

Microsoft® Research

Faculty Summit 2010

UK Digital Curation Centre : enabling research data management at the coalface

Dr Liz Lyon

Associate Director DCC / Director UKOLN

University of Bath, UK

Overview

Microsoft® Research

Faculty Summit 2010

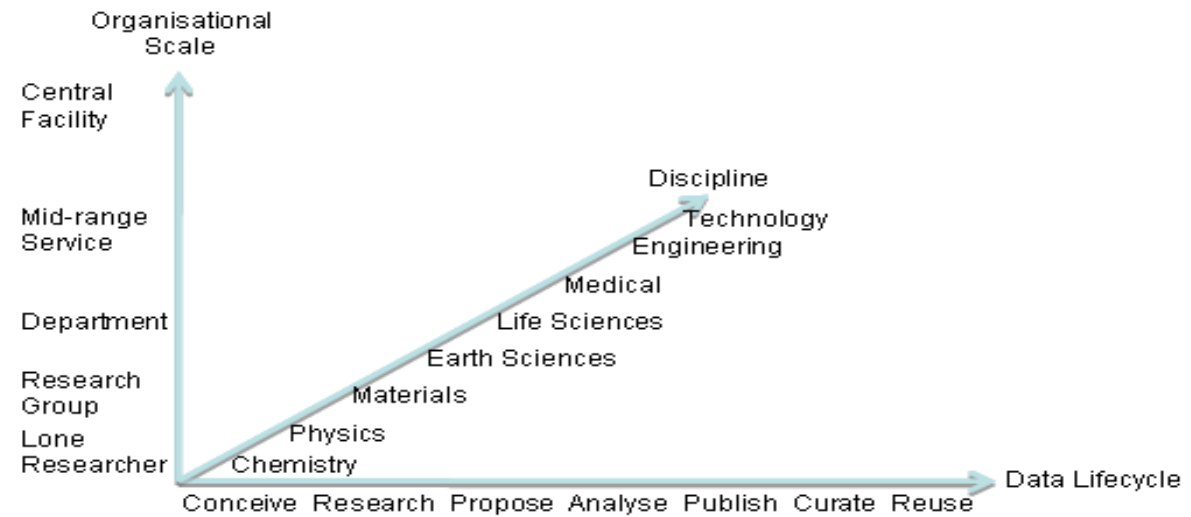
1. Moving data across boundaries : structural science
2. Managing data in institutions : emerging DCC tools
3. Making data count : publication and attribution



I₂S₂

Infrastructure for Integration in Structural Sciences

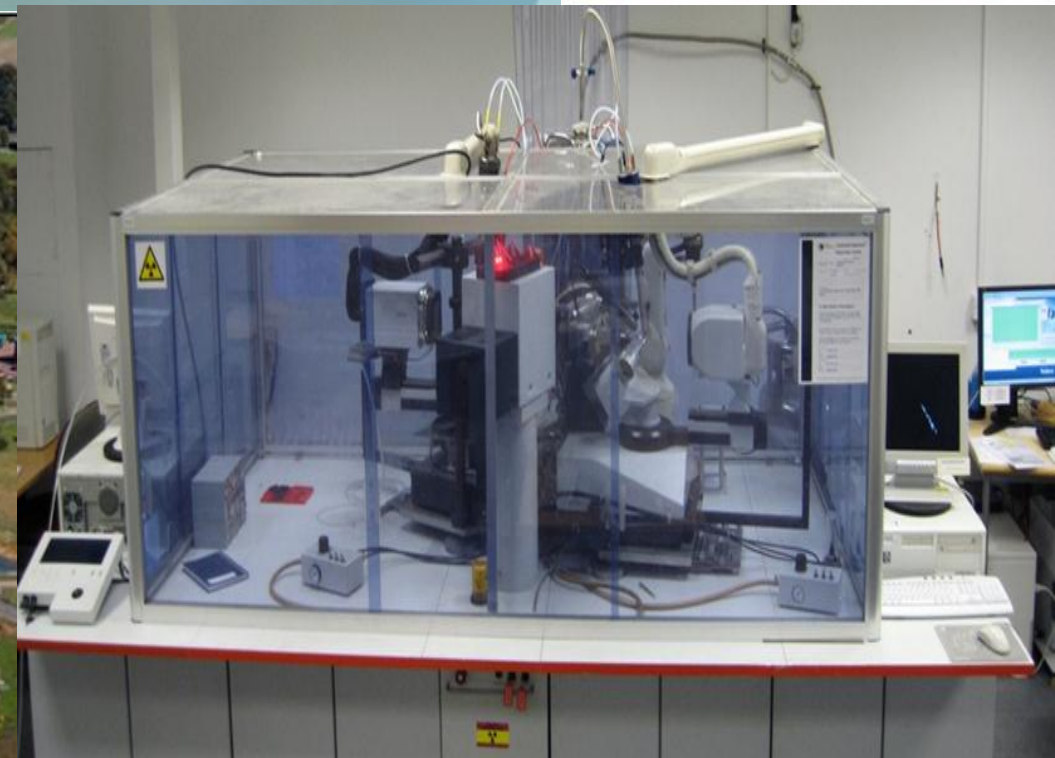
<http://www.ukoln.ac.uk/projects/I2S2/>




- “**Bridging the chasm**” between the local laboratory bench and large scale facilities e.g. DIAMOND synchotron
- Develop Integrated Information Model
- Use cases and Inter-disciplinary Pilots
- Cost-benefit analysis: before and after



