e-Science and Scholarly Communication

Tony Hey
Corporate VP for Technical Computing
Microsoft Corporation

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

Former Director General of Research Councils

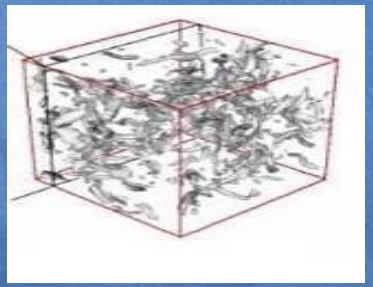
Office of Science and Technology, UK

A New Science Paradigm

- **Thousand years ago:**
  Experimental Science
  - description of natural phenomena
- **Last few hundred years:**
  Theoretical Science
  - Newton's Laws, Maxwell's Equations …

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$

- **Last few decades**:
  Computational Science
  - simulation of complex phenomena



- **Today:**
  e-Science or Data-centric Science
  - unify theory, experiment, and simulation
  - using data exploration and data mining
    - Data captured by instruments
    - Data generated by simulations
    - Data generated by sensor networks
    - Scientist analyzes databases/files

(With thanks to Jim Gray)

e-Science

- e-Science is about data-driven, multidisciplinary science and the technologies to support such distributed, collaborative scientific research
  - Many areas of science are now being overwhelmed by a ‘data deluge’ from new high-throughput devices, sensor networks, satellite surveys …
  - Areas such as bioinformatics, genomics, drug design, engineering and healthcare require collaboration between different domain experts
- ‘e-Science’ is a shorthand for a set of technologies to support collaborative networked science
- HPC and Information Management are key technologies to support this e-Science revolution

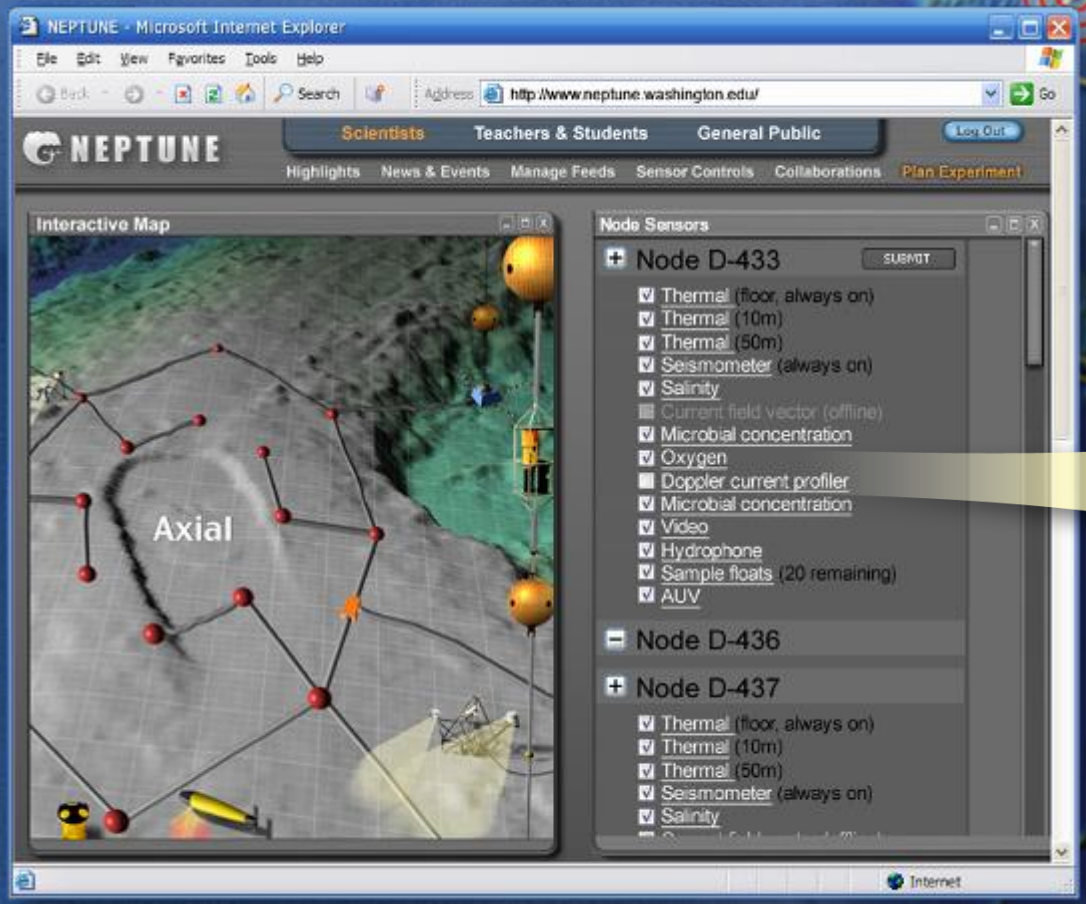
Vision For Scientific Workflow

Example: Project NEPTUNE

http://www.neptune.washington.edu/

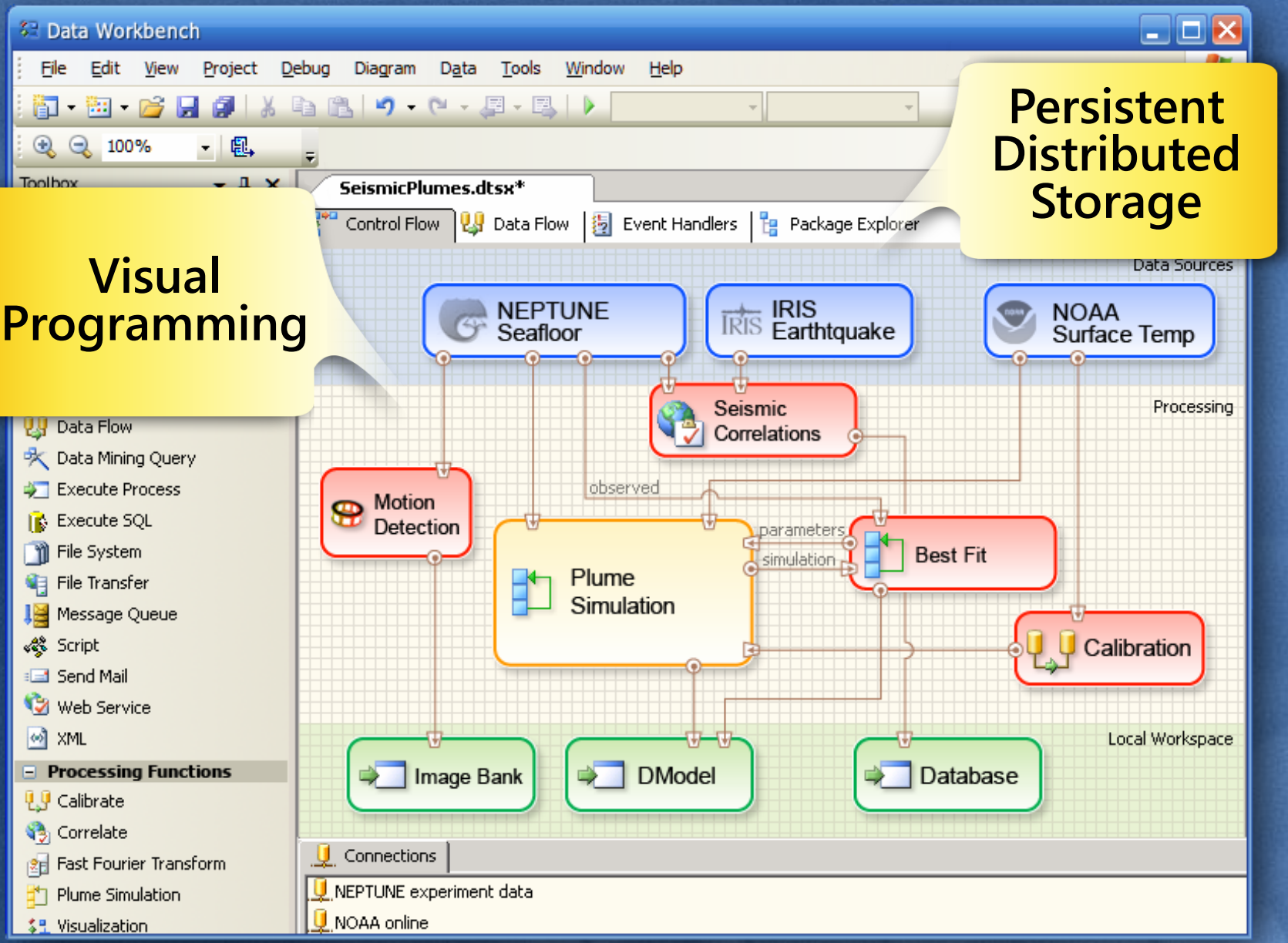
Programmable Sensors & Remote Instruments

