Multidisciplinary, Multiscale, Microtomographic Volume Imaging at Southampton

μ-VIS

UNIVERSITY OF Southampton

Storage and sharing of large 3D imaging datasets

Richard Boardman[1], **Ian Sinclair**[1], Simon Cox[1], Philippa Reed[1], Kenji Takeda[1,2], Jeremy Frey[1] and Graeme Earl[3]

[1] μ-VIS X-ray Imaging Centre
www.southampton.ac.uk/muvis
muvis@soton.ac.uk
[2] Microsoft Research Connections, Cambridge, UK
[3] Computational Archaeology, Southampton, UK

July 2012
3D Materials Science (3DMS)

UNIVERSITY OF Southampton

Content

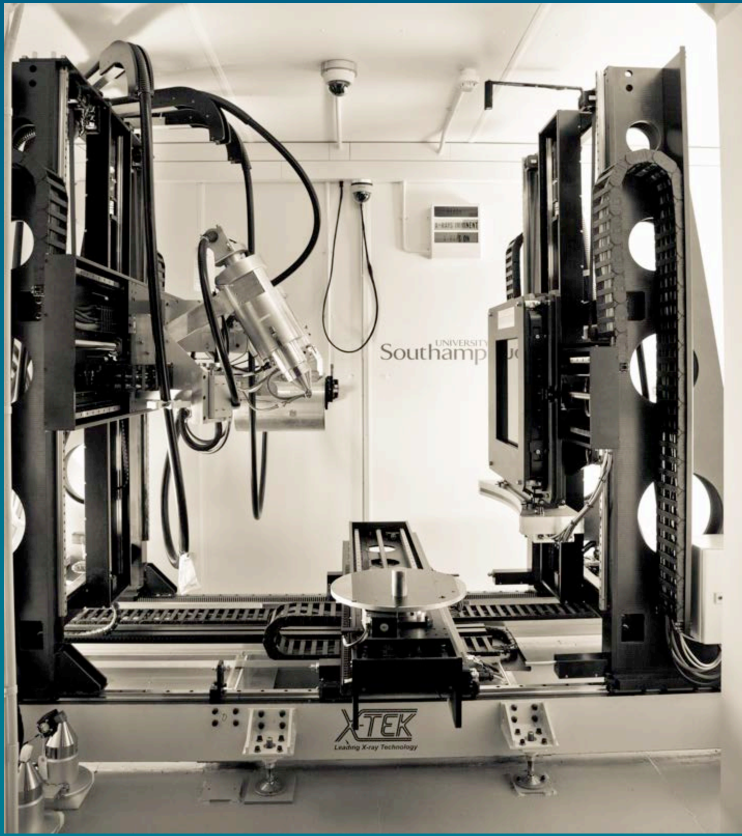
- Overview
  - Background
  - Delineation of scope & roles
  - Metadata quality
- Example engineering activities
- Metadata & database strategies
- Archiving practicalities
- Final thoughts

Motivation

- Significant investments in the generation of large voxel datasets (projects, scanning devices…)
  - High fidelity, large 3D datasets almost inevitably contain more potential than the original researcher/project intended
- To keep value, it is essential to retain the data and record parameters surrounding their acquisition and processing
  - Scientific diligence (e.g. experimental reproducibility…)
  - Sharing: extending the data life cycle
  - Funding body requirements
- Unshared data is a loss to science and engineering

Data: ownership & management

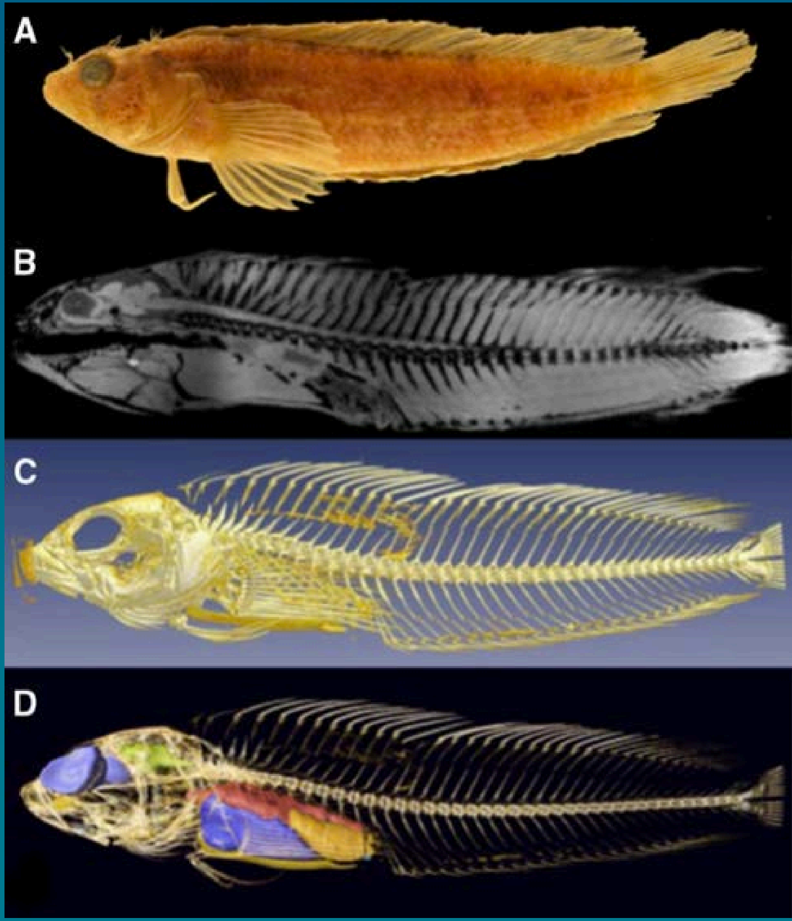
- The roles that are involved
  - Data authors & users
  - Supervisors
  - Facility managers
  - Computer scientists
  - Institutional leaders
  - Funders (government/others)
  - Open access ‘evangelists’
  - Salesmen
  - Legal aspects



µ-VIS X-ray Imaging Centre
www.southampton.ac.uk/muvis

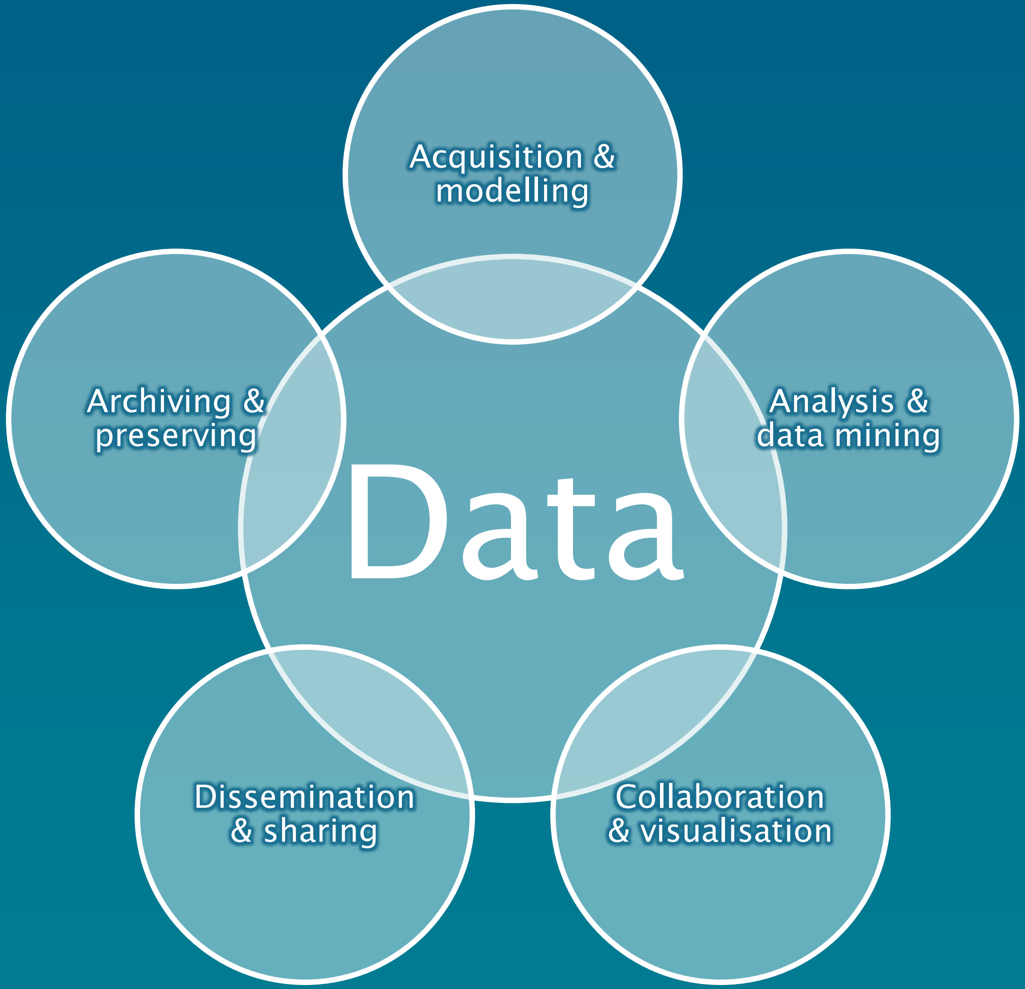
Early Developments

- www.digimorph.org
  - XCT data
  - >1000 bio/palaeo samples
- www.digitalfishlibrary.org
  - MRI data
  - >300 samples (fish!)
- Data reduced to 2D and animations
  - <5Mb
- Raw voxels not available as yet
- Recent example
  - 3D Materials Atlas



Island Kelpfish: MRI & CT data: DigiMorph & Digial Fish Library

Data-intensive science

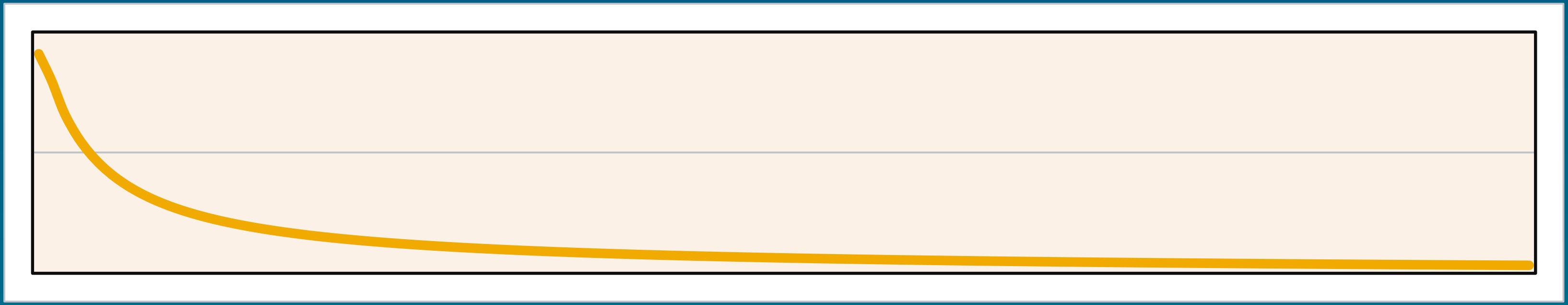


Acquisition: The data deluge

- We can generate data faster than we can consume it
  - Rate of generation now exceeds physical storage capacity (Feb. 2011)*
- Synchrotrons: terabytes per day
  - SLS: ~5TB/day (fast acquisition)
  - AS: 200TB/year (growing to 400TB/year with new beamlines)
  - ESRF (*ca.* 2010): O(100TB) 30-day storage, O(PB) for backups
- µ-VIS lab facility: up to two terabytes per day (robotic operation)
  - 20GB projections + 30GB reconstruction = 50GB in as little as 10-15 minutes
  - Plus O(10MB) metadata
- *LHC ~ 50-100PB/yr, ~20PB stored*

