Data Curation -
Past, Present and Future

Tony Hey
Senior Data Science Fellow
eScience Institute
University of Washington
tony.hey@live.com

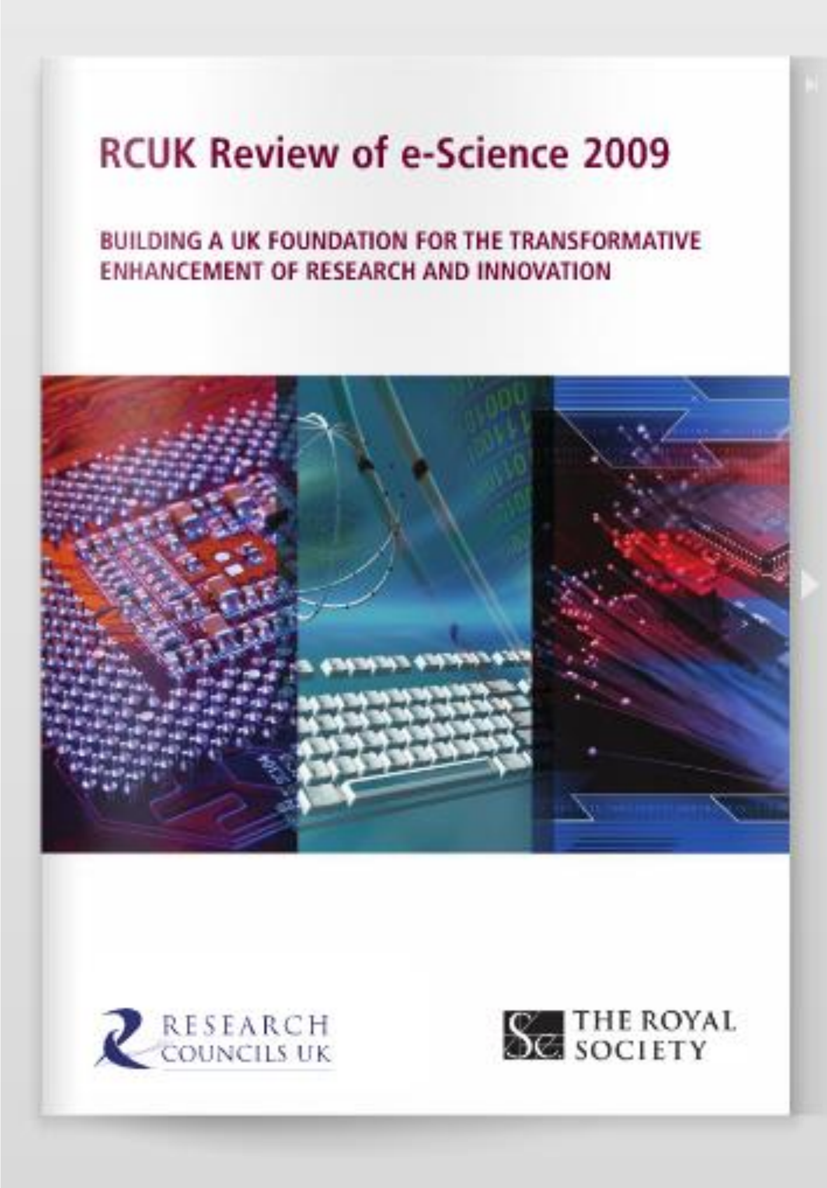
Warning and Acknowledgements

- This talk is a personal view of some of the projects and activities that I have learnt about over the past 10 years.
- The talk is not intended to be a comprehensive survey of all the significant projects and developments during the last decade.
- I am grateful to all those who helped me in the preparation of this talk – though they are not responsible for any errors or misunderstandings I may inadvertently introduced!
- Explicit thanks are due to:

Rolf Apweiler, Fran Berman, Simon Coles, Jeff Dozier, Christopher Erdmann, James Frew, Jeremy Frey, Francoise Genova, Carole Goble, Jessie Hey, Michael Kurtz, Bryan Lawrence, Bill Michener, Natasa Milic-Frayling, Carole Palmer, Beth Plale, and Alex Wade
- and to many others who have educated me about the importance of data curation, including, of course, Lee Dirks.

The Past

- The UK e-Science Initiative 2001 -2006
- Elements of a Global e-Infrastructure
- e-Science and the Fourth Paradigm
- Executable paper vision

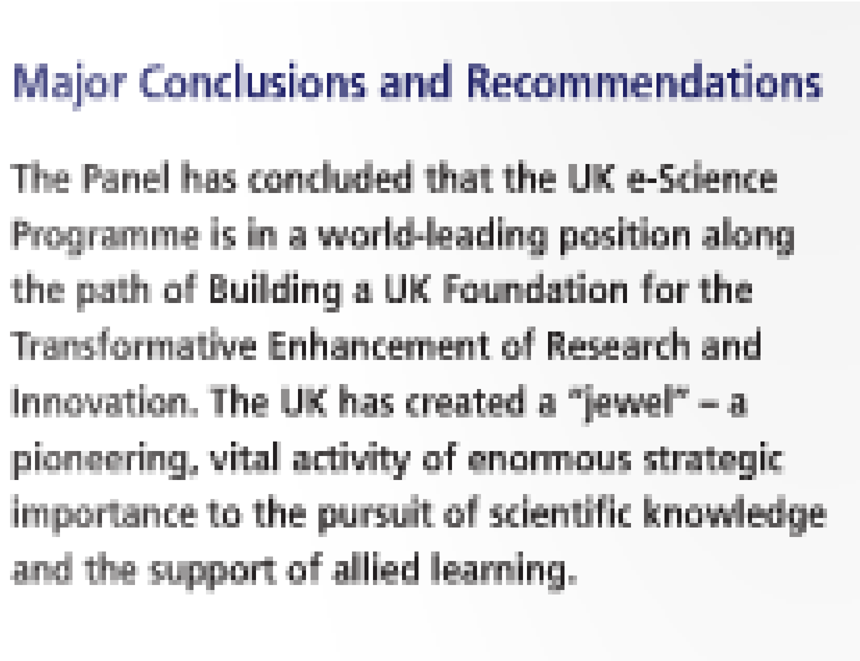


RCUK Review of e-Science 2009

BUILDING A UK FOUNDATION FOR THE TRANSFORMATIVE ENHANCEMENT OF RESEARCH AND INNOVATION

RESEARCH COUNCILS UK

THE ROYAL SOCIETY



Major Conclusions and Recommendations

The Panel has concluded that the UK e-Science Programme is in a world-leading position along the path of Building a UK Foundation for the Transformative Enhancement of Research and Innovation. The UK has created a "jewel" – a pioneering, vital activity of enormous strategic importance to the pursuit of scientific knowledge and the support of allied learning.

Chair: Dan Atkins

http://www.epsrc.ac.uk/research/intrevs/escience/Pages/default.aspx

The UK e-Science Initiative 2001 - 2006

A 5 year program of multidisciplinary data-intensive scientific research

- £200M (~ $320M) for university research
- £30M (~ $48M) for collaborative research with industry

e-Science Program covered all UK Research Councils

- Engineering and Science (EPSCRC)
- Biological and Biotechnology (BBSRC)
- Natural Environment (NERC)
- Medical (MRC)
- Particle Physics and Astronomy PPSRC
- Economics and Social Science (ESRC)
- Arts and Humanities (AHRC)

The Research Councils in the UK (RCUK)

- RCUK ≈ NSF + NIH

UK e-Science Program: Six Key Elements for a Global e-Infrastructure

1. High bandwidth Research Networks
2. Internationally agreed AAA Infrastructure
3. Development Centres for Open Software
4. **Technologies and standards for Data Provenance, Curation and Preservation**
5. **Open access to Data and Publications via Interoperable Repositories**
6. Discovery Services and Collaborative Tools

Slide from Tony Hey presentation 2004

DCC Digital Curation Centre

Established in 2004 with mission to work with librarians, scientists and computer scientists to examine all aspects of research data curation and preservation

- Actions needed to maintain and utilise digital data and research results over entire life-cycle
  - For current and future generations of users
- Digital Preservation
  - Long-run technological/legal accessibility and usability
- Data curation in science
  - Maintenance of body of trusted data to represent current state of knowledge in area of research
- Research in tools and technologies
  - Integration, annotation, provenance, metadata, security…..

eScience and the Fourth Paradigm

Thousand years ago – **Experimental Science**

- Description of natural phenomena

Last few hundred years – **Theoretical Science**

- Newton's Laws, Maxwell's Equations…

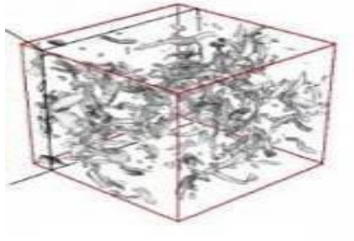
Last few decades – **Computational Science**

- Simulation of complex phenomena

Today – **Data-Intensive Science**

- Scientists overwhelmed with data sets from many different sources
  - Data captured by instruments
  - Data generated by simulations
  - Data generated by sensor networks

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

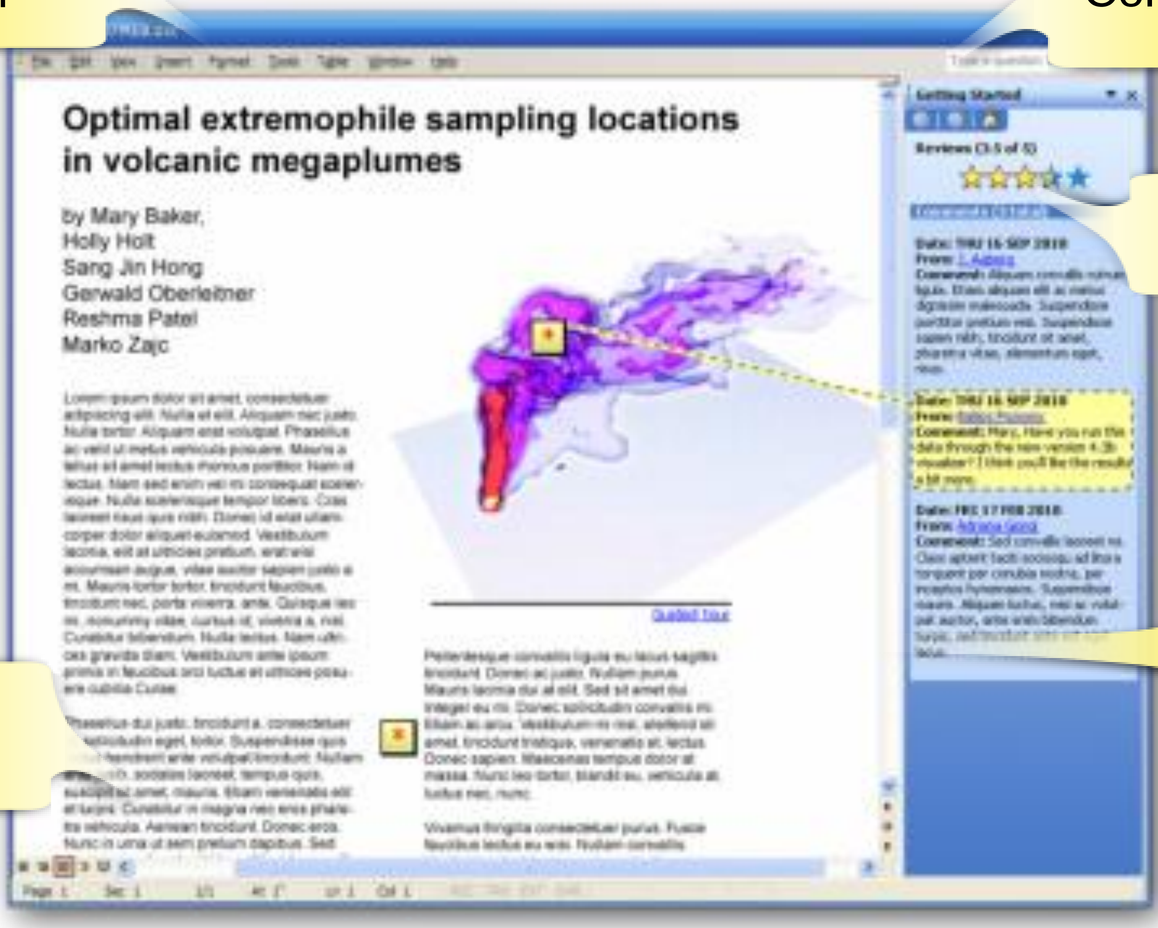


eScience is the set of tools and technologies to support data federation and collaboration

- For analysis and data mining
- For data visualization and exploration
- For scholarly communication and dissemination

(With thanks to Jim Gray)

Vision for a New Era of Research Reporting



File Edit View Insert Format Tools Table Window Help
Type a question
Getting Started
Reviews (1-5 of 5)
Guided Tour
Page 1 Sec 1 1/1 At 1" Ln 1 Col 1

Optimal extremophile sampling locations in volcanic megaplumes

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

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Nulla et elit. Aliquam nec justo. Nulla tortor. Aliquam erat volutpat. Phasellus ac velit ut metus vehicula posuere. Mauris a tellus sit amet lectus rhoncus porttitor. Nam id lectus. Nam sed enim vel mi consequat scelerisque. Nulla scelerisque tempor libero. Cras laoreet risus quis nibh. Donec id erat ullamcorper dolor aliquet euismod. Vestibulum lacinia, elit at ultricies pretium, erat nisl accumsan augue, vitae auctor sapien justo a mi. Mauris tortor tortor, tincidunt faucibus, tincidunt nec, porta viverra, ante. Quisque leo mi, nonummy vitae, cursus id, viverra a, nisl. Curabitur bibendum. Nulla lectus. Nam ultrices gravida diam. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae.

Phasellus dui justo, tincidunt a, consectetuer sollicitudin eget, tortor. Suspendisse quis hendrerit ante volutpat tincidunt. Nullam [illegible], sodales laoreet, tempus quis, suscipit sit amet, mauris. Etiam venenatis elit et turpis. Curabitur in magna nec eros pharetra vehicula. Aenean tincidunt. Donec eros. Nunc in urna ut sem pretium dapibus. Sed

Pellentesque convallis ligula eu lacus sagittis tincidunt. Donec ac justo. Nullam purus. Mauris lacinia dui at elit. Sed sit amet dui. Integer eu mi. Donec sollicitudin convallis mi. Etiam ac arcu. Vestibulum mi nisi, eleifend sit amet, tincidunt tristique, venenatis et, lectus. Donec sapien. Maecenas tempus dolor at massa. Nunc leo tortor, blandit eu, vehicula at, luctus nec, nunc.

Vivamus fringilla consectetuer purus. Fusce faucibus lectus eu wisi. Nullam convallis.

Date: THU 16 SEP 2010
From: J. Adams
Comment: Aliquam convallis rutrum ligula. Etiam aliquam elit ac metus dignissim malesuada. Suspendisse porttitor pretium wisi. Suspendisse sapien nibh, tincidunt sit amet, pharetra vitae, elementum eget, risus.

Date: THU 16 SEP 2010
From: [illegible]
Comment: Mary, Have you run the data through the new version 4.3b visualizer? I think you'll like the results a bit more.

Date: FRI 17 FEB 2010
From: Adriana [illegible]
Comment: Sed convallis laoreet mi. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Suspendisse mauris. Aliquam luctus, nisl ac volutpat auctor, ante erat bibendum turpis, sed tincidunt [illegible] lacus.

Reproducible Research

Collaboration

Reputation & Influence

Dynamic Documents

Interactive Data

(Thanks to Bill Gates SC05)

The Present – Disciplines, Programs and Policies

- Astronomy
- Bioinformatics
- Chemistry
- Environmental Science
- Ecology

- JISC and Jisc
- Australian eResearch Initiative
- Open Access, Open Data, Open Science

Astronomy

The 'Cosmic Genome Project': The Sloan Digital Sky Survey

- Two surveys in one
  - One quarter of the night sky
  - Photometric survey in 5 bands
  - Spectroscopic redshift survey
- Data is public
  - 2.5 Terapixels of images
  - 40 TB of raw data => 120TB processed data
  - 5 TB catalogs => 35TB in the end
- Started in 1992, finished in 2013
  - ➢SkyServer Web Service
    built at JHU by team led by
    Alex Szalay and Jim Gray

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

Long Term Access to Large Scientific Data Sets: The SkyServer and Beyond

NSF ACI Data Infrastructure Building Blocks (DIBBS) Program

- $7.6M project
- Started 2013 – end date 2018

Project goals

- Address curation issues arising from the data and service life-cycle
- Support small but complex data in the 'Long Tail' of science.

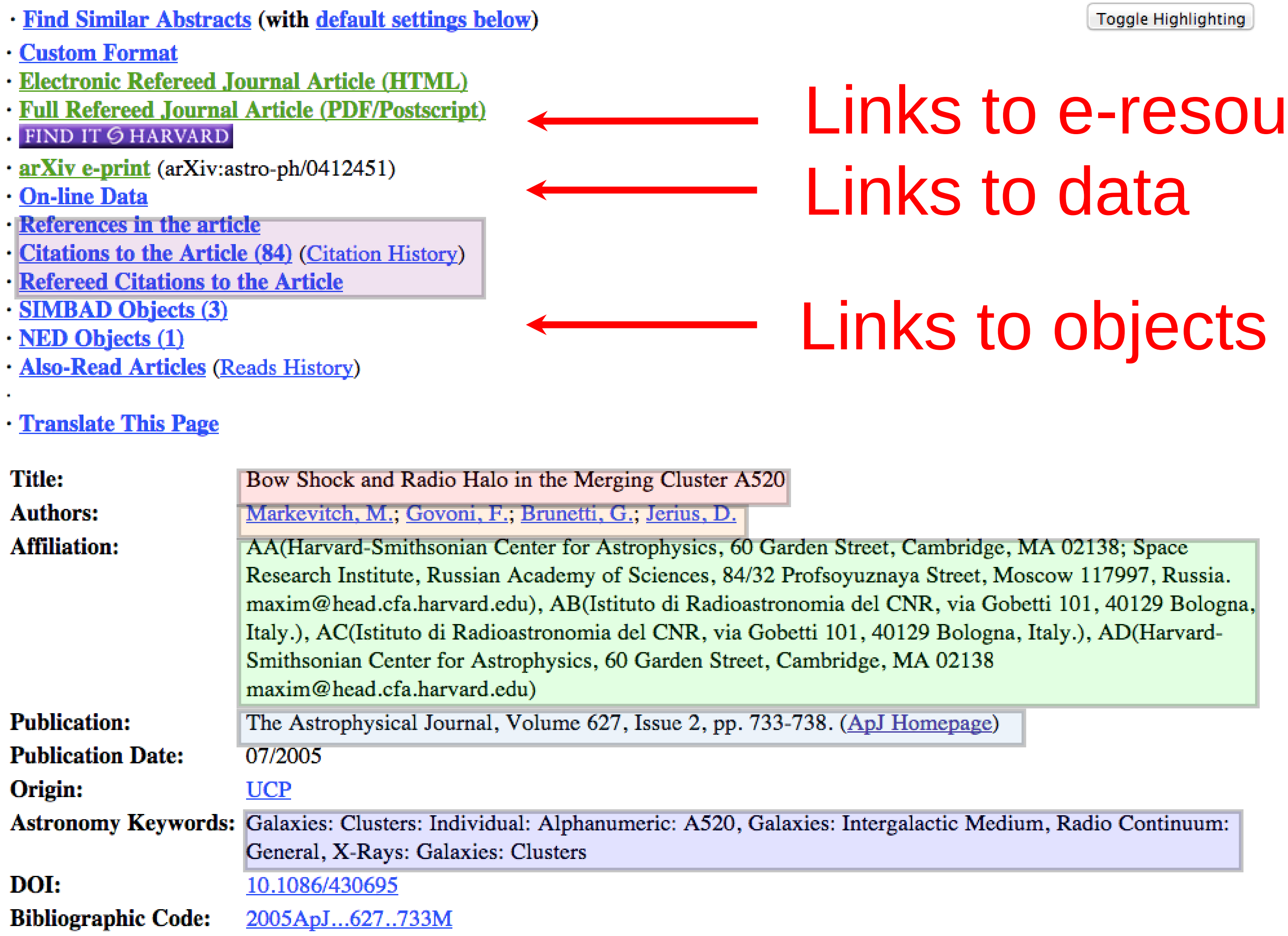
➢ Need to curate both Data and Services lifecycle

What happened to Virtual Observatories?

