Open Science
Open Data
Open Source

Tony Hey
Corporate Vice President
Microsoft Research

Microsoft Research Connections

Topics

The Fourth Paradigm

The Role of Open Source

Challenges and Opportunities of Open Data

The Emergence of Open Science

The Future of Data-Intensive Science

A Tidal Wave of Scientific Data

COMMUNICATIONS OF THE ACM

CACM.ACM.ORG

Surviving the Data Deluge

Open Information Extraction from the Web

The Economist

The data deluge

ieee spectrum

INSIDE TECHNOLOGY

AEROSPACE BIOMEDICAL COMPUTING CONSUMER ELECTRONICS ENERGY

AT WORK / INNOVATION

OPINION

The Coming Data Deluge

As science becomes more data intensive, so does our language

By PAUL MCFEDRIES / FEBRUARY 2011

Email Print Share

Illustration: gluekit

WIRED

SUBSCRIBE » SECTIONS » BLOGS » REVIEWS » VIDEO »

WIRED MAGAZINE: 16.07

SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

Illustration: Marian Bantjes

nature

THE BITER BIT
Viral infections for viruses

TROPICAL CYCLONES
The strong get stronger

BLACK HOLE PHYSICS
A new window on the Galactic Centre

BIG DATA

SCIENCE IN THE PETABYTE ERA

Emergence of a Fourth Research Paradigm

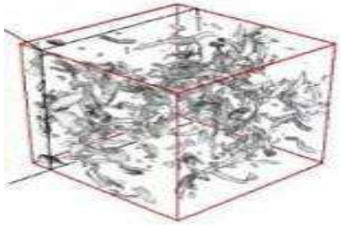
Thousand years ago – **Experimental Science**
- Description of natural phenomena

Last few hundred years – **Theoretical Science**
- Newton’s Laws, Maxwell’s Equations…

Last few decades – **Computational Science**
- Simulation of complex phenomena

Today – **Data-Intensive Science**
- Scientists overwhelmed with data sets from many different sources
  - Captured by instruments
  - Generated by simulations
  - Generated by sensor networks

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

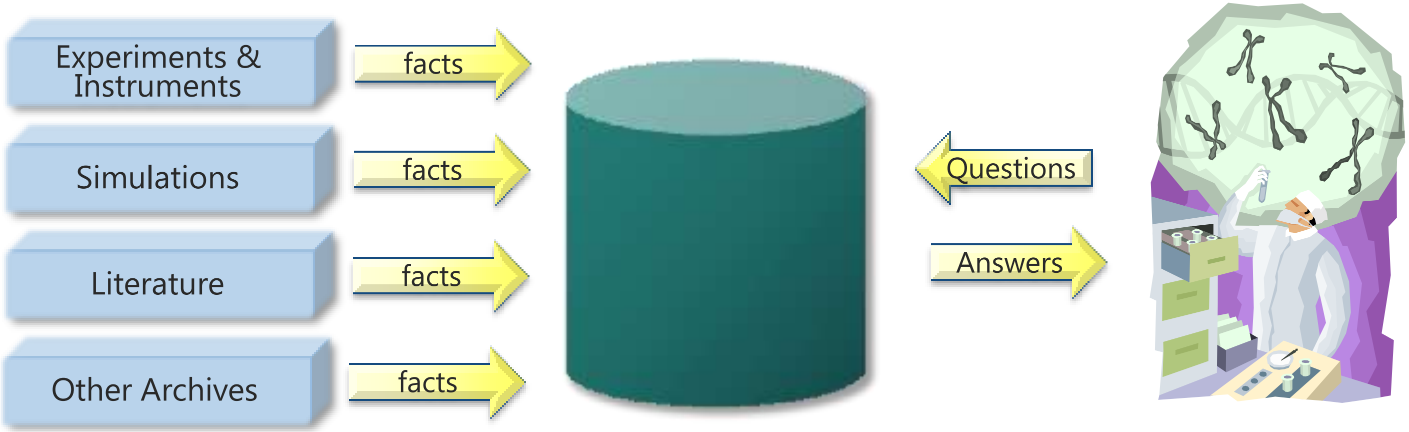


eScience is the set of tools and technologies to support data federation and collaboration
- For analysis and data mining
- For data visualization and exploration
- For scholarly communication and dissemination

(With thanks to Jim Gray)

X-Info

- The evolution of X-Info and Comp-X for each discipline X
- How to codify and represent our knowledge



Experiments & Instruments

Simulations

Literature

Other Archives

facts

facts

facts

facts

Questions

Answers

The Generic Problems

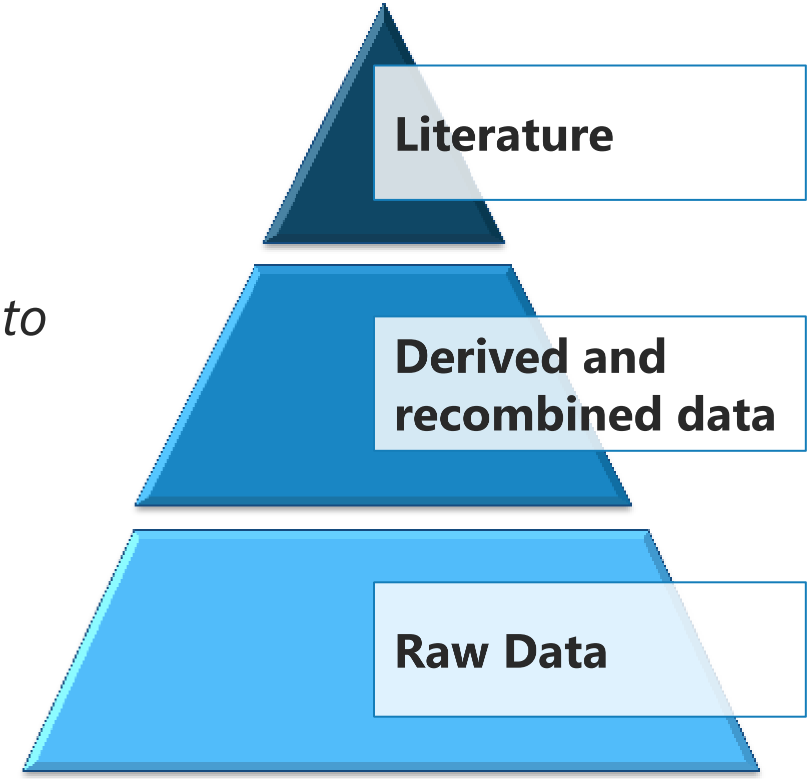
- Data ingest
- Managing a petabyte
- Common schema
- How to organize it
- How to *re*organize it
- How to share with others

- Query and Vis tools
- Building and executing models
- Integrating data and Literature
- Documenting experiments
- Curation and long-term preservation

(With thanks to Jim Gray)

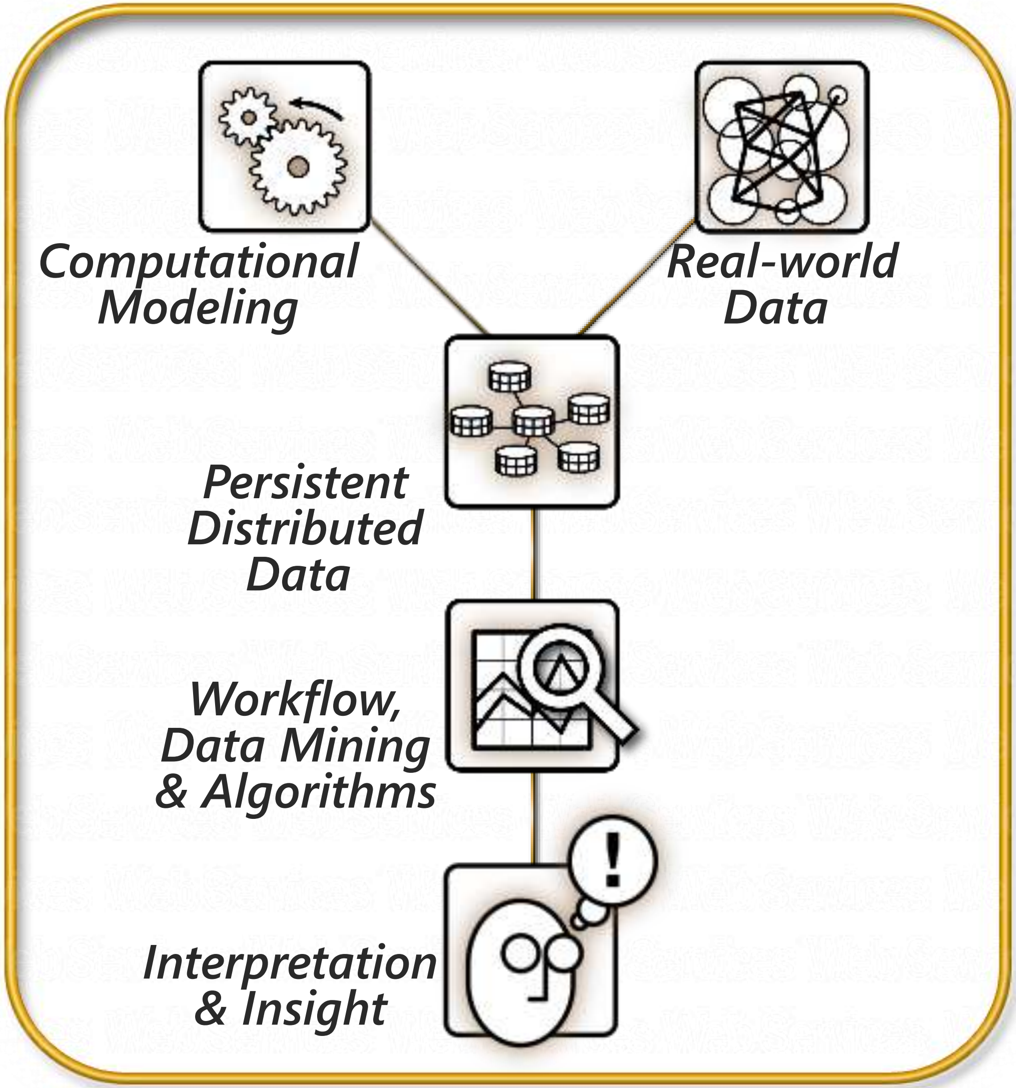
All Scientific Data Online

- Many disciplines overlap and use data from other sciences.
- Internet can unify all literature and data
- Go from literature *to* computation *to* data *back to* literature.
- Information at your fingertips – For everyone, everywhere
- Increase Scientific Information Velocity
- Huge increase in Science Productivity



(From Jim Gray's last talk)

Reduced Time to Insight


Computational Modeling
Real-world Data
Persistent Distributed Data
Workflow, Data Mining & Algorithms
Interpretation & Insight

(Thanks to Craig Mundie)

Topics

The Fourth Paradigm

The Role of Open Source

Challenges and Opportunities of Open Data

The Emergence of Open Science

The Future of Data-Intensive Science

Open Source Tools for Science

- Researchers often need to develop open source tools in the absence of commercial alternatives
- Funding Agencies often require that software developed with public funding be released as open source
- Important for researchers to understand choice of open source software license determines commercialization possibilities
- Most open source projects in SourceForge are inactive and have failed to grow a ‘critical mass’ of users and developers

open source initiative™

Example: UK JISC Funding Agency Report on Open Source Licenses (2006)

At the moment there are more than 50 OSI certified open source licenses. The following five are perhaps the most commonly used:

  - The GNU General Public License (GPL)
  - The GNU Lesser General Public License (LGPL)
  - Modified BSD (Berkeley Software Distribution) License (new BSD)
  - Apache License
  - Mozilla Public License (MPL)

The difference between them is the extent to which they control the way the code can be combined with other software.

