# e-Science and Cyberinfrastructure

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

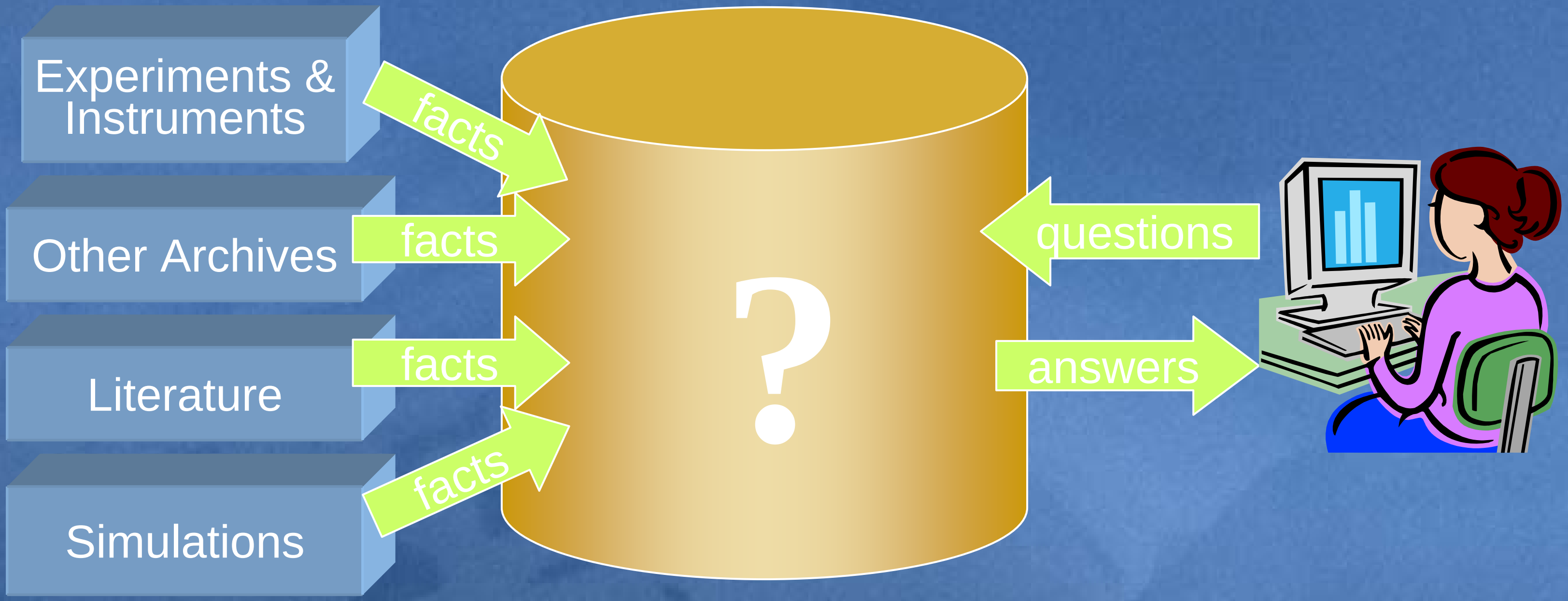
Former Director General of Research Councils

Office of Science and Technology, UK

# 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

# 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?

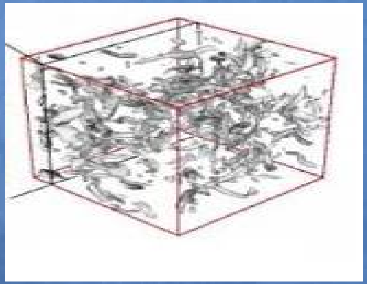
- Data Query and Visualization tools
- Support/training
- Performance
  - Execute queries in a minute
  - Batch (big) query scheduling

# 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 …
- **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

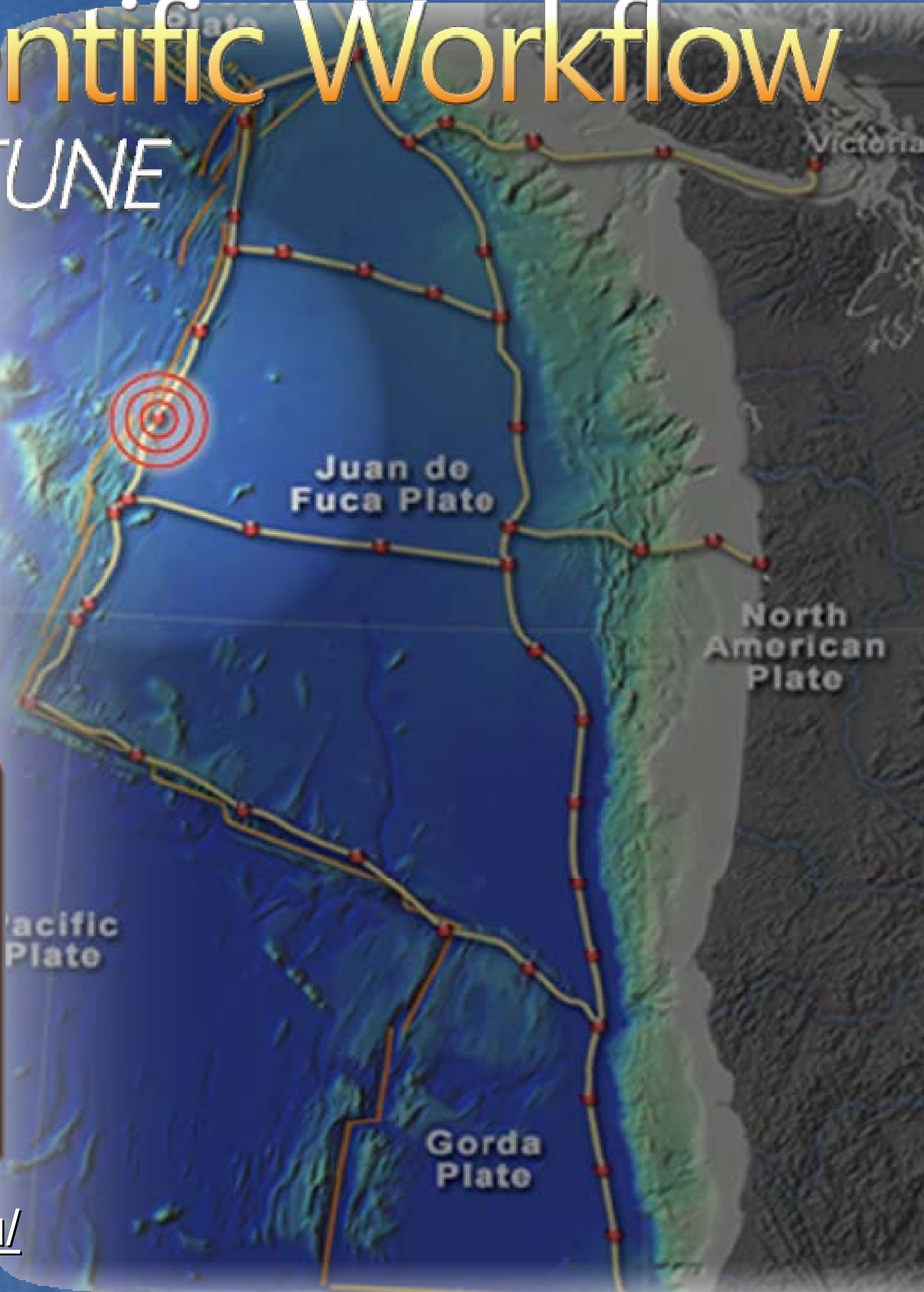
(With thanks to Jim Gray)

