e-Science and the Grid – Preserving the Data Deluge

Tony Hey

Director of UK e-Science Core Programme

Tony.Hey@epsrc.ac.uk

Licklider's Vision

"Lick had this concept – all of the stuff linked together throughout the world, that you can use a remote computer, get data from a remote computer, or use lots of computers in your job."

Larry Roberts – Principal Architect of the ARPANET

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

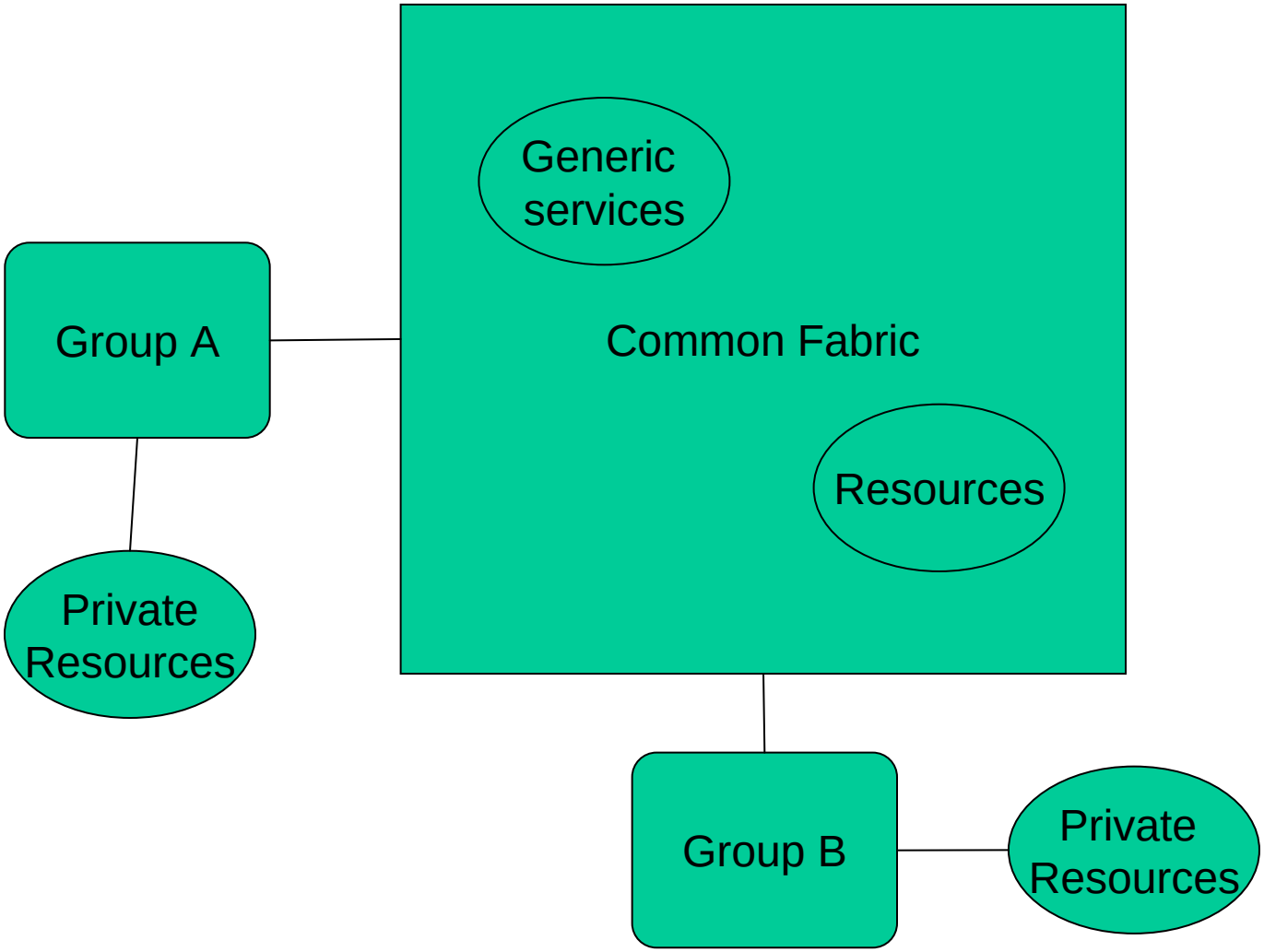
Office of Science and Technology

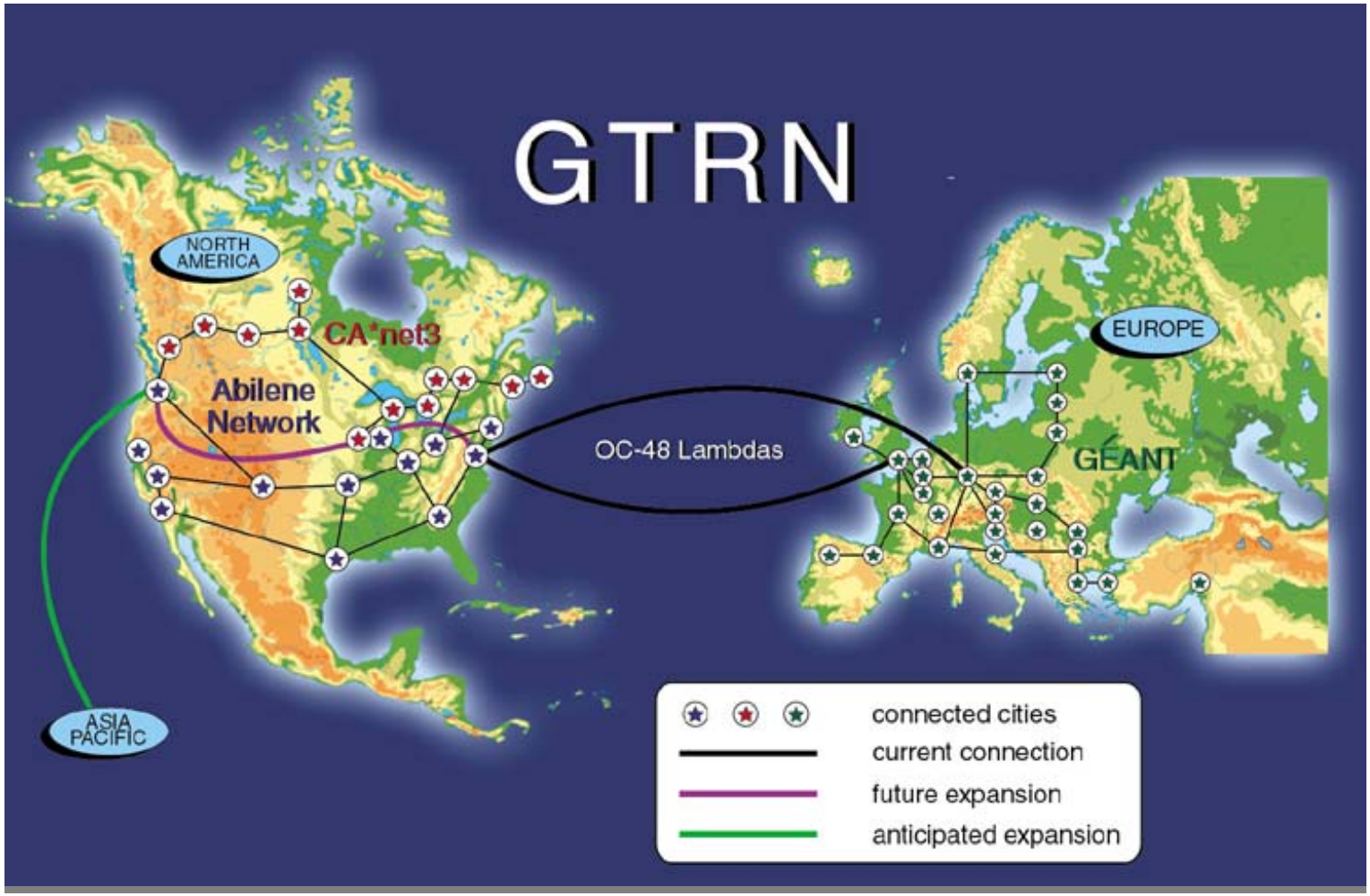
- Purpose of e-Science initiative is to allow scientists to do faster, different, better research