- UK AstroGrid project
  - Funding cancelled in 2008
- US Virtual Astronomy Observatory (VAO)
  - Project funding discontinued in 2014

➢ But much of the infrastructure, tools and technology still lives on with participation in the International Virtual Observatory Alliance (IVOA)

Astrophysics Data System ADS



- Find Similar Abstracts (with default settings below)
- Custom Format
- Electronic Refereed Journal Article (HTML)
- Full Refereed Journal Article (PDF/Postscript)
- FIND IT @ HARVARD
- arXiv e-print (arXiv:astro-ph/0412451)
- On-line Data
- References in the article
- Citations to the Article (84) (Citation History)
- Refereed Citations to the Article
- SIMBAD Objects (3)
- NED Objects (1)
- Also-Read Articles (Reads History)
-
- Translate This Page

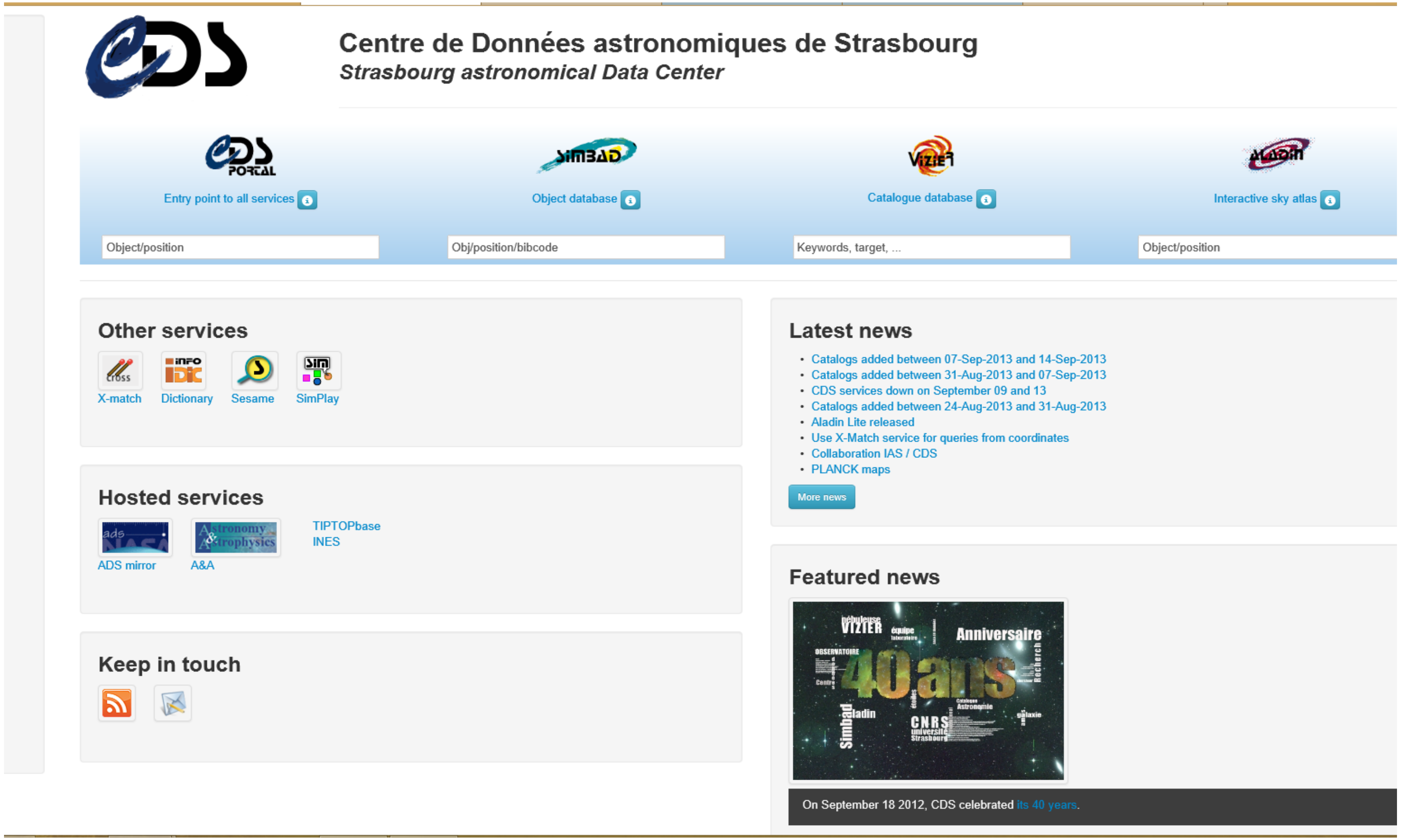
Title:	Bow Shock and Radio Halo in the Merging Cluster A520
Authors:	Markevitch, M.; Govoni, F.; Brunetti, G.; Jerius, D.
Affiliation:	AA(Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138; Space Research Institute, Russian Academy of Sciences, 84/32 Profsoyuznaya Street, Moscow 117997, Russia. maxim@head.cfa.harvard.edu), AB(Istituto di Radioastronomia del CNR, via Gobetti 101, 40129 Bologna, Italy.), AC(Istituto di Radioastronomia del CNR, via Gobetti 101, 40129 Bologna, Italy.), AD(Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138 maxim@head.cfa.harvard.edu)
Publication:	The Astrophysical Journal, Volume 627, Issue 2, pp. 733-738. (ApJ Homepage)
Publication Date:	07/2005
Origin:	UCP
Astronomy Keywords:	Galaxies: Clusters: Individual: Alphanumeric: A520, Galaxies: Intergalactic Medium, Radio Continuum: General, X-Rays: Galaxies: Clusters
DOI:	10.1086/430695
Bibliographic Code:	2005ApJ...627..733M

Links to e-resources

Links to data

Links to objects

Strasbourg CDS Datasets



Centre de Données astronomiques de Strasbourg

Strasbourg astronomical Data Center

Other services

X-match · Dictionary · Sesame · SimPlay

Hosted services

ADS mirror · A&A

TIPTOPbase
INES

Keep in touch

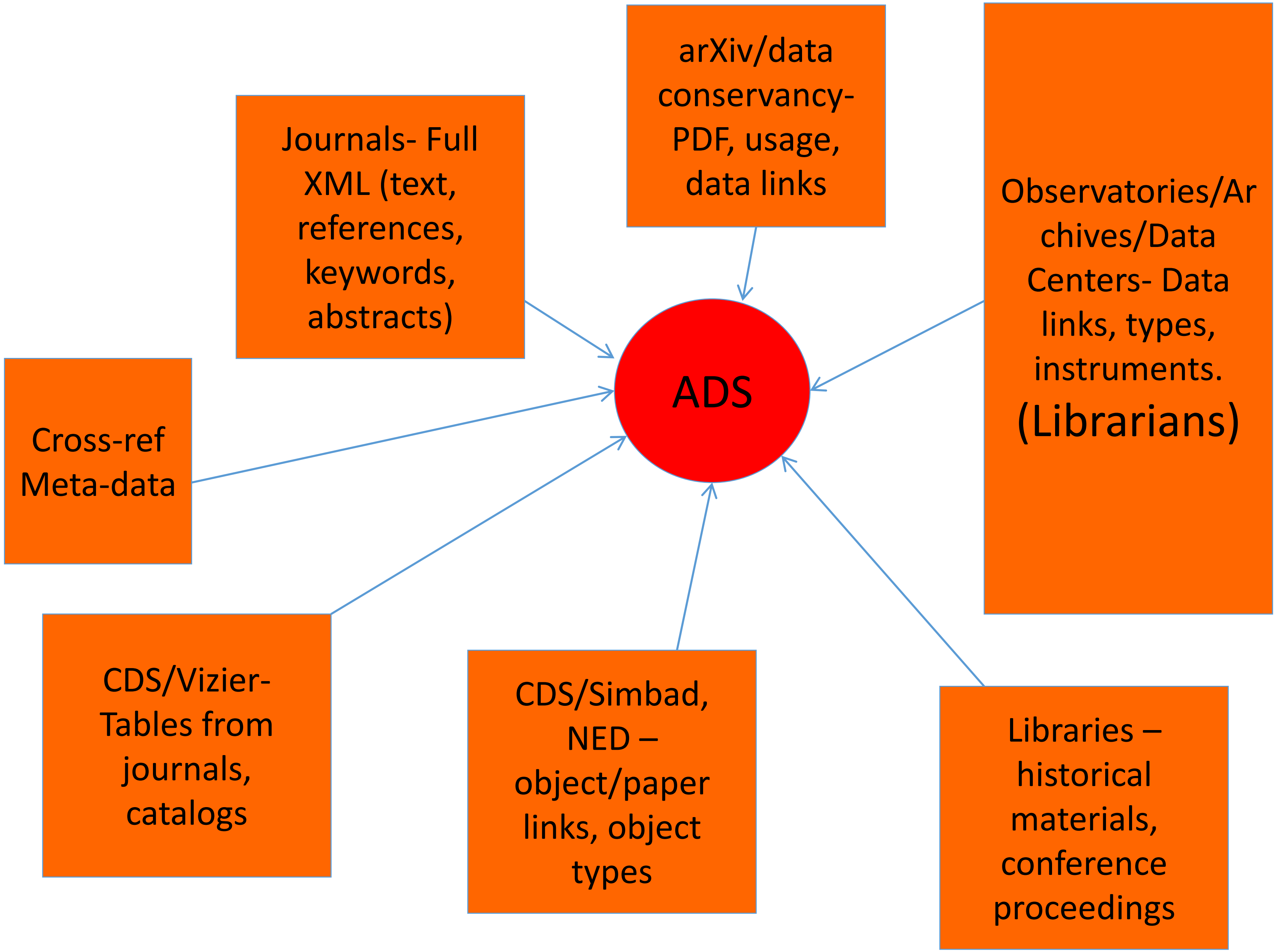
Latest news

- Catalogs added between 07-Sep-2013 and 14-Sep-2013
- Catalogs added between 31-Aug-2013 and 07-Sep-2013
- CDS services down on September 09 and 13
- Catalogs added between 24-Aug-2013 and 31-Aug-2013
- Aladin Lite released
- Use X-Match service for queries from coordinates
- Collaboration IAS / CDS
- PLANCK maps

More news

Featured news

On September 18 2012, CDS celebrated its 40 years.



ADS

Journals- Full XML (text, references, keywords, abstracts)

arXiv/data conservancy- PDF, usage, data links

Observatories/Archives/Data Centers- Data links, types, instruments.

(Librarians)

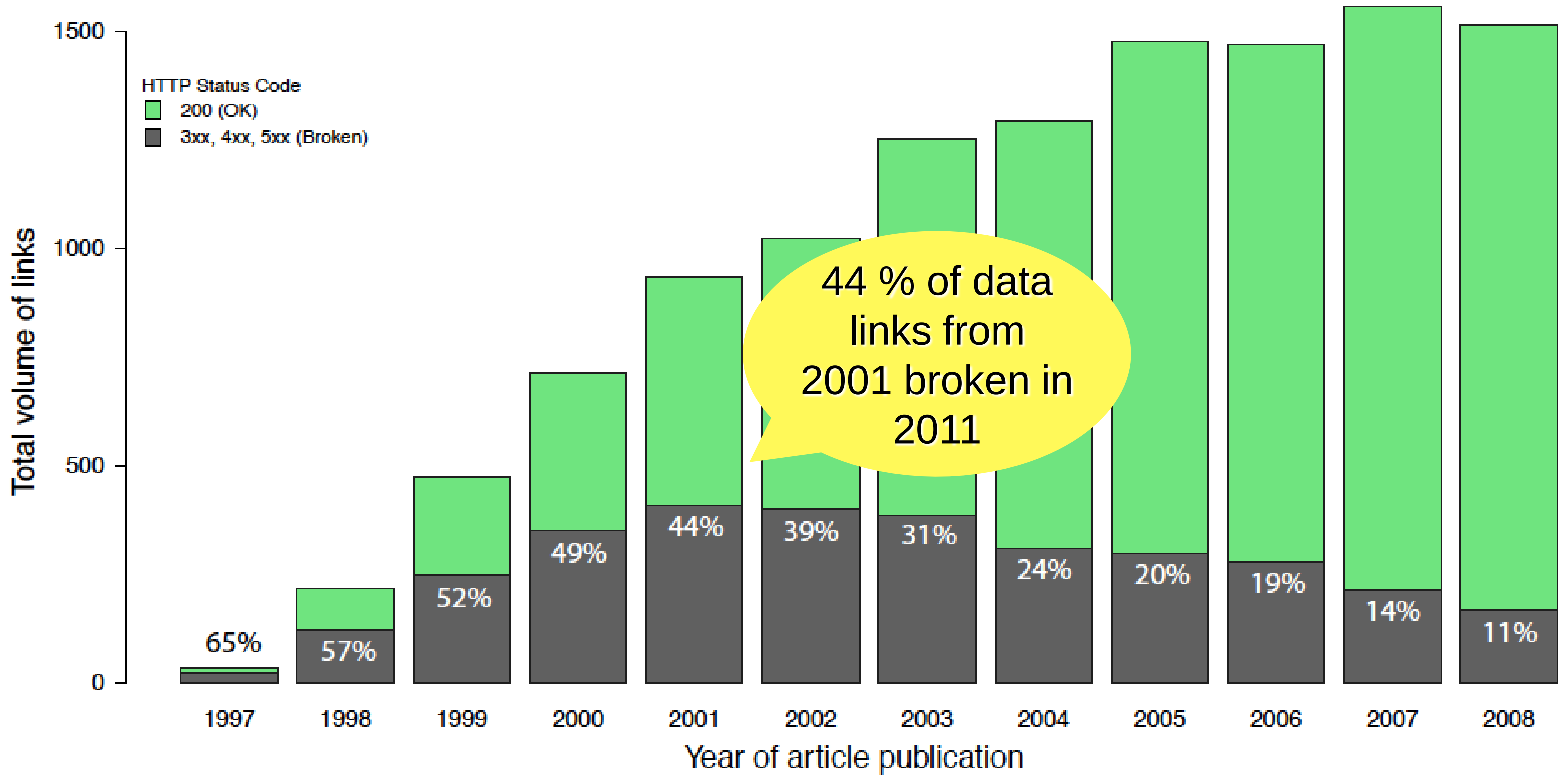
Cross-ref Meta-data

CDS/Vizier- Tables from journals, catalogs

CDS/Simbad, NED – object/paper links, object types

Libraries – historical materials, conference proceedings

But Sustainability of Data Links?



44 % of data links from 2001 broken in 2011

Figure 1. Volume of potential data links in astronomy publications. Total volume of external links in all articles published between 1997 and 2008 in the four main astronomy journals, color coded by HTTP status code. Green bars represent accessible links (200), grey bars represent broken links. .

Pepe et al. 2012

Progress since 2004?
The View from ADS (Michael Kurtz)

Comments:

- Does not see much progress in the last ten years: now one step back, waiting for the next two steps forward
- Ten years ago concerned that the Virtual Observatory would suffocate itself with bureaucracy: unfortunately this has now happened …
- New large repositories (Zenodo, Dataverse) are creating an infrastructure of almost entirely uncurated data

➢The problem with curation is that the funding is almost entirely local but in the digital world the use is mainly global. Leads to tragedy of the commons where no one will assume long-term obligation to curate and manage data which is mainly not from local sources.

Progress since 2004?
The View from CDS (Francoise Genova)

There are two major areas of progress:

- VO Framework
  - Interoperability framework with aspects on data description, formats, vocabularies, data models. Helps data producers share data so pay more attention to data curation and use elements of this framework.
- Long Tail Data
  - With the funding agency requirements on Data Management Plans and on making their data available, researchers more aware of importance of sharing data.
  - In astronomy, most original data from observations is in observatory archives, but at CDS we are seeing more data "attached to publications".

Evidence of Progress?

- Using only the numbers for A&A there were 200 papers with a "catalogue" attached in 2004, but 370 in 2013

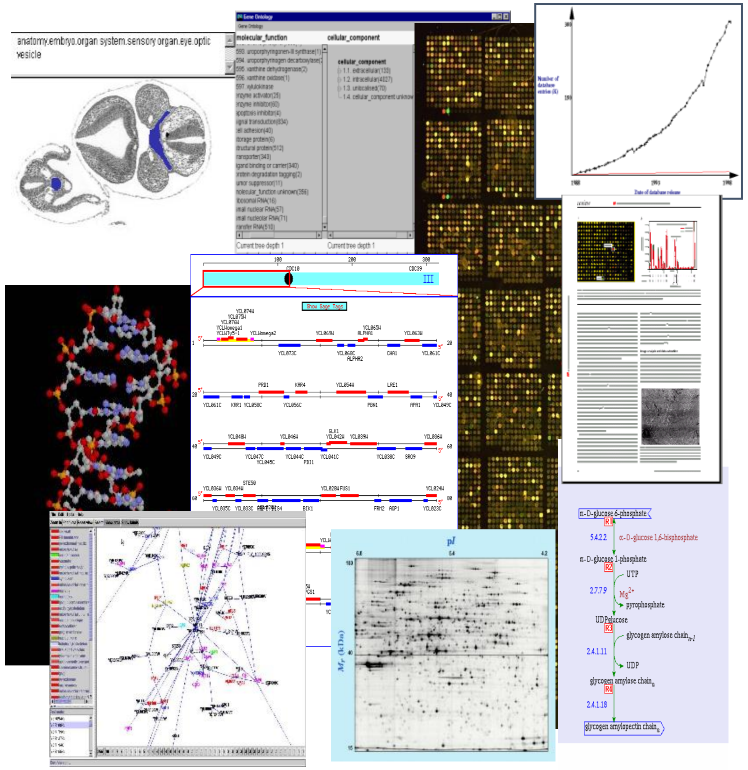
- Fewer than 40 catalogues with images, spectra and time series in 2004, but 195 in 2013.

- CDS developing added value services giving access to this new data resource

Bioinformatics

The myGrid Project

- Bioinformatics ‘Omics’ project
- Imminent ‘deluge’ of data
- Highly heterogeneous
- Highly complex and inter-related
- Convergence of data and literature archives



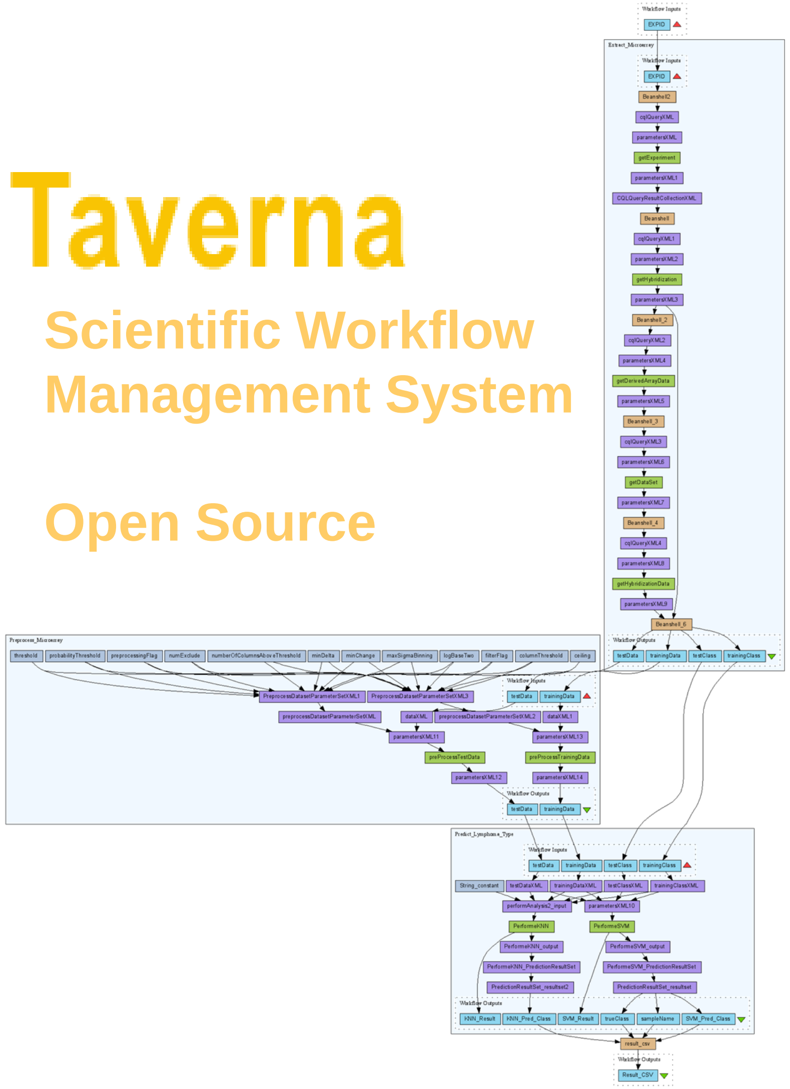
PI: Carole Goble, University of Manchester

Workflows: scientific computational pipelines for knowledge discovery

Taverna

Scientific Workflow Management System

Open Source



- Automated, repetitive discovery
- Agile, flexible platform
- Transparent, trackable processes
- Easy integration of

  data sets
- Advanced and scalable analytics
- End-user graphical interfaces
- Collaboration platform

http://www.taverna.org.uk

myexperiment

www.myexperiment.org

Socially share, discover and reuse workflows and other scientific methods.

Cooperative market place.

A scientific gateway.

