

*Citation for published version:* Patel, M 2010, 'Integrated Research Data Management in the Structural Sciences: Scaling up to integrated research data management' Paper presented at Scaling Up to Integrated Research Data Management, I2S2 Workshop, IDCC 2010, Chicago, 6/12/10, .

Publication date: 2010

**Document Version** Early version, also known as pre-print

Link to publication

Publisher Rights CC BY-SA

#### **University of Bath**

#### **General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.



### Infrastructure for Integration in Structural Sciences

### Manjula Patel

Scaling-up to Integrated Research Data Management Workshop 6<sup>th</sup> International Digital Curation Conference Holiday Inn, Mart Plaza Chicago, Illinois 6-8th December, 2010





This work is licensed under a Creative Commons Licence: Attribution-ShareAlike 3.0 http://creativecommons.org/licenses/by-sa/3.0/



- I2S2 Project overview and objectives
- Research Data & Infrastructure
- Requirements analysis
- A Scientific Research Activity Lifecycle Model
- An integrated information model
- Use cases
- Cost-Benefits Analysis

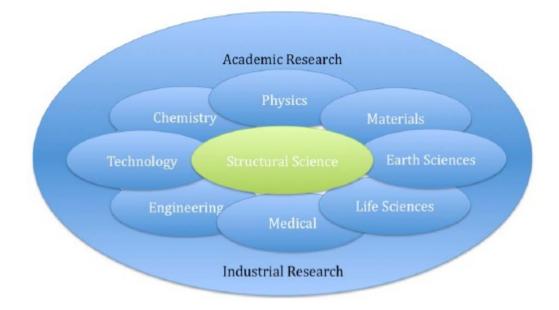


Diamond Light Source (DLS), Science & Technology Facilities Council, UK



### **I2S2 Project Overview**

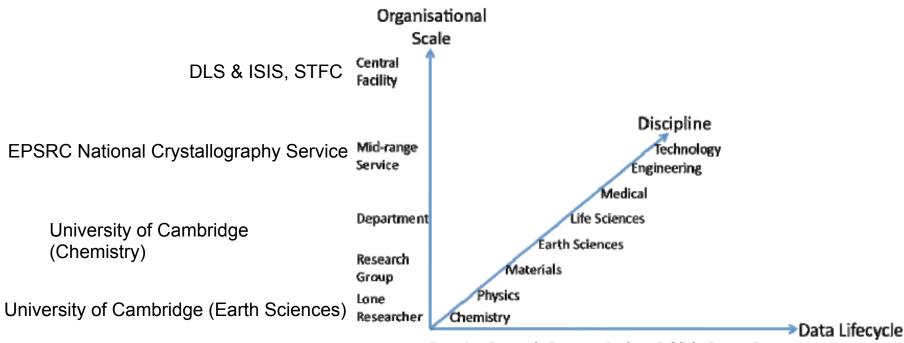
- Understand and identify requirements for a data-driven research infrastructure in the Structural Sciences
  - Examine localised data management practices
  - Investigate data management infrastructure in large centralised facilities
- Show how effective cross-institutional research data management can increase efficiency and improve the quality of research





Infrastructure for Integration in Structural Sciences

Scale and complexity: small laboratory to institutional installation to large scale facilities e.g. DLS & ISIS, STFC Interdisciplinary issues: research across domain boundaries Data lifecycle: data flows and data transformations over time



Conceive Research Propose Analyse Publish Curate Reuse



### **Research Data & Infrastructure**

- Research Data includes (all information relating to an experiment):
  - raw, reduced, derived and results data
  - research and experiment proposals
  - results of the peer-review process
  - laboratory notebooks
  - equipment configuration and calibration data
  - wikis and blogs
  - metadata (context, provenance etc.)
  - documentation for interpretation and understanding (semantics)
  - administrative and safety data
  - processing software and control parameters
- Infrastructure includes physical, technical, informational and human resources essential for researchers to undertake high-quality research:
  - Tools, Instrumentation, Computer systems and platforms, Software, Communication networks
  - Documentation and metadata
  - Technical support (both human and automated)
- Effective validation, reuse and repurposing of data requires
  - Trust and a thorough understanding of the data
  - Transparent contextual and provenance information detailing how the data were generated, processed, analysed and managed