Structural Sciences Infrastructure



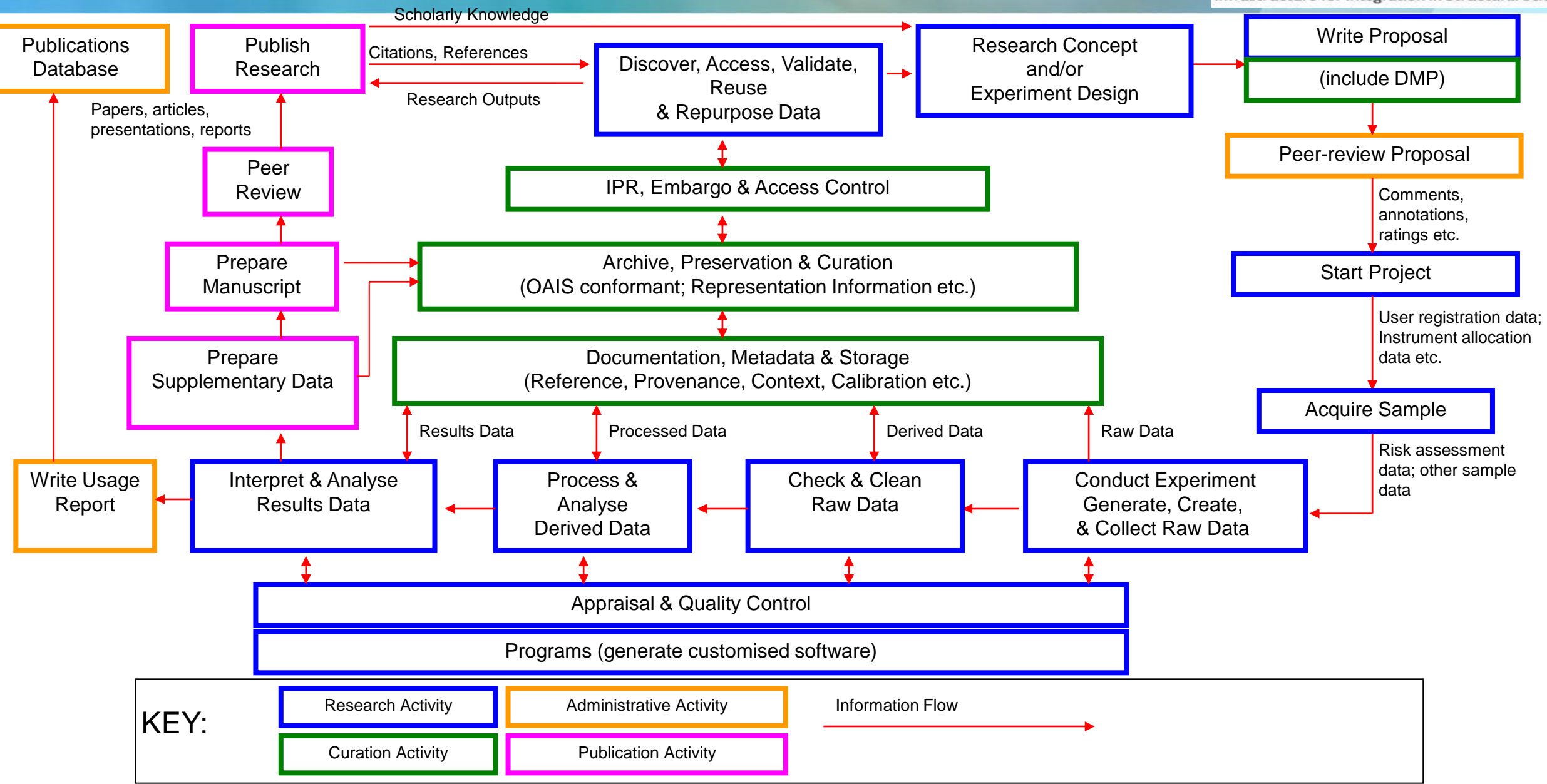
 UNIVERSITY OF CAMBRIDGE

Department of
Earth Sciences



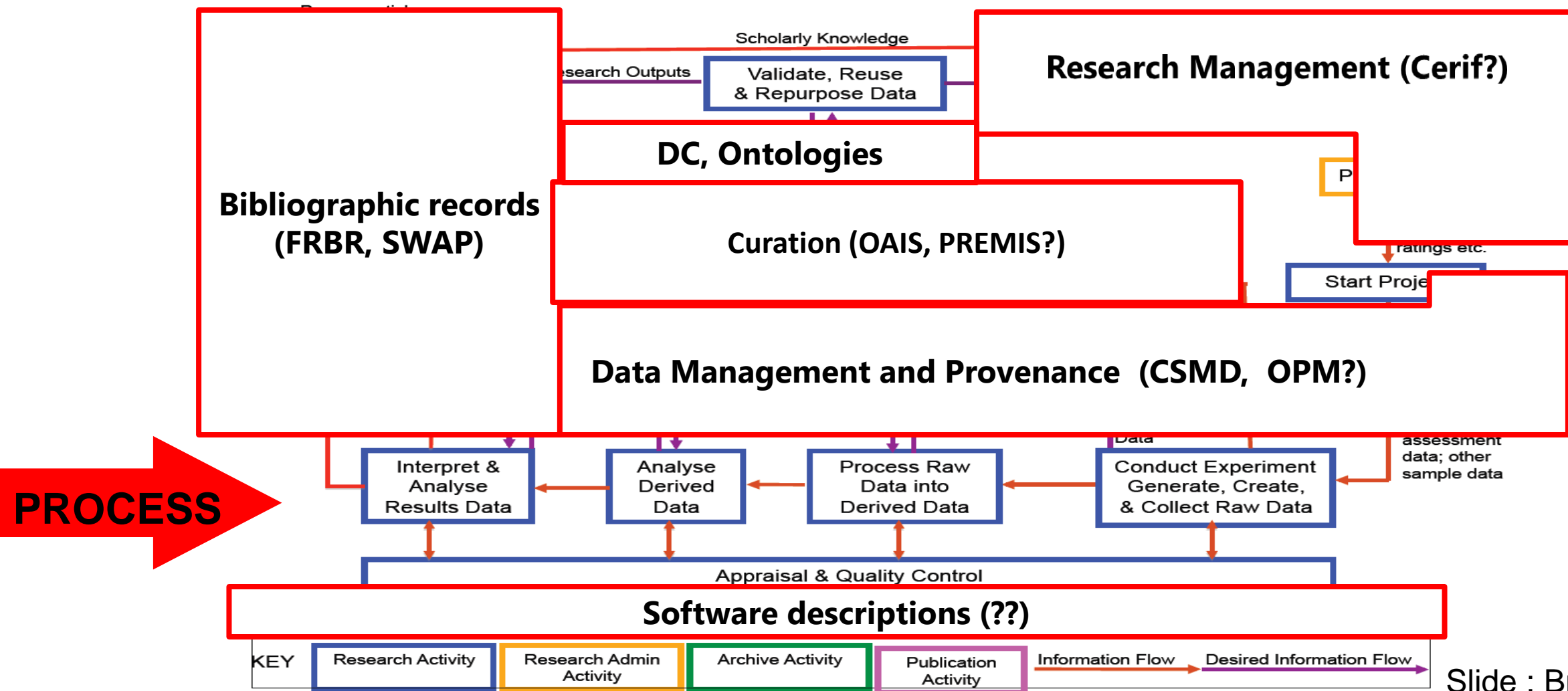
	Diamond Light Source Synchotron	National Crystallography Service University of Southampton	Local Earth Sciences Lab University of Cambridge
Function	International service -multiple communities	UK service - multiple institutions. Also uses Diamond	Lone researcher at institution - uses NCS and ISIS large-scale facility
Administration	Peer-reviewed proposal required	Vetted applications. Electronic & paper-based records –experiments, safety ERA, instrument time	Multiple proposals, multiple forms
Workflow	Formulaic and bespoke	Formulaic	Complex, unrecorded