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



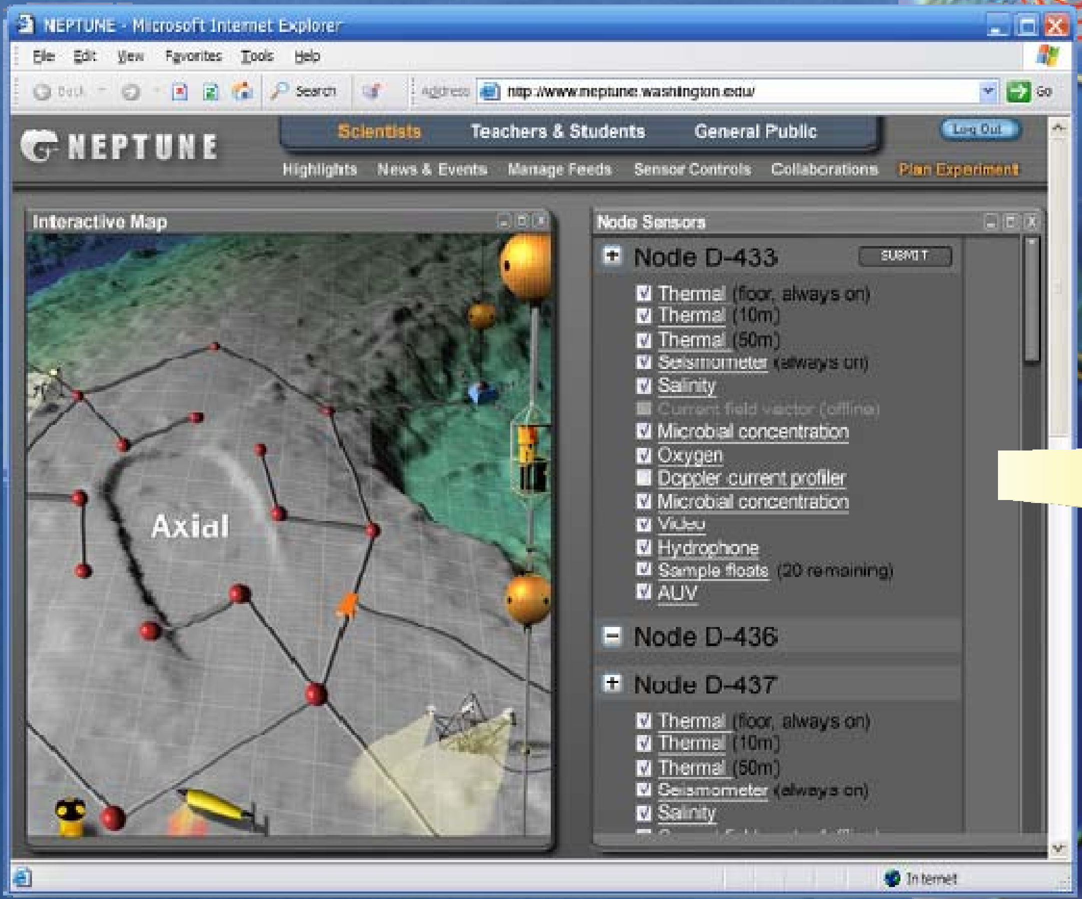
# Vision For Scientific Workflow

## *Example: Project NEPTUNE*



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

# Programmable Sensors & Remote Instruments



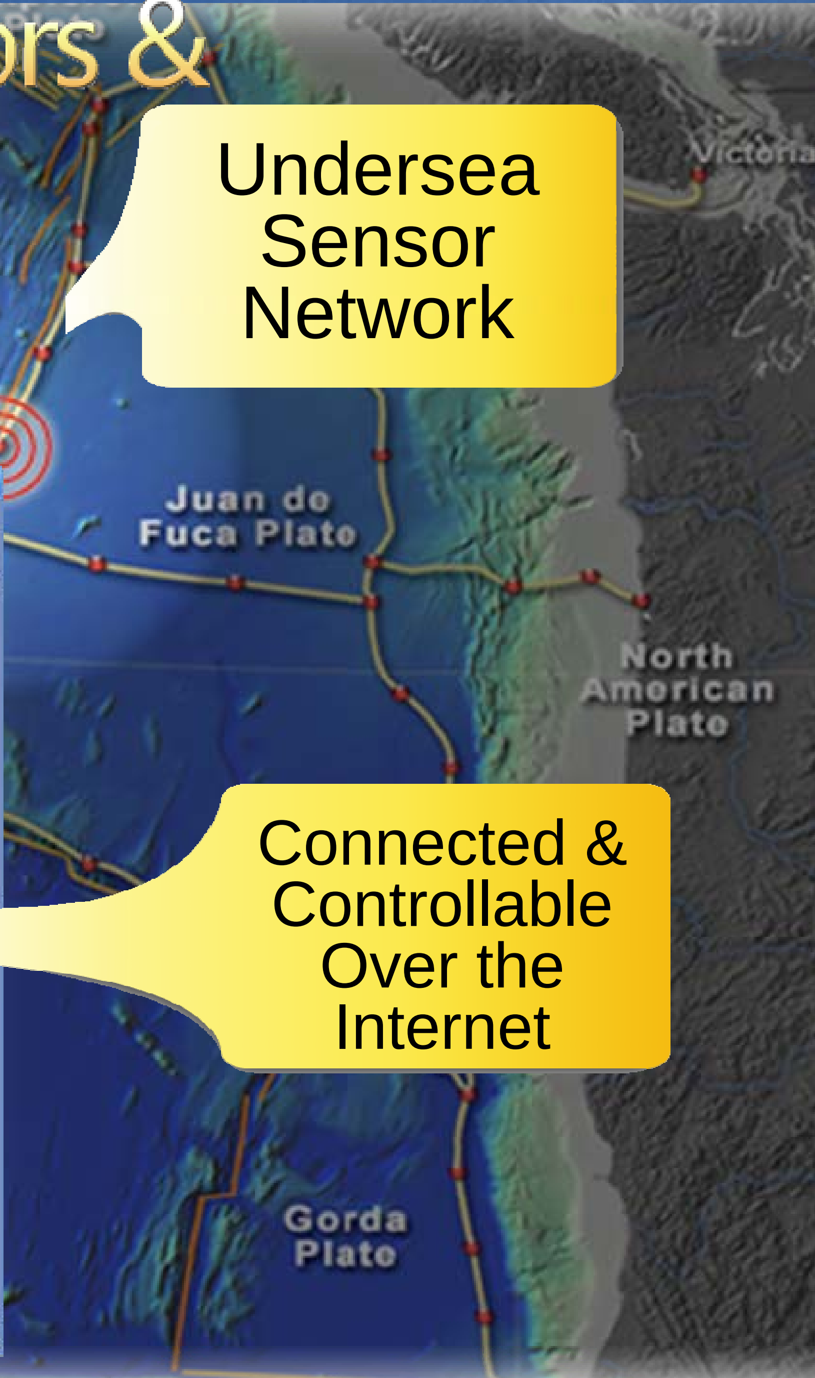
Node D-433

- ☑ Thermal (floor, always on)
- ☑ Thermal (10m)
- ☑ Thermal (50m)
- ☑ Seismometer (always on)
- ☑ Salinity
- ☐ Current field vector (offline)
- ☑ Microbial concentration
- ☑ Oxygen
- ☐ Doppler current profiler
- ☑ Microbial concentration
- ☑ Video
- ☑ Hydrophone
- ☑ Sample floats (20 remaining)
- ☑ AUV

Node D-436

Node D-437

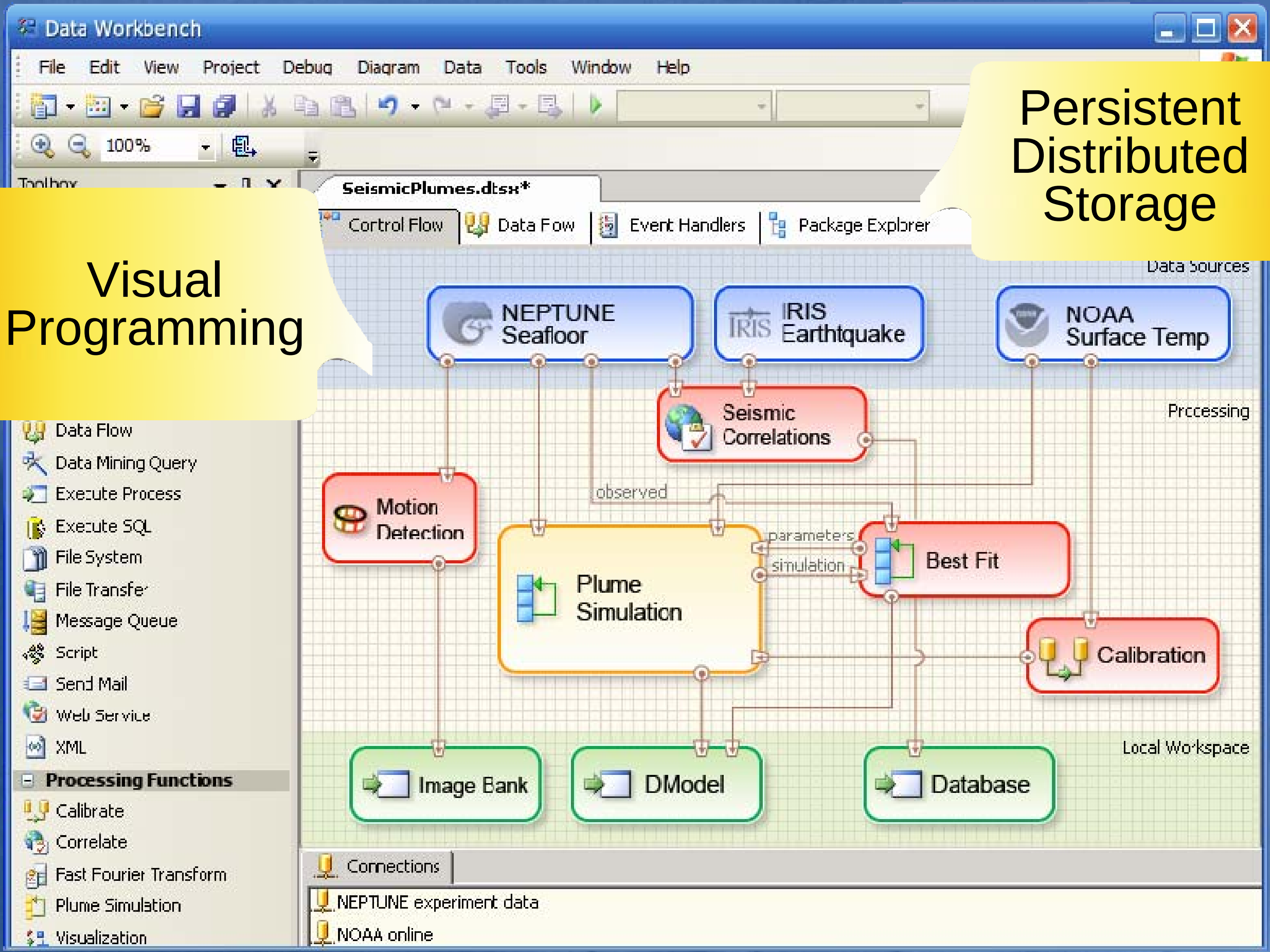
- ☑ Thermal (floor, always on)
- ☑ Thermal (10m)
- ☑ Thermal (50m)
- ☑ Seismometer (always on)
- ☑ Salinity



