innovating through e-Science



Standards and infrastructure

UK Light – US-UK lambda network

Security Task force:

UK e-Science Digital Certification authority

Grid Operations Support Centre

Security Access Markup Language SAML,

Oasis, Shibboleth

Open Middleware Infrastructure Initiative- OMII-UK

Digital Curation Centre

- 18 projects, 16 departments,
- 85% CS for e-Science projects in RAE 5*/5 departments
- 59 academics

HyOntUse
Secure location independent autonomic storage architectures
Grid-enabled numerical and symbolic services
Magikl: managing grids containing information and knowledge that are incomplete
A Semantic Firewall

Dynamic Ontologies: a Framework for Service Descriptions
Trusted Coordination in Dynamic Virtual Organisations
Service Level Agreement Based Scheduling Heuristics
A System for Publishing Scientific Data
PASOA: Provenance Aware Service Oriented Architecture
AMUSE: Autonomic Management of Ubiquitous Systems for e-Health
Dynamic Net Data: Theory and Experiment
Presenting Ontologies in Natural Language
Describing the Quality of Curated E-Science Information Resources
Pervasive Debugging
Inferring Quality of Service Properties for Grid Applications
Open Overlays: Component-Based Communications Support for the Grid
Virtual organisations

Key Elements of a UK e-Infrastructure

1. Research Network
2. National Grid and HEC Service
3. Open Middleware Infrastructure Institute
4. Digital Curation Centre
5. National e-Science Institute
6. Portals and Discovery Services
7. Access to facilities, data services and repositories
8. Tools and Services to support collaboration
9. National data archive
10. Support for International Standards

Open Middleware Infrastructure Institute



The Three OMII Goals

- -Set up Repository for WS-* generic Grid middleware for the UK e-Science community
 - -Capture generic middleware from UK e-Science Projects
 - -Commission middleware projects to fill specific 'gaps'
- All supported by Top Quality Software Engineering processes and standards
- To make components that interoperate
 - To make a set of Grid services that interoperate with everyone else
 - Globus GT4, EGEE gLite, ...



Digital Curation Centre

- Actions needed to maintain and utilise digital data and research results over entire life-cycle
 - For current and future generations of users
- Digital Preservation
 - Long-run technological/legal accessibility and usability
- Data curation in science
 - Maintenance of body of trusted data to represent current state of knowledge in area of research
- Research in tools and technologies
 - Integration, annotation, provenance, metadata, security.....
- Edinburgh with Glasgow, CLRC and UKOLN selected to run UK Digital Curation Centre

Collaboration & Col-laboratories

Distributed teams:

geographically, organisationally, functionally

Consenting, trusting to work together to create new knowledge and IP

NOT the same as open access to web pages!!

Virtual organisations – secure, casual, temporary, evanescent

Multiple disciplines



Collaboration & Col-laboratories

Teams with expertise and resources in different parts of the problem eg:

NASA, heart modelling

No longer limited by local mindset

– own lab, own computers, own data, own instruments, own codes, own experiments

Collaborations previously inconceivable can now become routine

and even casual – ie short lived, temporary, low bureaucracy

Can use collaborators' resources in situ – don't have to bring everything to one place

Because the infrastructure is done for you, cooperation can be quick, on-demand, affordable,.....

Audit trails.....

Curation of digital information

- In the next 5 years e-Science projects will produce more scientific data than has been collected in the whole of human history
- In 20 years it is almost certain that the operating systems, applications programs and the hardware used to store data will not exist as supported products
 - Need to research curation technologies and best practice
 - Need to liaise closely with individual research communities, data archives and libraries

Curation of digital information

Creation:

its about creating new data and IP, not just accessing existing data

Access:

huge distributed heterogeneous incompatible data sets

Provenance:

who created this data, what standing do they have?

is it refereed, peer-reviewed, cross-checked?

does it conflict with other data?

who has modified, annotated, critiqued this data set...?

Curation of digital information

Semantics:

how do I compare one group's version of a protein structure, galaxy, hurricane, heart model, with another group's ?

Search eg how do I dispatch search agents across heterogeneous distributed data sets to discover the data I might want

Resource management

How can (my agents) negotiate for rights to access, use, copy, modify, critique.....



My main message from the UK e-Science experience:

Applications pull is crucial

First class computer and communications research is crucial

You need a 2-D matrix of applications and infrastructure

“If you build the infrastructure, it will happen”.... - NO:

If you don't build the infrastructure, it won't happen - CORRECT

But if you don't get the researchers to really want to use it,

- IT REALLY WON'T HAPPEN...



My main message from the UK e-Science experience:

Applications pull is crucial

First class computer and communications research is crucial

- You need a 2-D matrix of applications and infrastructure

Computer science and communications research people are not second class citizens in this – “just hacking to provide services for real researchers”

Information Utilities are a massive computing and communications research area and research environment in their own right

To create information utilities, computing and communications researchers have to understand the real needs of many different applications domains by working with them as peers.

