e-Science and Open Access

Tony Hey

Director of UK e-Science Core Programme

Tony.Hey@epsrc.ac.uk

Berlin Declaration 2003

- ‘To promote the Internet as a functional instrument for a global scientific knowledge base and for human reflection’
- Defines open access contributions as including:
  - ‘original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material’

➢ This talk is mainly concerned with open access to data

NSF ‘Atkins’ Report on Cyberinfrastructure

- *‘the primary access to the latest findings in a growing number of fields is through the Web, then through classic preprints and conferences, and lastly through refereed archival papers’.*

- *‘archives containing hundreds or thousands of terabytes of data will be affordable and necessary for archiving scientific and engineering information’.*

Licklider's Vision for the Internet

"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

What is 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

Cyberinfrastructure/ e-Infrastructure and the Grid

- 'The Grid is a software infrastructure that enables flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions and resources' (Foster, Kesselman and Tuecke)
- ➢ Includes not only computers but also data storage resources and specialized facilities
- Long term goal is to develop the middleware services that allow scientists to routinely build the infrastructure for their 'Virtual Organisations'

Data Sharing in e-Science

- Particle Physics
  - global sharing of data and computation
- Astronomy
  - ‘Virtual Observatory’ for multi-wavelength astrophysics
- Chemistry
  - remote control of equipment and data archives
- Environment
  - federated data centres