* Hilbert & Lopez, (2011), Science

Long-tail science



- Small numbers of major projects/facilities responsible for a lot of output
  - Formal data management policies & resources
- Large number of smaller projects/facilities also do a lot!
  - Data management policies & resources very variable

UNIVERSITY OF Southampton

What do we mean by “large”

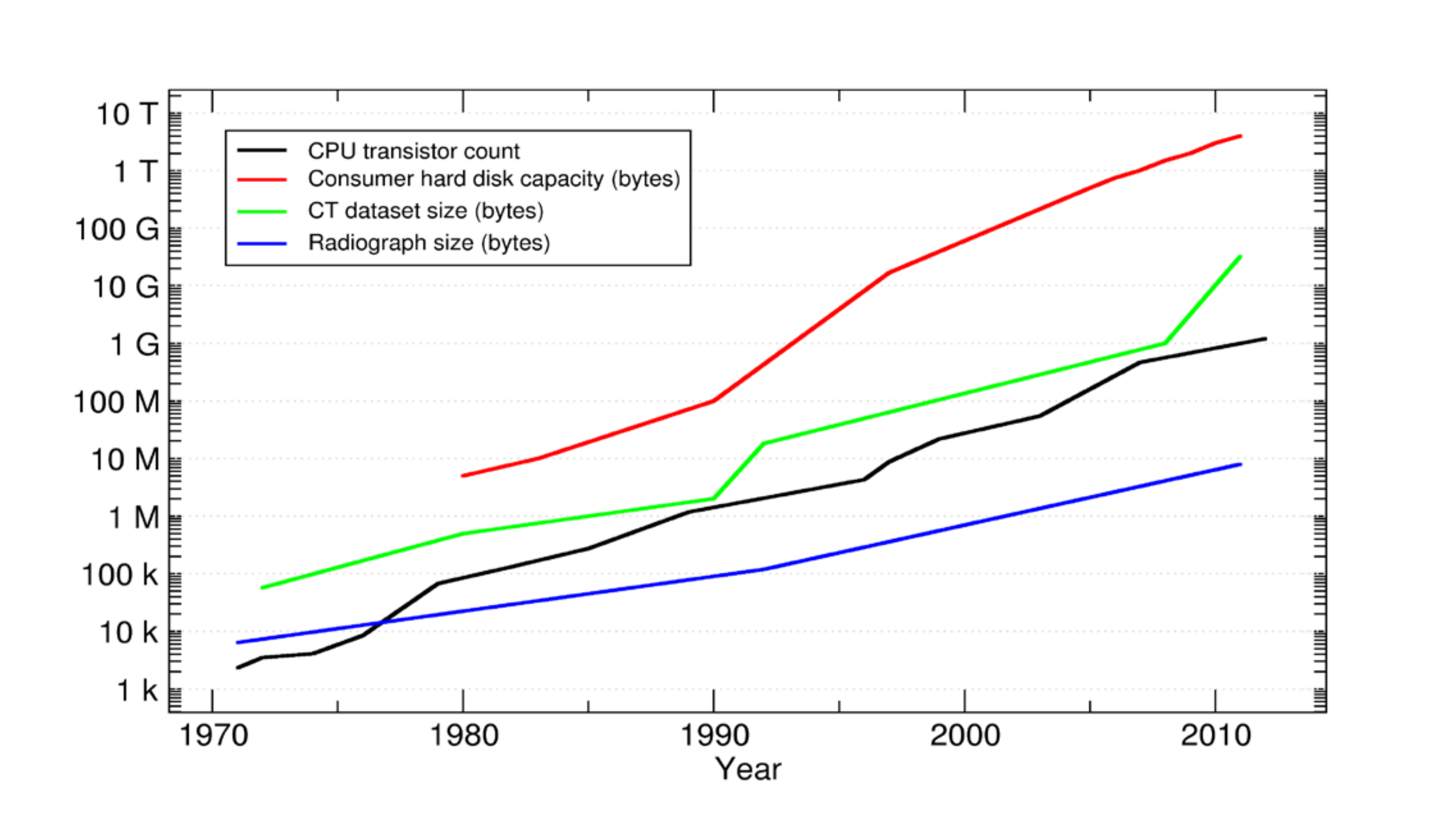
- For our purposes, right now we will say that it is datasets which are O(1GB – 100GB+)
- They won’t necessarily fit on CDs, DVDs, Blu-rays
- They will fit on hard drives
- Transferring them around even in isolation may present a challenge (portable drives, institutional networks, FTP/rsync/GridFTP?)
  - ➔ Grid/Cloud capabilities & tools, e.g. Globus, MS Azure, Amazon S3, **Dropbox**
  - ➔ ‘Never underestimate the bandwidth of a stationwagon full of tapes hurtling down the highway’?

What about Moore’s Law?

- Integrated circuit transistor count doubles every two years
- Broadly applicable to many areas of technology, including hard disk capacity
- Perhaps we can just wait a while and Moore’s Law will help us store data?
- Unfortunately, our large 3D datasets also “obey” Moore’s Law
  - This large 3D dataset problem will always be with us

UNIVERSITY OF Southampton

What about Moore's Law?



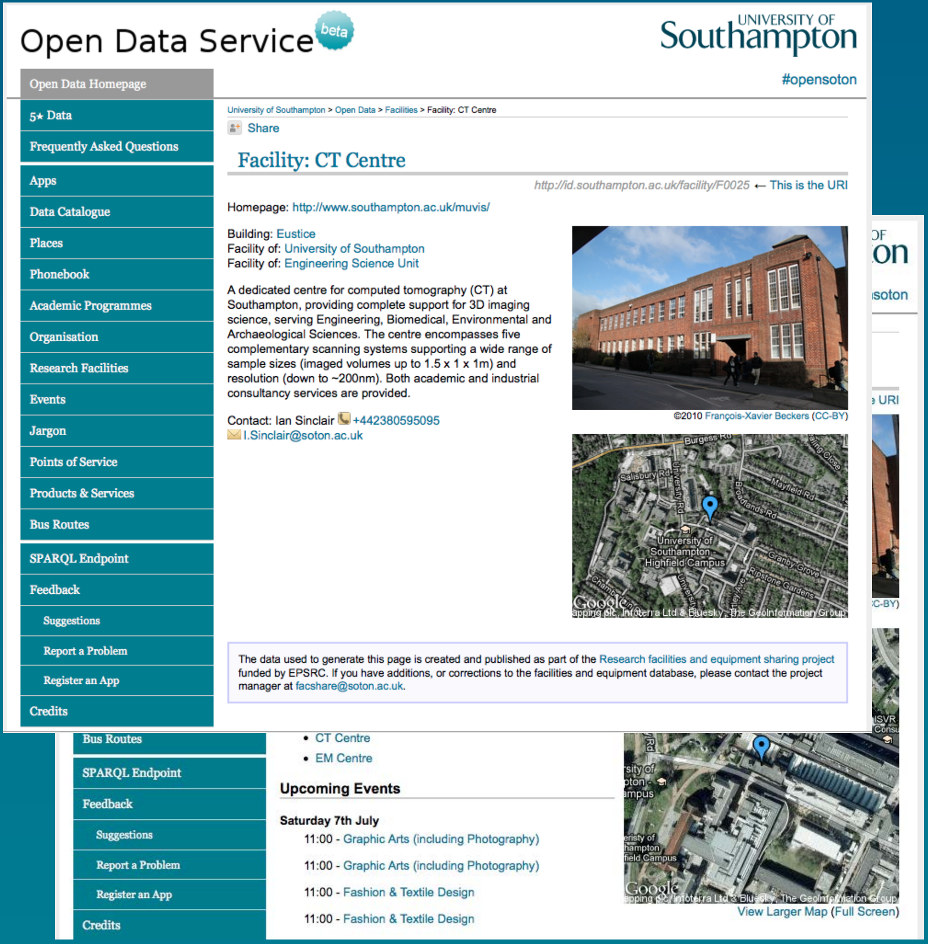
Sources: Kalendar, W (2011); μ-VIS X-ray Imaging Centre; IBM; AMD; Intel; DEC; Seagate; Western Digital

- Solution? Greater allocation of resources to storage/archiving – who pays?

UNIVERSITY OF Southampton

Southampton data sharing projects

- data.gov.uk – “Opening up Government”
  - – Founded by Nigel Shadbolt and Tim Berners-Lee
- Open Data Service
  - – data.southampton.ac.uk
- Research data access
  - – datapool.soton.ac.uk
  - – 10 year roadmap
  - ➢ *Recognising need for new services, policy framework and data management support*



Facility: CT Centre

http://id.southampton.ac.uk/facility/F0025 ← This is the URI

Homepage: http://www.southampton.ac.uk/muvis/

Building: Eustice
Facility of: University of Southampton
Facility of: Engineering Science Unit

A dedicated centre for computed tomography (CT) at Southampton, providing complete support for 3D imaging science, serving Engineering, Biomedical, Environmental and Archaeological Sciences. The centre encompasses five complementary scanning systems supporting a wide range of sample sizes (imaged volumes up to 1.5 x 1 x 1m) and resolution (down to ~200nm). Both academic and industrial consultancy services are provided.

Contact: Ian Sinclair ☎ +442380595095
✉ I.Sinclair@soton.ac.uk

©2010 François-Xavier Beckers (CC-BY)

The data used to generate this page is created and published as part of the Research facilities and equipment sharing project funded by EPSRC. If you have additions, or corrections to the facilities and equipment database, please contact the project manager at facshare@soton.ac.uk.

