

## Playing YOUR Part RDA Third Plenary Meeting

26 – 28 March 2014 Croke Park Conference Centre, Dublin, Ireland

research data sharing without barriers rd-alliance.org

**Twitter: @resdatall, #RDAPlenary** 





# Beyond Open Access to Open Data

Tony Hey Vice President Microsoft Research

## Ocean Modeling: A Case Study from a Data Scientist

Parker MacCready University of Washington School of Oceanography Coastal Modeling Group

### **Realistic Regional Ocean Hindcasts**







## How do we know the model is working?

# Comparisons were done to an extensive suite of in-situ observations

(see Sutherland et al. 2011)

sea surface height

12 NOAA tide gauges

#### salinity and temperature

over 2000 CTD casts from ECOHAB, RISE, DOE, PRISM, Hood Canal, IOS, King County, and NOAA

#### velocity and moored S,T

7 coastal ADCP / CTD moorings from the ECOHAB and RISE projects, 2 moorings from IOS

#### in addition: satellite and glider data

(Special thanks to R. Thomson, B. Peterson, B. Hickey, R. Kudela, N. Pelland, PRISM, ORCA, DOE, King County, Hood Canal citizen monitoring project, ORHAB, V. Trainer, and A. Odell for use of their data for model validation!!)





## Need to work with many different data formats

## Formats

- NetCDF
- .mat
- CSV
- XML
- Excel

### **Uses of the Data?**

## Software

- MATLAB
- Python
- C#
- Java

Physics: NSF Harmful Algal Blooms: NSF+NOAA Improve FW sources in CESM Climate Model: DOE Shellfish Health: EPA Ocean Acidification: WA State?

## **OData Protocol?**

- OASIS standard protocol for creating and consuming data APIs.
- Builds on core protocols like HTTP and common methodologies like REST.
- Result is a uniform way to expose full-featured data APIs.



# **Data Intensive Science**

# 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

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)

# **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)

# The US National Library of Medicine

- The <u>NIH Public Access Policy</u> 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 <u>PubMed Central</u> upon acceptance for publication.
- Policy requires that these papers are accessible to the public on PubMed Central no later than 12 months after publication.



### **Entrez cross-database search**

# **Open Data and Open Science**

## **Open Access to Research Publications: The Tipping Point**

- US White House Memorandum
- Global Research Council Action Plan
- G8 Science Ministers Joint Statement
- University of California OA Mandate

- 26 February 2013
- 30 May 2013
- 12 June 2013
  - 2 August 2013
- UC produces 40,000 publications per annum corresponding to about 2 – 3 % of all peer-reviewed articles in world each year
- "The faculty remains committed to working with publishers to transform the publishing landscape in ways that are sustainable and beneficial to both the University and the public."

# **Vision for a New Era of Research Reporting**



#### (Thanks to Bill Gates SC05)

# **The Berlin Declaration 2003**

- 'To promote the Internet as a functional instrument for a global scientific knowledge base and for human reflection'
- Defines open access contributions as including: 
   original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material'



# G8 Open Data Charter will 'increase transparency' and 'fuel innovation'

Five key principles outlines how governments should release datasets for economic and social benefits

#### Commission welcomes Parliam Open Data rules

Today the European Parliament formally adopte public sector information, thereby completing the

The EU Committee of Member States' Perman endorsed the new rules on 10 April and they will Council on 20<sup>t</sup> June. The agreement on the new 1 showing how all EU institutions are committed to goldmine.

European Commission Vice-President Neelie Kr efforts to bring government data closer to citi, finally getting the much needed legal framewor jobs.

After this final endorsement, the Commission will the most relevant elements addressed in the D charging arrangements

Member States will have 24 months from the Directive to transpose it into national laws Once the data market in Europe by making all the gen available for re-use. Developers, programmers, able to get and re-use public sector data at zero also have access to more exciting and inspirat museums, libraries and archives now fall under th

#### JAMES VINCENT 🛨 | WEDNESDAY 19 JUNE 2013



http://www.independent.co.uk/life-style/gadgets-and-tech/g8-open-data-charter-will-increase-transparency-and-fuel-innovation-8665696.html

# Collaboration and Sharing of Data is Expected and Growing





... expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work.