The e-Science Paradigm

- The Integrative Biology Project involves the University of Oxford (and others) in the UK 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 to be able to easily build a secure 'Virtual Organisation' allowing access to each group's resources
  - Will enable researchers to do different science

e-Infrastructure/Cyberinfrastructure for Research





The Global Grid =

A set of core middleware services running on top of Global Terabit Research Networks

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

RCUK e-Science Funding

First Phase: 2001 –2004

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

Second Phase: 2003 –2006

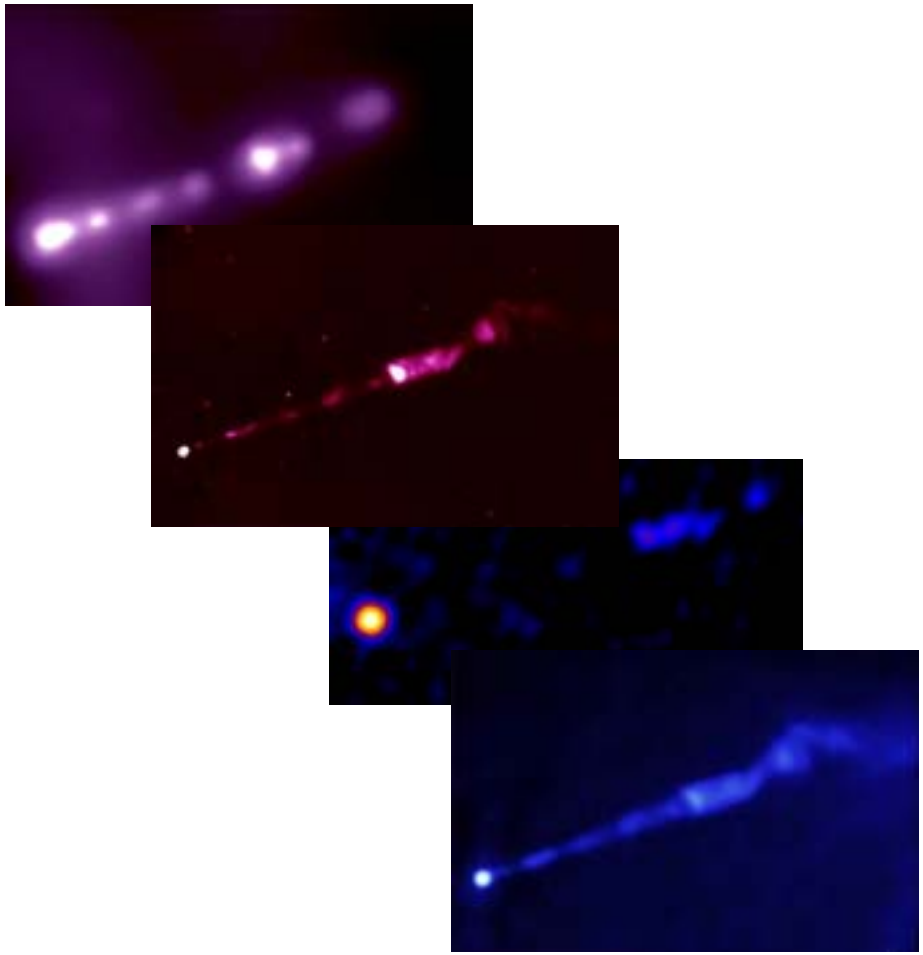
- Application Projects
  - £96M
  - All areas of science and engineering
- Core Programme
  - £16M Research Infrastructure
  - DTI Technology Fund

Some Example e-Science Projects

- Particle Physics
  - **global sharing of data and computation**
- Astronomy
  - **a 'Virtual Observatory' for multi-wavelength astrophysics**
- Chemistry
  - **automatic annotation of data, remote control of equipment, simulation, database access and electronic logbooks**
- Engineering
  - **industrial healthcare, data mining and virtual organisations**
- Bioinformatics
  - **data integration and knowledge discovery**
- Healthcare
  - **sharing normalized mammograms**