Steve Lloyd
Tony Doyle
John Gordon

GRIDPP
UK Particle Physics

GridPP Presentation to PPARC Grid Steering Committee
26 July 2001


THE UNIVERSITY OF BIRMINGHAM


University of Bristol


BRUNEL UNIVERSITY WEST LONDON


CERN


CLRC


UNIVERSITY OF CAMBRIDGE


University of Durham


THE UNIVERSITY OF EDINBURGH


UNIVERSITY of GLASGOW

Imperial College
OF SCIENCE, TECHNOLOGY AND MEDICINE


LANCASTER UNIVERSITY


THE UNIVERSITY of LIVERPOOL


THE UNIVERSITY of MANCHESTER


Queen Mary University of London


Royal Holloway University of London

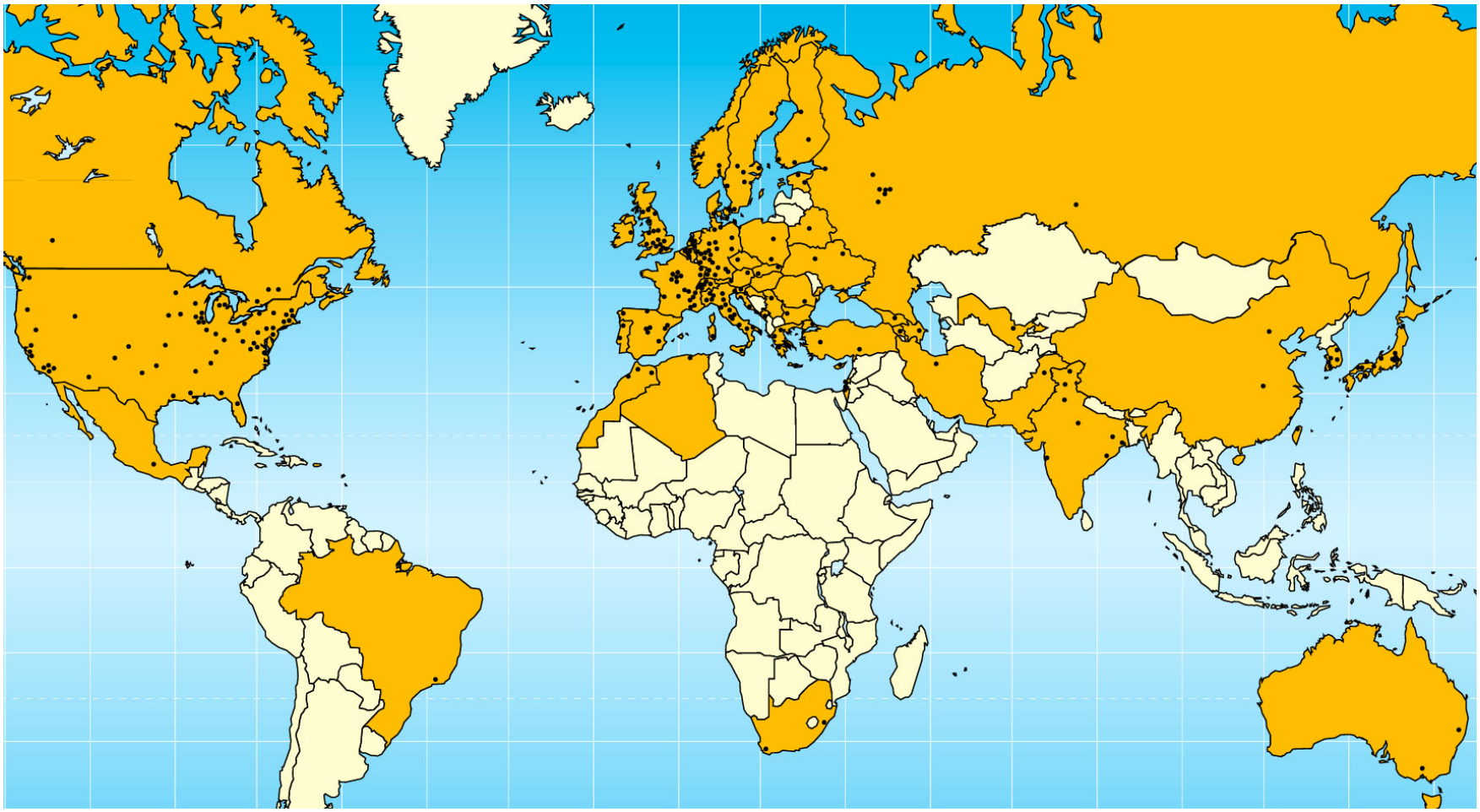

THE UNIVERSITY OF SHEFFIELD


THE UNIVERSITY OF SUSSEX

PRIFYSGOL CYMRU ABERTAWE
UNIVERSITY OF WALES SWANSEA


UCL

CERN Users in the World – A Global VO



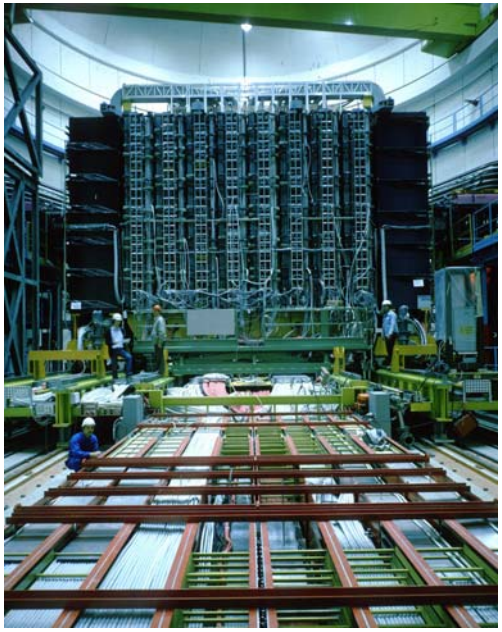
Europe: 267 institutes, 4603 users
Elsewhere: 208 institutes, 1632 users

Particle Physics and Open Access

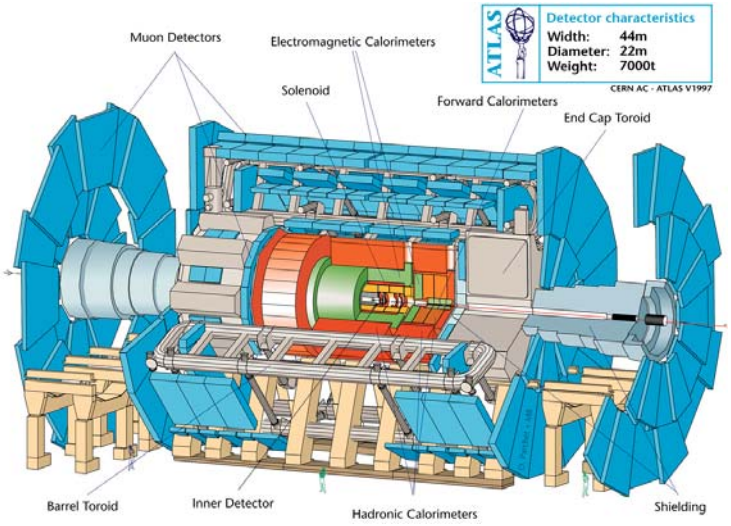
- Tradition of preprint circulation on submission of paper to journal
- Paul Ginsparg set up Los Alamos arXiv as electronic version of this widely used system
- Now relocated to Cornell and content now includes other areas of physics and related subjects

➢ But what about the primary data?