- At the one extreme, the BSD license permits open source software to be merged with closed-source code and then sold under a conventional license.
- At the other, the GPL license insists that if the software is combined with other code then that too must be under a GPL license.

Example: US National Institutes of Health (NCI) Summary of OSS Licenses

Open Source Licensing Principles: Three major models exist:

GNU Public License (GPL)

- Developed for use with UNIX operating system
- Open referred to as “copyleft”: Work can be freely distributed under the same licensing terms as the original
- Often called the “viral” open source license

Lesser General Public License (LGPL)

- For use with data libraries and other collections
- Applies to the program itself, but not to linking programs

Berkeley Software Development (BSD)

- Developed for Berkeley UNIX
- Is non-viral: Derivative works not subject to the original open source terms
- May therefore be more attractive for commercial use

OUTERcurve FOUNDATION

Enable the exchange of code and understanding among software companies and open source communities.

> *"Whatever the future holds for Kinect, Microsoft has (over the last 18 months at least) open sourced most of its community developed projects and technologies via the Outercurve Foundation — the not-for-profit software IP management and project development organization."*
>
> *Adrian Bridgwater*
> *Dr. Dobbs*
> *April 25, 2011*

Outercurve Foundation and Open Source

The Museum As A Metaphor

- Sponsors create "Galleries" based on technology or industry themes
- Gallery Managers and the Foundation encourage project assignments into Galleries
- Individual Projects are complementary with the theme of the Gallery

Research Accelerators Gallery

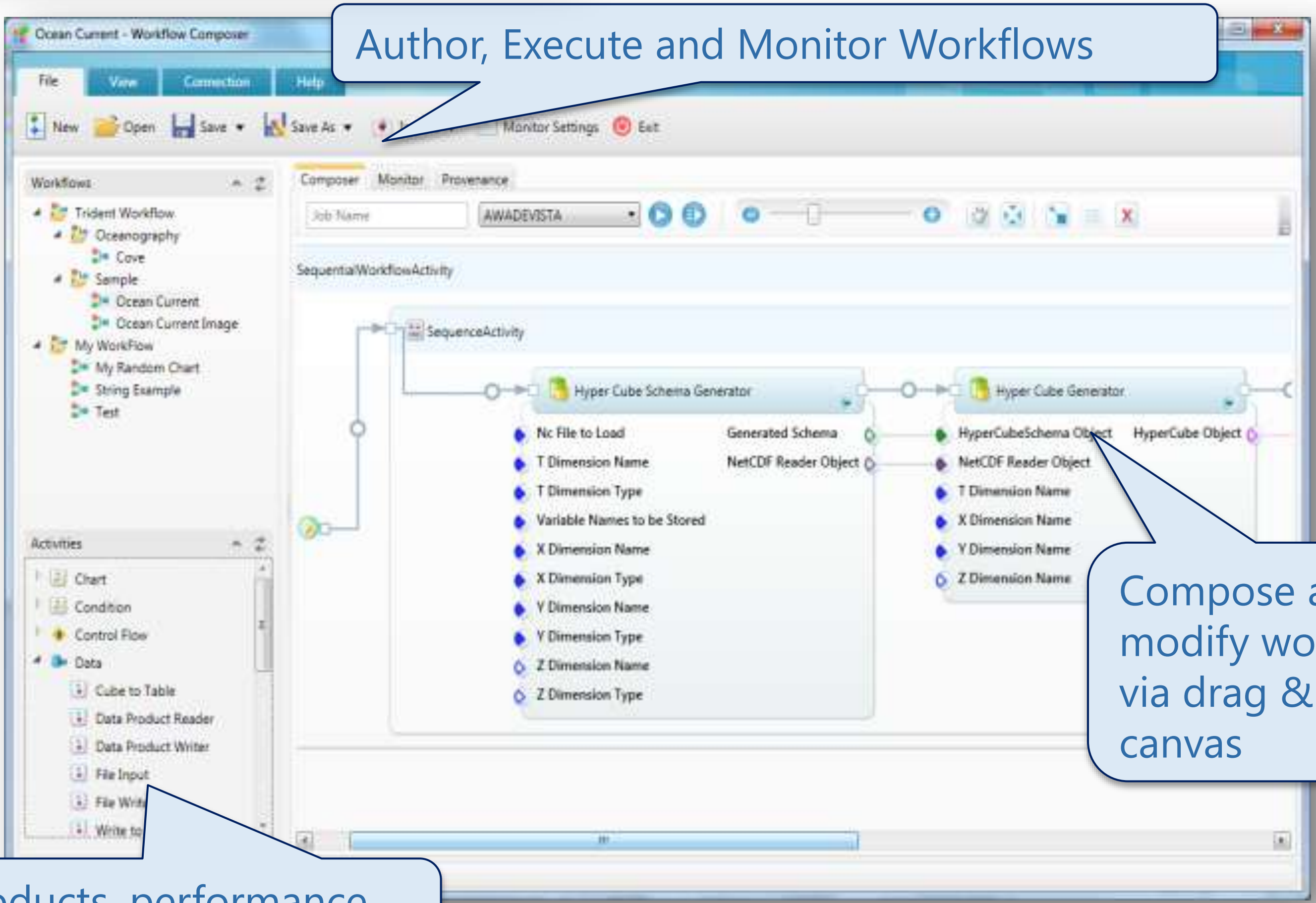
Project Trident: Toolset based on Windows Workflow Foundation that provides scientists' need for a flexible, powerful way to analyze large, diverse datasets.

Chemistry Add-in for Word: Chem4Word is an add-in for Microsoft Word that enables semantic authoring of chemical structures.

ConferenceXP: Platform for real-time collaboration that seamlessly connects people or groups over a network, providing high-quality, low-latency videoconferencing and a rich set of collaboration capabilities.

Project Trident – Scientific Workflow Workbench

Microsoft® .NET | Windows Workflow Foundation



Author, Execute and Monitor Workflows

Compose and modify workflows via drag & drop canvas

View data products, performance metrics, and provenance data

http://tridentworkflow.codeplex.com/

Microsoft Research Connections

Chem4Word– Chemical Drawing in Word

Semantic chemistry for students and publishers

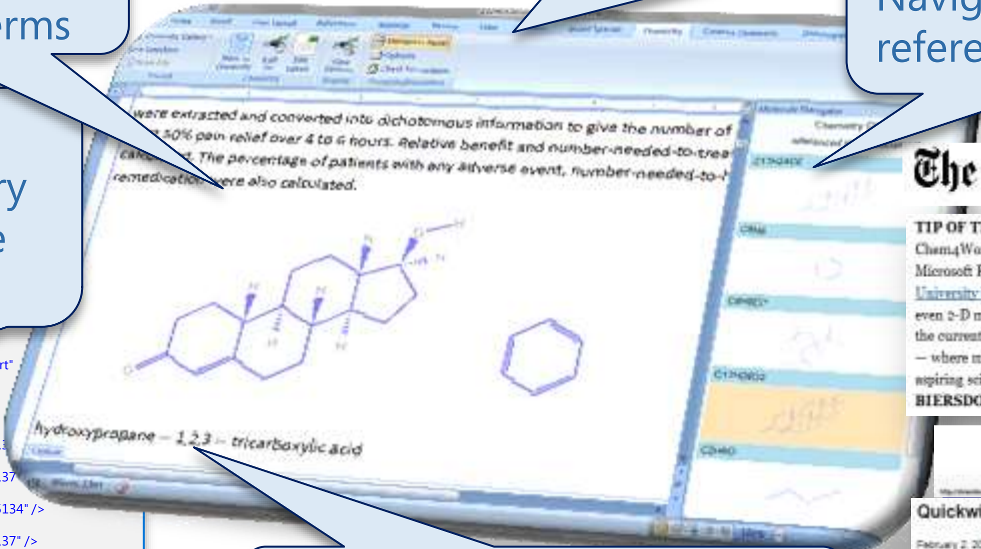
UNIVERSITY OF CAMBRIDGE

Intent: Recognizes chemical dictionary and ontology terms

Author/edit 1D and 2D chemistry. Change chemical layout styles.

Relationships: Navigate and link referenced chemistry

Data: Semantics stored in Chemistry Markup Language (CML)



were extracted and converted into dichotomous information to give the number of … 50% pain relief over 4 to 6 hours. Relative benefit and number-needed-to-trea… calc… The percentage of patients with any adverse event, number-needed-to-h… remedication were also calculated.

hydroxypropane – 1,2,3 – tricarboxylic acid

```
<?xml version="1.0" ?>
<cml version="3" conver   ="org-synth-report"
xmlns="http://www.xml-cml.org/schema">
  <molecule id="m1">
    <atomArray>
      <atom id="a1" elementType="C" x2="-
2.9149999618530273" y2="0.769999980926513
      <atom id="a2" elementType="C" x2="-
1.5813208400249916" y2="1.5399999809265137
      <atom id="a3" elementType="O" x2="-
0.24764171819695613" y2="0.7699999809265134" />
      <atom id="a4" elementType="O" x2="-
1.5813208400249912" y2="3.0799999809265137" />
      <atom id="a5" elementType="H" x2="-
4.248679083681063" y2="1.5399999809265137" />
      <atom id="a6" elementType="H" x2="-
2.914999961853028" y2="-0.7700000190734864" />
      <atom id="a7" elementType="H" x2="-
4.248679083681063" y2="-1.907348645691087E-8" />
      <atom id="a8" elementType="H"
x2="1.0860374036310796" y2="1.5399999809265132" />
    </atomArray>
    <bondArray>
      <bond atomRefs2="a1 a2" order="1" />
      <bond atomRefs2="a2 a3" order="1" />
      <bond atomRefs2="a2 a4" order="2" />
      <bond atomRefs2="a1 a5" order="1" />
      <bond atomRefs2="a1 a6" order="1" />
      <bond atomRefs2="a1 a7" order="1" />
      <bond atomRefs2="a3 a8" order="1" />
    </bondArray>
  </molecule>
</cml>
```

Intelligence: Verifies validity of authored chemistry

The New York Times Personal Tech

TIP OF THE WEEK Chemistry students and teachers might want to check out the new Chem4Word add-on for Microsoft Word. The free software, which was developed by Microsoft Research and the Unilever Centre for Molecular Science Informatics at the University of Cambridge, allows Word users to insert chemical symbols, formulas and even 2-D models of molecules into documents. Chem4Word works with Word 2007 and the current beta version of Word 2010, and is listed as a beta version itself at bit.ly/r1K33 — where more information and a demonstration video are also available for scientists, aspiring scientists and those who have chemistry papers due soon. **J. D. BIERSDORFER**