Commons-based Production

Social curation of scientific assets

Social networking about content

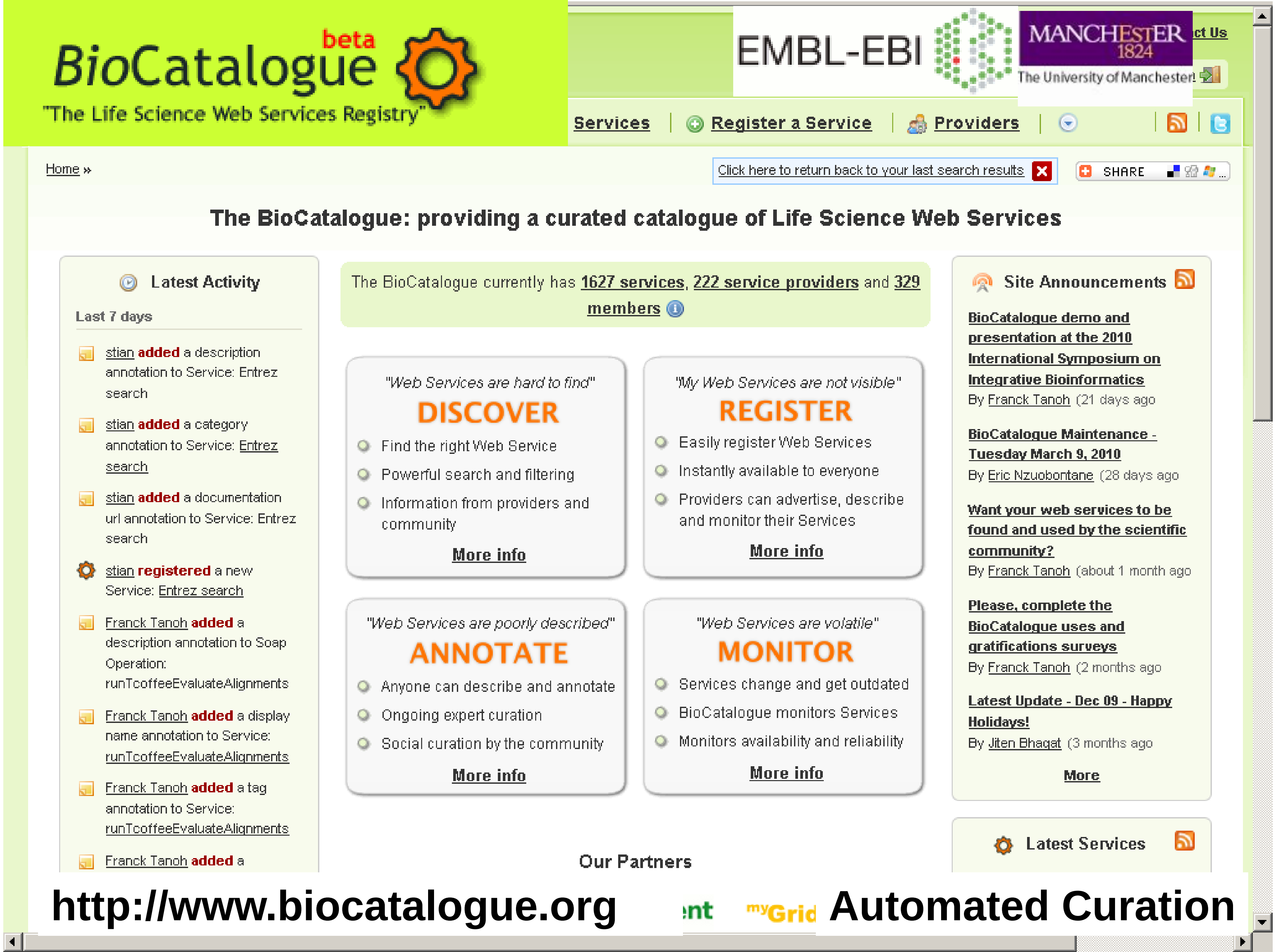
Where do I find workflows?
And other methods.

How do I connect with other authors and users?

Shared best practice.
Variant design.

Standing on the shoulders of peers.

Towards reproducible science.



Home »

Click here to return back to your last search results

The BioCatalogue: providing a curated catalogue of Life Science Web Services

Latest Activity

Last 7 days

- stian **added** a description annotation to Service: Entrez search
- stian **added** a category annotation to Service: Entrez search
- stian **added** a documentation url annotation to Service: Entrez search
- stian **registered** a new Service: Entrez search
- Franck Tanoh **added** a description annotation to Soap Operation: runTcoffeeEvaluateAlignments
- Franck Tanoh **added** a display name annotation to Service: runTcoffeeEvaluateAlignments
- Franck Tanoh **added** a tag annotation to Service: runTcoffeeEvaluateAlignments
- Franck Tanoh **added** a

The BioCatalogue currently has 1627 services, 222 service providers and 329 members

"Web Services are hard to find"

DISCOVER

- Find the right Web Service
- Powerful search and filtering
- Information from providers and community

More info

"My Web Services are not visible"

REGISTER

- Easily register Web Services
- Instantly available to everyone
- Providers can advertise, describe and monitor their Services

More info

"Web Services are poorly described"

ANNOTATE

- Anyone can describe and annotate
- Ongoing expert curation
- Social curation by the community

More info

"Web Services are volatile"

MONITOR

- Services change and get outdated
- BioCatalogue monitors Services
- Monitors availability and reliability

More info

Our Partners

Site Announcements

BioCatalogue demo and presentation at the 2010 International Symposium on Integrative Bioinformatics
By Franck Tanoh (21 days ago

BioCatalogue Maintenance - Tuesday March 9, 2010
By Eric Nzuobontane (28 days ago

Want your web services to be found and used by the scientific community?
By Franck Tanoh (about 1 month ago

Please, complete the BioCatalogue uses and gratifications surveys
By Franck Tanoh (2 months ago

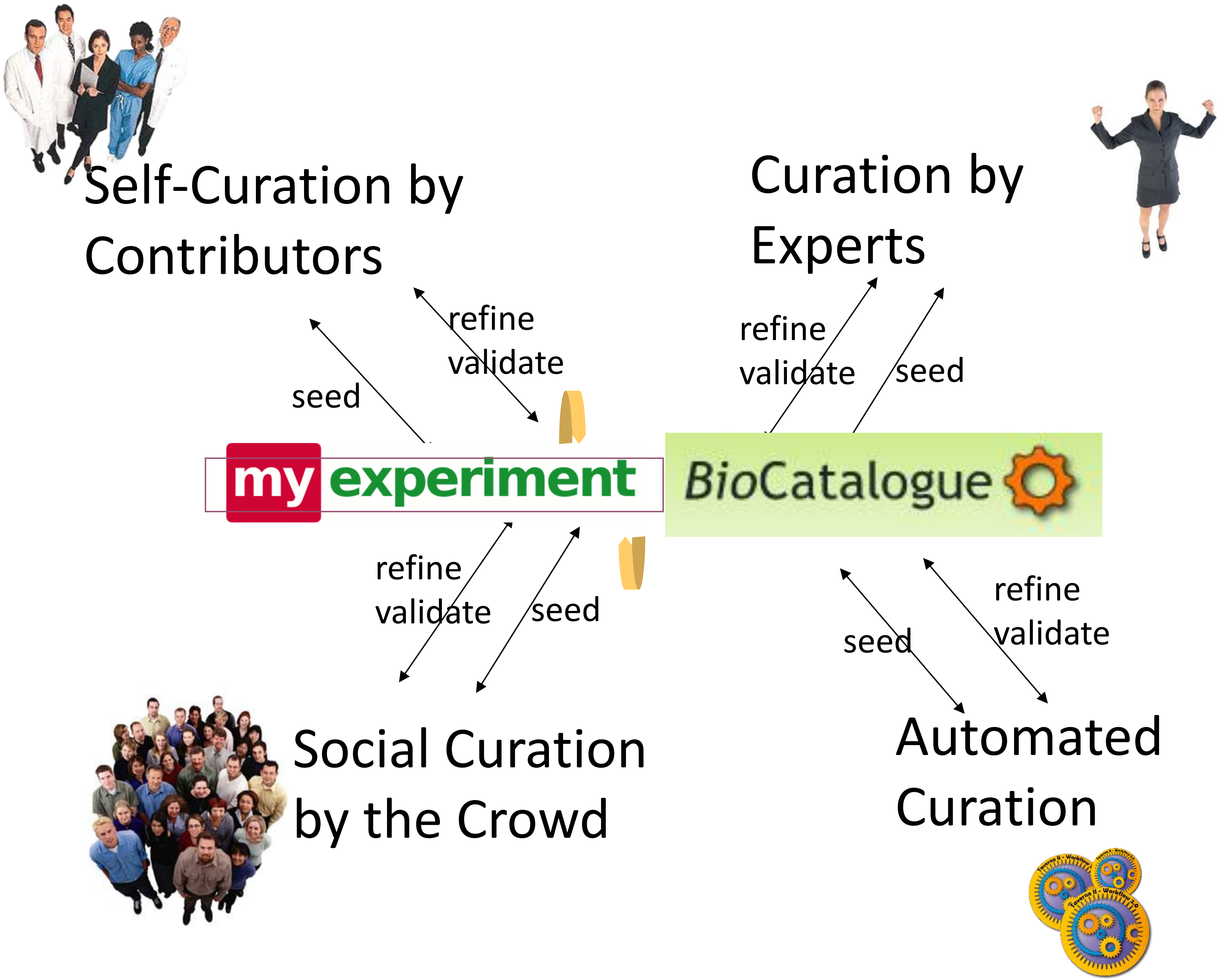
Latest Update - Dec 09 - Happy Holidays!
By Jiten Bhagat (3 months ago

More

Latest Services

http://www.biocatalogue.org

Automated Curation



Self-Curation by Contributors

Curation by Experts

seed

refine validate

refine validate

seed

myexperiment

BioCatalogue

refine validate

seed

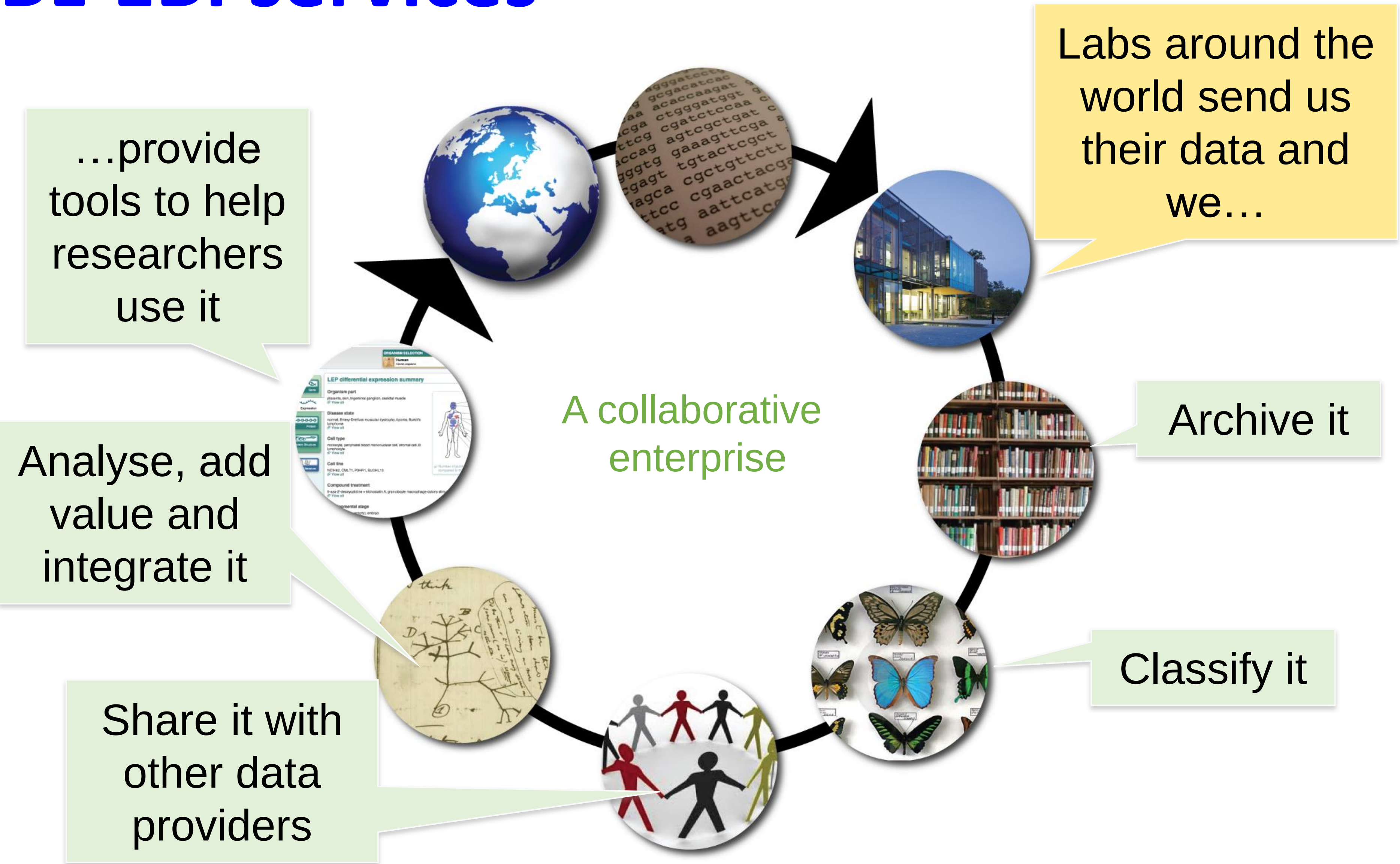
seed

refine validate

Social Curation by the Crowd

Automated Curation

EMBL-EBI services



Labs around the world send us their data and we…

Archive it

Classify it

Share it with other data providers

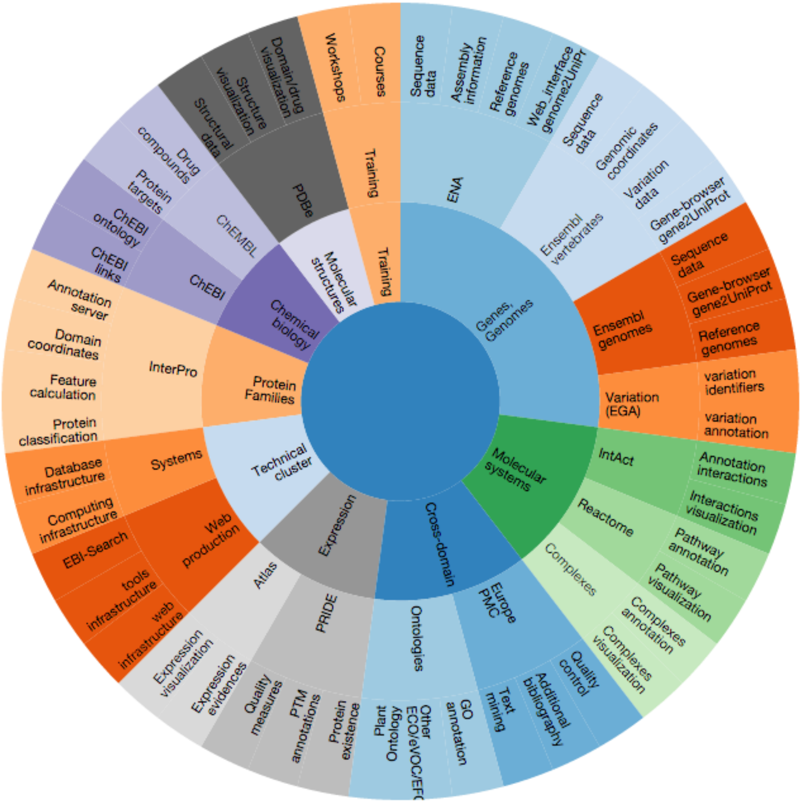
Analyse, add value and integrate it

…provide tools to help researchers use it

A collaborative enterprise

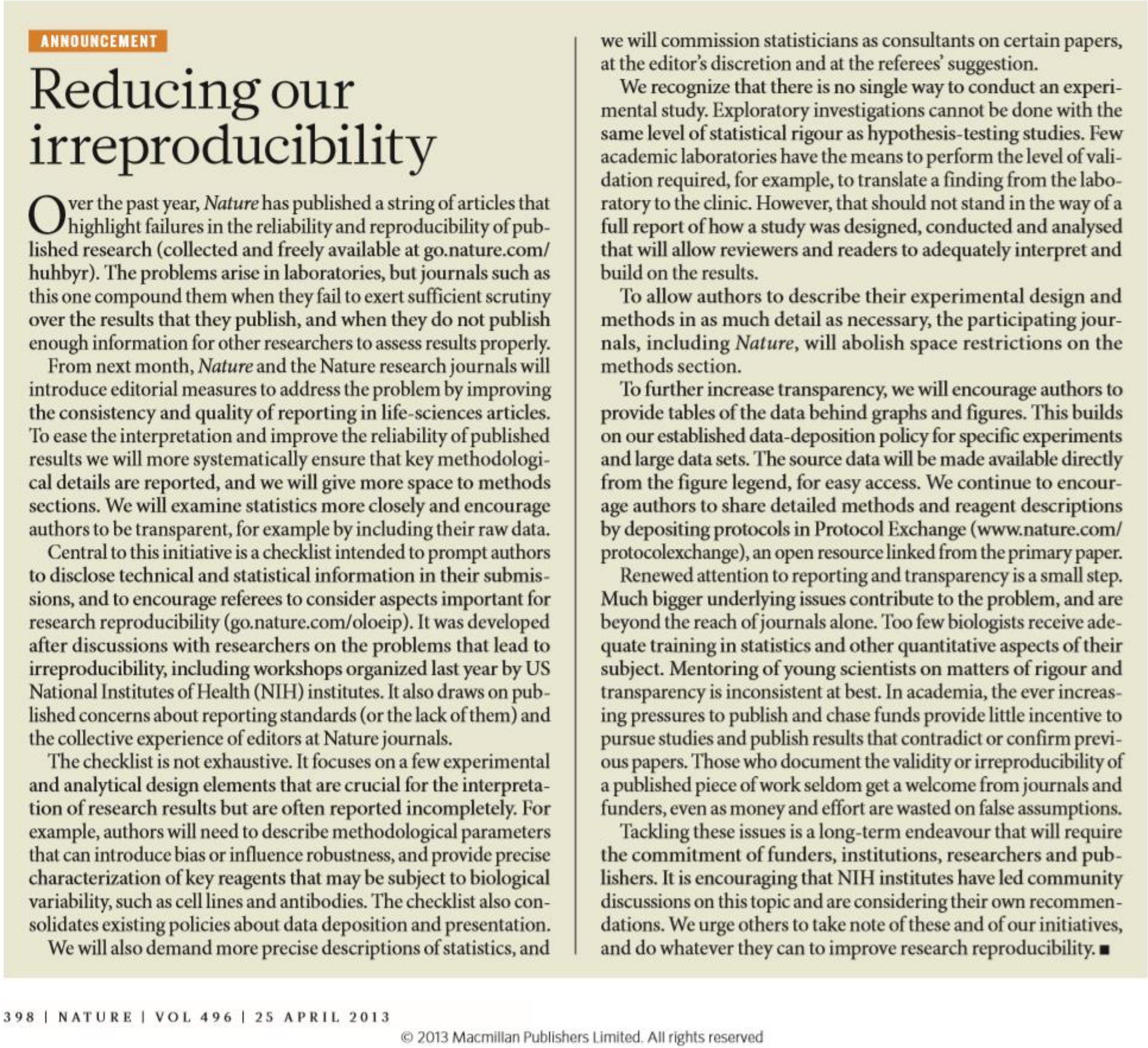
Sharing Annotation: The UniProt example

- Total UniProt operation has 100 FTEs with around 40 curators
- Purely archival databases have tens of thousands of users, value-added ones like UniProt have millions of users.
- The literature used for curation in UniProt is around 10,000 papers a year, valued at about €1B of research funding
- Full costs of the UniProt - around €12M a year – are small compared to cost of research and number of scientists served



But now serious problems of research reproducibility in bioinformatics

- Review of 2,047 retracted articles indexed in PubMed in May of 2012 concluded that:
  - 21.3% were retracted because of errors,
  - 67.4% were retracted because of scientific misconduct
    - Fraud or suspected fraud (43.4%)
    - Duplicate publication (14.2%)
    - Plagiarism (9.8%)
- Study by pharma companies Bayer and Amgen concluded that between 60% and 70% of biomedicine studies may be non-reproducible
  - Amgen scientists were only able to reproduce 7 out of 53 cancer results published in Science and Nature



ANNOUNCEMENT

Reducing our irreproducibility

Over the past year, *Nature* has published a string of articles that highlight failures in the reliability and reproducibility of published research (collected and freely available at go.nature.com/huhbyr). The problems arise in laboratories, but journals such as this one compound them when they fail to exert sufficient scrutiny over the results that they publish, and when they do not publish enough information for other researchers to assess results properly.

From next month, *Nature* and the Nature research journals will introduce editorial measures to address the problem by improving the consistency and quality of reporting in life-sciences articles. To ease the interpretation and improve the reliability of published results we will more systematically ensure that key methodological details are reported, and we will give more space to methods sections. We will examine statistics more closely and encourage authors to be transparent, for example by including their raw data.

Central to this initiative is a checklist intended to prompt authors to disclose technical and statistical information in their submissions, and to encourage referees to consider aspects important for research reproducibility (go.nature.com/oloeip). It was developed after discussions with researchers on the problems that lead to irreproducibility, including workshops organized last year by US National Institutes of Health (NIH) institutes. It also draws on published concerns about reporting standards (or the lack of them) and the collective experience of editors at Nature journals.

The checklist is not exhaustive. It focuses on a few experimental and analytical design elements that are crucial for the interpretation of research results but are often reported incompletely. For example, authors will need to describe methodological parameters that can introduce bias or influence robustness, and provide precise characterization of key reagents that may be subject to biological variability, such as cell lines and antibodies. The checklist also consolidates existing policies about data deposition and presentation.

We will also demand more precise descriptions of statistics, and we will commission statisticians as consultants on certain papers, at the editor's discretion and at the referees' suggestion.

We recognize that there is no single way to conduct an experimental study. Exploratory investigations cannot be done with the same level of statistical rigour as hypothesis-testing studies. Few academic laboratories have the means to perform the level of validation required, for example, to translate a finding from the laboratory to the clinic. However, that should not stand in the way of a full report of how a study was designed, conducted and analysed that will allow reviewers and readers to adequately interpret and build on the results.

To allow authors to describe their experimental design and methods in as much detail as necessary, the participating journals, including *Nature*, will abolish space restrictions on the methods section.