Modern Particle Physics Detectors

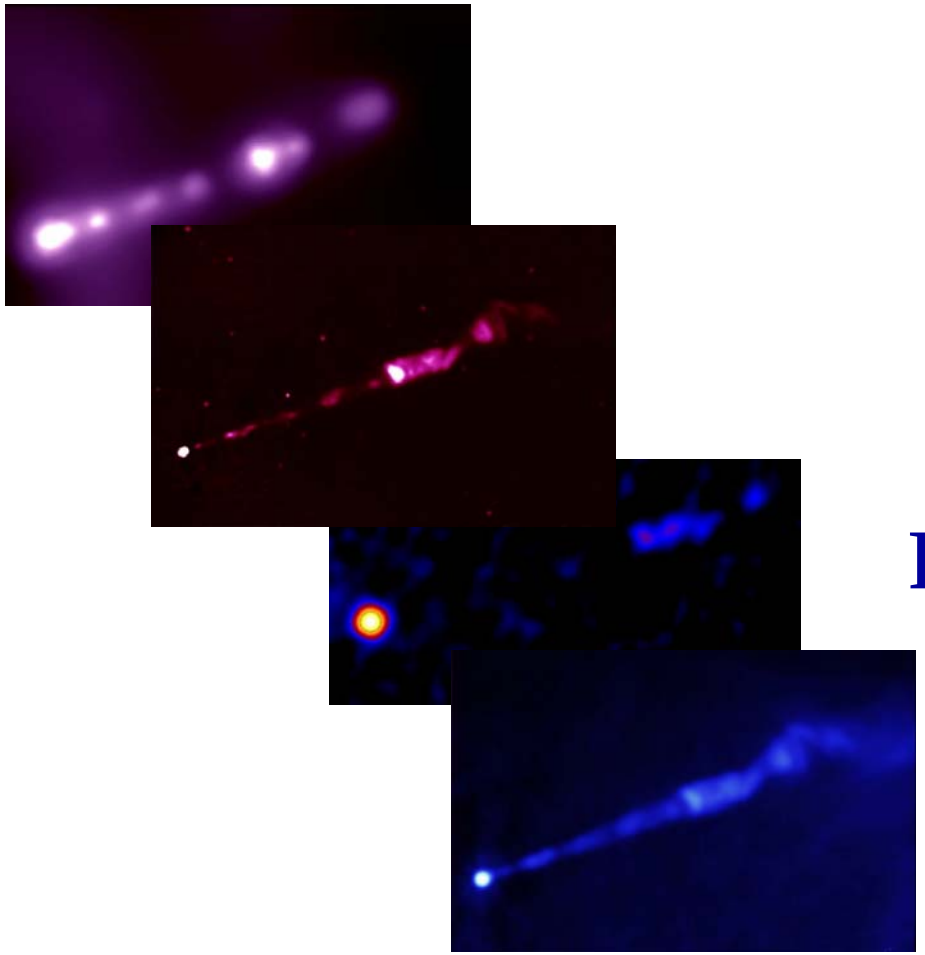


UA1 at CERN 1981-1989
"hermetic detector"



ATLAS at LHC, 2006-2020
$150 * 10^6$ sensors

➢ Can data from such complex experiments be made 'open access'?



AstroGrid

Powering the Virtual Universe

www.astrogrid.ac.uk

Multi-wavelength showing the jet in M87: from top to bottom – X-ray, Optical, Infra-Red and Radio

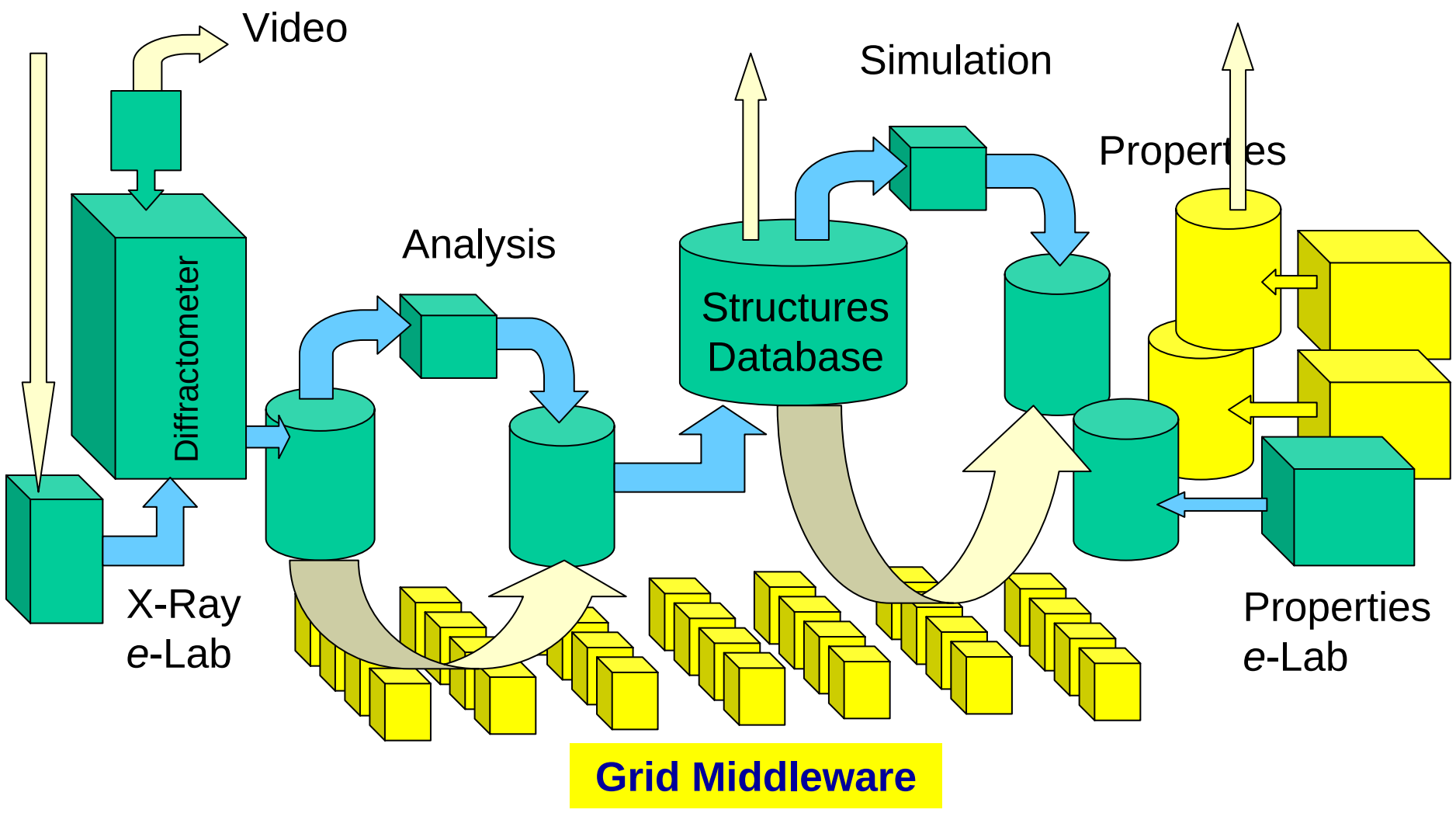
US NVO Project

- NVO development schedules are closely tied to Working Group activities of the International Virtual Observatory Alliance (IVOA)
  - Standards development is international
  - Work is collaborative *and* productive
- NVO tools and applications development linked to demonstrations and public software releases
- On track for having complete IT infrastructure for the NVO by the end of the five-year project