CERN Users in the World – A Global VO

Europe: 267 institutes, 4603 users
Elsewhere: 208 institutes, 1632 users



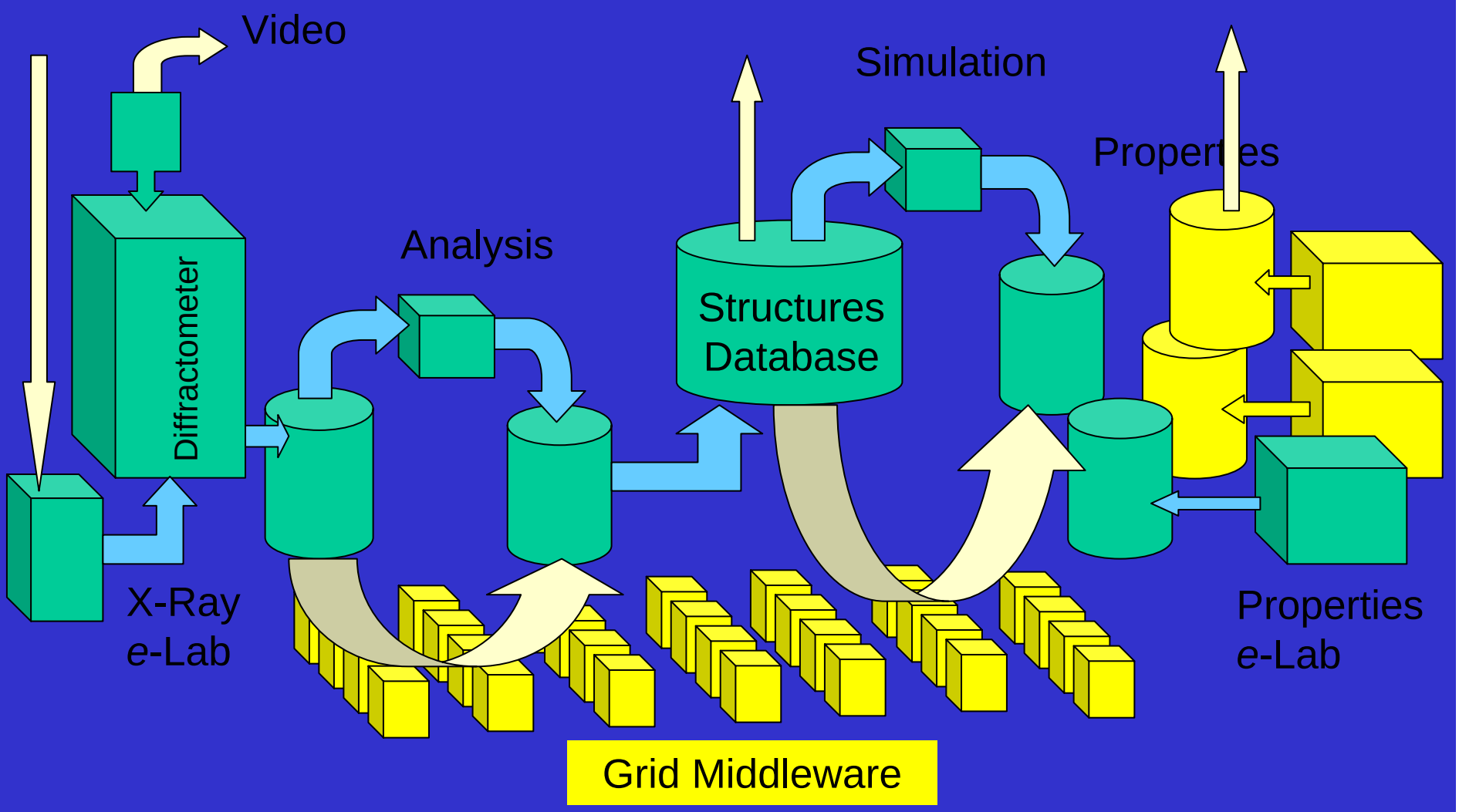
Powering the Virtual Universe

http://www.astrogrid.ac.uk

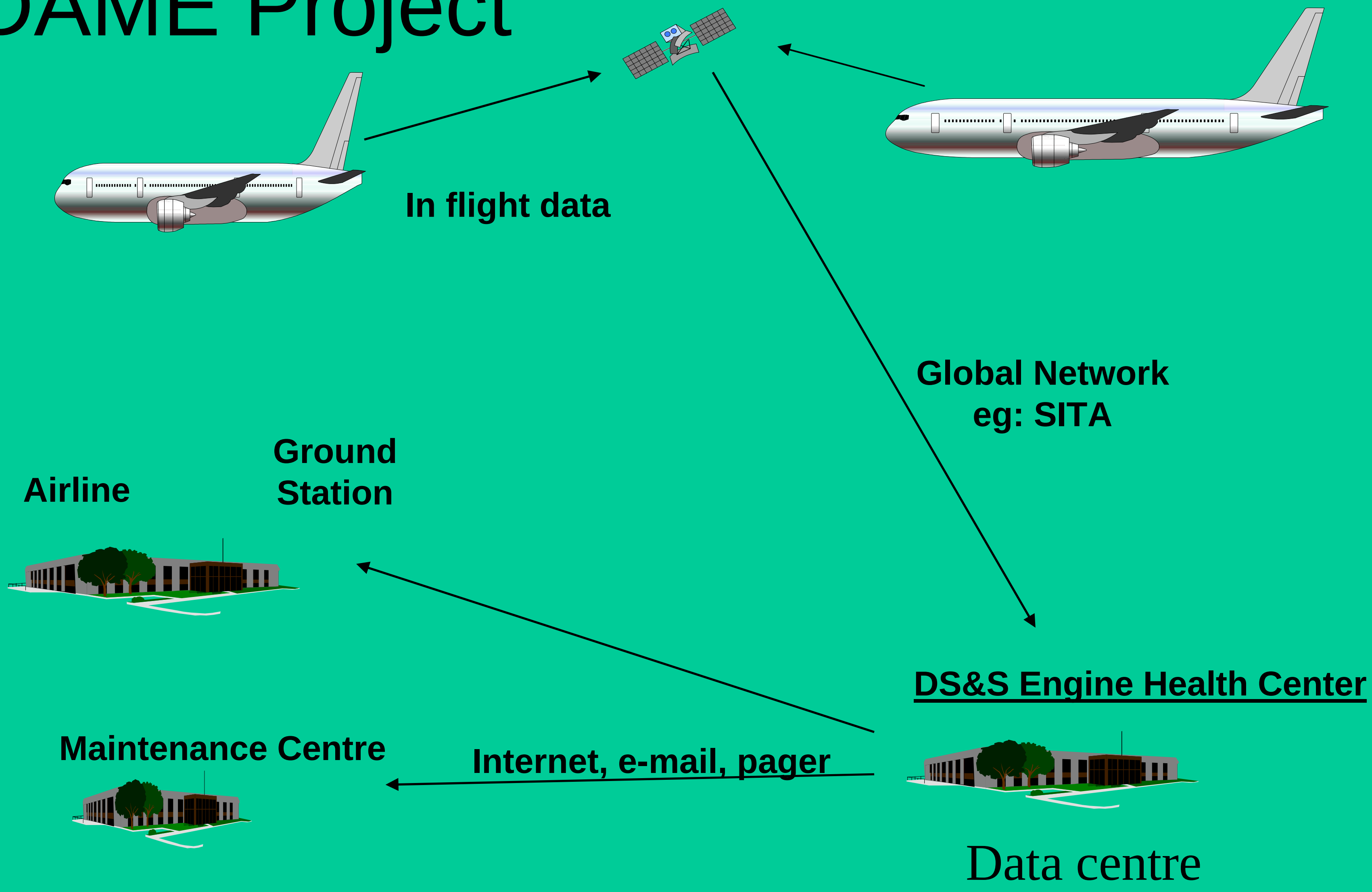
(Edinburgh, Belfast, Cambridge, Leicester, London, Manchester, RAL)

Multi-wavelength showing the jet in M87: from top to bottom – Chandra X-ray, HST optical, Gemini mid-IR, VLA radio.