Software	In-house scripts	In-house scripts + open-source suite	In-house scripts + open-source suite
Raw data storage	In-house GDA store	ATLAS data-store	Laptop / local server
Derived data storage	Taken offsite on laptop / USB stick	eCrystals repository	Laptop / local server / USB stick
Metadata	Core Scientific MetaData Model	eBank/eCrystals schema	?
Identifiers	Beam-line number	DOI InChI	?

An Idealised Scientific Research Activity Lifecycle Model



Existing work : mappings and gaps

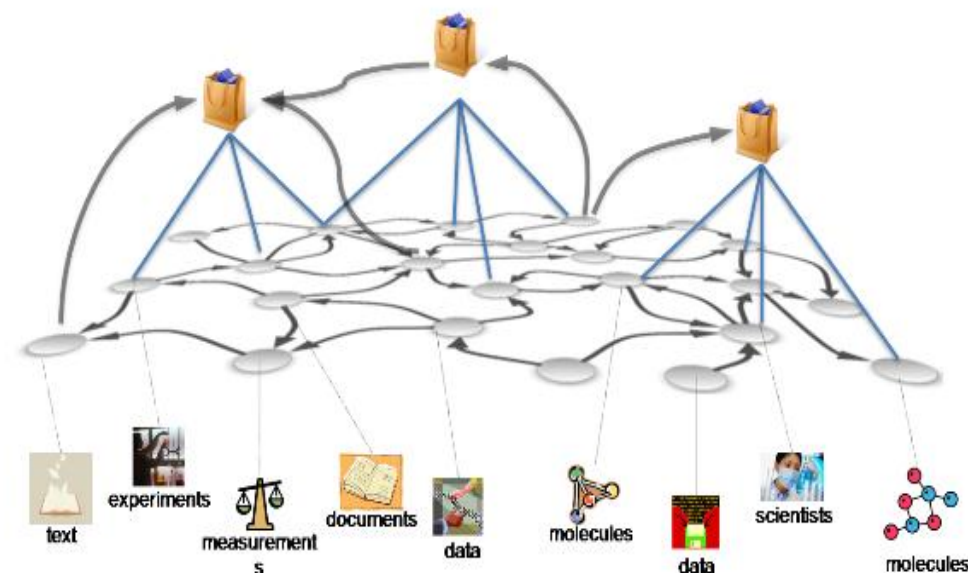
An Idealised Scientific Research Data Lifecycle Model



- Focus on Open Methodology
- Develop Data Model
- Join up to other Data Model work
 - OreChem
 - Data Conservancy
- Linked data approach
- <http://www.ukoln.ac.uk/projects/I2S2/>