International Virtual Observatory Alliance

- Reached international (IVOA) agreements on Astronomical Data Query Language, VOTable 1.1, UCD 1+, Resource Metadata Schema
- Image Access Protocol, Spectral Access Protocol and Spectral Data Model, Space-Time Coordinates definitions and schema
- Interoperable registries by Jan 2005 (NVO, AstroGrid, AVO, JVO) using OAI publishing and harvesting

Comb-*e*-Chem Project



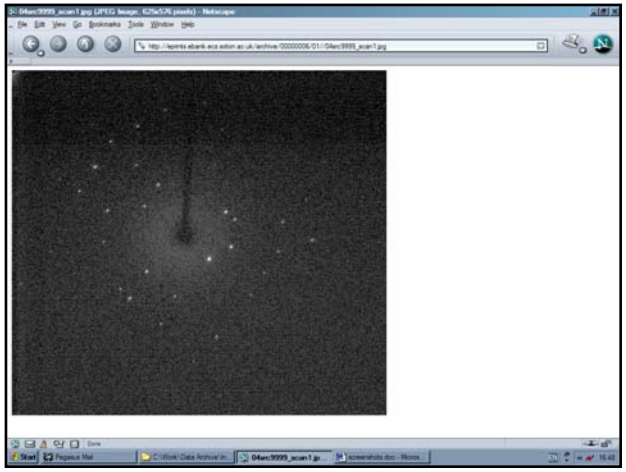
Referee@source or Referee on demand?

- 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

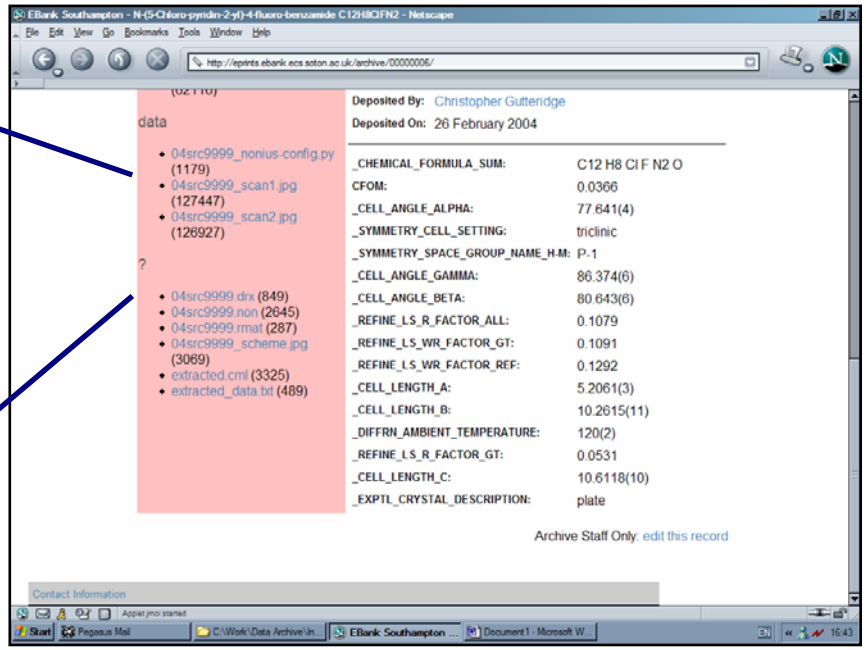
Goals of *e*-Bank Project

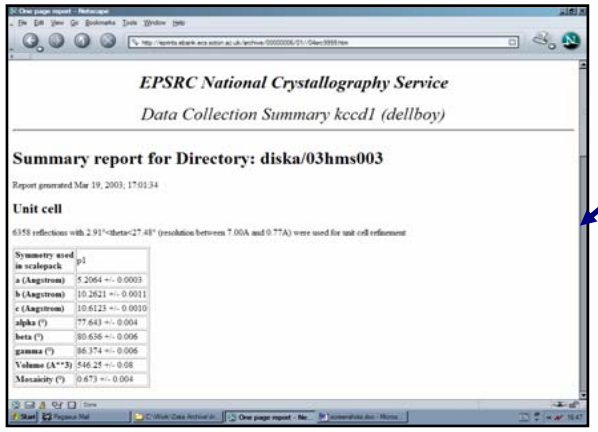
- Provide self archive of results plus the raw and analysed data
- Links from traditionally published work provides the provenance to the work
- Disseminate for “Public Review” – raw data provided so that users can check themselves
- Avoid the “publication bottleneck” but still provide the quality check

