Reproducible Research and Data-Intensive Scientific Discovery

Tony Hey
Corporate Vice President
Microsoft Research

Microsoft Research Connections

Topics

- The Scientific Data Deluge
- Data-Intensive Scientific Discovery
- NSF OCI Data/Viz Task Force Report
- Sharing Research Data
- Reproducible Research
- Supporting the Data Life Cycle
- The Future?

Topics

The Scientific Data Deluge

Data-Intensive Scientific Discovery

NSF OCI Data/Viz Task Force Report

Sharing Research Data

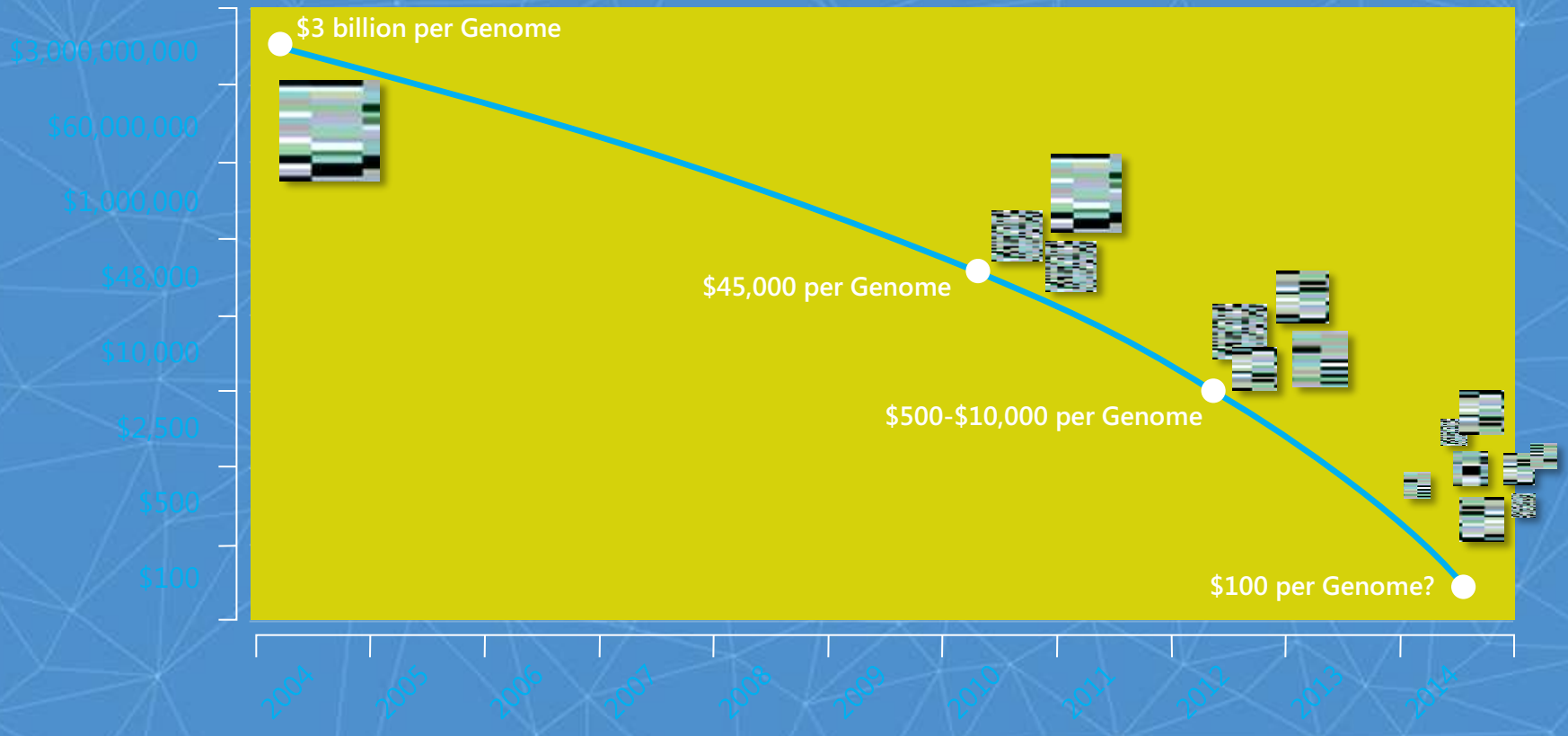
Reproducible Research

Supporting the Data Life Cycle

The Future?

A Tidal Wave of Scientific Data

Gene Sequencing Explosion



Source: George Church, Harvard Medical School, as reported in IEEE Spectrum, Feb '10. Figures represented in USD

Genomics and Personalized Medicine

Adapting treatments to a person's specific genetic make-up:

- Targeting patients who can benefit (*e.g.* 10% of people cannot respond to codeine), and not develop toxicities (*e.g.* Abacavir for HIV).
- Appropriate dosage of a drug by using genetic variants to understand drug metabolism (*e.g.* anti-depressants, beta blockers, opioid analgesics)
- More drug approvals (re-approvals) because can now target the right sub-group based on genetics.

Astronomy and Particle Physics

In 2000 the Sloan Digital Sky Survey collected more data in its 1st week than was collected in the entire history of Astronomy

By 2016 the New Large Synoptic Survey Telescope in Chile will acquire 140 terabytes in 5 days - more than Sloan acquired in 10 years

The Large Hadron Collider at CERN generates 40 terabytes of data every second

Sources: The Economist, Feb '10; IDC

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

Public Use of the SkyServer Data

- **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

- **New paradigm for scientific publishing**
  - Data are published before analysis by scientists