Comb-*e*-Chem Project

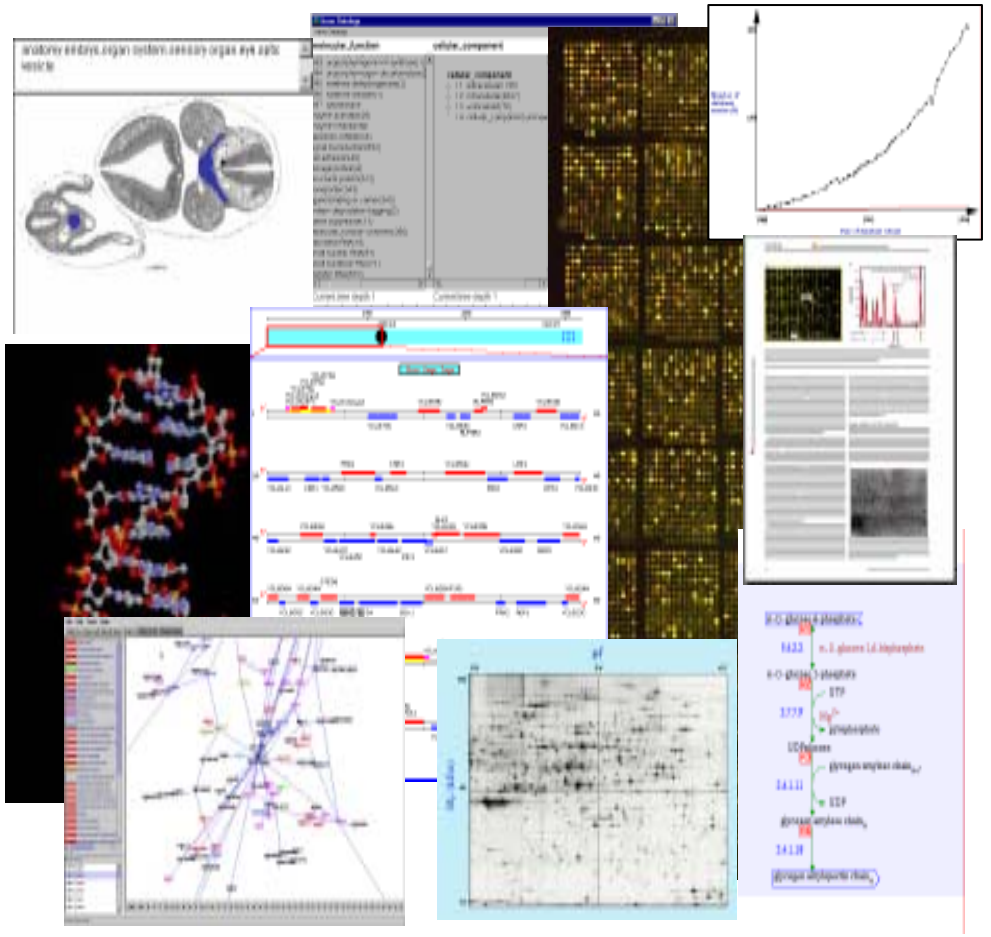


DAME Project



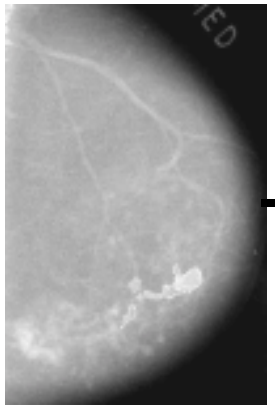
myGrid Project

- Imminent ‘deluge’ of data
- Highly heterogeneous
- Highly complex and inter-related
- Convergence of data and literature archives

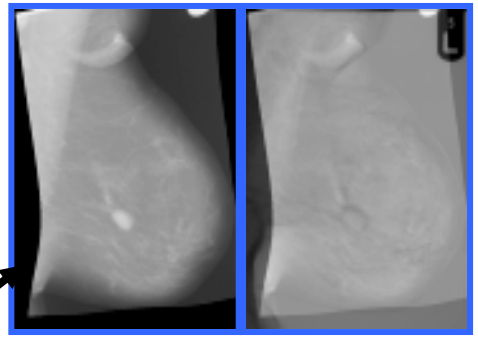


eDiaMoND Project

Mammograms have different appearances, depending on image settings and acquisition systems

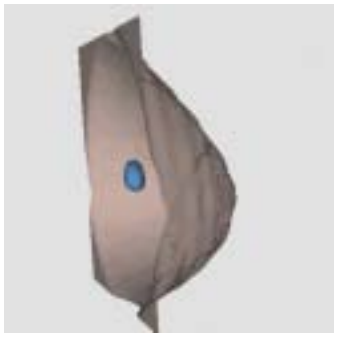


Standard Mammo Format



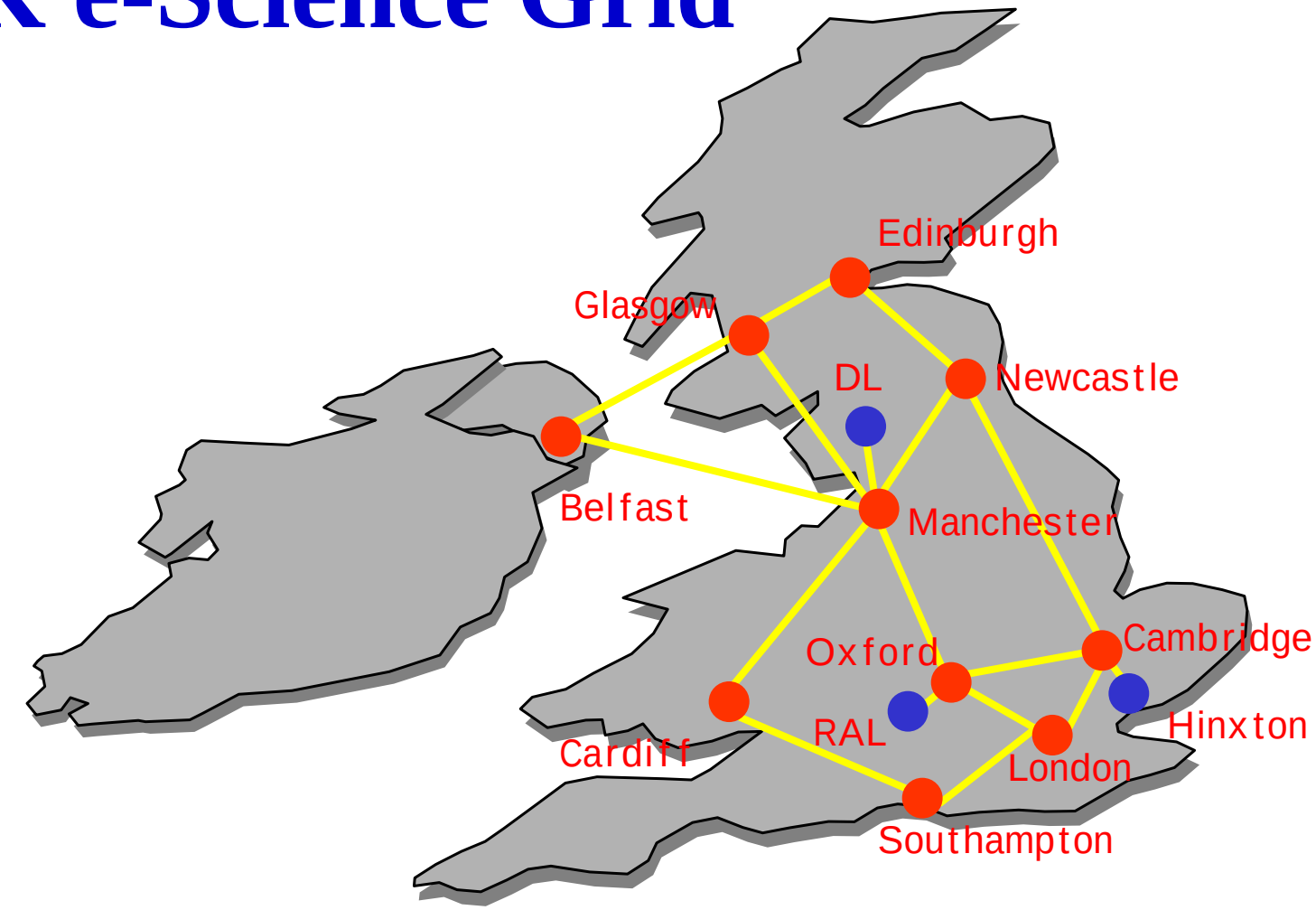
Temporal mammography

Computer Aided Detection



3D View

UK e-Science Grid



A Status Report on UK e-Science

- An exciting portfolio of Research Council e-Science projects
  - Beginning to see e-Science infrastructure deliver some early ‘wins’ in several areas
  - TeraGyroid success at SC03: ‘heroic’ achievement
  - Astronomy, Chemistry, Bioinformatics, Engineering, Environment, Healthcare ….
- 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, …