NIH reaffirms its support for the concept of data sharing. We believe that data sharing is essential for expedited translation of research results into knowledge, products, and procedures to improve human health ... The NIH expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers.



A primary goal of Data.gov is to improve access to Federal data and expand creative use of those data beyond the walls of government by encouraging innovative ideas (e.g., web applications). Data.gov strives to make government more transparent and is committed to creating an unprecedented level of openness in Government.

# **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.

## Key driver from a UK Research Council

EPSRC Policy Framework on research data (May 2011)

- "all institutions in receipt of their funding should develop a clear roadmap for research data management, which should be implemented by May 1st 2015"
- "organisations will ensure that EPSRC-funded research data is securely preserved for a minimum of 10 years"



# PLOS' New Data Policy: Public Access to Data

By Liz Silva Posted: February 24, 2014

**UPDATE:** A flurry of interest has arisen around the revised PLOS data policy that we <u>announced in December</u> and which will come into effect for research papers submitted next month. We are gratified to see a huge swell of support for the ideas behind the policy, but we note some concerns about how it will be implemented and how it will affect those preparing articles for publication in PLOS journals. We'd therefore like to clarify a few points that have arisen and once again encourage those with concerns to check the <u>details of the policy</u> or our <u>FAQs</u>, and to <u>contact us</u> with concerns if we have not covered them.

#### Is the policy about what to share, or about how and where to share it?

There is nothing new in the policy about what types and forms of data should be shared. As we said <u>in December</u>, "PLOS journals have requested data be available since their inception, but we believe that providing more specific instructions for authors regarding appropriate data deposition options, and providing more information in the published article as to how to access data, is important for readers and users of the research we publish." As we have further <u>clarified</u>, "the Data Policy states the 'minimal dataset' consists "of the dataset used to reach the conclusions drawn in the manuscript with related metadata and methods, and any additional data required to replicate the reported study findings in their entirety. This does not mean that authors must submit all data collected as part of the research, but that they must provide the data that are relevant to the specific analysis presented in the paper." The 'minimal dataset' Linking Publications to Data: The State of the Art

## **Astrophysics Data System ADS**

Find Similar Abstract	ts (with default settings below) Toggle Highlighting
Custom Format	
• Electronic Refereed Jo	ournal Article (HTML)
Full Refereed Journal	Article (PDF/Postscript)
<ul> <li>FIND IT G HARVARD</li> </ul>	
<ul> <li><u>arXiv e-print</u> (arXiv:as</li> </ul>	stro-ph/0412451)
• <u>On-line Data</u>	LINKS to data
• References in the artic	
<u>Citations to the Article</u> Defensed Citations to the	e (84) (Citation History)
SIMBAD Objects (3)	
• NED Objects (1)	Links to objects
Also-Read Articles (Re	eads History)
• <u>Translate This Page</u>	
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. ( <u>ApJ Homepage</u> )
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
<b>Bibliographic Code:</b>	2005ApJ627733M

## **Strasbourg CDS Datasets**

Centro Strasbo	e de Données astronom ourg astronomical Data Cente	iques de Strasbourg <sup>r</sup>	
CDS	SHITE AD		
Object/position	Obj/position/bibcode	Keywords, target,	Object/position
Other services Image: Sesame       Image: Sesame         Dictionary       Sesame         Hosted services		Latest news - Catalogs added between 07-Sep-2013 and 14-S - Catalogs added between 31-Aug-2013 and 07-S - CDS services down on September 09 and 13 - Catalogs added between 24-Aug-2013 and 31-Au - Aladin Lite released - Use X-Match service for queries from coordinate - Collaboration IAS / CDS - PLANCK maps More news	ep-2013 ep-2013 ug-2013 s
ADS mirror A&A		Featured news	1
Keep in touch		WIZTER winn Anniversaire Station Gladin Gladin CMRS	
		On September 18 2012, CDS celebrated its 40 year	ars.