Undersea Sensor Network

Connected & Controllable Over the Internet

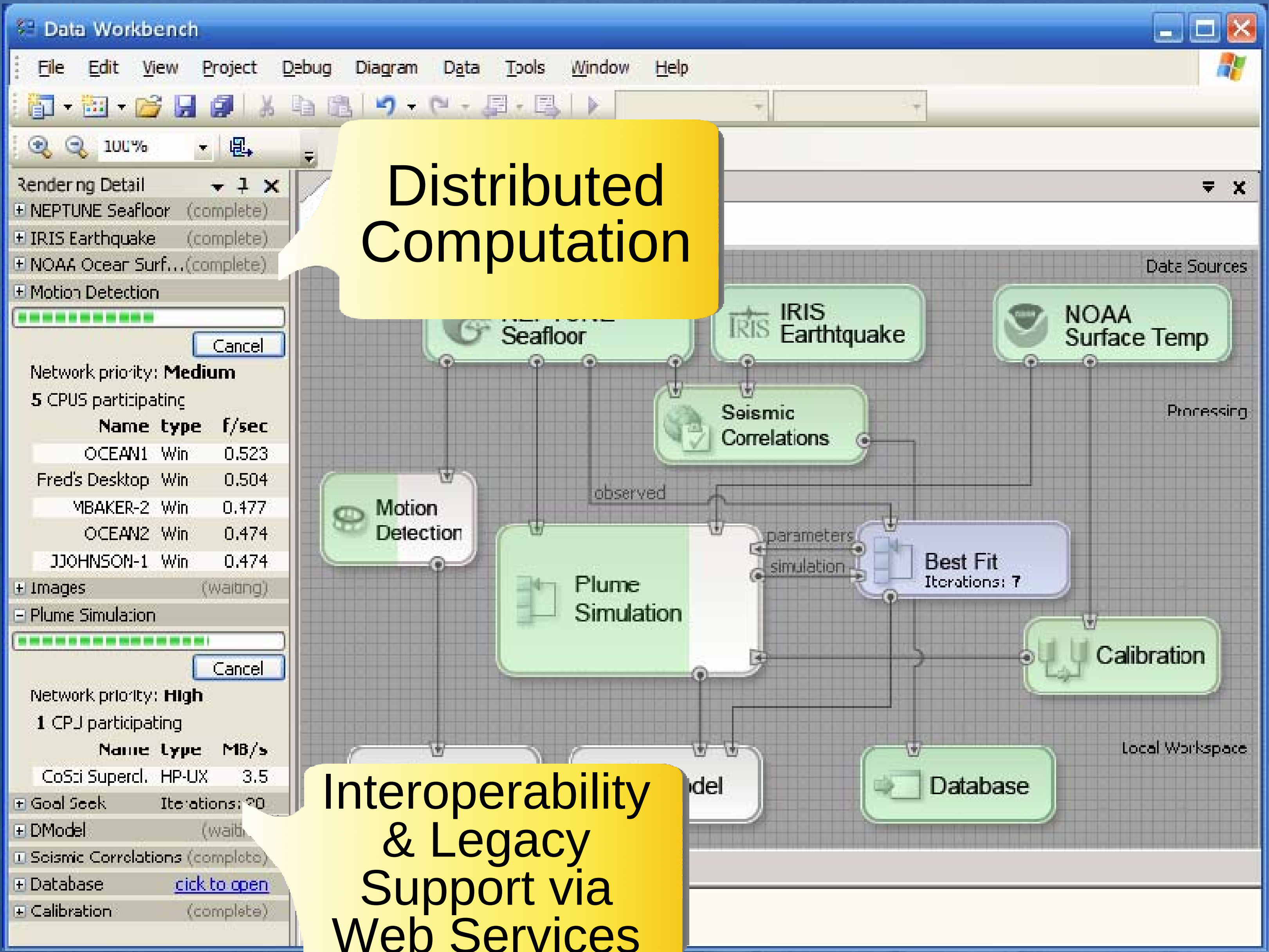
# Data Workbench



Visual Programming

Persistent Distributed Storage

# Data Workbench



Rendering Detail

⊞ NEPTUNE Seafloor (complete)

⊞ IRIS Earthquake (complete)

⊞ NOAA Ocean Surf... (complete)

⊞ Motion Detection

Cancel

Network priority: **Medium**

**5** CPUS participating

| Name | type | f/sec |
|---|---|---|
| OCEAN1 | Win | 0.523 |
| Fred's Desktop | Win | 0.504 |
| MBAKER-2 | Win | 0.477 |
| OCEAN2 | Win | 0.474 |
| JJOHNSON-1 | Win | 0.474 |

⊞ Images (waiting)

⊟ Plume Simulation

Cancel

Network priority: **High**

**1** CPU participating

| Name | type | MB/s |
|---|---|---|
| CoSci Supercl. | HP-UX | 3.5 |

⊞ Goal Seek Iterations: 20

⊞ DModel (waiti

⊞ Seismic Correlations (complete)

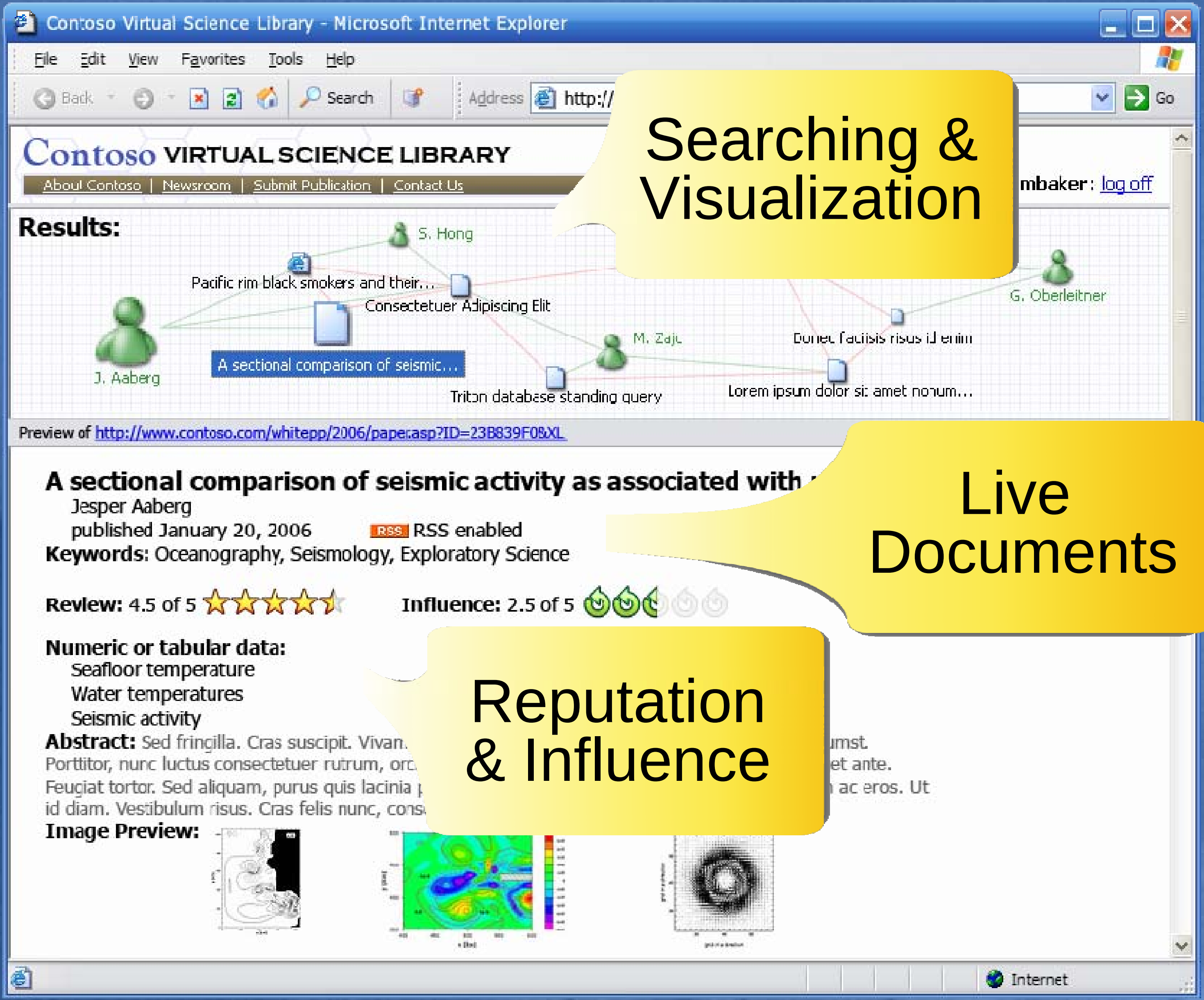
⊞ Database click to open

⊞ Calibration (complete)

Distributed Computation

Interoperability & Legacy Support via Web Services

# Research


Contoso Virtual Science Library - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Address
http://
Go
S. Hong
G. Oberleitner
M. Zaju
J. Aaberg
Pacific rim black smokers and their...
Consectetuer Adipiscing Elit
A sectional comparison of seismic...
Donec facilisis risus id enim
Triton database standing query
Lorem ipsum dolor sit amet nonum...
Internet

Contoso VIRTUAL SCIENCE LIBRARY

About Contoso | Newsroom | Submit Publication | Contact Us

mbaker: log off

**Results:**

Preview of http://www.contoso.com/whitepp/2006/paper.asp?ID=23B839F0&XL

**A sectional comparison of seismic activity as associated with**

Jesper Aaberg

published January 20, 2006 RSS RSS enabled

**Keywords:** Oceanography, Seismology, Exploratory Science

**Review:** 4.5 of 5 ★★★★☆ **Influence:** 2.5 of 5

**Numeric or tabular data:**

Seafloor temperature

Water temperatures

Seismic activity

**Abstract:** Sed fringilla. Cras suscipit. Vivan umst. Porttitor, nunc luctus consectetuer rutrum, orc et ante. Feugiat tortor. Sed aliquam, purus quis lacinia ac eros. Ut id diam. Vestibulum risus. Cras felis nunc, cons