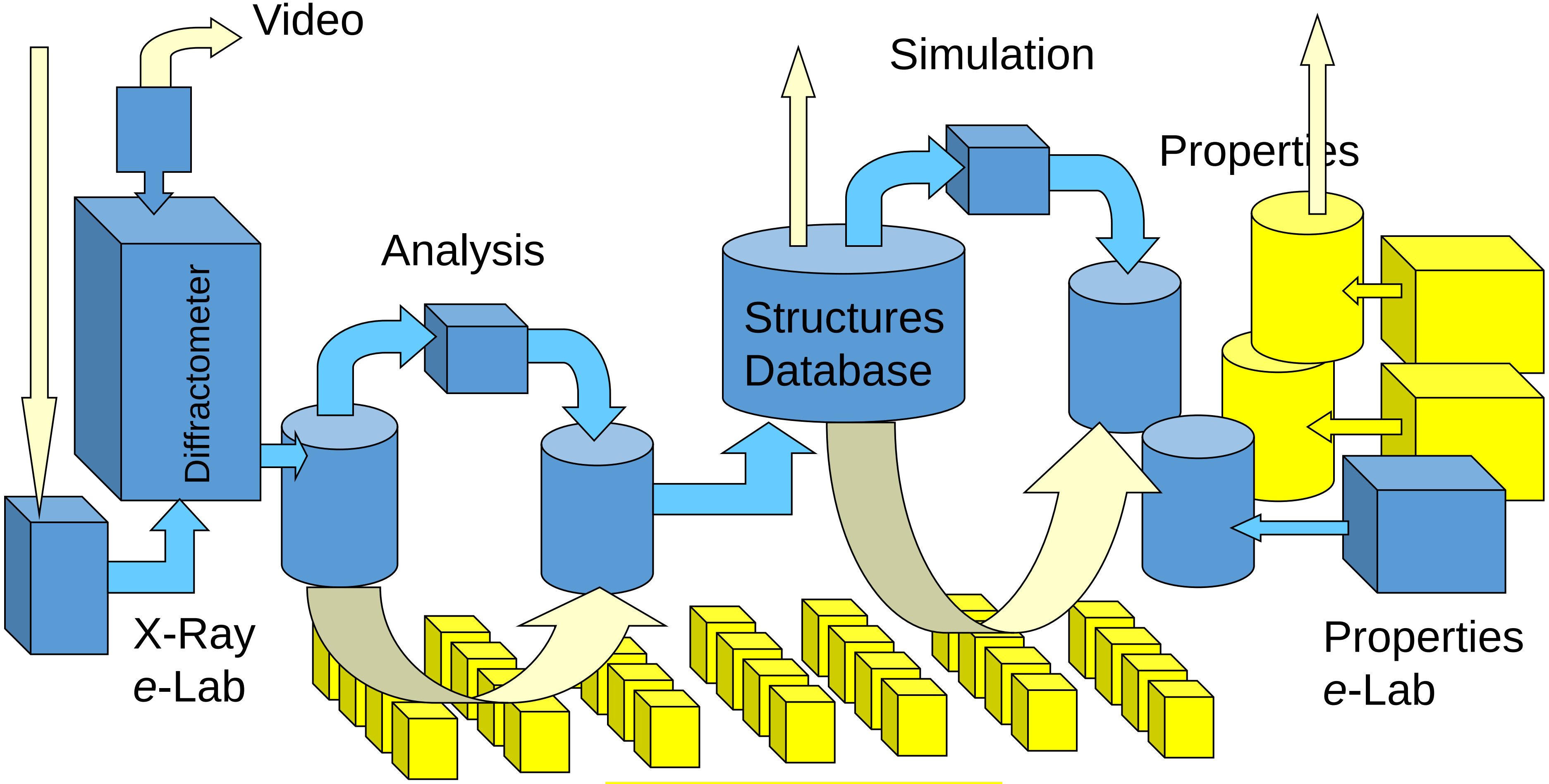
To further increase transparency, we will encourage authors to provide tables of the data behind graphs and figures. This builds on our established data-deposition policy for specific experiments and large data sets. The source data will be made available directly from the figure legend, for easy access. We continue to encourage authors to share detailed methods and reagent descriptions by depositing protocols in Protocol Exchange (www.nature.com/protocolexchange), an open resource linked from the primary paper.

Renewed attention to reporting and transparency is a small step. Much bigger underlying issues contribute to the problem, and are beyond the reach of journals alone. Too few biologists receive adequate training in statistics and other quantitative aspects of their subject. Mentoring of young scientists on matters of rigour and transparency is inconsistent at best. In academia, the ever increasing pressures to publish and chase funds provide little incentive to pursue studies and publish results that contradict or confirm previous papers. Those who document the validity or irreproducibility of a published piece of work seldom get a welcome from journals and funders, even as money and effort are wasted on false assumptions.

Tackling these issues is a long-term endeavour that will require the commitment of funders, institutions, researchers and publishers. It is encouraging that NIH institutes have led community discussions on this topic and are considering their own recommendations. We urge others to take note of these and of our initiatives, and do whatever they can to improve research reproducibility. ■

Chemistry

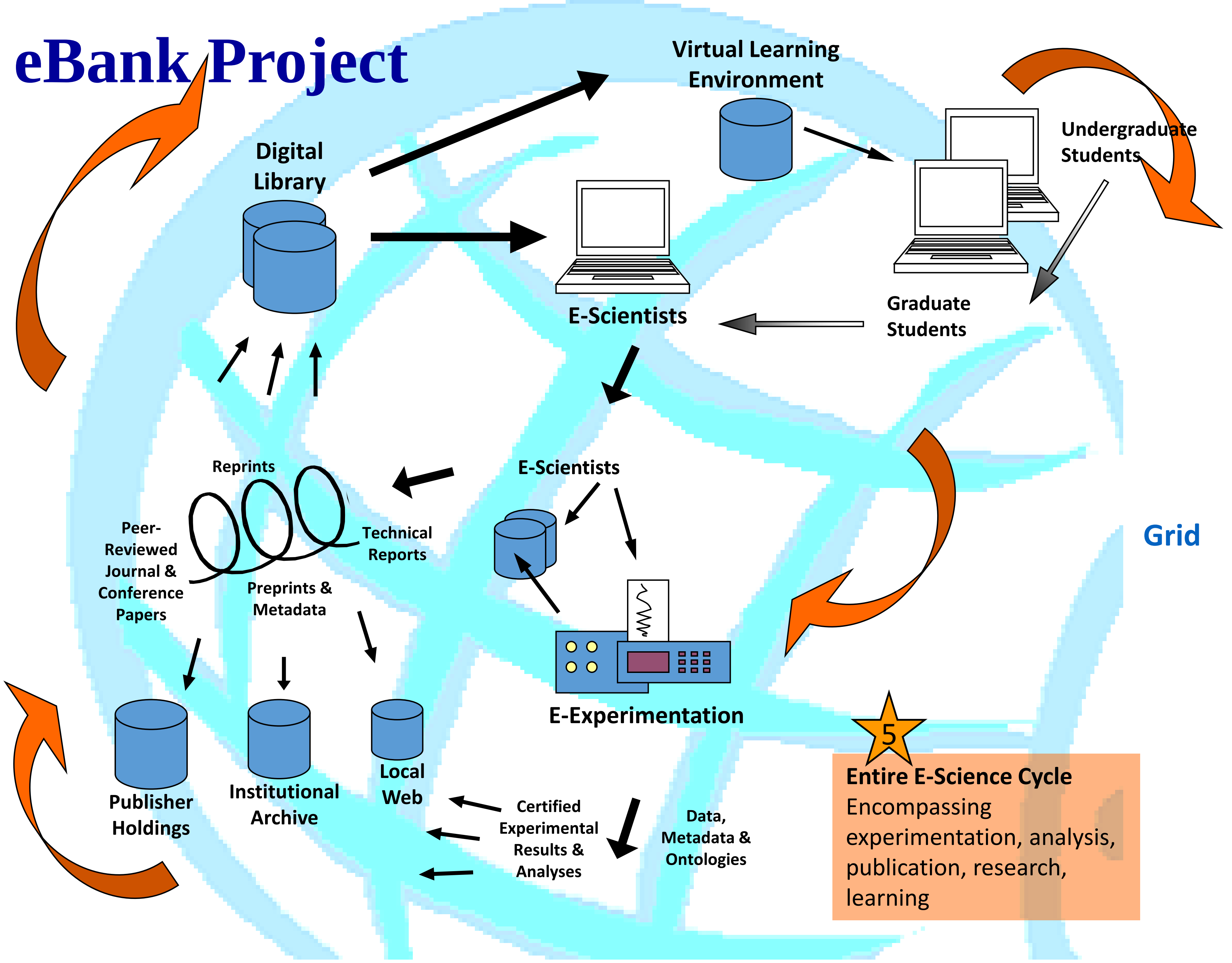
The Comb-*e*-Chem Project



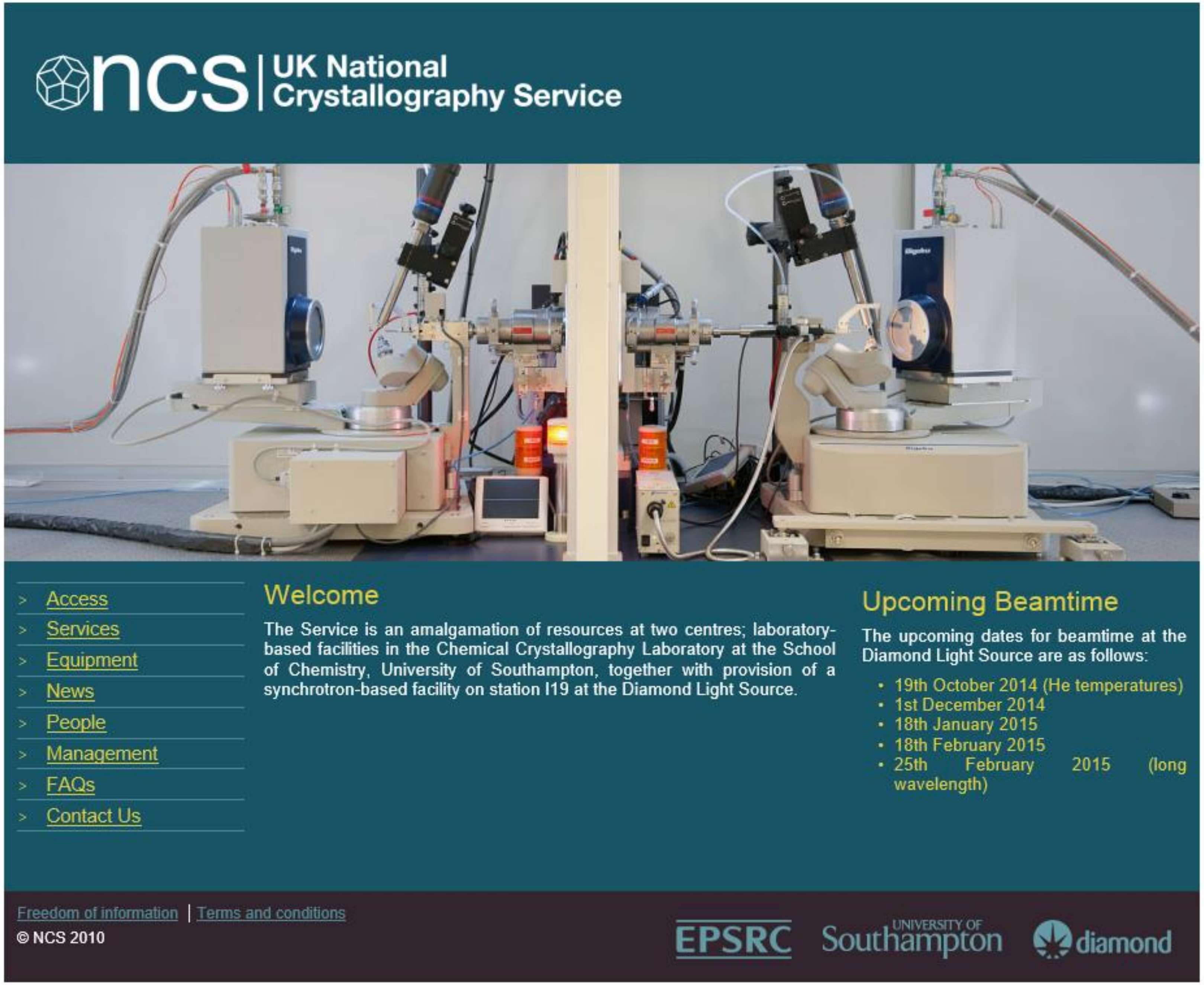
Grid Middleware

PI: Jeremy Frey, University of Southampton

eBank Project



Entire E-Science Cycle
Encompassing experimentation, analysis, publication, research, learning


ncs | UK National Crystallography Service
> Access
> Services
> Equipment
> News
> People
> Management
> FAQs
> Contact Us
Freedom of information | Terms and conditions
EPSRC
UNIVERSITY OF Southampton
diamond

Welcome

The Service is an amalgamation of resources at two centres; laboratory-based facilities in the Chemical Crystallography Laboratory at the School of Chemistry, University of Southampton, together with provision of a synchrotron-based facility on station I19 at the Diamond Light Source.

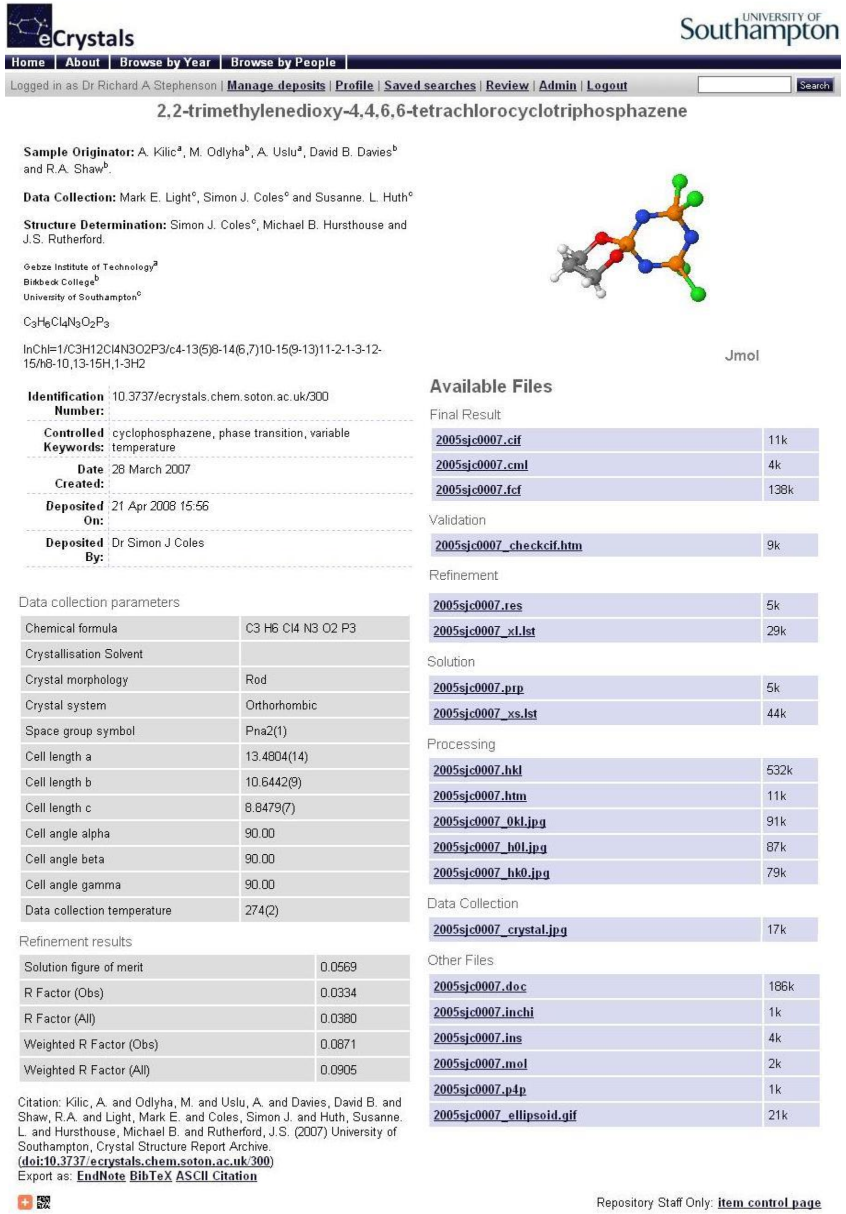
Upcoming Beamtime

The upcoming dates for beamtime at the Diamond Light Source are as follows:

- 19th October 2014 (He temperatures)
- 1st December 2014
- 18th January 2015
- 18th February 2015
- 25th February 2015 (long wavelength)

© NCS 2010

The eCrystals Data Repository


eCrystals
UNIVERSITY OF Southampton
Home | About | Browse by Year | Browse by People
Search
Jmol

Logged in as Dr Richard A Stephenson | **Manage deposits** | **Profile** | **Saved searches** | **Review** | **Admin** | **Logout**

2,2-trimethylenedioxy-4,4,6,6-tetrachlorocyclotriphosphazene

Sample Originator: A. Kilic[a], M. Odlyha[b], A. Uslu[a], David B. Davies[b] and R.A. Shaw[b].

Data Collection: Mark E. Light[c], Simon J. Coles[c] and Susanne. L. Huth[c]

Structure Determination: Simon J. Coles[c], Michael B. Hursthouse and J.S. Rutherford.

Gebze Institute of Technology[a]
Birkbeck College[b]
University of Southampton[c]

$C_3H_6Cl_4N_3O_2P_3$

InChI=1/C3H12Cl4N3O2P3/c4-13(5)8-14(6,7)10-15(9-13)11-2-1-3-12-15/h8-10,13-15H,1-3H2

Identification Number:	10.3737/ecrystals.chem.soton.ac.uk/300
Controlled Keywords:	cyclophosphazene, phase transition, variable temperature
Date Created:	28 March 2007
Deposited On:	21 Apr 2008 15:56
Deposited By:	Dr Simon J Coles

Data collection parameters

Chemical formula	C3 H6 Cl4 N3 O2 P3
Crystallisation Solvent	
Crystal morphology	Rod
Crystal system	Orthorhombic
Space group symbol	Pna2(1)
Cell length a	13.4804(14)
Cell length b	10.6442(9)
Cell length c	8.8479(7)
Cell angle alpha	90.00
Cell angle beta	90.00
Cell angle gamma	90.00
Data collection temperature	274(2)

Refinement results

Solution figure of merit	0.0569
R Factor (Obs)	0.0334
R Factor (All)	0.0380
Weighted R Factor (Obs)	0.0871
Weighted R Factor (All)	0.0905

Citation: Kilic, A. and Odlyha, M. and Uslu, A. and Davies, David B. and Shaw, R.A. and Light, Mark E. and Coles, Simon J. and Huth, Susanne. L. and Hursthouse, Michael B. and Rutherford, J.S. (2007) University of Southampton, Crystal Structure Report Archive. (**doi:10.3737/ecrystals.chem.soton.ac.uk/300**)
Export as: **EndNote** **BibTeX** **ASCII Citation**

Available Files

Final Result

2005sjc0007.cif	11k
2005sjc0007.cml	4k
2005sjc0007.fcf	138k

Validation

2005sjc0007_checkcif.htm	9k

Refinement

2005sjc0007.res	5k
2005sjc0007_xl.lst	29k

Solution

2005sjc0007.prp	5k
2005sjc0007_xs.lst	44k

Processing

2005sjc0007.hkl	532k
2005sjc0007.htm	11k
2005sjc0007_0kl.jpg	91k
2005sjc0007_h0l.jpg	87k
2005sjc0007_hk0.jpg	79k

Data Collection

2005sjc0007_crystal.jpg	17k

Other Files

2005sjc0007.doc	186k
2005sjc0007.inchi	1k
2005sjc0007.ins	4k
2005sjc0007.mol	2k
2005sjc0007.p4p	1k
2005sjc0007_ellipsoid.gif	21k

Repository Staff Only: **item control page**

- Quick & simple to deposit
- Software tools
- Laboratory archive
- Community involvement
- ‘Embargo’ facility
- Structured foundations
- Discoverable & harvestable

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

If only I knew exactly how she did this experiments

I wish I had recorded things at the start the way I do now…..

I know all this supplementary information could be useful but will people really remember the format? Is it worth all the hassle?

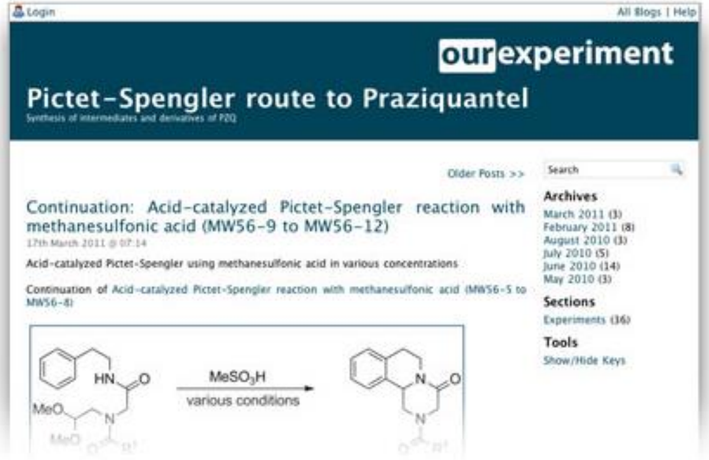
I wish I could get the numbers from this graph - the pdf is not much use.