# Reinforcing the Link between Research Publications and Research Data

The Dataverse project at Harvard has been awarded an Alfred Sloan Foundation grant for the next 2 years to enhance the link between journals and data



Seamless integration between the two systems:



Deposit Data to Dataverse through standard API (based on SWORD)

IQSS Dataverse Network >	POWENED BYTHE Dataverse Provi
Chris Anderson Dataverse	Q 🛐 💭 Create Account Log I
Chris Anderson	Search Studies Go Advanced Search Tips
Sort By: \$	Studies: 7
Replication data for: Norway: Monthly party support, 1984-82 by Christ Anderson Abstract: Conventional wisdom has it that the state of the economy drives public support for gove between economic performance and mass opinion appears to vary in strength and direction	hdt:1902.1/10234 annents. Yet the relationship hdt:1902.1/10234 hdt:1902.
Replication data for: Netherlands: Monthly party support, 1978-92 by Christ Anderson Abstract: Conventional wisdom has it that the state of the economy drives public support for gov between economic performance and mass opinion appears to vary in strength and direction	httl:1902.1/10233
Replication data for: Germany: Monthly party support, 1960-89 by Christ Anderson Abstract: Conventional wisdom has it that the state of the economy drives public support for gove between economic performance and mass opinion appears to vary in strength and direction	hdi:1902.1/10222
Replication data for: France: Monthly Prime Minister and Presidential Popularity, 1960-92 by Chris Anderson Abstract: Conventional wisdom has it that the state of the economy drives public support for gove between economic performance and mass ophion agrees to vary in strength and direction	hdi:1902.1/10221 A 8 downloads + analyses Last Released: Jan 21,2008
Replication data for: EU: Public Support for European Integration, 1973-93 by Chris Anderson Abstract: Conventional wisdom has it that the state of the economy drives public support for gov butteen economic and entermore on tense activities under to such a state and the direction	Indi:1902.1/10232 Image: 53 downloads + analyses Last Released: Jan 21, 2009

#### The Dataverse Network

**Open Journal System** 

## Linking to All Data, Big and Small

+ + E S - - - - - - - - - - - - - - - - - -			ESO Telescope Bibliog	graphy		
				telbib Sta	tistics   API   Help    Libraries Hom	e   Archive Home   ESO Home 🔝
Remove Filters	New Searc	H Q EDIT SEARCI	H 14-			VISUALIZE C EXPORT ±
Year: 2013 (x)	Results 1 -	25 of 284				« Previous Next »
Journal	YEAR V	AUTHOR	Tmue	INSTRUMENTS	ACCESS TO DATA	FULLTEXT ADS
A&A (121)	2013	Levenhagen, R. S. et al.	Physical parameters and chemical abundances of $\delta$ Ceti	FEROS1.5		₽2013NewA1855L
ApJ (66)	2013	Gullikson, Kevin et al.	Detection of Low-Mass-ratio Stellar Binary Systems	CRIRES	60.A-9051	₽2013AJ1453G
Nature (5)	2013	Schaerer, D. et al.	Properties of z $\sim$ 3-6 Lyman break galaxies. I. Testing star formation histories and the SFR-mass relation with ALMA and near-IR spectroscopy	ISAAC	168.A-0485, 64.O-0643, 66.A-0572, 68.A-0544	₽2013A&A549A4S
Instrument	2013	Penev, K. et al.	HATS-1b: The First Transiting Planet Discovered by the HATSouth Survey	FEROS2.2, VIS_GROND	087.A-9014, 087.C-0508, 088.A-9008, 089.A-9006, 089.A-9008	₽2013AJ1455P
FORS2 (34) HARPS (32)	2013	Willott, Chris J. et al.	An Exponential Decline at the Bright End of the $z$ = 6 Galaxy Luminosity Function	VIRCAM	179.A-2005, 179.A-2006	₽2013AJ1454W
FLAMES-GIRAFFE	2013	Kraus, M. et al.	Molecular emission from GG Carinae's circumbinary disk	CRIRES, SINFONI	088.D-0442, 384.D-0613	₽2013A&A549A28K
EFOSC2_NTT (19)	2013	Fang, M. et al.	Young stars in $\epsilon$ Chamaleontis and their disks: disk evolution in sparse associations	VISIR	076.C-0470	₽2013A&A549A15F
	2013	Kleyna, J. et al.	P/2010 A2 LINEAR. II. Dynamical dust modelling	EFOSC2_NTT	184.C-1143	₽2013A&A549A13K
	2013	Mathur, S. et al.	Study of HD 169392A observed by CoRoT and HARPS	HARPS	185.D-0056	₽2013A&A549A12M
	2013	Kamiński, T. et al.	Aluminium oxide in the optical spectrum of VY Canis Majoris	UVES	266.D-5655, 67.B-0504	₽2013A&A549A6K
	2013	Persson, M. V. et al.	Warm water deuterium fractionation in IRAS 16293-2422. The high-resolution ALMA and SMA view $% \lambda = 0.01111111111111111111111111111111111$	ALMA_Bands	2011.0.00007.SV	₽2013A&A549L3P
	2013	Fynbo, J. P. U. et al.	Optical/Near-infrared Selection of Red Quasi-stellar Objects: Evidence for Steep Extinction Curves toward Galactic Centers?	EFOSC2_NTT	8600-A.880	■2013ApJS2046F
	2013	MacGregor, Meredith A. et	Millimeter Emission Structure in the First ALMA Image of the AU Mic Debris Disk	ALMA_Bands	2011.0.00142.S	₽2013ApJ762L21M

