

eScience, Semantic Computing and the Cloud

Towards a Smart Cyberinfrastructure for eScience

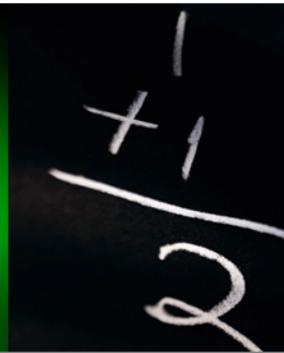
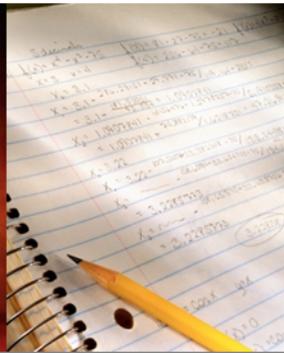
Tony Hey

Corporate Vice President

Microsoft Research



eScience



A Data Deluge in Science

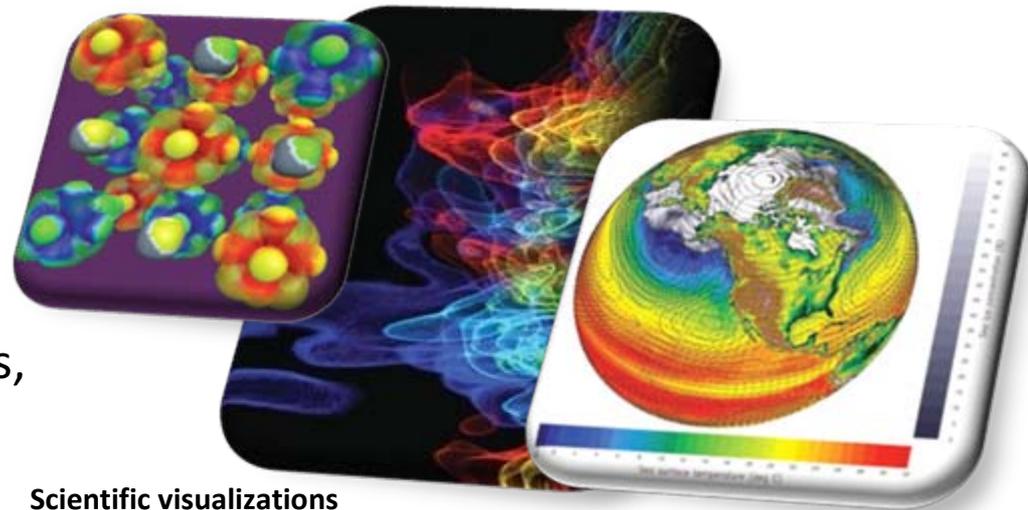
- Data collection
 - Sensor networks, satellite surveys, high throughput laboratory instruments, observation devices, supercomputers, LHC ...
- Data processing, analysis, visualization
 - Legacy codes, workflows, data mining, indexing, searching, graphics ...
- Archiving
 - Digital repositories, libraries, preservation, ...



SensorMap

Functionality: Map navigation

Data: sensor-generated temperature, video camera feed, traffic feeds, etc.



Scientific visualizations

NSF Cyberinfrastructure report, March 2007

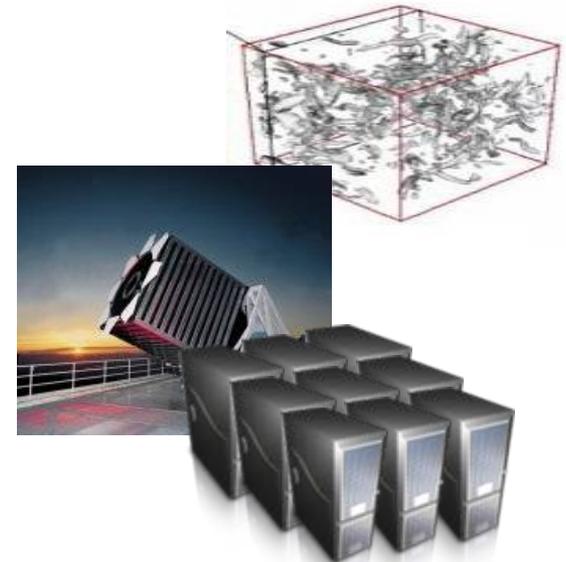


Emergence of a New Research Paradigm?

- 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 – **eScience 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
 - Scientists overwhelmed with data
 - Computer Science and IT companies have technologies that will help



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



(With thanks to Jim Gray)



Today

Web users...

- Generate content on the Web
 - Blogs, wikis, podcasts, videocasts, etc.
- Form communities
 - Social networks, virtual worlds
- Interact, collaborate, share
 - Instant messaging, web forums, content sites
- Consume information and services
 - Search, annotate, syndicate

Scientists...

- Annotate, share, discover data
 - Custom, standalone tools
- Conferences, Journals
 - Publication process is long, subscriptions, discoverability issues
- Collaborate on projects, exchange ideas
 - Email, F2F meetings, video-conferences
- Use workflow tools to compose services
 - Domain-specific services/tools



Data can be easily produced

University of Southampton Crystal Structure Report Archive

Home
About
Browse
Search
Register
User Area
Help

6, 7, 9, 10, 12, 13, 15, 16-Octahydro-benzo-1, 4, 7, 10, 13-pentaoxacyclopentadecin

Simon J Coles, Michael B Hursthouse,
Jeremy G Frey and Esther Rousay,
University of Southampton

C14H20O5

InChI=1/C14H20O5/c1-2-4-14-13(3-1)18-11-9-16-7-5-15-6-8-17-10-12-19-14/h1-4H,5-12H2

DOI: 10.594/ecrystals.chem.soton.ac.uk/145

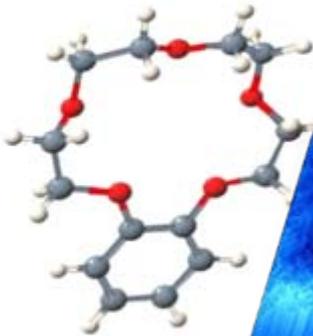
Compound Class: Organic

Keywords: crown ethers

Creation Date: 07 October 2004

Deposited By: A.N. Admin

Deposited On: 20 February 2006



Available Files

Refinement results	
Solution figure of merit	0.0409
R Factor (Obs)	0.0487
R Factor (All)	0.0977
Weighted R Factor (Obs)	0.1005
Weighted R Factor (All)	0.1192

Data Collection	
04yc0831.hkl	702k
04yc0831.hem	10k
04yc0831_0k.jpg	57k
04yc0831_hkl.jpg	85k
04yc0831_hkl.jpg	88k
04yc0831_crystal.jpg	17k

Other Files	
04yc0831.doc	78k
04yc0831.txt	155k

Depositor Comments	
Structure already known, but accurately redetermined for a local research project	04yc0831.cif 13k 04yc0831.cmr 6k

Data collection parameters	
Chemical formula	C14H20O5
Crystallisation Solvent	
Crystal morphology	Plate
Crystal system	Orthorhombic
Space group symbol	Pbca
Cell length a	16.4963(18)
Cell length b	8.325(3)
Cell length c	20.061(6)
Cell angle alpha	90.00
Cell angle beta	90.00

Validation	
04yc0831_checkcif.htm	7k

Refinement	
04yc0831.res	6k
04yc0831_at.txt	34k

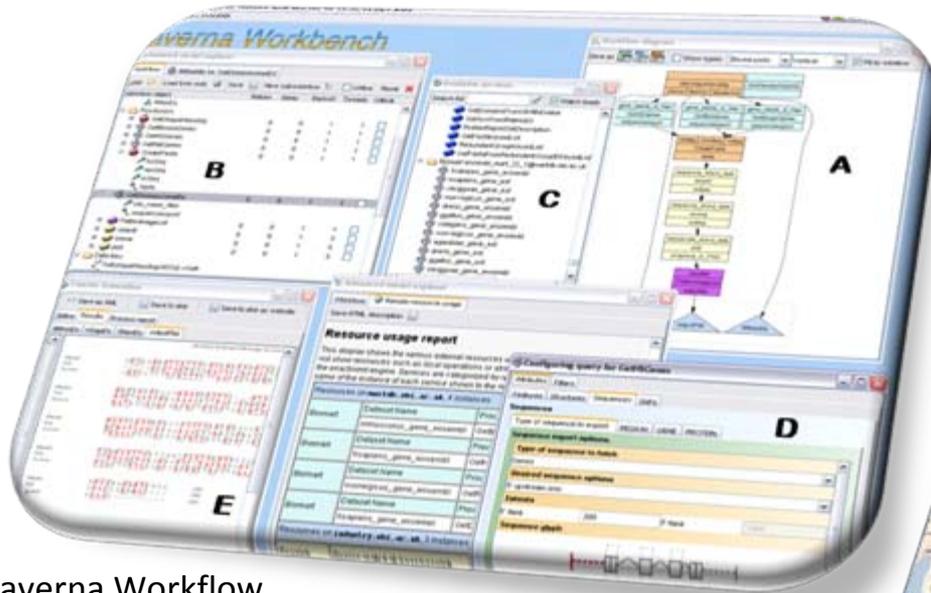
Solution	
04yc0831.prp	6k
04yc0831_xe.txt	30k

Chalkley, S.J., Hursthouse, M.B., Frey, J.G. and Rousay, E. (2004), Southampton, UK, University of Southampton, Crystal Structure Report Archive (doi:10.594/ecrystals.chem.soton.ac.uk/145)

<http://ecrystals.chem.soton.ac.uk>

Thanks to Jeremy Frey

Data and services can be easily composed



Taverna Workflow

Compose services from the Web



SensorMap

Functionality: Map navigation

Data: sensor-generated temperature, video camera feed, traffic feeds, etc.

Data is easily accessible



With thanks to
Catharine van Ingen

Today...

Computers are great **tools** for



huge amounts of **data**

For example, Google and Microsoft both have copies of the Web for indexing purposes



Tomorrow...