Undersea Sensor Network



Connected & Controllable Over the Internet

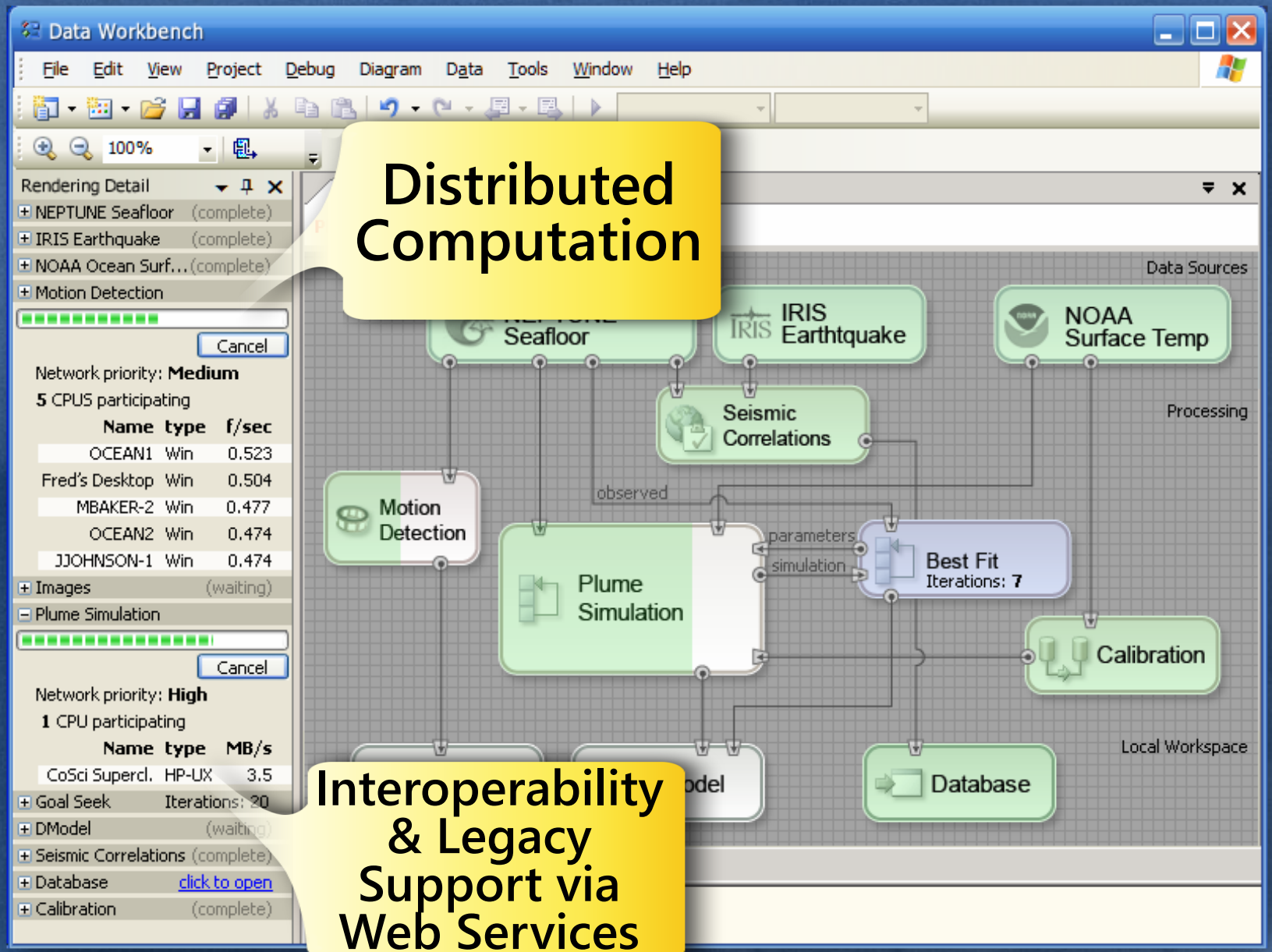
Data Workbench



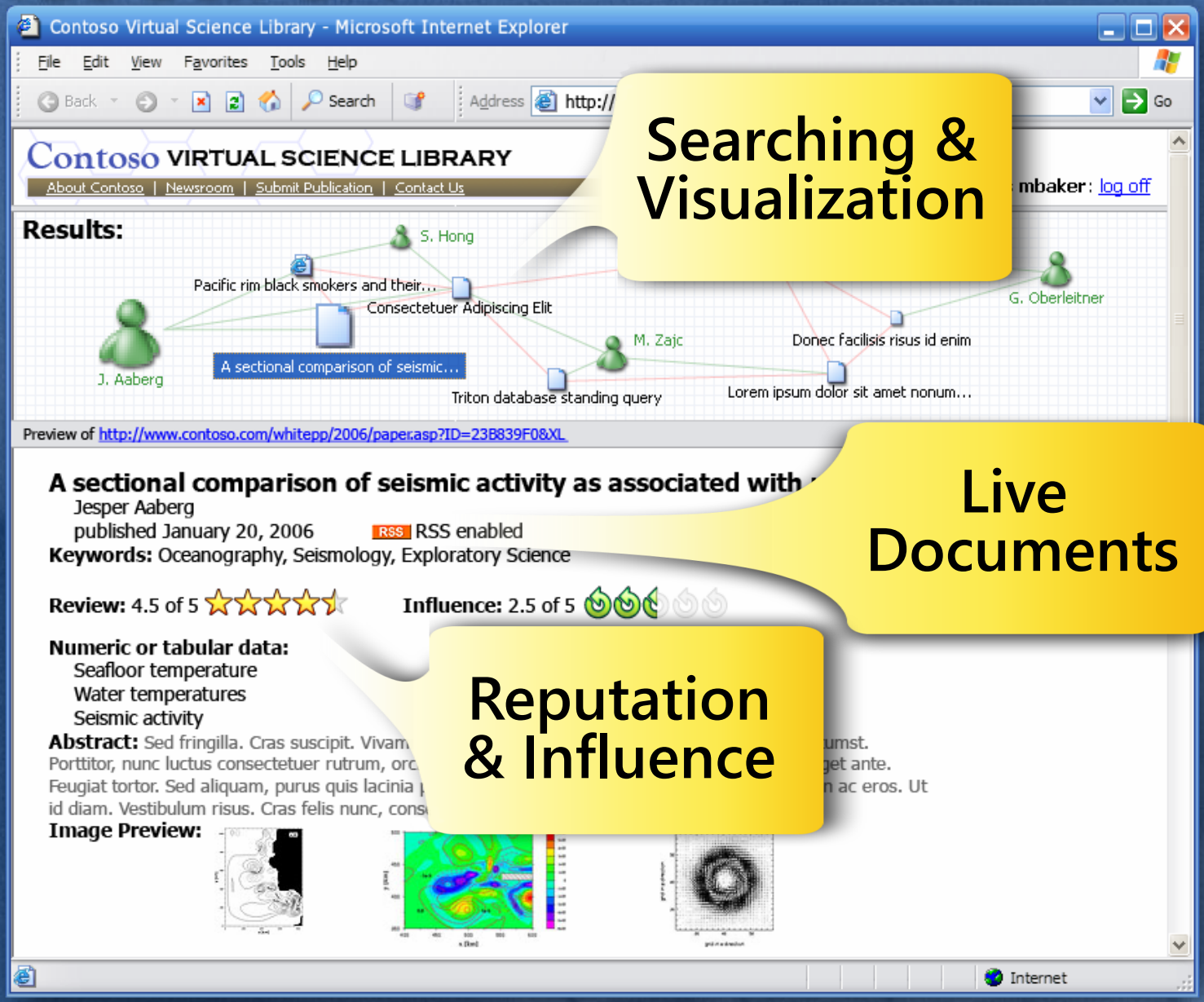
Visual Programming

Persistent Distributed Storage

Data Workbench



Research



Searching & Visualization

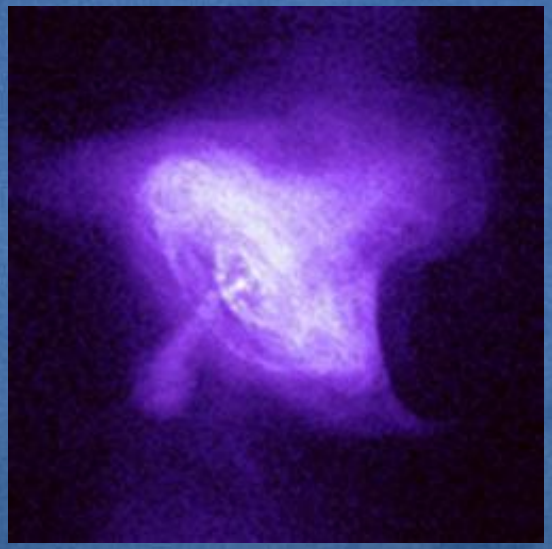
Live Documents

Reputation & Influence

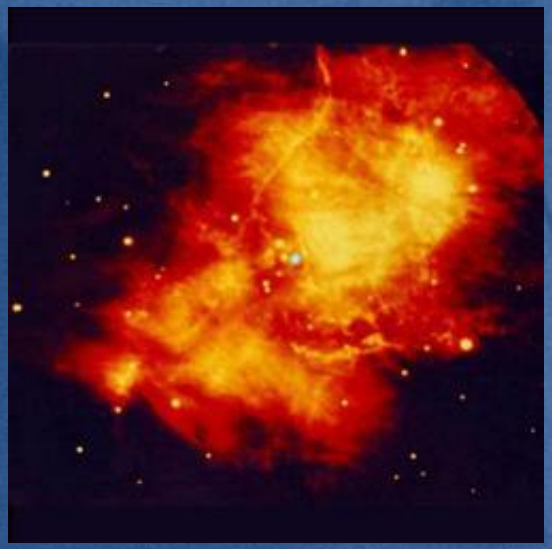
Two examples of e-Science

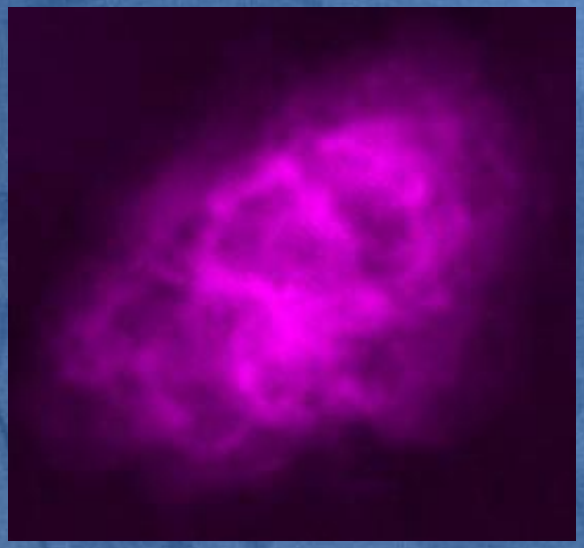
- Astronomy – The International Virtual Observatory
- Chemistry – The Comb-e-Chem Project

The Multiwavelength Crab Nebulae



Crab star
1053 AD



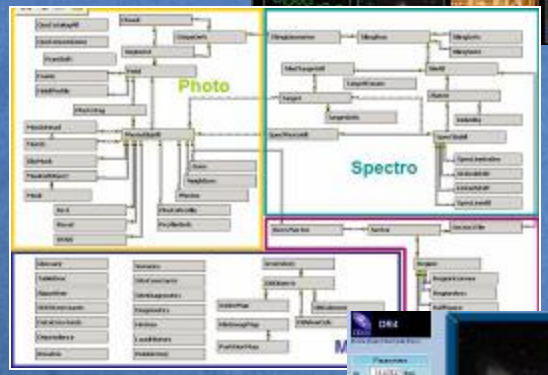


X-ray, optical, infrared, and radio views of the nearby Crab Nebula, which is now in a state of chaotic expansion after a supernova explosion first sighted in 1054 A.D. by Chinese Astronomers.