#### Slide courtesy of Christopher Erdmann

	Smithsonian ER FOR ASTROPHYSICS	EXPL	ORING	G THE UNIVE	RSE
			POWER	Datavers Network	8 9 MORECT V. 3.4
Astronomy Dataver	se Network	Q	POWER	Datavers Network	v.34

This is the Astronomy data repository for Harvard affiliates. Administration and support is provided by the Harvard-Smithsonian Center for Astrophysics (CfA) in collaboration with Harvard Library (HL) and the institute for Quantitative Social Science (IQSS). Infrastructure is provided by Harvard University Information Technology Services.

The Astronomy Dataverse Network plays an important role in fulfilling your Data Management Plan requirements (e.g. as mandated by NSF), and for providing data re-use and obtaion opportunities. Find out more about our team by exploring the Seamless Astronomy and Woltsach Library teams at the CIA. We are also connecting the Astronomy Dataverse to the indexing services provided ...more as

Da	ataverses
13	Dataverses

#### Studies

#### 80 Studies, 598 Files, 54,385 Downloads

 A Dataverse is a container for research data studies, customized and managed by its owner.

RECENTLY RELEASED DATAVERSES
SPT Garay Cluster Spectroscopy
 Apr 9, 2013
 ArrR
 Apr 9, 2013
 ArrR
 Apr 9, 2013
 Laboratory for Vsual Learning
 Mar 9, 2013
 AstroClusters
 Mar 5, 2013
 View More >

() A study is a container for a research data set. It includes cataloging information, data files and complementary files.

RECENTLY RELEASED STUDIES

2011 SPT-GMOS 1D and 2D Spectra by Stubbs, Christopher;	Bayliss,
Matthew; Ruel, Jonathan	Apr 30, 2013
M51 VLA Observations of HI (1984) by Arnold Rots	Apr 10, 2013
All-sky Galactic radiation at 45 MHz and spectral index betwee	en 45 and 408
MHz by Guzman, Andres	Apr 9, 2013
Handling, archiving, and citing data in astronomy by Alberto P	epe; August
Muerch; Merce Crosas; Christopher Erdmann; Alyssa Goodman	Mar 5, 2013
Replication data for: Simulating the X-ray emission from accre	ition shocks
on T Tauri stars by Günther, Hans Moritz	Feb 13, 2013
View More >	
MOST DOWNLOADED STUDIES	
NH3 and CCS at the GBT in Perseus by COMPLETE team	11331
IRAS Based Thermal Emission Maps of Taurus by COMPLET	E team 9260
IRAS Based Thermal Emission Maps of Serpens by COMPLE	TE teamopes
IRAS Based Thermal Emission Maps of Ophiuchus by COMP	LETE team