**Image Preview:**

Searching & Visualization

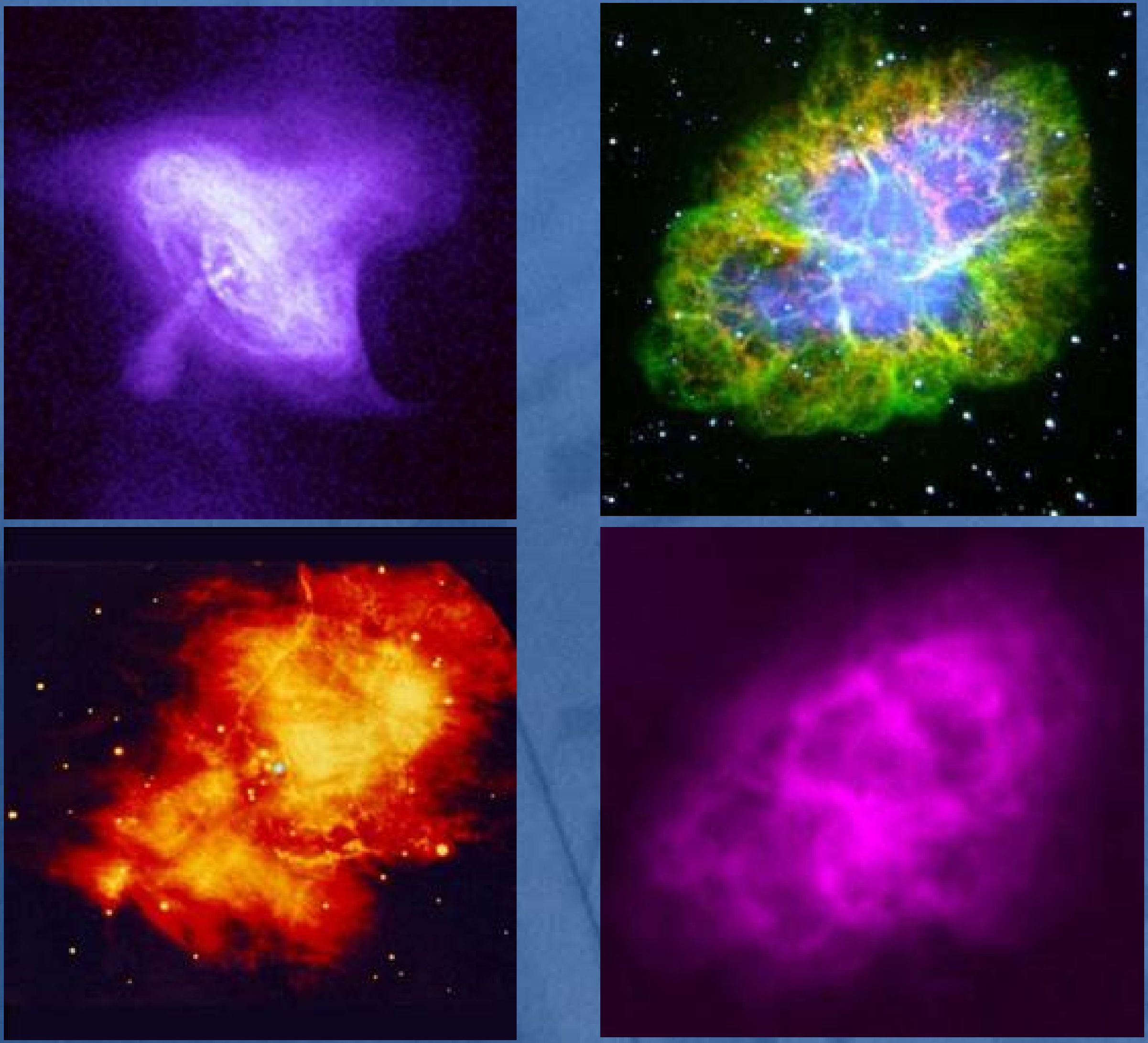
Live Documents

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

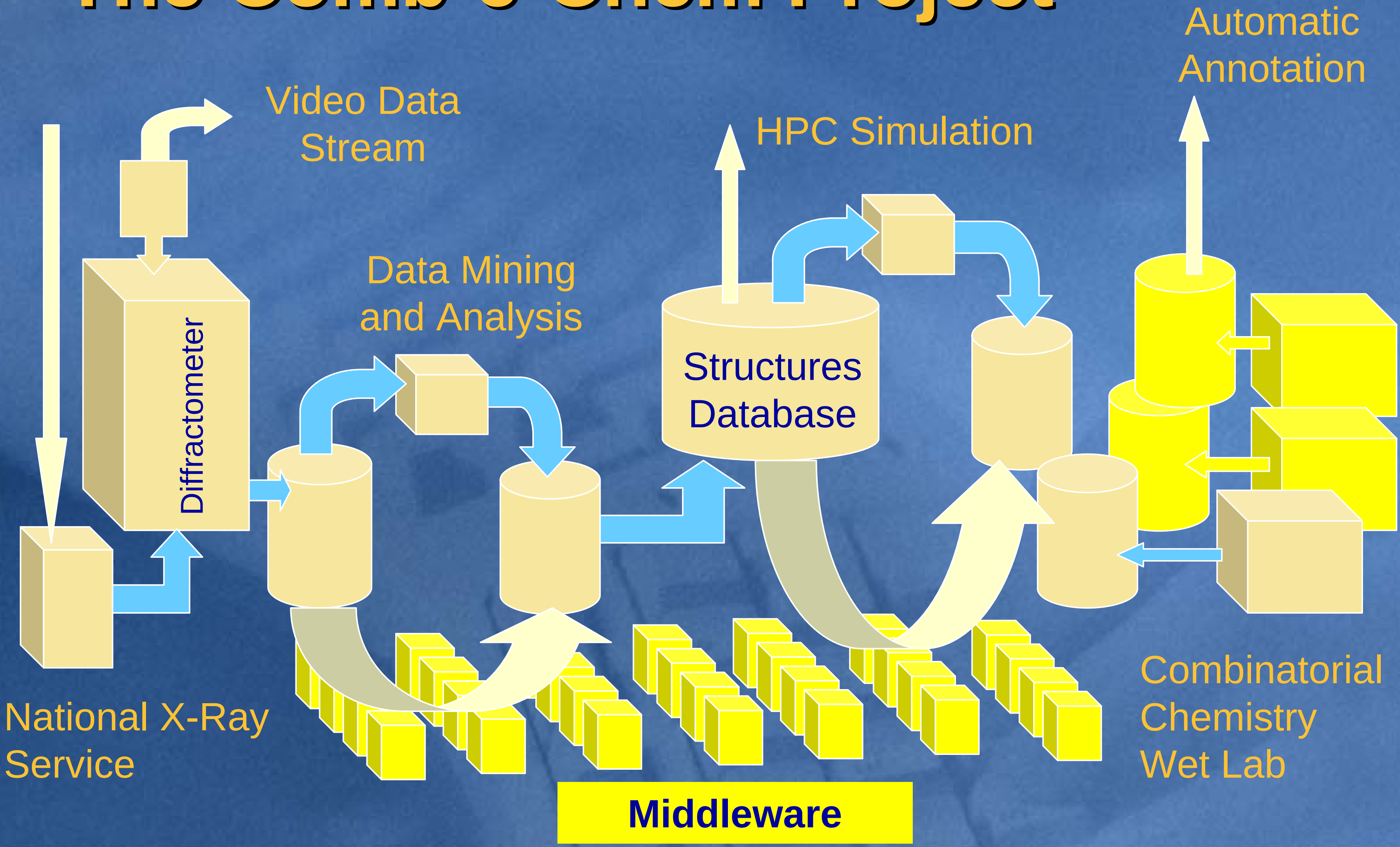
X-ray,
optical,
infrared, and
radio
views of the nearby
Crab Nebula, which is
now in a state of chaotic
expansion after a
supernova explosion
first sighted in 1054
A.D. by Chinese
Astronomers.

Slide courtesy of Robert Brunner @ CalTech.