Typical Laboratory

LabTrove

©flickr.com/julia_manzerova

- > About Us
- > Get LabTrove
- > Documentation
- > Support
- > Publications
- > Users
- > Contact Us



Pictet-Spengler route to Praziquantel

Synthesis of intermediates and derivatives of PZQ

Older Posts >>

Continuation: Acid-catalyzed Pictet-Spengler reaction with methanesulfonic acid (MW56-9 to MW56-12)

17th March 2011 @ 07:14

Acid-catalyzed Pictet-Spengler using methanesulfonic acid in various concentrations

Continuation of Acid-catalyzed Pictet-Spengler reaction with methanesulfonic acid (MW56-5 to MW56-8)

Archives

March 2011 (3)
February 2011 (8)
August 2010 (3)
July 2010 (5)
June 2010 (14)
May 2010 (3)

Sections

Experiments (36)

Tools

Show/Hide Keys

LabTrove enables the formation of a Smart Research Framework, helping the creation and preservation of the record

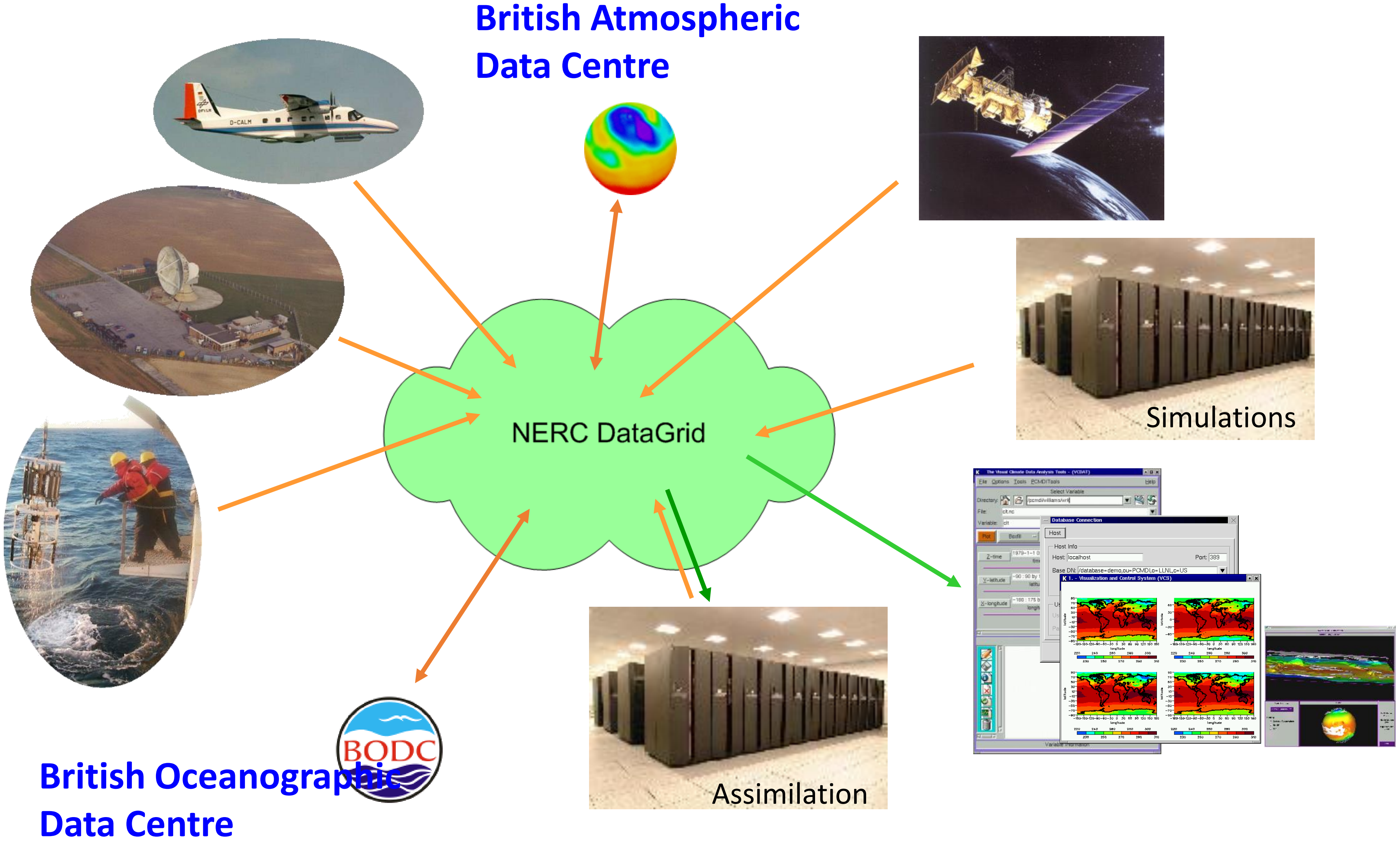
Progress?

Comments from Simon Coles:

- Major advance is usability which changes user attitude and behaviour
- Most scientific digital infrastructures only now adopting ‘Web2.0’ technologies
- This culture change has made adoption of standards to enable integration, interoperability and functionality
- Everything can now be linked up in logical and easy to use workflows
- Mobile devices are on the verge of transforming the ‘doing and recording’ of science

Environmental Science

The NERC DataGrid Project



British Atmospheric Data Centre

NERC DataGrid

Simulations

Assimilation

British Oceanographic Data Centre

PI: Bryan Lawrence, CLRC

Science & Technology
Facilities Council

www.stfc.ac.uk

From Citation & Location to Data Publication & Research Object: the CLADDIER Legacy

Catherine Jones, Shirley Crompton, Brian Matthews, Antony Wilson, Sarah Callaghan, Sam Pepler and Bryan Lawrence
Scientific Computing Department, Science and Technology Facilities Council, Centre for Environmental Data Archival Science and Technology Facilities Council and NCAS, Department of Meteorology University of Reading

CLADDIER: Citation, Location, And Deposition in Discipline and Institutional Repositories (2005-2007

USE CASE: A scientist is researching the biology of seawater. As part of her analysis she uses existing publications and data together with data from her project.

On completion she publishes a paper, citing the publications and datasets used and lodges her own publication and datasets in appropriate repositories. This new work is of interest to another scientist and he will be able to find both the publication and the datasets used through the mutual citations held in the two repositories.

OUTPUTS:

- A format for citing data was proposed and the issues of data publication, dataset definition and citation of dynamic data were considered.
- The CLADDIER Discovery Service enabled federated searches across both publication and data repositories
- A citation notification scheme based on the TrackBack protocol creating direct bidirectional links between repositories

Welcome to the CLADDIER Discovery Service 0.0 alpha

Search

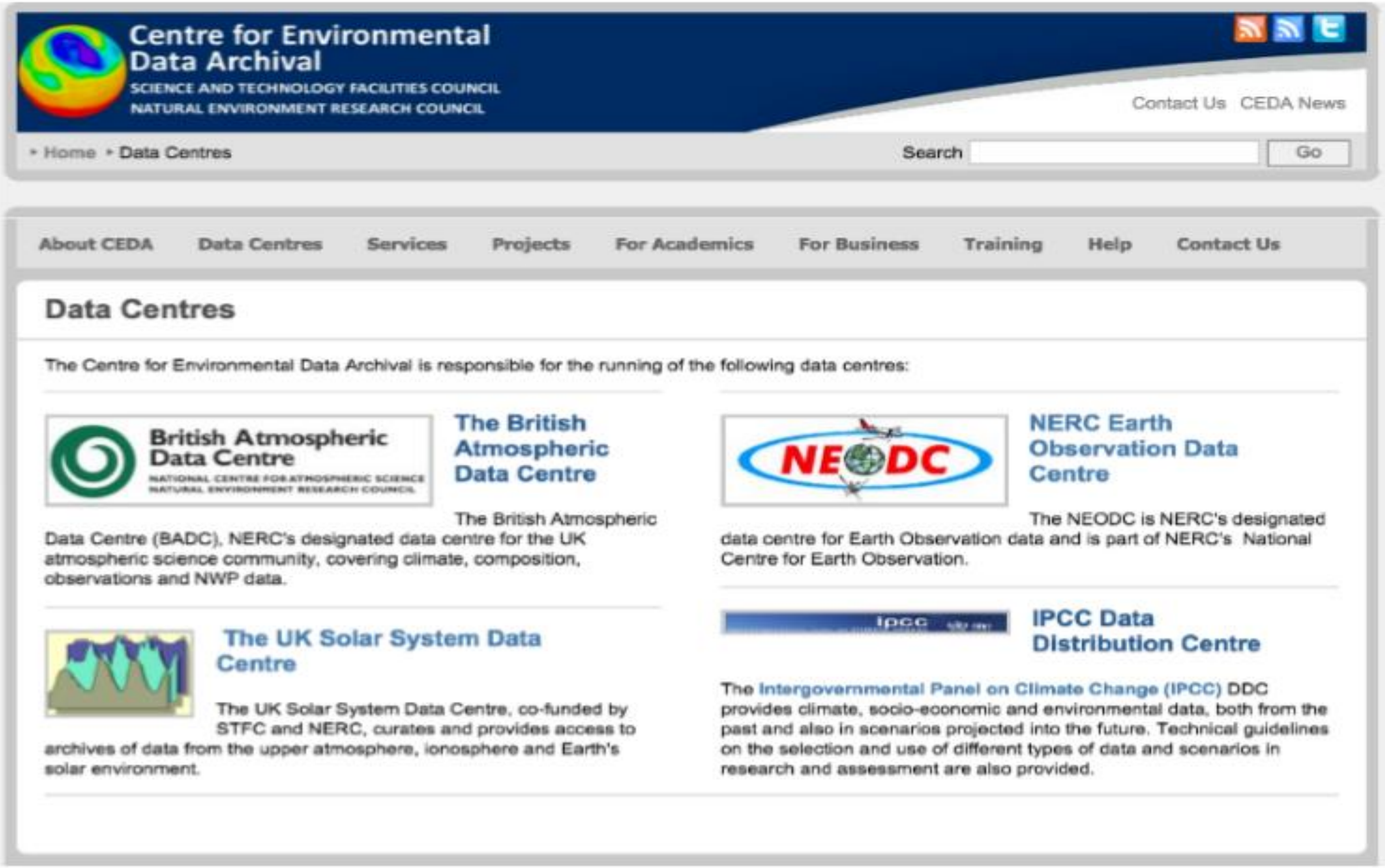
Global ocean mean wave period data: validation and description

Reducing uncertainty in calculations of air-sea flux of carbon dioxide using scatterometer, passive microwave, and model re-analysis wind data

ECMWF Operational Analyses

ECMWF 40 year re-analysis data (ERA-40)

CEDA


Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL
Contact Us
CEDA News
Home
Data Centres
Search
Go
About CEDA
Data Centres
Services
Projects
For Academics
For Business
Training
Help
Contact Us
British Atmospheric Data Centre
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL
NEODC
ipcc

Data Centres

The Centre for Environmental Data Archival is responsible for the running of the following data centres:

The British Atmospheric Data Centre

The British Atmospheric Data Centre (BADC), NERC's designated data centre for the UK atmospheric science community, covering climate, composition, observations and NWP data.

NERC Earth Observation Data Centre

The NEODC is NERC's designated data centre for Earth Observation data and is part of NERC's National Centre for Earth Observation.

The UK Solar System Data Centre

The UK Solar System Data Centre, co-funded by STFC and NERC, curates and provides access to archives of data from the upper atmosphere, ionosphere and Earth's solar environment.

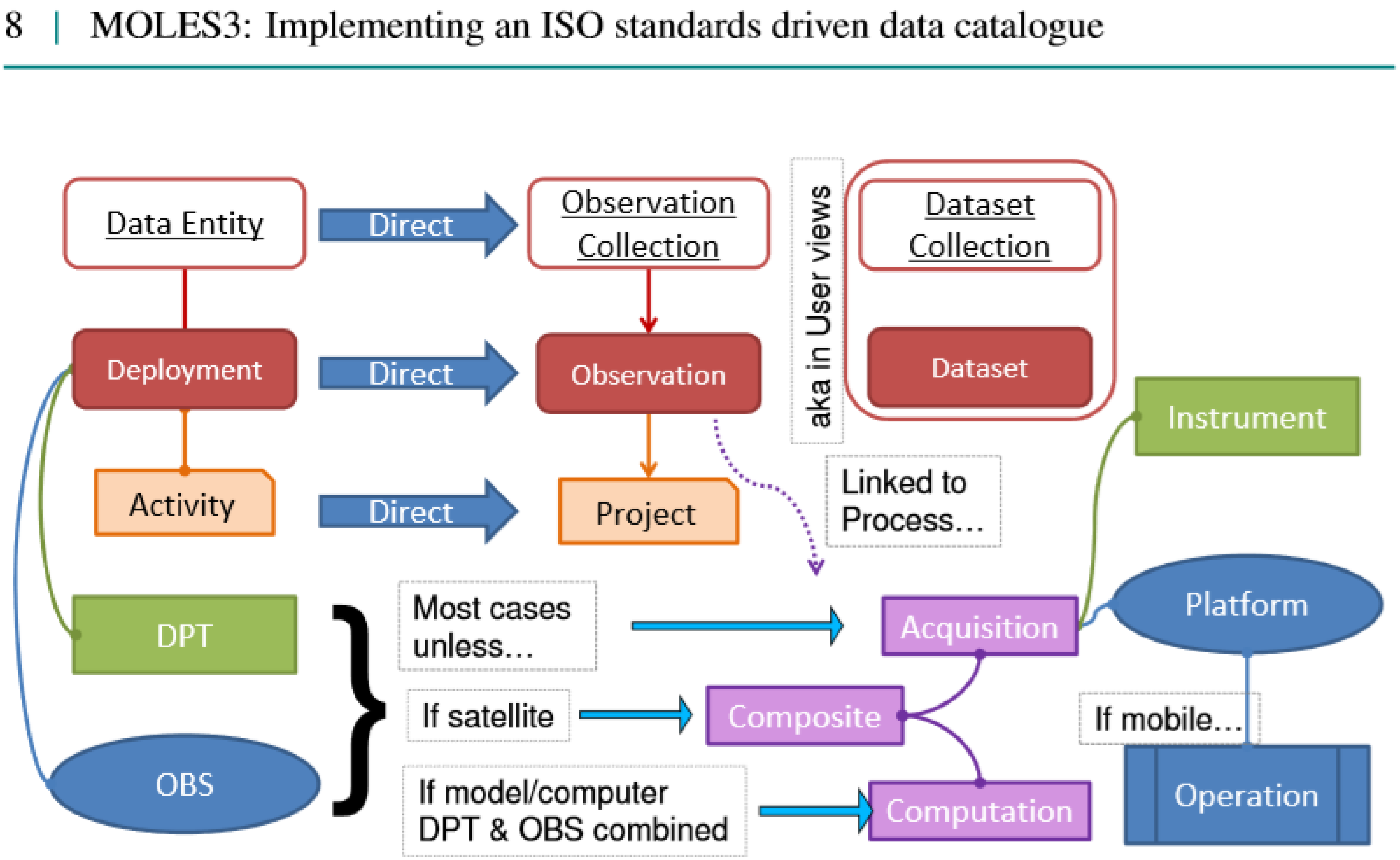
IPCC Data Distribution Centre

The Intergovernmental Panel on Climate Change (IPCC) DDC provides climate, socio-economic and environmental data, both from the past and also in scenarios projected into the future. Technical guidelines on the selection and use of different types of data and scenarios in research and assessment are also provided.

Four data centres: http://ceda.ac.uk

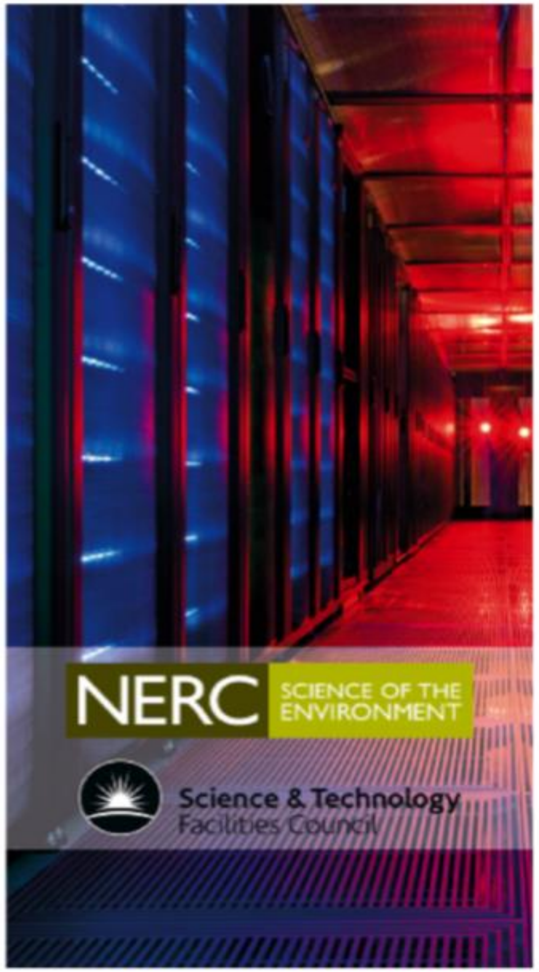
Providing Curation (of the archive)

MOLES: Metadata Objects for Linking Environmental Sciences



8 | MOLES3: Implementing an ISO standards driven data catalogue

So we have built a Data-Intensive HPC cloud: JASMIN


NERC SCIENCE OF THE ENVIRONMENT
Science & Technology Facilities Council

- 12 PB Fast Storage
- 1 PB Bulk Storage
- Elastic Tape
- 4000 cores: half deployed as hypervisors, half as the "Lotus" batch cluster.

Progress in Environmental Data Curation?

Professor James Frew (UCSB):

- Biggest change is funding agency mandate.
- NSF’s Data Management Plan for all proposals has made scientists (pretend?) to take data curation seriously.
- There are better curated databases and metadata now - but not sure that quality fraction is increasing!
- Frew’s first law: scientists don’t write metadata
- Frew’s second law: any scientist can be forced to write bad metadata

➢ Should automate creation of metadata as far as possible

➢ Scientists need to work with metadata specialists with domain knowledge

Ecology

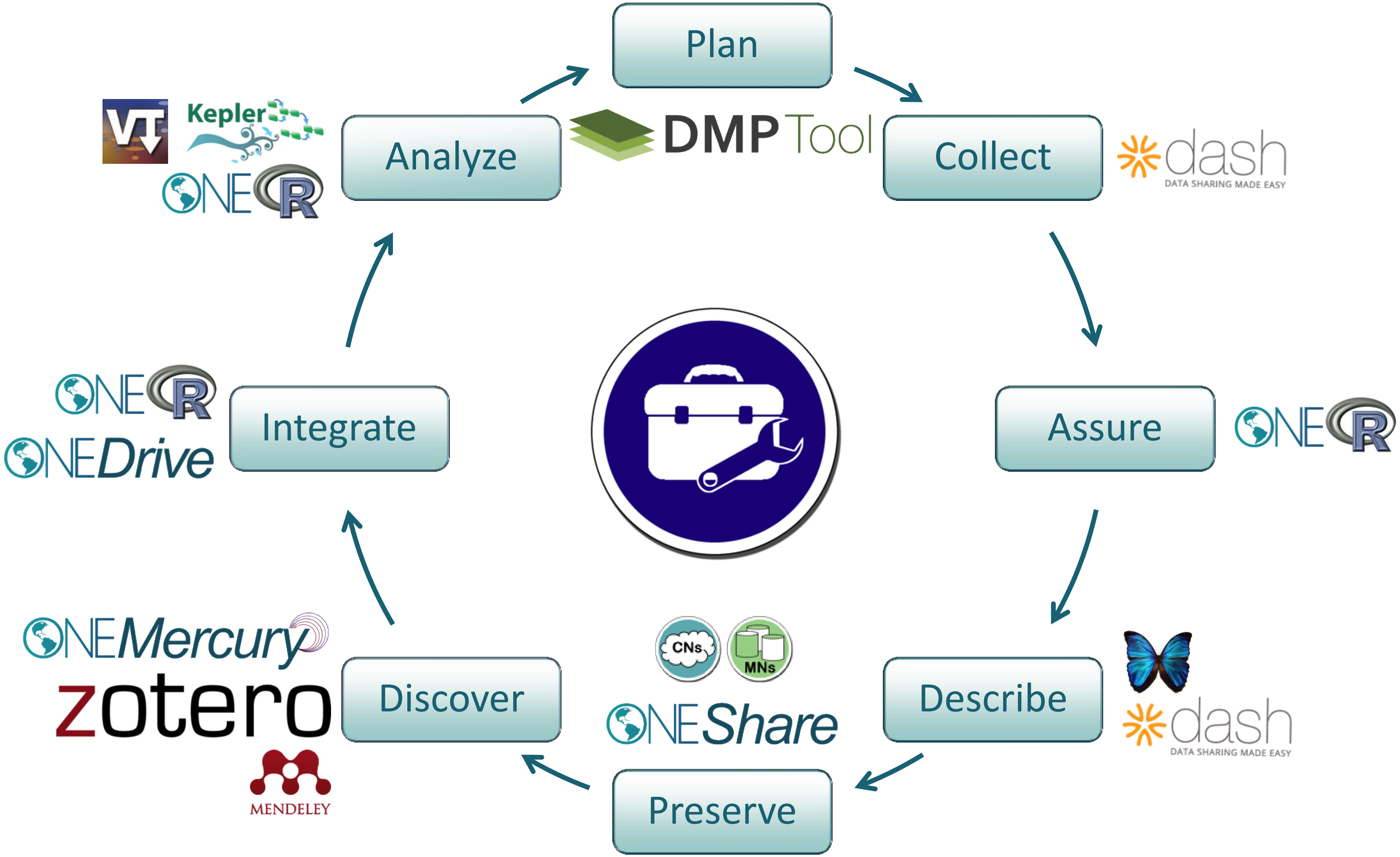
DataONE: Data Observation Network for Earth