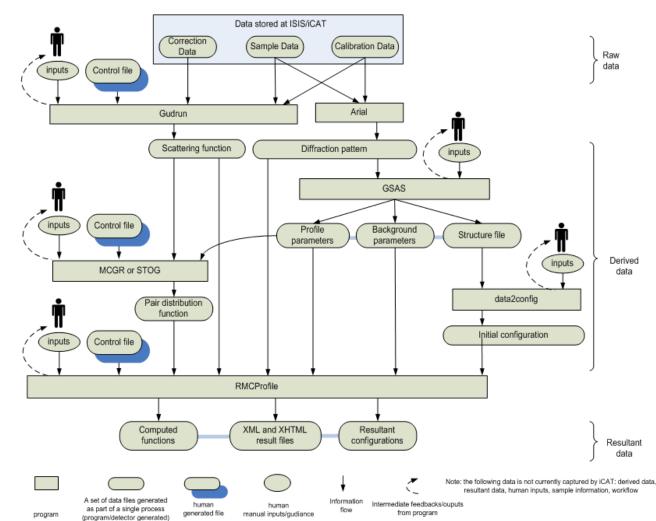
# Earth Sciences, Cambridge

- Construct large scale atomic models of matter that best match experimental data; using Reverse Monte-Carlo Simulation techniques
- Experiment and data collection conducted at ISIS Neutron Source (GEM)
- Little or no shared infrastructure
  - Data sharing with colleagues via email, ftp, memory stick etc.
  - Data received from ISIS is currently stored on laptops or WebDAV server
- Management of intermediate, derived and results data a major issue
  - Data managed by individual researcher on own laptop
  - No departmental or central institutional facility
- Data management needs largely so that
  - Data can be shared internally
  - A researcher (or another team member) can return to and validate results in the future
  - External collaborators can access and use the data
- Any changes should be embedded into scientist's workflow and be nonintrusive



### Earth Sciences, Cambridge: Typical workflow

Infrastructure for Integration in Structural Sciences

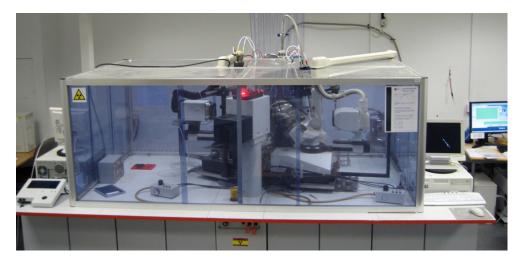


Martin Dove & Erica Yang, July 2010



# Chemistry, Cambridge

- Implementation and enhancement of a pilot repository for crystallography data underway (CLARION Project)
- Need for IPR, embargo and access control to facilitate the controlled release of scientific research data
- Information in laboratory notebooks need to be shared (ELN)
- Importance of data formats and encodings (RDF, CML) to maximise potential for data reuse and repurposing



EPSRC National Crystallography Service, University of Southampton, UK

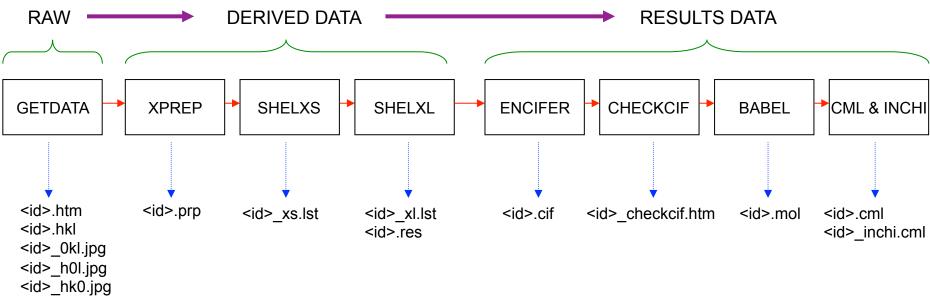


# **EPSRC NCS**, Southampton

- Service provision function (operates nationally across institutions)
  - Local x-ray diffraction instruments + use of DLS (beamline 119)
  - Retain experiment data
  - Maintain administrative data
- Raw data generated in-house is stored at ATLAS Data Store (STFC)
- Local institutional repository (eCrystals) for intermediate, derived and results data
  - Metadata application profile
  - Public and private parts (embargo system)
  - Digital Object Identifier, InChi
- Experiments conducted and data collected by NCS scientists either in-house or at DLS
- Labour-intensive paper-based administration and records-keeping
  - Paper-based system for scheduling experiments
  - Paper copies of Experiment Risk Assessment (ERA) get annotated by scientist and photocopied several times
  - Several identifiers per sample
- Administrative functions require streamlining between NCS and DLS
  - e.g. standardisation of ERA forms, identifiers