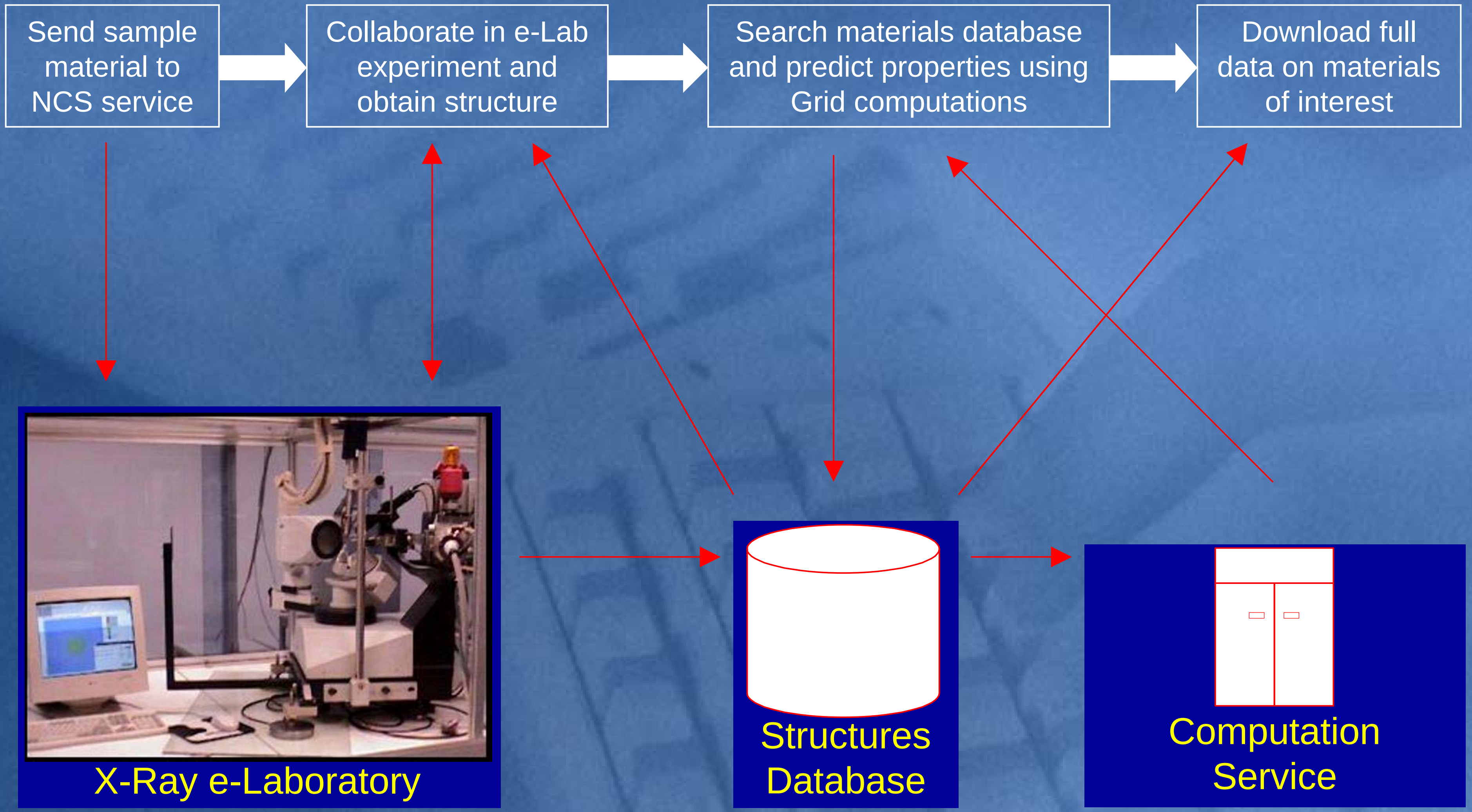
# IVO: An Astronomy Data Grid

- 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

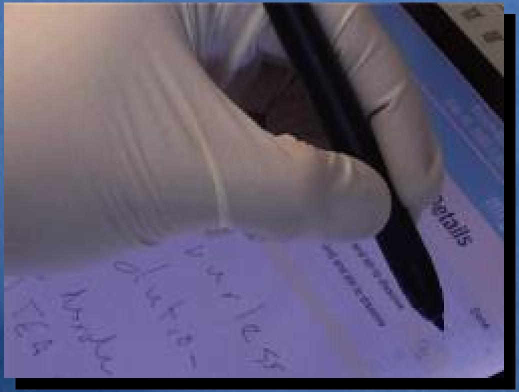
Search materials database and predict properties using Grid computations

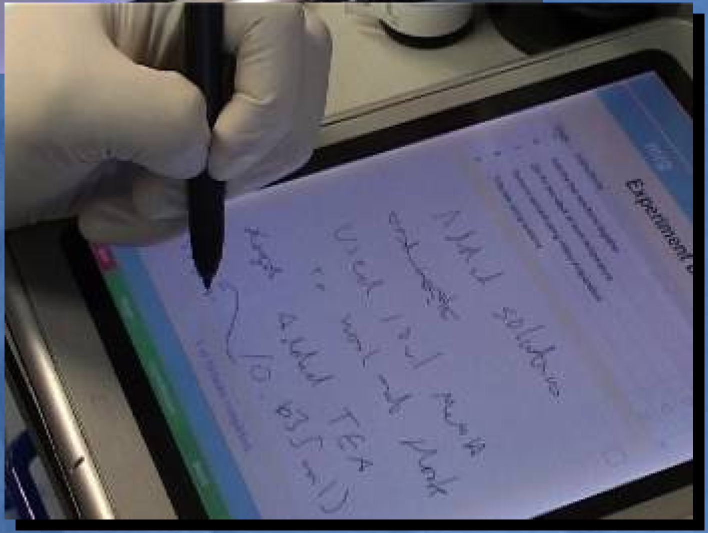
Download full data on materials of interest

X-Ray e-Laboratory

Structures Database

Computation Service



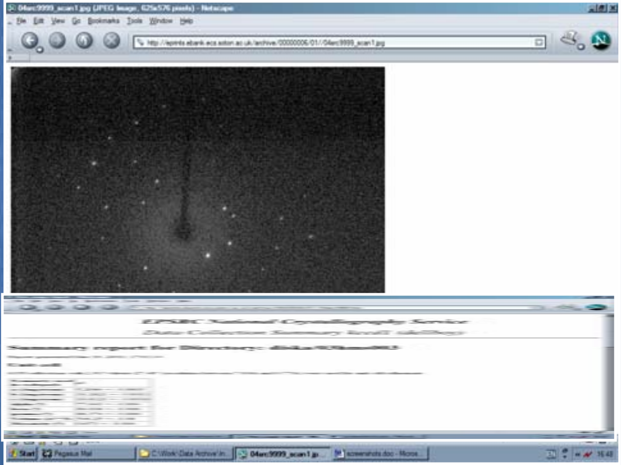


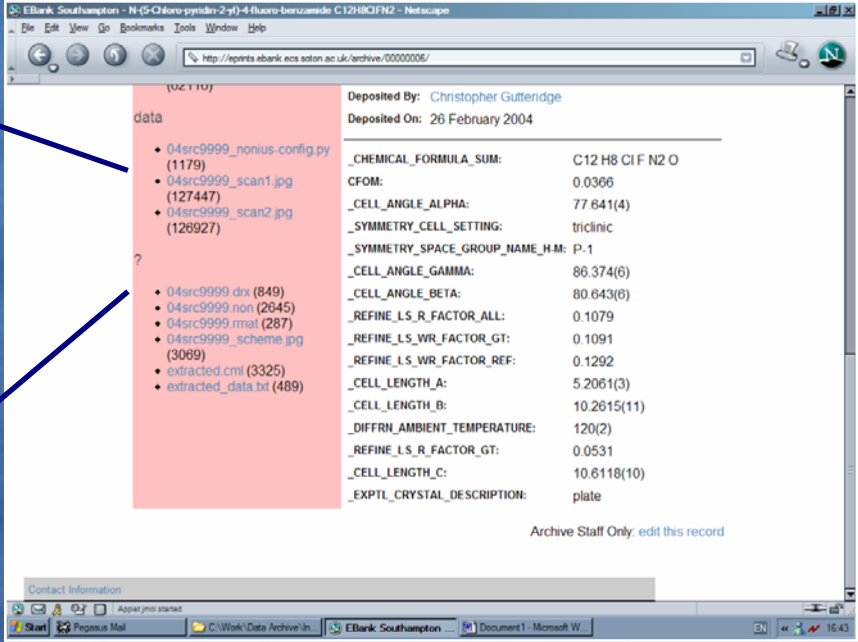
A digital lab book replacement that chemists were able to use, and liked

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

# Crystallographic e-Prints

## Direct Access to Raw Data from scientific papers


04src9999_scan1.jpg (JPEG Image, 625x576 pixels) - Netscape
File Edit View Go Bookmarks Tools Window Help
http://eprints.ebank.ecs.soton.ac.uk/archive/00000006/01/04src9999_scan1.jpg
Start
Pegasus Mail
C:\Work\Data Archive\n...
04src9999_scan1.jp...
screenshots.doc - Micros...
16:48


EBank Southampton - N-(5-Chloro-pyridin-2-yl)-4-fluoro-benzamide C12H8ClFN2 - Netscape
File Edit View Go Bookmarks Tools Window Help
http://eprints.ebank.ecs.soton.ac.uk/archive/00000006/
Applet jmol started
Start
Pegasus Mail
C:\Work\Data Archive\n...
EBank Southampton ...
Document1 - Microsoft W...
16:43

data

- 04src9999_nonius-config.py (1179)
- 04src9999_scan1.jpg (127447)
- 04src9999_scan2.jpg (126927)

?

- 04src9999.drx (849)
- 04src9999.non (2645)
- 04src9999.rmat (287)
- 04src9999_scheme.jpg (3069)
- extracted.cml (3325)
- extracted_data.txt (489)

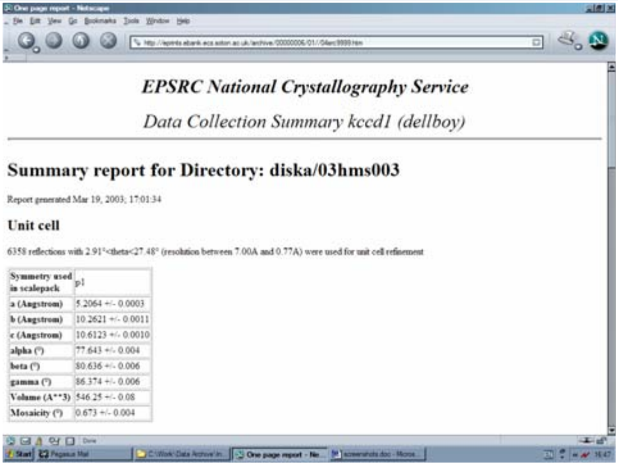
**Deposited By:** Christopher Gutteridge

**Deposited On:** 26 February 2004

| | |
|---|---|
| _CHEMICAL_FORMULA_SUM: | C12 H8 Cl F N2 O |
| CFOM: | 0.0366 |
| _CELL_ANGLE_ALPHA: | 77.641(4) |
| _SYMMETRY_CELL_SETTING: | triclinic |
| _SYMMETRY_SPACE_GROUP_NAME_H-M: | P-1 |
| _CELL_ANGLE_GAMMA: | 86.374(6) |
| _CELL_ANGLE_BETA: | 80.643(6) |
| _REFINE_LS_R_FACTOR_ALL: | 0.1079 |
| _REFINE_LS_WR_FACTOR_GT: | 0.1091 |
| _REFINE_LS_WR_FACTOR_REF: | 0.1292 |
| _CELL_LENGTH_A: | 5.2061(3) |
| _CELL_LENGTH_B: | 10.2615(11) |
| _DIFFRN_AMBIENT_TEMPERATURE: | 120(2) |
| _REFINE_LS_R_FACTOR_GT: | 0.0531 |
| _CELL_LENGTH_C: | 10.6118(10) |
| _EXPTL_CRYSTAL_DESCRIPTION: | plate |