Computers will still be great **tools** for



huge amounts of **data**

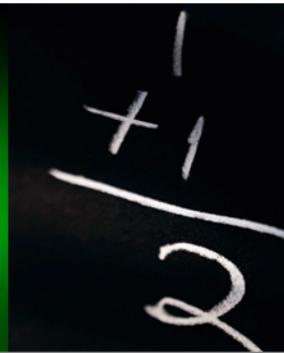
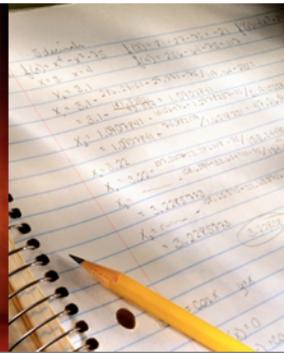
We would like computers to also help with the **automatic**



of the world's **information**

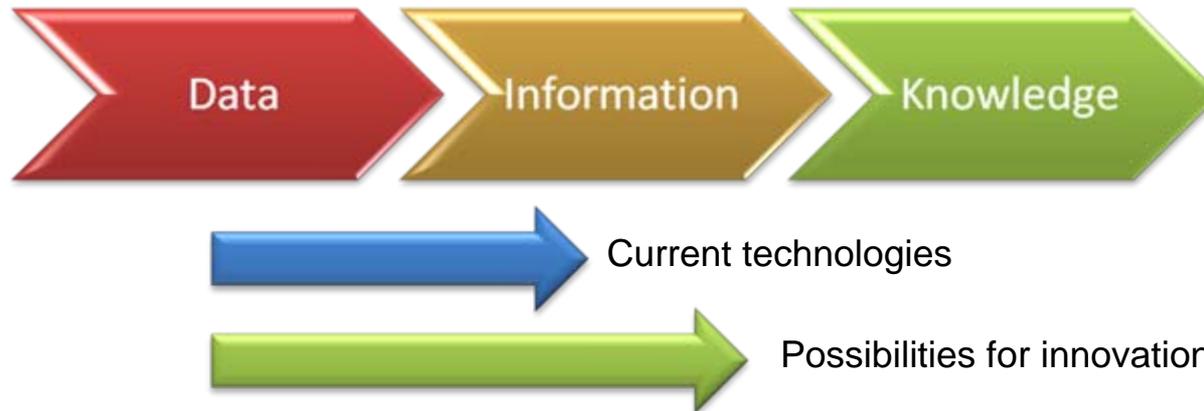


Semantic Computing



Need for Semantic Computing?

- Semantic computing combines concepts and technologies that
 - Enable data modeling
 - Capture relationships
 - Allow communities to define ontologies
 - Exploit machine learning
- Will empower computers to reason about the data



Semantic Computing

- Some efforts are driven by the traditional “knowledge engineering” community
 - Engaged in building well-controlled ontologies
 - Important for domain-specific vocabularies with data formats and relationships specific to a community
 - Model does not easily scale to the Internet
- Some efforts are driven by the Web 2.0 community
 - Focus on the pervasiveness of Web protocols/standards
 - Emphasis on microformats (small, flexible, embeddable structures)
 - Exploit evolving and ever-expanding vocabularies such as folksonomies and tag clouds



Semantic Web as the platform?



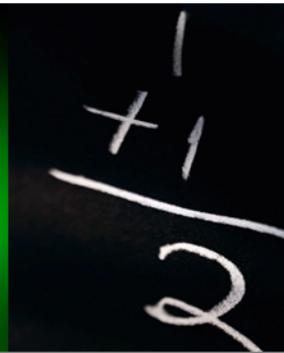
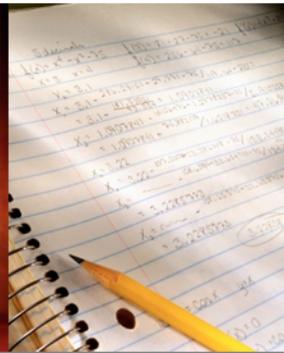
[Mark Butler \(2003\) Is the semantic web hype?](http://www.hpl.hp.com/personal/marbut/isTheSemanticWebHype.pdf)

<http://www.hpl.hp.com/personal/marbut/isTheSemanticWebHype.pdf>

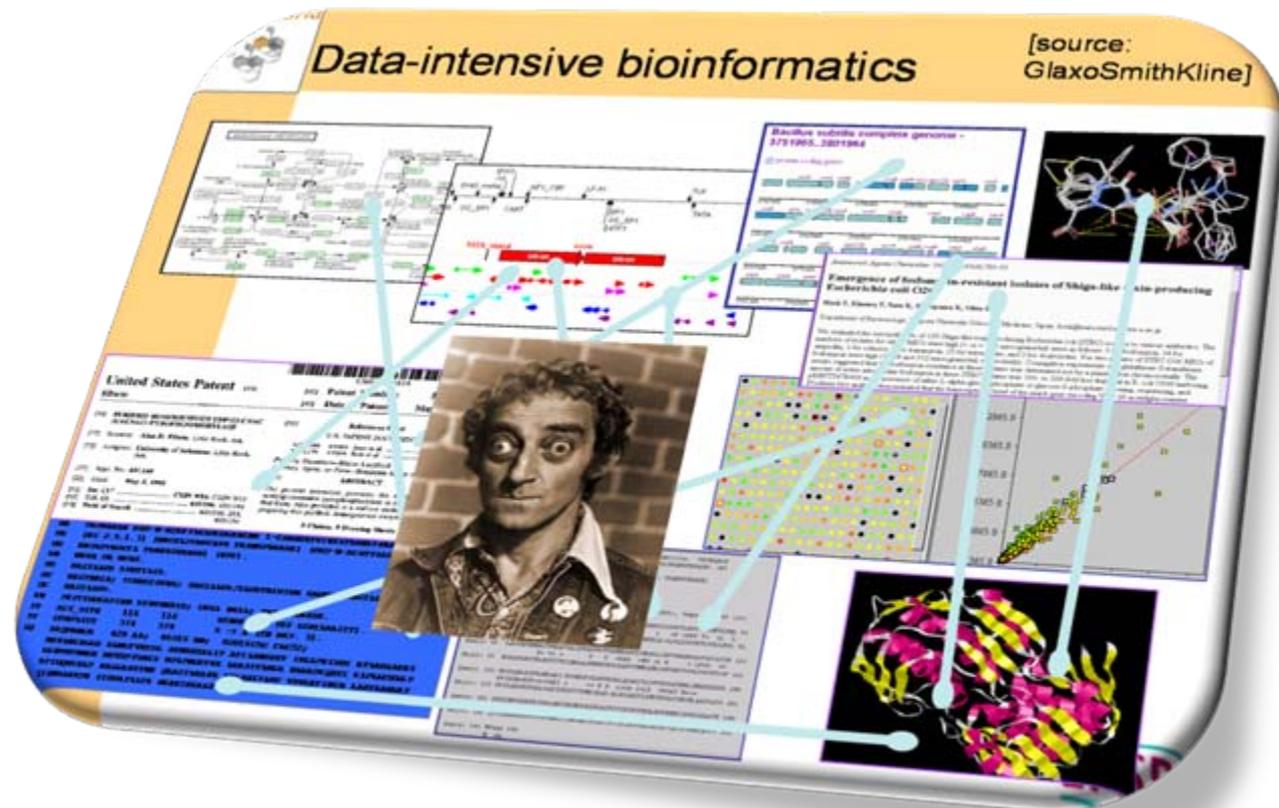


eScience and Semantic Computing

in action



- Semantic relationships between different data
- Semantic descriptions of services
- Annotations
- Provenance
- Repositories
- Ontologies



myGrid: Semantic Web Technologies

- myGrid built on Web Services, Workflows AND Semantic Web technologies
- Semantic Web technologies are used to:
 - Find appropriate services during workflow design
 - Find similar workflows for reuse and repurposing
 - Record the process and outcome of an experiment, in context
 - the experimental provenance





- Goal
 - Apply the capabilities of the AKT and MIAS IRCs to collaborative medical problem solving in the domain of breast cancer screening and diagnosis.
- Focus
 - Provide support to the Multi-Disciplinary Meetings (MDMs) that take place between various medical practitioners of different expertise, in coming to a collaborative diagnosis and plan of action in symptomatic focal breast disease.
- Services and technologies
 - Ontology Services
 - Annotation and Enrichment Services
 - Reasoning using services and GRID-services



MIAKT

MIAKT Demonstrator

Concept Browser Patient Timeline IIP V Special Nodes Zoom: Locality:

Tree List Buttons

Name

- Medical_Exam
- Medical_Image
- MetaData
- MetaData_Image
- MetaData_MRI
- MetaData_ROI
- Morphologic_Descriptor
- Non_Mass_Enha_Patt...
- Non_Mass_Enhancem...
- Nonspace_Occupying...
- Normal_Breast_Tissue
- Oncologist
- Other_Entity
- Other_Findings
- Patient
- Patient_Man_Category
- Position_Descriptor
- Radiologist
- Region_Of_Interest
- Right_CC_Image
- Right_Lateral_Image
- Shell_Descriptor
- Spread_Descriptor
- Stage_Descriptor
- Surgeon
- Symptom_Descriptor
- Tangential_Image

Concept editing

Instance 00281_patient Annotation View

Help Instances of Patient

Instance #ta-soton-1076933068214

type [Triple_Assessment_Proc](#)

consist_of_subproc [00281_mammography](#)

involve_patient [00281_patient](#)

Instance #00281_mammography

produce_result [image_00281_left_cc](#)

produce_result [image_00281_right_mlo](#)

produce_result [image_00281_right_cc](#)