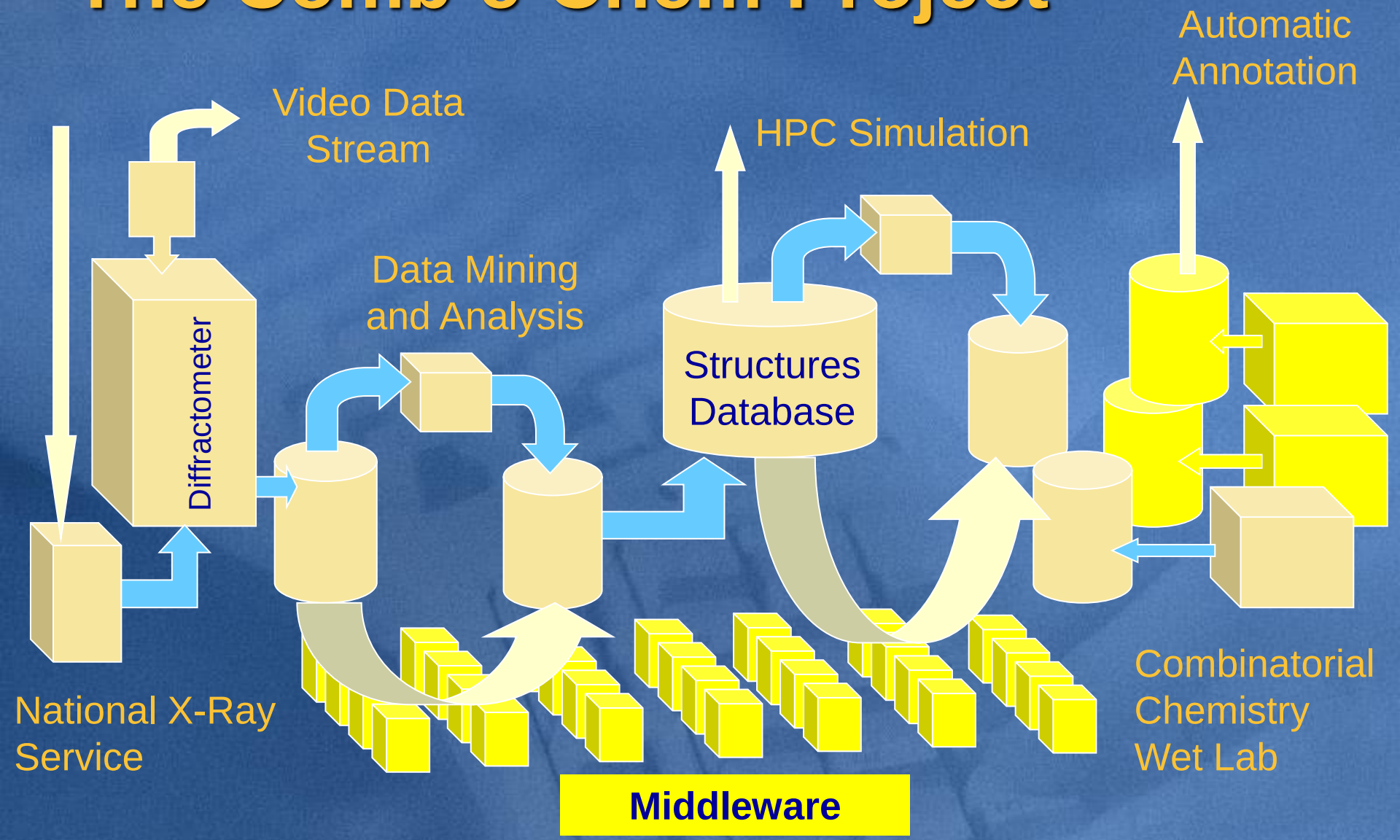
Slide courtesy of Robert Brunner @ CalTech.

IVO: An Astronomy Data Grid

- Working to build world-wide telescope
  - All astronomy data and literature
  - online and cross indexed
  - Tools to analyze it
- Built SkyServer.SDSS.org
- Built Analysis system
  - MyDB
  - CasJobs (batch job)
- OpenSkyQuery
  Federation of ~20 observatories.
- Results:
  - It works and is used every day
  - Spatial extensions in SQL 2005
  - A good example of Data Grid
  - A good example of Web Services



The Comb-e-Chem Project



National Crystallographic Service

Send sample material to NCS service

Collaborate in e-Lab experiment and obtain structure

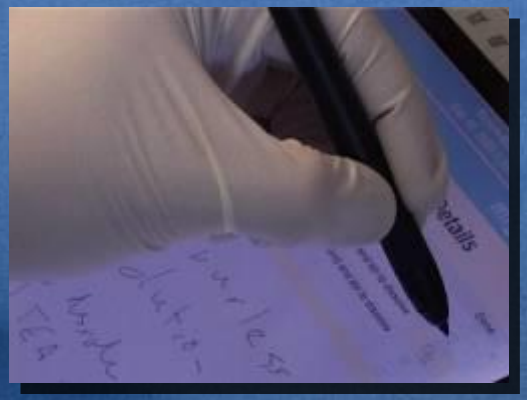
Search materials database and predict properties using Grid computations

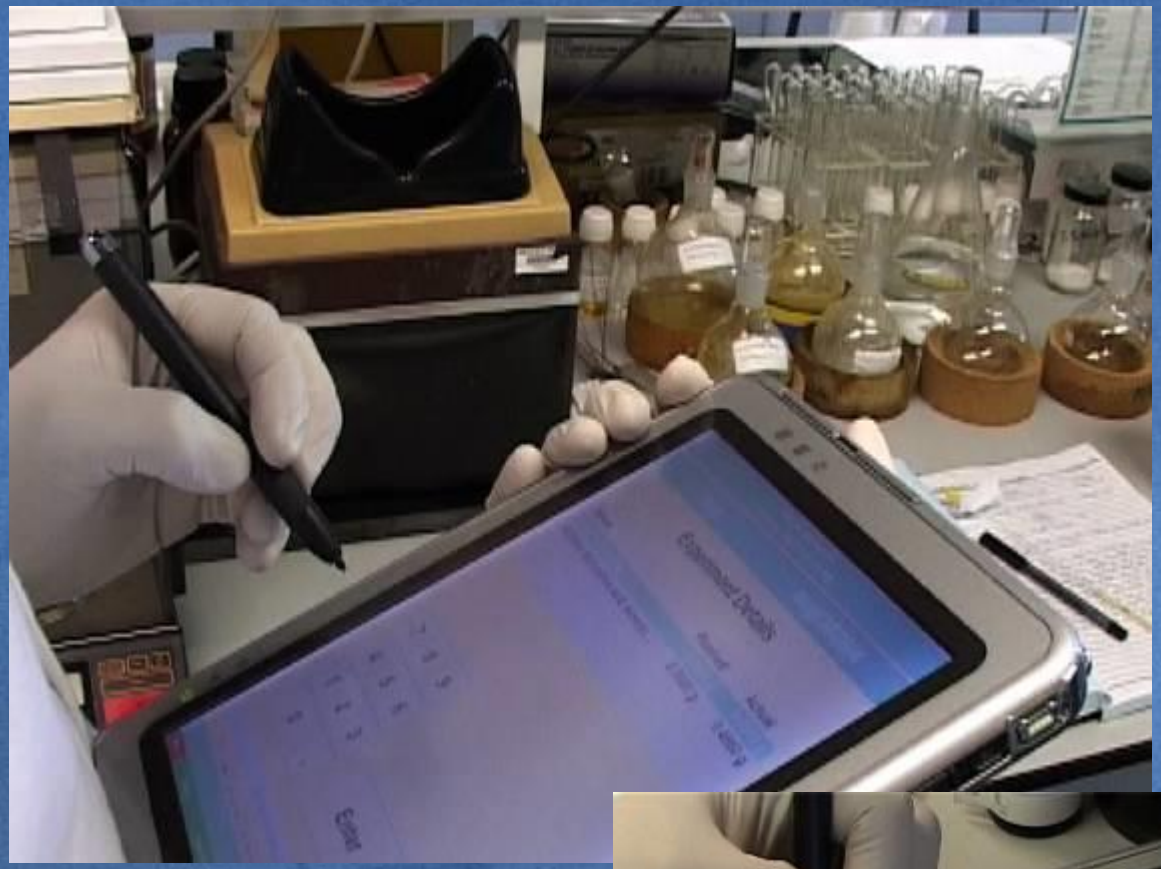
Download full data on materials of interest