New ideas coming from such interaction will surprise and delight both sides

The contribution of serious computer science and maths to eg biology, medicine, engineering design and manufacture, are going to be huge and crucial

Some possible objectives for a national program

Better participation in world class research communities - for many areas of research this style of doing business is becoming essential

– its not going to go away

Accelerate migration of e-Science ->e-Research > e-Industry & e-Business

Accelerate the emergence of next generation global standards platforms
(cf TCP-IP for internet)

- need to make a contribution to have a seat at the table

e-Research program itself provides a giant lab for potential adopters to observe, explore, participate in how people like them might be affected and get involved

Key route for training new kinds of knowledge workers not just

e-Research is about doing qualitatively new things

HPC – provides massive compute power

Grids – allow sharing of compute, data and specialised resources
– borrow cycles, ship jobs to where the cycles are (see Foster Kessleman) – infrastructure that allows sharing and collaboration

e- Research – next step – how to collaborate

- collaborate with other people and their facilities
- use other peoples' codes at their facilities
- share and create huge data resources

Data - deep, subtle, complex, huge, heterogeneous, distributed

- provenance, semantics – astronomy, proteins, meteorology, health...
- audit, quality assurance, persistence, integrity,

e-Research is about doing qualitatively new things

- ❖ A sufficient increase in scale becomes something *qualitatively new*
eg: in silico experimentation – with huge resources
giant valuable data collections
- ❖ New kinds of collaborations:
distributed, temporary, cross-organisation, cross-discipline
- ❖ Different teams have different parts of the problem, and the solution
- ❖ No longer have to bring all these to one place, one machine
- ❖ Consenting mutually trusting teams and communities
- not same as wide open webs
- ❖ Expand the mindset of researchers – no longer limited by own lab, own instruments, own data, own experiments, doing it all in one place
- ❖ *New domains for collaboration and knowledge creation
eg social sciences, healthcare, law, finance, media*

Needs

Matrix of generic infrastructure and demanding applications domains

- application domains with real operational needs
- innovate in generic IT infrastructure
 - including on-campus
 - next step to information utilities
- enable fast cross discipline and cross-organisation collaboration

Director(s) with budgets and discretion



Needs

- Change culture of leading researchers – they **MUST HAVE** e-research capabilities to do their next stage of research.....
- Get applications areas to really need and value e-Research
- Change researcher mindsets – no longer just limited by own lab, own instruments, own experiments
- Revitalise computer science research – next generation information utilities have challenging research problems
 - scale, security, data curation.....
- e-Research is not just about communications, networks, infrastructure - not even just about Grids!

Needs:

- Training and awareness for a new generation of research professionals in most disciplines,
 - outreach, centres of expertise and support with marginal cost for access and entry
- Technology research:
 - Security, authentication, integrity,...
 - Curation, persistence, repositories, search,...
 - IP management, audit,...
 - Agents, negotiation, trading, pricing

What is a policy framework for e-Research?

leadership, vision, objectives

funding

stakeholders

directors, management

empowerment of talented people

doing new things and new ways of doing things

monitoring and oversight – closing the loop, not fire and forget

quality, excellence, surprises

What is a policy framework for e-Research?

Must be researcher-lead,

-uncompromising on excellence and real need

get champions to lead a critical mass of community to “get it”

let them get on with it, but monitor quality hands on

insist on real “production” mindset – not just offline testbeds

trust them to find the sweetspots for “must have” applications

expect them to surprise you

What is a policy framework for e-Research?

Role of Champions

Build trust, excitement, motivation and excellence:

- trust - get potential participants to “suspend disbelief” while you all take the risk of going once round the loop
- excitement – “this could be really neat”
- motivation – “there’s something in this for me – with this I could really do something I couldn’t do before..”
- excellence – “this is serious new stuff not, just easy re-badging”



eResearch in Australia: next steps

What is a policy framework for e-Research?

Intellectual property & peer review in the e-Research era
-incremental creation and publishing of IP

IT is most unregulated technology ever – but our society is now
critically dependent on it, vulnerable to it

Grids, e-Research etc are steps on the way to robust global
information utilities for all of society



eResearch in Australia: next steps

Key questions:

What's the vision?

What's the high level focus, priorities, metrics?

how will you know if you've succeeded in 5-7 years ?

and if you're on track for success in 2-3 years?

Where's the specific earmarked funding *for applications pull*?

Who can allocate it speedily with excellent taste and judgment?

Who's in charge of what?

Who is allowed to take some risks, ie make some (new) mistakes?

To move at Grid speeds you need to be able to make crisp decisions, do a few things well, build critical mass, multidisciplinary communities quickly



eResearch in Australia

eResearch Open Forum eResearch – Col-laboratories and Curation

John Taylor

Presentation to DEST researchers

Canberra, Australia

20 September 2005

john @taylorfamily.tv