- CT Centre
- EM Centre

Upcoming Events

Saturday 7th July

11:00 - Graphic Arts (including Photography)

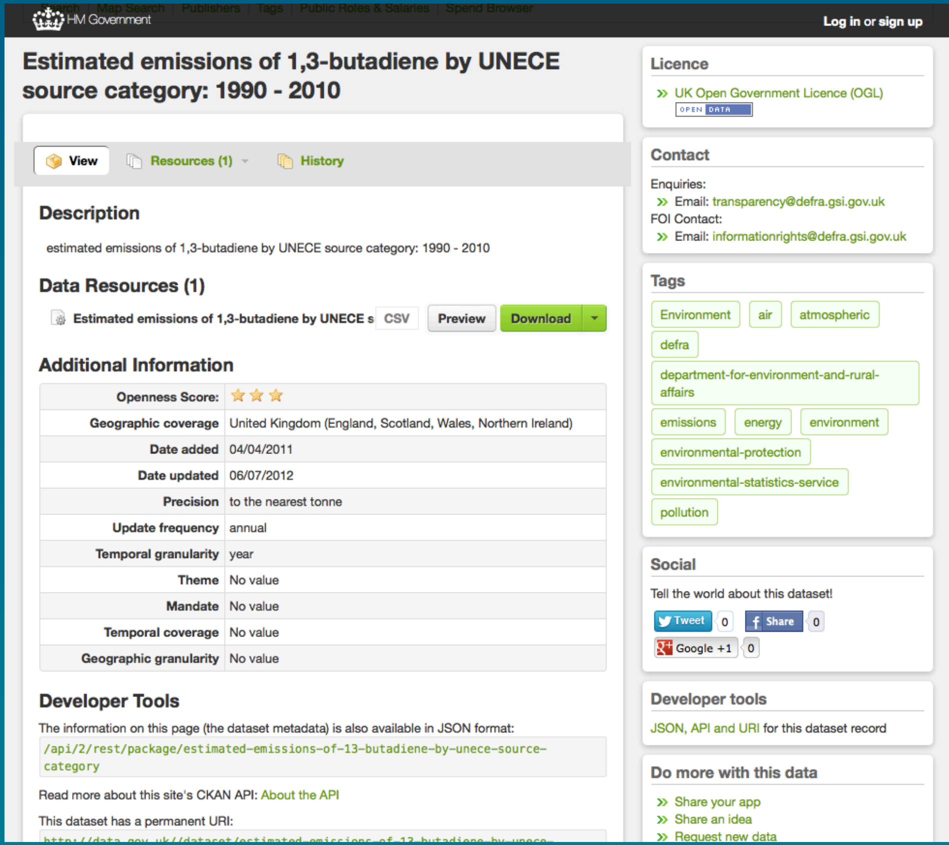
11:00 - Graphic Arts (including Photography)

11:00 - Fashion & Textile Design

11:00 - Fashion & Textile Design

Southampton data sharing projects

- Data grades (stars)
  1. Anything/‘stuff’
  2. Structured data, e.g. Excel file instead of jpeg of a data table
  3. Open format, e.g. CSV vs. Excel files
  4. Provide persistent link
  5. Links to others data/information, to provide context



Estimated emissions of 1,3-butadiene by UNECE source category: 1990 - 2010

Description

estimated emissions of 1,3-butadiene by UNECE source category: 1990 - 2010

Data Resources (1)

Estimated emissions of 1,3-butadiene by UNECE s

Additional Information

Openness Score:	★★★
Geographic coverage	United Kingdom (England, Scotland, Wales, Northern Ireland)
Date added	04/04/2011
Date updated	06/07/2012
Precision	to the nearest tonne
Update frequency	annual
Temporal granularity	year
Theme	No value
Mandate	No value
Temporal coverage	No value
Geographic granularity	No value

Developer Tools

The information on this page (the dataset metadata) is also available in JSON format:

```
/api/2/rest/package/estimated-emissions-of-13-butadiene-by-unece-source-category
```

Read more about this site's CKAN API: About the API

This dataset has a permanent URI:

```
http://data.gov.uk//dataset/estimated-emissions-of-13-butadiene-by-unece-
```

Licence

» UK Open Government Licence (OGL)

OPEN DATA

Contact

Enquiries:
» Email: transparency@defra.gsi.gov.uk
FOI Contact:
» Email: informationrights@defra.gsi.gov.uk

Tags

Environment · air · atmospheric · defra · department-for-environment-and-rural-affairs · emissions · energy · environment · environmental-protection · environmental-statistics-service · pollution

Social

Tell the world about this dataset!

Developer tools

JSON, API and URI for this dataset record

Do more with this data

» Share your app
» Share an idea
» Request new data

- 1* is great, but must aim for 5*

Typical laboratory conversations?

If only I knew exactly how she did this experiments

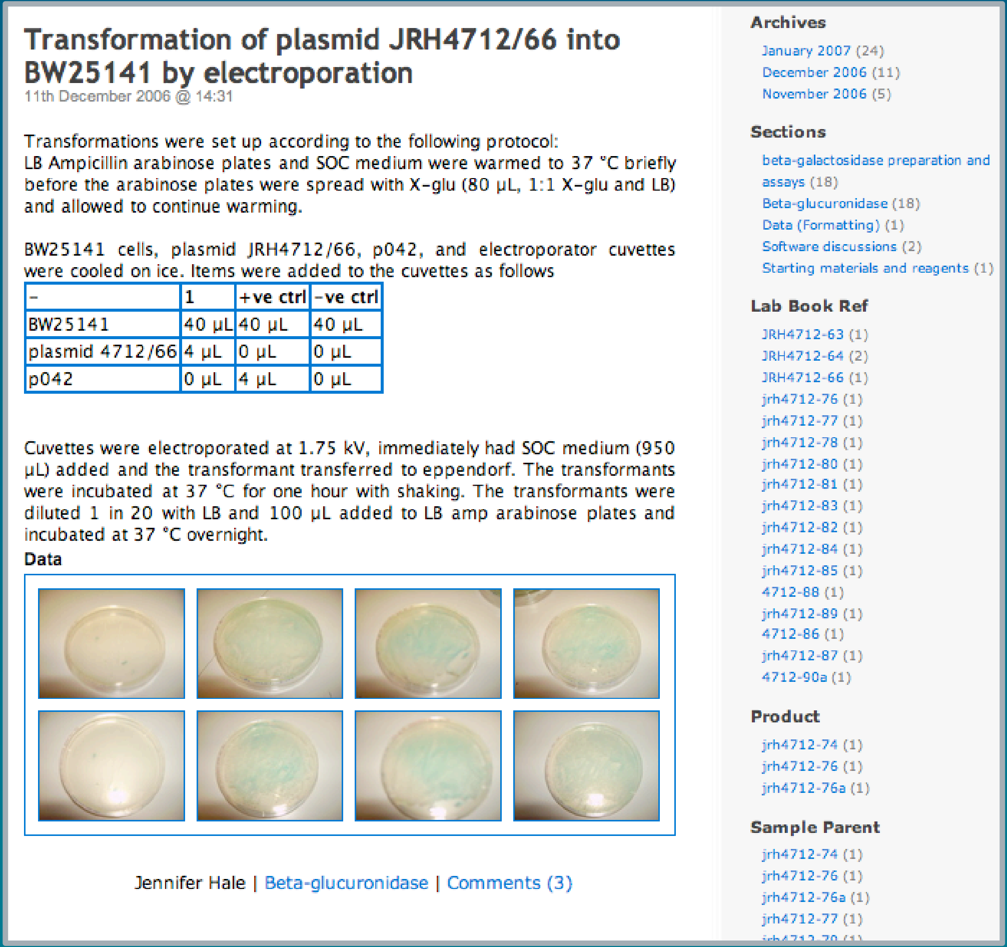
I wish I had recorded things at the start the way I do now…..

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.

Implementation of e-lab book

- Blog based format
- Purpose built engine
- Fully flexible system with arbitrary metadata
- Full record of changes



Transformation of plasmid JRH4712/66 into BW25141 by electroporation

11th December 2006 @ 14:31

Transformations were set up according to the following protocol:
LB Ampicillin arabinose plates and SOC medium were warmed to 37 °C briefly before the arabinose plates were spread with X-glu (80 µL, 1:1 X-glu and LB) and allowed to continue warming.

BW25141 cells, plasmid JRH4712/66, p042, and electroporator cuvettes were cooled on ice. Items were added to the cuvettes as follows

-	1	+ve ctrl	-ve ctrl
BW25141	40 µL	40 µL	40 µL
plasmid 4712/66	4 µL	0 µL	0 µL
p042	0 µL	4 µL	0 µL

Cuvettes were electroporated at 1.75 kV, immediately had SOC medium (950 µL) added and the transformant transferred to eppendorf. The transformants were incubated at 37 °C for one hour with shaking. The transformants were diluted 1 in 20 with LB and 100 µL added to LB amp arabinose plates and incubated at 37 °C overnight.

Data

Jennifer Hale | Beta-glucuronidase | Comments (3)

Archives

January 2007 (24)
December 2006 (11)
November 2006 (5)

Sections

beta-galactosidase preparation and assays (18)
Beta-glucuronidase (18)
Data (Formatting) (1)
Software discussions (2)
Starting materials and reagents (1)

Lab Book Ref

JRH4712-63 (1)
JRH4712-64 (2)
JRH4712-66 (1)
jrh4712-76 (1)
jrh4712-77 (1)
jrh4712-78 (1)
jrh4712-80 (1)
jrh4712-81 (1)
jrh4712-83 (1)
jrh4712-82 (1)
jrh4712-84 (1)
jrh4712-85 (1)
4712-88 (1)
jrh4712-89 (1)
4712-86 (1)
jrh4712-87 (1)
4712-90a (1)

Product

jrh4712-74 (1)
jrh4712-76 (1)
jrh4712-76a (1)

Sample Parent

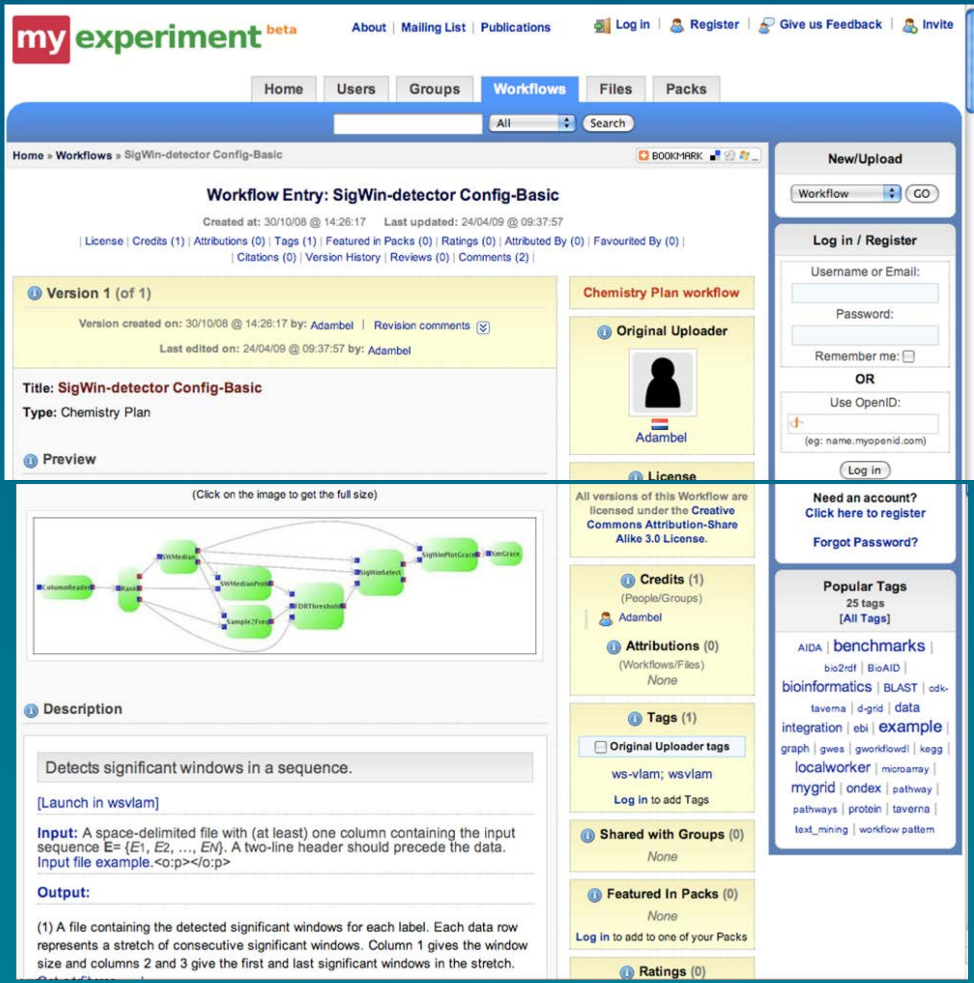
jrh4712-74 (1)
jrh4712-76 (1)
jrh4712-76a (1)
jrh4712-77 (1)

http://chemtools.chem.soton.ac.uk/projects/blog/ “Bio Blogs”

http://blogs.openwetware.org/scienceintheopen Discussion

Implementation of e-lab book

- "Facebook for Scientists" ...but different to Facebook!
- A repository of research methods
- A community social network of people and things
- Machinery for coordinating the execution of (scientific) services and linking together (scientific) resources
- Open source (BSD) Ruby on Rails application with HTML, REST and SPARQL interfaces