Topics

The Scientific Data Deluge

Data-Intensive Scientific Discovery

NSF OCI Data/Viz Task Force Report

Sharing Research Data

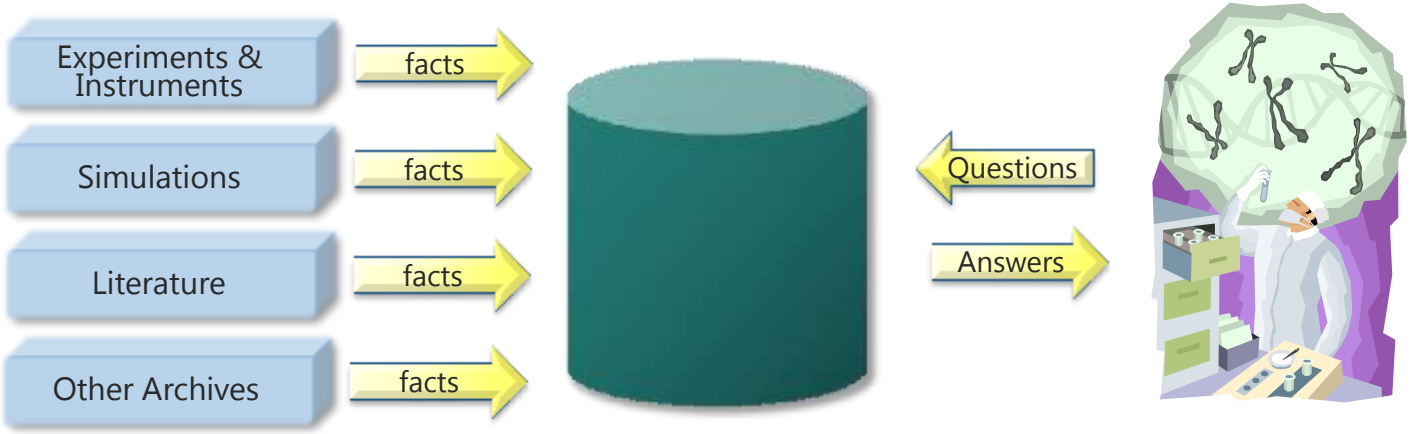
Reproducible Research

Supporting the Data Life Cycle

The Future?

X-Info

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



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)

Emergence of a Fourth Research Paradigm

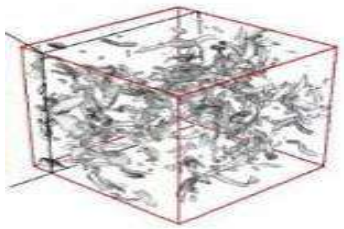
Thousand years ago – **Experimental Science**
- Description of natural phenomena

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)

Machine Learning and eScience

Tackling societal challenges

Identifying genetic and environmental causes of disease

Fighting HIV/AIDS

Increasing energy yield of sugar cane through genome assembly

World Wide Telescope

www.worldwidetelescope.org

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

Participants

- Alyssa Goodman; Harvard University
- Alex Szalay; Johns Hopkins University
- Curtis Wong, Jonathan Fay; Microsoft Research

- Integration of data sets and one-click contextual access
- Easy access and use

- As of May 2010, over 4M unique users (someone that has downloaded, installed, and successfully used WWT)
- The average number of WWT users over 8K per day

ChronoZoom – The 'Big History' Agenda

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."***

http://chronozoom.cloudapp.net/firstgeneration.aspx

Topics

The Scientific Data Deluge

Data-Intensive Scientific Discovery

NSF OCI Data/Viz Task Force Report

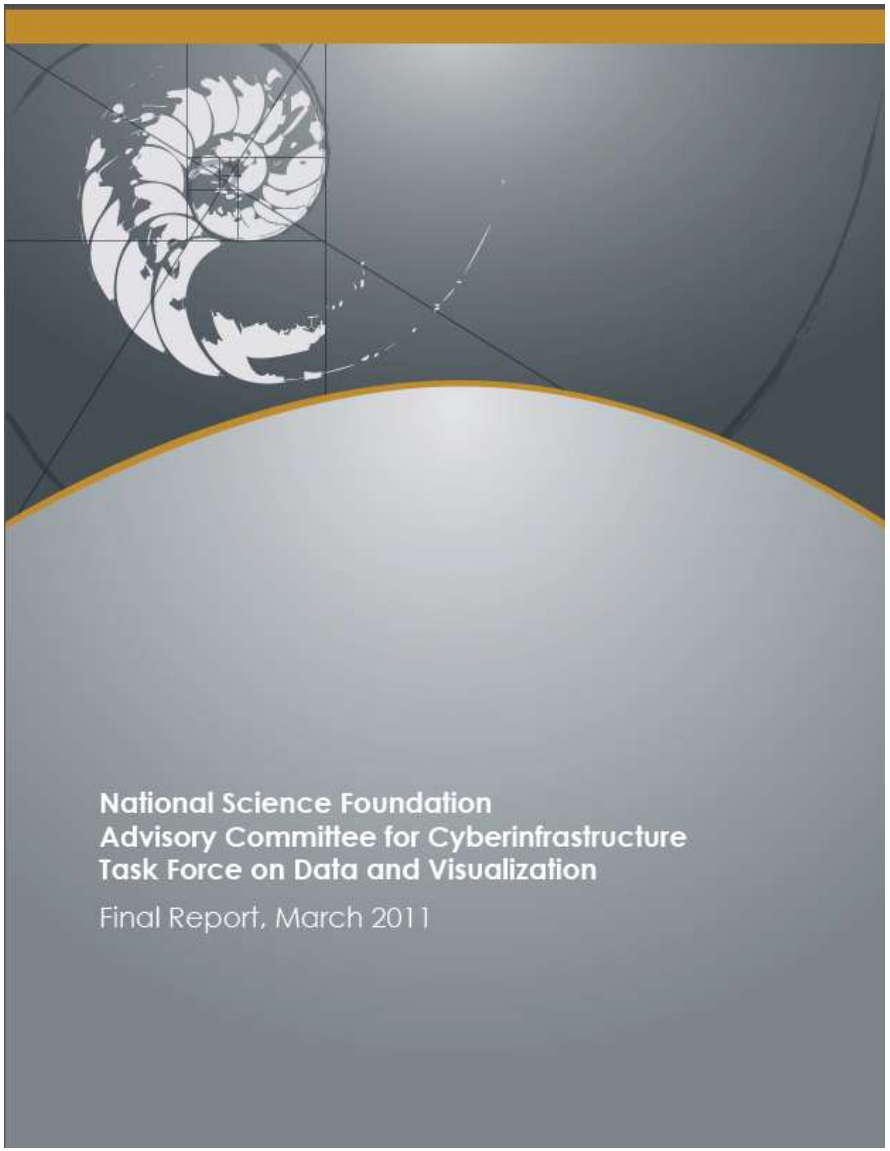
Sharing Research Data

Reproducible Research

Supporting the Data Life Cycle

The Future?

NSF-OCI Task Force on Data and Visualization



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.

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

Topics

The Scientific Data Deluge

Data-Intensive Scientific Discovery

NSF OCI Data/Viz Task Force Report

Sharing Research Data

Reproducible Research

Supporting the Data Life Cycle

The Future?

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF MANAGEMENT AND BUDGET
WASHINGTON, D.C. 20503
THE DIRECTOR
M-10-30
MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES
SUBJECT: Science and Technology Priorities for the FY 2012 Budget

"Agencies, in cooperation with OSTP and OMB, should develop and sustain datasets to better document Federal science, technology, and innovation investments and to make these data open to the public in accessible, useful formats. Agencies should develop and regularly update their data sharing policies for research performers and create incentives for sharing data publicly in interoperable formats to ensure maximum value, consistent with privacy, national security, and confidentiality concerns."