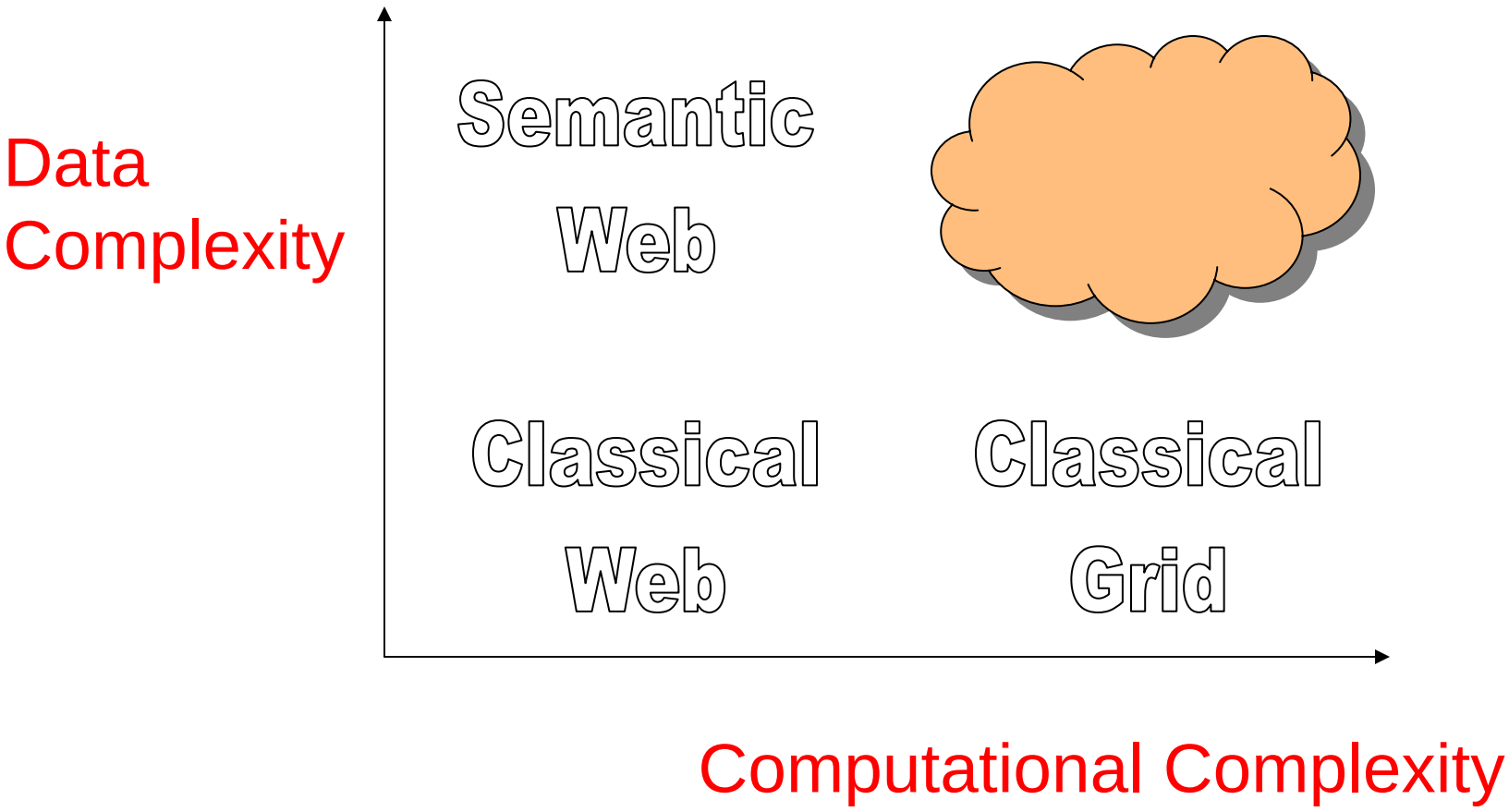
Identifiable UK Focus

- Data Access and Integration
  - OGSA-DAI and DAIT project
- Key grid data services
  - Workflow, Provenance, Notification
  - Distributed Query, Knowledge Management
- Data Curation and Data Handling
  - Digital Curation Centre
- Security, AA and all that
  - Digital Certificates and Single Sign-On
  - Federated Shibboleth framework for universities

Metadata & Ontologies

- Metadata – computationally accessible data about the services
- Ontologies – the shared and common understanding of a domain
  - A vocabulary of terms
  - Definition of what those terms mean.
  - A shared understanding for people and machines
  - Usually organised into a taxonomy.

The Semantic Grid



JISC Committee for Support of Research (JCSR)

- Ensure JISC addresses the needs of the HE research community
  - Members representing each of the Research Councils plus the AHRB
- Recurrent budget of £3M p.a.
  - Strategy to co-fund some of the JCSR activities with other relevant funding bodies
  - Projects with BBSRC, CLRC, EPSRC, ESRC and the e-Science Core Programme

JISC/JCSR e-Science Support

- Digital Curation Centre
  - Joint funding with e-Science Core Programme
- Text Mining Centre
  - Led by UMIST
- The *e*-Bank Project
  - Uses Comb-e-Chem Project as exemplar

Digital Curation Centre (DCC)

- In next 5 years e-Science projects will produce more scientific data than has been collected in the whole of human history
- In 20 years can guarantee that the operating and spreadsheet program and the hardware used to store data will not exist
  - Research curation technologies and best practice
  - Need to liaise closely with individual research communities, data archives and libraries
- Edinburgh with Glasgow, CLRC and UKOLN selected as site of DCC

Terminology: Digital Curation

Digital Curation = Digital Preservation and Data Curation

- 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

DCC

Digital Preservation: The issues

- Long-term preservation
  - Preserving the bits for a long time (“digital objects”)
  - Preserving the interpretation (emulation vs. migration)
- Political/social
  - Appraisal - what to keep?
  - Responsibility - who should keep it?
  - Legal - can you keep it?
- Size
  - Storage of/access to Petabytes of regular data
  - Grid issues
- Finding and extracting metadata
  - Descriptions of digital objects

Data Publishing: The Background

In some areas – notably biology – databases are replacing (paper) publications as a medium of communication

- These databases are built and maintained with a great deal of human effort
- They often do not contain source experimental data.Sometimes just annotation/metadata
- They borrow extensively from, and refer to, other databases
- You are now judged by your databases as well as your (paper) publications!
- Upwards of 1000 (public databases) in genetics