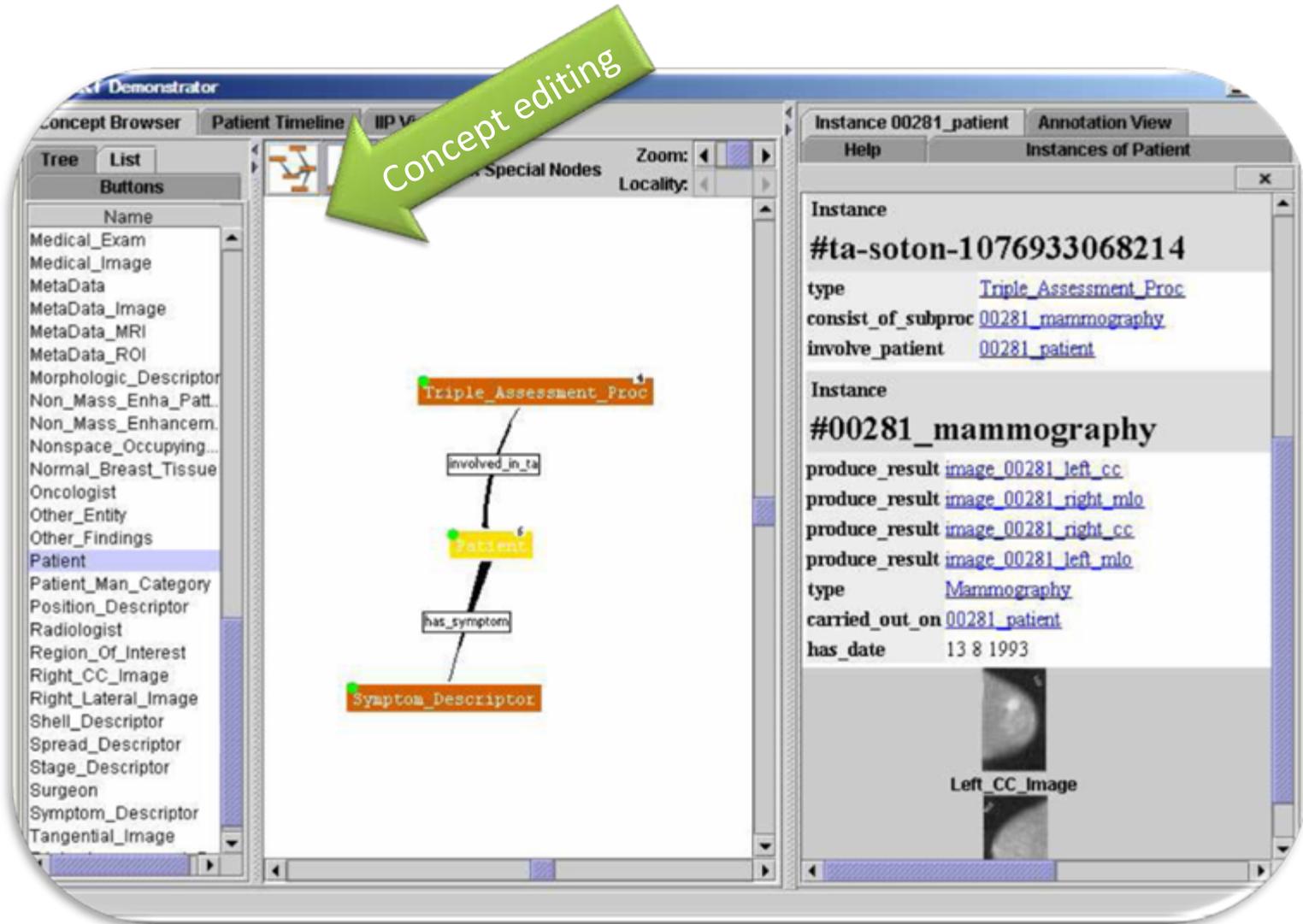
produce_result [image_00281_left_mlo](#)

type [Mammography](#)

carried_out_on [00281_patient](#)

has_date 13 8 1993

Left_CC_Image



MIAKT

The screenshot displays the MIAKT software interface. On the left, a medical image (likely a mammogram) is shown with a white outline around a region of interest. The image is labeled 'Rcc' in the top left corner. A 'Brightness %' slider is visible above the image, ranging from 0 to 4550. The right side of the interface shows a panel for 'Instance 00061_patient' in 'Annotation View'. This panel contains several sections:

- Image Features:** A table with columns 'Feature' and 'Value'. It lists 'Monochrome Histogram Feature' with a value of '[191]'.
- Regions of Interest Low Level Features:** A table with columns 'Feature' and 'Value'. It lists 'PerimeterFeature' (16.38 cm), 'AreaFeature' (439979.0 pixels), 'BoundaryFeature' ((1624:3400),(1629:3400)), and 'Monochrome Histogram Feature' ([191]).
- Region of Interest Semantics:** A table with columns 'Concept' and 'Value'. It lists 'Lesion Type' (ralc lucent centered) and 'Lesion Shape' (chano mamma stellate).

Below these tables is a 'Delete Annotation' button. At the bottom of the panel, there is a section for 'New Region of Interest: roi_0' with the following text:

- Region of Interest 0 type assertion
- Feature Vector assertion: PerimeterFeature = 16.38 cm
- Feature Vector 0 type assertion
- Link Feature Vector 0 with ROI 0
- Feature Vector assertion: AreaFeature = 439979.0 pixels
- Feature Vector 1 type assertion
- Link Feature Vector 1 with ROI 0

At the very bottom of the panel are two buttons: 'Assert these' and 'Remove Selected'.

Two green arrows point to the right side of the interface. The top arrow is labeled 'Semantic associations' and points to the 'Region of Interest Semantics' table. The bottom arrow is labeled 'Annotations' and points to the 'New Region of Interest' section.



SWAN Project

- Semantic Web Applications of Neuromedicine
- Project led by Tim Clark and the IIC

SWAN Project

Semantic Web Applications in Neuromedicine



About

SWAN is a project to develop effective specialist knowledge bases for the Alzheimer Disease research community, using the energy and self-organization of that community enabled by Semantic Web technology.

Project Documents & Links

- [SWAN Project Abstract](#)
- [Pathways Knowledgebase Presentation \(SWLS 2004\)](#)
- SWAN: Knowledge Infrastructure for Alzheimer Disease Research (contact PIs)
- SWAN Alzheimer Disease Research Use Cases (contact PIs)
- [SWAN Talk, Biological Data Standards Normalization Meeting, Oct 2005](#)
- SWAN RDF Schema at purl.org/swan/0.1/

These and all other SWAN documents and software are provided under terms of the [W3C® Document License](#) applicable to W3C® public documents, except where other terms are specifically noted.

Funding

Funding for the SWAN project has been generously provided by the [Ellison Medical Foundation](#), and other charitable organizations.

The MIND Center for Interdisciplinary Informatics (CII) develops and applies integrative computational methods in biomedical and brain research, working with leading clinicians and researchers to understand and cure neurological disorders.

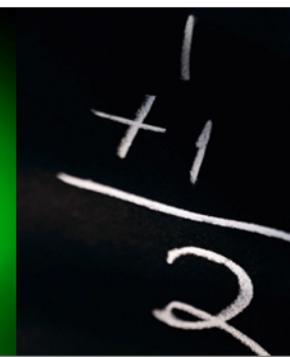
We are funded by the [MassGeneral Institute for Neurodegenerative Disease](#), the [Martinos Center for Biomedical Imaging](#), and by Federal and private charitable grants.

MassGeneral Institute for Neurodegenerative Disease
114 16th Street, CNY 114
Charlestown, MA 02129

email:
mahmind@partners.org
phone: +1 617 724 3585
+1 617 726 1278



Cloud Computing



Rationale for Cloud computing

- Outsourcing of IT infrastructure
- Minimize costs
 - Large cloud/utility computing provides can have relatively very small ownership and operational costs due to the huge scale of deployment and automation
- Small businesses have access to large scale resources
 - The acquisition, operation, and maintenance costs would have been prohibiting



Example: Amazon Web Services

Simple Storage Service (S3)

- storage for the Internet
- Simple Web Services interface to store and retrieve any amount of data from anywhere on the Web

SimpleDB

- Structured data

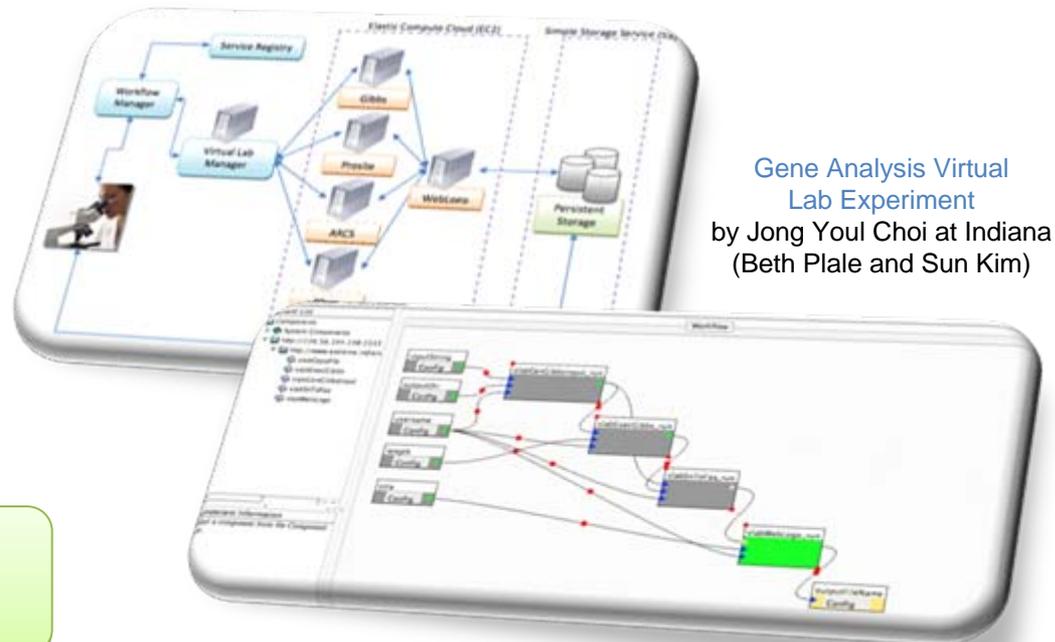
Simple Queue Service

- Scalable message queuing

Standards-based REST and SOAP
Web Service interfaces

Elastic Compute Cloud (EC2)

- Compute on demand
- Virtualization
- Integration with S3



Gene Analysis Virtual
Lab Experiment
by Jong Youl Choi at Indiana
(Beth Plale and Sun Kim)



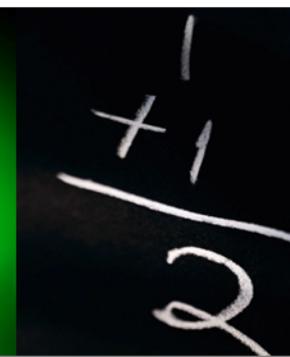
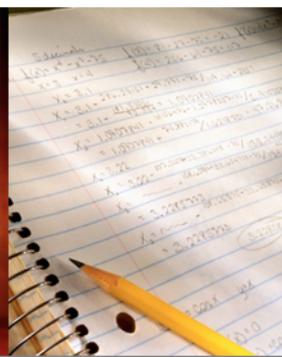
Microsoft Cloud Services