IRAS Based Thermal Emission Maps of Perseus by COMPLETE team 2172

View More >

# **AstroCurator: Telescope Bibliographies**

Scientific literature	Semi-automated search tool FUSE	Telbib database (back-end)
ads s	None       Manual Sector         Name       Manual Sector	Importe       Importe         Importe       Importe
Telbib database (front-end)	Data archive	Tags + proIDs for telbib records
	ESO Observing Programmes	
Science         Autor of a comparison of the science of the scie	171.3-3085A), Service Mode, UT3-Melipal	

Meakins & Grothkopf, 2011: Linking Publications and Observations: The ESO Telescope Bibliography. ASP Conf. Proc. 461, 767 <u>http://adsabs.harvard.edu/abs/2012ASPC..461..767M</u>

Slide courtesy of Christopher Erdmann

## **Sustainability of Data Links?**



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

# **Data Science Tools**

## **Tools to Support the Entire Data Life Cycle**



FOURTH PARADIGM

DATA-INTENSIVE SCIENTIFIC DISCOVERY

DINCENTORY HEY, STEWART TANKEY, AND SRIDTIN TOLLS

DATA-INTENSIVE SCIENTIFIC DISCOVERY

fourthparadigm.org



### **From Data to Decisions**



### **From Data to Decisions**



## **FetchClimate**

- Intelligent environmental information service
- Automatically:
  - Selects best data source to answer the query
  - Re-grids results
  - Calculates uncertainty
- Windows Azure grants for FetchClimate

White House Climate Data Initiative addresses climate change







## **Open UK Weather Forecast Data**

- Daily forecast + 5 days
- 3 hourly forecast
  + 5 days
- Observation data

Web search: *"open weather data azure"* 





<b>₩</b> DATA <u>U</u> P								
			DataUp is a data doci sharing tool that allow • Upload tabular data to your	umenting, archiving a vs you to: own repository in the cloud	and		_	
*DATAUP						Welcome. Sequim Test   Settings	Sign Out	
Tiles	Add I	New Rule					_ [ _	
Repositories	Rule Name *	Node Atlas						
Quality Checks	Description	Files that are processe of <u>DateTime</u> Latitude a	d in Node Atlas need a minimum and Longitude in any order					
		<b>₩</b> DATAUP					Welcome, Kristin Tol	lle   Sign Out
	Header Names	My Files	View All * Date Uploaded *	③ Post ④ Do	ownload 🛞 Delete			
		Add a New File     Ot     Drag & Drop Files Here #	Ireland Water Sa cov Expires in 65hrs.	excel_good_exam xtsx deptest_sd_imp	Ferns cov Expires in 65hrs.	Tibető cv Expires in 42hrs.	Tibet5 cov Expires in 42hrs.	
Privacy Statement   Terms of Use   Con		File Name excel_good_example2			Repository Name deptest_sd_imp			
		File Extension						
Privacy Statement   Terms of Use	Contact About Dataup	File Size 45.92 KB						
		Uploaded Date 2/26/2014 13:36:3			<sup>Status</sup> Inqueue			
University of California		Tibet4.csv Expires in 42hrs.	Tibet3 cov Expires in 42hrs.	Tibet2.csv Expires in 42hrs.	Tibet1 csv Expires in 42hrs.	Masterlist.cv Expires in 42hrs.	Configuration.cv Expres in 42hrs.	
California Digital Library	20	Chirano Narootics dev	Earthquake Data f vizv					ad (201)
Microsoft Research							Research	••• Windows Azure