Crystallographic e-Prints



- **Direct Access to Raw Data from scientific papers**

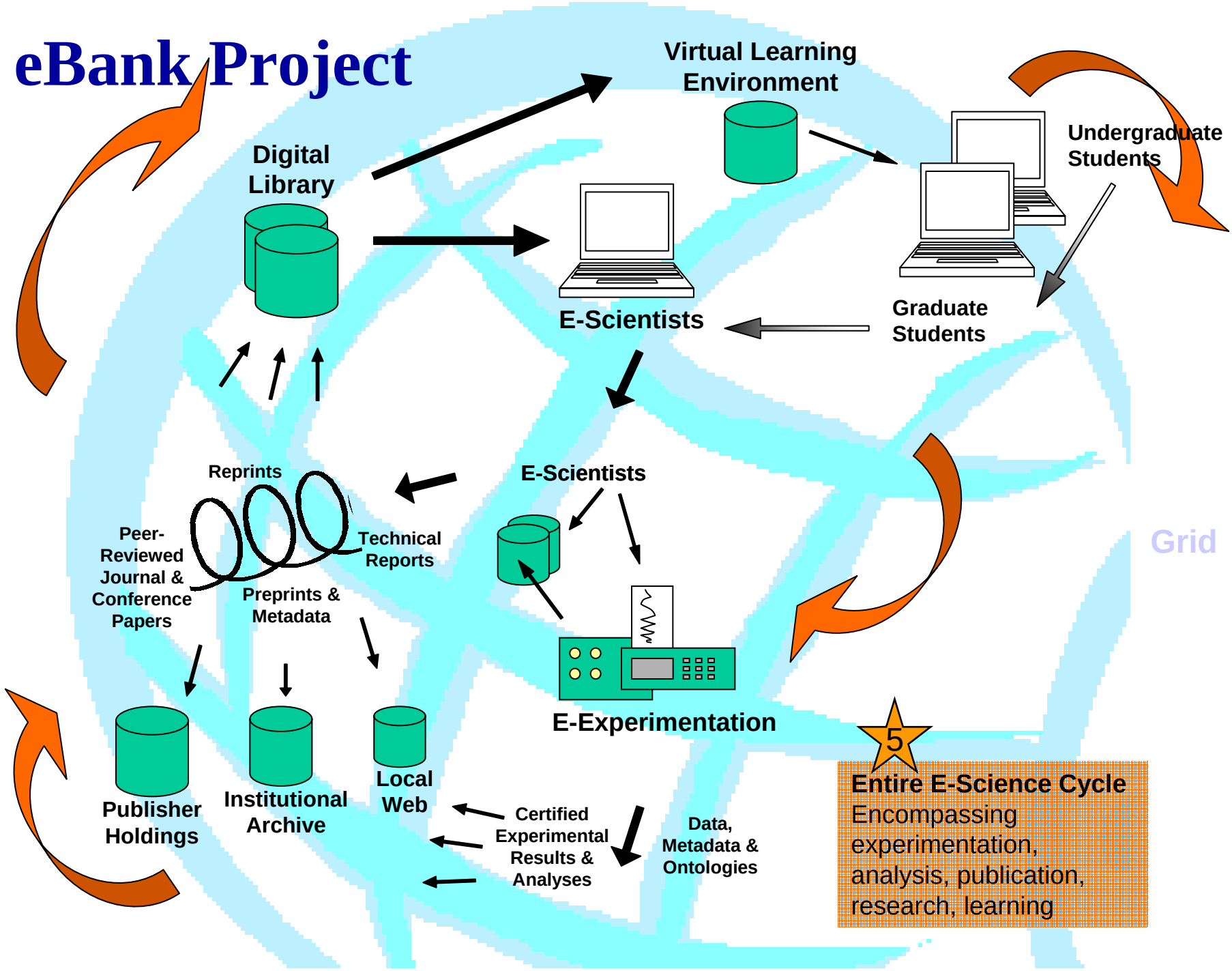




Raw data sets can be very large and these are stored at National Datastore using SRB server

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 extend to provision of real spectra - rather than very reduced summaries - for chemistry publications

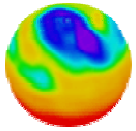

eBank Project
Virtual Learning Environment
Undergraduate Students
Digital Library
E-Scientists
Graduate Students
Reprints
E-Scientists
Peer-Reviewed Journal & Conference Papers
Technical Reports
Preprints & Metadata
Grid
E-Experimentation
Publisher Holdings
Institutional Archive
Local Web
Certified Experimental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

NERC Data Grid Project

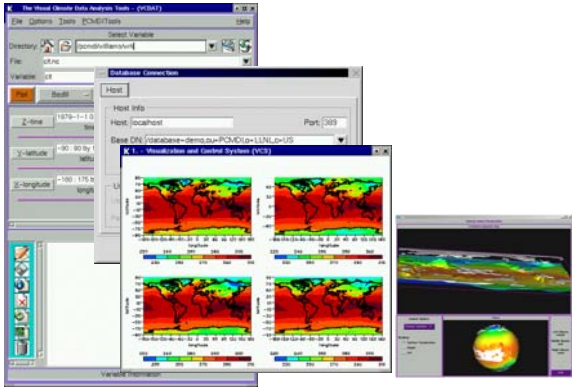
- Objective is to build a Grid that makes data discovery, delivery and use much easier than it is at present
- *Standards compliant* (ISO 19115, 19118), *semantic* data model for maximum interoperability
- Data can be stored in many different ways (flat files, databases…)
- Clear separation between *discovery* and *use* of data.