NSF Data Sharing Policy 2010

"Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing."

All future grant proposals now require a two-page Data Management Plan that addresses the above requirement and the Plan will be subject to peer review.

The Conundrum of Sharing Research Data

Christine Borgman, UCLA

Paper submitted to JASIST, June 21, 2011

NSB Report (2005)

"Long-Lived Digital Data Collections"

identifies 4 Categories of Data:

- Observational Data
- Computational Data
- Experimental Data
- Records

The Conundrum of Sharing Research Data (2)

Why Share Research Data?

1. To reproduce or to verify research
   - Problematic, only applicable to some data and some types of research
2. To make results of publicly funded research available to the public
   - "Public monies for public good" argument
3. To enable others to ask new questions of extant data
   - New results from scientific data mash-ups
4. To advance the state of research and innovation
   - Make research process more efficient

Funding Data Storage, Curation and Analysis

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

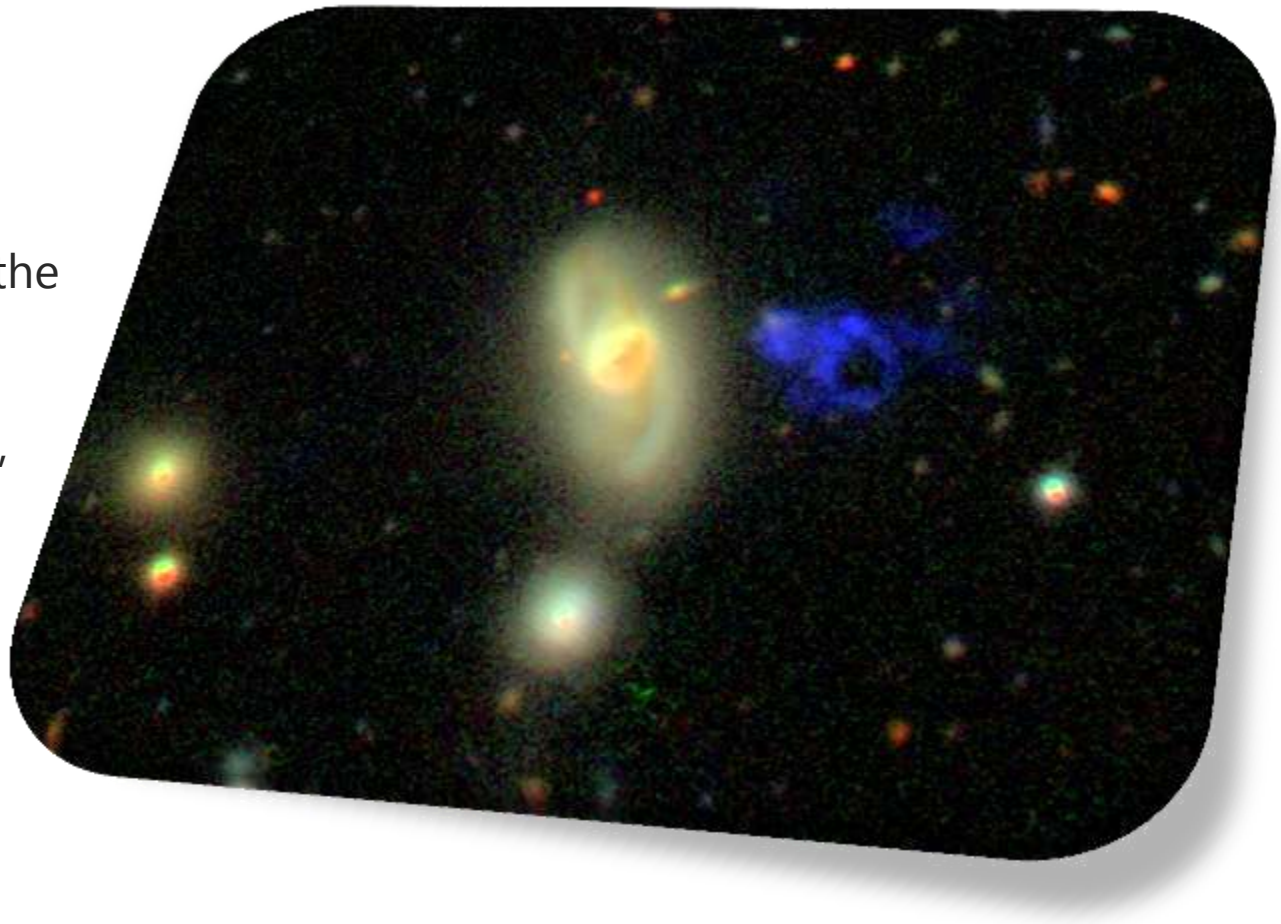
Galaxy Zoo activities give a useful indication of the latent appetite for scientific engagement in society. This is a collection of online astronomy projects which invite members of the public to assist in classifying galaxies.

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".



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.

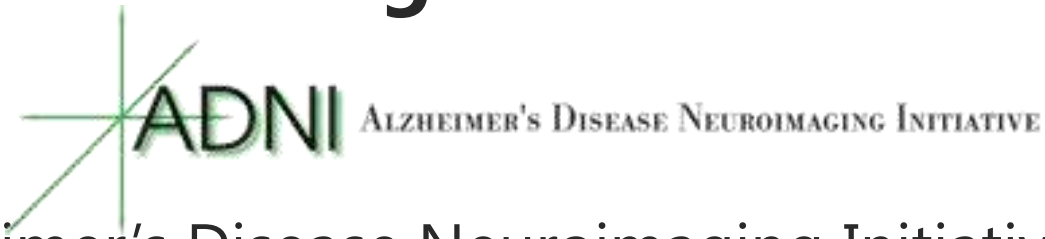
USGS is aiming for a VOI calculation that can inform decisions that maximize agricultural production by:

- 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/

Rapid Data Sharing for Alzheimer Biomarkers



- 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/

Topics

The Scientific Data Deluge

Data-Intensive Scientific Discovery

NSF OCI Data/Viz Task Force Report

Sharing Research Data

Reproducible Research

Supporting the Data Life Cycle

The Future?

Science November 9th 2010

POLICYFORUM

COMPUTER SCIENCE

Accessible Reproducible Research

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

Jill P. Mesirov

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.