## **Power Query (discover & access data)**



#### gdp per capita

#### 2376 results

- Real GDP Per Capita, All States 1... From Bureau of Economic Analysis on Febr... Real GDP by state is an inflation-adjusted measure of each state's gross...
- W Central Intelligence Agency (1993...
   From Wikipedia on January 20, 2014.
   List of countries by GDP (PPP) per capita Wikipedia, the free encyclopedia
- W Methodology List of countries... From Wikipedia on January 20, 2014. List of countries by GDP (PPP) per capita -Wikipedia, the free encyclopedia
- World Bank (2005–2012) Metho...
   From Wikipedia on January 20, 2014.
   List of countries by GDP (PPP) per capita Wikipedia, the free encyclopedia

W CIA World Eactbook (2003–2012)



From SharePoint List

Sp Import data from a Microsoft

SharePoint site.

### **Power Map for Excel**

### Explore geospatial data with Power Map

Power Map for Excel is a three-dimensional (3D) data visualization tool for Excel 2013 which provides a powerful method for people to look at information in new ways. It enables the data discoveries that might not be seen in traditional two-dimensional (2D) tables and charts. With Power Map you can plot geographic and temporal data visually, analyze that data on a 3D globe and over time, and create visual tours to share with others.

See Power Map in action  $\bigcirc$ 



#### http://www.microsoft.com/powerBI/solutions/demo/business-demos.aspx

### **Excel PowerBI**



# Jisc

## Research Data Registry and Discovery Service

### UK pilot project

Stand up working system Explore metadata harvesting Test metadata harvesting Collect feedback

UK Data Archive, NERC Data Catalogue, nine universities Based on ANDS platform and modified



http://www.dcc.ac.uk/projects/research-data-registry-pilot





because good research needs good data

# **Data Repository Registries**



Purdue University 603 repositories http://databib.org



Humboldt-Universität zu Berlin 603 repositories <u>http://re3data.org</u> Identify and locate online repositories of research data:

- What repositories are appropriate for a researcher to submit his or her data to?
- How do users find appropriate data repositories and discover datasets that meet their needs?
- How can librarians help patrons locate and integrate data into their research or learning?

The goal of re3data.org is to create a global registry of research data repositories. The registry will cover research data repositories from different academic disciplines. re3data.org will present repositories for the permanent storage and access of data sets to researchers, funding bodies, publishers and scholarly institutions. In the course of this mission re3data.org aims to promote a culture of sharing, increased access and better visibility of research data.

# The Role of the Cloud?

# Industry is building out massive Cloud Infrastructure



# Bing Speech Recognition Service: The Cloud Changes the Game



## **Data Repository software examples**

د <u>ن:</u> eCrystals					impton
Home About Browse by Year Browse by People					
Login   Create Ad	count			Search	
		C22 H18 N4 O	4 - C <sub>22</sub> H <sub>18</sub> N <sub>4</sub> O <sub>4</sub>		
Sample Originator: John Spencer <sup>b</sup> .			Available Files		
Data Collection	n: Graham J. Tizzard <sup>a</sup> and Mic	hael B. Hursthouse <sup>a</sup>	Final Result		
Structure Dete	rmination: Graham J. Tizzard	•.	2009src1152.cif		18k
University of South	ampton*		2009src1152.cml		6k
Carl-Inversity of Green	witch-		2009src1152.fcf		100k
622H18H404	91404627 20/20 15 18 9 8 7	12 22 48\44 25 49 44	Validation		
4-5-12-19(18)20 (H 24 28 29)	3(22(29)24-21(25)28)17-8-2-1-	3-10-17/h1-13H,14-15H2,	2009src1152_checkcif.htm		8k
(1,21,20,20)	10 5259/sec.stels/1227		Refinement		
Number:	10.5258/edrystals/1227		2009src1152.res		8k
Controlled Keywords:	Oxindole		Solution		
Date Created:	16 December 2009		2009src1152.prp		6k
Deposited On:	16 Feb 2010 11:14		Processing		
Deposited	Dr G J Tizzard		2009src1152.hkl		573k
By:			2009src1152.htm		12k
Data collection	on parameters		2009src1152_0kl.jpg		44k
Chemical form	ula	C22 H18 N4 O4	2009src1152_h0l.jpg		32k
Crystal morph	plogy	Plate	2009src1152_hk0.jpg		47k
			0.0 50		