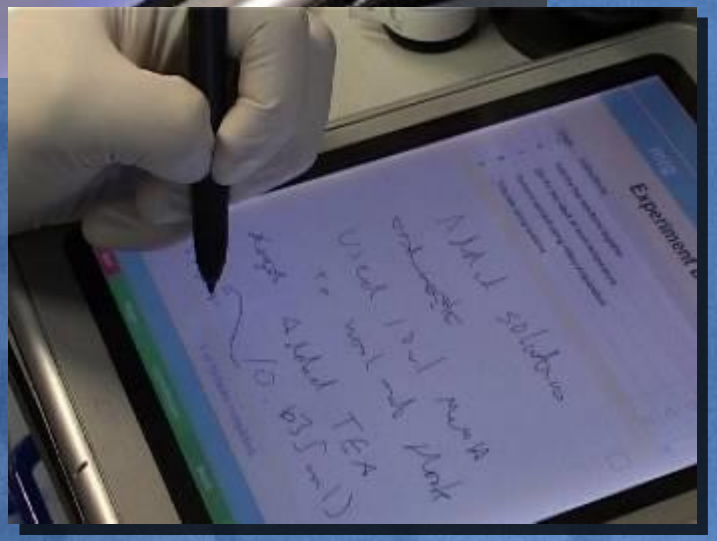
X-Ray e-Laboratory

Structures Database

Computation Service



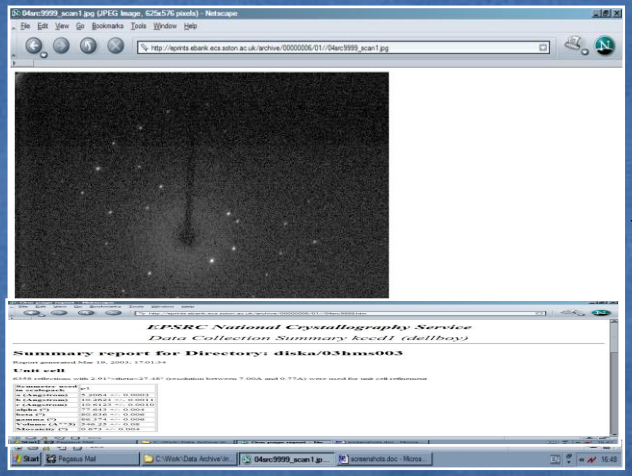




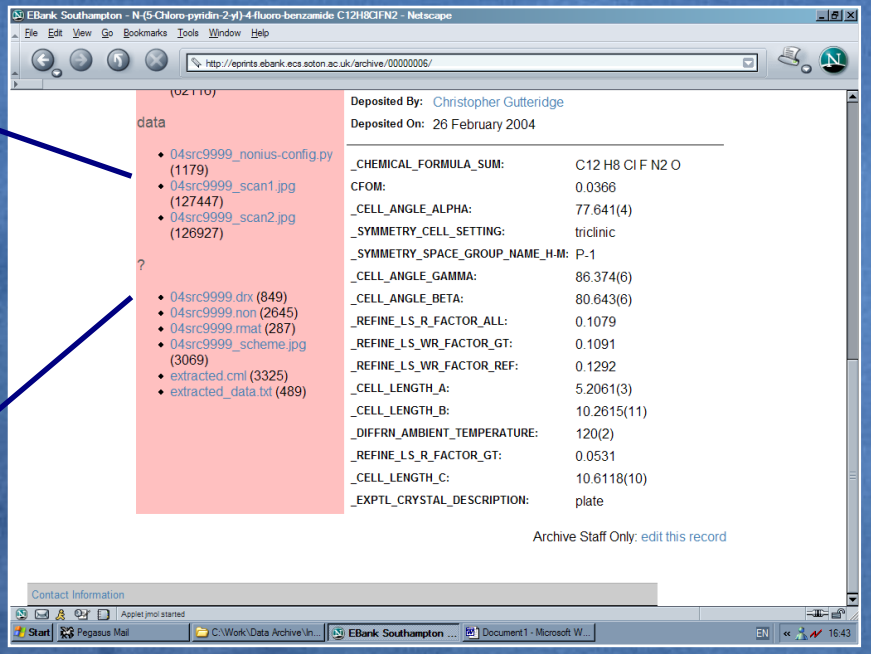
A digital lab book replacement that chemists were able to use, and liked

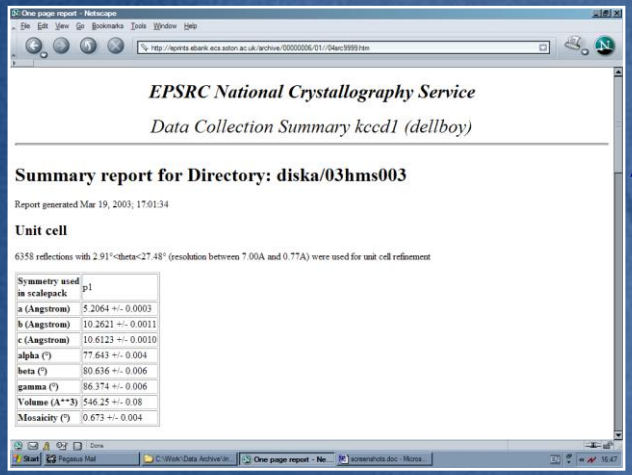
Monitoring laboratory experiments using a broker delivered over GPRS on a PDA

Crystallographic e-Prints



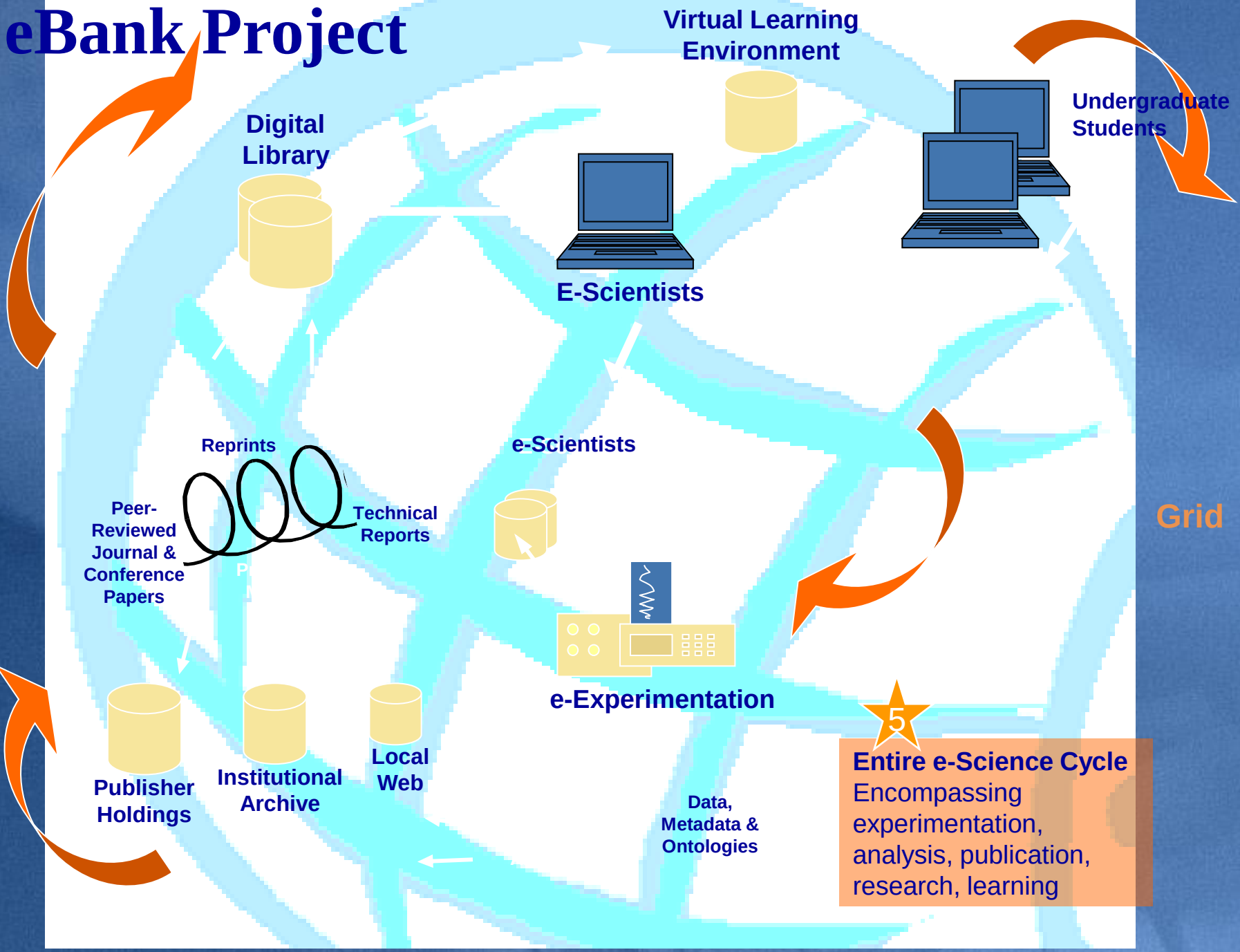
Direct Access to Raw Data from scientific papers





Raw data sets can be very large - stored at UK National Datastore using SRB software