Home » Workflows » SigWin-detector Config-Basic

Workflow Entry: SigWin-detector Config-Basic

Created at: 30/10/08 @ 14:26:17 Last updated: 24/04/09 @ 09:37:57

| License | Credits (1) | Attributions (0) | Tags (1) | Featured in Packs (0) | Ratings (0) | Attributed By (0) | Favourited By (0) | Citations (0) | Version History | Reviews (0) | Comments (2) |

Version 1 (of 1)

Version created on: 30/10/08 @ 14:26:17 by: Adambel | Revision comments

Last edited on: 24/04/09 @ 09:37:57 by: Adambel

Title: SigWin-detector Config-Basic

Type: Chemistry Plan

Preview

(Click on the image to get the full size)

Description

Detects significant windows in a sequence.

[Launch in wsvlam]

Input: A space-delimited file with (at least) one column containing the input sequence E= {E1, E2, ..., EN}. A two-line header should precede the data. Input file example.<o:p></o:p>

Output:

(1) A file containing the detected significant windows for each label. Each data row represents a stretch of consecutive significant windows. Column 1 gives the window size and columns 2 and 3 give the first and last significant windows in the stretch.

Chemistry Plan workflow

Original Uploader

Adambel

License

All versions of this Workflow are licensed under the Creative Commons Attribution-Share Alike 3.0 License.

Credits (1)
(People/Groups)
Adambel

Attributions (0)
(Workflows/Files)
None

Tags (1)

□ Original Uploader tags

ws-vlam; wsvlam

Log in to add Tags

Shared with Groups (0)
None

Featured In Packs (0)
None
Log in to add to one of your Packs

Ratings (0)

New/Upload

Log in / Register

Username or Email:

Password:

Remember me: □

OR

Use OpenID:

(eg: name.myopenid.com)

Need an account?
Click here to register

Forgot Password?

Popular Tags
25 tags
[All Tags]

AIDA | benchmarks | bio2rdf | BioAID | bioinformatics | BLAST | cdk-taverna | d-grid | data integration | ebi | example | graph | gwes | gworkflowdl | kegg | localworker | microarray | mygrid | ondex | pathway | pathways | protein | taverna | text_mining | workflow pattern

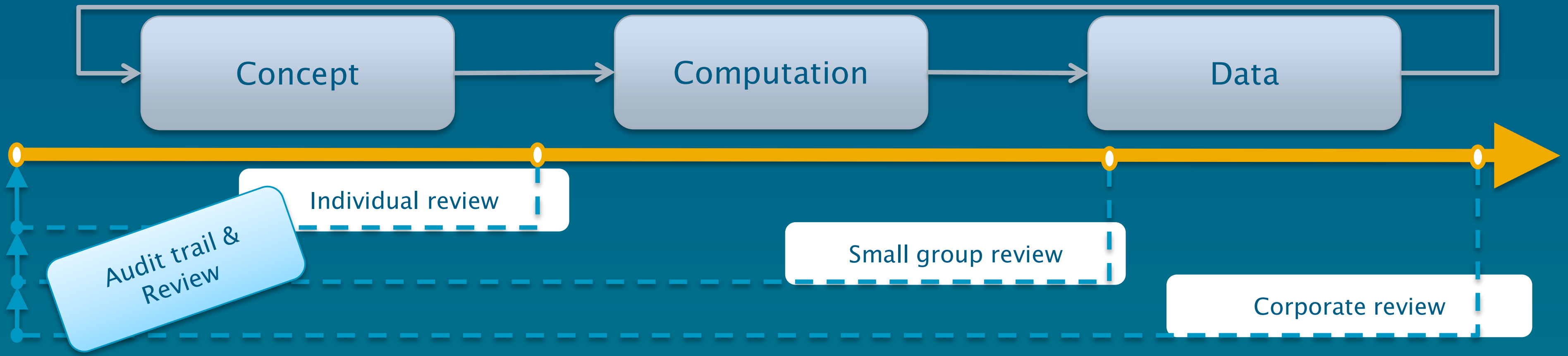
www.myexperiment.org

Preserving the record

- Key goal: record the whole experimental process prior to and during, rather than after
  - Ensures we efficiently generate a traceable, complete record of the work
  - Foundation for high quality sharing & reuse of data, extending data life
  - Scalable from a single lab to whole communities

TAGtivity	Wiki	BAE Systems	Rolls-Royce	Airbus
Organizing thoughts and reference material... experience	Semantic structure, search and reasoning	Software + Services for connecting engineers and experts to users and data	Orchestration of gas turbine design calculations	Robust, reliable and scalable data intensive collaboration

Centre for Fluid Dynamics Simulation Project



Data sources

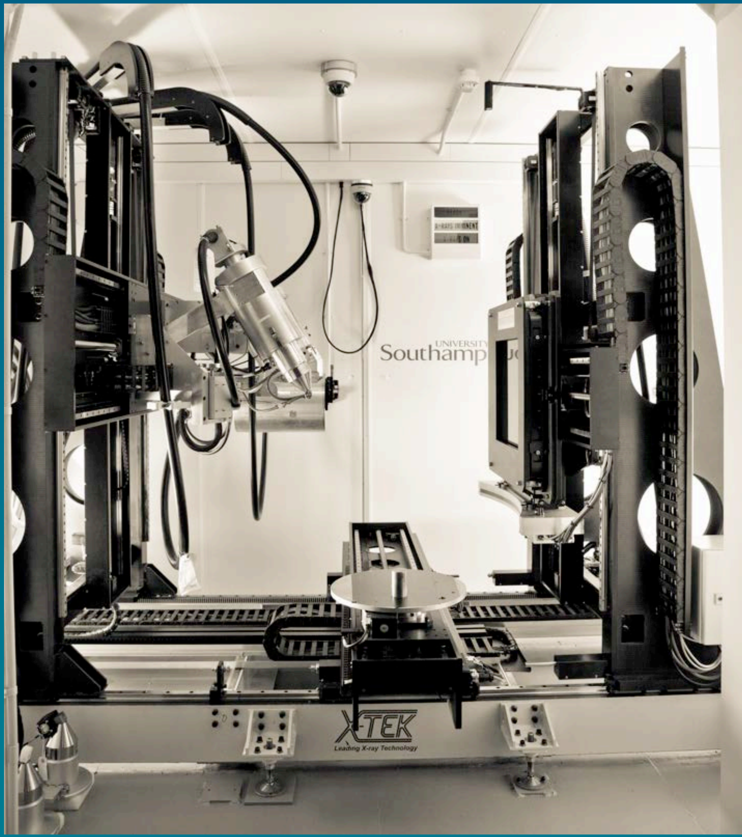
Tagtivity database Filesystem	Wiki database Workflow database Knowledge database Corporate database	Task database Conversation database Workflow tracking Simulation database	Workflow templates Workflow tracking Filesystem	Sharepoint database Active Directory

Technology

Microsoft Office 2007 SQL Server 2005/ 2008 Windows Presentation Foundation; Matlab	MediaWiki SQL Server 2008 D2R Server ARQ/SPARQL	Windows Server 2008 Hyper-V RC0 Windows HPC Server 2008 Beta2 SQL Server 2008 CTP6 Office Communication Server	Windows Workflow Foundation Windows CCS 2003 Linux (Interop) Visual Studio 2005 SQL Server 2005	Sharepoint Server Active Directory HP-UX (Interop) Windows Communication Foundation

μ-VIS X-ray Imaging Centre

- Five CT scanners, including
  - 225/450kV custom “hutch”, imaging up to 1x2m, panel shift and line detector
  - 225kV Nikon/Metris HMX with rototic sample exchange
  - Largest single scan >1TB
  - 60TB online data store, 10GbE connectivity
  - Workstations up to 32 CPU cores/128GB RAM/nVidia Tesla GPU rack
  - >100 users/year