THE CHRONICLE of Higher Education. Wired Campus

Quickwire: Microsoft Word Goes Chemical

February 2, 2011, 2:50 pm

By Josh Fischman

Chem4Word, a free, open-source plug-in that lets authors draw intricate chemical structures—and store information about molecules—within their Word documents, has been released by Microsoft Research (the company's unit that collaborates with universities), the University of Cambridge, and the Outercurve Foundation.

http://chronicle.com/blogs/wiredcampus/quickwire-microsoft-word-goes-chemical/29423

http://research.microsoft.com/chem4word/

Biodex (MBF):
An Open Source Bioinformatics Library for .NET

- Simplifies the creation of bioinformatics applications on the Microsoft platform
  - Consists of file parsers and writers, algorithms and webservice connectors
- Focuses on the assembly, manipulation and comparison of next-generation DNA sequencing data
  - Project is steered by a Technical Advisory Board including commercial and academic users
- Ownership is being transferred to the Outercurve Foundation
  - Version 1.0 is already available under the MS-PL license
  - Version 2.0 will be released in July 2011 under the Apache 2.0 license

http://research.microsoft.com/bio

Example Project: Increasing energy yield of sugar cane through genome assembly

- Sugar cane energy yield is 6x that of corn
- Yield increases predicted to be as much as 3x
- Sugar-cane genome needed to achieve such increases
- Working with researchers in Brazil to assemble the genome
- Basic idea: Leverage genome of sorghum, which is similar and known
- Algorithms to be integrated into Biodex (MBF)

Topics

The Fourth Paradigm

The Role of Open Source

Challenges and Opportunities of Open Data

The Emergence of Open Science

The Future of Data-Intensive Science

Example: Sloan Digital Sky Survey

"The Cosmic Genome Project"

- Two surveys in one
  - Photometric survey in 5 bands
  - Spectroscopic redshift survey
- Data is public
  - 2.5 Terapixels of images
  - 40 TB of raw data => 120TB processed
  - 5 TB catalogs => 35TB in the end
- Started in 1992, finished in 2008
- Database and spectrograph built at JHU (SkyServer)

The University of Chicago
Princeton University
The Johns Hopkins University
The University of Washington
New Mexico State University
Fermi National Accelerator Laboratory
US Naval Observatory
The Japanese Participation Group
The Institute for Advanced Study
Max Planck Inst, Heidelberg

Sloan Foundation, NSF, DOE, NASA

Open Data: Public Use of the Sloan Data

Posterchild in 21st century data publishing

- Set up SkyServer web service
- 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
- New publishing paradigm: data
  published before analysis
  by astronomers

Welcome!!!

This website presents data from the Sloan Digital Sky Survey, a project to make a map of a large part of the universe. We would like to show you the beauty of the universe, and share with you our excitement as we build the largest map in the history of the world.

News

The site contains new data from the SDSS Data Release 1 (DR1)

More...

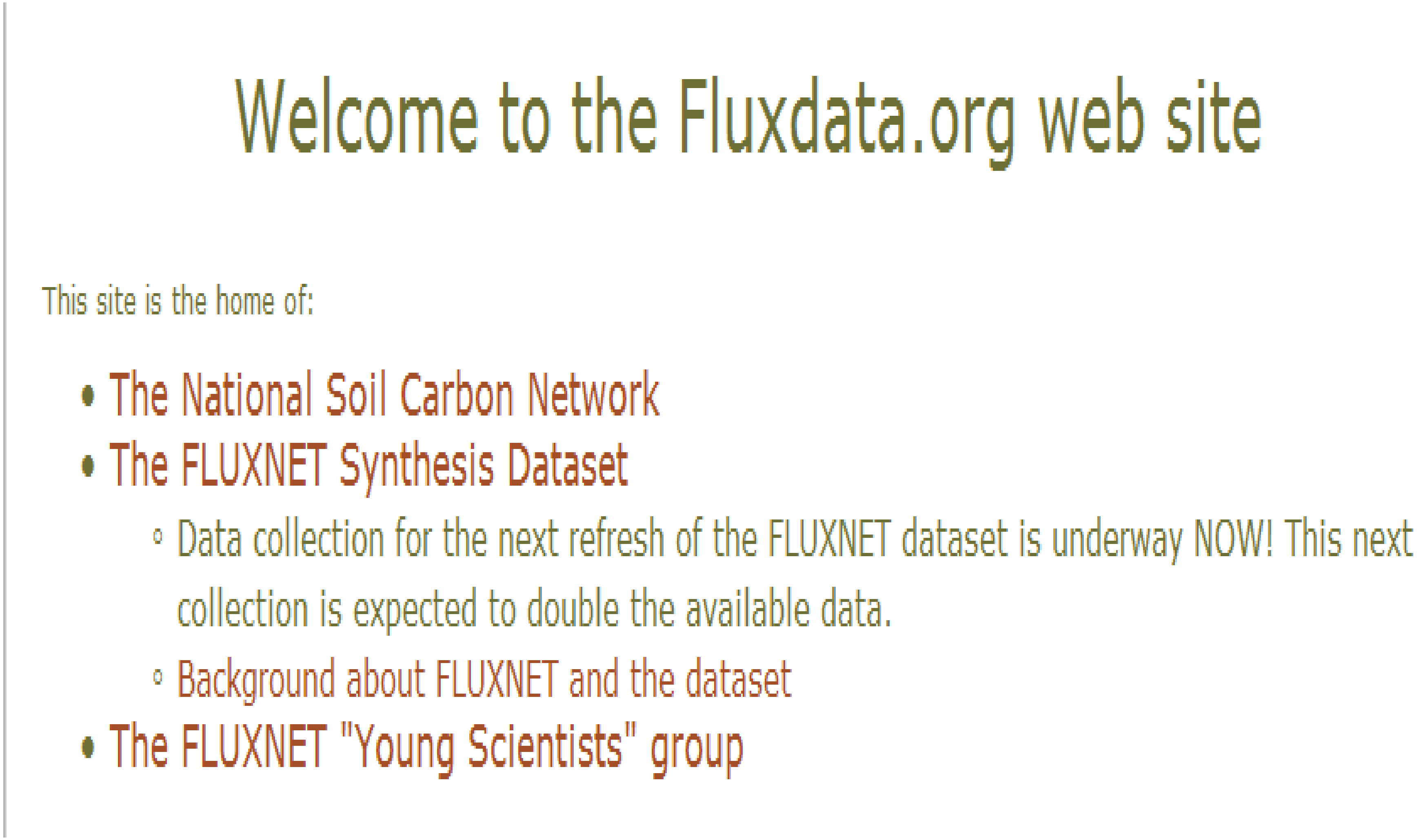
For Astronomers

A separate branch of the website for professional astronomers (English)

More...

http://cas.sdss.org/dr7/en/

An Environmental DataServer

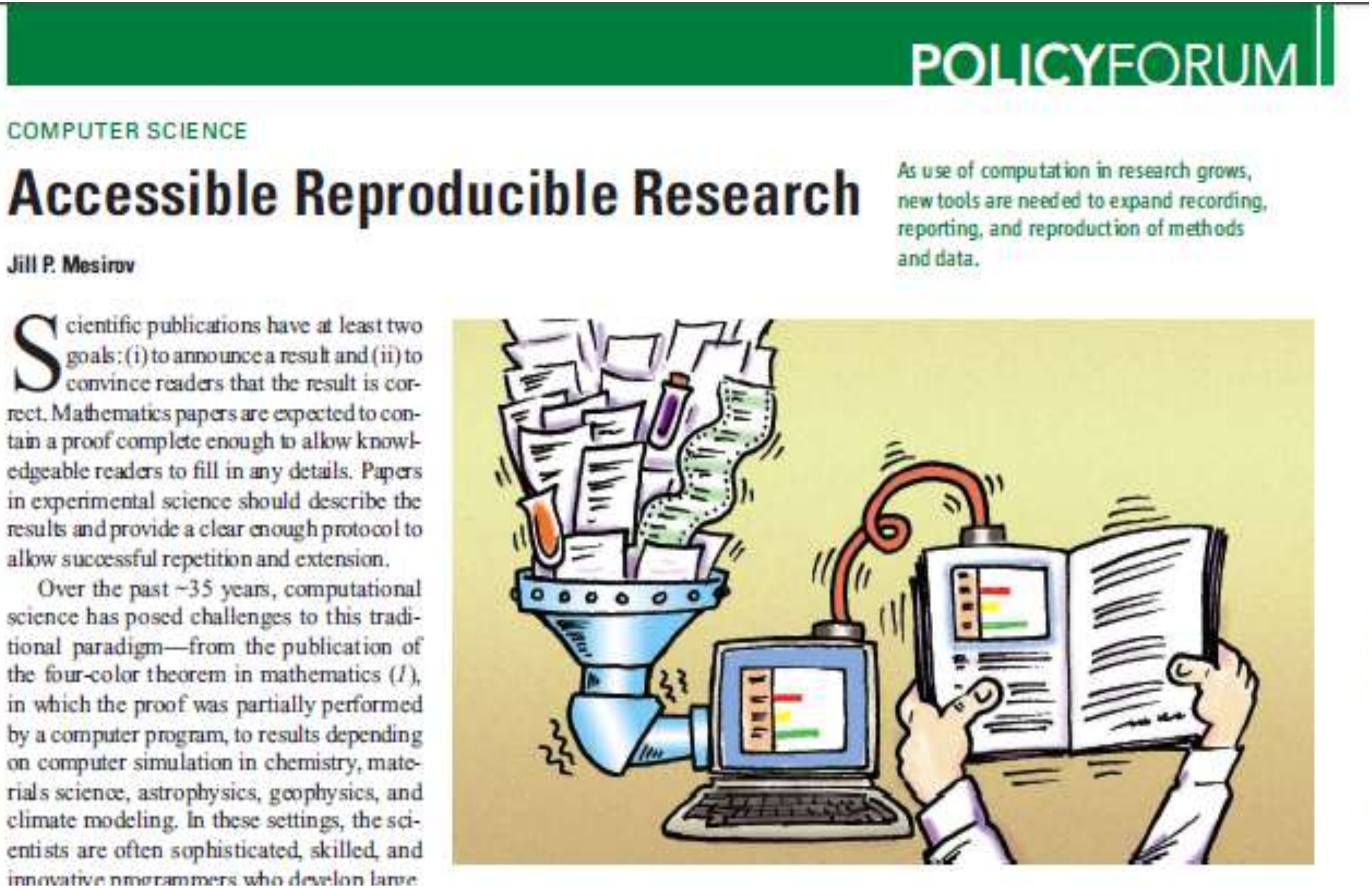


Welcome to the Fluxdata.org web site

This site is the home of:

- The National Soil Carbon Network
- The FLUXNET Synthesis Dataset
  - Data collection for the next refresh of the FLUXNET dataset is underway NOW! This next collection is expected to double the available data.
  - Background about FLUXNET and the dataset
- The FLUXNET "Young Scientists" group

Science November 9th 2010



POLICYFORUM

COMPUTER SCIENCE

Accessible Reproducible Research

Jill P. Mesirov

As use of computation in research grows, new tools are needed to expand recording, reporting, and reproduction of methods and data.