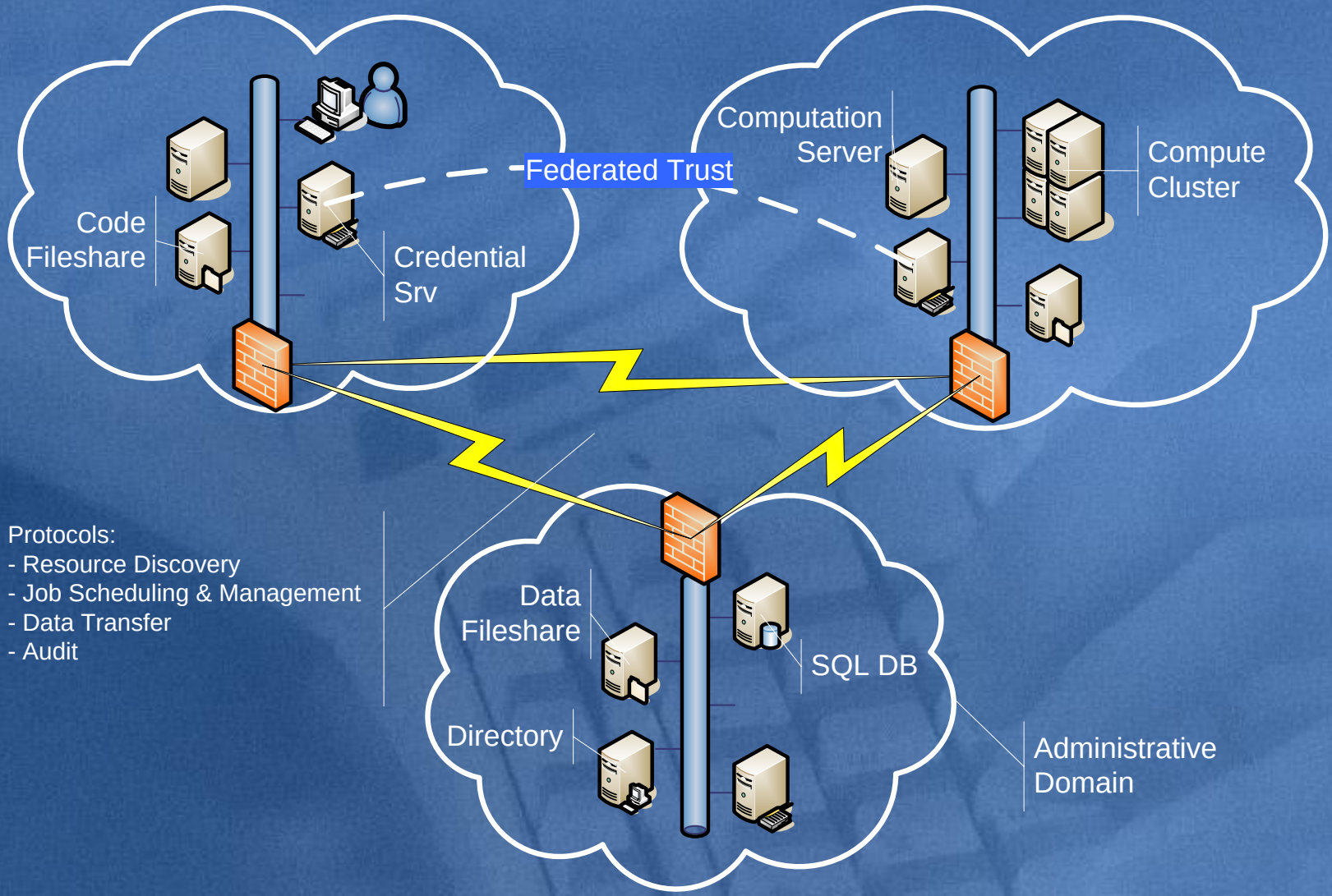
eBank Project



Cyberinfrastructure

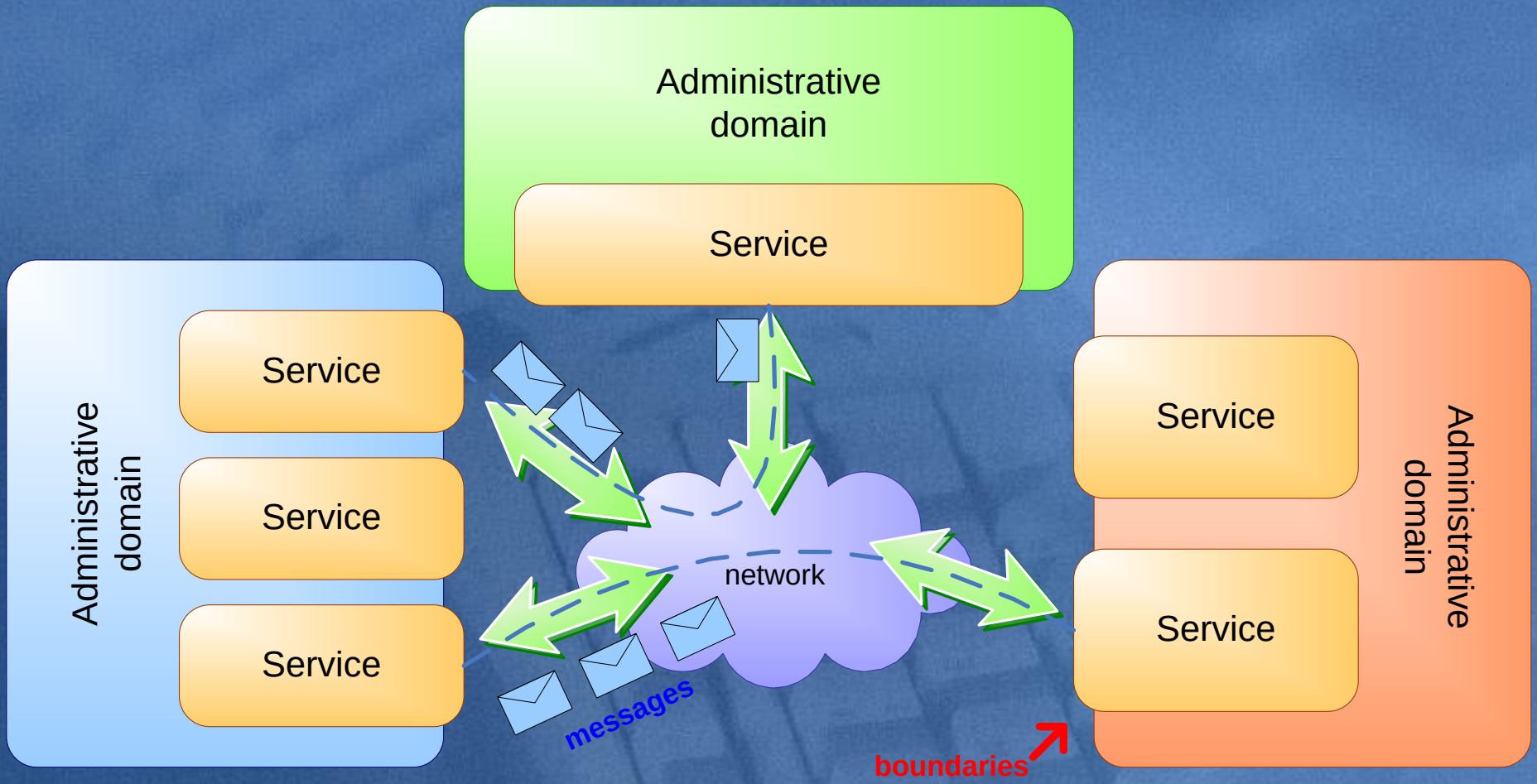
- In the US, Europe and Asia there is a common vision for the ‘cyberinfrastructure’ required to support the e-Science revolution
- Set of Grid Middleware Services supported on top of high bandwidth academic research networks
- Opportunity for Computer Science community to provide scientists with powerful new tools to analyze their data
- Open access federation of research repositories containing full text and data

Grids for Virtual Organizations

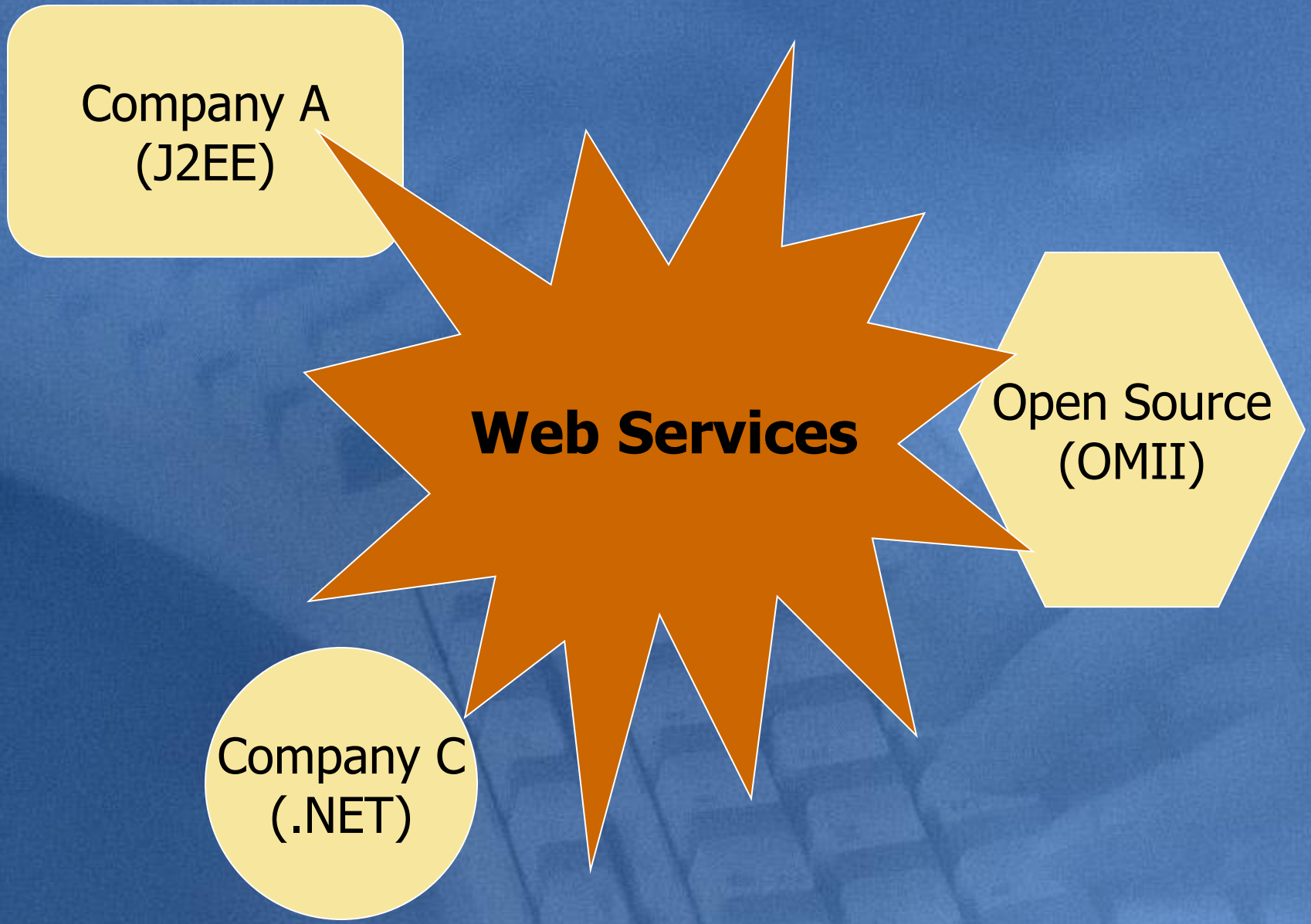


Protocols:
- Resource Discovery
- Job Scheduling & Management
- Data Transfer
- Audit