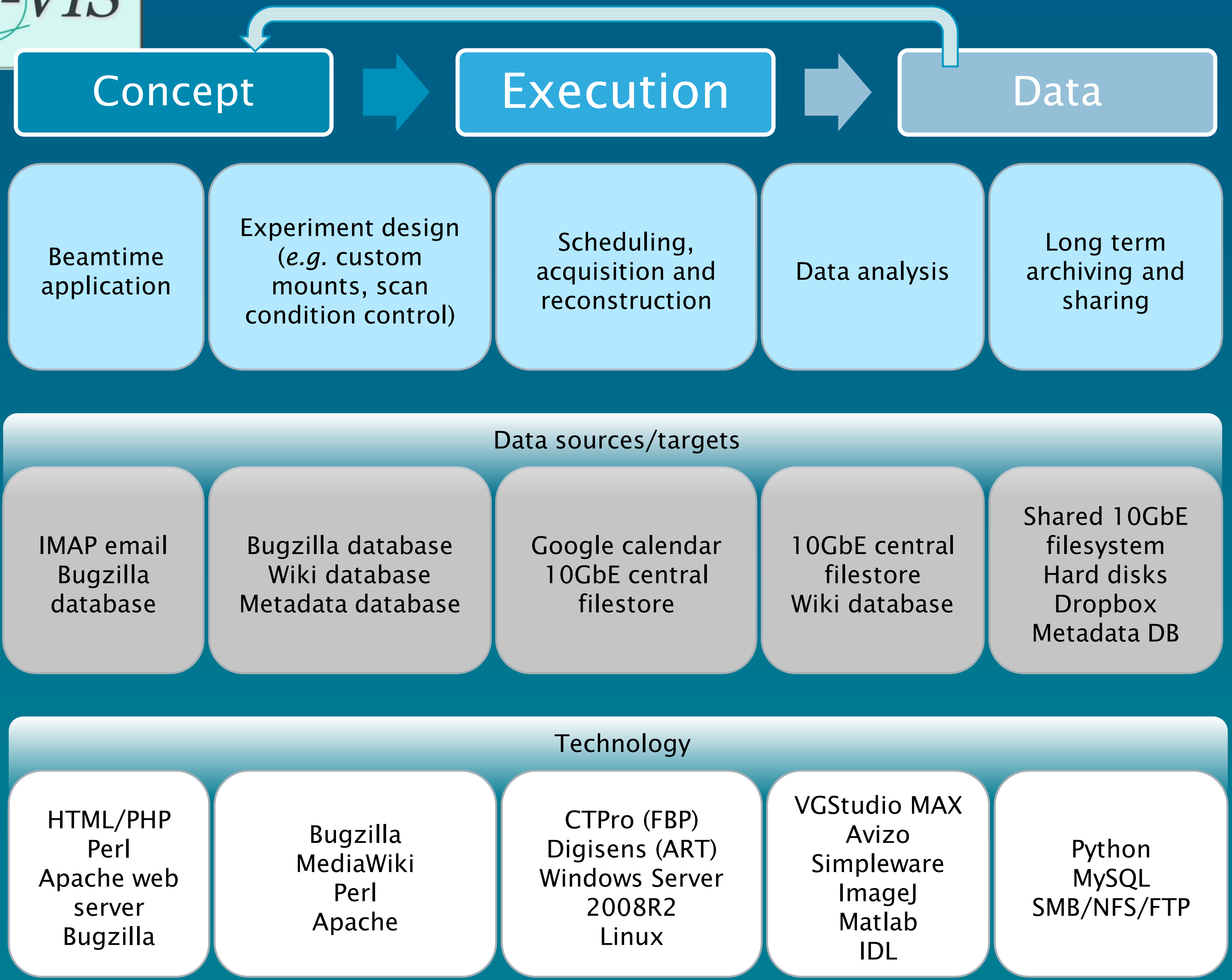


μ-VIS X-ray Imaging Centre
www.southampton.ac.uk/muvis

Multidisciplinary, Multiscale, Microtomographic Volume Imaging at Southampton

μ-VIS

UNIVERSITY OF Southampton



Concept → Execution → Data

Concept		Execution	Data	
Beamtime application	Experiment design (*e.g.* custom mounts, scan condition control)	Scheduling, acquisition and reconstruction	Data analysis	Long term archiving and sharing

Data sources/targets

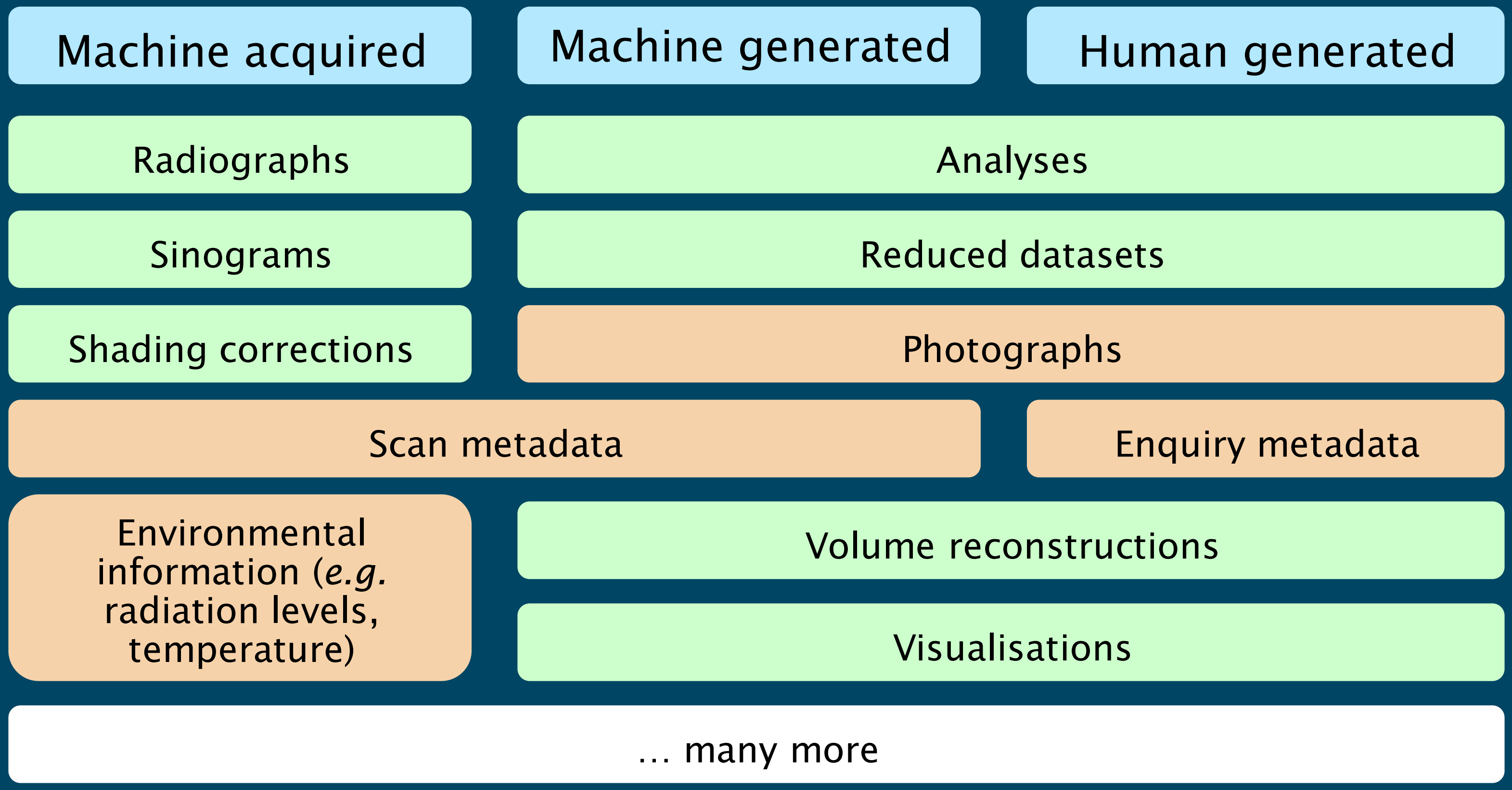
IMAP email Bugzilla database	Bugzilla database Wiki database Metadata database	Google calendar 10GbE central filestore	10GbE central filestore Wiki database	Shared 10GbE filesystem Hard disks Dropbox Metadata DB

Technology

HTML/PHP Perl Apache web server Bugzilla	Bugzilla MediaWiki Perl Apache	CTPro (FBP) Digisens (ART) Windows Server 2008R2 Linux	VGStudio MAX Avizo Simpleware ImageJ Matlab IDL	Python MySQL SMB/NFS/FTP

What data storage and sharing means

Data storage: database, central file store



Machine acquired

Machine generated

Human generated

Radiographs

Analyses

Sinograms

Reduced datasets

Shading corrections

Photographs

Scan metadata

Enquiry metadata

Environmental information (*e.g.* radiation levels, temperature)

Volume reconstructions

Visualisations

… many more

What metadata are relevant?

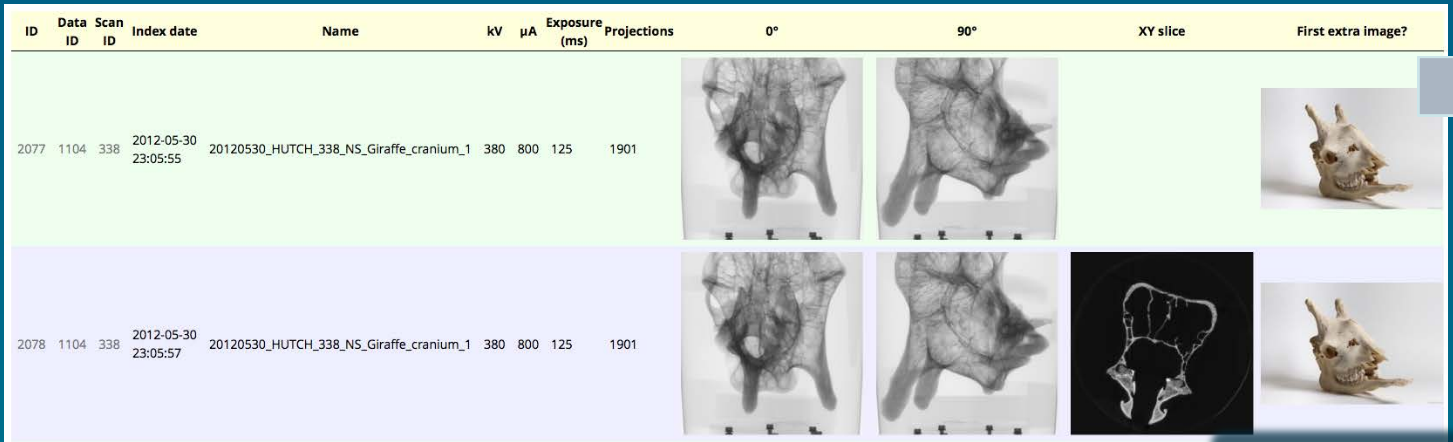
- In CT, a sensible minimum is:
  - Two projections (at 0 and 90 degrees)
  - A central slice of the reconstructed volume
  - All the available acquisition condition metadata (filters, kV, µA, source to detector distance *&c.*)
- For one CT scan, this might be 30 or 40MB; much more manageable than 50GB
- Once metadata are stored, a web interface provides tools to review and search

Further metadata

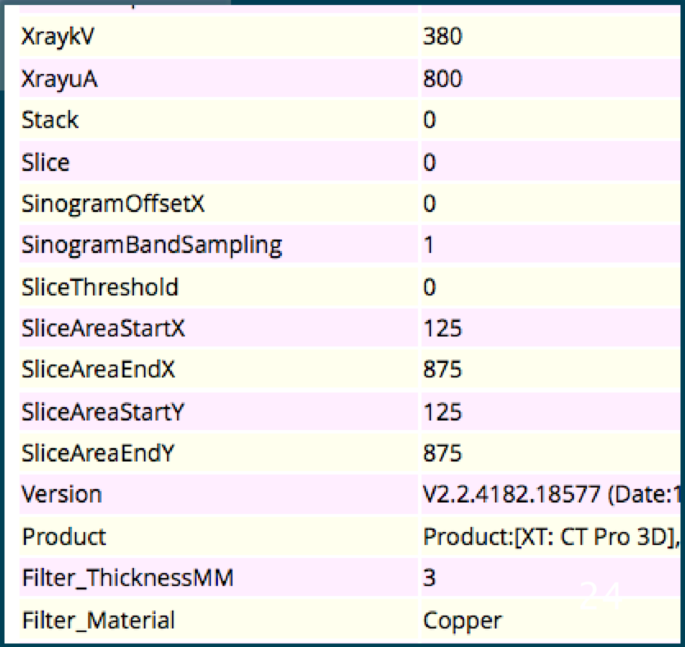
- In our case, if a user adds something “extra” to a CT scan directory, then this is also captured
  - Photographs of the sample, special sample mounting rigs, documentation, charts, videos, anything else
- … and “Smart Pen” output (operator notes) is added (→ searchable pdf file)
- E-lab books TBC

UNIVERSITY OF Southampton

Metadata: browsing and searching



ID	Data ID	Scan ID	Index date	Name	kV	µA	Exposure (ms)	Projections	0°	90°	XY slice	First extra image?
2077	1104	338	2012-05-30 23:05:55	20120530_HUTCH_338_NS_Giraffe_cranium_1	380	800	125	1901				
2078	1104	338	2012-05-30 23:05:57	20120530_HUTCH_338_NS_Giraffe_cranium_1	380	800	125	1901				
			2012-05-30									