Archive Staff Only: edit this record

Contact Information


One page report - Netscape
File Edit View Go Bookmarks Tools Window Help
http://eprints.ebank.ecs.soton.ac.uk/archive/00000006/01/04src9999.htm
Done
Start
Pegasus Mail
C:\Work\Data Archive\n...
One page report - Ne...
screenshots.doc - Micros...
16:47

*EPSRC National Crystallography Service*

*Data Collection Summary kccd1 (dellboy)*

**Summary report for Directory: diska/03hms003**

Report generated Mar 19, 2003; 17:01:34

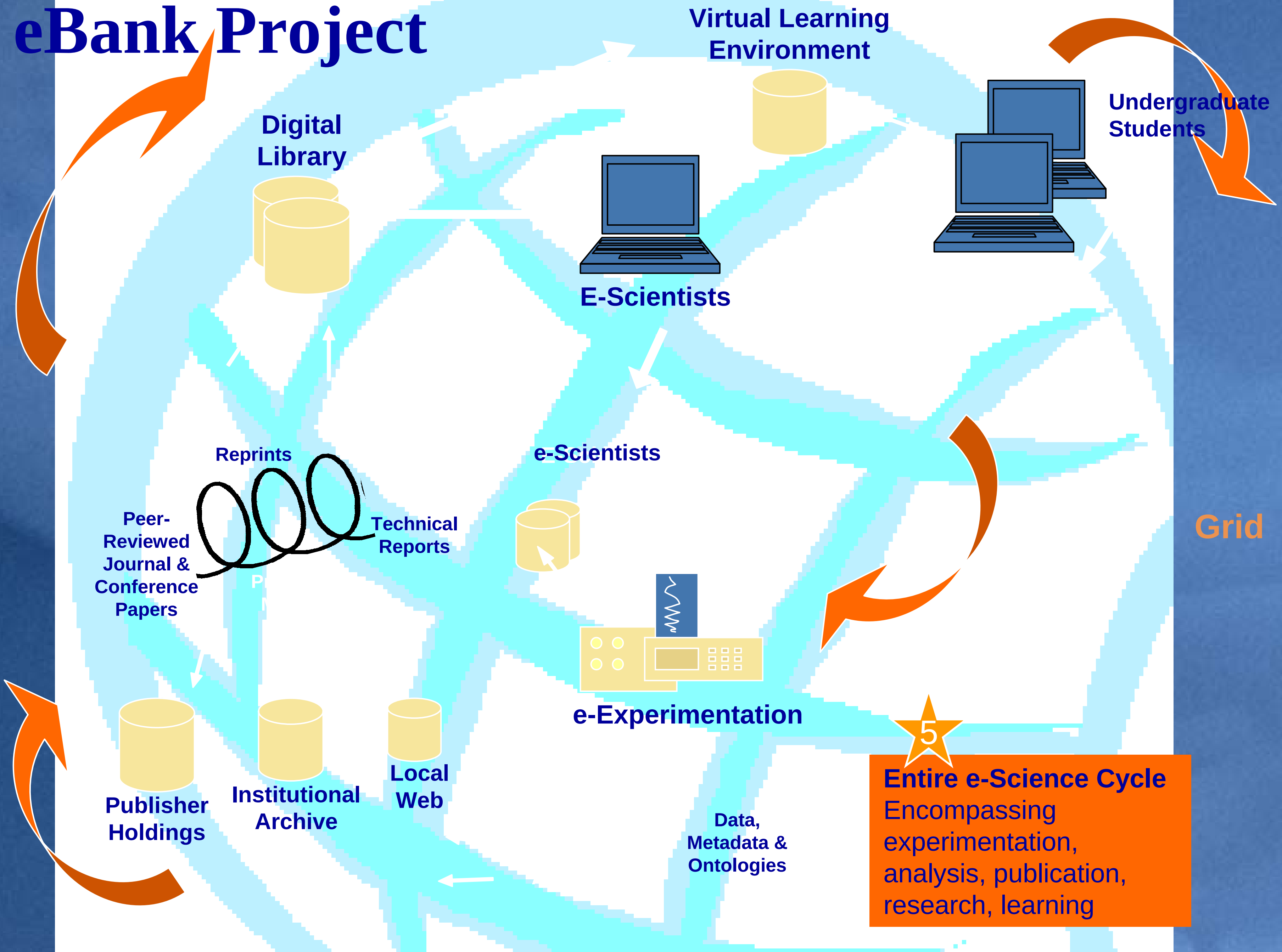
**Unit cell**

6358 reflections with 2.91°<theta<27.48° (resolution between 7.00A and 0.77A) were used for unit cell refinement

| | |
|---|---|
| **Symmetry used in scalepack** | p1 |
| **a (Angstrom)** | 5.2064 +/- 0.0003 |
| **b (Angstrom)** | 10.2621 +/- 0.0011 |
| **c (Angstrom)** | 10.6123 +/- 0.0010 |
| **alpha (°)** | 77.643 +/- 0.004 |
| **beta (°)** | 80.636 +/- 0.006 |
| **gamma (°)** | 86.374 +/- 0.006 |
| **Volume (A**3)** | 546.25 +/- 0.08 |
| **Mosaicity (°)** | 0.673 +/- 0.004 |