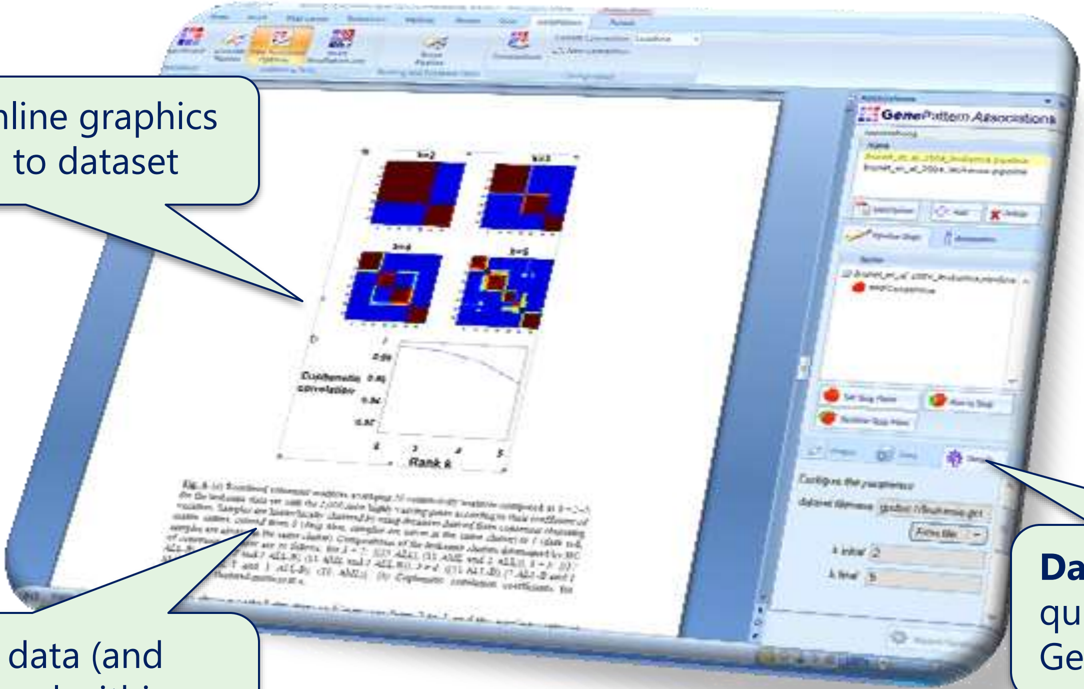
Scientific publications have at least two goals: (i) to announce a result and (ii) to convince readers that the result is correct. Mathematics papers are expected to contain a proof complete enough to allow knowledgeable readers to fill in any details. Papers in experimental science should describe the results and provide a clear enough protocol to allow successful repetition and extension.

Over the past ~35 years, computational science has posed challenges to this traditional paradigm—from the publication of the four-color theorem in mathematics (1), in which the proof was partially performed by a computer program, to results depending on computer simulation in chemistry, materials science, astrophysics, geophysics, and climate modeling. In these settings, the scientists are often sophisticated, skilled, and innovative programmers who develop large

GenePattern Reproducible Research Add-in

BROAD INSTITUTE

Services: Connects to GenePattern database



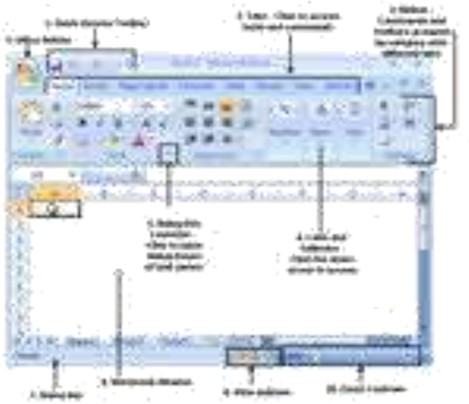
Relationships: Inline graphics are synchronized to dataset

Data: Control and execute query pipelines into GenePattern

Data: Resulting data (and provenance) stored within Word document

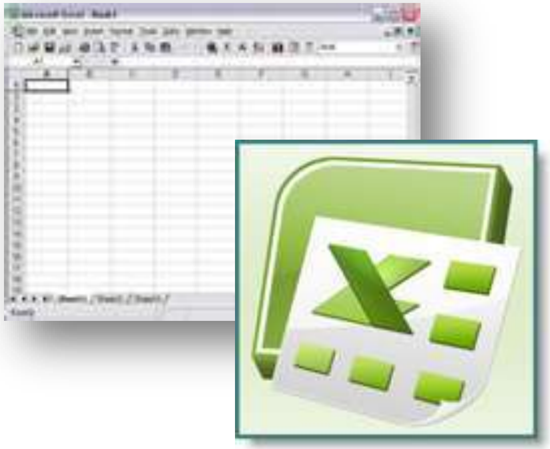
http://GenepatternWordAddin.codeplex.com

Data Curation Add-in for Microsoft Excel

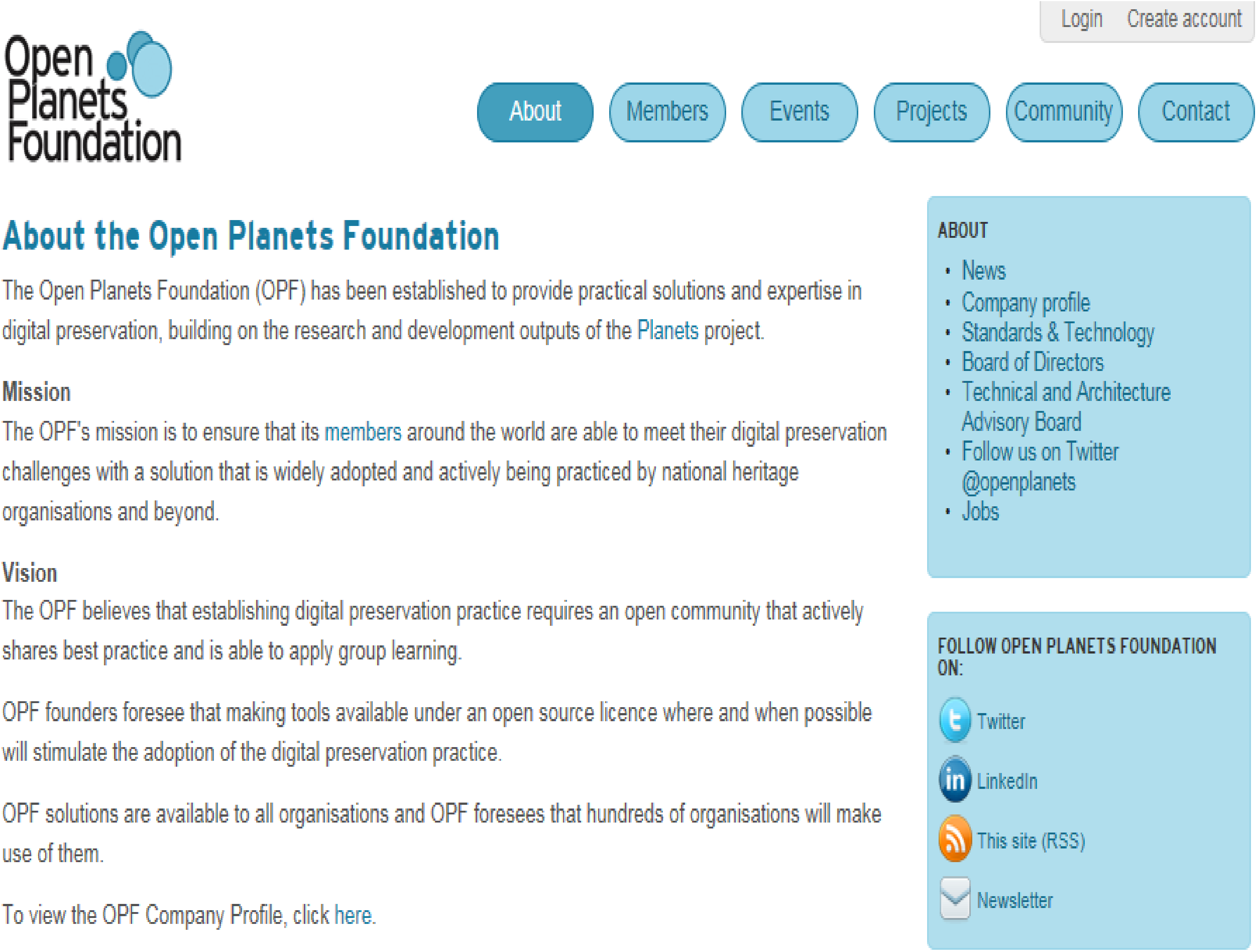


University of California CDL California Digital Library

DataONE



- **Microsoft Research, in partnership with California Digital Library's Curation Center**
  - Collaboration with **Trisha Cruse & John Kunze**
  - Part of the **DataONE** (an NSF DataNet Project)
- **Proposed functionality *under consideration*:**
  - **Support for versioning**, so that revision history and the original raw data can be easily protected and recovered,
  - **Standardized date/time stamps** so that researchers can easily determine when the data were created and last updated.
  - **A "workbook builder"** allowing researchers to select from globally shared standardized layouts for capturing data,
  - **Ability to export metadata in a standard format** (e.g., a DataCite citation or an EML document that describes the dataset(s) in a workbook) so that researchers can readily share their data,
  - **Ability to select from a globally shared vocabulary of terms for data descriptions** (e.g., column names), and as needed to add new terms to the globally shared vocabulary, to enable wide collaboration between researchers
  - **Ability to import term descriptions from the shared vocabulary and annotate them locally** to refine their definitions as used in the dataset,
  - **"Speed bumps" to discourage use of macros and customizations** that would impede interoperation of data imported from Excel into other applications, and
  - **Ability to deposit data and metadata directly into a data archive** to enable compliance with funding agency requirements to preserve and publish research data.



About the Open Planets Foundation

The Open Planets Foundation (OPF) has been established to provide practical solutions and expertise in digital preservation, building on the research and development outputs of the Planets project.

Mission

The OPF's mission is to ensure that its members around the world are able to meet their digital preservation challenges with a solution that is widely adopted and actively being practiced by national heritage organisations and beyond.

Vision

The OPF believes that establishing digital preservation practice requires an open community that actively shares best practice and is able to apply group learning.

OPF founders foresee that making tools available under an open source licence where and when possible will stimulate the adoption of the digital preservation practice.

OPF solutions are available to all organisations and OPF foresees that hundreds of organisations will make use of them.

To view the OPF Company Profile, click here.

http://www.openplanetsfoundation.org/

Open Data Award

BioMed Central
The Open Access Publisher

Data sharing is an increasingly important part of the research and publication process. But there are many challenges associated with openly sharing scientific data, particularly when sharing goes against cultural or community norms.