### **EPSRC NCS:** typical workflow



- <id>\_\_\_\_\_\_rig
- Initialisation: mount new sample
- Collection: collect data
- Processing: process and correct images
- Solution: solve structures
- Refinement: refine structure

- CIF: produce Crystallographic Information File
- Validation: chemical & crystallographic checks
- Report: generate Crystal Structure Report
- CML, INChl



### DLS & ISIS, STFC

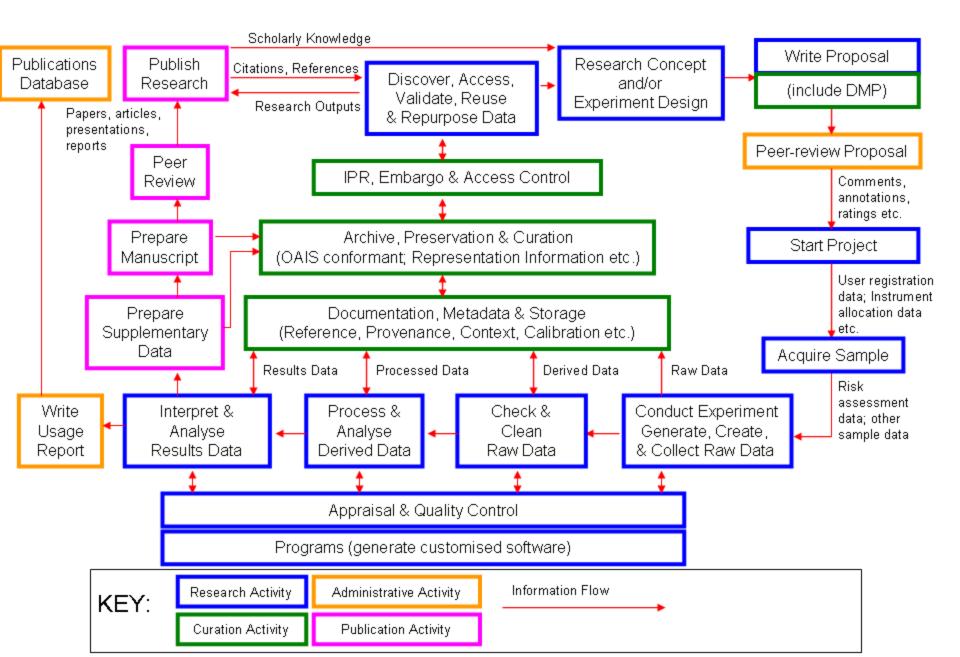
- Operate on behalf of multiple institutions and communities
- Scientific (peer) and technical review of proposals for beam time allocation
- User offices manage administrative and safety information
- Service function implies an obligation to retain raw data
- Large infrastructure, engineered to manage raw data
  - Designed to describe facilities based experiments in Structural Science
     e.g. ISIS Neutron Source, Diamond Light Source.
  - ICAT implementation of Core Scientific Metadata Model (CSMD)
- No storage or management of derived and results data
  - Derived data taken off site on laptops, removable drives etc.
  - Results data independently worked up by individual researchers
- Experiment/Sample identifiers based on beam line number