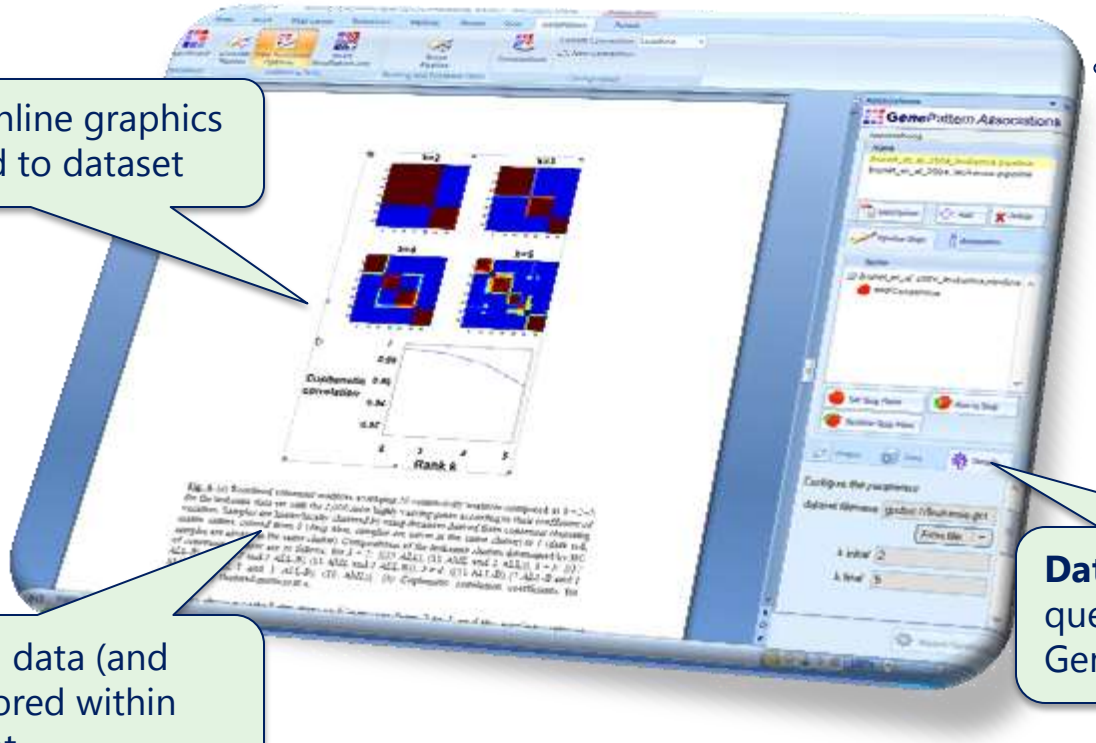
Over the past ~35 years, computational science has posed challenges to this traditional paradigm—from the publication of the four-color theorem in mathematics (*1*), in which the proof was partially performed by a computer program, to results depending on computer simulation in chemistry, materials science, 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



Thanks to Jill Mesirov and her team at the Broad Institute and to Barbara Hill and Christopher Lewis for the demo/video

Data: Control and execute query pipelines into GenePattern

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

http://www.broadinstitute.org/cancer/software/genepattern/grrd/WordAddInDemo.mov

Topics

The Scientific Data Deluge

Data-Intensive Scientific Discovery

NSF OCI Data/Viz Task Force Report

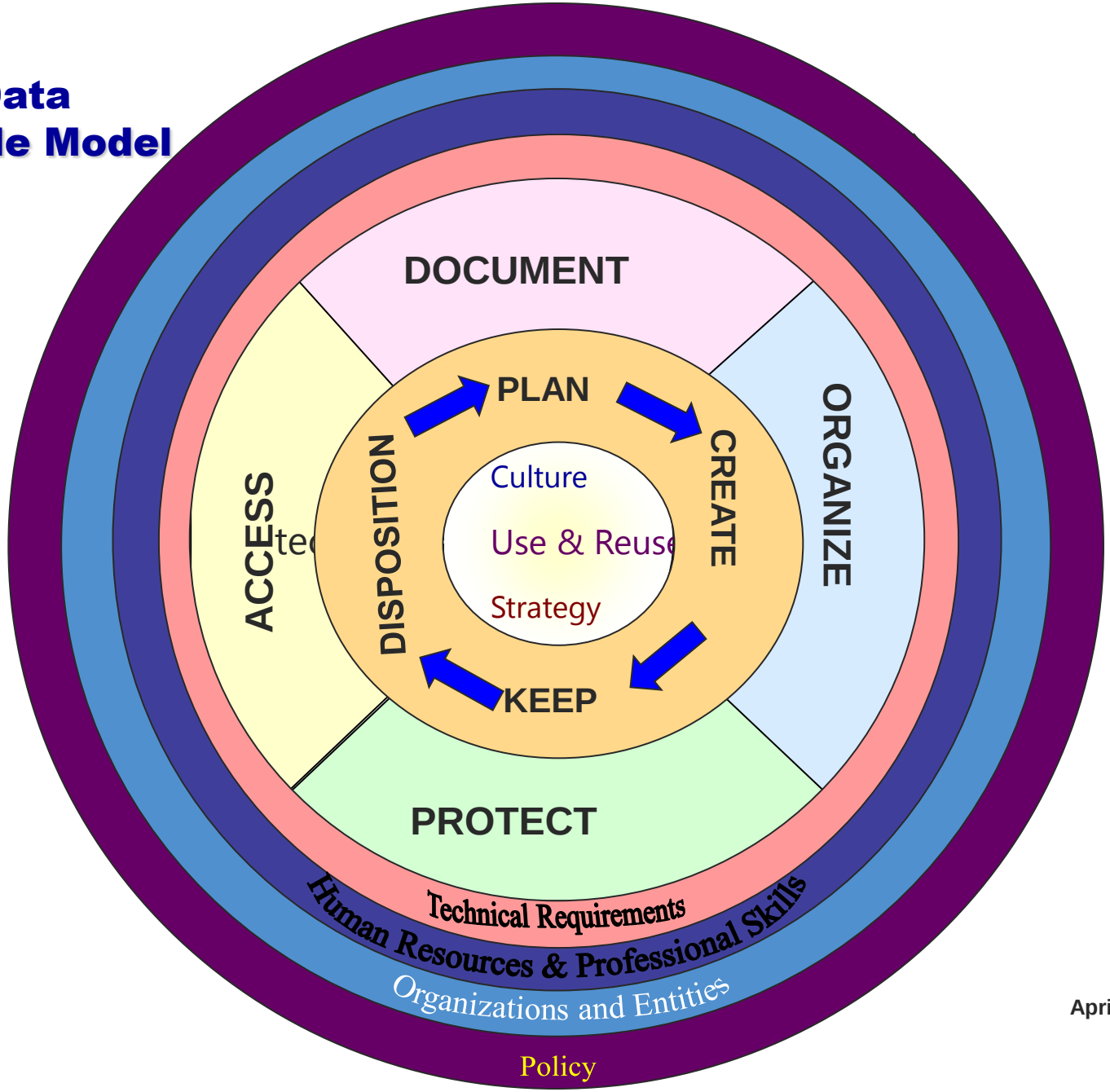
Sharing Research Data

Reproducible Research

Supporting the Data Life Cycle

The Future?