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

# eBank Project

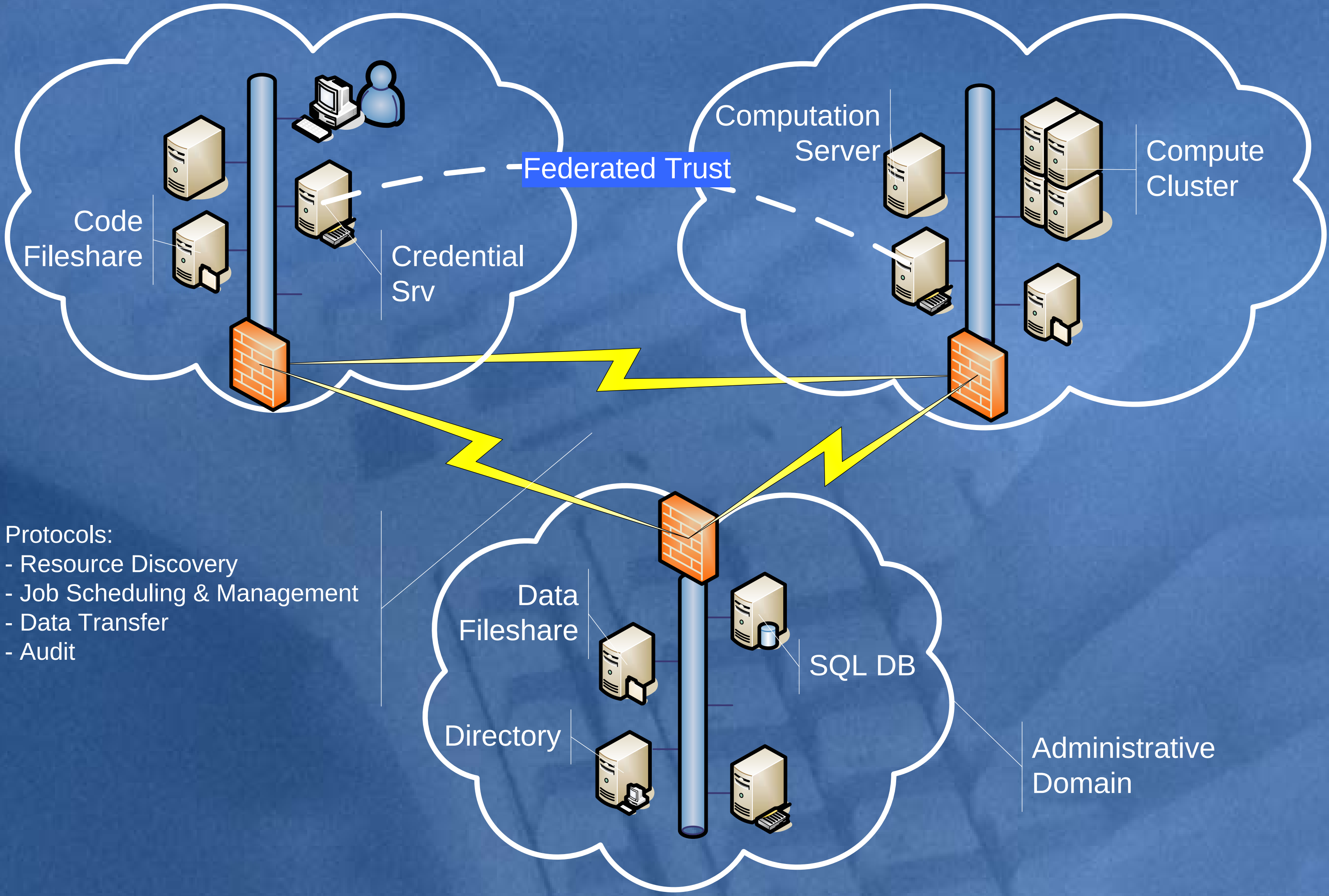


**Entire e-Science Cycle**
Encompassing experimentation, analysis, publication, research, learning

# Cyberinfrastructure

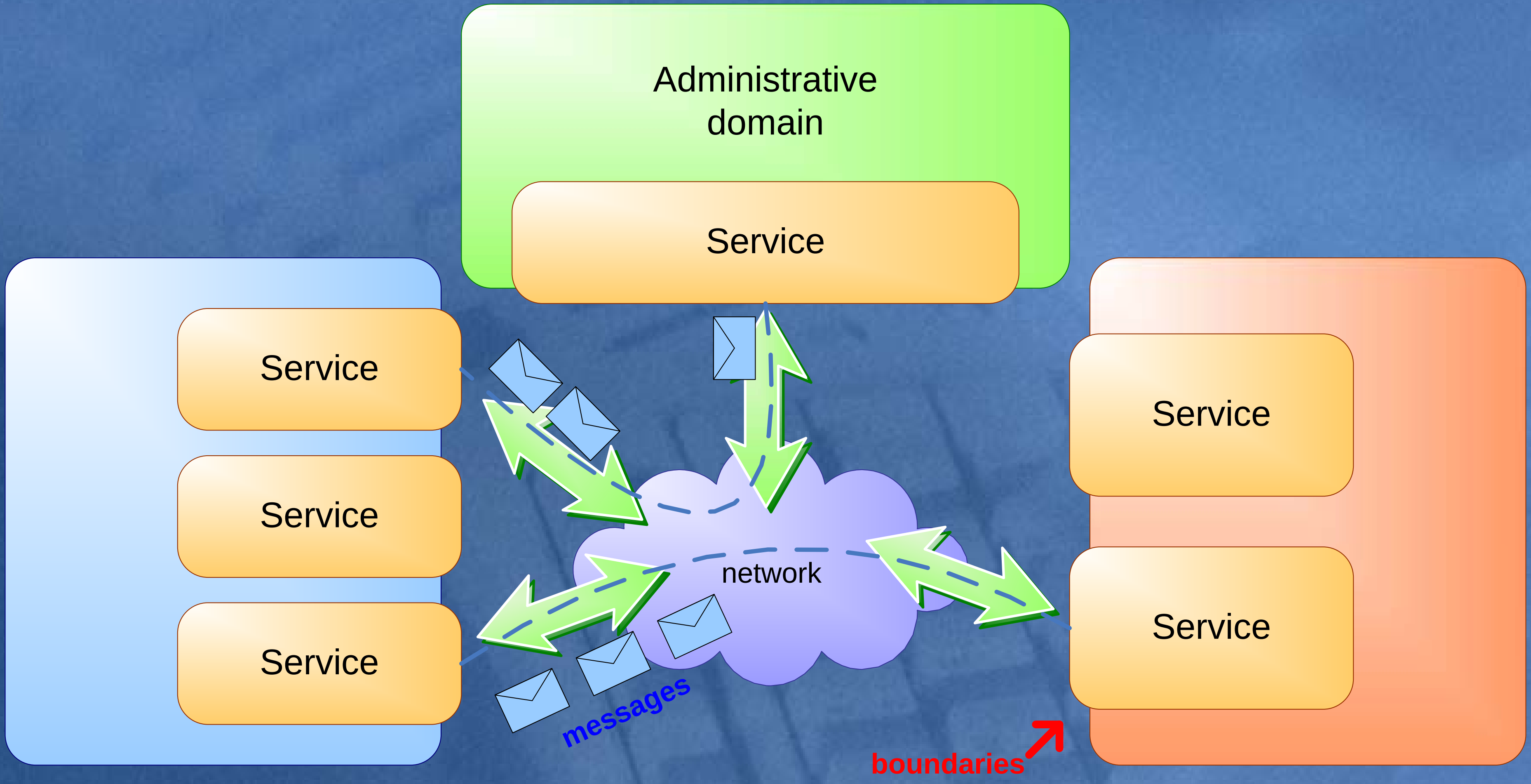
- In the US, Europe and Asia there is a common vision for the ‘cyberinfrastructure’ required to support the e-Science revolution
- Set of Grid Middleware Services supported on top of high bandwidth academic research networks
- Opportunity for Computer Science community to provide scientists with powerful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

# Grids for Virtual Organizations

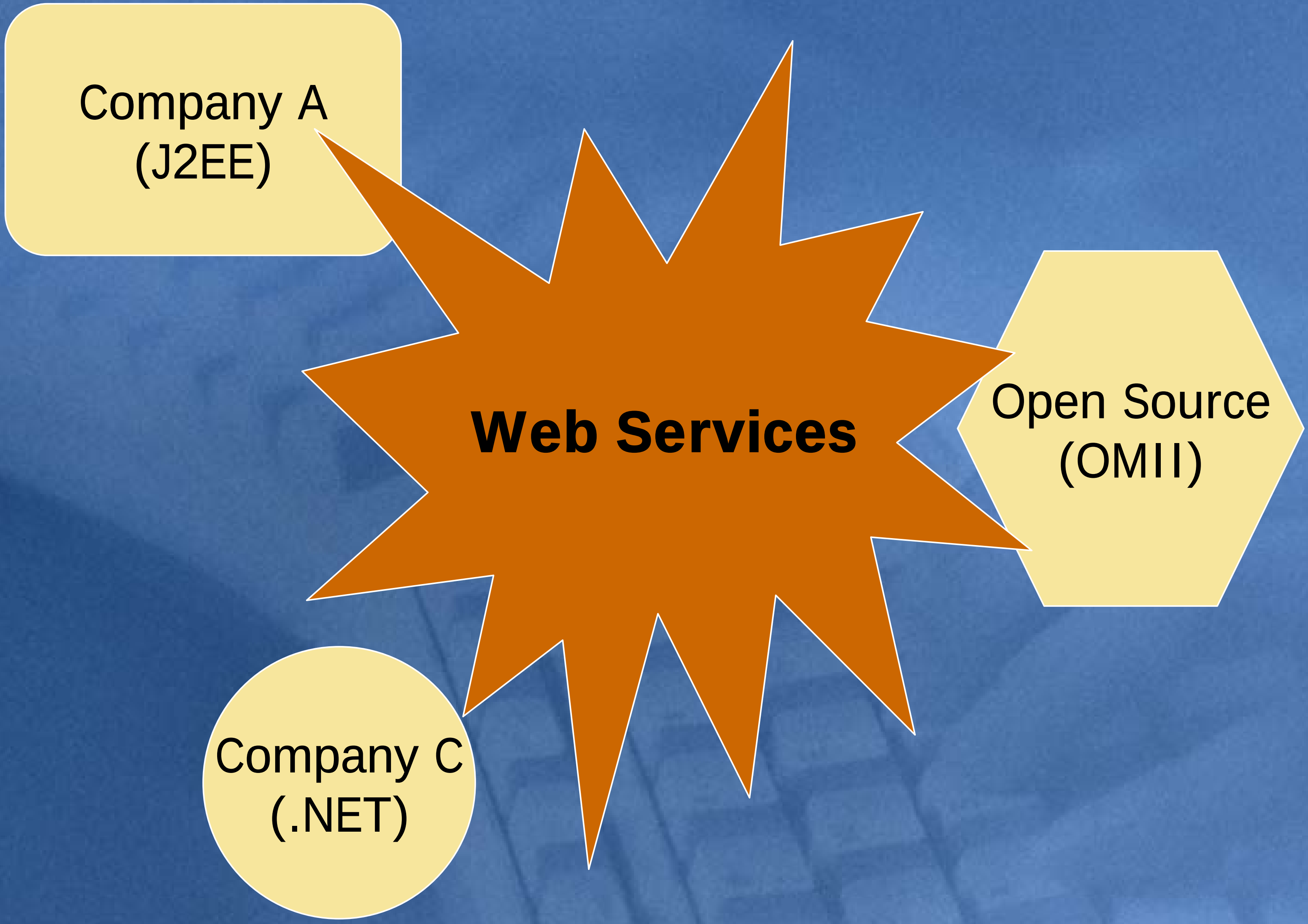


Protocols:
- Resource Discovery
- Job Scheduling & Management
- Data Transfer
- Audit

# Service-Orientation for building Distributed Systems



# Web Services and Interoperability



Company A
(J2EE)

Web Services

Open Source
(OMII)

Company C
(.NET)

# 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

# The e-Science Data Life Cycle

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

# 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

# OECD Declaration on Access to Research Data from Public Funding

- 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’

# Open Access and Scholarly Publishing

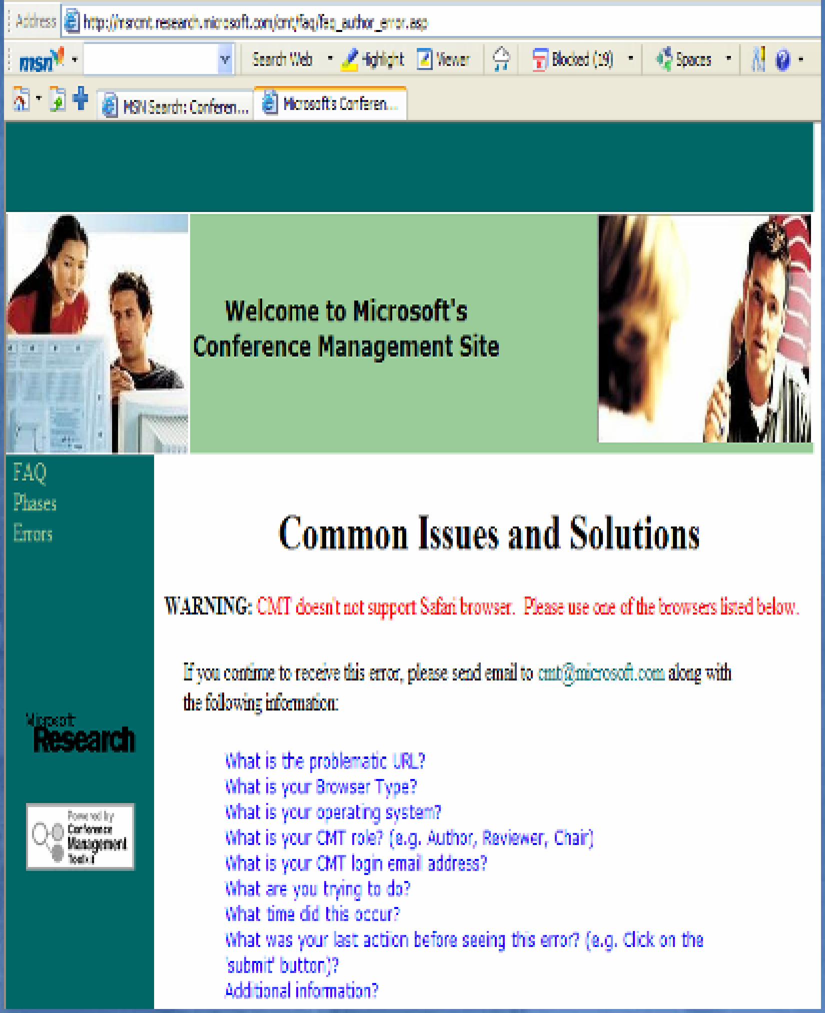
- Goal is to work with the research community to assist them in developing open and interoperable frameworks for scholarly publishing
- Two aspects
  - ‘Community publishing’ toolset
  - Service Oriented Framework for Interoperable Repositories

# Community Publishing

- Develop toolset for ‘self-publishing’ of workshop and conference proceedings based on MSR Conference Management Tool
  - 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



Welcome to Microsoft's Conference Management Site

## Common Issues and Solutions

WARNING: CMT doesn't not support Safari browser. Please use one of the browsers listed below.