Service-Orientation for building Distributed Systems



Web Services and Interoperability



Microsoft Open Specification Promise (September 12 2006)

- Covers Web Services specifications
  - SOAP, WSDL, WS-I, WS-Security, WS-Management, WS-Eventing, WS-Addressing ….
- Q: How does the Open Specification Promise work? Do I have to do anything in order to get the benefit of this OSP?
- A: No one needs to sign anything or even reference anything. Anyone is free to implement the specification(s), as they wish and do not need to make any mention of or reference to Microsoft. Anyone can use or implement these specification(s) with their technology, code, solution, etc. You must agree to the terms in order to benefit from the promise; however, you do not need to sign a license agreement, or otherwise communicate your agreement to Microsoft.

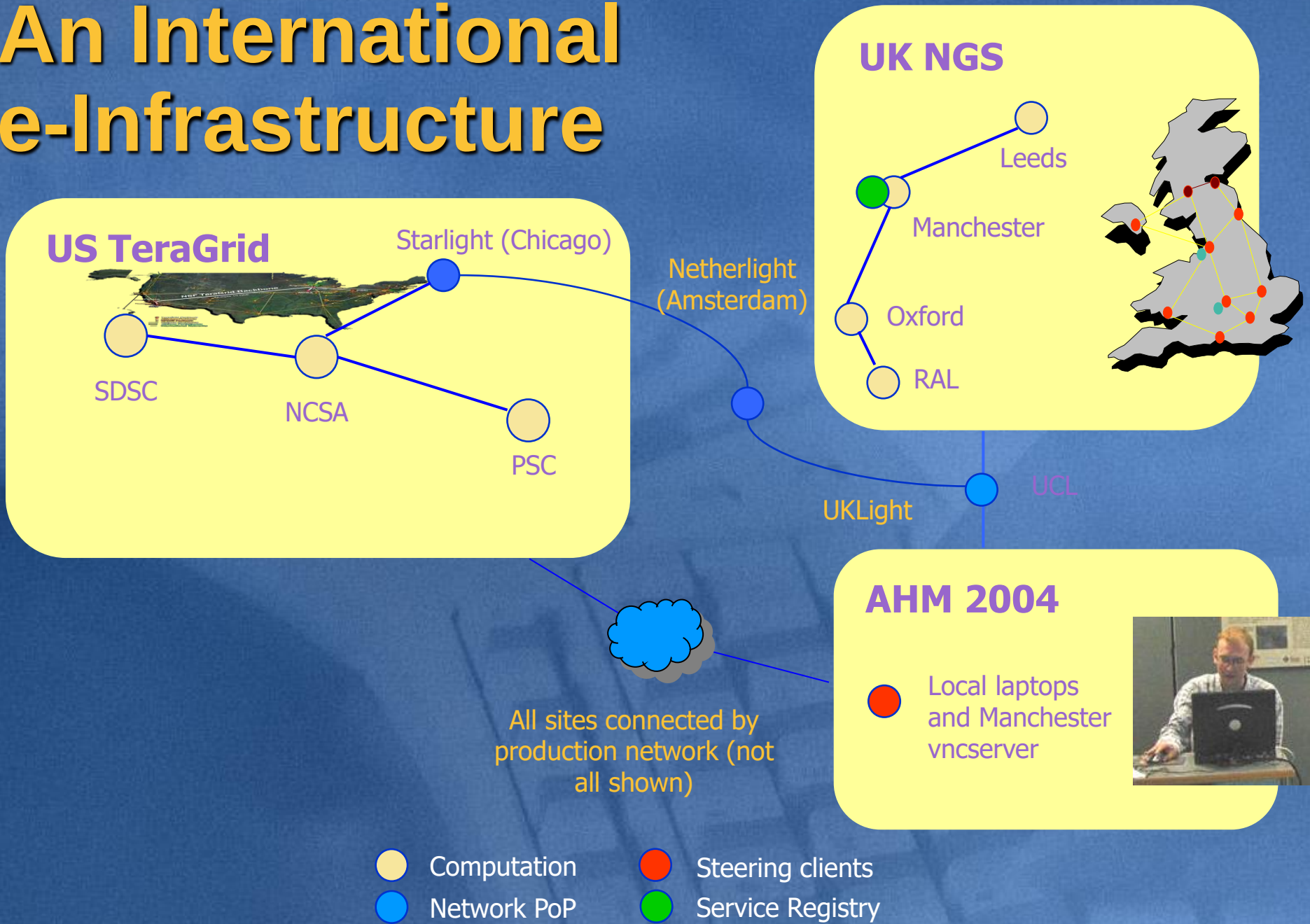
Progress in Grid Standards?

- The GGF/EGA merger gives great opportunity for the new Open Grid Forum (OGF) to standardize a small set of basic Grid services based on generally accepted Web Services
  - Harness the power of the world-wide Grid community to develop robust open source reference implementations
- Grid research community needs to propose and explore new features in real experiments
  - OGF can reassure industry about progress in Grid standards and grow the market for all

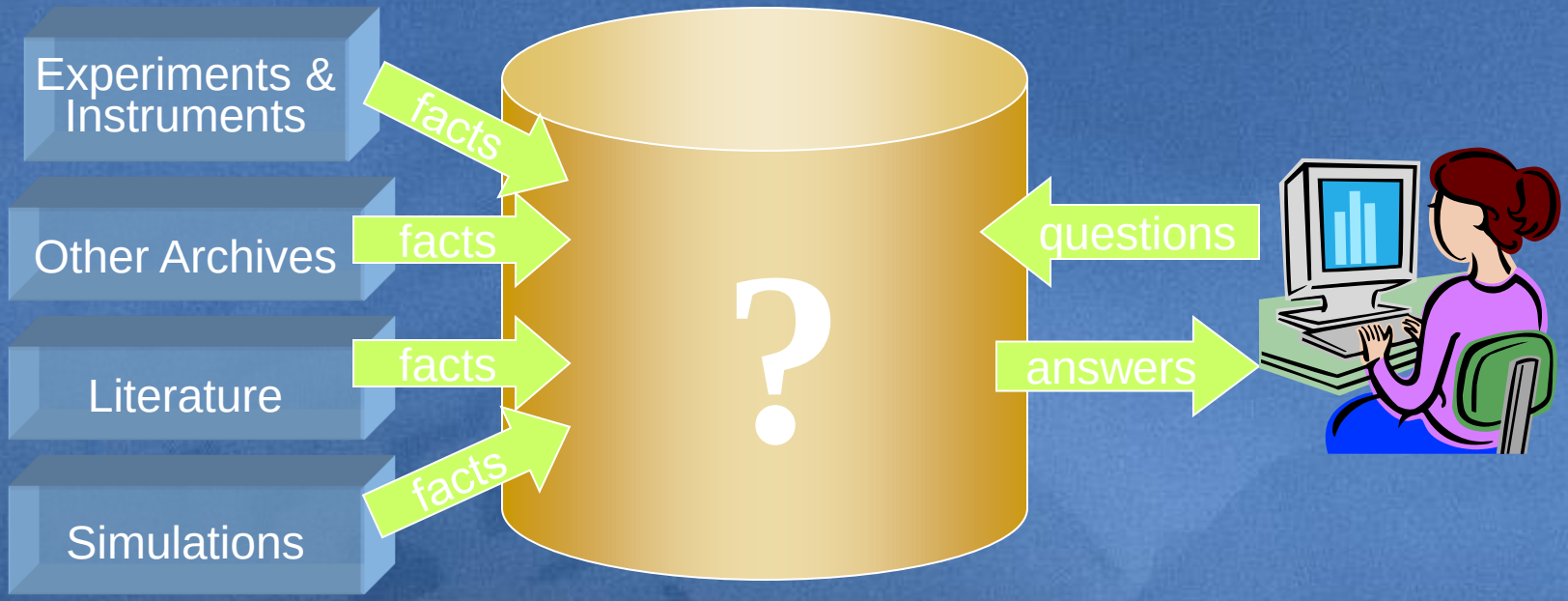
Key Data Issues for e-Science

- Networks
  - Lambda technology
- The Data Life Cycle
  - From Acquisition to Preservation
- Scholarly Communication
  - Open Access to Data and Publications

An International e-Infrastructure



The Problem for the e-Scientist



- Data ingest
- Managing a petabyte
- Common schema
- How to organize it?
- How to *re*organize it?
- How to coexist & cooperate with others?