Bill Michener

College of University Libraries & Learning Sciences, University of New Mexico

DataONE
Data Observation Network for Earth

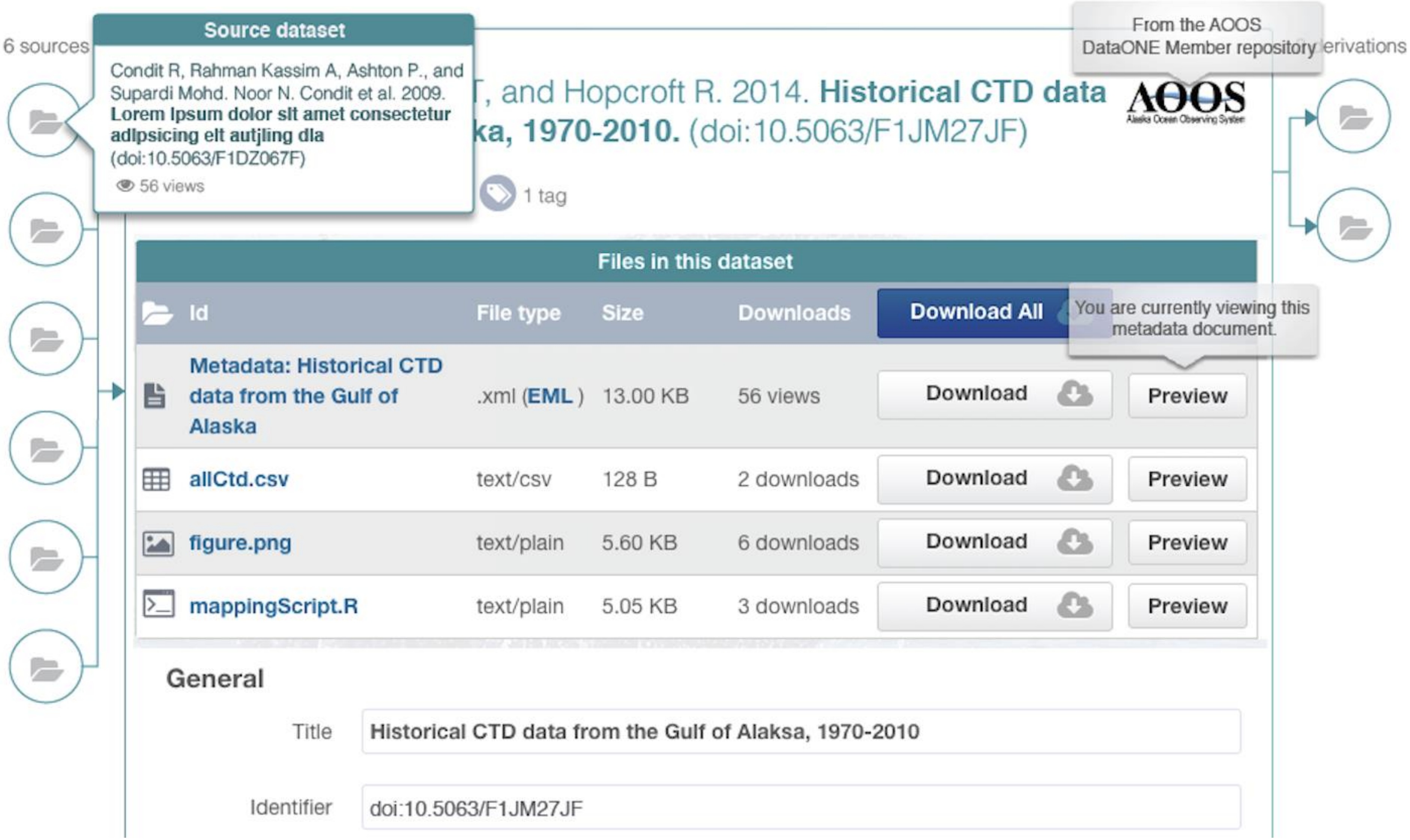
Enabling Science through Tools and Services



DataONE

Data Observation Network for Earth

Provenance tracking and display



6 sources

Source dataset

Condit R, Rahman Kassim A, Ashton P., and Supardi Mohd. Noor N. Condit et al. 2009. **Lorem Ipsum dolor slt amet consectetur adlpsicing elt autjling dla** (doi:10.5063/F1DZ067F)

56 views

From the AOOS DataONE Member repository

erivations

T, and Hopcroft R. 2014. **Historical CTD data** **ka, 1970-2010.** (doi:10.5063/F1JM27JF)

AOOS Alaska Ocean Observing System

1 tag

Files in this dataset

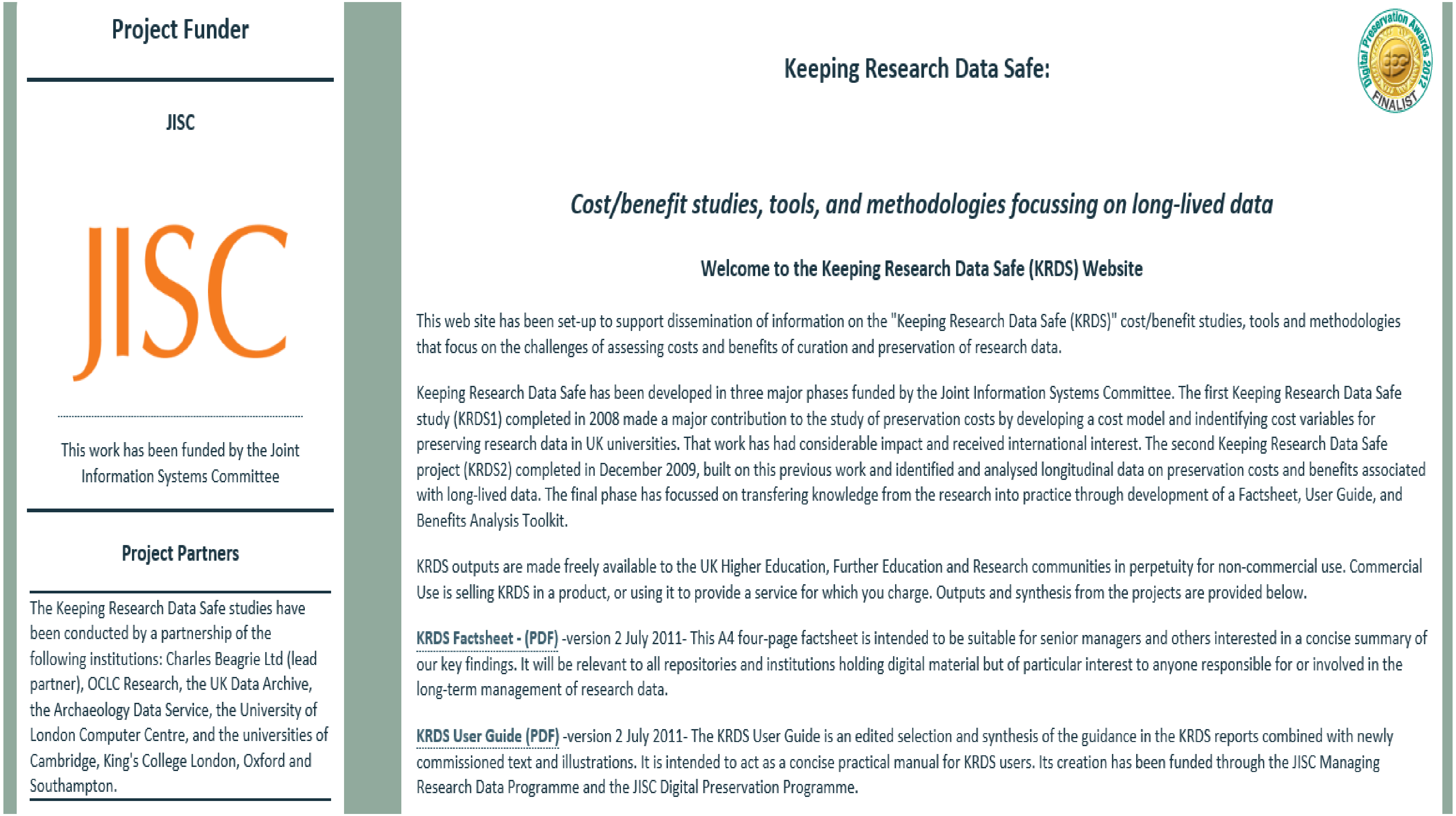
Id	File type	Size	Downloads	Download All	
Metadata: Historical CTD data from the Gulf of Alaska	.xml (EML)	13.00 KB	56 views	Download	Preview
allCtd.csv	text/csv	128 B	2 downloads	Download	Preview
figure.png	text/plain	5.60 KB	6 downloads	Download	Preview
mappingScript.R	text/plain	5.05 KB	3 downloads	Download	Preview

You are currently viewing this metadata document.

General

Title	Historical CTD data from the Gulf of Alaksa, 1970-2010
Identifier	doi:10.5063/F1JM27JF

JISC and Jisc



Project Funder

JISC

JISC

This work has been funded by the Joint Information Systems Committee

Project Partners

The Keeping Research Data Safe studies have been conducted by a partnership of the following institutions: Charles Beagrie Ltd (lead partner), OCLC Research, the UK Data Archive, the Archaeology Data Service, the University of London Computer Centre, and the universities of Cambridge, King's College London, Oxford and Southampton.

Keeping Research Data Safe:

Cost/benefit studies, tools, and methodologies focussing on long-lived data

Welcome to the Keeping Research Data Safe (KRDS) Website

This web site has been set-up to support dissemination of information on the "Keeping Research Data Safe (KRDS)" cost/benefit studies, tools and methodologies that focus on the challenges of assessing costs and benefits of curation and preservation of research data.

Keeping Research Data Safe has been developed in three major phases funded by the Joint Information Systems Committee. The first Keeping Research Data Safe study (KRDS1) completed in 2008 made a major contribution to the study of preservation costs by developing a cost model and indentifying cost variables for preserving research data in UK universities. That work has had considerable impact and received international interest. The second Keeping Research Data Safe project (KRDS2) completed in December 2009, built on this previous work and identified and analysed longitudinal data on preservation costs and benefits associated with long-lived data. The final phase has focussed on transfering knowledge from the research into practice through development of a Factsheet, User Guide, and Benefits Analysis Toolkit.

KRDS outputs are made freely available to the UK Higher Education, Further Education and Research communities in perpetuity for non-commercial use. Commercial Use is selling KRDS in a product, or using it to provide a service for which you charge. Outputs and synthesis from the projects are provided below.

KRDS Factsheet - (PDF) -version 2 July 2011- This A4 four-page factsheet is intended to be suitable for senior managers and others interested in a concise summary of our key findings. It will be relevant to all repositories and institutions holding digital material but of particular interest to anyone responsible for or involved in the long-term management of research data.

KRDS User Guide (PDF) -version 2 July 2011- The KRDS User Guide is an edited selection and synthesis of the guidance in the KRDS reports combined with newly commissioned text and illustrations. It is intended to act as a concise practical manual for KRDS users. Its creation has been funded through the JISC Managing Research Data Programme and the JISC Digital Preservation Programme.

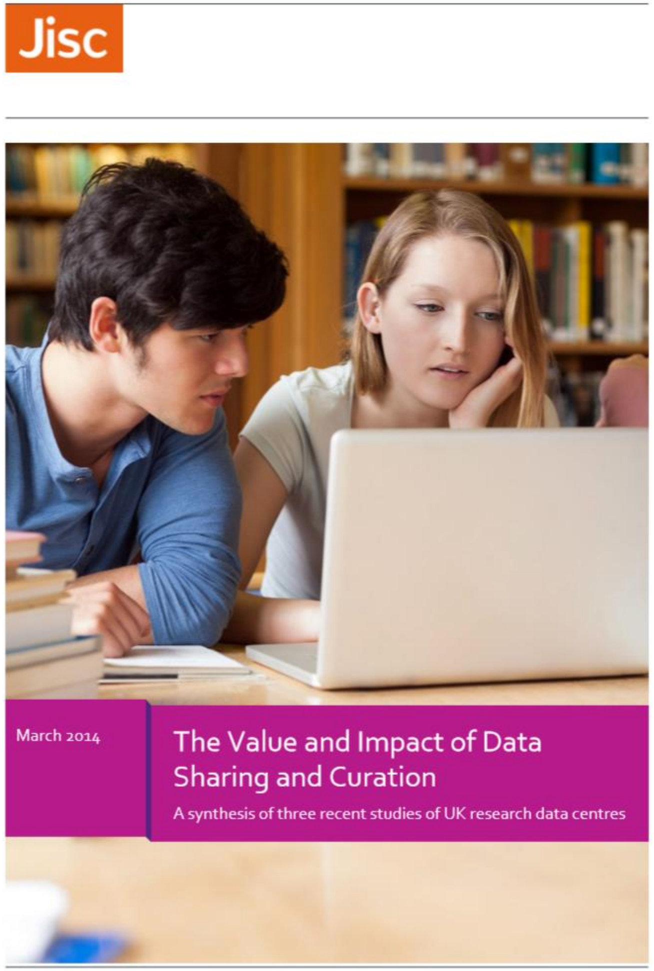
The Value and Impact of Data Sharing and Curation

- Studies covered the Economic and Social Data Service (ESDS), the Archaeology Data Service (ADS), and the British Atmospheric Data Centre (BADC)
- All three studies combined quantitative and qualitative analytical approaches in order to quantify value and impacts in economic terms and explore other, non-economic benefits.
- The economic analysis indicated that:

  - ➢ Very significant increases in research, teaching and studying efficiency were realised by the users as a result of their use of the data centres

  - ➢ The value to users exceeds the investment made in data sharing and curation via the centres in all three cases

  - ➢ By facilitating additional use, the data centres significantly increase the measurable returns on investment in the creation/collection of the data hosted.

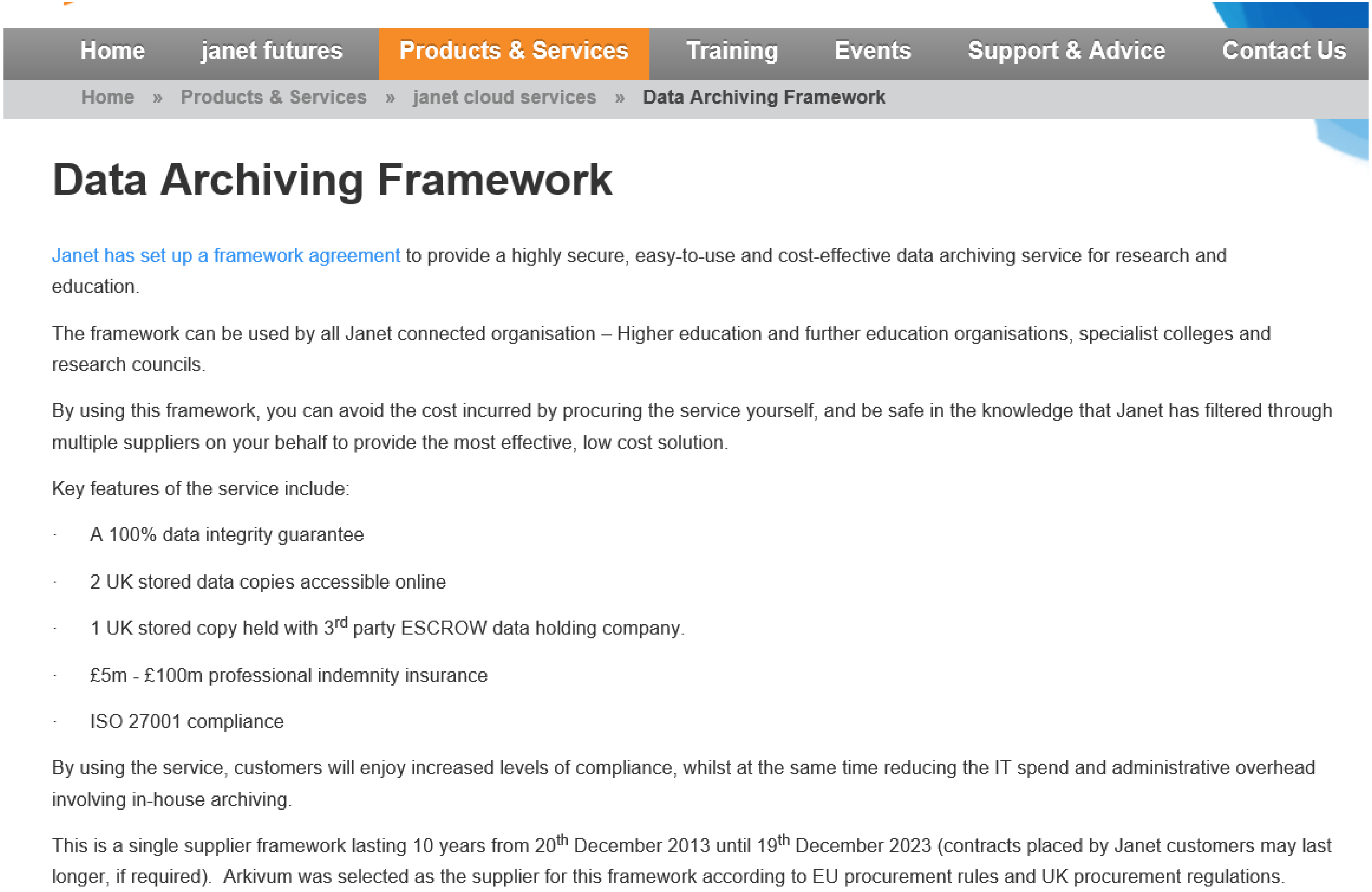


Jisc

March 2014

The Value and Impact of Data Sharing and Curation

A synthesis of three recent studies of UK research data centres


Home
janet futures
Products & Services
Training
Events
Support & Advice
Contact Us
Home » Products & Services » janet cloud services » Data Archiving Framework

Data Archiving Framework

Janet has set up a framework agreement to provide a highly secure, easy-to-use and cost-effective data archiving service for research and education.

The framework can be used by all Janet connected organisation – Higher education and further education organisations, specialist colleges and research councils.

By using this framework, you can avoid the cost incurred by procuring the service yourself, and be safe in the knowledge that Janet has filtered through multiple suppliers on your behalf to provide the most effective, low cost solution.

Key features of the service include:

- A 100% data integrity guarantee
- 2 UK stored data copies accessible online
- 1 UK stored copy held with 3rd party ESCROW data holding company.
- £5m - £100m professional indemnity insurance
- ISO 27001 compliance

By using the service, customers will enjoy increased levels of compliance, whilst at the same time reducing the IT spend and administrative overhead involving in-house archiving.

This is a single supplier framework lasting 10 years from 20th December 2013 until 19th December 2023 (contracts placed by Janet customers may last longer, if required). Arkivum was selected as the supplier for this framework according to EU procurement rules and UK procurement regulations.

Australian eResearch Initiative

Australian National Data Service: ANDS

To make Australia’s research data assets more valuable for its researchers, research institutions and the nation

Brief history of ANDS

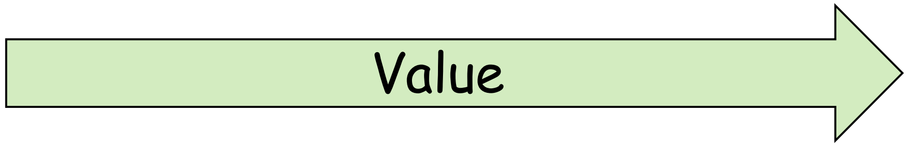
ands AUSTRALIAN NATIONAL DATA SERVICE

- 2006: Review of international eResearch support, in particular UK eScience program
- 2007: Report 'Towards an Australian Data Commons'
- 2009: ANDS commences operations with initial funding of $24M to:
  - influence national policy in the area of data management in the Australian research community
  - inform best practice for the curation of data
  - transform the disparate collections of research data around Australia into a cohesive collection of research resources

ANDS Goal is to transform:

ands
AUSTRALIAN NATIONAL DATA SERVICE

Data that are:	To Structured Collections that are:
Unmanaged	Managed
Disconnected	Connected
Invisible	Findable
Single use	***Reusable***



Value

… so that researchers can easily publish, discover, access and use research data.

ANDS Major Open Data Program

Provides funds, and ANDS support, to create out of existing data resources:

- an internationally significant open data collection
- that helps drive the institutional research strategy
- builds institutional partnerships, and
- builds institutional reputation

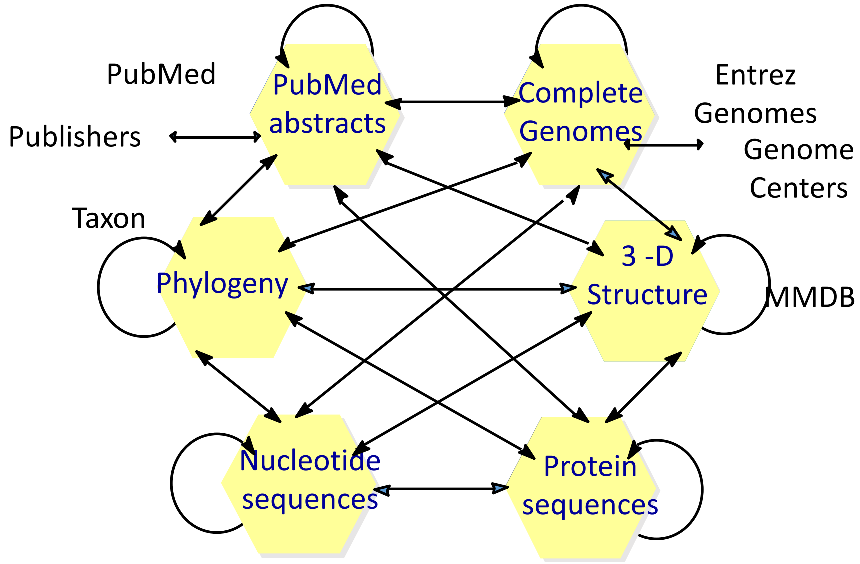
➢ E.g. James Cook University Tropical Data Hub

Open Access, Open Data, Open Science

The US National Library of Medicine

- The NIH Public Access Policy ensures that the public has access to the published results of NIH funded research.
- Requires scientists to submit final peer-reviewed journal manuscripts that arise from NIH funds to the digital archive PubMed Central *upon acceptance for publication*.
- 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

NIH Open Access Compliance?

- PMC Compliance Rate
  - Before legal mandate compliance was 19%
  - Signed into law by George W. Bush in 2007
  - After legal mandate compliance up to 75%
- NIH have taken a further step of announcing that, ‘sometime in 2013’ they stated that they

  ‘… will hold processing of non-competing continuation awards if publications arising from grant awards are not in compliance with the Public Access Policy.’