- Exchange Server hosting
 - Live@EDU
 - BizTalk Services
 - Mail (Live Mail, Hotmail)
 - Identity (Live ID)
 - Dataflow (PopFly)
 - Xbox Live
 - SQL Server Data Services
 - Office Live Workspaces
 - Windows Live
- Many more coming



eScience and Cloud Computing

in action



The SkyServer Project

Jim Gray (MSR) and Alex Szalay (JHU)



- The Sloan Digital Sky Survey (SDSS):
The “Cosmic Genome Project”
 - 5 color images of $\frac{1}{4}$ of the sky
 - Pictures of 300 million celestial objects
 - Distances to the closest 1 million galaxies
- Built the public archive for the SDSS
- Interesting challenge in digital publishing
 - Have to publish first in order to analyze



Public Use of the SkyServer

- **Posterchild in 21st century data publishing**

- 380 million web hits in 6 years
- 930,000 distinct users vs 10,000 astronomers
- 1600 refereed papers!
- Delivered 50,000 hours of lectures to high schools
- Delivered 100B rows of data



➤ **World's most used astronomy facility for last 2 years**

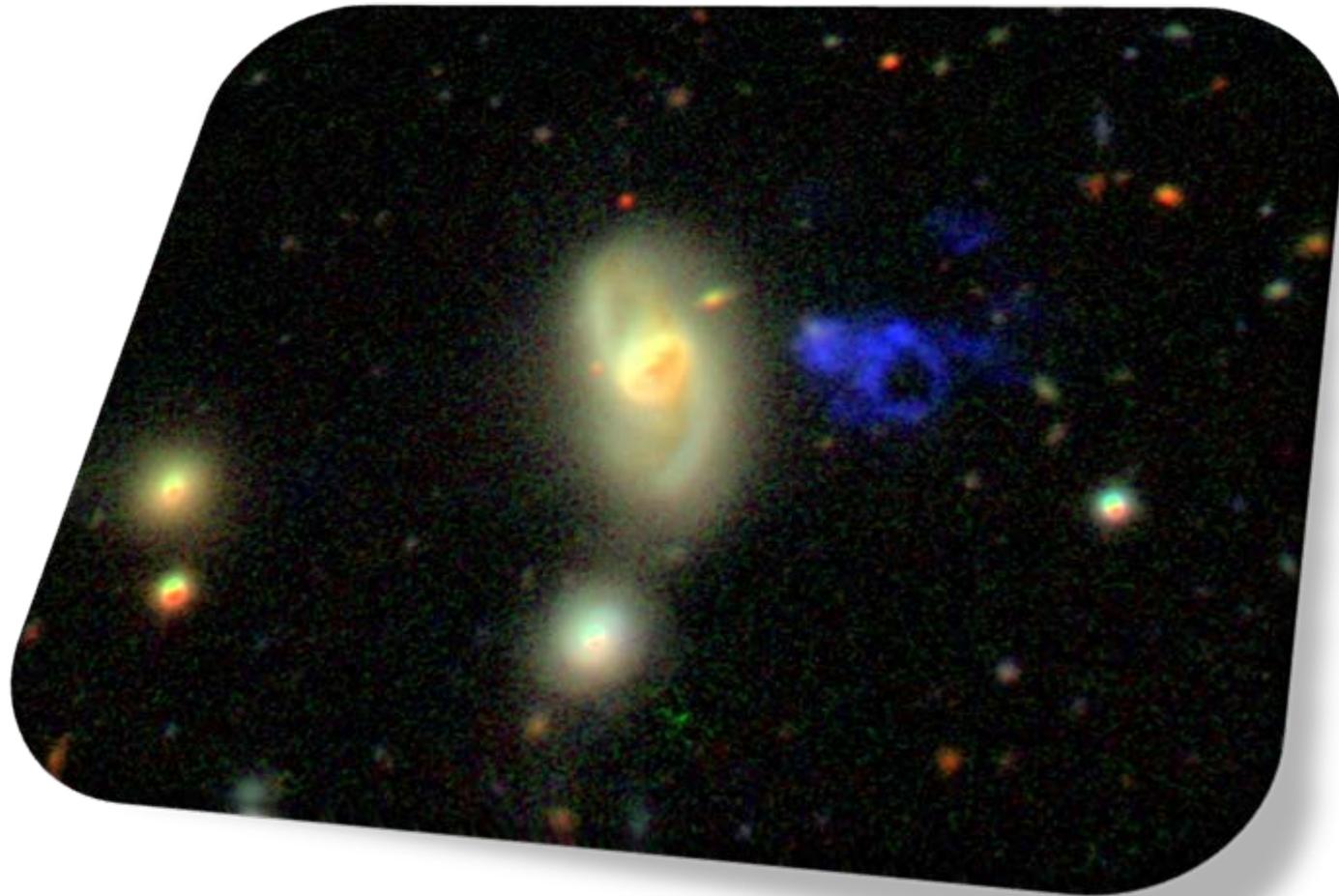


GalaxyZoo

- Goal of 1 million visual galaxy classifications by the public
- Enormous publicity (CNN, Times, Washington Post, BBC)
- 100,000 people participating, blogs, poems ...
- Application is like Amazon's 'Mechanical Turk' Web Service that allows users to search for photographs ...



Hanny's Voorwerp



World Wide Telescope

Seamless Rich Social Media Virtual Sky
Web application for science and education

Project organization

- Alyssa Goodman (Harvard)
- Alex Szalay (JHU)
- Curtis Wong, Jonathan Fay (MSR)

Project Goals

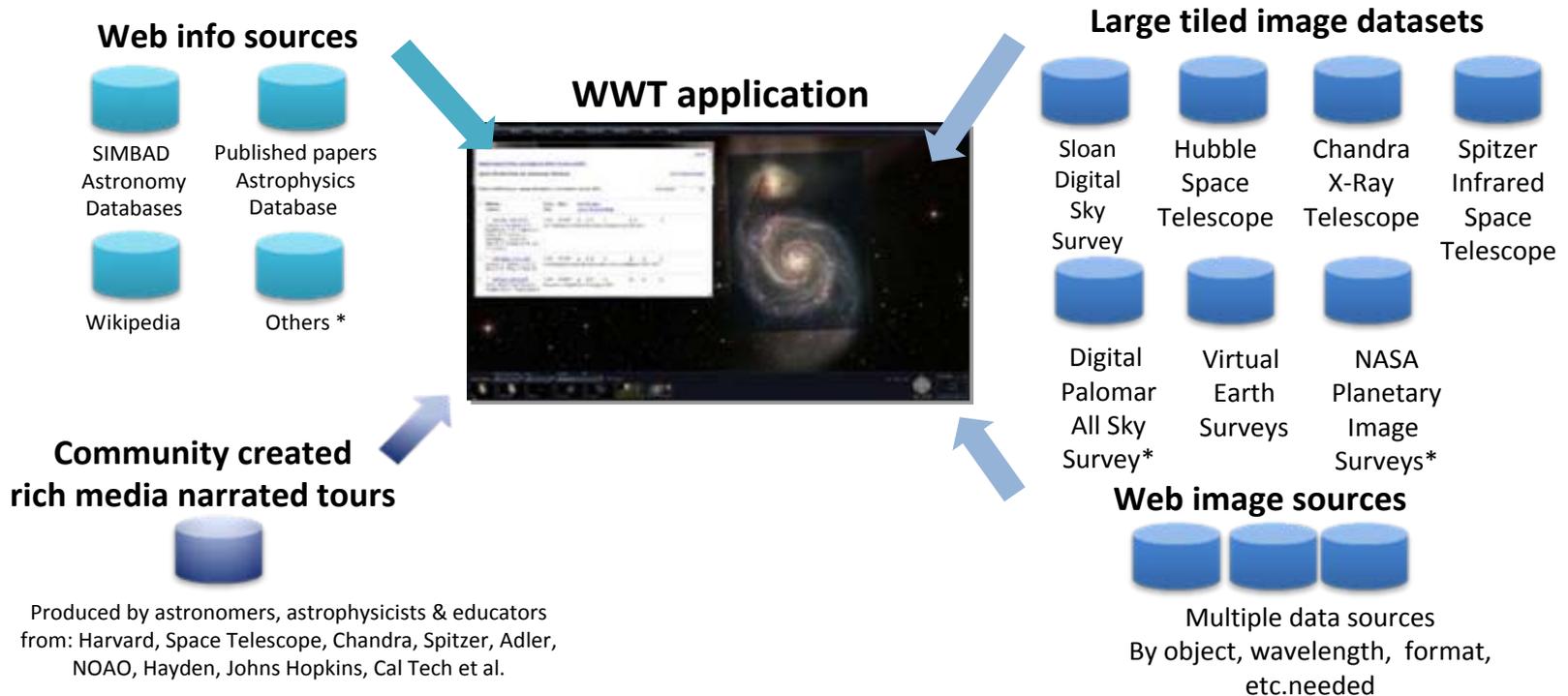
- Science- Seamless integration of data sets and one click contextual access
- Education- Easy as Powerpoint

Soon: <http://www.worldwidetelescope.org/>



WWT Architecture

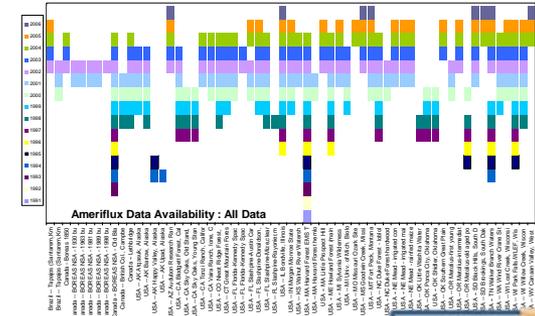
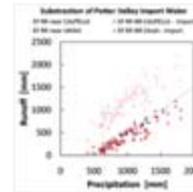
- Web 2.0 browsing environment for large distributed image and information data sets with integrated rich media authoring and annotation sharing
- Integrated easy to use rich social media authoring environment
- Geospatial Tiled Multiresolution Image Browser - distributed massive image and data sets
- Links to deep web information source



Berkeley Water Center



Understanding regional hydrology



Project Organization

- Jim Hunt, Dennis Baldocchi, UC Berkeley
- Deb Agarwal, Lawrence Berkeley Laboratory
- Catharine van Ingen, MSR

Goals

- Enable rapid scientific data browsing for availability and applicability
- Enable environmental science via data synthesis from multiple sources