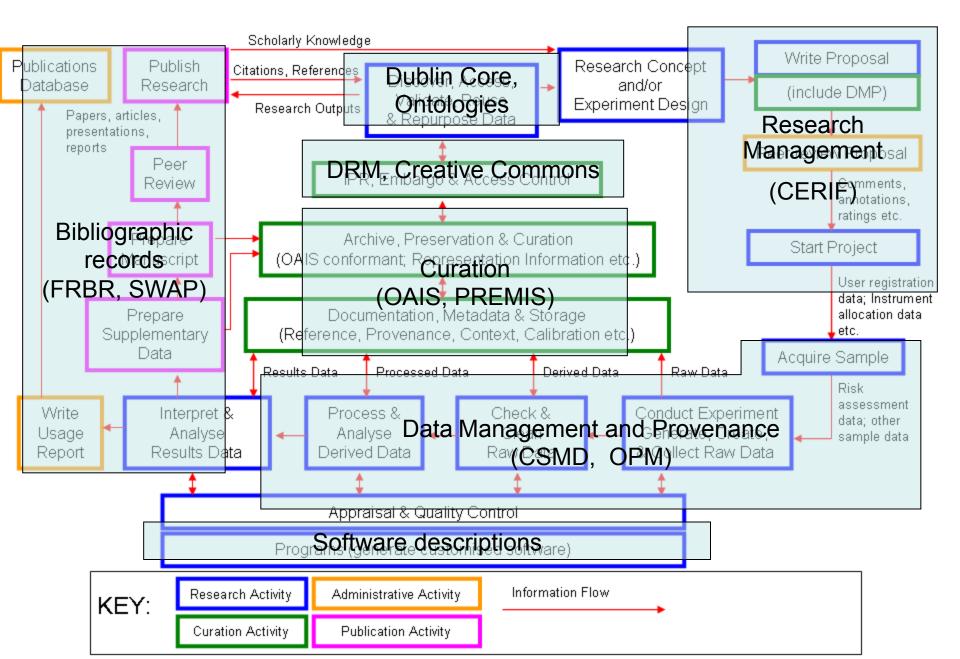
### **Generalised Requirements**

- Basic requirement for data storage and backup facilities to sophisticated needs such as structuring and linking together of data
- Contextual information is not routinely captured
- The actual workflow or processing pipeline is not recorded
- Processing pipeline is dependent on a suite of software
- Adequate metadata and contextual information to support:
  - Maintenance and management
  - Linking together of all data associated with an experiment
  - Referencing and citation
  - Authenticity
  - Integrity
  - Provenance
  - Discovery, search and retrieval
  - Curation and preservation
  - IPR, embargo and access management
  - Interoperability and data exchange
- Simplification of inter-organisational communications and tracking, referencing and citation of datasets
  - Standardised ERA forms
  - Unique persistent identifiers
- Solutions should be as non-intrusive as possible

### An Idealised Scientific Research Activity Lifecycle Model



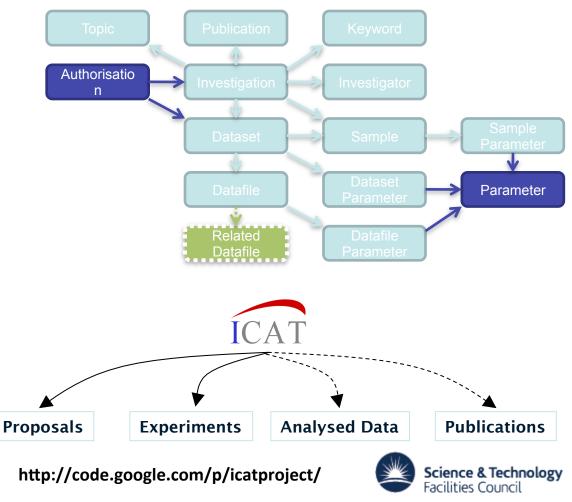
### An Idealised Scientific Research Activity Lifecycle Model





### **Core Scientific Metadata Model**

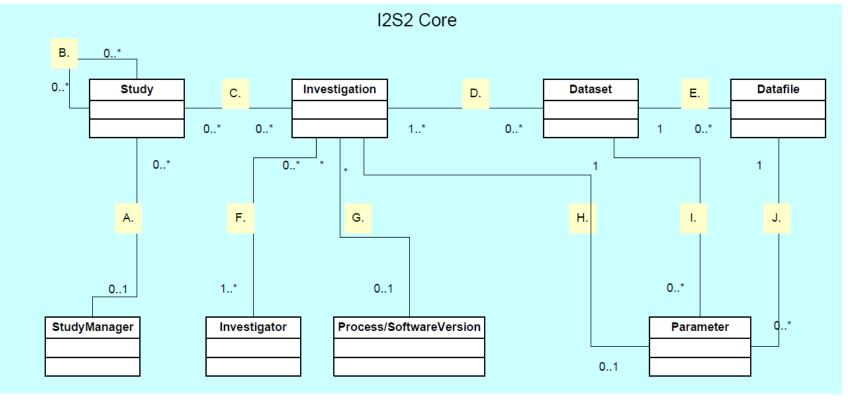
- Designed to describe facilities based
   experiments in Structural Science
- CSMD is the basis of I2S2 integrated information model
- Forms a basis for extension to:
  - Laboratory based science
  - Derived data
  - Secondary analysis data
  - Preservation information
  - Publication data
- Aim to cover the scientist's research lifecycle as well as facilities data