- NIH now implemented their policy about continuation awards
  - Compliance rate increasing ½% per month
  - By November 2014, compliance rate had reached 86%

Press Release 10-077

Scientists Seeking NSF Funding Will Soon Be Required to Submit Data Management Plans

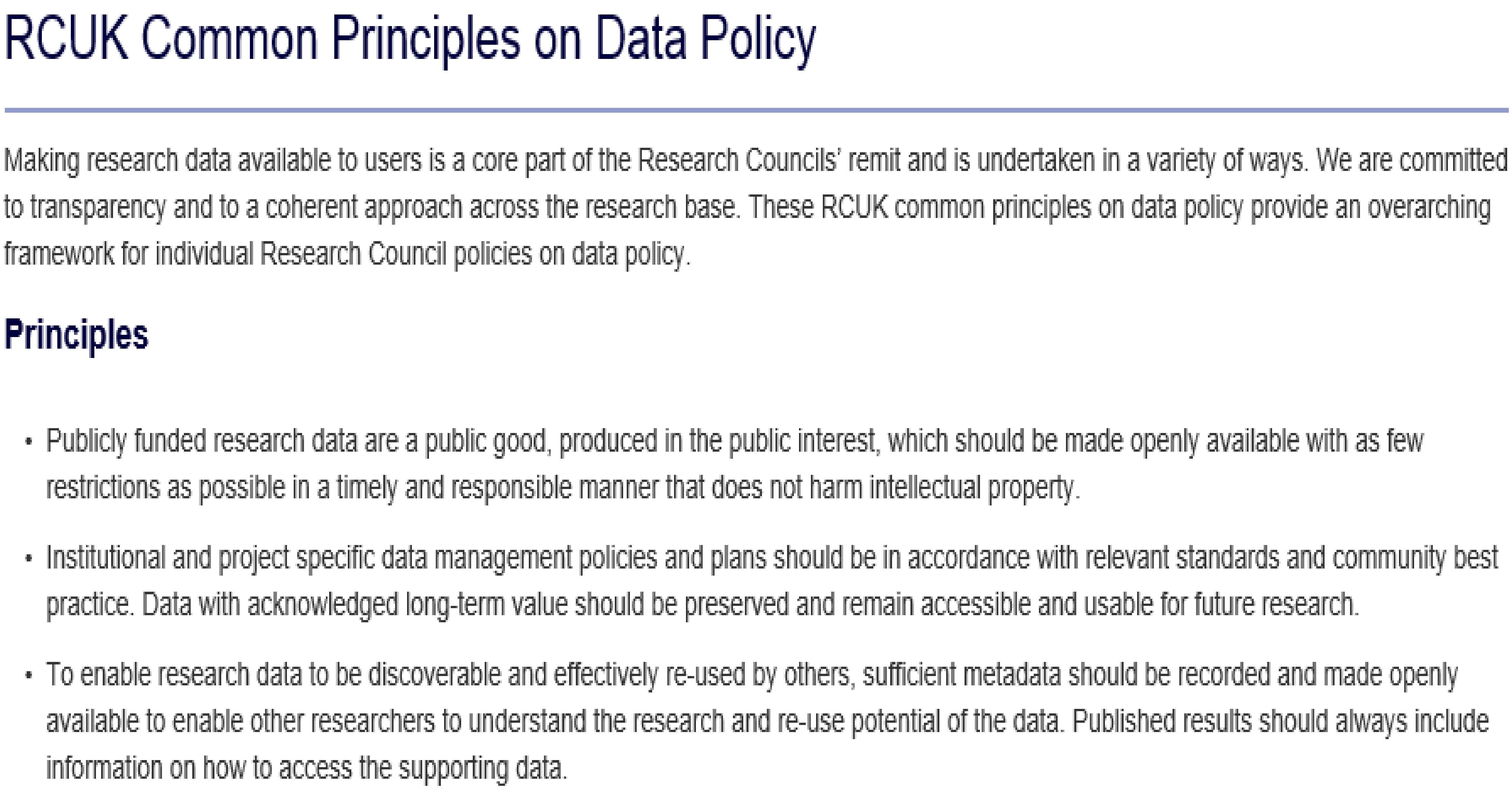
Government-wide emphasis on community access to data supports substantive push toward more open sharing of research data

May 10, 2010

During the May 5th meeting of the National Science Board, National Science Foundation (NSF) officials announced a change in the implementation of the existing policy on sharing research data. In particular, on or around October, 2010, NSF is planning to require that all proposals include a data management plan in the form of a two-page supplementary document. The research community will be informed of the specifics of the anticipated changes and the agency's expectations for the data management plans.

The changes are designed to address trends and needs in the modern era of data-driven science.

"Science is becoming data-intensive and collaborative," noted Ed Seidel, acting assistant director for NSF's Mathematical and Physical Sciences directorate. "Researchers from



RCUK Common Principles on Data Policy

Making research data available to users is a core part of the Research Councils' remit and is undertaken in a variety of ways. We are committed to transparency and to a coherent approach across the research base. These RCUK common principles on data policy provide an overarching framework for individual Research Council policies on data policy.

Principles

- Publicly funded research data are a public good, produced in the public interest, which should be made openly available with as few restrictions as possible in a timely and responsible manner that does not harm intellectual property.
- Institutional and project specific data management policies and plans should be in accordance with relevant standards and community best practice. Data with acknowledged long-term value should be preserved and remain accessible and usable for future research.
- To enable research data to be discoverable and effectively re-used by others, sufficient metadata should be recorded and made openly available to enable other researchers to understand the research and re-use potential of the data. Published results should always include information on how to access the supporting data.

US White House Memorandum

- Directive requiring the major Federal Funding agencies *“to develop a plan to support increased public access to the results of research funded by the Federal Government.”*

- The memorandum defines digital data *“as the digital recorded factual material commonly accepted in the scientific community as necessary to validate research findings including data sets used to support scholarly publications, but does not include laboratory notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects, such as laboratory specimens.”*

22 February 2013

EPSRC Expectations for Data Preservation

- Research organisations will ensure that EPSRC-funded research data is securely preserved for a minimum of 10 years from the date that any researcher ‘privileged access’ period expires

- Research organisations will ensure that effective data curation is provided throughout the full data lifecycle, with ‘data curation’ and ‘data lifecycle’ being as defined by the Digital Curation Centre

The Future

- Research Data Alliance?
- NIH Commons and Hybrid Cloud?
- Research libraries and data scientists?
- 2013 as the Tipping Point for Open Science?

Research Data Alliance

Funded by the European Commission, NSF and ANDS

RDA

RESEARCH DATA ALLIANCE

Research Data Sharing without barriers

Research Data Alliance Created to Accelerate Development of Research Data Sharing Infrastructure Worldwide

- RDA community focuses on building **social, organizational and technical infrastructure** to
  - reduce barriers to data sharing and exchange
  - accelerate the development of coordinated global data infrastructure

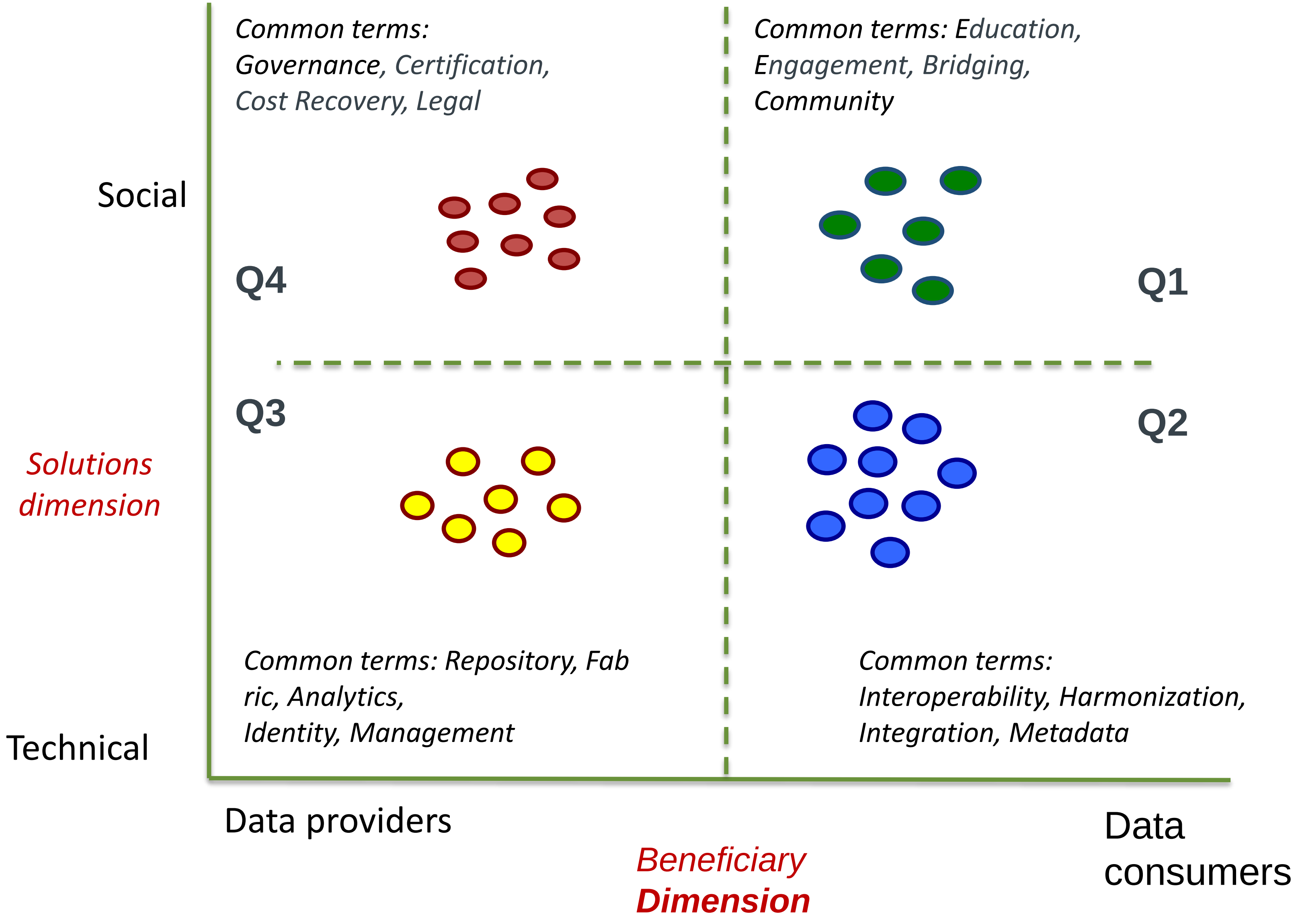
Plenary 2
Washington, DC

CREATE → ADOPT → USE

RDA Working Group Infrastructure Deliverables are:

- **Focused pieces of adopted code, policy, infrastructure, standards, or best practices** that enable data to be shared and exchanged
- **"Harvestable" efforts** for which 12-18 months of work can eliminate a roadblock for a substantial community
- **Efforts that have substantive applicability** to "chunks" of the data community, but may not apply to everyone
- **Efforts for which working scientists and researchers can start today** while more long-term or far-reaching solutions are appropriately discussed in other venues

RDA Working Groups and Interest Groups



Social

Q4

Common terms:
Governance, Certification,
Cost Recovery, Legal

Common terms: Education,
Engagement, Bridging,
Community

Q1

Q3

Q2

Solutions
dimension

Technical

Common terms: Repository, Fab
ric, Analytics,
Identity, Management

Common terms:
Interoperability, Harmonization,
Integration, Metadata

Data providers

Beneficiary
Dimension

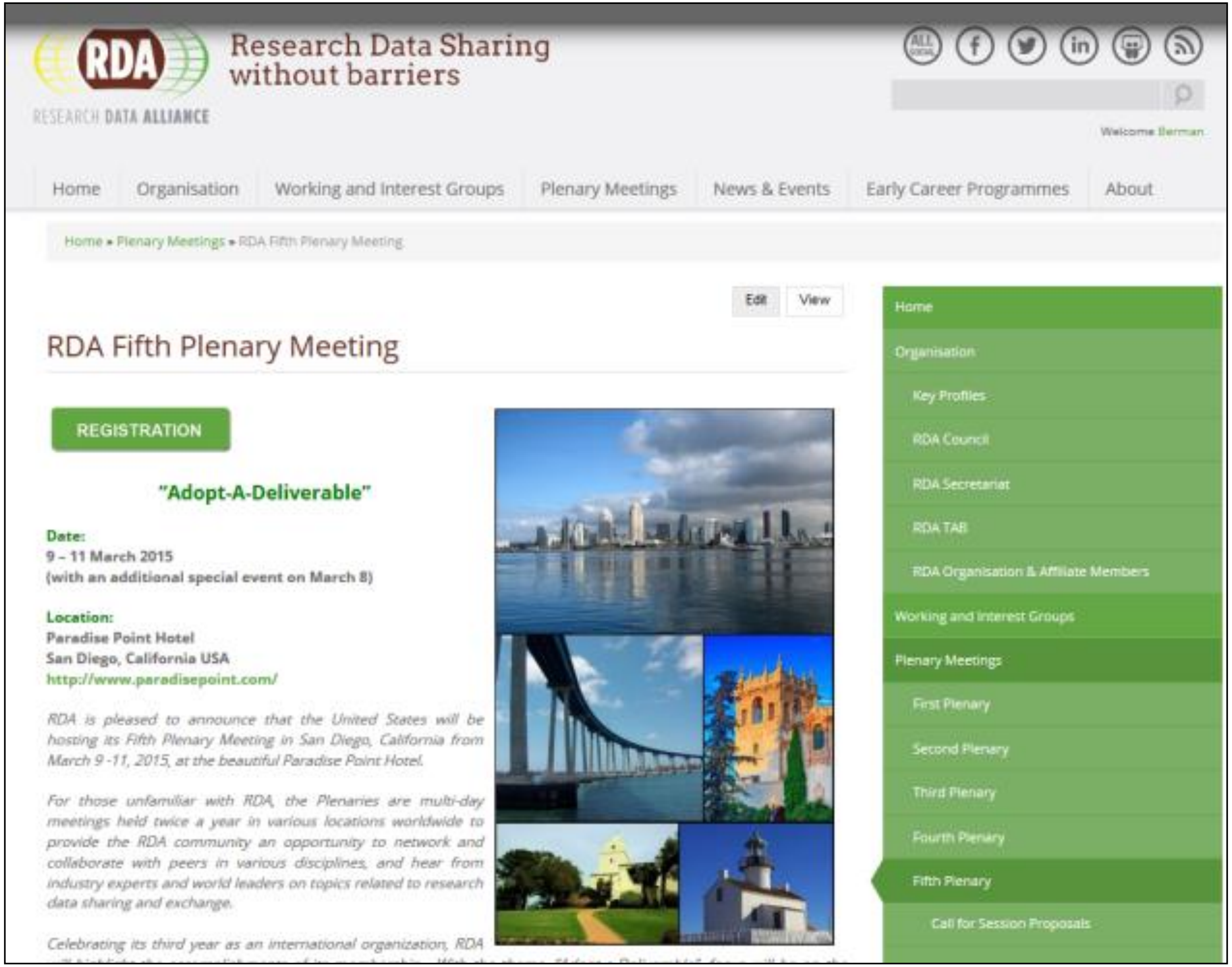
Data
consumers

RDA Plenary 5: San Diego

March 8:
RDA Adoption Day *(open)*

March 9-11:
RDA Plenary 5

- *Registration is now open*



RDA Fifth Plenary Meeting

"Adopt-A-Deliverable"

Date:
9 – 11 March 2015
(with an additional special event on March 8)

Location:
Paradise Point Hotel
San Diego, California USA
http://www.paradisepoint.com/

RDA is pleased to announce that the United States will be hosting its Fifth Plenary Meeting in San Diego, California from March 9 -11, 2015, at the beautiful Paradise Point Hotel.

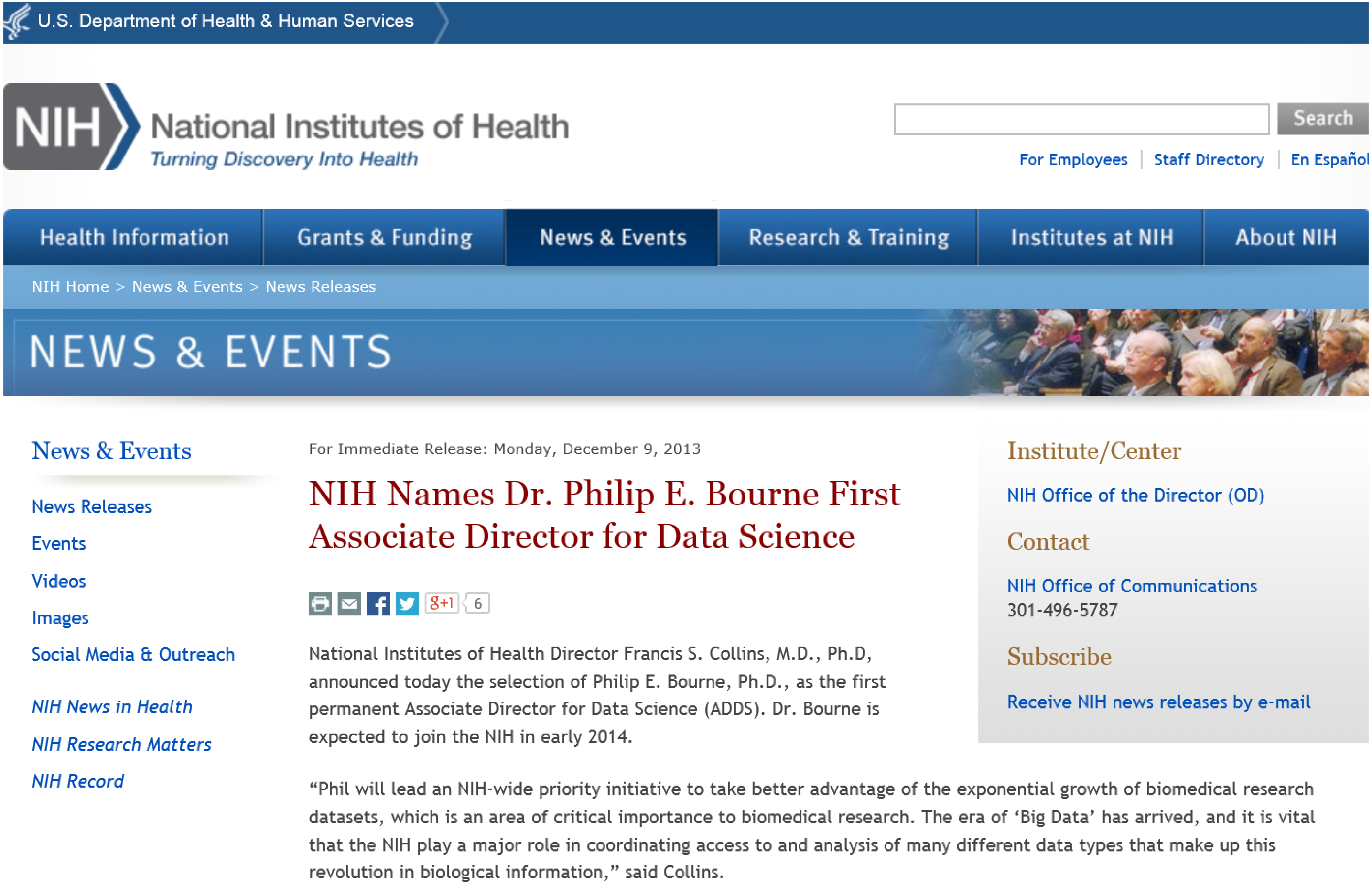
For those unfamiliar with RDA, the Plenaries are multi-day meetings held twice a year in various locations worldwide to provide the RDA community an opportunity to network and collaborate with peers in various disciplines, and hear from industry experts and world leaders on topics related to research data sharing and exchange.

Celebrating its third year as an international organization, RDA

The Data Harvest Report

How sharing research data can yield knowledge, jobs and grow

A RDA Europe Report



U.S. Department of Health & Human Services

NIH National Institutes of Health
Turning Discovery Into Health

Search

For Employees | Staff Directory | En Español

NIH Home > News & Events > News Releases

NEWS & EVENTS

News & Events

- News Releases
- Events
- Videos
- Images
- Social Media & Outreach
- *NIH News in Health*
- *NIH Research Matters*
- *NIH Record*

For Immediate Release: Monday, December 9, 2013

NIH Names Dr. Philip E. Bourne First Associate Director for Data Science

National Institutes of Health Director Francis S. Collins, M.D., Ph.D, announced today the selection of Philip E. Bourne, Ph.D., as the first permanent Associate Director for Data Science (ADDS). Dr. Bourne is expected to join the NIH in early 2014.

“Phil will lead an NIH-wide priority initiative to take better advantage of the exponential growth of biomedical research datasets, which is an area of critical importance to biomedical research. The era of ‘Big Data’ has arrived, and it is vital that the NIH play a major role in coordinating access to and analysis of many different data types that make up this revolution in biological information,” said Collins.