- Data Query and Visualization tools
- Support/training
- Performance
  - Execute queries in a minute
  - Batch (big) query scheduling

The e-Science Data Life Cycle

- Data Acquisition
- Data Ingest
- Metadata
- Annotation
- Provenance
- Data Storage
- Data Cleansing
- Data Mining
- Curation
- Preservation

Publishing Data & Analysis Is Changing

Roles	*Traditional*	*Emerging*
Authors	**Scientists**	**Collaborations**
Publishers	**Journals**	**Project web site**
Curators	**Libraries**	**Data+Doc Archives**
Archives	**Archives**	**Digital Archives**
Consumers	**Scientists**	**Scientists**

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
- Provenance
  - ‘Where did this data come from?’
- Exporting/publishing in agreed formats
  - To other programs as well as people
- Security
  - Specifying/enforcing read/write access to *parts* of your data

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’

OECD Declaration on Access to Research Data from Public Funding (January 2004)

Supported by 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 UK and the United States

OECD Declaration recognizes:

- 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

NIH Data Sharing

- **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 Plan (2005)**
  - The reasonableness of the data sharing plan or the rationale for not sharing research data will be assessed by the reviewers
  - The presence of a data sharing plan will be part of the terms and conditions of the award

Scholarly Communication

- Global Movement towards permitting 'Open Access' to scholarly publications
  - Libraries can no longer afford publisher subscriptions
  - Principle that results of publicly funded research should be available to all
- Mandates for Open Access
  - US Proposal – Cornyn-Lieberman Bill
    - Supported by most top US research universities
  - EU Proposals
    - UK, France and German initiatives

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’

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.’

Open Access and Scholarly Publishing

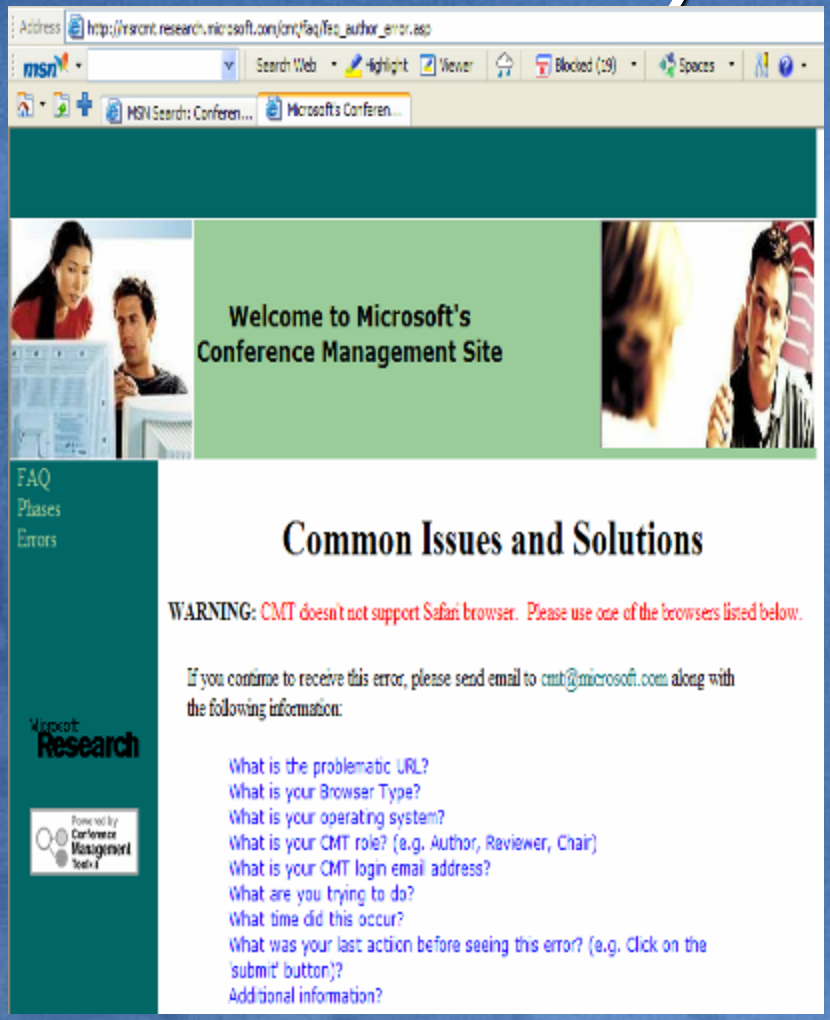
- Goal is to work with the research community to assist them in developing open and interoperable frameworks for scholarly publishing
- Two aspects
  - ‘Community publishing’ toolset
  - Service Oriented Framework for Interoperable Repositories

Community Publishing

- Develop toolset for ‘self-publishing’ of workshop and conference proceedings
  - Base development around existing MSR Workshop tool ‘CMT’
  - Work with forward-looking publishers to develop new publishing models
- Offer Microsoft as one site where such academic publications can be kept ‘in perpetuity’?
  - Important that Microsoft is not only repository – cf LOCKSS and Portico

CMT: Conference Management Tool

- **Currently support a conference** peer-review **system (~300 conferences)**
  - **Form committee**
  - **Accept Manuscripts**
  - **Declare interest**
  - **Review**
  - **Decide**
  - **Form program**
  - **Notify**
  - **Revise**



CMT++: eJournal Management Tool

- **Add publishing steps**
  - **Form committee**
  - **Accept Manuscripts**
  - **Declare interest**
  - **Review**
  - **Decide**
  - **Form program**
  - **Notify**
  - **Revise**
  - **Publish**

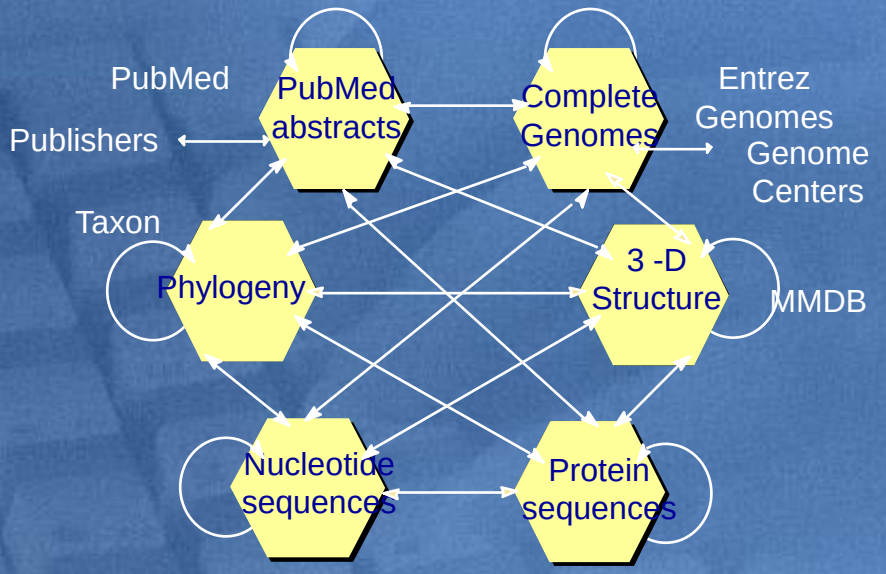
- Connect to Archives
- Manage archive document versions
- Capture Workshop
  - presentations
  - proceedings
- Capture classroom ConferenceXP
- Moderated discussions of published articles

The Three Prophets of Open Access

- Paul Ginsparg's arXiv at Cornell has demonstrated a new model of scientific publishing
  - Pioneered electronic version of 'preprints' hosted on the Web now used routinely by the physics community
- David Lipman of the NIH National Library of Medicine has developed PubMedCentral as repository for NIH funded research papers
  - Microsoft funded development of 'portable PMC' now being deployed in UK and other countries
- Stevan Harnad's 'self-archiving' EPrints project in Southampton provides a basis for OAI-compliant 'Institutional Repositories'
  - JISC-funded TARDis Project at Southampton is hybrid of full-text open access and links to publisher sites

The NLM Example: Entrez-GenBank

- Sequence data deposited with Genbank
- Literature references Genbank ID
- BLAST searches Genbank
- Entrez integrates and searches
  - PubMedCentral
  - PubChem
  - Genbank
  - Proteins, SNP,
  - Structure,..
  - Taxononomy…



Portable PubMedCentral

- “Information at your fingertips”
- Helping build PortablePubMedCentral
- Deployed US, China, England, Italy, South Africa, (Japan soon).
- Each site can accept documents
- Archives replicated
- Federate thru web services
- Working to integrate Word/Excel/… with PubmedCentral
- To be clear: NCBI is doing 99% of the work.

Routes to Open Access

Stevan Harnad identifies 2 roads to OA:

(1) OA Journal publishing – 'Gold'

- "author pays" rather than present subscription model
- E.g. PLoS journals

(2) Self-Archiving in Repository – 'Green'

- Author provides OA by putting e-print of paper submitted to journal in repository or on own web site
- 94% of journals are 'Green' and permit self-archiving

Key results from TARDis project in UK FAIR programme

- ‘Hybrid’ research publications database building up to represent full range of types of research in all disciplines across the institution
- Embed in research recording process with institutional commitment
- Add more full text as climate improves/authors become familiar with practice
- Library checks metadata, adds DOI or other link to publisher version

- Provided feedback to EPrints software to give good citation format: providing tools for recording once – many outputs eg export to research group web pages

- Hey, Jessie M.N., Simpson, Pauline and Carr, Leslie A. (2005) The TARDis Route Map to Open Access: developing an Institutional Repository Model. In, Dobreva, Milena and Engelen, Jan (eds.) ELPUB2005 From Author to Reader: Challenges for the Digital Content Chain: Proceedings of the 9th ICCC International Conference on Electronic Publishing, Katholieke Universiteit Leuven, Leuven-Heverlee, Belgium, 8-10 June 2005. Leuven, Belgium, Peeters Publishing, 179-182.

  http://eprints.soton.ac.uk/16262/
- Simpson, Pauline and Hey, Jessie (2006) Repositories for research: Southampton’s evolving role in the knowledge cycle. Program, 40, (3), 224-231.