Infrastructure for Integration in Structural Sciences

- Removal of facility specific information
- A simple model to describe datasets



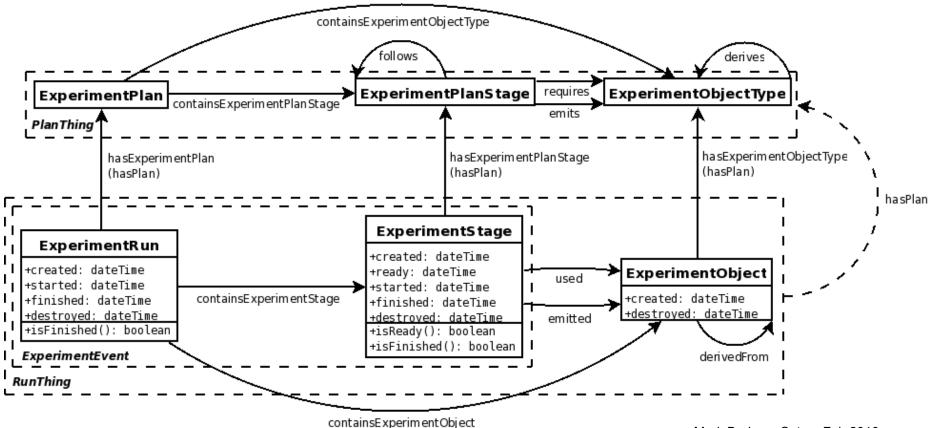
Erica Yang, STFC, 2010



### oreChem Model

Infrastructure for Integration in Structural Sciences

- An abstract model for planning and enacting chemistry experiments
- Enables exact replication of methodology in a machine-readable form
- Allows rigorous verification of reported results

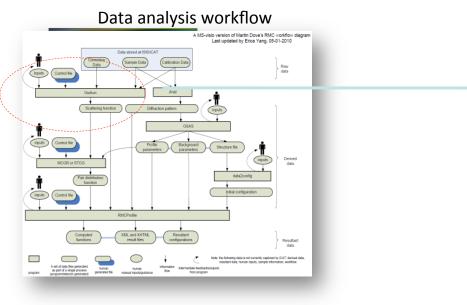


Mark Borkum, Soton, Feb 2010



### **I2S2 Integrated Information Model**

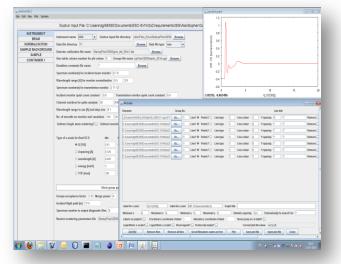
- I2S2-IM = CSMD-Core + oreChem Model
- Use oreChem to describe planning and enactment of scientific process
- Use CSMD to describe the data-sets from the experiment
- I2S2-IM being implemented at STFC in the form of ICAT-Lite
  - A personal workbench for managing data flows
  - Allows the user to commit data
  - Enables capture of provenance information