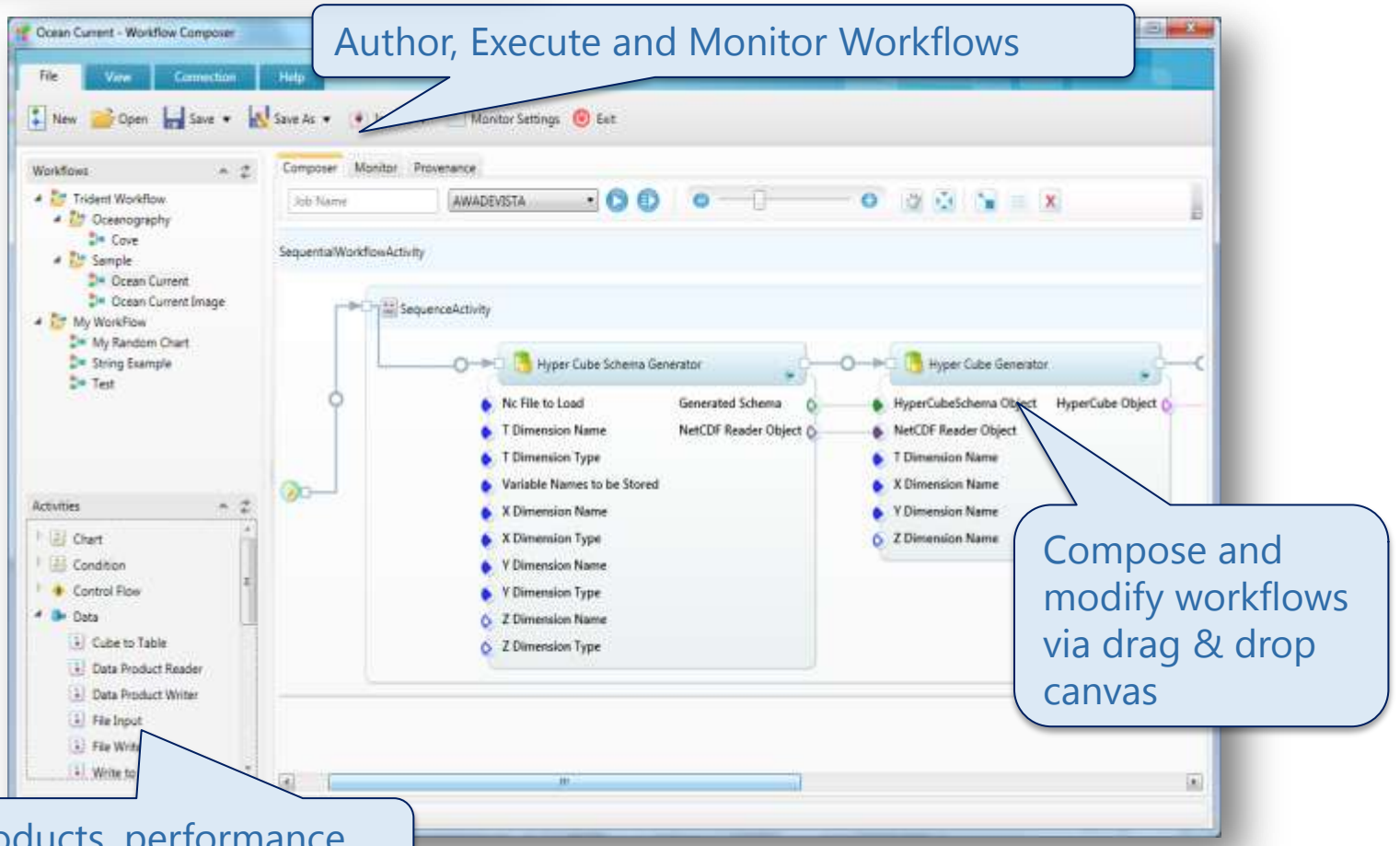
IWGDD Digital Data Life Cycle Model



April 2008

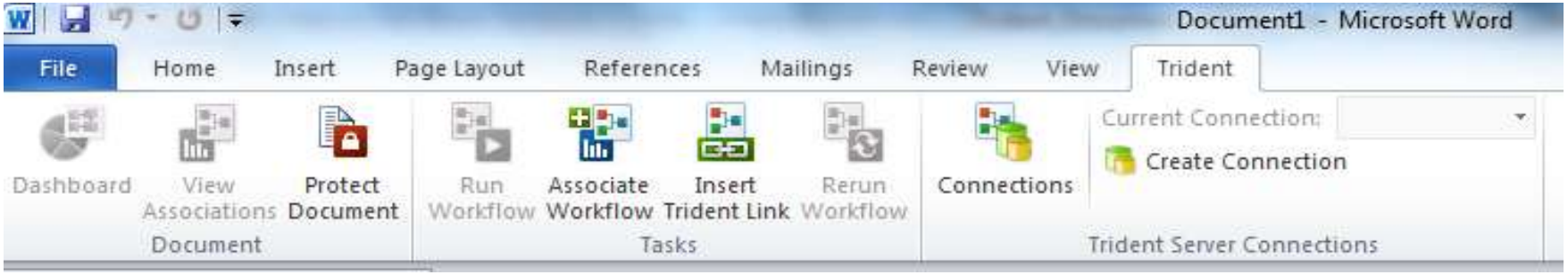
Project Trident – Scientific Workflow Workbench