Complexity + Volume + Remote Access = Grid Challenge

British Atmospheric Data Centre



NERC DataGrid

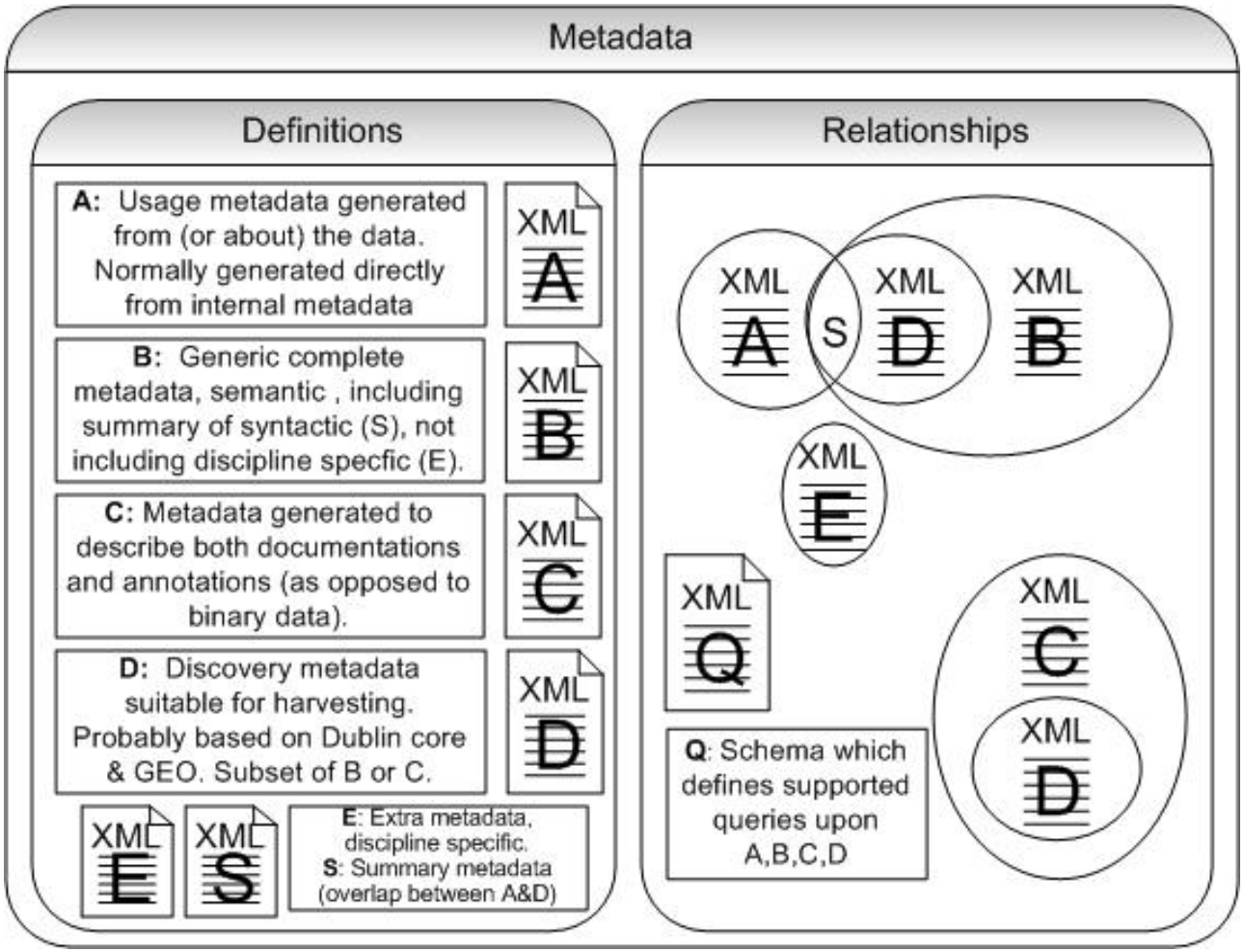


British Oceanographic Data Centre

NDG Key Goals

- Further development of prototype NDG
  - Needs to be much easier to use tools and become a data provider
- Deployment of NDG in
  - HIGEM, BDAN, RAPID, EcoGRID, DEWS (DTI funded), MOTIIVE
  - Possibly: QUEST, e-Jules and other NERC initiatives under discussion.
- International
  - NDG Peering with Earth System Grid (US), SeeGrid (Australia), Hamburg MPI (WDC-A Climate), GO-ESSP
  - Become more involved in OGC/ISO community to get CSML subsumed in GML.
- Sustainability
  - Evaluating the requirements of long term deployment of NDG
  - Split the software and schema development and maintenance from grid infrastructure management.

NDG Metadata Taxonomy



CSML in MarineXML

MarineXML Draft Final report:

"... there is a momentum from organisations such as IHO and WMO to adopt consistent approaches for the vocabulary of their data along the reference implementation of ISO Standards prescribed by the [Open Geospatial Consortium]..."

"The NDG format proved a robust recipient for the data from each community. It produced economical files with few redundant elements, striking about the right balance between weak and strong typing."

Other Key NDG Relationships

- Earth Simulator project depending on NDG information management …
- NCAS Big Data Analysis Network depending on NDG information management
- ECOGrid and RAPID data handling building on NDG foundations
- Met Office, Delivering Environmental Web Services DEWS project (DTI Funded) using NDG guidance, as is their re-engineering project.
- Commercial Geographic Information Systems via Open Geospatial Consortium membership

Digital Curation?

- 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

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
- Research in tools and technologies
  - Integration, annotation, provenance, metadata, security…..

Digital Preservation: The issues

- Long-term preservation
  - Preserving the bits for a long time (“digital objects”)
  - Preserving the interpretation (emulation/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 – but will they count in the UK RAE?
- 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
- 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

Policies for Open Access to Data?