oreChem Project

Integrating Chemistry Scholarship with the Semantic Web

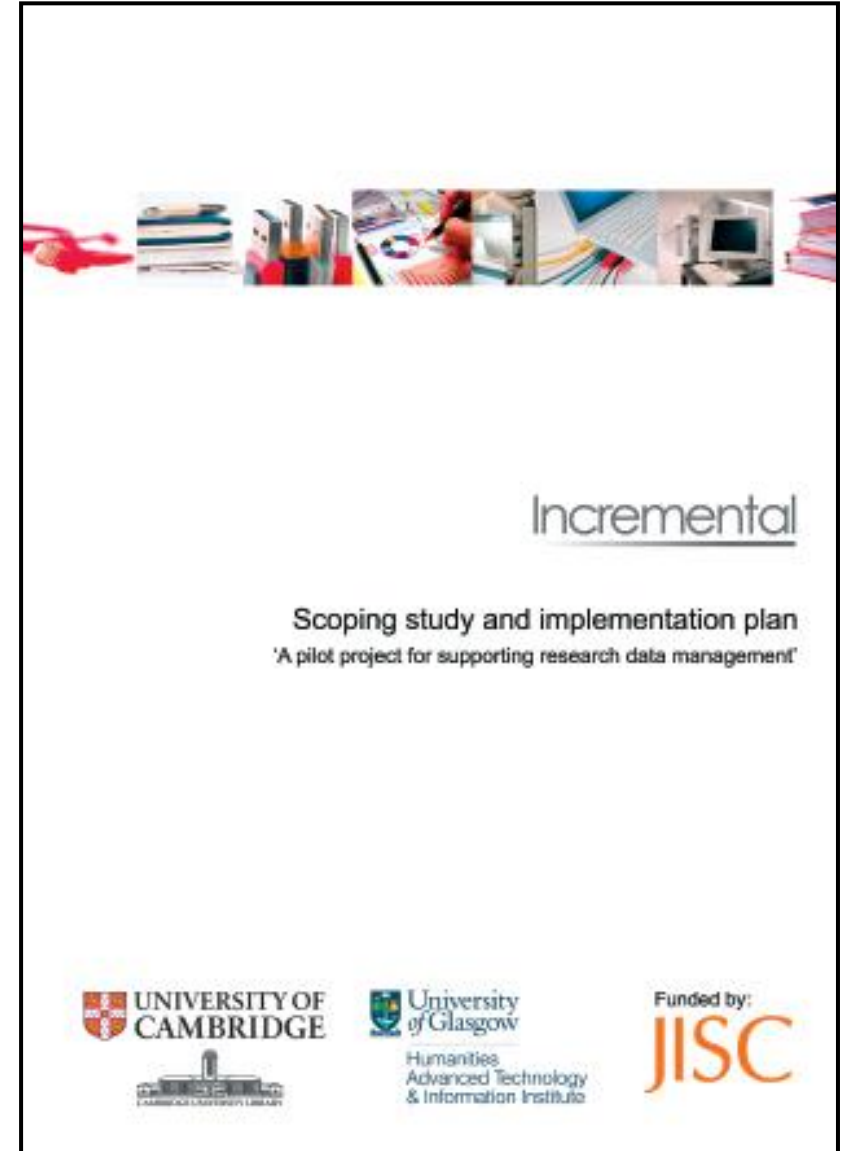


dataconservancy.org

*“...it is apparent that **the greatest need is for a robust data management infrastructure which supports each researcher in capturing, storing, managing and working with all the data generated during an experiment.** Internal sharing of research data amongst collaborating scientists ... is also a primary concern as is a requirement for access to research data in the long run so that a researcher ... can return to and validate the results well into the future.”*

Institutional perspective : Scoping study

- Creating & organising data
- Storage and access
- Back-up
- Preservation
- Sharing and re-use



"It's hard to overcome your personal investment... it's like giving away your baby"

"While many researchers are positive about sharing data in principle, they are almost universally reluctant in practice. using these data to publish results before anyone else is the primary way of gaining prestige in nearly all disciplines."

"I just back everything up onto data sticks. I didn't even know you could back-up to servers".

<http://www.flickr.com/photos/mattimattila/3003324844/>



"The policy was huge and not very clear. It took a few attempts to understand, whereas you just want a quick yes/no answer."

The majority of people felt that some form of policy or guidance was needed....

Emerging funder requirements



National Science Foundation
WHERE DISCOVERIES BEGIN

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

welcometrust

require that the applicants provide a data management and sharing plan as part of their application; and