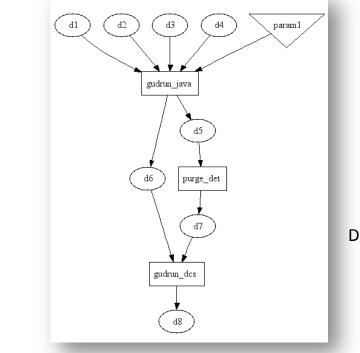


#### Data analysis folders

<ul> <li>Include in library</li> <li>Share</li> </ul>	with • B	urn New folder			
🔹 dist		ame	Date modified	Туре	Size
🎍 1252-Demo		SLS39534.mut01	26/01/2010 11:19	MUT01 File	1 KB
nbproject		\$L\$39534.rat	18/08/2010 10:49	Rating System File	1 KB
🥩 src		SLS39534.rawmon	26/01/2010 11:19	RAWMON File	39 KB
🏂 test		SLS39534.rawtrans	26/01/2010 11:19	RAWTRANS File	39 KB
🤹 .svn		SLS39534.somrat	26/01/2010 11:19	SAMRAT File	31 KB
💰 GudrunFilles_Erica		SLS39534.smomon	26/01/2010 11:19	SMOMON File	39 KB
Jun .svn		SLS39534.trans01	26/01/2010 11:19	TRANS01 File	2 KB
💰 Gudrun GUL2		SLS39542.bad	26/01/2010 11:16	BAD File	12 KB
🛃 RawData	- III - 4	SLS39542.grp	18/06/2010 10:49	Microsoft Progra	5 KB
🥧 run		SLS39542.rat	18/06/2010 10:49	Rating System File	1 KB
avn		SLS39621.bad	09/06/2010 11:42	BAD File	0 KB
SANDALS	1 4	SLS39621.grp	09/06/2010 11:42	Microsoft Progra	5 KB
svn		SLS39621.rat	09/06/2010 11:42	Rating System File	1 KB
🕺 Water	- 6	SLS39629.mul01	26/01/2010 11:19	MUL01 File	5 KB
StartupFiles		SLS39629.mut01	26/01/2010 11:19	MUT01 File	1 KB
🤹 icat		SLS39629.pla01	26/01/2010 11:19	PLA01 File	7 KB
🛃 testFolder		SLS39629.rawmon	26/01/2010 11:19	RAWMON File	39 KB
🛃 xml-resources		SLS39629.rawtrans	26/01/2010 11:19	RAWTRANS File	39 KB
ICAT_LITE_WS		SLS39629.smomon	26/01/2010 11:19	SMOMON File	39 KB
Maven		SLS39629.trans01	26/01/2010 11:19	TRANS01 File	2 KB
myGWTApp1		SLS39629.vanrat	26/01/2010 11:19	VANRAT File	31 KB
💰 schema		SLS39630.bad	26/01/2010 11:16	BAD File	24 KB
💋 testfiles		SLS39630.grp	18/06/2010 10:49	Microsoft Progra	5 KB
CostBenefitAnalysis		SLS39630.rat	18/06/2010 10:49	Rating System File	1 KB
DataManagementTools		SLS39630.rawmon	26/01/2010 11:19	RAWMON File	39 KB
Day2DayProgress		SL\$39630.rawtrans	26/01/2010 11:19	<b>RAWTRANS File</b>	39 KB
all deliverables		SLS39630.smomon	26/01/2010 11:19	SMOMON File	39 KB
a graphviz		SLS39631.abs01	26/01/2010 11:19	ABS01 File	7 KB
GridMP		SLS39631.bad	26/01/2010 11:16	BAD File	12 KB
gridmp-isisSVN		SLS39631.bak	26/01/2010 11:19	BAK File	45 KB
I252-actitivity-model		SLS39631.chksum	26/01/2010 11:19	CHKSUM File	31 KB
🗼 icatgoogle		SLS39631.d cd01	26/01/2010 11:19	DCS01 File	548 KB
information-model		SLS39631.grl	26/01/2010 11:19	GR1 File	45 KB
💰 meetings		SLS39631.gr2	26/01/2010 11:19	GR2 File	45 KB
💰 metadata		SLS39631.grp	26/01/2010 11:16	Microsoft Progra	4 KB
Notes		SLS39631.gud	26/01/2010 11:19	GUD File	3 KB
💰 proposal		\$L\$39631.mgor01	26/01/2010 11:19	MGOR01 File	45 KB
🤹 references		SLS39631.mint01	26/01/2010 11:19	MINT01 File	48 KB
o requirements		SLS39631.mul01	26/01/2010 11:19	MUL01 File	5 KB
o talks		SLS39631.mut01	26/01/2010 11:19	MUT01 File	1 KB
82 items					