- OECD Statement
- US Examples
  - Bromley
  - NASA
  - NSF
  - NIH
- UK Examples
  - Research Councils
  - Wellcome, Nature
  - FoI

OECD
DECLARATION ON ACCESS TO RESEARCH DATA FROM PUBLIC FUNDING
adopted on 30 January 2004 in Paris

The governments of Australia, Austria, Belgium, Canada, China, the Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, the Russian Federation, the Slovak Republic, the Republic of South Africa, Spain, Sweden, Switzerland, Turkey, the United Kingdom, and the United States recognize that:

- Optimum international exchange of data, information and knowledge contributes decisively to the advancement of scientific research and innovation
- Open access to, and unrestricted use of, data promotes scientific progress and facilitates the training of researchers
- Open access will maximise the value derived from public investments in data collection efforts
- Substantial benefits that science, the economy and society at large could be gained from the opportunities that expanded use of digital data resources
- The risk that undue restrictions on access to and use of research data from public funding could diminish the quality and efficiency of scientific research and innovation

The ‘Bromley Principles’ (1)

- The Global Change Research Program requires an early and continuing commitment to the establishment, maintenance, validation, description, accessibility, and distribution of high-quality, long-term data sets.
- Full and open sharing of the full suite of global data sets for all global change researchers is a fundamental objective.
- Preservation of all data needed for long-term global change research is required. For each and every global change data parameter, there should be at least one explicitly designated archive. Procedures and criteria for setting priorities for data acquisition, retention, and purging should be developed by participating agencies, both nationally and internationally. A clearinghouse process should be established to prevent the purging and loss of important data sets.
- Data archives must include easily accessible information about the data holdings, including quality assessments, supporting ancillary information, and guidance and aids for locating and obtaining the data.

The ‘Bromley Principles’ (2)

- National and international standards should be used to the greatest extent possible for media and for processing and communication of global data sets.
- Data should be provided at the lowest possible cost to global change researchers in the interest of full and open access to data. This cost should, as a first principle, be no more than the marginal cost of filling a specific user request. Agencies should act to streamline administrative arrangements for exchanging data among researchers.
- For those programs in which selected principal investigators have initial periods of exclusive data use, data should be made openly available as soon as they become widely useful. In each case, the funding agency should explicitly define the duration of any exclusive use period.

Source: Data Management for Global Change Research Policy Statements, U.S. Global Change Research Program, July 1991.

NASA's Data Policy (1999): Earth Science Enterprise

- NASA is committed to full and open sharing of ESE data from its funded and owned systems
  - No period of exclusive access
  - Access is extended to the full scientific community and the general public
- NASA is committed to non-discriminatory access to data
  - This principle allows for the possibility that categories of users can be established, with preferential treatment between the categories
  - All users in a clearly defined user category can obtain data on the same terms and conditions, and the categories are defined in such a way that all potential users can be assigned to a category
  - All data required for long-term global change research shall be archived

NASA - International Issues

- Cooperative activities with international partners present unique circumstances because different laws and policies govern each country
- NASA's international agreements must address data policy: each party's rights and responsibilities for data acquisition, distribution and archiving
- As a minimum requirement for any mission, NASA-affiliated users should be able to obtain and use data from cooperative missions
- Access to data should be made as broad and inexpensive as possible to facilitate maximum utilization by the scientific community

National Science Foundation
Office of Polar Programs

General Guidelines for all OPP supported projects:

- All data and derived data products collected under OPP-awards which are appropriate for submission to a national data center or OPP specified data repository should be promptly submitted within a reasonable amount of time, as described below
- OPP considers the documentation of data sets, known as metadata, as vital to the exchange of information on polar research and to a data set's accessibility and longevity for reuse.
- Data archives of OPP-supported projects should include easily accessible information about the data holdings, including quality assessments, supporting ancillary information, and guidance for locating and obtaining the data.
- National and international standards should be used to the greatest extent possible for the collection, processing and communication of OPP-sponsored data sets.

NIH Data Sharing Policy (2003)

- ‘Data should be made as widely and freely available as possible while safeguarding the privacy of participants, and protecting confidential and proprietary data’
- Data Sharing Methods:
  - Publishing – scientific publications
  - Researcher’s efforts – CD, Web site
  - Data Enclave – secure, access controlled
  - Data Archive – ingest, curation and distribution
  - Mixed Mode – multiple levels of access

NIH cancer Biomedical Informatics Grid (caBIG) Project

Software, data, standards, infrastructure directly supported by caBIG resources must be open source and open access (i.e. licensed to the government with the government having no restrictions with regard to redistribution).

NIH Grant Condition on Sharing Data

Data Sharing Plan:

- The reasonableness of the data sharing **plan** or the rationale for not sharing research data **will** be assessed by the reviewers
- Reviewers will not factor the proposed data sharing plan into the determination of scientific merit or the priority score
- The presence of a data sharing **plan** will be part of the terms and conditions of the award
- The funding organization will be responsible for monitoring the data sharing policy

UK Research Councils

- NERC, ESRC and the AHRB have well developed policies on data management
- MRC have a draft policy on Data Sharing and Preservation
- CCLRC are taking lead for RCUK on data archives and curation
- At present BBSRC has no stated policy but have established a Data Policy Working Group
- PPARC astronomy community are providing the infrastructure and common standards for world wide Virtual Observatories
- EPSRC are not aware of ‘any big problems within the EPSRC community regarding access to scientific data’

Joint Data Standards Study

- Sponsors: MRC, BBSRC, Wellcome Trust, NERC, DTI, JISC JCSR
- Study will examine large-scale data sharing in the life sciences: data standards, incentives, barriers and funding models
- Explore models of best practice for creation and implementation of data-sharing tools and standards
- Identify incentives and barriers to data sharing