EPrints (Southampton) <u>http://eprints.org</u> eCrystals example CKAN (OKFN) <u>http://ckan.org</u> On Azure via VM Depot Dataverse (Harvard IQSS) <u>http://thedata.org</u> Coming soon to Azure

## **Hosted data repositories**





fig**share** is a repository where users can make all of their research outputs available in a citable, shareable and discoverable manner.

#### http://figshare.com

Windows Azure Marketplace is an online market for buying and selling finished Software as a Service (SaaS) applications and premium datasets.

#### http://datamarket.azure.com



Windows Azure Marketplace

### Windows Azure for Research Accelerate the Speed of Scientific Discovery

Windows Azure provides researchers with the power and scalability of cloud computing for collaboration, computation, and data-intensive processing. This open and flexible global cloud platform supports any language, tool, or framework.

### Windows Azure

#### The Windows Azure for Research program:

- Free access to Windows Azure cloud computing and storage (submit proposals for Windows Azure Research Awards)
- Windows Azure for Research training classes
- Support and technical resources

S

Apply the power of cloud computing to your computational and data challenges. Experiment at **azure4research.com**.

# **Data Science in the Future?**



'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.thedatascientist.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 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** People who think to 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



Slide courtesy of Jian Qin

### **Some Resources**

- Microsoft Research
  - http://research.microsoft.com
- Microsoft Research Connections
  - <u>http://research.microsoft.com/en-us/collaboration/</u>
- Science at Microsoft
  - http://www.microsoft.com/science
- Scholarly Communications
  - <u>http://www.microsoft.com/scholarlycomm</u>
- Azure Cloud for Research
  - <u>http://research.microsoft.com/en-us/projects/azure/default.aspx</u>
- Outercurve Foundation
  - http://www.outercurve.org/
- Tony Hey on eScience
  - <u>http://tonyhey.net/</u>



© 2013 Microsoft Corporation. All rights reserved. Microsoft, Windows and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries. The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation. MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.

Appendix

# The arXiv Sustainability Model

- Operation of arXiv is currently funded by Cornell University Library.
- In 2010, Cornell broadened funding support for arXiv by asking institutions to make an annual contribution based on the amount downloaded by each institution.
- Annual donations vary in size between \$2,300 to \$4,000, based on usage.
- As of February 2010, 27 institutions have pledged support on this basis.
- The annual budget for arXiv was \$400,000 for 2010.

## **Problems of Research Reproducibility**

- A detailed review of 2,047 retracted articles indexed in PubMed conducted in May of 2012 by Fang, Streen and Casadevall concluded that barely 21.3% were retracted because of errors, while 67.4% were retracted because of scientific misconduct, which included fraud or suspected fraud (43.4%), duplicate publication (14.2%) and plagiarism (9.8%).
- Studies carried out by the pharmaceutical companies Bayer (Germany) and Amgen (USA) concluded that between 60% and 70% of studies in the field of biomedicine may include non-reproducible results.

### Alzheimer's Disease Neuroimaging Initiative

- Alzheimer's Disease Neuroimaging Initiative (ADNI) launched in 2004 specifically to improve clinical trials by different centers agreeing to share data
  - Data fro the 14 different centers involved in the initiative be combined and compared
  - 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/

### National Database for Autism Research

Federated sources of data, tools, & specimens from major US autism research funders and investigators

➢partners adopting NDAR standards, e.g.:

- Global Unique Identifier (GUID)
- Data dictionary (29,000 elements defined)
- Data definition & validation tools
- Authentication scheme



Science"





## Playing YOUR Part RDA Third Plenary Meeting

26 – 28 March 2014 Croke Park Conference Centre, Dublin, Ireland

research data sharing without barriers rd-alliance.org

**Twitter: @resdatall, #RDAPlenary**