  http://eprints.soton.ac.uk/41240/

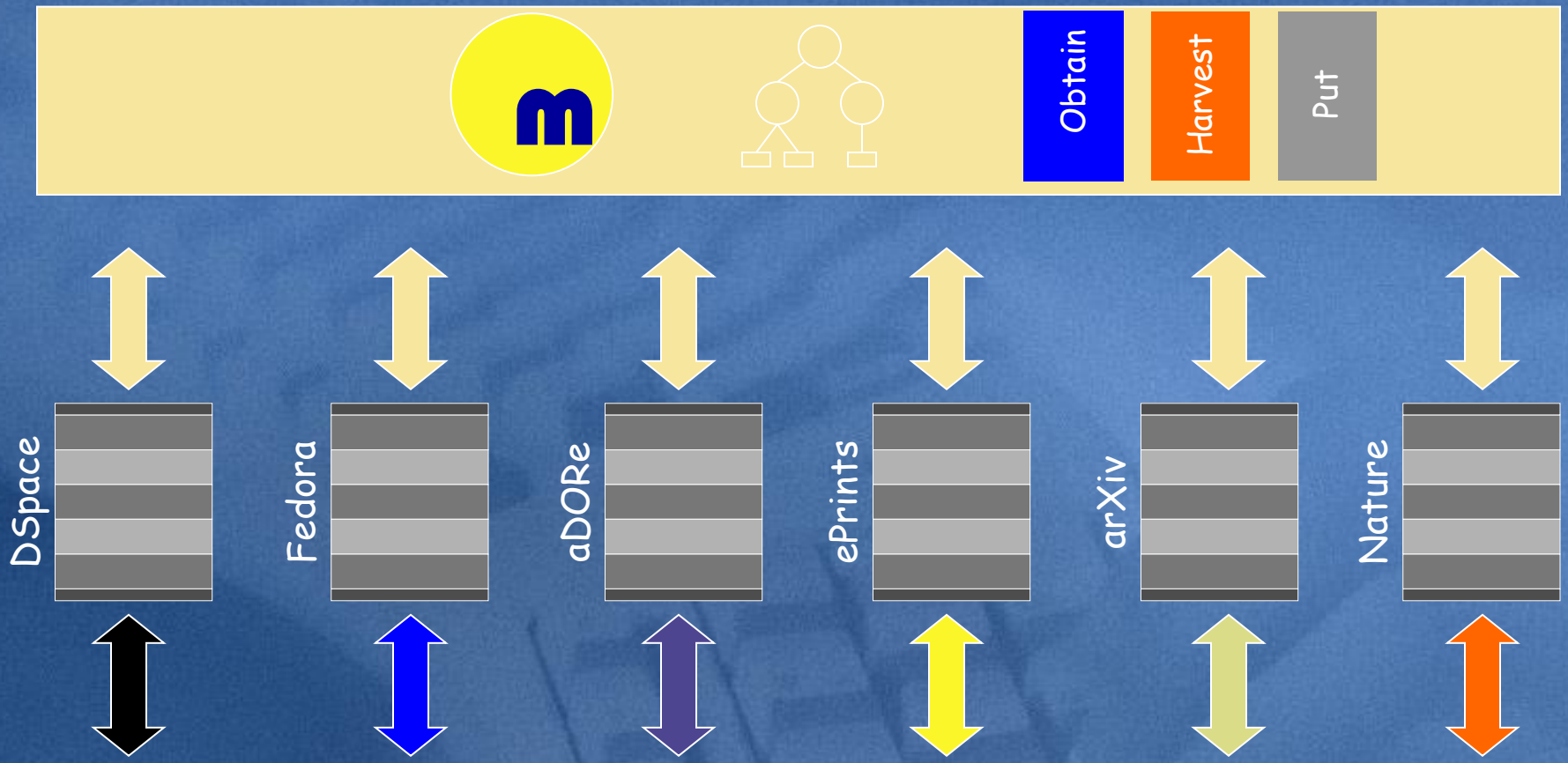
http://tardis.eprints.org/

TARDis

OA and Institutional Repositories

- Registry of OA Repositories records:
  - 213 archives using EPrints software
  - 174 archives using DSpace software
- OAIster records:
  - ~10M records from ~700 institutions
- Sources of information about ‘Green Route’ to OA
  - www.jisc.ac.uk/publications
  - www.eprints.org
  - www.openarchives.org
  - oaister.umdl.umich.edu/o/oaister
  - www.OpenDOAR.org

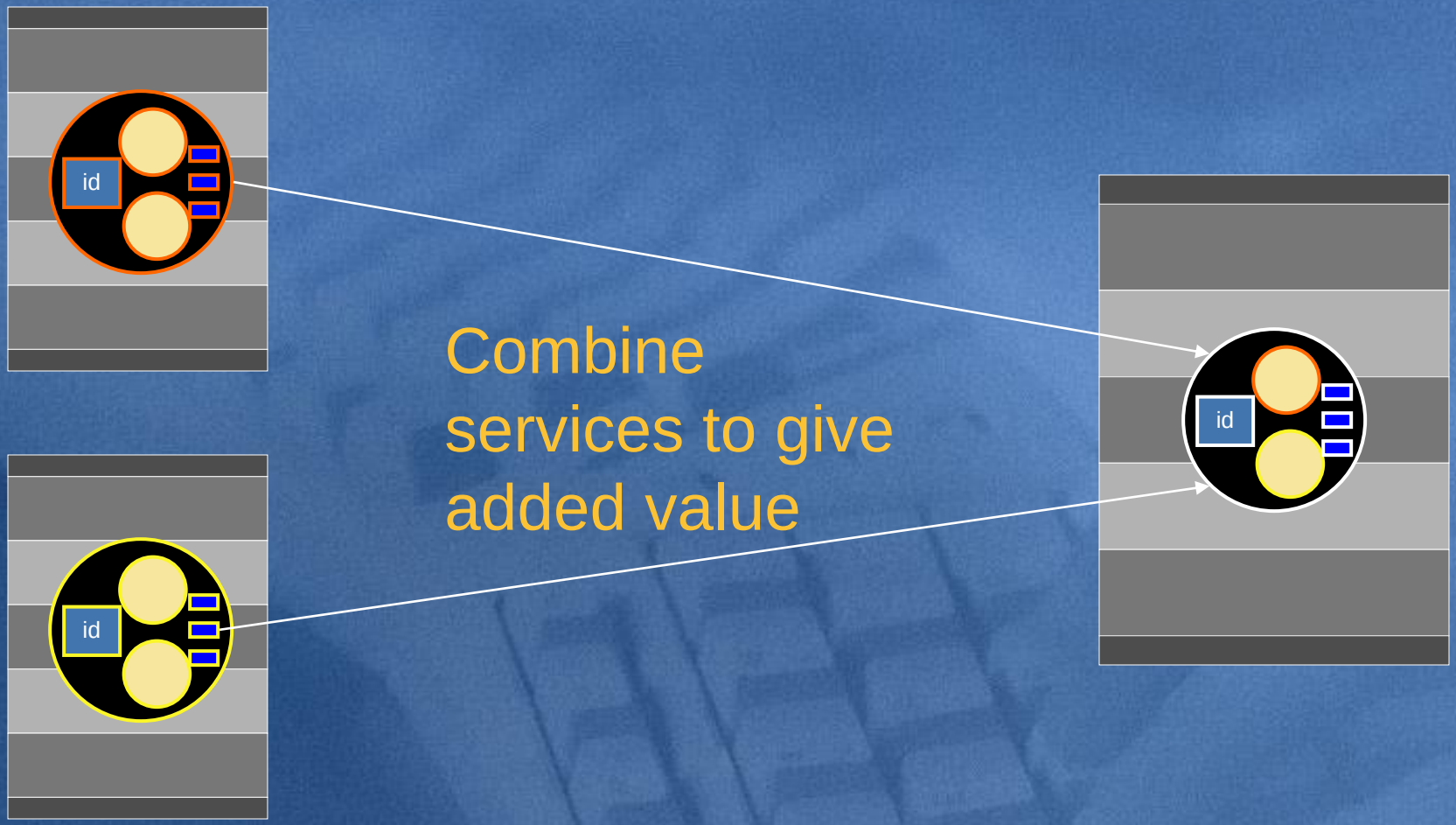
Augmenting interoperability



The Service Revolution

- Web 2.0
  - Social networks, tagging for sharing e.g. e.g. Flikr, Del.icio.us, MySpace, CiteULike, Connotea …
  - Wikis, Blogs, RSS, folksonomies …
- Software delivered as a service
  - Microsoft Live services
    - Office Live
    - Xbox Live
    - Windows Live Academic
  - Mashups
    - SensorWeb + VirtualEarth
    - http://mashupcamp.com

e-Science Mashups?



Combine services to give added value

‘As We May Think’ Vannevar Bush, 1945

- Still grappling with the data preservation issues he raised:
  - “A record if it is to be useful to science, must be continuously extended, it must be stored, and above all it must be consulted.”
- Can now realize his idea of the ‘memex’
  - “a future device for individual use, which is a sort of mechanized private file and library”
  - Search by following ‘trails’ through data
- Now Paul Ginsparg’s ‘As We May Read’ …

Interoperability

The right approach for the right situation

Uniformity

- Early De Jure Standards
- Works well for the physical world

Translatability

- De Facto Standards

Danke

Grazie

Thank you

Obrigado

Gracias

Merci

Спасибо

Microsoft Office Open XML Formats (OOXML)

- Documents in Office 2007 will be based on new XML-based file formats
  - Open, royalty-free file format specification will allow interoperability
- OOXML submitted to ECMA International Standards Organization
  - Microsoft also offering ‘Covenant Not to Sue’
- OpenXML Translator Project
  - Microsoft backing open source project to create translation tool between OOXML and Open Document Format ODF

Technical Computing at Microsoft

- Advanced Computing for Science and Engineering
  - Application of new algorithms, tools and technologies to scientific and engineering problems
- High Performance Computing
  - Application of high performance clusters and database technologies to industrial and scientific applications
- Radical Computing
  - Research in potential breakthrough technologies

Summary

Microsoft wishes to work with the university research and library communities to:

- develop interoperable high-level services, work flows, tools and data services
- accelerate progress in a small number of societally important scientific applications
- assist in the development of interoperable repositories and new models of scholarly publishing
- explore radical new directions in computing and ways and applications to exploit on-chip parallelism

➢ How can Microsoft best collaborate with the scientific community?

Microsoft®

Your potential. Our passion.™