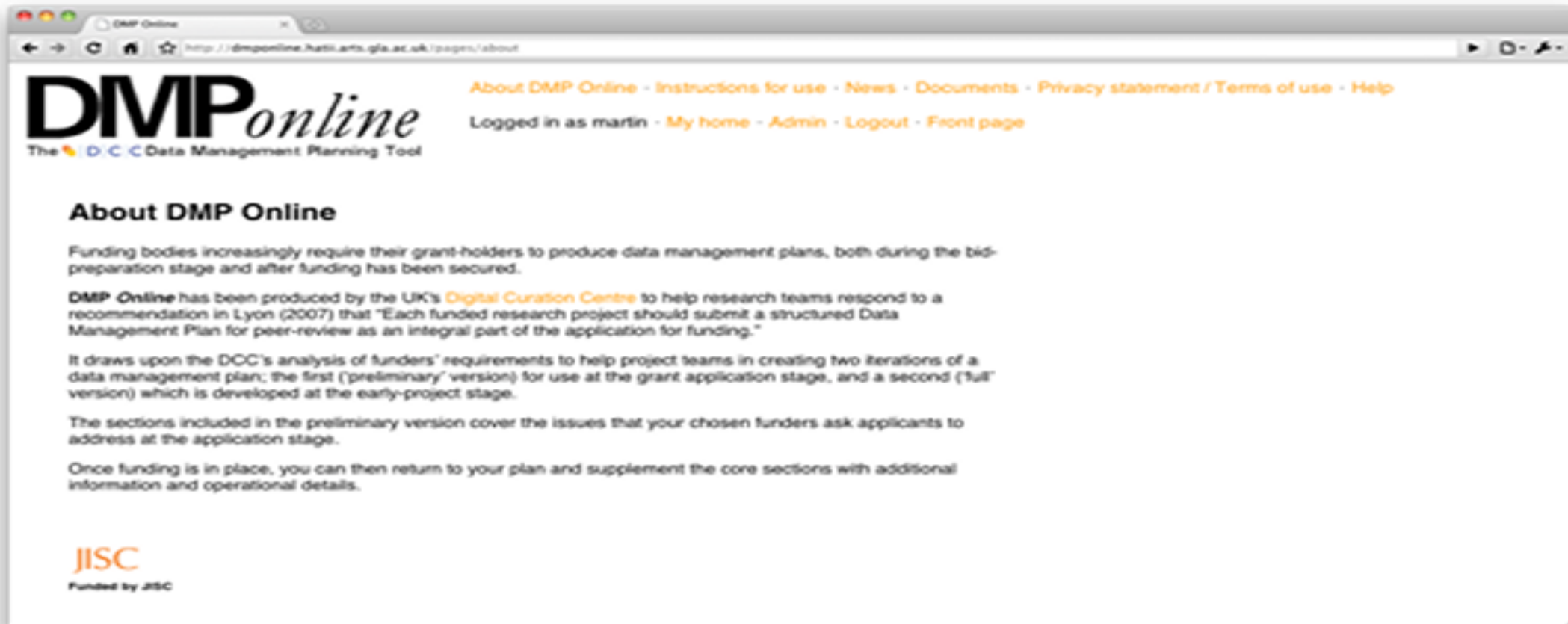


Checklist for a Data Management Plan

Post-Consultation (v3.0: 19 May 2010)

Martin Donnelly (University of Edinburgh, martin.donnelly@ed.ac.uk)
Sarah Jones (University of Glasgow, s.jones@hatii.arts.gla.ac.uk)

- Data types, formats, standards, capture
- Ethics and Intellectual Property
- Access, sharing and re-use
- Short-term storage & data management
- Deposit & long-term preservation
- Adherence and review



DMP Online
Currently updating Version 2.0
Version 3.0 summer 2010

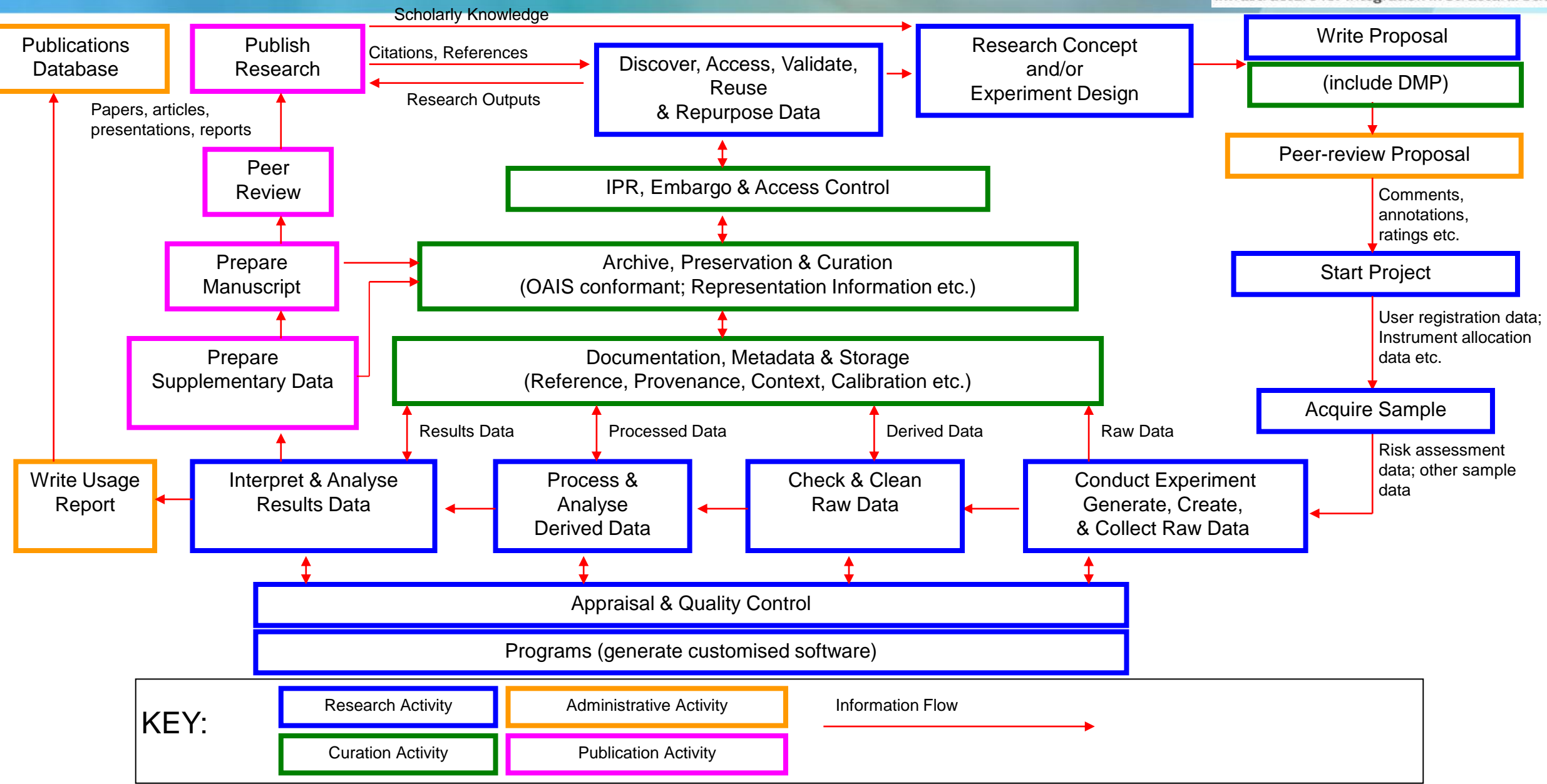
<http://www.dcc.ac.uk/dmponline>

Making DMPs work : the start of a long process...