The Open Data Award (sponsored by Microsoft Research) recognizes researchers who have demonstrated leadership in the sharing, standardization, publication, or re-use of biomedical research data.

Excellence in Open Access Research
http://www.biomedcentral.com/researchawards/

Science is based on building on, reusing and openly criticising the published body of scientific knowledge.

For science to effectively function, and for society to reap the full benefits from scientific endeavours, it is crucial that science data be made open.

The Panton Principles
Principles for Open Data in Science
http://pantonprinciples.org/

Dr. Peter Murray-Rust; Jean-Luc Bouvé, accepting the Open Data Award on behalf of Dr. Tommi Nyman; and Alex Wade

This year's award recognized biologist Tommi Nyman from Finland for the article, "How common is ecological speciation in plant-feeding insects? A 'Higher' Nematinae perspective," published in the open-access journal, *BMC Evolutionary Biology*.

The data are well labelled and readily understandable by other scientists; moreover, the authors showed great transparency in their work, particularly in their first additional data file, which fully documents how they sampled their insects. This level of openness is not commonly seen and it demonstrates real leadership.

5th Annual
BioMed Central
5th Annual Research Awards

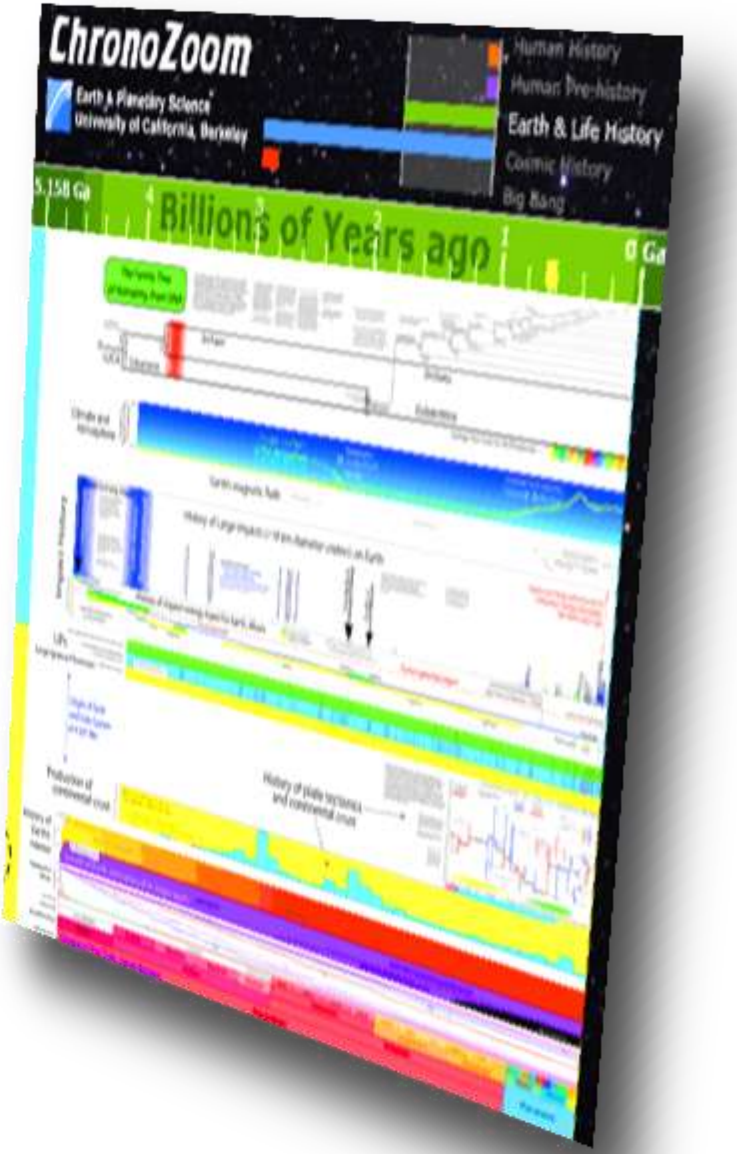
Microsoft Research

ChronoZoom – History in its broadest possible context …

The challenge: exploration of all known time series data with the ability to smoothly transition from billions of years down to individual nanoseconds…

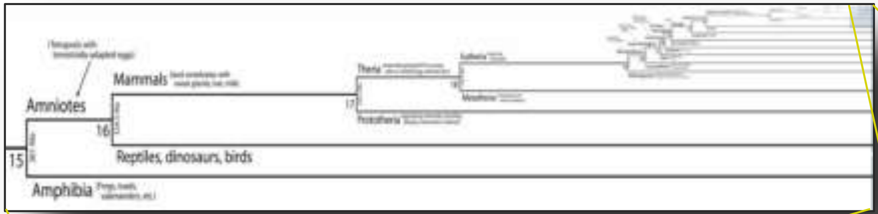
This is what Walter Alvarez, Professor of Earth and Planetary Science at University of Berkeley set out to do.

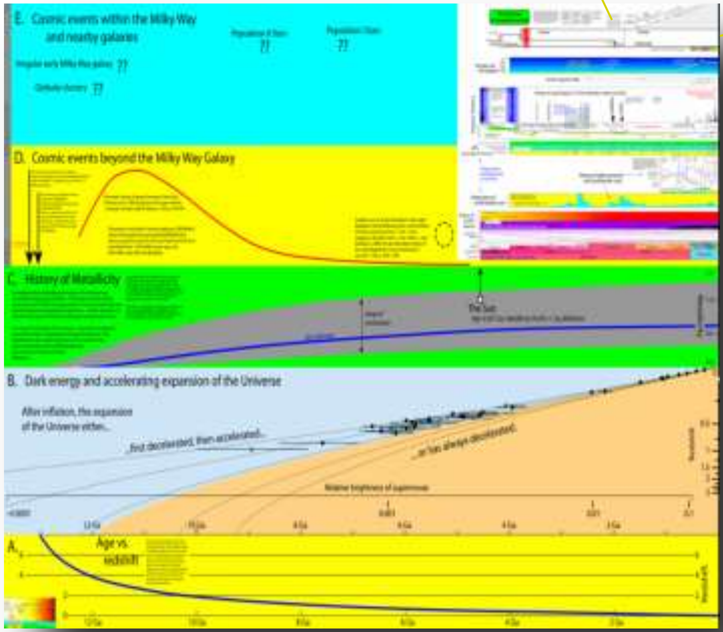
Our vision is to create an application that allows researchers to browse, overlay, and explore interdisciplinary data sources.

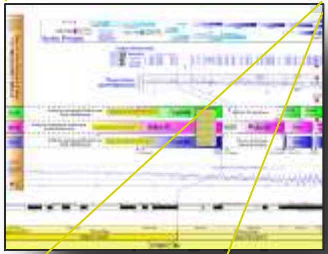


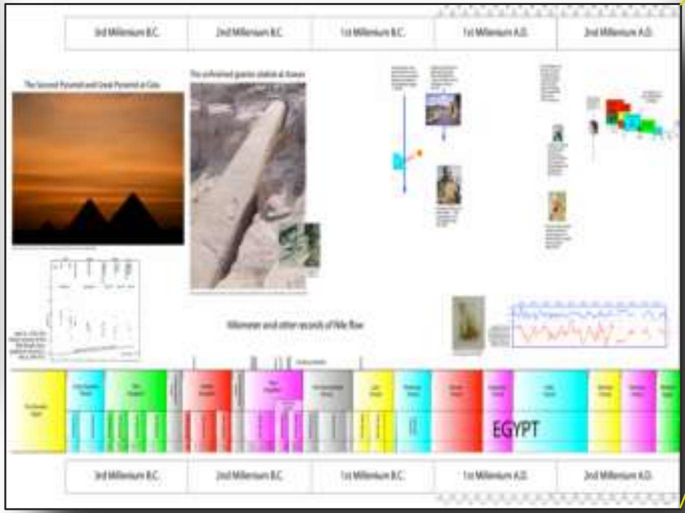
www.chronozoomtimescale.org

‘Big History’




E. Cosmic events within the Milky Way and nearby galaxies
D. Cosmic events beyond the Milky Way Galaxy
C. History of Metallicity
The Sun
B. Dark energy and accelerating expansion of the Universe
After inflation, the expansion of the Universe either...
...first decelerated, then accelerated...
...or has always decelerated
A. Age vs. redshift





See the demo live at
www.chronozoomtimescale.org

Zoom Technology for Big History

http://www.youtube.com/watch?v=V2q-U02HLb0

Topics

The Fourth Paradigm

The Role of Open Source

Challenges and Opportunities of Open Data

The Emergence of Open Science

Future of Data-Intensive Science

Rapid Data Sharing for Alzheimer Biomarkers

ADNI Alzheimer's Disease Neuroimaging Initiative

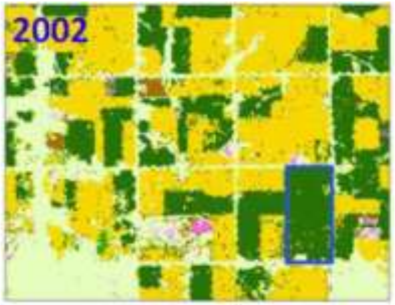
- Alzheimer's Disease Neuroimaging Initiative (ADNI) launched in 2004 specifically to improve clinical trials by different centers agreeing to share data.
- Not only can the data fro the 14 different centers involved in the initiative be combined and compared, but the data is typically made publicly available within a week of being collected.
- Hundreds of scientists have made tens of thousands of downloads from the ADNI website.
- Of several dozen papers that have so far been published using ADNI data, a significant number were authored by researchers who are not even directly funded by the project.

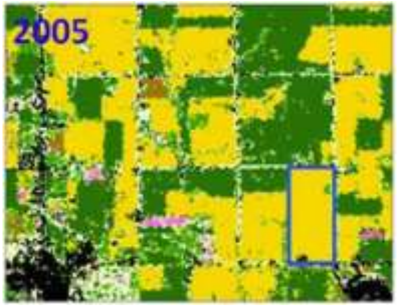
http://www.adni-info.org/

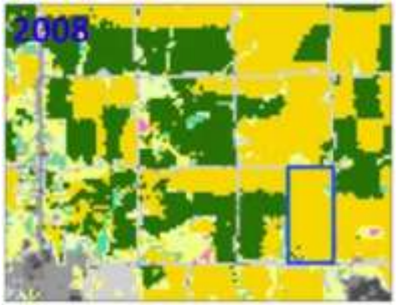
Satellite Data providing Value Of Information

Scientists at the U.S. Geological Survey (USGS)

- Developing an economic framework to measure what they call the "VOI" or **Value Of Information**
- Using storehouse of Land Use / Land Cover maps created from Landsat's moderate resolution land imagery since the early 1970s.







USGS is aiming for a VOI calculation that can inform decisions that maximize agricultural production by:

- Reconciling groundwater pollution hazards with the region's agricultural needs
- Thereby lowering mitigation and treatment costs necessary to avoid human health and other consequences of contaminated groundwater.