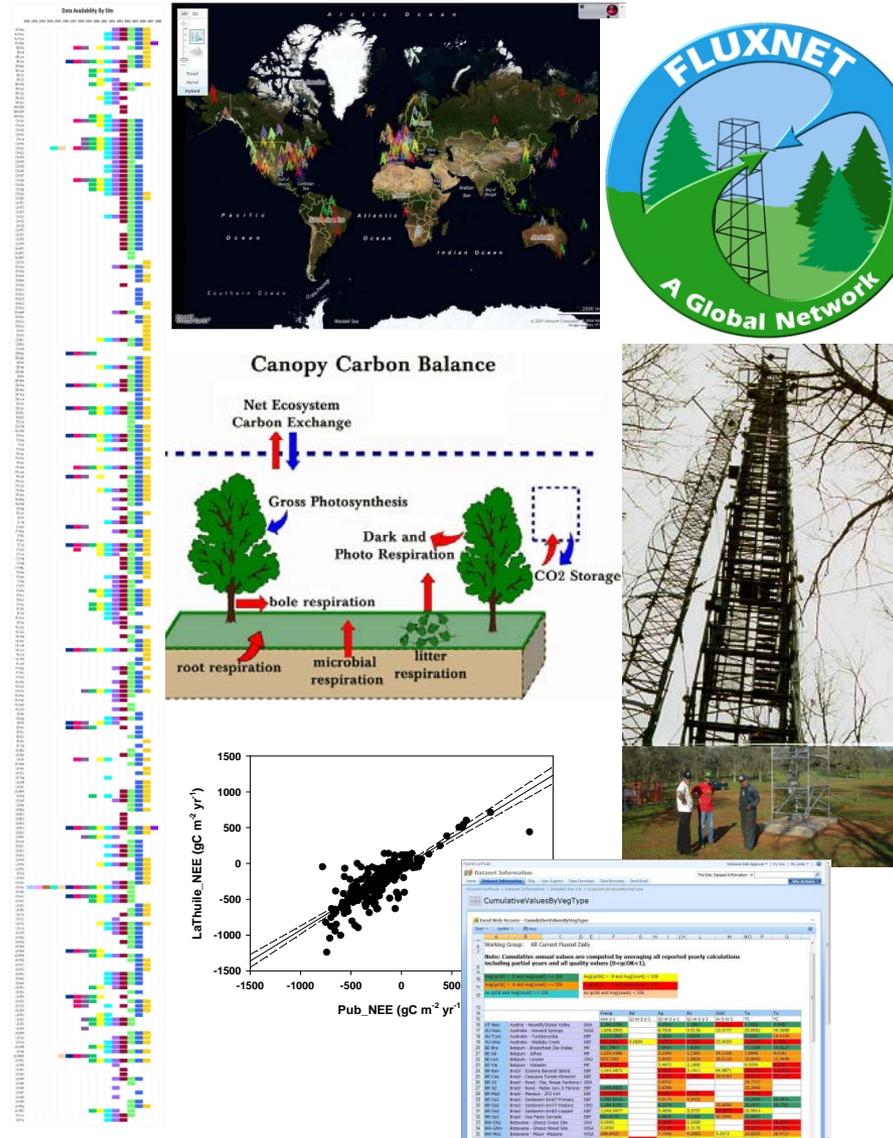
Proof Points

- Environmental Data Server, www.fluxdata.org (SharePoint), serves **921 site years** of carbon-climate field data from 160+ field teams to 60+ paper writing teams (800M values)
- Multiple projects now **leveraging** same SQL Server database and data cube approach
- CUAHSI consortium: **100 universities collaborating** on hydrology

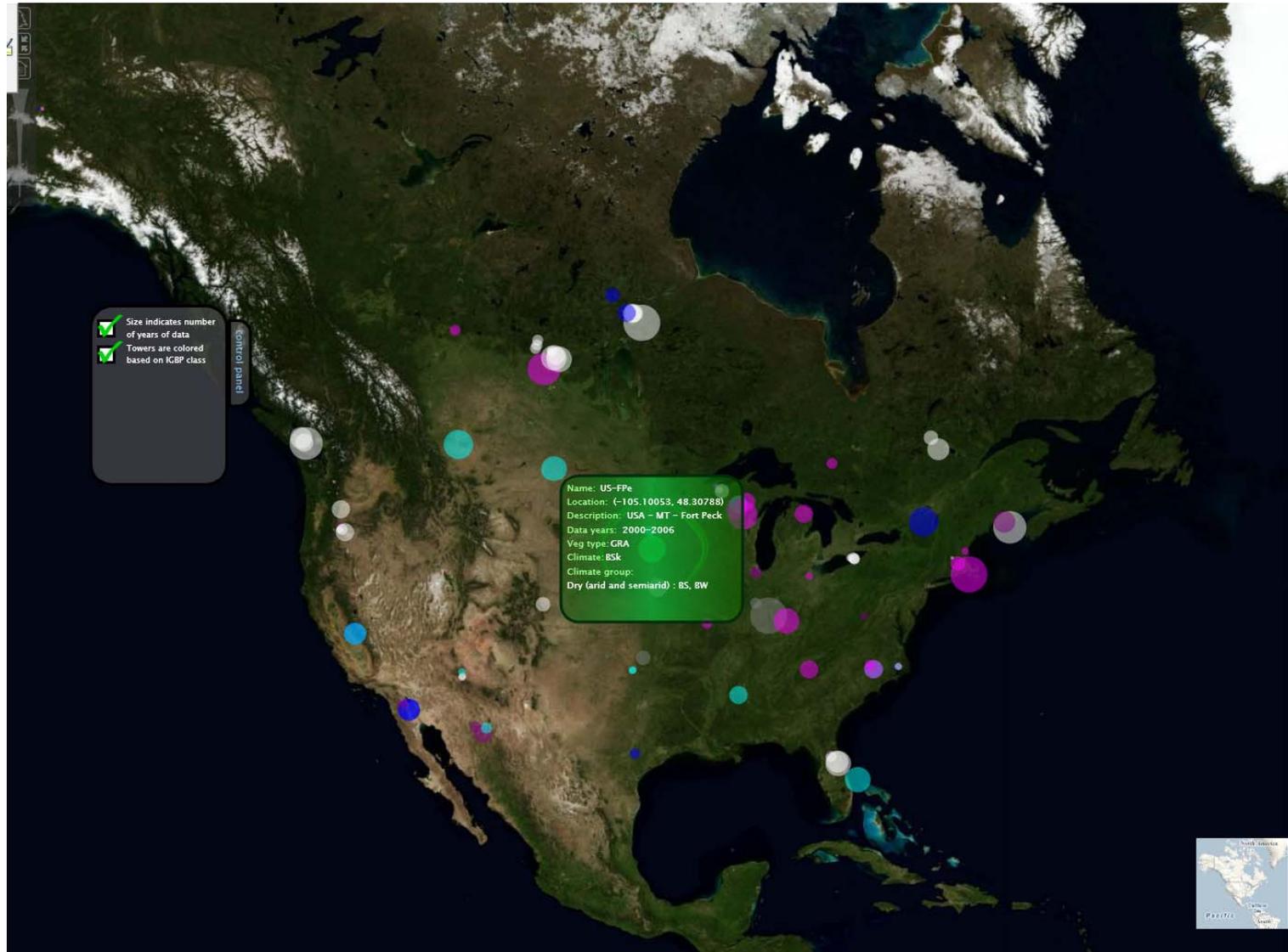


Carbo-Climates Synthesis (BWC Dennis Baldocchi et al)

- What is the role of photosynthesis in global warming?
 - Measurements of CO₂ in the atmosphere show 16-20% less than emissions estimates predict
 - The difference is either due to plants or ocean absorption.
- Communal field science – each investigator acts independently.
- Cross site studies and integration with modeling increasingly important
- Sharepoint site www.fluxnet.org
 - 921 site-years of data from 240 sites around the world; 80+ site-years now being added
 - 60+ paper writing teams
 - American data subset is public and served more widely
 - Summary data products greatly simplify initial data discovery



Mashup of Ameriflux Sites



Digital Watersheds (BWC, James Hunt)

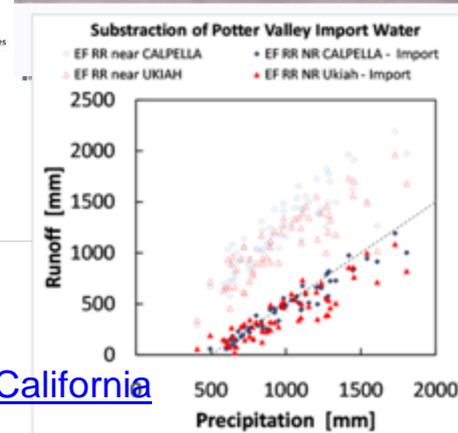
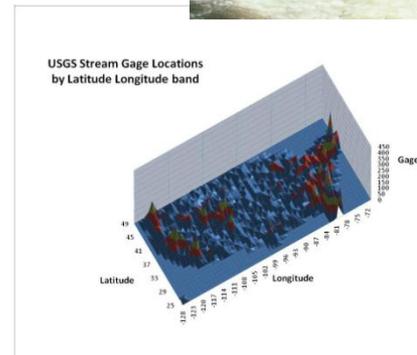
- Russian River watershed challenges: forestry, farming, urbanization, gravel mining, and fish habitat restoration.
 - Can we understand historic and on-going changes using only publically available data sources such as USGS, NOAA, Sonoma Ecology Center, etc?
- Early studies examined overall water balance and changes in suspended sediment
 - scientific data “mashups” are leading to useful results.
- Recent engagement with National Marine Fisheries and USBR expanding this to other watersheds across Northern California
[http://www-esd.lbl.gov/BWC/ California](http://www-esd.lbl.gov/BWC/California)
- Sharing technology with CUAHSI (100 universities)

“We see water through a fish eye lens”

<http://bwc.berkeley.edu>

<http://www-esd.lbl.gov/BWC/California>

<http://www.cuahsi.org/>



Supporting researchers worldwide

A Software + Services vision

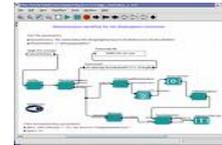


Research Pipeline



- Data Acquisition and Modeling

- Data capture from source, cleaning, storage, etc.
- SQL Server, SSIS, Windows WF



- Support Collaboration

- Allow researchers to work together, share context, facilitate interactions
- SharePoint Server, One Note 2007 (shared)