Institute/Center

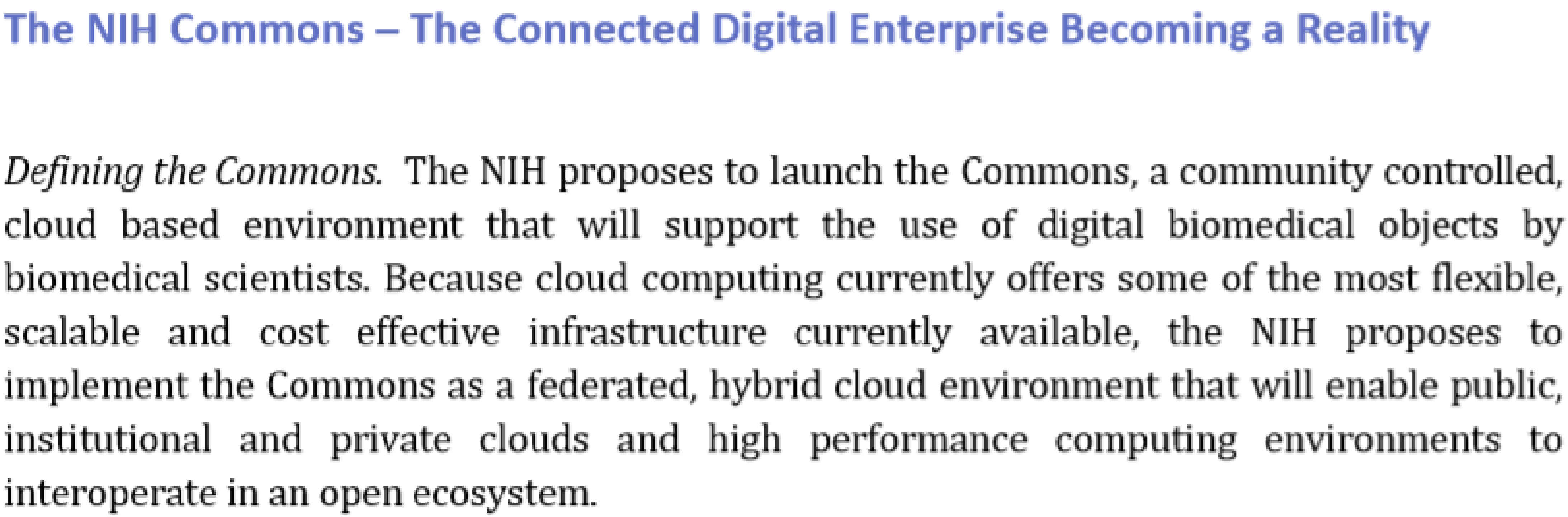
NIH Office of the Director (OD)

Contact

NIH Office of Communications
301-496-5787

Subscribe

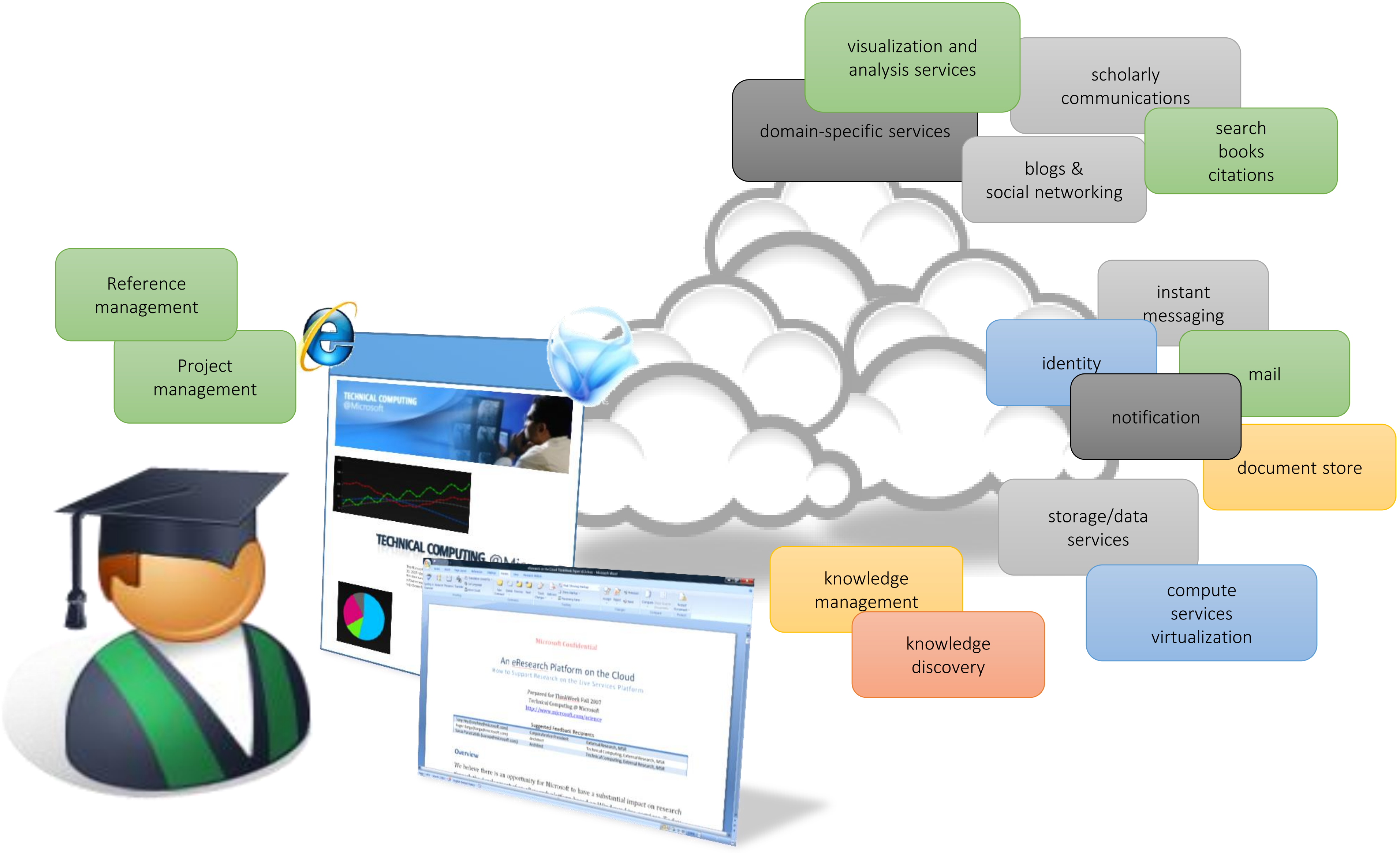
Receive NIH news releases by e-mail



The NIH Commons – The Connected Digital Enterprise Becoming a Reality

Defining the Commons. The NIH proposes to launch the Commons, a community controlled, cloud based environment that will support the use of digital biomedical objects by biomedical scientists. Because cloud computing currently offers some of the most flexible, scalable and cost effective infrastructure currently available, the NIH proposes to implement the Commons as a federated, hybrid cloud environment that will enable public, institutional and private clouds and high performance computing environments to interoperate in an open ecosystem.

What is the Role for Research Libraries?



visualization and analysis services

scholarly communications

domain-specific services

search books citations

blogs & social networking

Reference management

Project management

instant messaging

identity

mail

notification

document store

storage/data services

knowledge management

compute services virtualization

knowledge discovery

Microsoft Confidential

An eResearch Platform on the Cloud

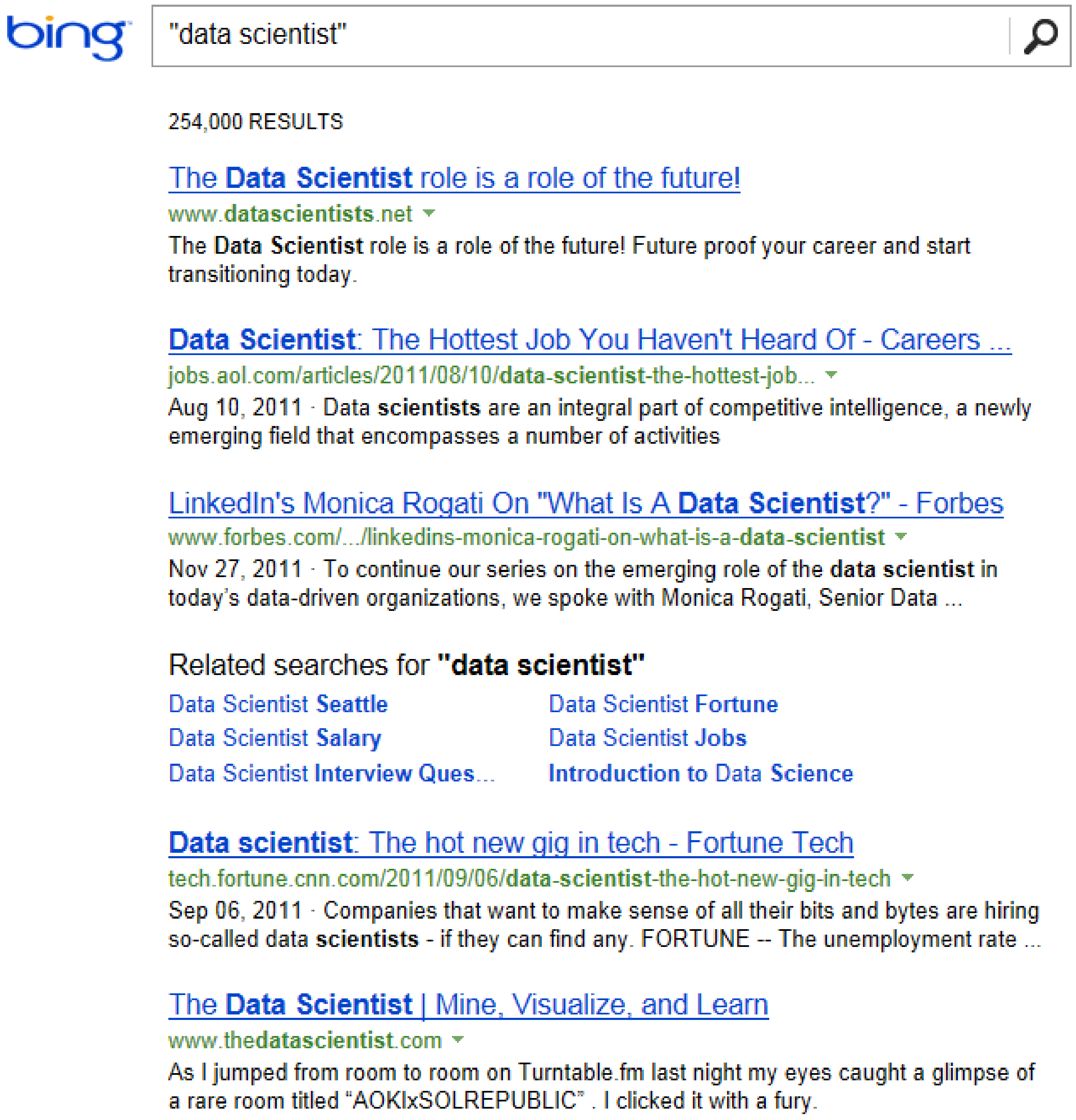
How to Support Research on the Live Services Platform

Prepared for ThinkWeek Fall 2007
Technical Computing @ Microsoft
http://www.microsoft.com/science

Suggested Feedback Recipients		
Tony Hey (tonyhey@microsoft.com)	Corporate Vice President	External Research, MSR
Roger Barga (barga@microsoft.com)	Architect	Technical Computing, External Research, MSR
Savas Parastatidis (savas@microsoft.com)	Architect	Technical Computing, External Research, MSR

Overview

We believe there is an opportunity for Microsoft to have a substantial impact on research



"data scientist"

254,000 RESULTS

The **Data Scientist** role is a role of the future!
www.**datascientists**.net
The **Data Scientist** role is a role of the future! Future proof your career and start transitioning today.

Data Scientist: The Hottest Job You Haven't Heard Of - Careers ...
jobs.aol.com/articles/2011/08/10/**data-scientist**-the-hottest-job...
Aug 10, 2011 · Data **scientists** are an integral part of competitive intelligence, a newly emerging field that encompasses a number of activities

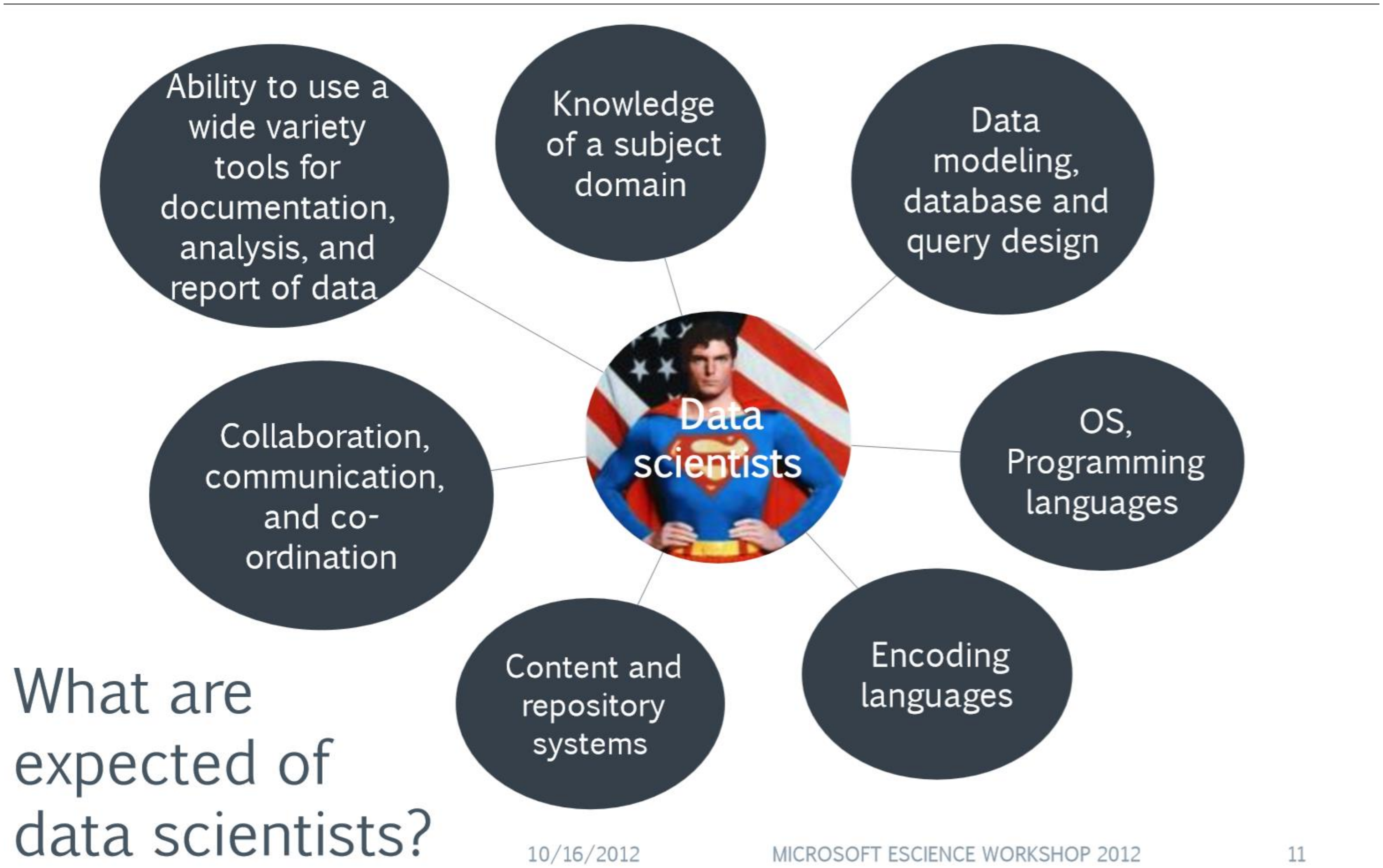
LinkedIn's Monica Rogati On "What Is A **Data Scientist**?" - Forbes
www.forbes.com/.../linkedins-monica-rogati-on-what-is-a-**data-scientist**
Nov 27, 2011 · To continue our series on the emerging role of the **data scientist** in today's data-driven organizations, we spoke with Monica Rogati, Senior Data ...

Related searches for **"data scientist"**

Data Scientist **Seattle**	Data Scientist **Fortune**
Data Scientist **Salary**	Data Scientist **Jobs**
Data Scientist **Interview Ques**...	**Introduction to** Data **Science**

Data scientist: The hot new gig in tech - Fortune Tech
tech.fortune.cnn.com/2011/09/06/**data-scientist**-the-hot-new-gig-in-tech
Sep 06, 2011 · Companies that want to make sense of all their bits and bytes are hiring so-called data **scientists** - if they can find any. FORTUNE -- The unemployment rate ...

The **Data Scientist** | Mine, Visualize, and Learn
www.the**datascientist**.com
As I jumped from room to room on Turntable.fm last night my eyes caught a glimpse of a rare room titled "AOKIxSOLREPUBLIC" . I clicked it with a fury.



What are expected of data scientists?

Data scientists

Ability to use a wide variety tools for documentation, analysis, and report of data

Knowledge of a subject domain

Data modeling, database and query design

OS, Programming languages

Encoding languages

Content and repository systems

Collaboration, communication, and co-ordination

10/16/2012 MICROSOFT ESCIENCE WORKSHOP 2012 11

Slide courtesy of Jian Qin

What is a Data Scientist?

Data Engineer

People who are expert at

- Operating at low levels close to the data, write code that manipulates
- They may have some machine learning background.
- Large companies may have teams of them in-house or they may look to third party specialists to do the work.

Data Analyst

People who explore data through statistical and analytical methods

- They may know programming; May be an spreadsheet wizard.
- Either way, they can build models based on low-level data.
- They eat and drink numbers; They know which questions to ask of the data. Every company will have lots of these.

Data Steward

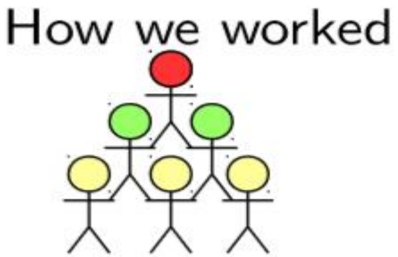
CONCEPTUALISE
TRANSFORM
CREATE OR RECEIVE
DISPOSE
CURATE
ACCESS, USE & REUSE
APPRAISE & SELECT
PRESERVATION PLANNING
DESCRIPTION
REPRESENTATION INFORMATION
COMMUNITY WATCH & PARTICIPATION
STORE
INGEST
PRESERVE
PRESERVATION ACTION
MIGRATE
REAPPRAISE

People who think about managing, curating, and preserving data.

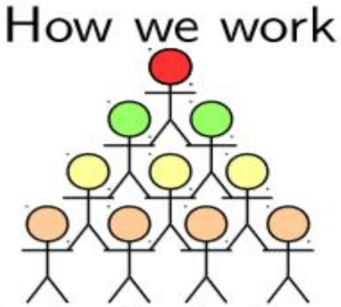
- They are information specialists, archivists, librarians and compliance officers.
- This is an important role: if data has value, you want someone to manage it, make it discoverable, look after it and make sure it remains usable.

What is a data scientist? Microsoft UK Enterprise Insights Blog, Kenji Takeda
http://blogs.msdn.com/b/microsoftenterpriseinsight/archive/2013/01/31/what-is-a-data-scientist.aspx

How do we work?



PI stands on the shoulders of her postdocs and students (and as Newton would have said, the giants.)



PI stands on the shoulders of her postdocs, students, software engineers and data scientists. (Are the giants down with the turtles?).

- It's fair to say that our institutions have not really caught onto the necessity to have careers for everyone in that stack.
- From the people managing vocabularies and manually entering metadata, to the software engineers and data scientists, we have new careers appearing, and we're not really ready for it.
- Mercifully we're not alone, bioinformatics is blazing a similar trail, but we have much to do.

Open Access to Scholarly Publications and Data: 2013 as the Tipping Point?

• US OSTP Memorandum	26 February 2013
• Global Research Council Action Plan	30 May 2013
• G8 Science Ministers Joint Statement	12 June 2013
• European Union Parliament	13 June 2013

Two final comments:

Someone praising Helen Berman, Head of the Protein Data Bank PDB:

‘One of the remarkable things about Helen is that her life has been devoted to service within science rather than, as some might call it, doing real science.’

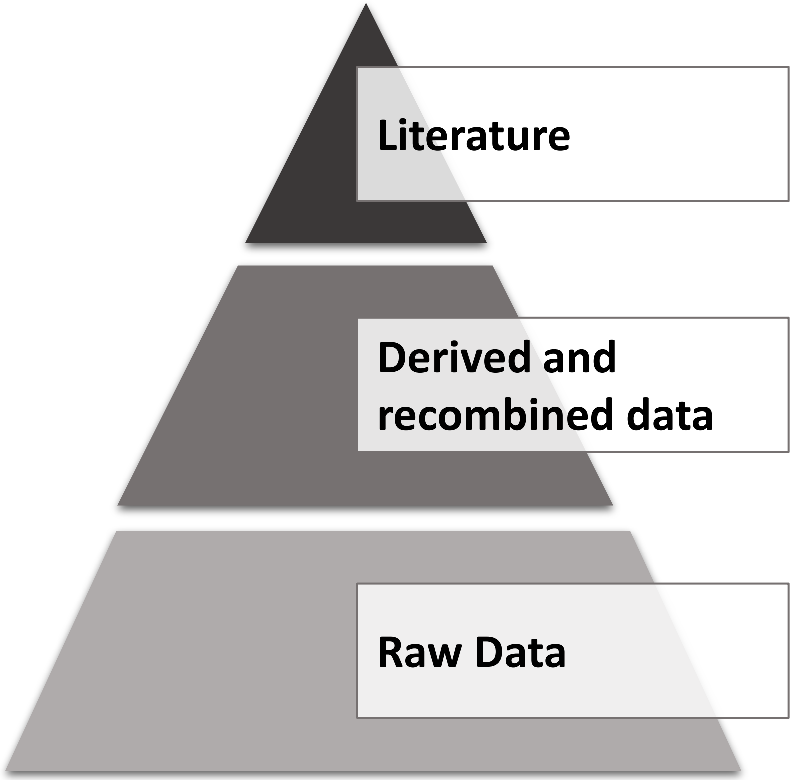
- Culture change is still ongoing!

Michael Lesk on Just-in-time instead of Just-in-case?

‘Most of the cost of archiving is spent at the start, before we know whether the articles will be read or the data used. With data, with no emotional investment in peer review, it might be easier to do a simpler form of deposit, where as much as possible is postponed till the data are called for. There is of course some risk that a just-in-time system will leave us, some years down the road, with a data set which we wish we had curated while the creator was still alive. However, the longer the data has gone unused, the more likely it is to never be used.’

Jim Gray's Vision: 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)

Thank you for listening!