Data Publishing: The issues

- Data integration
  - Tying together data from various sources
- Annotation
  - Adding comments/observations to existing data
  - Becoming a new form of communication among scientists
- Provenance
  - Where did this data come from?
- Exporting/publishing in agreed formats
  - To other program as well as people
- Security
  - Specifying/enforcing read/write access to *parts* of your data


Edinburgh has research positions in databases, digital curation, XML, web technology, fundamentals.
Edinburgh is a great place to live!!!
Contact
Peter Buneman
opb@inf.ed.ac.uk
Top-rated department. World-class database group. Good connections with logical foundations, scientific DBs, distributed computation (Grid)

The *e*-Bank JISC e-Science Project

- School of Chemistry and
  School of Electronics and Computer Science
  University of Southampton
- UKOLN
  University of Bath
- Psigate
  University of Manchester

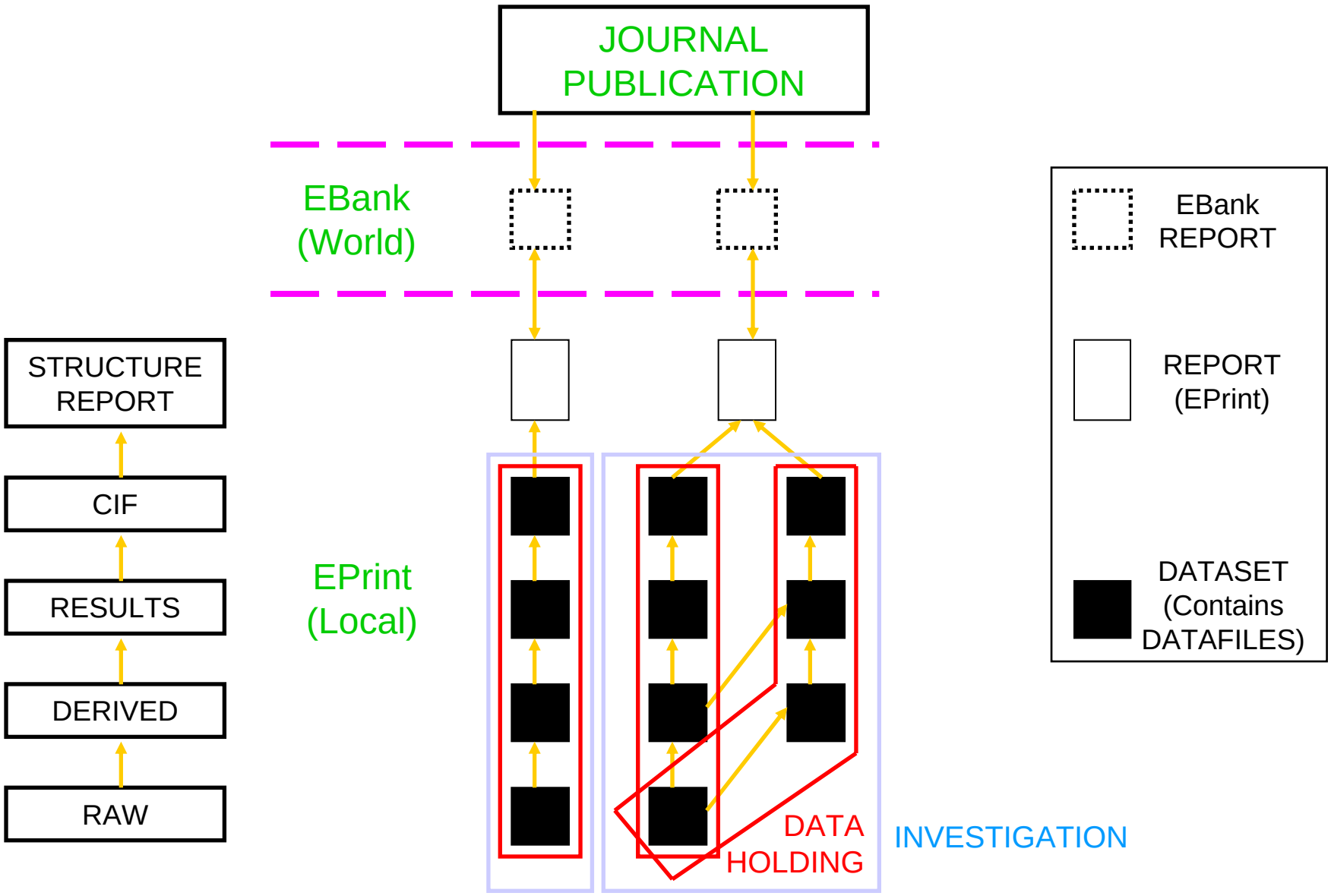
Referee@source or Referee on demand?

- High data throughout
- Any given data set is not that important
- Cannot justify a full referee process for each
- Better to make data available rather than simply leave it alone
- Need to have access to raw data to allow users to check