- Data Analysis, Modeling, and Visualization

- Mining techniques (OLAP, cubes) and visual analytics
- SQL Analysis Services, BI, Excel, Optima, SILK (MSR-A)



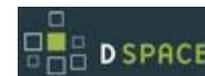
- Disseminate and Share Research Outputs

- Publish, Present, Blog, Review and Rate
- Word, PowerPoint



- Archiving

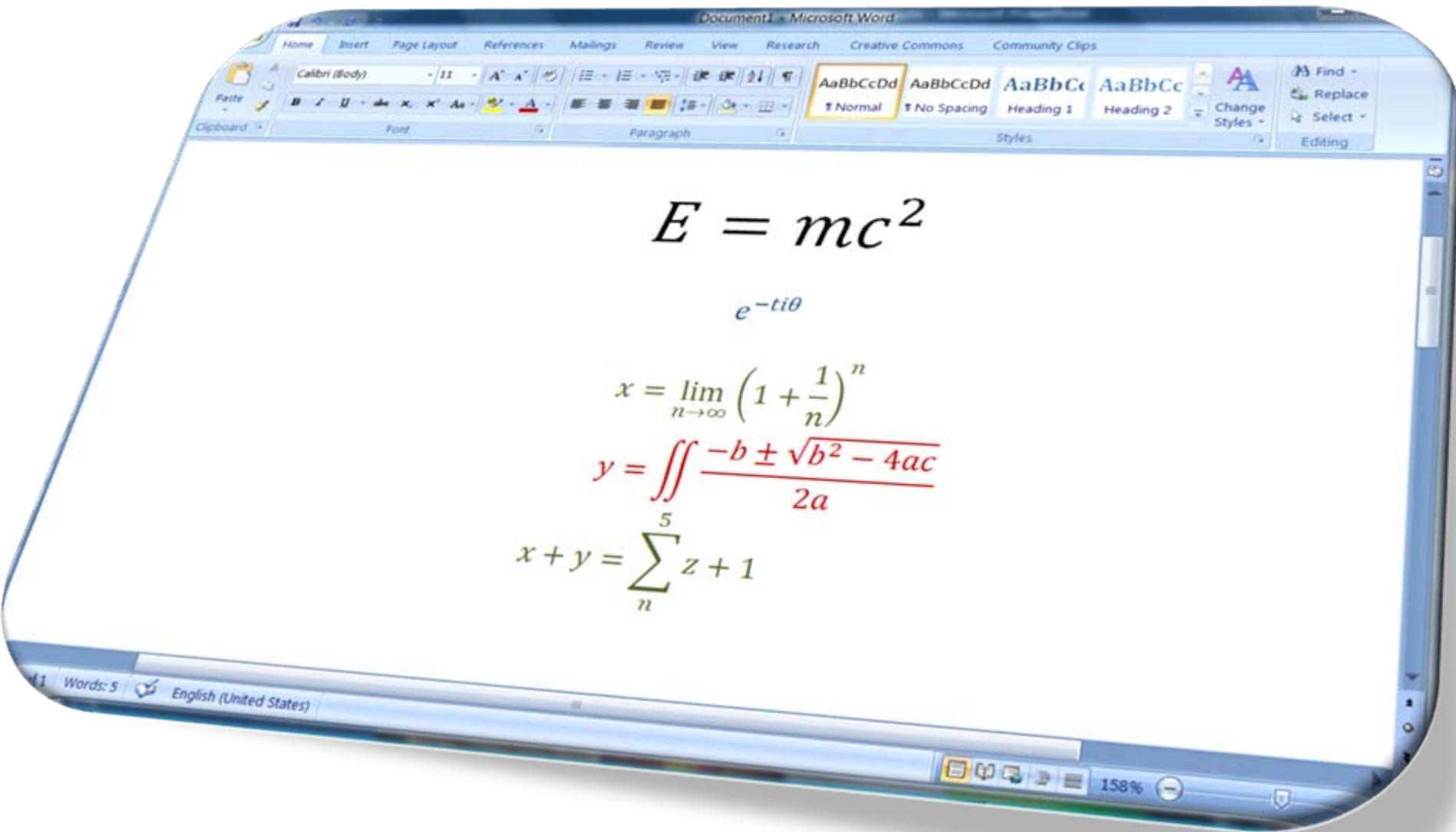
- Published literature, reference data, curated data, etc.
- SQL Server



Microsoft has technologies that can offer end-to-end support



Math in Word 2007





Trident Scientific Workflow Workbench

Univ. of Washington and Monterey Bay Aquarium Research Institute

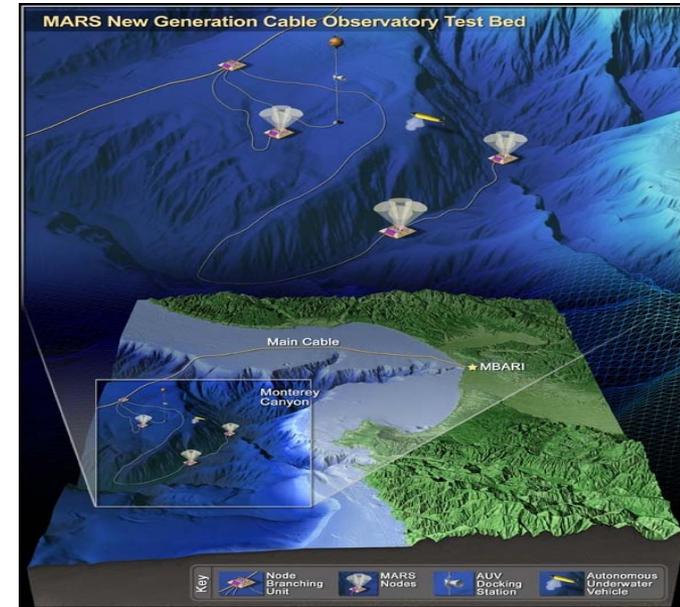
Scientific workflow workbench to automate the data processing pipelines of the world's first plate-scale undersea observatory

Goals

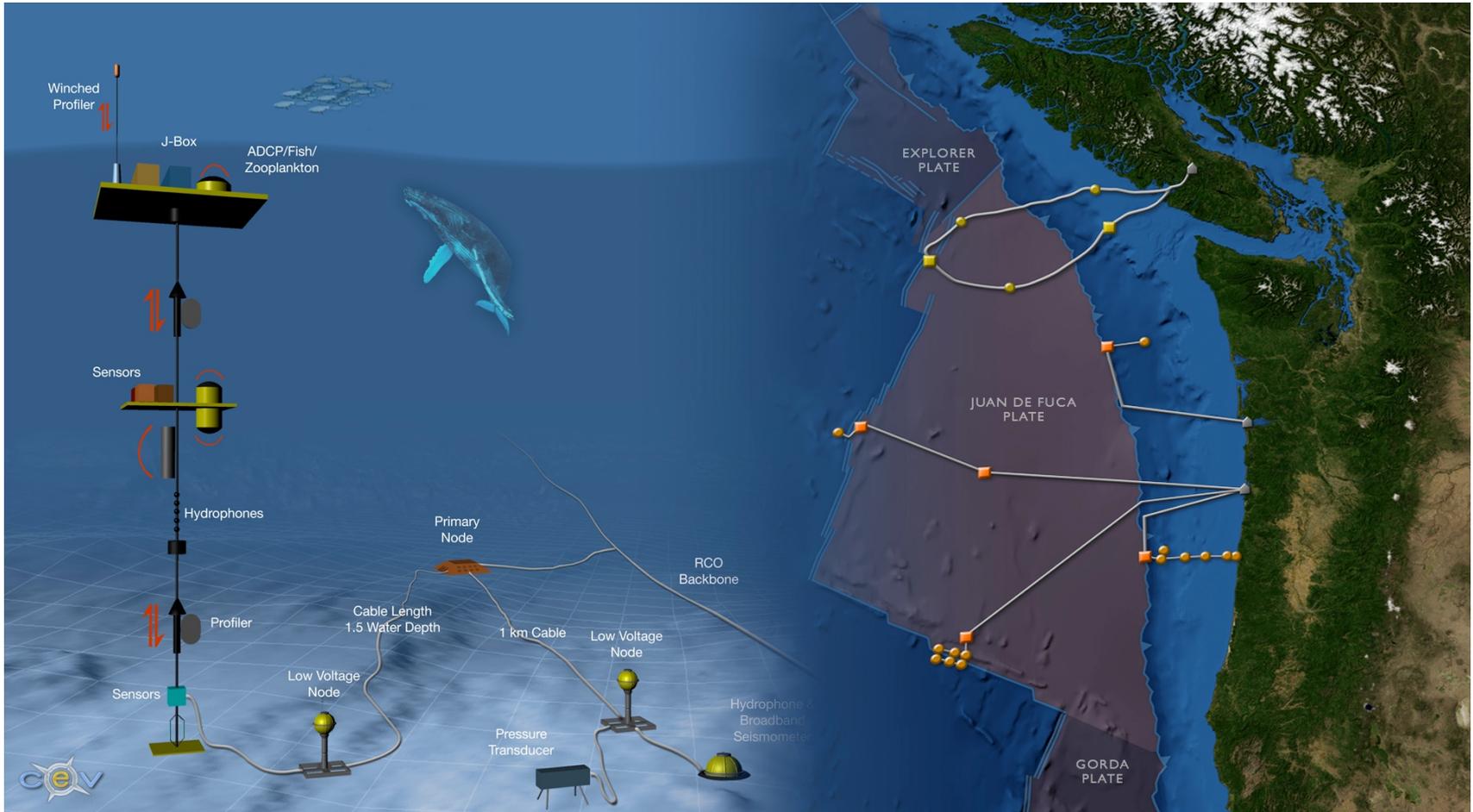
- From raw data to useable data products (visualizations)
- Focusing on cleaning, analysis, regridding, interpolation
- Support real time, on-demand visualizations
- Custom activities and workflow libraries for authoring
- Visual programming accessible via a browser

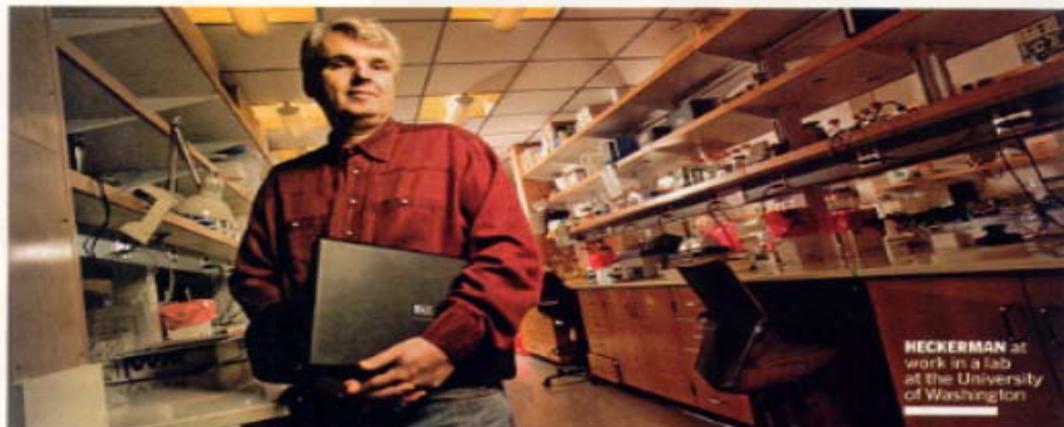
Proof Points

- A **scientific workflow workbench** for a number of science projects, reusable workflows, automatic provenance capture.
- **Demonstrate scientific use** of Windows WF and SQL Server
- Demo at TechFest 2008, in collaboration with the UW Neptune team.