➢ Completion due in ‘early 2005’

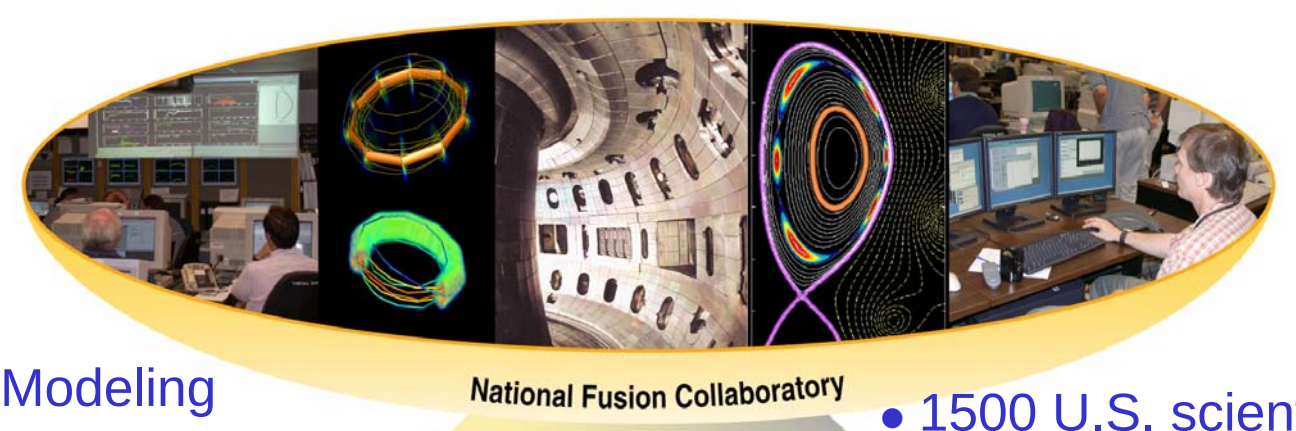
NERC Information Strategy

- Information management and dissemination is recognised as a priority task in all NERC-funded work
- Responsibility for NERC's data lies with its specialist data centres
- Data management plans will be mandatory component of all project plans
- NERC Data Grid will be key technology for enabling seamless access to NERC's data holdings
- NERC will make its information fully accessible as required by the Baker Report in 1999

Data Sharing in the Fusion Community

- Detailed, analysed, results from JET plus metadata are available to all members of the EFDA and to certain others (in US, Japan, Russia)
- As for particle physics the raw data is not available since it would not be useful
- Publication of results based on analysis of these data requires clearance from EFDA-JET
- All major tokamak machines around the world submit data to the ITPA group whose data base is publicly available

THE NATIONAL FUSION COLLABORATORY PROJECT SEEKS TO UNIFY SCIENTISTS ACROSS THE UNITED STATES



- Theory & Modeling
  - Realistic non–linear 3D models
- Experimental Facilities
  - $1B capital investment

- 1500 U.S. scientists
  - 90 sites, 37 states
  - USDOE OFES funded

VISION FOR THE FUSION GRID

- Data, Codes, Analysis Routines, Visualization Tools should be thought of as network accessible services
- Shared security infrastructure
- Collaborative nature of research requires shared visualization applications and widely deployed collaboration technologies
  - Integrate geographically diverse groups
- Not focused on CPU cycle scavenging or “distributed” supercomputing (typical Grid justifications)
  - Optimize the most expensive resource - people’s time

Wellcome Trust's Policy on Data

The Trust's guidelines on good research practice state:

- While recognizing the need for scientists to protect their own research interests, the Trust encourages the researchers whom it funds to be as open as possible in discussing their work with other scientists and with the public
- Once results have been published, the Trust expects researchers to make available relevant data and materials to other researchers, on request, provided that this is consistent with any ethics approvals and consents which cover the data and materials and any intellectual property rights in them
- The Trust recognizes that publication of the results of research may need to be delayed for a reasonable period pending protection of intellectual property arising from the research. However, any such periods of delay in publication should be kept to a minimum

➢ Trust now considering requiring a Data Sharing Plan similar to that required by NIH

Nature’s Policy on Data Availability

- An inherent principle of publication is that others should be able to replicate and build upon the authors' published claims
- A condition of publication in *Nature* is that authors are required to make materials, data and associated protocols available to readers on request
- Any restrictions on the availability of materials or information must be disclosed at the time of submission of the manuscript

UK Freedom of Information Act

- Principle is that all publicly held information should be available to the public unless not in public interest
- Sets out number of exemptions such as the security services and personal data
- Statistical information is NOT covered by an exemption but unfortunately there is no definition of ‘statistical information’
- Not clear how this applies to scientific data!

Conclusions

- The principle of open access to publicly funded scientific data is widely supported internationally
- The individual Research Councils have data sharing policies in different stages of maturity
- e-Science has the potential to transform the way universities and industry pursue research

e-Government and the Grid

‘[The Grid] intends to make access to computing power, scientific data repositories and experimental facilities as easy as the Web makes access to information.’

Tony Blair, 2002

Acknowledgements

With special thanks to Ken Buetow, Peter Burnhill, Peter Buneman, Jeremy Frey, Nancy Hey, Jon Hoare, Liz Lyon, Bryan Lawrence and Mark Thorley