#### Scientific software: Gudrun



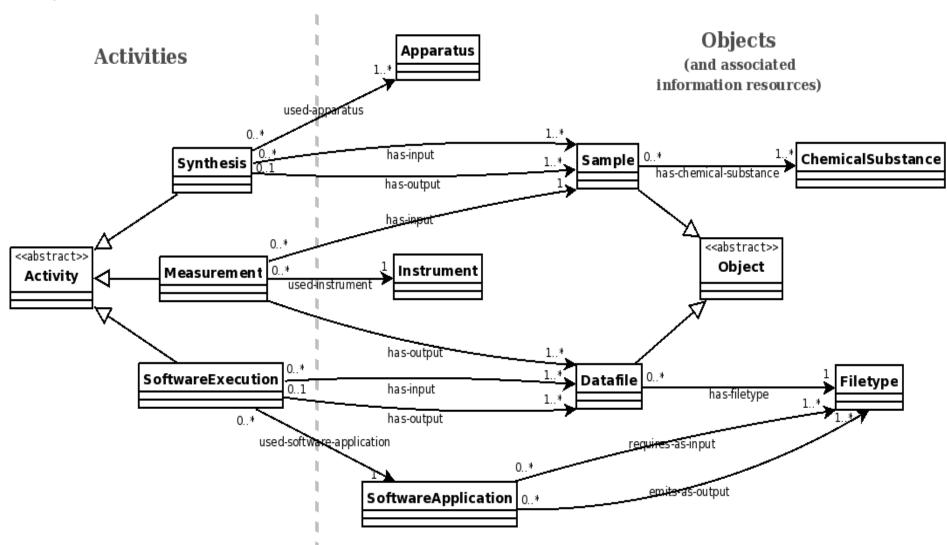


Archive Browse Restore

**Derived** Data









# Testing the I2S2-IM

Infrastructure for Integration in Structural Sciences

### Case study 1: Scale and Complexity

- Data management issues spanning organisational boundaries in Chemistry
- Interactions between a lone worker or research group, the EPSRC NCS and DLS
- Traversing administrative boundaries between institutions and experiment service facilities
- Aim to probe both cross-institutional and scale issues

#### Case Study 2: Inter-disciplinary issues

- Collaborative group of inter-disciplinary scientists (university and central facility researchers) from both Chemistry and Earth Sciences
- Use of ISIS neutron facility and subsequent modelling of structures based on raw data
- Identification of infrastructural components and workflow modelling
- Aim to explore the role of XML for data representation to support easier sharing of information content of derived data



# **Cost-Benefits Analysis**

- A before and after cost-benefit analysis using the Keeping Research Data Safe model
- Extending the KRDS Model
  - Focus has been on extensions and elaboration of activities in the research (KRDS "pre-Archive") phase
- Metrics and assigning costs
  - Identification of activities in research activity lifecycle model that will represent significant cost savings or benefits
  - Work to identify non-cost benefits and possible metrics
- 2 use case studies
  - Quantitative -cost-benefits in terms of service efficiencies (NCS)
  - Qualitative -researcher benefits (improvement in tools; ease of making data accessible)



Integration in Structural Sciences

### Conclusions

- Considerable variation in requirements between differing scales of science
- At present individual researcher, group, department, institution, facilities all working within their own frameworks
- Merit in adopting an integrated approach which caters for all scales of science:
  - Aggregation and/or cross-searching of related datasets
  - Efficient exchange, reuse and repurposing of data across disciplinary boundaries
  - Data mining to identify patterns or trends
- I2S2 Integrated information model aims to:
  - Support the scientific research activity lifecycle model
  - Capture processes and provenance information
  - Interoperate with and complement existing models and frameworks
- Before and after cost-benefits analysis to assess impact



- Liz Lyon (UKOLN, University of Bath & Digital Curation Centre)
- Manjula Patel (UKOLN, University of Bath & Digital Curation Centre)
- Simon Coles (EPSRC National Crystallography Centre, University of Southampton)
- Brian Matthews (Science & Technology Facilities Council)
- Erica Yang (Science & Technology Facilities Council)
- Martin Dove (Earth Sciences, University of Cambridge)
- Peter Murray-Rust (Chemistry, University of Cambridge)
- Neil Beagrie (Charles Beagrie Ltd.)

<u>m.patel@ukoln.ac.uk</u> http://www.ukoln.ac.uk/projects/I2S2/