Trident – Scientific Workbench





Using Spam Blockers To Target HIV, Too

A Microsoft researcher and his team make a surprising new assault on the AIDS epidemic

BY STEPHEN BAKER AND JAY GREENE

CUT-RATE PAINKILLERS! Unclaimed riches in Nigeria! Most of us quickly identify such e-mail messages as spam. But how would you teach that skill to a machine? David Heckerman needed to know. Early this decade, Heckerman was leading a spam-blocking team at Microsoft Research. To build their tool, team members meticulously mapped out thousands of signals that a message might be junk. An e-mail featuring "Viagra," for example, was a good bet to be spam—but things got complicated in a hurry.

If spammers saw that "Viagra" messages were getting zapped, they switched to Viagra, or Vi agra. It was almost as if spam, like a living thing, were mutating.

This parallel between spam and biology resonated for Heckerman, a physician as well as a PhD in computer science. It didn't take him long to realize that his spam-blocking tool could extend far beyond junk e-mail, into the realm of life science. In 2003, he surprised colleagues in Redmond, Wash., by refocusing the spam-blocking technology on one of the world's deadliest, fastest-mutating conundrums: HIV, the virus that leads to AIDS.

Heckerman was plunging into medicine—and carrying Microsoft with him. When he brought his plan to Bill Gates, the company chairman "got really excited," Heckerman says. Well versed on HIV

from his philanthropy work, Gates lined up Heckerman with AIDS researchers at Massachusetts General Hospital, the University of Washington, and elsewhere.

Since then, the 50-year-old Heckerman and two colleagues have created their own biology niche at Microsoft, where they build HIV-detecting software. These are research tools to spot infected cells and correlate the viral mutations with the individual's genetic profile. Heckerman's team runs mountains of data through enormous clusters of 320 computers, operating in parallel. Thanks to smarter algorithms and more powerful machines, they're sifting through the data 480 times faster than a year ago. In June, the team released its first batch of tools for free on the Internet.

A new industry for the behemoth to conquer? Not exactly. Heckerman's nook in Redmond represents just one small node in a global AIDS research effort marked largely by cooperation. "The Microsoft group has a different perspective and a good statistical background," says Bette Korber, an HIV researcher at Los Alamos National Laboratories. The key quarry they all face is the virus itself, which is proving wlier than any of Microsoft's corporate foes. While Heckerman has high hopes that his tools will lead to vaccines that can be tested on humans within three years, his research

Similar mutations may crop up in computer and medical viruses

PhyloD: Leveraging phylogeny for associations studies

PhyloD (4)

Add Dist:

Distribution:

Computation has finished!

Elapsed time: 2 s

Results:

LeafDistribution	rowIndex	rowCount	pieceIndex	NullIndex				
ConditionalEscape	0	1	0	-1	A02	326	360	68

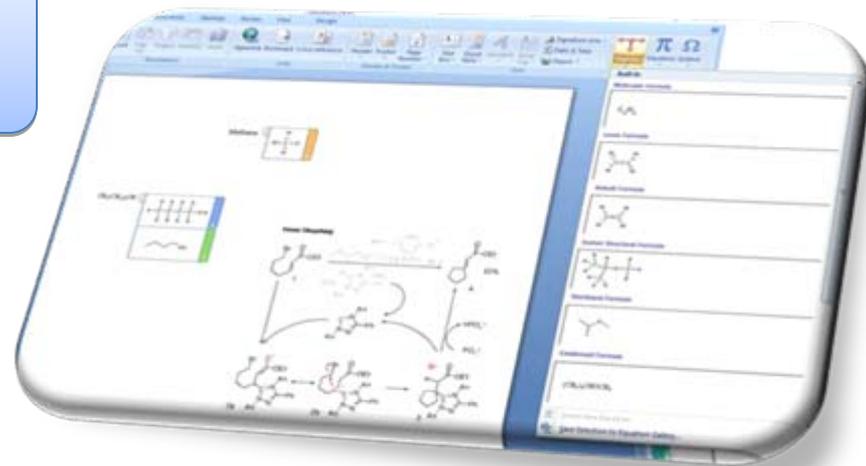
- Problem: Find associations between genotypes and phenotypes (e.g., genetic causes of disease) using data from a set of “individuals” (humans, HIV viruses, etc.)
- Previous solutions ignore phylogenetic structure (i.e., assume the data to be IID). This solution leverages phylogenetic structure of the individuals.
- Applications:
 - Identify effects of immune pressure on HIV evolution (papers in *Science* and *Nature Medicine*)
 - Inferring protein structure (in collaboration with David Baker)
 - Genome-wide association studies (personalized medicine)





Chemistry Drawing for Office

- Peter Murray Rust, Univ. of Cambridge
- Murray Sargent, Office
- Geraldine Wade, Advanced Reading Technologies



Goals

- Support students/researchers in simple chemistry structure authoring/editing
- Enable ecosystem of tools around lifecycle of chemistry-related scholarly works
- Support the Chemistry Markup Language
- Proof of concept plug-in

Execution

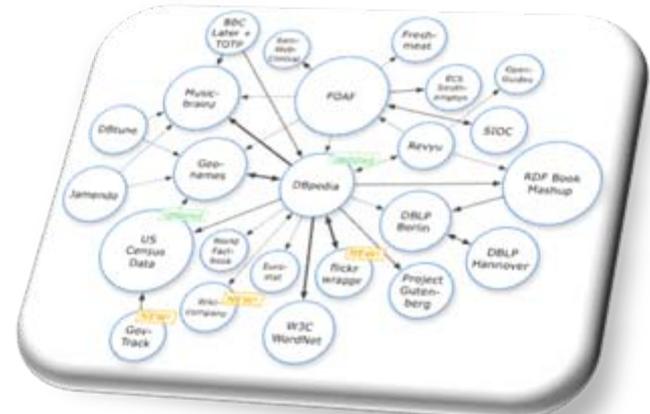
- MSR Developer to work on the proof of concept
- Post-doc in Cambridge to use plug-in and give feedback and move their chemistry tools to .NET and Office
- Advanced Reading Technologies to create necessary glyphs





Semantic Annotations in Word

- Phil Bourne and Lynn Fink, UCSD



Attribution: [Richard Cyganiak](#)

Goals

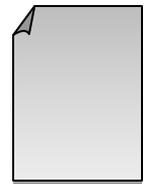
- Semantic mark-up using ontologies and controlled vocabularies
- Facilitate/automate referencing to PDB (and other resources) from manuscript
- Conversion of manuscript to NLM DTD for direct submission to publisher

Scenario

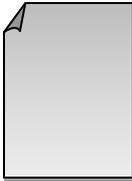
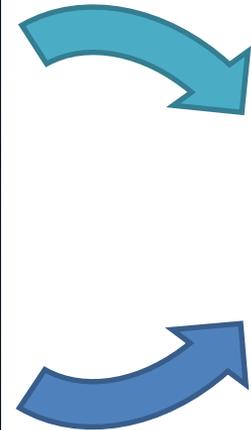
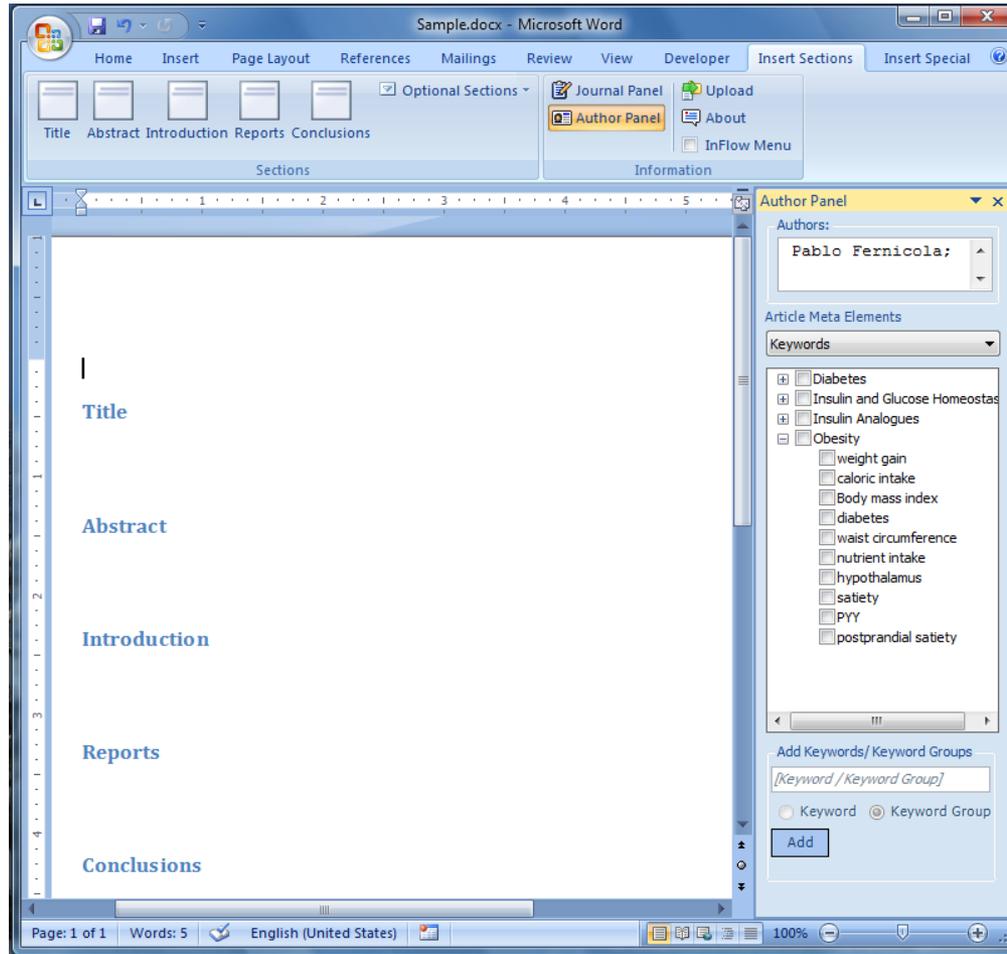
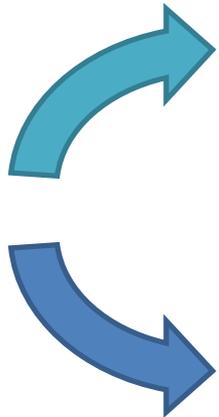
- Authors do not need to be aware of the use of semantic technologies
- A domain-specific ontology is downloaded and made available from within Microsoft Word 2007
- Authors can record their intention, the meaning of the terms they use based on their community's agreed vocabulary