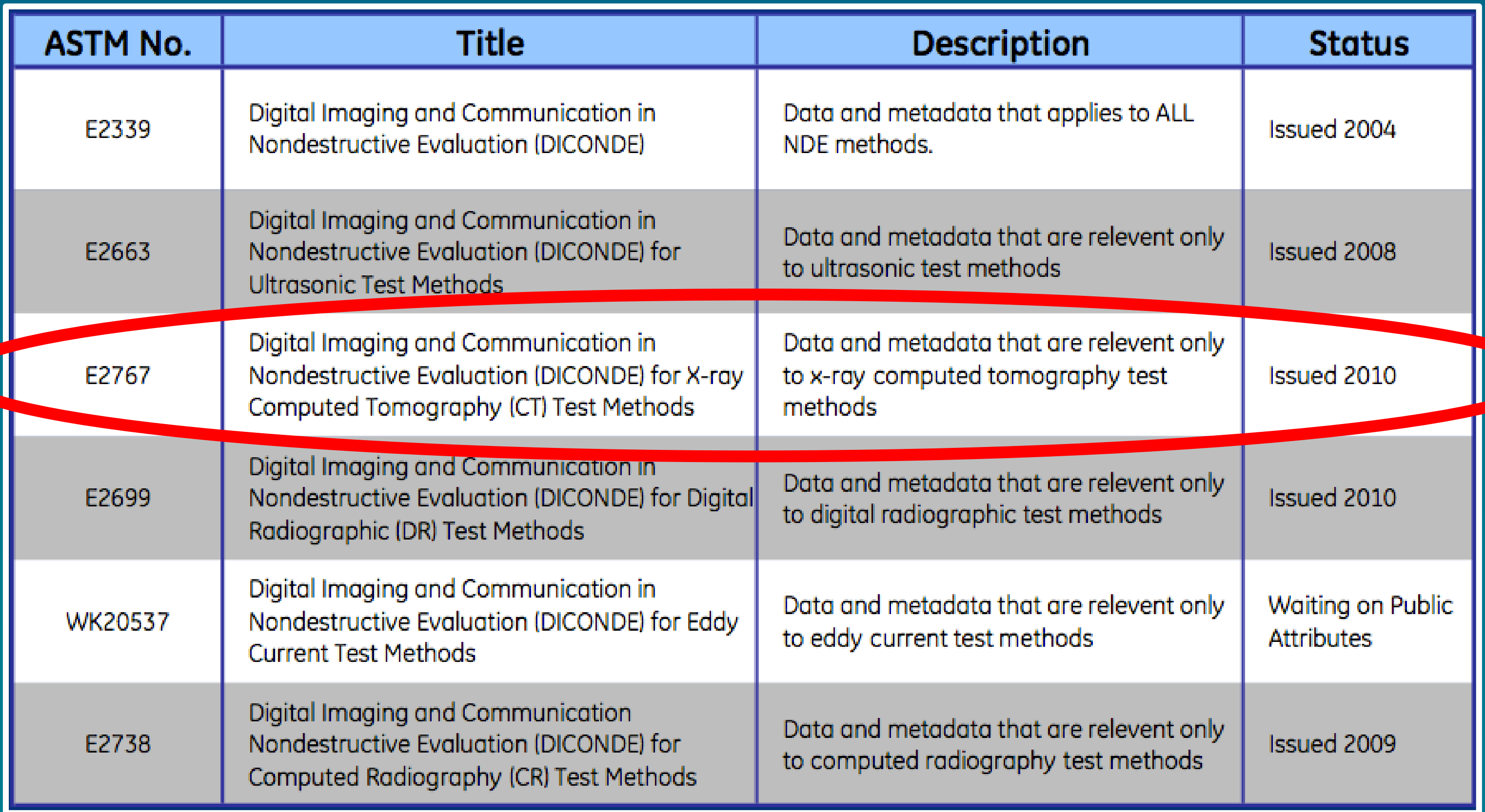


XraykV	380
XrayuA	800
Stack	0
Slice	0
SinogramOffsetX	0
SinogramBandSampling	1
SliceThreshold	0
SliceAreaStartX	125
SliceAreaEndX	875
SliceAreaStartY	125
SliceAreaEndY	875
Version	V2.2.4182.18577 (Date:1
Product	Product:[XT: CT Pro 3D],
Filter_ThicknessMM	3
Filter_Material	Copper

Web browser interface
- 2 projections & central slice
- Extensive metadata
- Dataset names, IDs, times
- NetApp/archive location
- Original proposal, emails…

Metadata standards: DICONDE

- Direct mapping of DICOM to industrial CT
- Firmly established approach, detailed

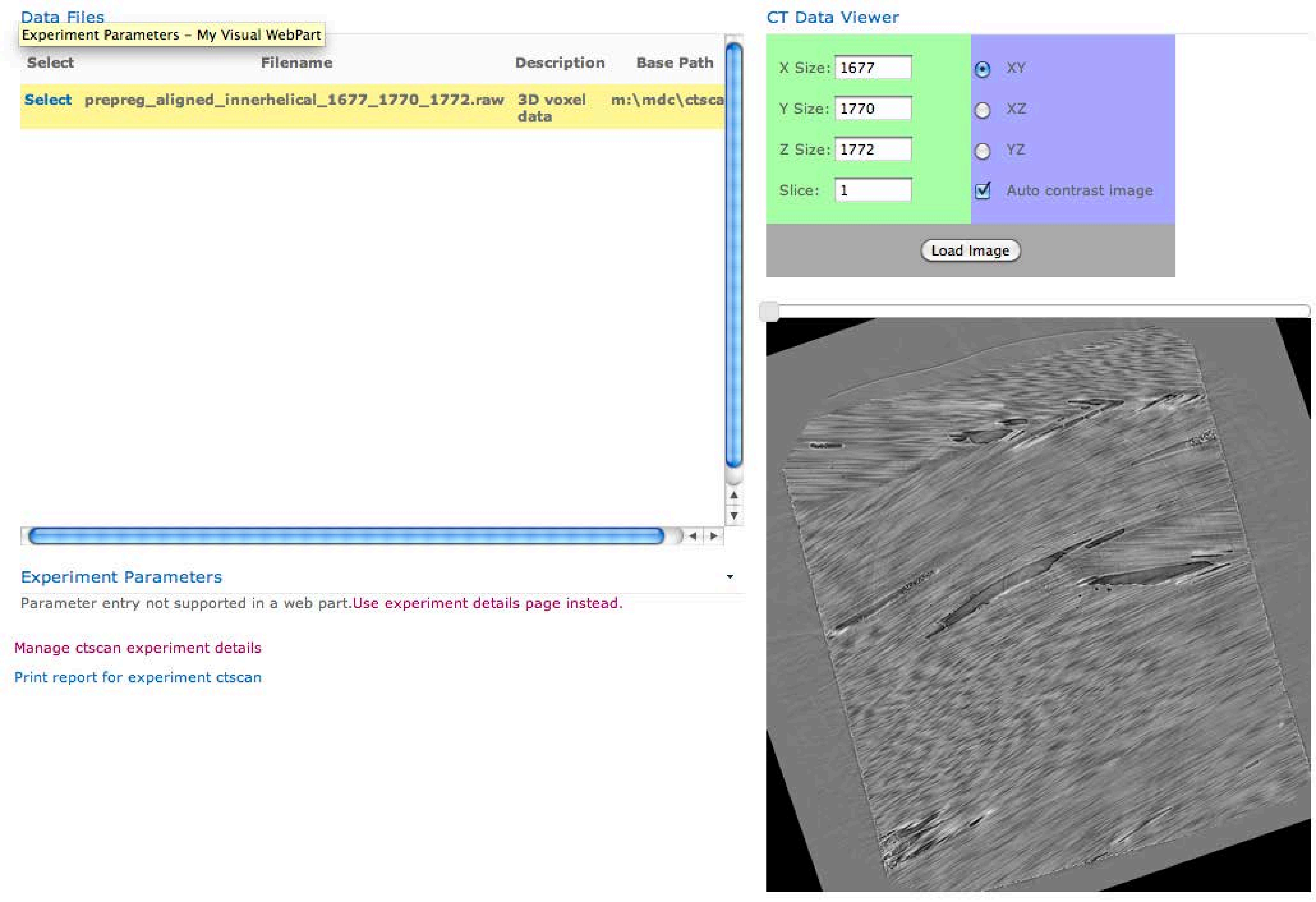


ASTM No.	Title	Description	Status
E2339	Digital Imaging and Communication in Nondestructive Evaluation (DICONDE)	Data and metadata that applies to ALL NDE methods.	Issued 2004
E2663	Digital Imaging and Communication in Nondestructive Evaluation (DICONDE) for Ultrasonic Test Methods	Data and metadata that are relevent only to ultrasonic test methods	Issued 2008
E2767	Digital Imaging and Communication in Nondestructive Evaluation (DICONDE) for X-ray Computed Tomography (CT) Test Methods	Data and metadata that are relevent only to x-ray computed tomography test methods	Issued 2010
E2699	Digital Imaging and Communication in Nondestructive Evaluation (DICONDE) for Digital Radiographic (DR) Test Methods	Data and metadata that are relevent only to digital radiographic test methods	Issued 2010
WK20537	Digital Imaging and Communication in Nondestructive Evaluation (DICONDE) for Eddy Current Test Methods	Data and metadata that are relevent only to eddy current test methods	Waiting on Public Attributes
E2738	Digital Imaging and Communication Nondestructive Evaluation (DICONDE) for Computed Radiography (CR) Test Methods	Data and metadata that are relevent only to computed radiography test methods	Issued 2009

Heterogeneous Data Centre

- To provide a user-centric software system for users to store and share their data and metadata in a usable way
- The user decides their own metadata structures
  - Stored as name-value pairs and can be hierarchical, providing a flexible approach to data management
- To support a wide variety of data, from small text files to large voxel data files.
- Provide the ability for users to tag data sets with **any** relevant metadata

CT Dataset browsing in HDC



Data Files

Select	Filename	Description	Base Path
Select	prepreg_aligned_innerhelical_1677_1770_1772.raw	3D voxel data	m:\mdc\ctsca

Experiment Parameters

Parameter entry not supported in a web part.Use experiment details page instead.