USGS
science for a changing world

ftp://ftpext.usgs.gov/

Funding Data Storage, Curation and Analysis

RUTGERS

JERSEY ROOTS, GLOBAL REACH

COASTAL OCEAN OBSERVATION LAB

About RU COOL Contact RU COOL Visiting RU COOL DOE Meeting

Historically, after a boating or aircraft accident at sea, the U.S. Coast Guard historically has relied on current charts and wind gauges to figure out where to hunt for survivors.

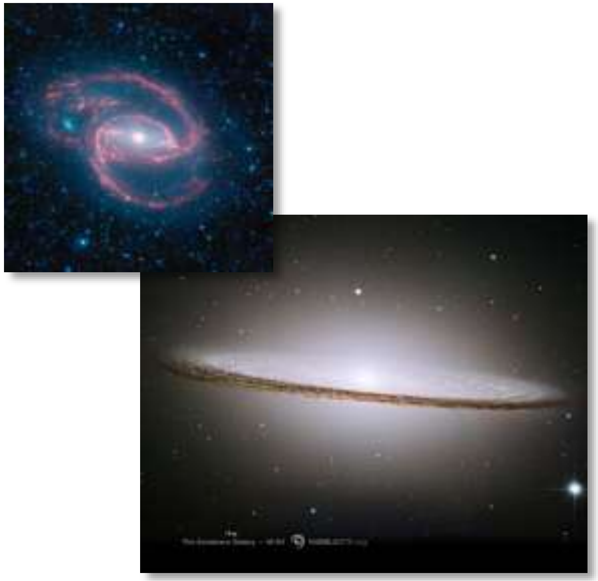
Scientists have been collecting high frequency radar data that can remotely measure ocean surface waves and currents – it is now available to the USCG for rescue operations.

However, a large fraction of the data the Rutgers team collects has to be thrown out because there is no room to store it and no support within existing research projects to better curate and manage the data. **"I can get funding to put equipment into the ocean, but not to analyze that data on the back end,"**

Professor Oscar Schofield
Bio-Optical Oceanography

Citizen Scientists and Data Analysis

Galaxy Zoo activities give a useful indication of the latent appetite for scientific engagement in society. This is a collection of online astronomy projects which invite members of the public to assist in classifying galaxies.

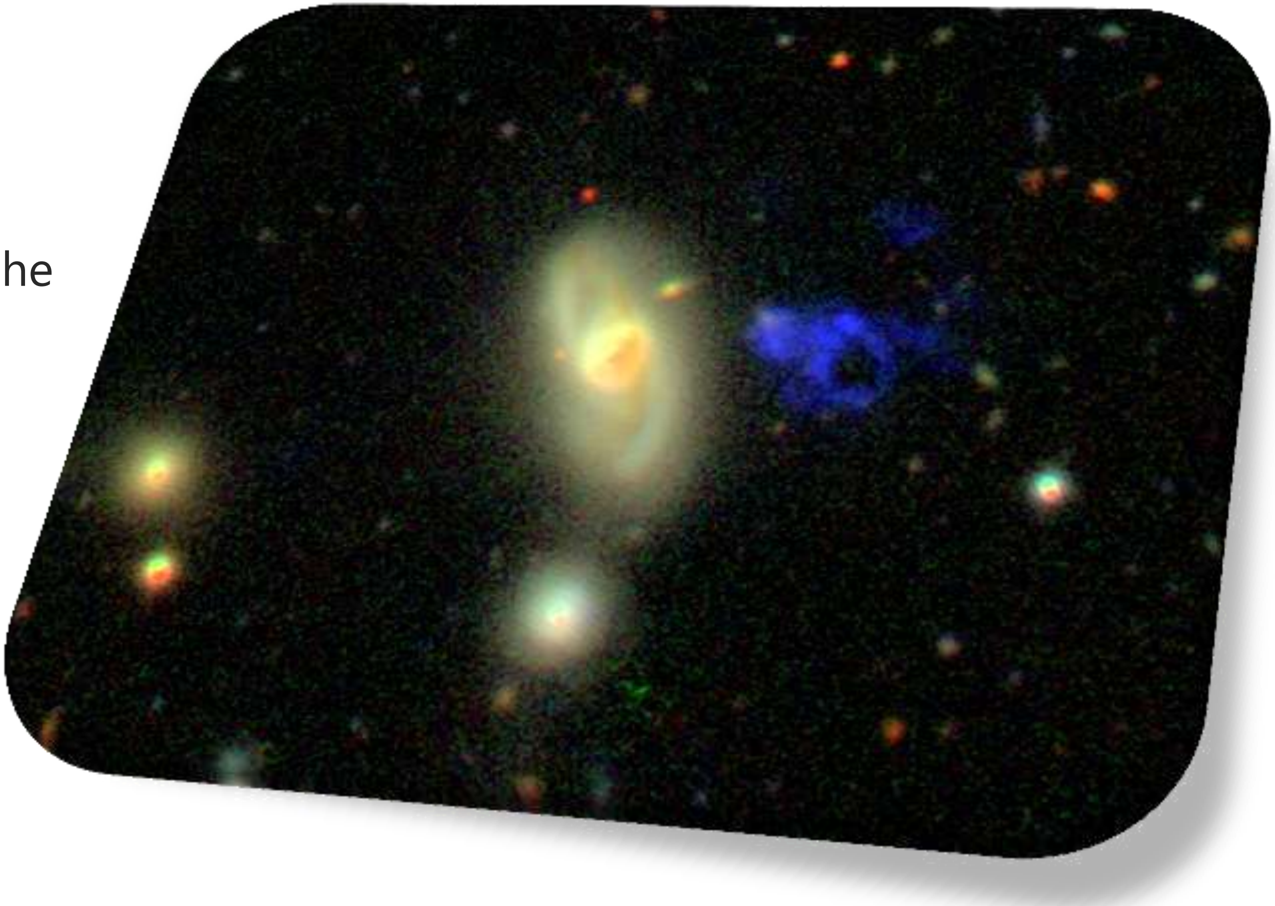


In the first year, **50 million classifications were made by 150,000 individuals in the general public** – it quickly became the world's largest database of galaxy shapes. The original project that it spawned Galaxy Zoo 2 in February 2009 to classify another 250,000 SDSS galaxies. The project included unique scientific discoveries such as Hanny's Voorwerp and 'Green Pea' galaxies.

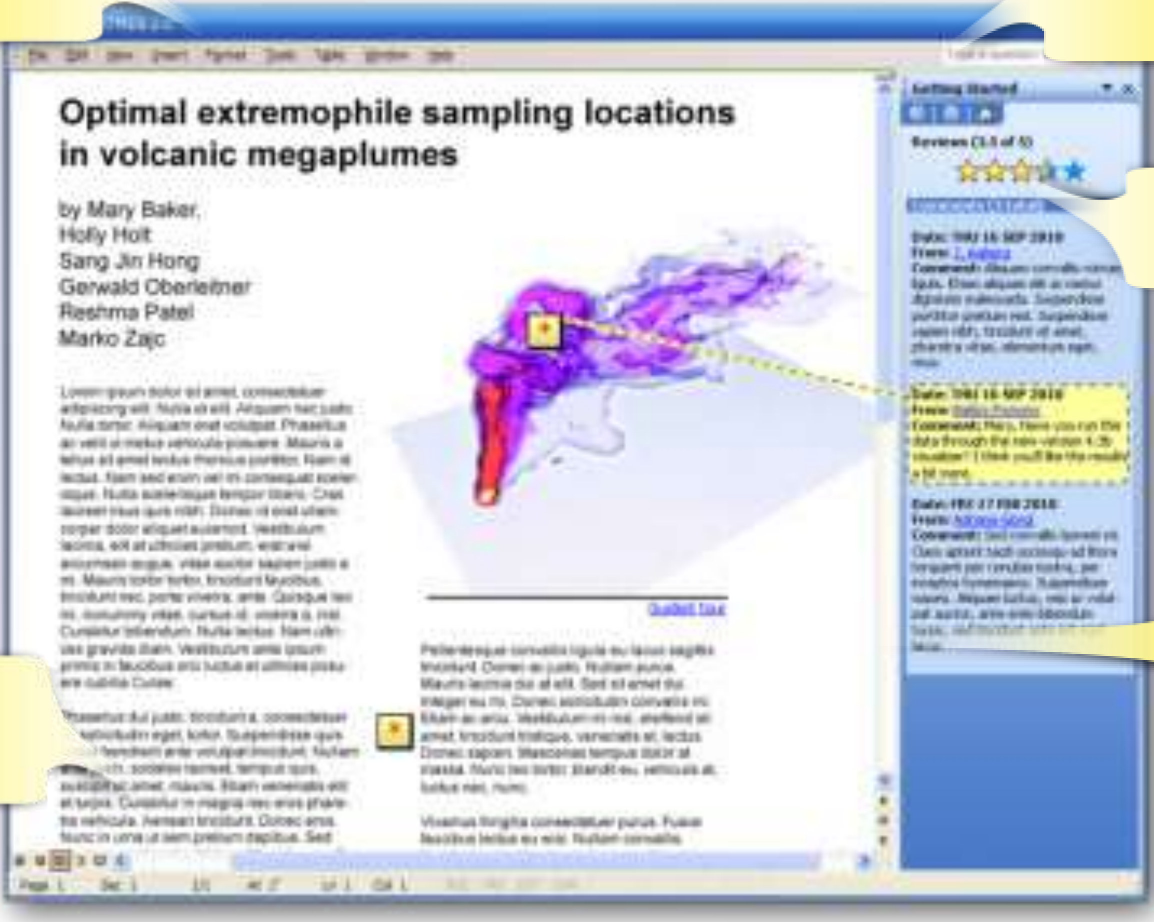
GALAXY ZOO.org

Hanny van Arkle’s Voorwerp

Hanny Van Arkel, a Dutch schoolteacher and Galaxy Zoo volunteer, posted an image to the Galaxy Zoo forum and asked "What's the blue stuff below?" No one knew. The object became known as the "**Voorwerp**“, Dutch for "object".



Envisioning a New Era of Research Reporting



Optimal extremophile sampling locations in volcanic megaplumes

by Mary Baker,
Holly Holt
Sang Jin Hong
Gerwald Oberleitner
Reshma Patel
Marko Žajc

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. [illegible]

Reproducible Research

Collaboration

Reputation & Influence

Dynamic Documents

Interactive Data

(Thanks to Craig Mundie)

Microsoft Research Connections

Datacite and ORCID

DataCite

- International consortium to establish easier access to scientific research data
- Increase acceptance of research data as legitimate, citable contributions to the scientific record
- Support data archiving that will permit results to be verified and re-purposed for future study.

ORCID - Open Research & Contributor ID

- Aims to solve the author/contributor name ambiguity problem in scholarly communications
- Central registry of unique identifiers for individual researchers
- Open and transparent linking mechanism between ORCID and other current author ID schemes.
- Identifiers can be linked to the researcher’s output to enhance the scientific discovery process

The Future of Research Repositories?