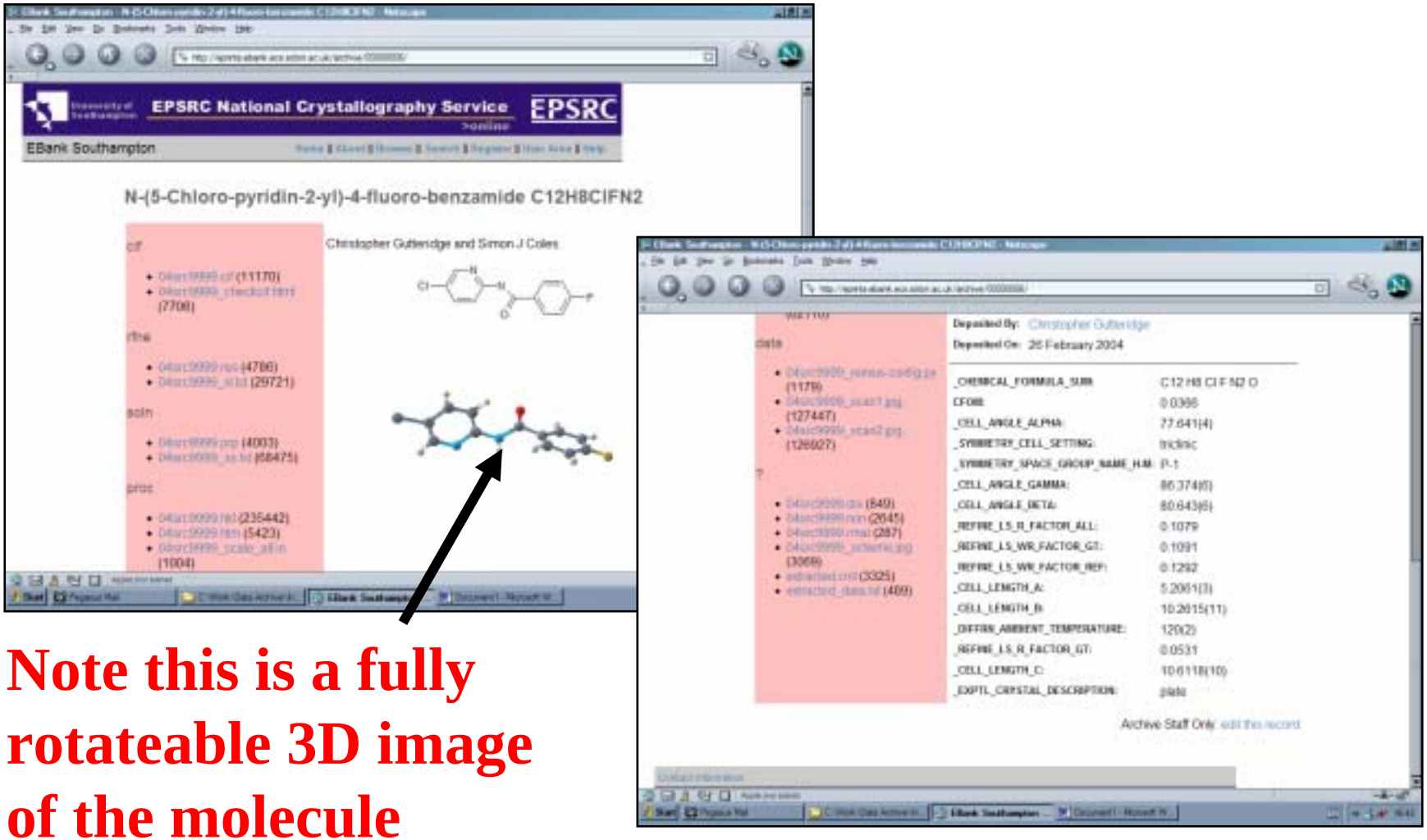
Goals of *e*-Bank Project

- Provide self archive of results plus the raw and analysed data
- Provide a route to disseminate these results
- Links from traditionally published work provides the provenance to the work
- Disseminate for “Public Review” – raw data provided so that users can check themselves
- Avoid the “publication bottleneck” but still provide the quality check

Crystallographic *e*-Prints

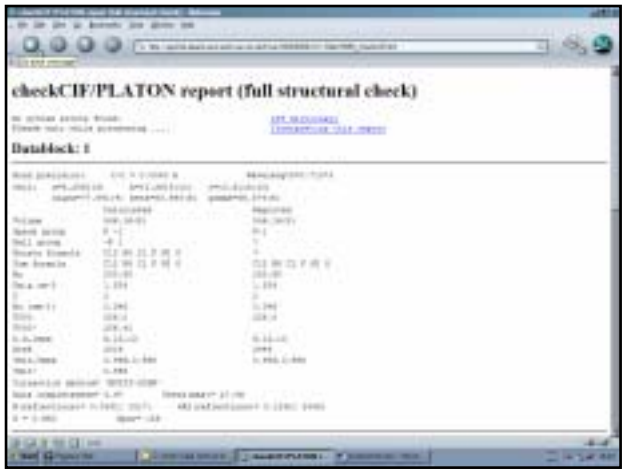


Crystallographic *e*-Prints

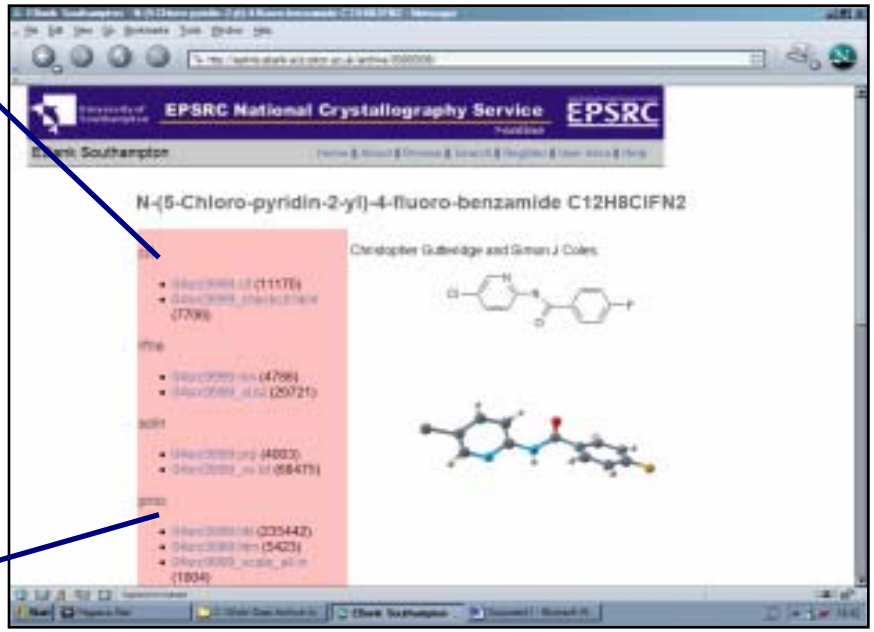


Note this is a fully rotateable 3D image of the molecule

Direct access to data



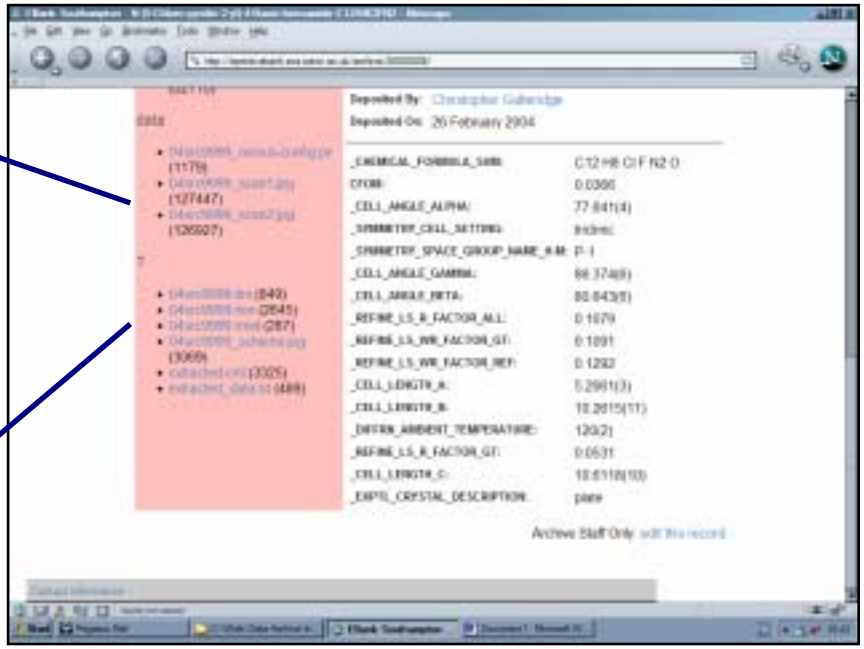
- **DERIVED DATA**

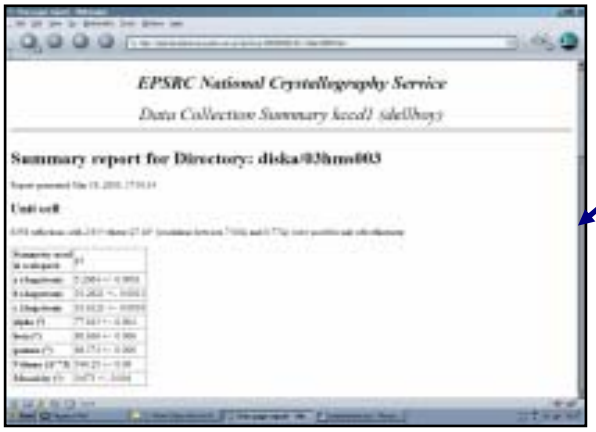


Links to download the raw and processed data

Direct access to data

- **RAW DATA**





Raw data sets can be very large and these are stored at the Atlas Datastore (using SRB server) and made available via a URI resolver