Manage ctscan experiment details

Print report for experiment ctscan

CT Data Viewer

X Size: 1677

Y Size: 1770

Z Size: 1772

Slice: 1

● XY

○ XZ

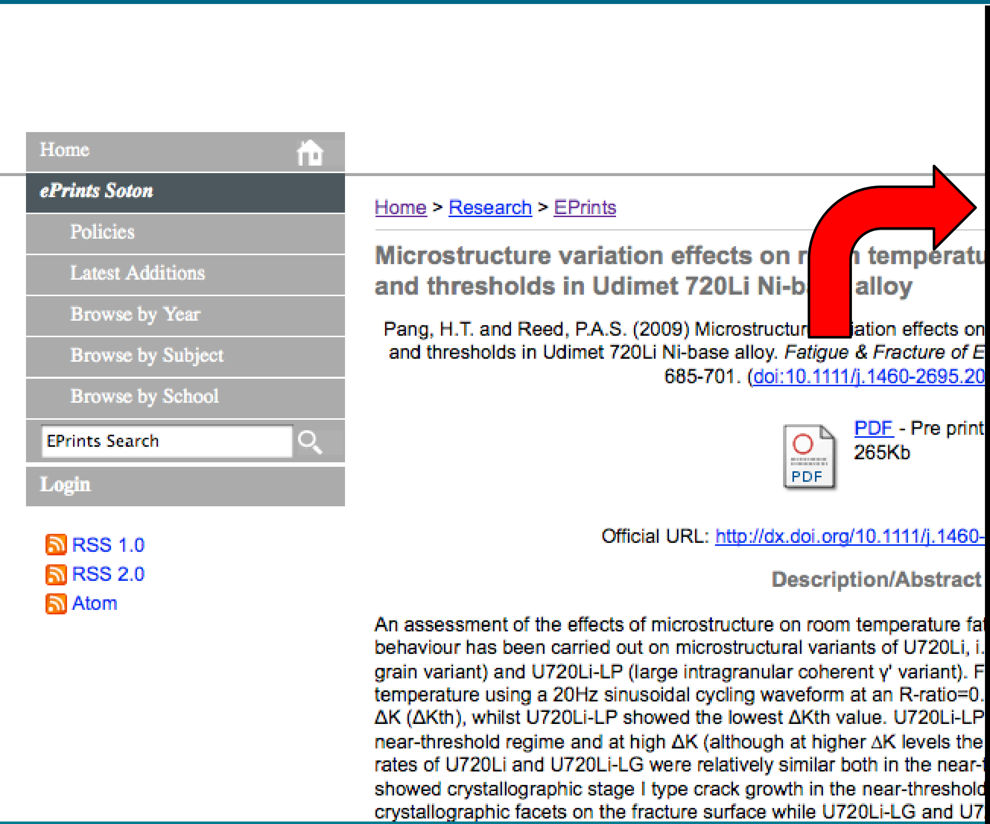
○ YZ

☑ Auto contrast image

UNIVERSITY OF Southampton

Heterogeneous Data Centre

- Data can be uploaded via EPrints with the EP2DC service or directly



Home > Research > EPrints

Microstructure variation effects on r n temperatu and thresholds in Udimet 720Li Ni-b alloy

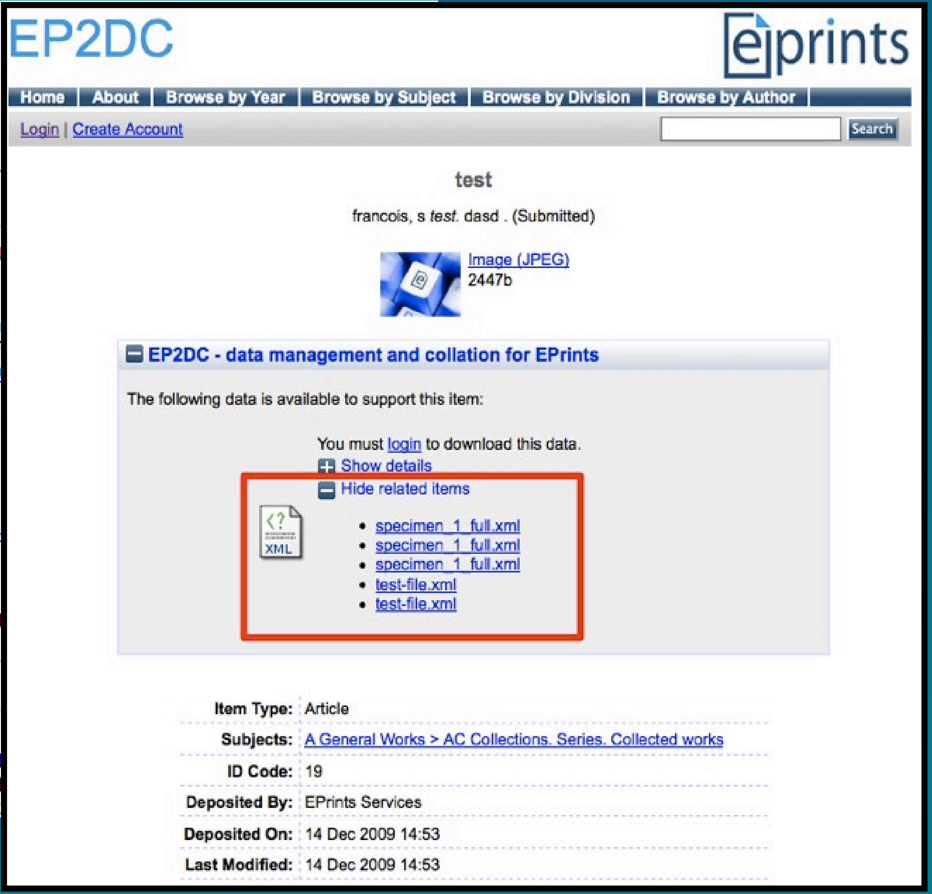
Pang, H.T. and Reed, P.A.S. (2009) Microstructur iation effects on and thresholds in Udimet 720Li Ni-base alloy. *Fatigue & Fracture of E* 685-701. (doi:10.1111/j.1460-2695.20

PDF - Pre print
265Kb

Official URL: http://dx.doi.org/10.1111/j.1460-

Description/Abstract

An assessment of the effects of microstructure on room temperature fa behaviour has been carried out on microstructural variants of U720Li, i. grain variant) and U720Li-LP (large intragranular coherent γ' variant). F temperature using a 20Hz sinusoidal cycling waveform at an R-ratio=0. ΔK (ΔKth), whilst U720Li-LP showed the lowest ΔKth value. U720Li-LP near-threshold regime and at high ΔK (although at higher ΔK levels the rates of U720Li and U720Li-LG were relatively similar both in the near- showed crystallographic stage I type crack growth in the near-threshold crystallographic facets on the fracture surface while U720Li-LG and U7



test

francois, s *test.* dasd . (Submitted)

Image (JPEG)
2447b

EP2DC - data management and collation for EPrints

The following data is available to support this item:

You must login to download this data.

Show details

Hide related items

- specimen_1_full.xml
- specimen_1_full.xml
- specimen_1_full.xml
- test-file.xml
- test-file.xml

Item Type:	Article
Subjects:	A General Works > AC Collections. Series. Collected works
ID Code:	19
Deposited By:	EPrints Services
Deposited On:	14 Dec 2009 14:53
Last Modified:	14 Dec 2009 14:53

Archiving practicalities

- Many options available: carefully indexed disks or tapes, online NAS, cloud storage
- It is impossible to *100% guarantee* that data will never be lost
  - We can get close (90%, 99%, 99.9%...)
- Cost scales with reliability

Archiving practicalities *(continued)*

- One copy on one hard disk: ~10-20% chance of data loss over 5 years
  - Approximate cost in 2012: ~$10/TB/year
- Two copies on two separate disks: ~1-4% chance of data loss over 5 years
  - Approximate cost in 2012: ~$20/TB/year
- "Enterprise" class storage (*e.g.* NetApp): <1% chance of data loss over 5 years
  - Approximate cost in 2012: ~$500/TB/year

Cloud storage

- Provides a scalable and reliable option to store data, e.g. Amazon S3
  - ‘11 nines’ reliability levels
- Typical pricing is O($0.10)/GB/month (2012)
  - around $1200/TB/year; additional charges for uploading and downloading
- Recently, providers have been waiving upload charges
  - 30GB download O($5)
  - May make storing large amounts of data with relatively few downloads more attractive

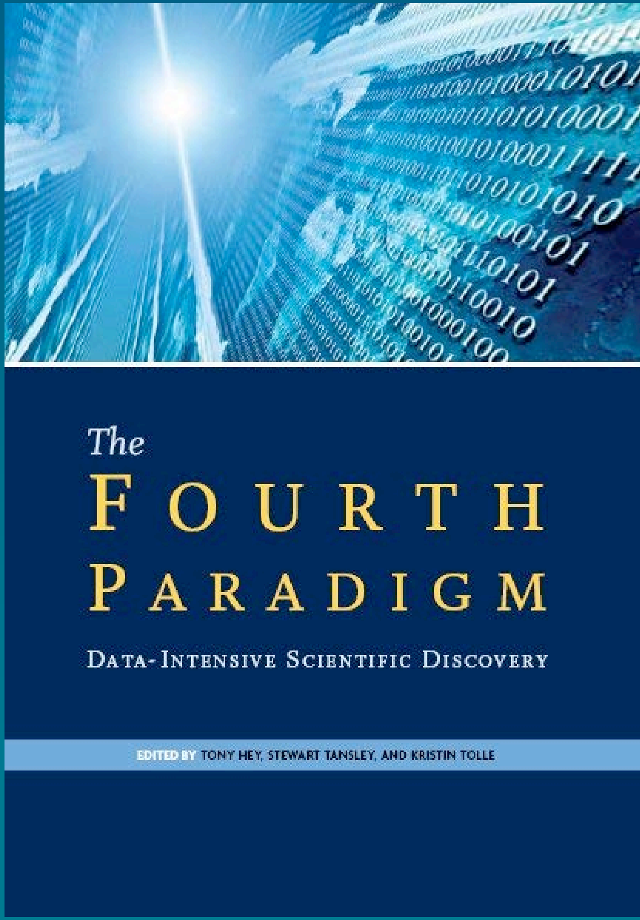
Final thoughts

- Local behaviour – what's going on in your lab? Are people carefully looking after their datasets?
  - The generation of quality data and metadata is best done *concurrently*
  - If we look after it, we can make better use of it
- Look for technology that will work well with your, requirements, current systems and budgets
  - *Many* strategies & tools are already in place
  - BioSimGrid, ROOT…
- The first step is, start now…
  - The sooner, the better

UNIVERSITY OF Southampton

Final thoughts

- Contribution to the ‘4th Paradigm’?
  - Science driven by the capture, curation, analysis of large data
  - *All* data becomes publically **available** and **usable**, like books in the library
- *1st: Empirical description of nature (~1000 yrs ago)*
- *2nd: Mathematical theory (~100yrs)*
- *3rd: Large simulation (~30 yrs)*