- Repositories will contain not only full text versions of research papers but also ‘grey’ literature such as workshop papers, presentations, technical reports and theses
- In the future repositories will also contain data, images and software
- Need for both centralized and federated databases of scientific information and cross database search tools
  - Centralized: NIH National Library of Medicine
  - Federated: WorldWideScience.org

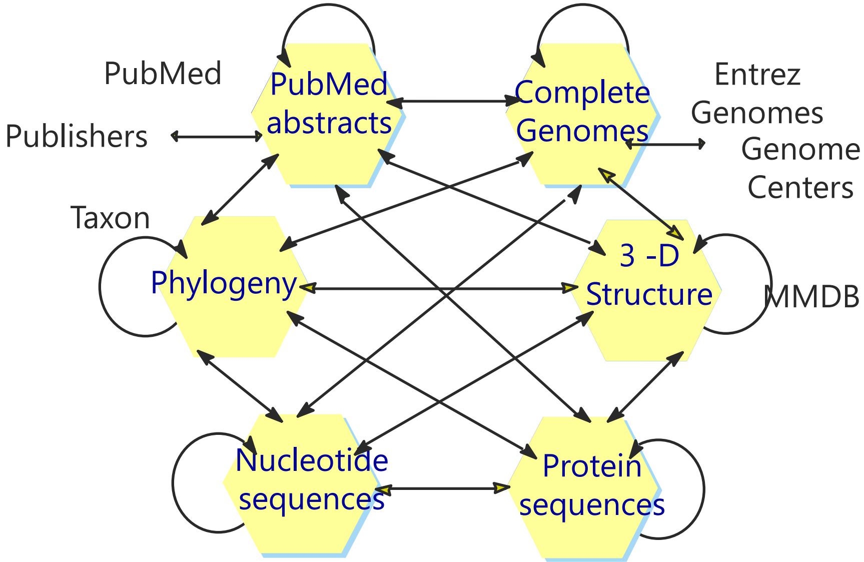
The US National Library of Medicine and PubMed Central

- The NIH Public Access Policy ensures that the public has access to the published results of NIH funded research.

- It requires scientists to submit final peer-reviewed journal manuscripts that arise from NIH funds to the digital archive PubMed Central

- To help advance science and improve human health, the Policy requires that these papers are accessible to the public on PubMed Central no later than 12 months after publication.

PubMed Central



Entrez cross-database search

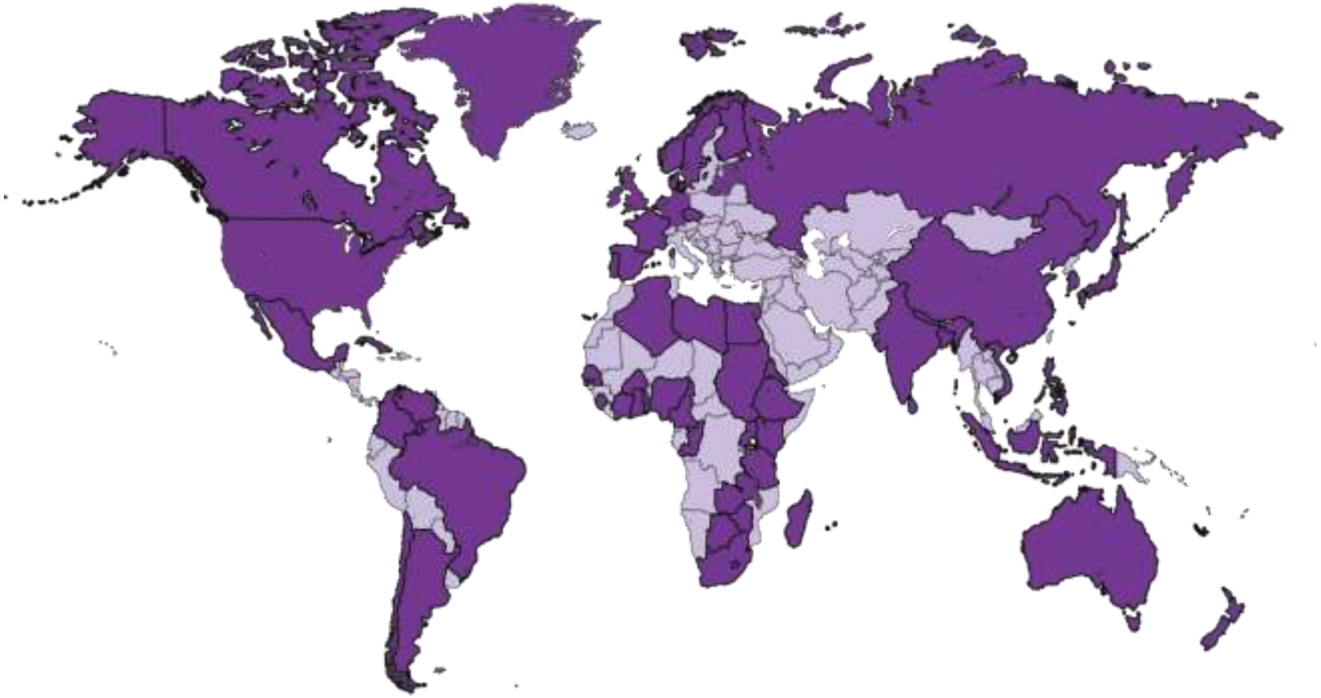
http://www.ncbi.nlm.nih.gov/pmc/

WorldWideScience – Facts and Figures

Tremendous growth in search content: from 10 nations to 65 nations in 3 years

> 400 million pages

- From well-known sources: *e.g.*, PubMed, Science.gov, Scielo
- To more obscure sources: *e.g.*, Bangladesh Journals Online



Powered by

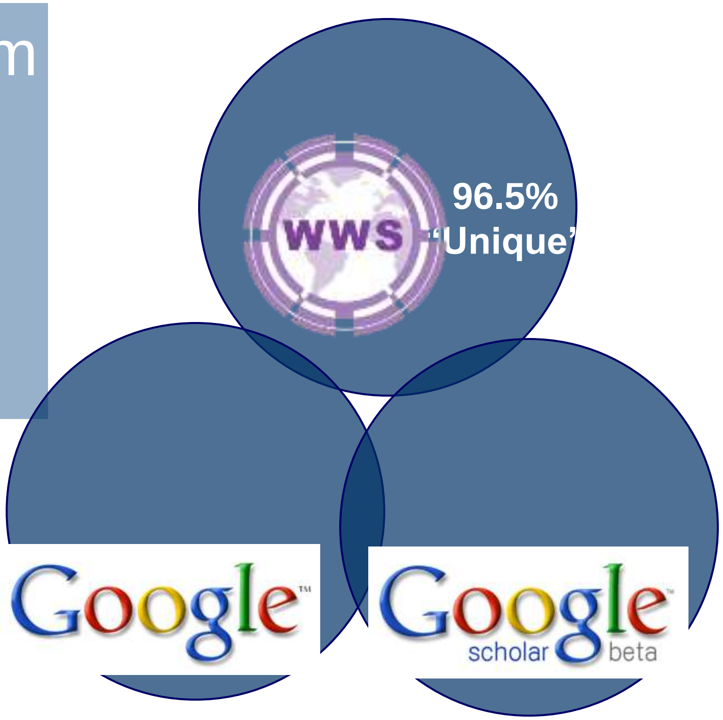
Microsoft® Translator

http://worldwidescience.org/

Microsoft Research Connections

WorldWideScience and the Invisible Web

- In comparison of search results from identical queries on WWS, Google, and Google Scholar, only 3.5% overlap (i.e., WorldWideScience is 96.5% unique)



Accelerated access → Accelerated discovery:
The case for WorldWideScience.org

Slide courtesy of Walt Warnick DOE OSTI

Need for Semantic Computing

Computers are great **tools** for

Storing	Computing
Managing	Indexing

huge amounts of **data**

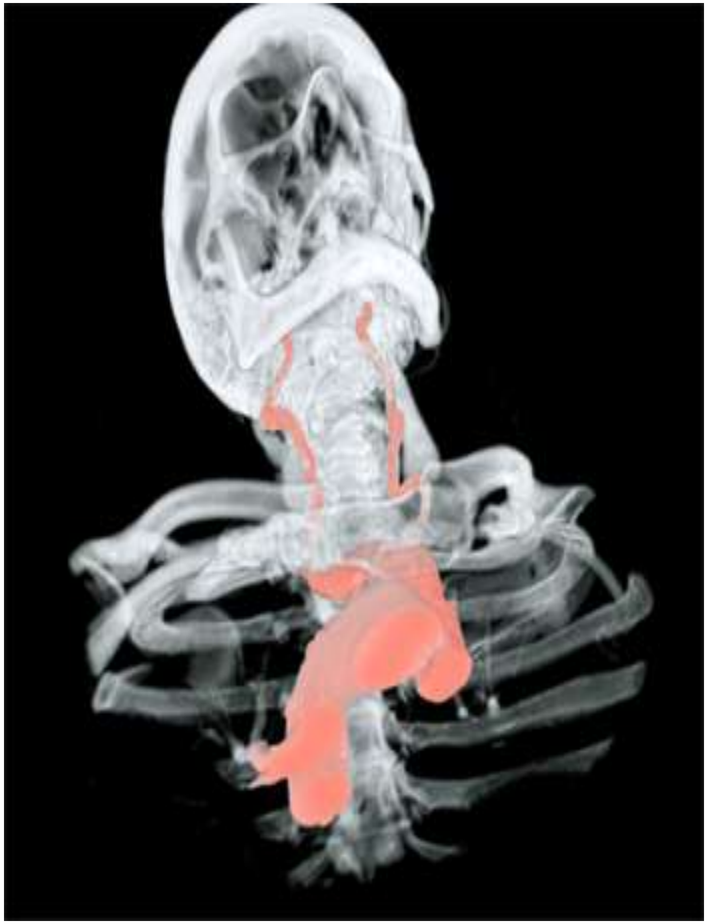
In the future we will need computers to help with the **automatic**

Acquisition	Discovery
Aggregation	Organization
Correlation	Analysis
Interpretation	Inference

of the world's **information**

InnerEye: Semantic Understanding of Medical Images

- InnerEye focuses on the analysis of patient scans using machine learning techniques for automatic detection and segmentation of healthy anatomy as well as anomalies.
- In this image, we see InnerEye can separate a carotid artery visually from adjacent parts of a human body