Traditional Publishers

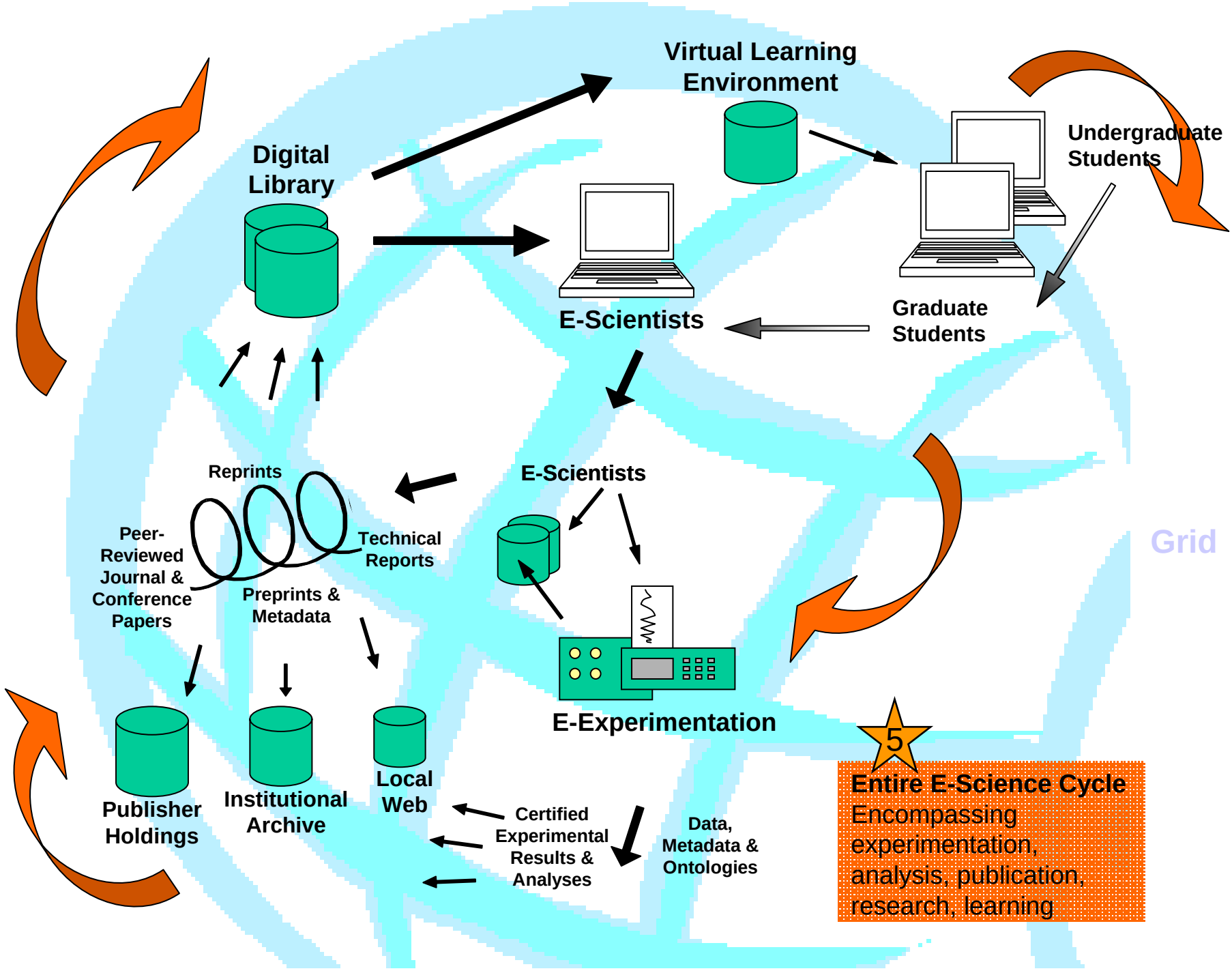
- Papers should contain ideas not so much data
- Original / Raw data needs to be accessible
- Very useful and productive discussions with several publishers of crystallographic data
  - IUCr (International Union of Crystallography)
  - Elsevier, RSC

Moving on from Crystallography

- Crystallography only a start
  - Chosen due to suitability of data
  - International agreement on representation of much of the data
- Next stage spectroscopic data
  - Interest of several instrument manufacturers
  - Again use international standards

e-Bank: Some Comments

- Data as well as traditional bibliographic information is made available via an OAI interface
- Can construct high level search on data – aggregate data from many e-print systems
- Build new data services
- Will make provision of real spectra (rather than very reduced summaries) for chemistry publications (see recent House of Commons Committee question from Dr Iddon MP)


Virtual Learning Environment
Digital Library
Undergraduate Students
E-Scientists
Graduate Students
Reprints
E-Scientists
Peer-Reviewed Journal & Conference Papers
Technical Reports
Preprints & Metadata
Grid
E-Experimentation
Publisher Holdings
Institutional Archive
Local Web
Certified Experimental Results & Analyses
Data, Metadata & Ontologies
5
Entire E-Science Cycle
Encompassing experimentation, analysis, publication, research, learning

JCSR Text Mining Centre

- UK Partners:
  - UMIST/UManchester
  - University of Liverpool
  - University of Salford
- Self-funded international partners:
  - UC Berkeley California
  - University of Geneva
  - University of Tokyo
  - San Diego Supercomputing Centre (SDSC)

JCSR Text Mining Centre

Remit:

- To drive the associated national and international research agenda
- To establish a service for the wider academic community
- To connect with industry

Initial focus is biology/biomedicine domain.

- Growth of biomedical knowledge means users need new tools to deal with an increasingly large body of biomedical articles.
- Potential users of text mining services include both academic and governmental/corporate organisations.

Text Mining

- Attempt to discover new, previously unknown information by applying techniques from natural language processing, data mining, and information retrieval:

  (1) To identify and gather relevant textual sources

  (2) To analyse these to extract facts involving key entities and their properties

  (3) To combine the extracted facts to form new facts or to gain valuable insights

- Text mining results can be used either directly by the individual scientist or indirectly to validate and complement (currently) manually curated scientific databases

Proposed Centre Activities

- **Develop text mining infrastructure**
- **Support information retrieval and harvesting**
- **Support for terminology management and information extraction**
- **Support for Data Grid technologies**
- **User Interface development**
- **Visualization and knowledge representation technologies**

MIT DSpace Vision

‘Much of the material produced by faculty, such as datasets, experimental results and rich media data as well as more conventional document-based material (e.g. articles and reports) is housed on an individual’s hard drive or department Web server. Such material is often lost forever as faculty and departments change over time.’

A Definition of e-Research?

The invention and exploitation of advanced IT

- to generate, curate and analyse research data
- to develop and explore models and simulations
- to enable *dynamic* distributed virtual organisations

Acknowledgements

With special thanks to Peter Buneman, Peter Burnhill, Jeremy Frey, David Gavaghan, Carole Goble and Liz Lyon