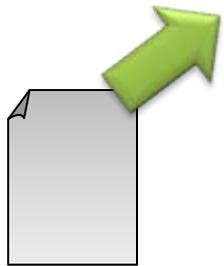
NLM DTD plug in



XML



nlmx



Journal Template





Research Output Repository

A platform for building services and tools for research output repositories

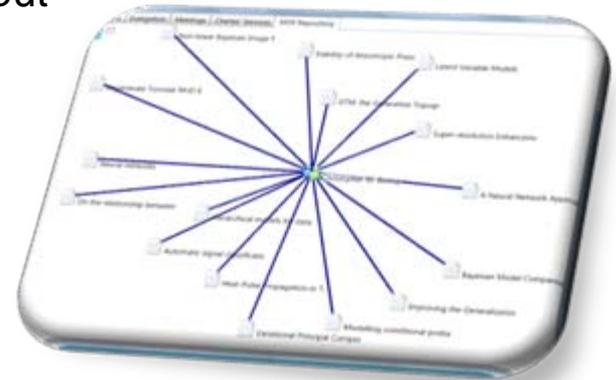
- Papers, Videos, Presentations, Lectures, References, Data, Code, etc.
- Relationships between stored entities

Goals

- Support the MSR publishing and dissemination platform for all researcher outputs
- Enable a tools and services ecosystem for “research output” repositories on MS technologies

Execution

- Support Eprints and Dspace front ends
- Deployment within MSR early Q2
- Release to the community late Q2
- Built on SQL Server 2008 + Entity Framework



Research Output Repository Platform

- A Semantic Computing platform
- A hybrid between a relational database and a triple store

Triple stores

- Evolution friendly
- Poor performance
- No need to model everything in advance
- Semantic interpretation at the application level

Relational schema

- Evolution not so easy
- Great opportunities for optimization
- Model everything in advance

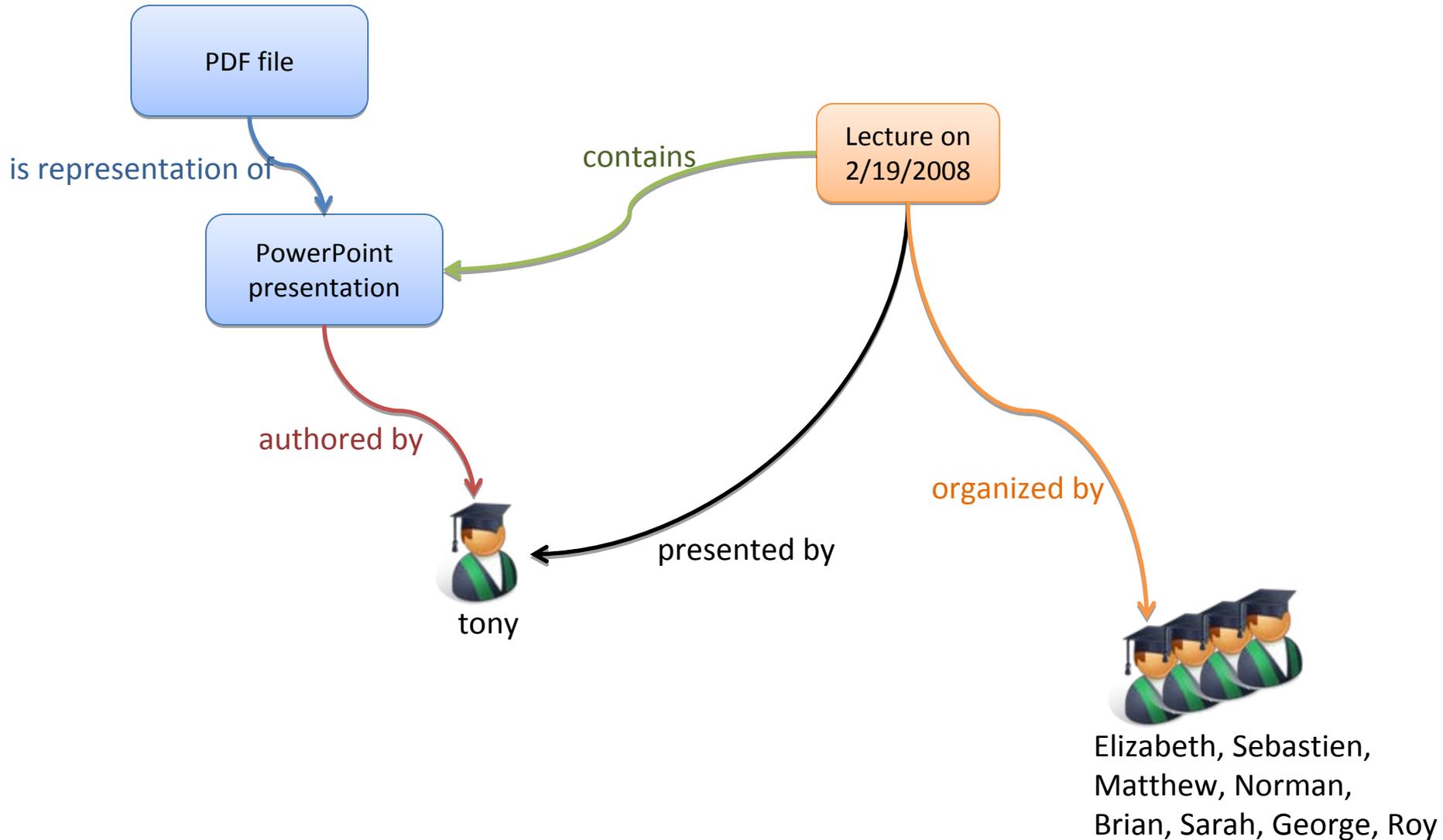


Research Output Repository Platform

- Maintain a balance
- Try to model the frequently used entities in our app domain
- Try to capture the frequently used relationships
- Allow for extensibility (Relationships, Properties)



Research Output Repository Platform



A Digital Dark Age?





PLANETS

Tools and methods for sustainable long-term preservation of digital objects

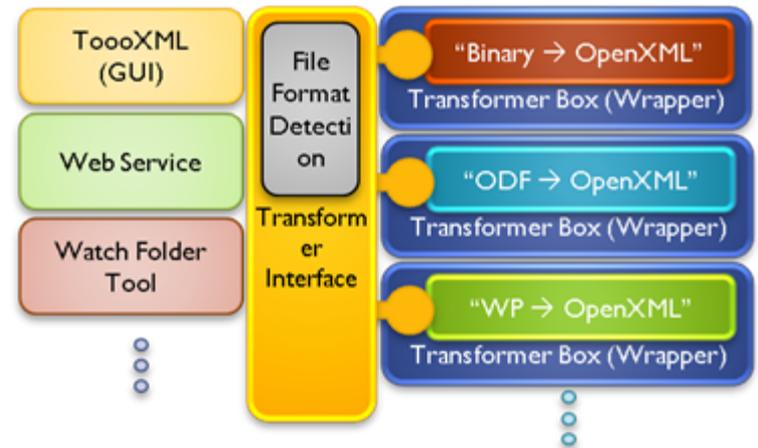
Organization

- High-profile EU Commission Project, €14M for 4 years
- Consortium of 5 national libraries, 4 national archives, 4 universities and 4 industry partners



Goals

- Preservation of Office Documents based on OpenXML
- Deliver converters for MS Office binary formats
- Funded open source project for ODF to/from OpenXML converter
- Deliver Preservation Toolkit



eScience and Semantic Computing meet the Cloud

The cyberinfrastructure for the next
generation of researchers



The Future: Software plus Services for Science?

- Expect scientific research environments will follow similar trends to the commercial sector
 - Leverage computing and data storage in the cloud
 - Scientists already experimenting with Amazon S3 and EC2 services, with mixed results;
- For many of the same reasons
 - Siloed research teams, no resource sharing across labs
 - High storage costs
 - Low resource utilization
 - Excess capacity
 - High costs of reliably keeping machines up-to-date
 - Little support for developers, system operators



A smart cyberinfrastructure

- Collective intelligence
 - If [last.fm](#) can recommend what song to broadcast to me based on what my friends are listening to, why cannot the cyberinfrastructure of the future recommend articles of potential interest based on what the experts in the field that I respect are reading?
 - Already examples emerging but the process is manual (Connotea, BioMedCentral Faculty of 1000 ...)
- Automatic correlation of scientific data
- Smart composition of services and functionality
- Cloud computing to aggregate, process, analyze and visualize data



Acknowledgements

- The ideas presented here were developed with input from many colleagues in the community and at Microsoft Research:
 - Thanks are due to David De Roure, Jeremy Frey, Carole Goble, Peter Murray-Rust, Alan Rector, Nigel Shadbolt and Alex Szalay
 - And special thanks to Roger Barga, Savas Parastatidis and Evelyne Viegas at Microsoft Research who have tried to educate me ...
- See www.microsoft.com/science for some more details of Microsoft's activities in Scientific and Technical Computing



Microsoft[®]

Your potential. Our passion.[™]