Microsoft® .NET | Windows Workflow Foundation

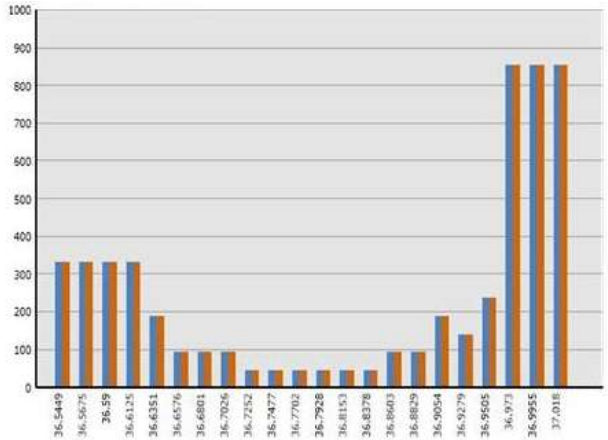


http://tridentworkflow.codeplex.com/

Trident Word Add-in for Reproducible Research



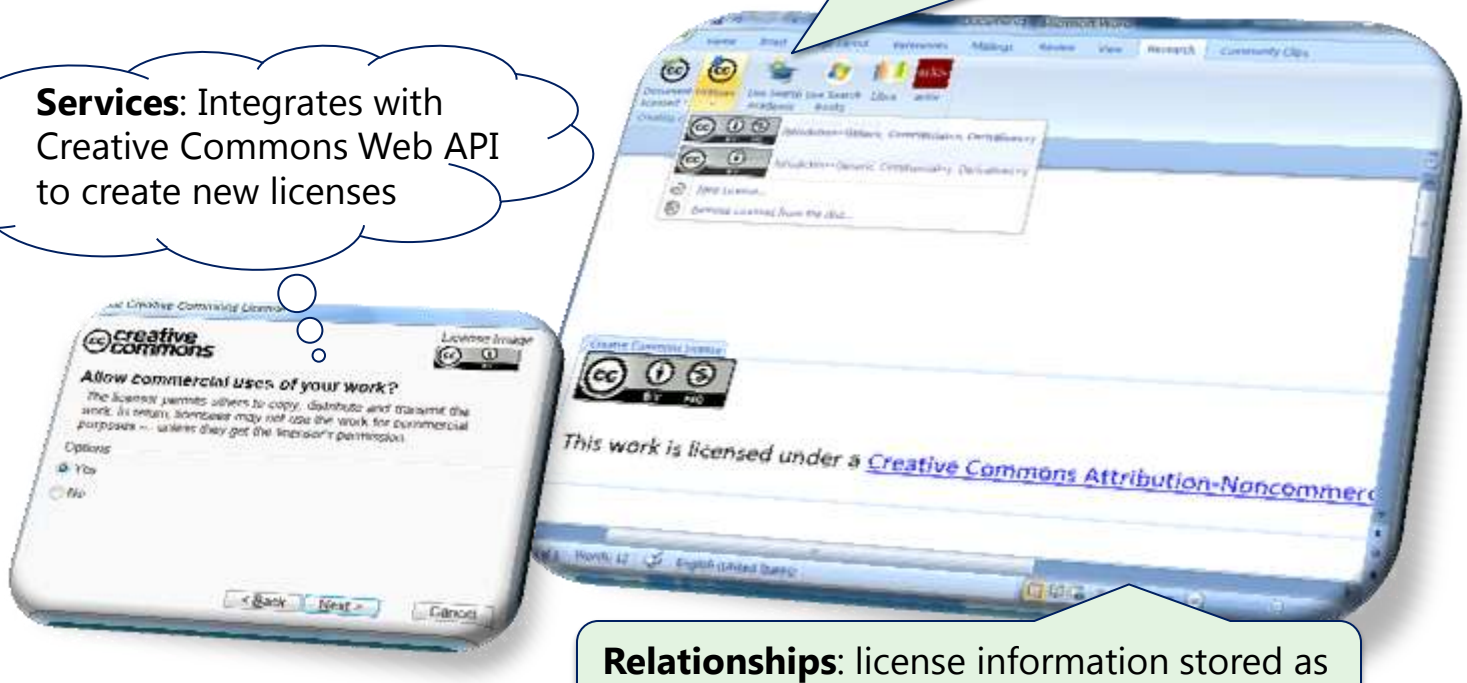
- Embed a Trident workflow package inside a Word document by associating with an image or text
- View inputs and outputs of an embedded workflow
- Rerun a workflow to reproduce the results while remaining in the Word application



Creative Commons Add-in for Office

Intent: Insert Creative Commons licenses from within Office 2007

Services: Integrates with Creative Commons Web API to create new licenses



Relationships: license information stored as RDF XML within the document OOXML

Downloads = 146,000+

Source code + binary:
http://ccaddin2007.codeplex.com

Article Authoring Add-in for Word

PubMed National Library of Medicine NLM

Hindawi

PLoS PUBLIC LIBRARY of SCIENCE

Aries systems Publish Faster. Publish Smarter.