- **Embed DMPs** in research lifecycles / activity model as the norm
- **Code of Conduct for Research**
- **Assess & review DMPs** (not just the science content of proposals)
- **Educate reviewers** (DCC guidance for social science in prep)
- **Manage compliance**
- **Infrastructure to share DMPs**
- **Analyse cost-benefits**



An Idealised Scientific Research Activity Lifecycle Model





Editorial

Nature Cell Biology **11**, 1273 (2009)
doi:10.1038/ncb1109-1273a

Sharing data

Reference datasets should be accessible independently of scientific papers in a citable form, allowing attribution.

Incentives?



Scholar Factor (SF)

Philip E. Bourne[✉], J. Lynn Fink

Correspondence

Nature Biotechnology **27**, 984 - 985 (2009)
doi:10.1038/nbt1109-984b



Accreditation and attribution in data sharing

Gudmundur A Thorisson¹

1. Department of Genetics, University of Leicester, UK.

Credit where credit is overdue

EDITORIAL

A universal tagging system that links data sets with the author(s) that generated them is essential to promote data sharing within the proteomics and other research communities.

Data citation, credit, metrics, attribution

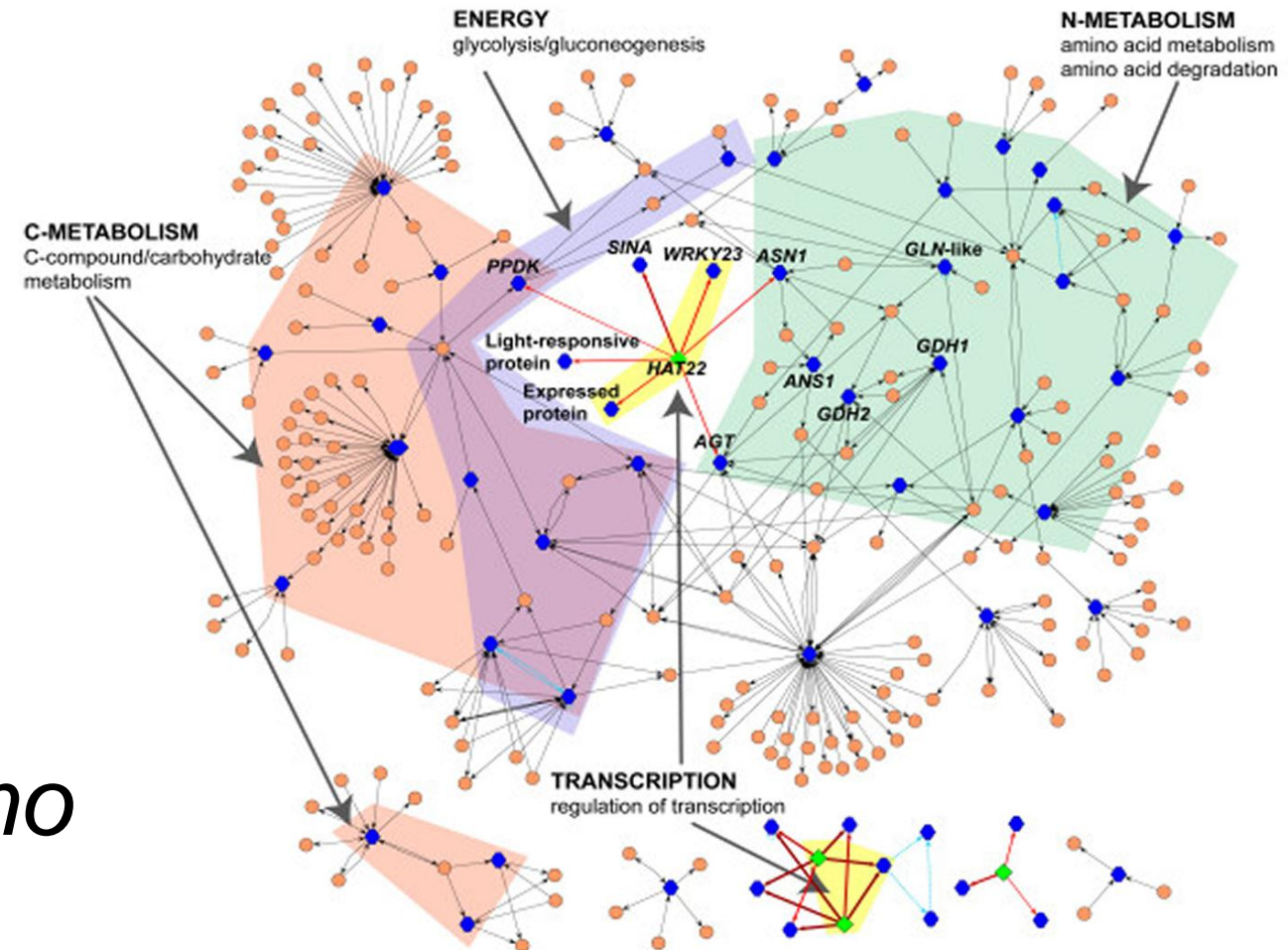
Complexity : what are we citing?