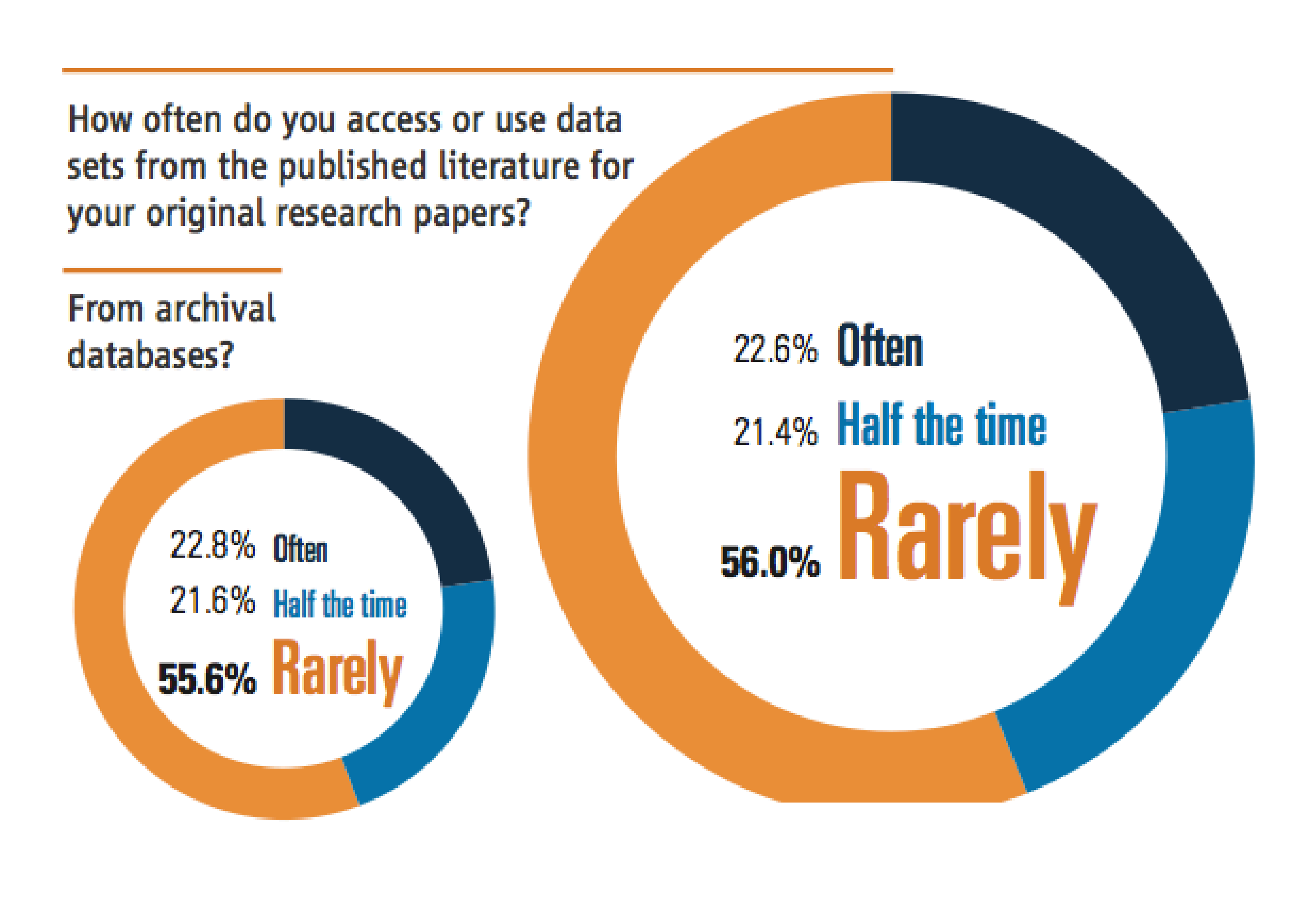

The FOURTH PARADIGM

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

Acknowledgements

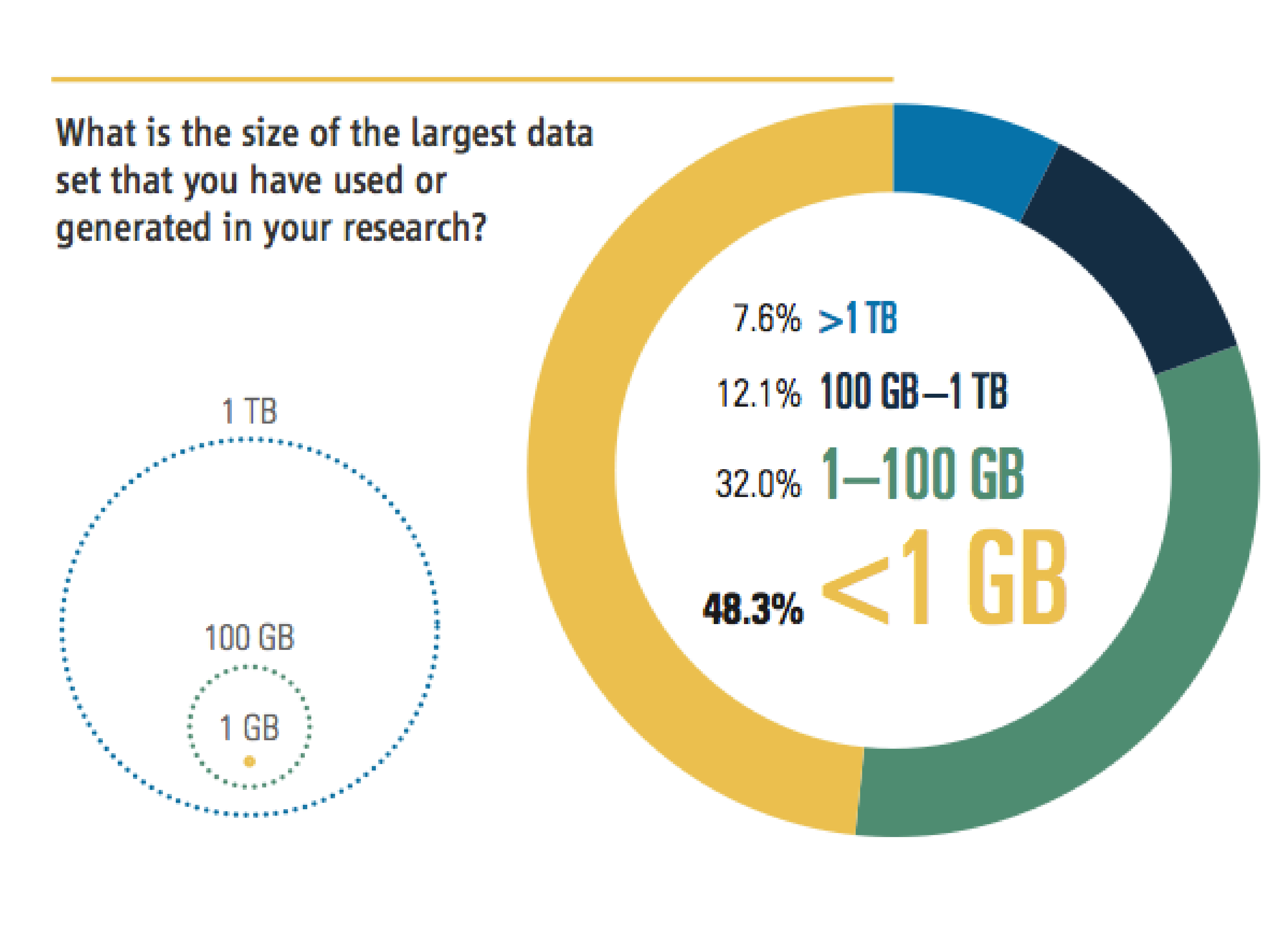
- The µ-VIS team
- Mark Scott (HDC)
- Oliver Bunk (Swiss Light Source)
- Uli Felzmann (Australian Synchrotron)



How often do you access or use data sets from the published literature for your original research papers?

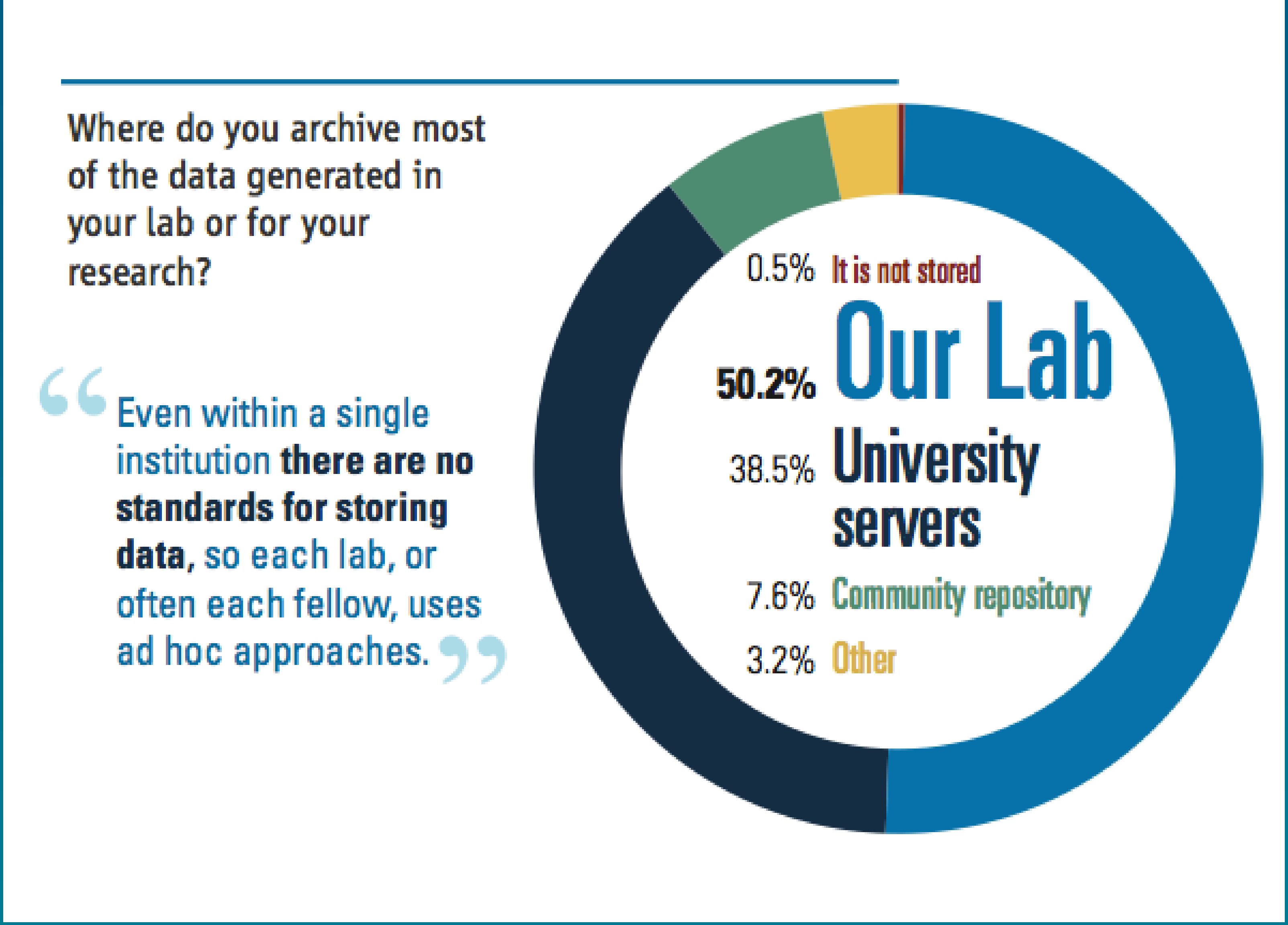
From archival databases?

www.sciencemag.org/special/data/



What is the size of the largest data set that you have used or generated in your research?

www.sciencemag.org/special/data/



Where do you archive most of the data generated in your lab or for your research?

> Even within a single institution **there are no standards for storing data**, so each lab, or often each fellow, uses ad hoc approaches.

www.sciencemag.org/special/data/