Services: repository deposit via SWORD

arXiv.org

Fedora Commons

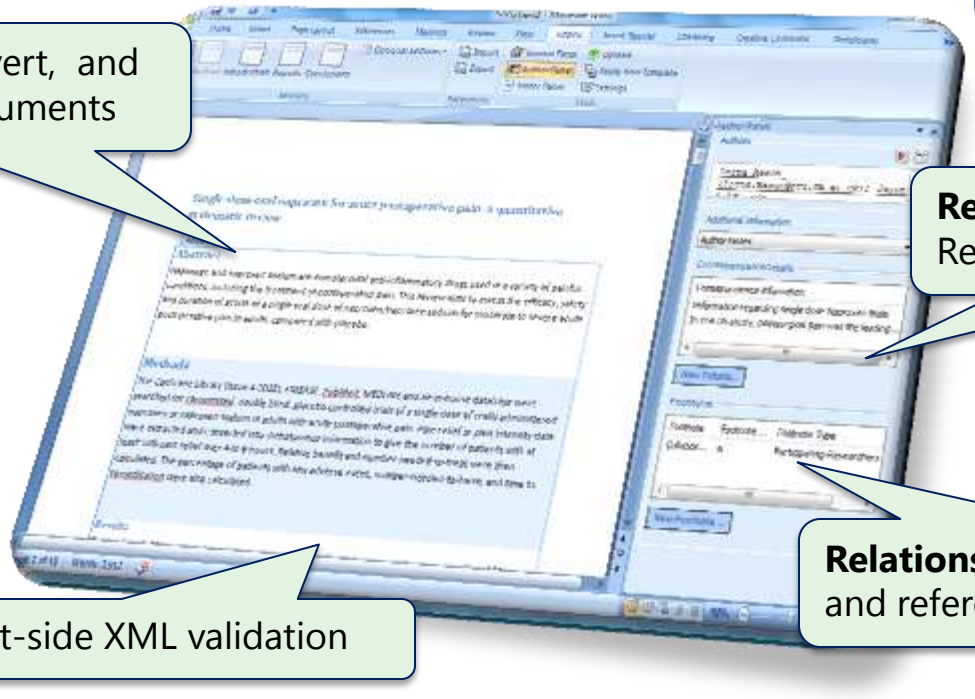
PubMed

eprints

Zentity

DSPACE

Structure: Read, convert, and author NLM XML documents



Relationships: ORE Resource Map creation

Relationships: Citation lookup and reference management

Structure: Client-side XML validation

Downloads = 4,000+

Binary (version 2.0):
http://research.microsoft.com/authoring/

Ontology Add-in for Word

THE NATIONAL CENTER FOR BIOMEDICAL ONTOLOGY

science commons

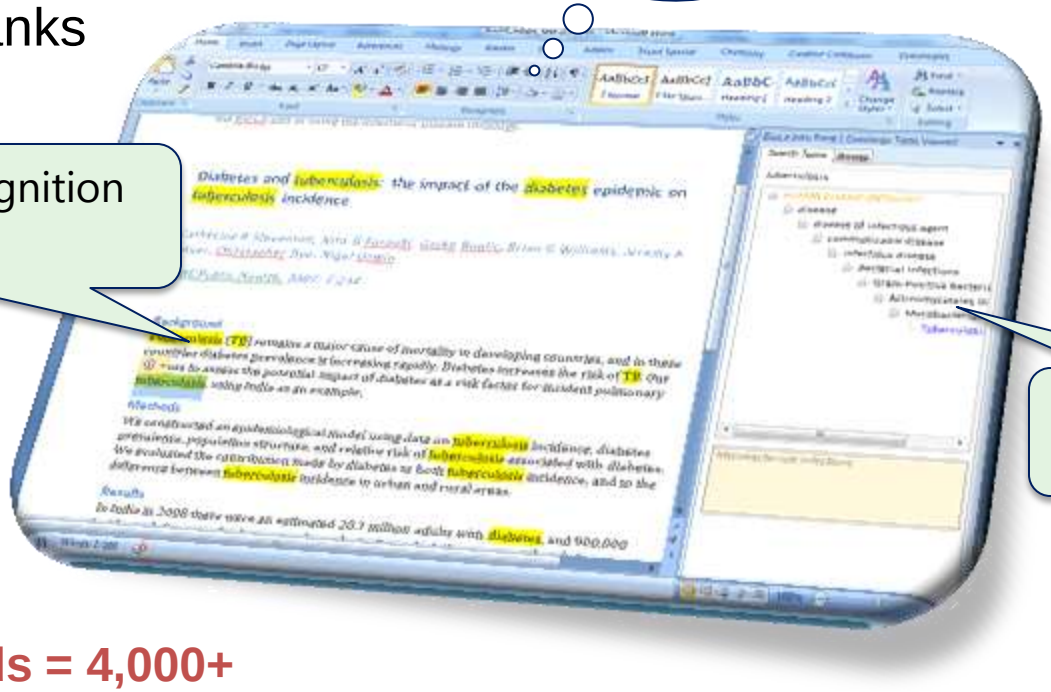
- John Wilbanks

Services: Ontology download web service

University of California San Diego

- Phil Bourne
- Lynn Fink

Intent: Term recognition & disambiguation

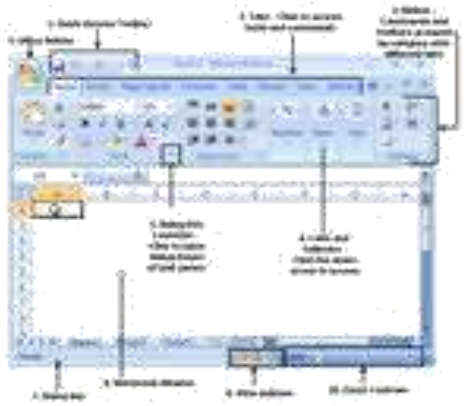


Relationships: Ontology browser

Downloads = 4,000+

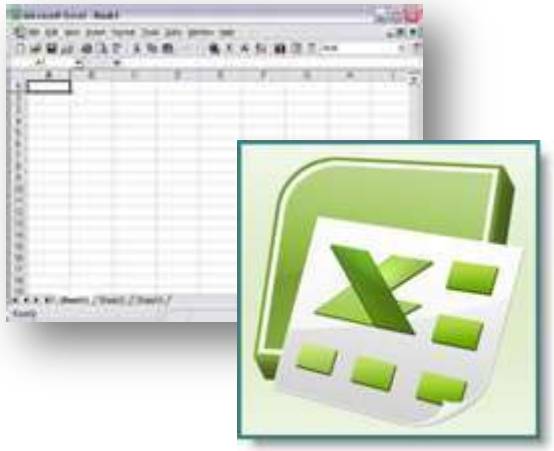
Source code + binary:
http://research.microsoft.com/ontology/

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.

Zentity: Semantically-enabled repository software

Built on top of SQL Server & Entity Framework

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

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

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

UNIVERSIDAD DE BOGOTÁ JORGE TADEO LOZANO

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

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.