Macro

- Journal
- Article
- Workflow
- Visualisation
- Model
- Data
- Annotation
- Concept

Micro / Nano

Attribution granularity

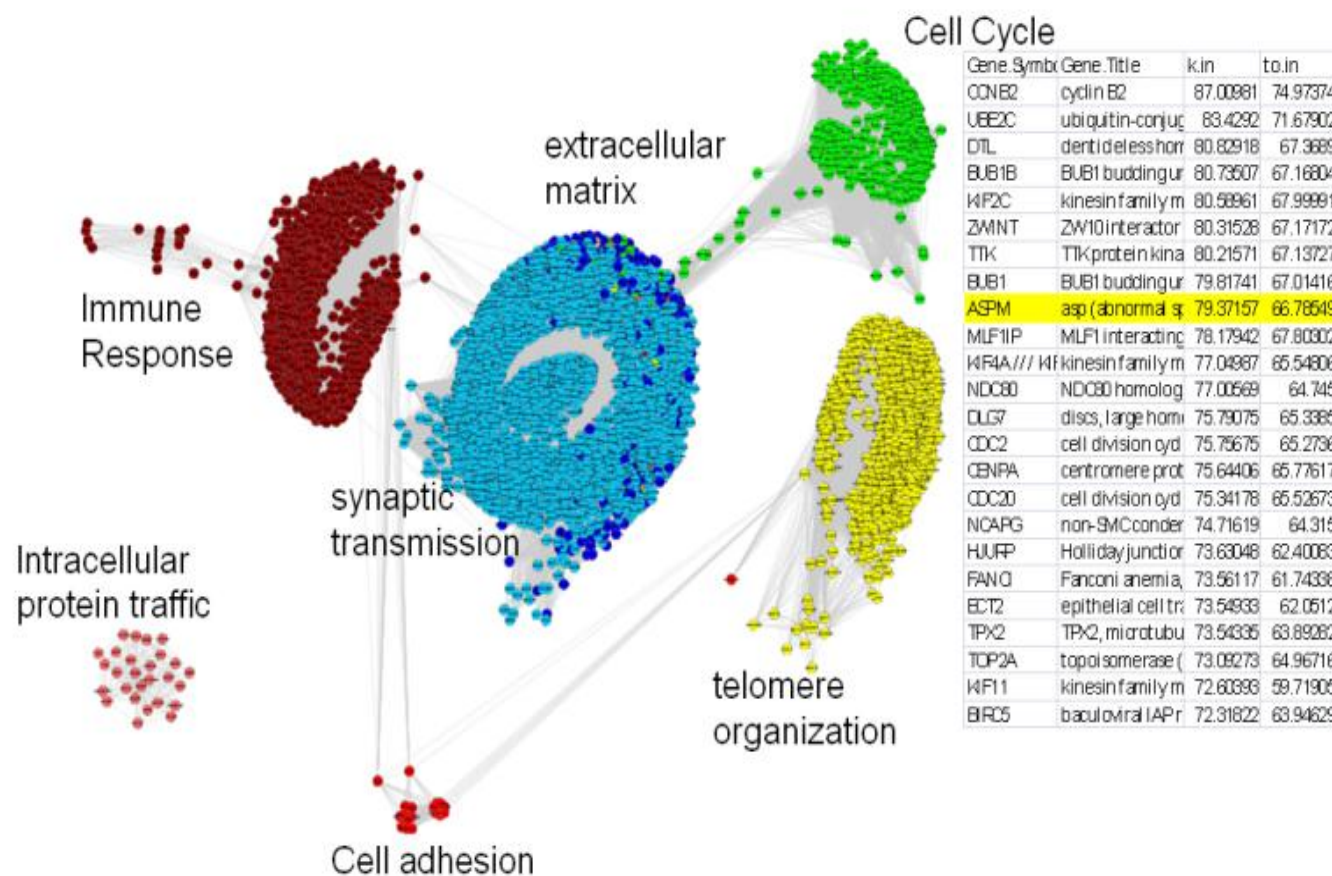


Gene hunters uncover networks behind disease

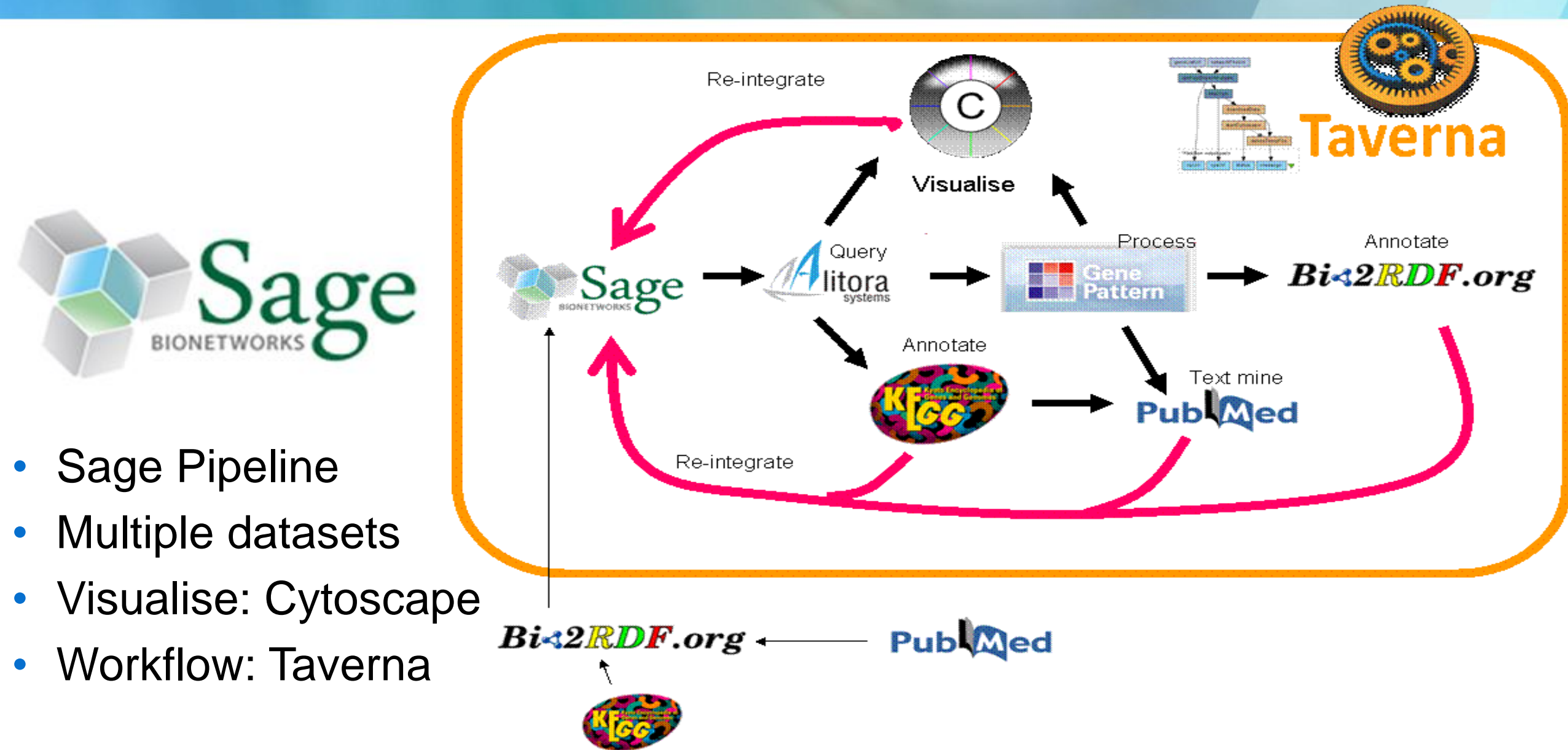
New technique offers different route to drug targets.