If you continue to receive this error, please send email to cmt@microsoft.com along with the following information:

What is the problematic URL?
What is your Browser Type?
What is your operating system?
What is your CMT role? (e.g. Author, Reviewer, Chair)
What is your CMT login email address?
What are you trying to do?
What time did this occur?
What was your last action before seeing this error? (e.g. Click on the 'submit' button)?
Additional information?

# 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

# Portable PubMedCentral

PubMed Central

- “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

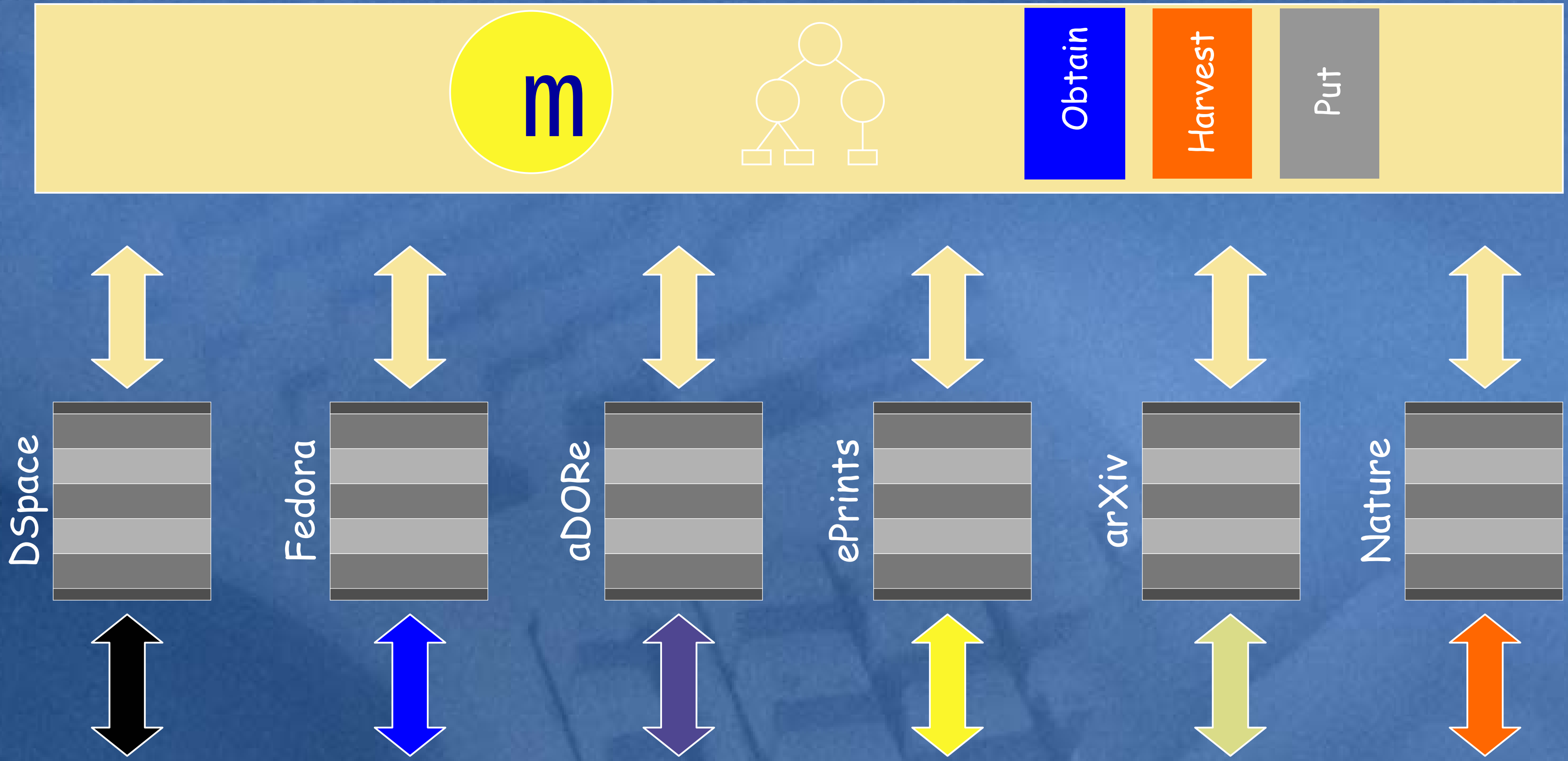
(2) Self-Archiving in Repository – ‘Green’

- 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

# 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

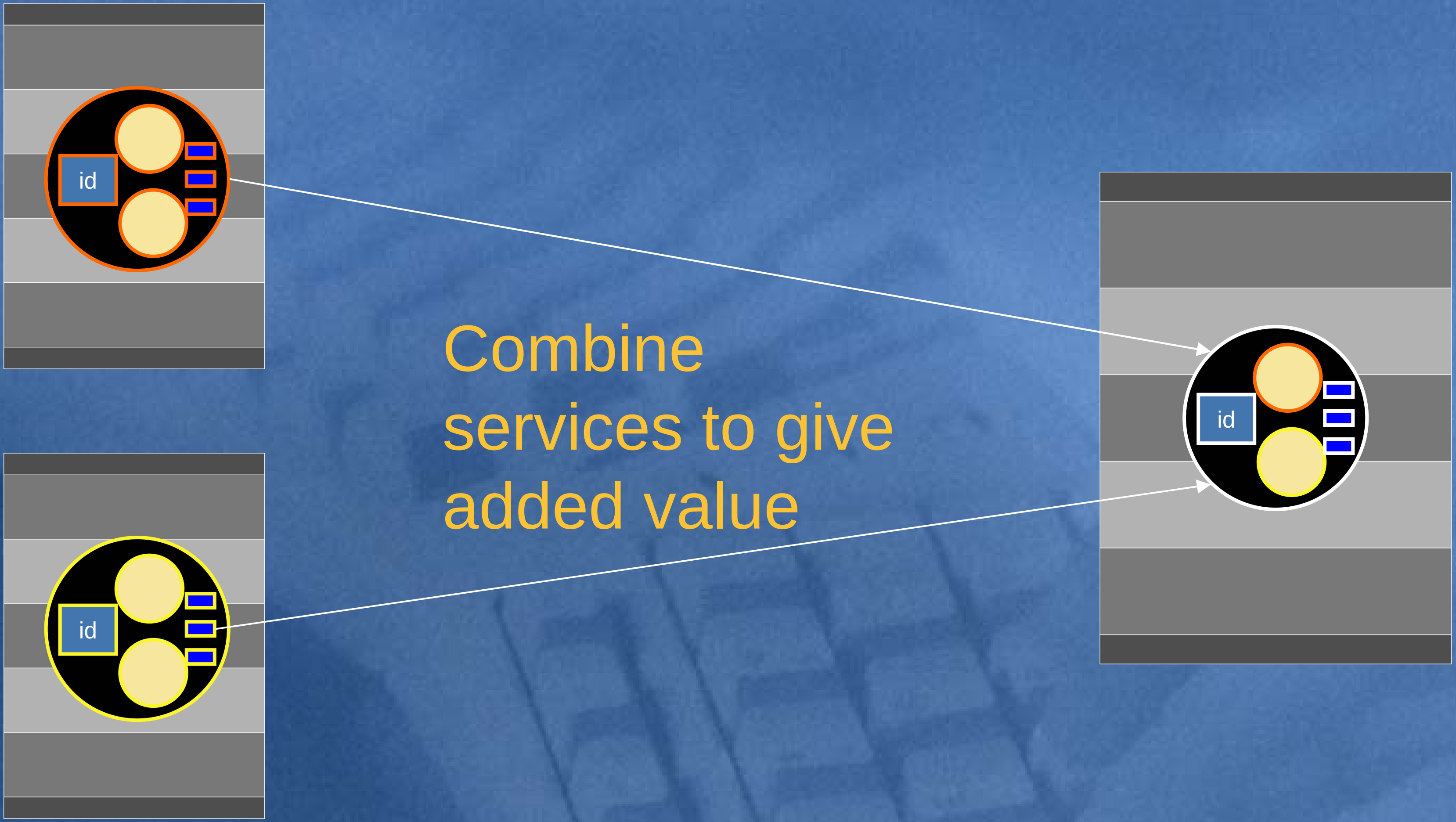


Individual Data Models and Services

# 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

Merci

Gracias

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

# 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.™

This presentation is for informational purposes only. Microsoft makes no warranties, express or implied, in this summary.