Chem4Word– Chemical Drawing in Word

Semantic chemistry for students and publishers

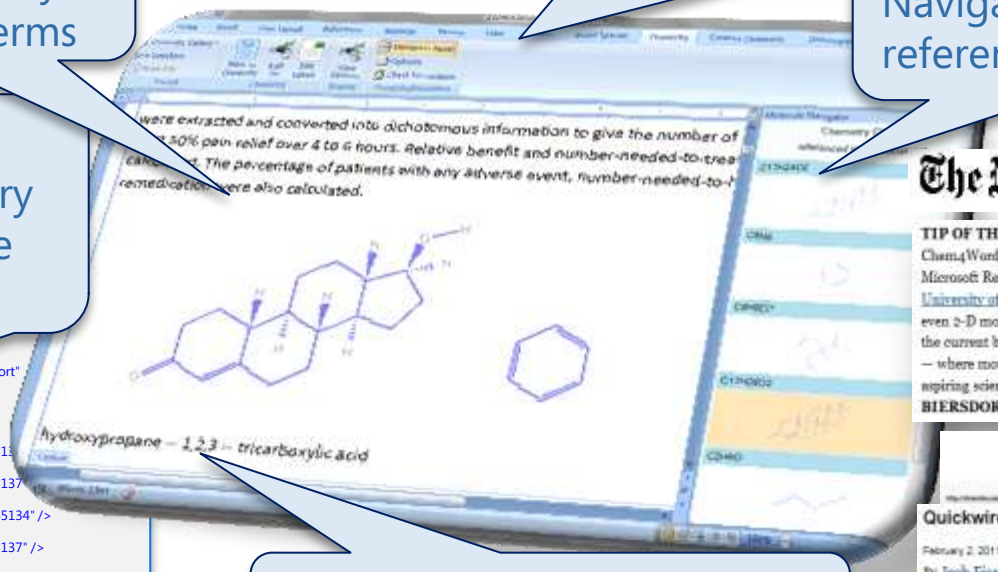
UNIVERSITY OF CAMBRIDGE

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)




```
<?xml version="1.0" ?>
<cml version="3" convention="org-synth-report"
xmlns="http://www.xml-cml.org/schema">
  <molecule id="m1">
    <atomArray>
      <atom id="a1" elementType="C" x2="-
2.9149999618530273" y2="0.7699999809265137" />
      <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="-
2.9149999618530273" y2="-0.7700000190734864" />
      <atom id="a7" elementType="H" x2="-
4.248679083681063" y2="-1.9073486459691087E-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/rjK33 — 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/

Microsoft Research Connections

Topics

The Scientific Data Deluge

Data-Intensive Scientific Discovery

NSF OCI Data/Viz Task Force Report

Sharing Research Data

Reproducible Research

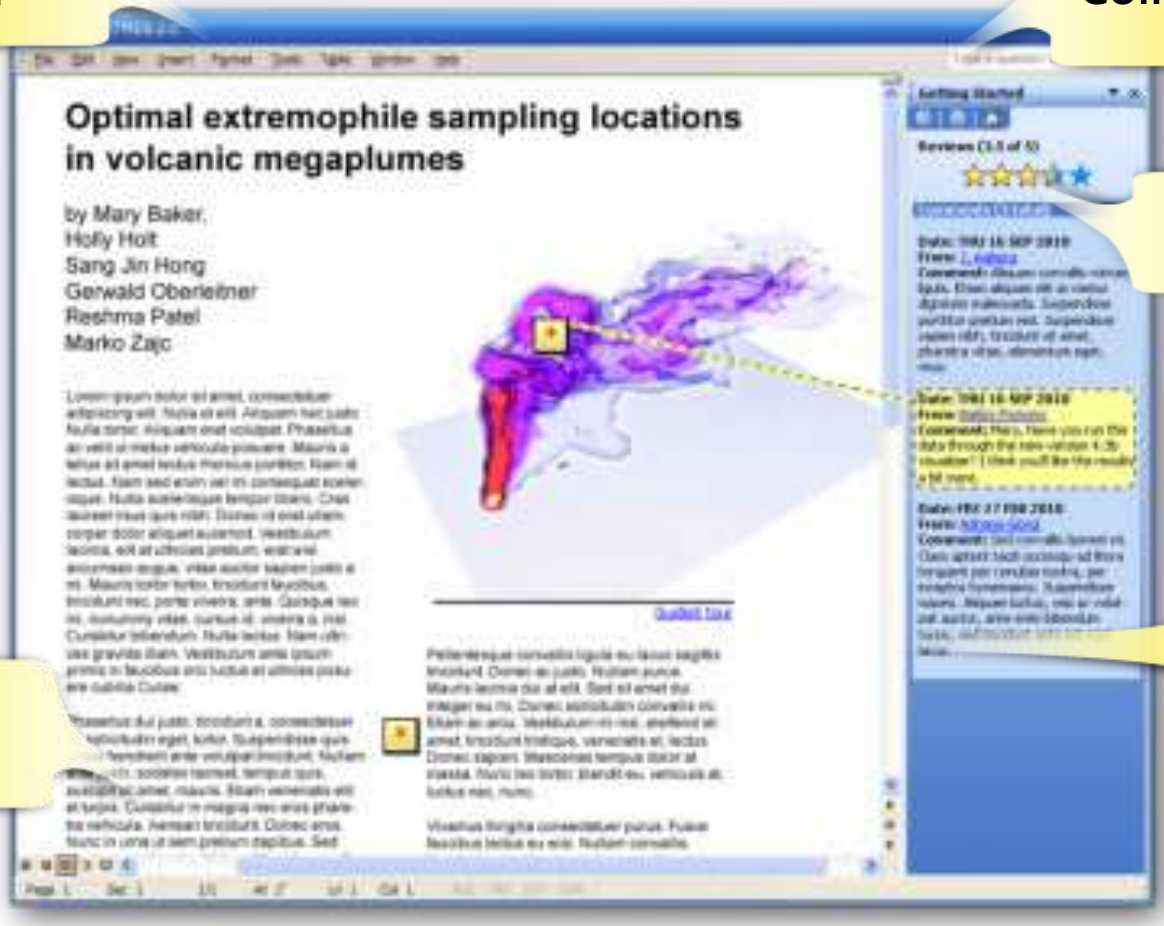
Supporting the Data Life Cycle

The Future?

Envisioning a New Era of Research Reporting

Reproducible Research

Collaboration



Reputation & Influence

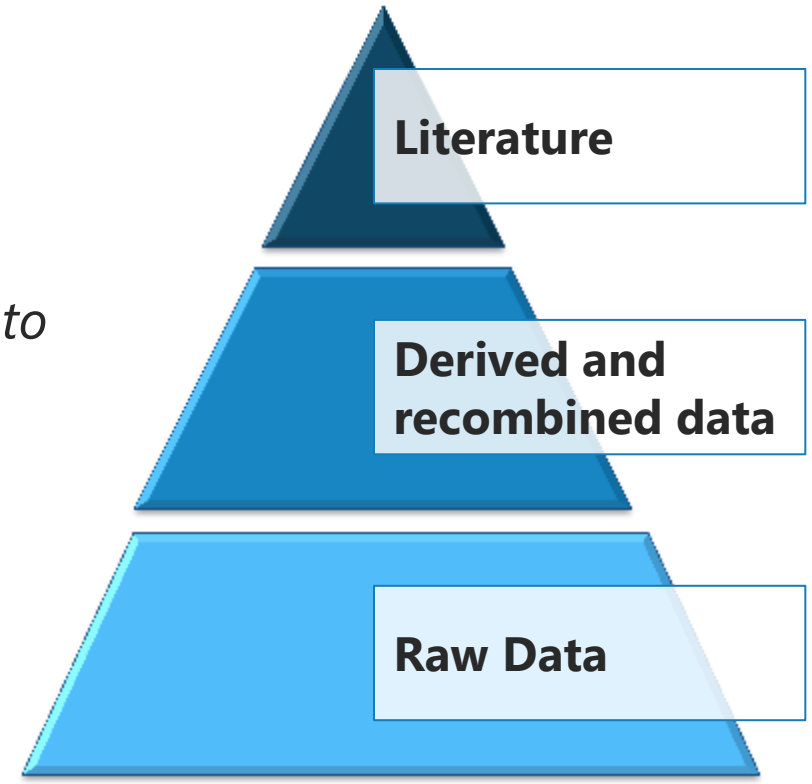
Interactive Data

Dynamic Documents

(Thanks to Bill Gates SC05)

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)

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.™