- Integrative genomics
- Gene expression & clinical traits data in Sage Commons
- Genome-Wide Association Studies (GWAS)
- Large-scale predictive network models of disease
- Co-expression and Bayesian (probabilistic graph) networks
- Complex data analysis pipelines

TCGA GBM Coexpression Network

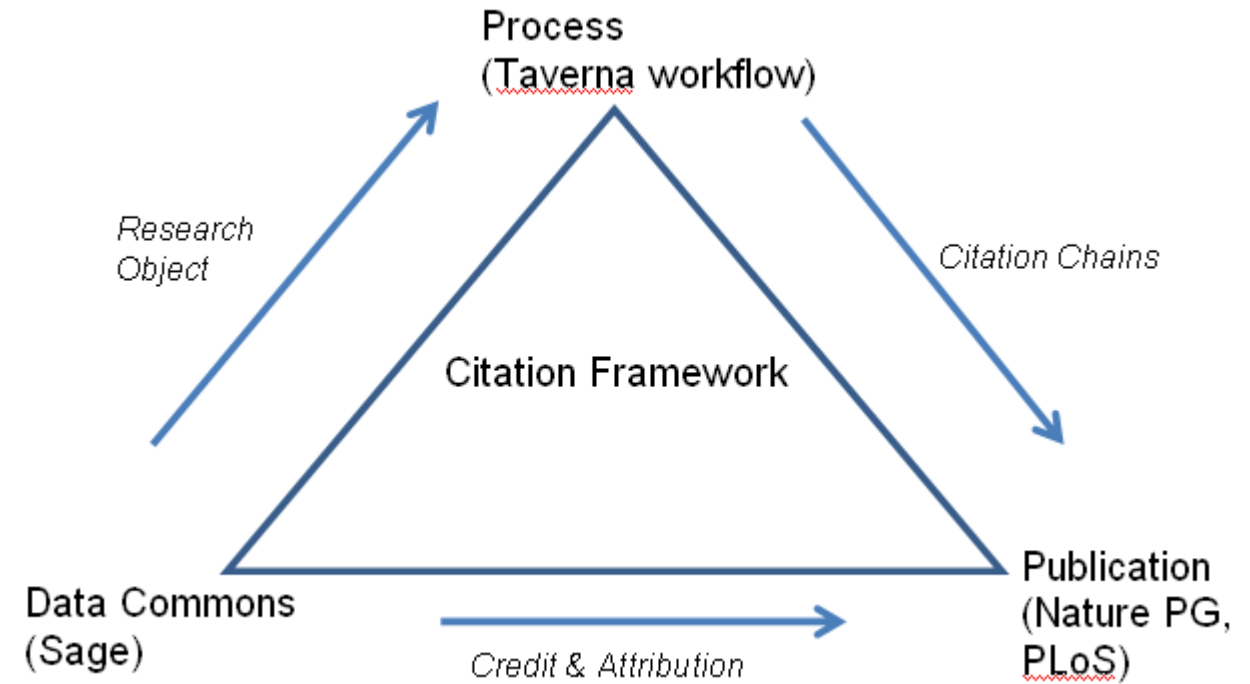


Large-scale predictive network models of disease



Functionality? How do we cite?

- Persistent identification - URIs
- Identifier-agnostic framework
- Resilient resolution service
- Multi-directional linking e.g. to peer-reviewed paper, to datasets
- Version control, provenance



The Open Provenance Model
Core Specification (v1.1)



Open Research
& Contributor ID



PLoS COMPUTATIONAL
BIOLOGY

An official journal of the International Society for Computational Biology

- **Infrastructure : seamless & cost-effective**
- **Open Methodology** : emerging Data Model
- **Researchers need help** with data management
- **Data Management Plans** : DCC DMP online tool
- **We need to incentivise data management**
- **Citation Framework** : assure credit & attribution



Thank you...



6th International Digital Curation Conference
"Participation & Practice: growing the Curation Community
through the data decade"

www.dcc.ac.uk

Chicago Mart Plaza, 6-8 December 2010



because good research needs good data

The Microsoft logo is centered on the page. It features the word "Microsoft" in a bold, italicized, sans-serif typeface. A registered trademark symbol (®) is positioned at the top right of the word. The background of the slide is white, with a blue decorative header at the top consisting of overlapping geometric shapes.

© 2010 Microsoft Corporation. All rights reserved. Microsoft, Windows, Windows Vista 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.

Microsoft® Research

Faculty Summit 2010