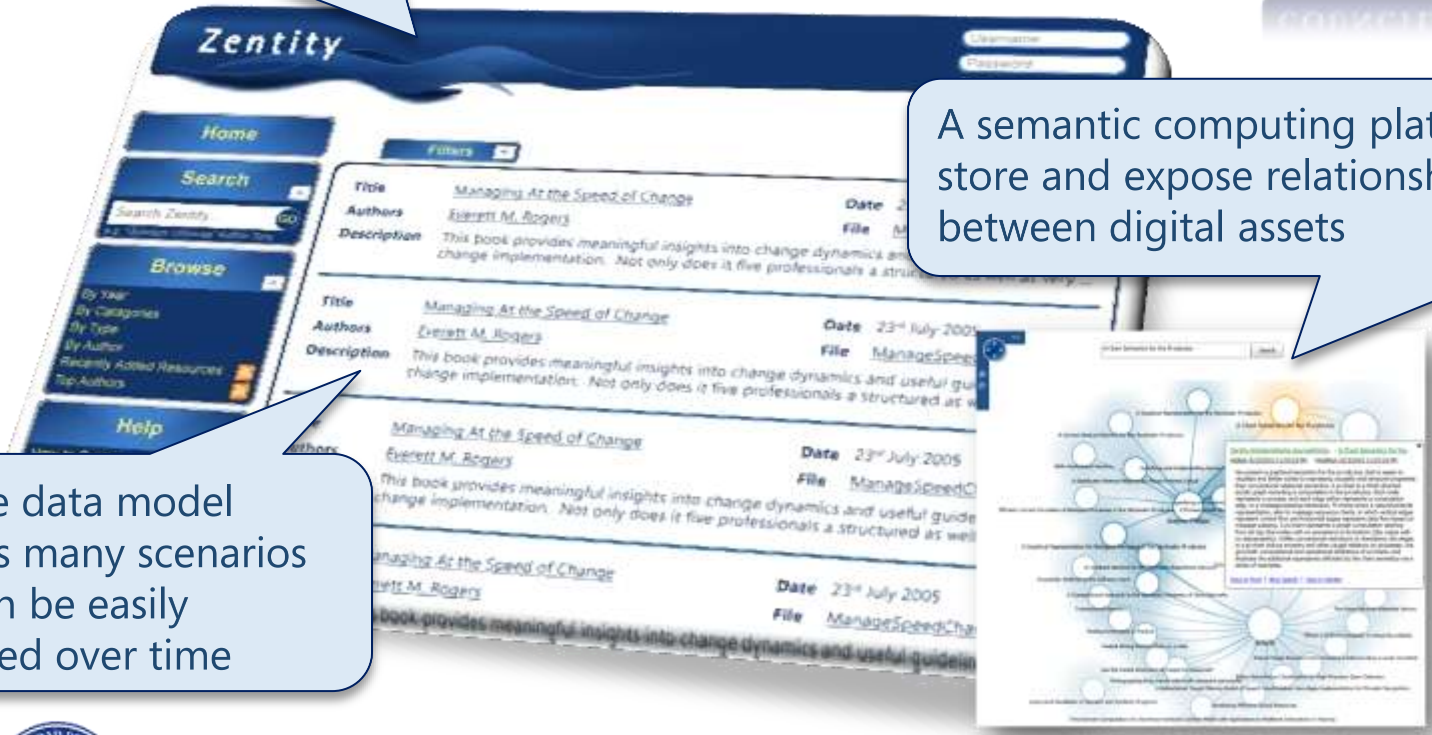


Zentity: Semantically-enabled repository software

Built on top of SQL Server & Entity Framework

Default web UI with CSS support and custom ASP.Net controls

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL



Title	Managing At the Speed of Change		
Authors	Everett M. Rogers	Date	23rd July 2005
Description	This book provides meaningful insights into change dynamics and useful guide… change implementation. Not only does it five professionals a structured as well…	File	ManageSpeedCh…
Title	Managing At the Speed of Change		
Authors	Everett M. Rogers	Date	23rd July 2005
Description	This book provides meaningful insights into change dynamics and useful gui… change implementation. Not only does it five professionals a structured as w…	File	ManageSpeed…
Title	Managing At the Speed of Change		
Authors	Everett M. Rogers	Date	23rd July 2005
Description	This book provides meaningful insights into change dynamics and useful guide… change implementation. Not only does it five professionals a structured as well…	File	ManageSpeedC…
	Managing At the Speed of Change		
	Everett M. Rogers	Date	23rd July 2005
	This book provides meaningful insights into change dynamics and useful guidelin…	File	ManageSpeedCha…

A semantic computing platform to store and expose relationships between digital assets

Flexible data model enables many scenarios and can be easily extended over time

Universidad de Bogotá
JORGE TADEO LOZANO

http://research.microsoft.com/zentity/

Microsoft Research Connections

Topics

The Fourth Paradigm

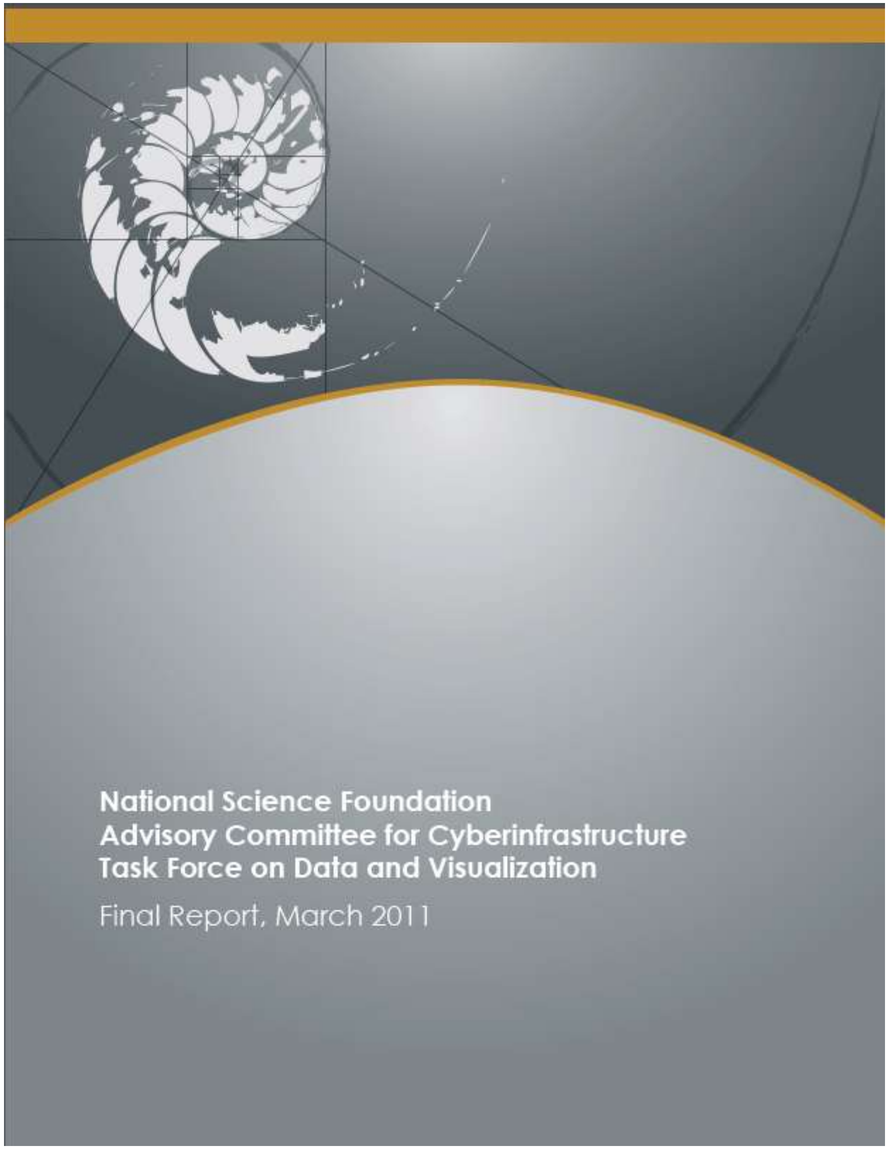
The Emergence of Open Science

Challenges and Opportunities of Open Data

The Role of Open Source

Future of Data-Intensive Science

NSF-OCI Task Force on Data and Visualization



National Science Foundation
Advisory Committee for Cyberinfrastructure
Task Force on Data and Visualization

Final Report, March 2011

Advisory Committee on Cyberinfrastructure

March 2011

Tony Hey, Co-Chair
Microsoft Corporation
Dan Atkins, Co-Chair
University of Michigan
Margaret Hedstrom
University of Michigan

http://www.nsf.gov/od/oci/taskforces/TaskForceReport_Data.pdf

Principal Recommendations

The Task Force strongly encourages the NSF to create a sustainable data infrastructure fit to support world-class research and innovation. It believes that such infrastructure is essential to sustain the USA's long-term leadership in scientific research and a legacy which can drive future discoveries, innovation and national prosperity.

To help realize this potential the Task Force identified challenges and opportunities which will require focused and sustained investment with clear intent and purpose; these are clustered into six main areas:

- **Infrastructure Delivery**
- **Culture and Sociological Change**
- **Roles and Responsibilities**
- **Economic Value and Sustainability**
- **Data Management Guidelines**
- **Ethics, Privacy and Intellectual Property**

- **Infrastructure Delivery** - Acknowledge that data infrastructure and services are essential research assets fundamental to today's science and worthy of long-term investments.
  - **Make specific budget allocations for the establishment and maintenance of research data sets and services and associated software and visualization tools.**

- **Culture and Sociological Change** - Introduce new funding models that reinforce expectations and institute specific conditions for data sharing.
  - **Create new norms and practices for citation and attribution so that data producers, software and tool developers, and data curators are credited with their contributions to scientific research.**

- **Roles and Responsibilities** - Recognize that responsibility for data stewardship is shared among:
  - Principal Investigators
  - Research centers
  - University research libraries
  - Discipline-based libraries and archives
  - National scientific agencies
  - Commercial service providers.

- **Economic Value and Sustainability** - Develop and publish realistic cost models to underpin institutional/national business plans for research repositories/data services.

- **Data Management Guidelines** - Identify and share best practices for critical areas of data management.

- **Ethics, Privacy and Intellectual Property** - Invest in the research and training of the research community in *privacy-preserving data-access* so that PIs can embrace privacy by design.

Paul Ginsparg: “As We May Read”

“On the one-decade time scale, it is likely that more **research communities will join some form of global unified archive system without the current partitioning and access restrictions** familiar from the paper medium, for the simple reason that it is the best way to communicate knowledge and hence to create new knowledge.”

“Ironically, it is also possible that the technology of the 21st century will allow the **traditional players from a century ago, namely the professional societies and institutional libraries, to return to their dominant role in support of the research Enterprise**.”

‘Openness’ will be critical for reducing the time-to-impact of Data-Intensive Science

1. Researchers must cooperate on standards for data provenance, curation and preservation
2. Scientific research needs to move to a default expectation of data-sharing
3. Publication processes and social behavior must be more flexible, real-time, and collaborative
4. Funding Agencies, Academia and Industry must share the costs of creating tools and technologies to publish, maintain, and consume open data sets
5. Data infrastructure and services must be recognized as core assets for scientific research

Resources

- Microsoft Research
  - http://research.microsoft.com
  - Microsoft Research downloads: http://research.microsoft.com/research/downloads
- Microsoft External Research
  - http://research.microsoft.com/en-us/collaboration/
- Science at Microsoft
  - http://www.microsoft.com/science
- Scholarly Communications
  - http://www.microsoft.com/scholarlycomm
- CodePlex
  - http://www.codeplex.com

Microsoft®

Your potential. Our passion.™