Citation for published version:

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.
Manjula Patel
Scaling-up to Integrated Research Data Management Workshop
6th International Digital Curation Conference
Holiday Inn, Mart Plaza
Chicago, Illinois
6-8th December, 2010
Outline

• 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
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
Objectives

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
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
• 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 non-intrusive
Earth Sciences, Cambridge: Typical workflow

[Diagram showing a typical workflow in Earth Sciences]

Note: The following data is not currently captured by ICAT: derived data, resultant data, human inputs, sample information, workflow.
• 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
• **Service provision function** (operates nationally across institutions)
  – Local x-ray diffraction instruments + use of DLS (beamline I19)
  – 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

- **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
- **CML, INChI**
• 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

- **Publications Database**
  - Papers, articles, presentations, reports

- **Publish Research**
  - Citations, References
  - Research Outputs

- **Discover, Access, Validate, Reuse & Repurpose Data**

- **Research Concept and/or Experiment Design**

- **Write Proposal**
  - (include DMP)

- **Peer-review Proposal**
  - Comments, annotations, ratings etc.

- **Start Project**
  - User registration data; Instrument allocation data etc.
  - Risk assessment data; other sample data

- **Acquire Sample**

- **IPR, Embargo & Access Control**

- **Archive, Preservation & Curation**
  - (OAIS conformant; Representation Information etc.)

- **Documentation, Metadata & Storage**
  - (Reference, Provenance, Context, Calibration etc.)

- **Prepare Manuscript**
  - Prepare Supplementary Data

- **Prepare Manuscript**

- **Write Usage Report**

- **Interpret & Analyse Results Data**

- **Process & Analyse Derived Data**

- **Check & Clean Raw Data**

- **Conduct Experiment Generate, Create, & Collect Raw Data**

- **Appraisal & Quality Control**

- **Programs (generate customised software)**

**KEY:**
- Research Activity
- Administrative Activity
- Curation Activity
- Publication Activity
- Information Flow
An Idealised Scientific Research Activity Lifecycle Model

- **Bibliographic records (FRBR, SWAP)**
  - Prepare Supplementary Data
  - Peer Review
  - Write Usage Report

- **Data Management and Provenance (CSMD, OPM)**
  - Collect Raw Data
  - Process & Analyse Derived Data
  - Check & Analyse Raw Data
  - Conduct Experiment (Reference, Provenance, Context, Calibration etc.)

- **Curation (OAIS, PREMIS)**
  - Archive, Preservation & Curation (OAIS conformant; Representation Information etc.)
  - Documentation, Metadata & Storage (Reference, Provenance, Context, Calibration etc.)

- **Dublin Core, Ontologies**
  - DRM, Creative Commons
  - IPR, Embargo & Access Control

- **Research Concept and/or Experiment Design**
- **Write Proposal (include DMP)**
- **Start Project**
  - User registration data; Instrument allocation data etc.

- **Research Management (CERIF)**
  - Comments, annotations, ratings etc.

**Software descriptions**
- Programs (generate customised software)

**KEY:**
- Research Activity
- Administrative Activity
- Curation Activity
- Publication Activity
- Information Flow

- Scholarly Knowledge
- Citations, References
- Research Outputs

- Publications Database
- Publish Research
- Peer Review
- Publish Supplemental Data
- Interpret & Analyse Results Data

- **Publishers**
- Papers, articles, presentations, reports

- **Software**
- Bibliographic records (FRBR, SWAP)
- Data Management and Provenance (CSMD, OPM)
- Curation (OAIS, PREMIS)
- Dublin Core, Ontologies
- DRM, Creative Commons
- Research Concept and/or Experiment Design
- Write Proposal (include DMP)
- Research Management (CERIF)
- Start Project
- Acquire Sample
- Collect Raw Data
- Process & Analyse Derived Data
- Conduct Experiment (Reference, Provenance, Context, Calibration etc.)
- Check & Analyse Raw Data
- Archive, Preservation & Curation (OAIS conformant; Representation Information etc.)
- Documentation, Metadata & Storage (Reference, Provenance, Context, Calibration etc.)
- Dublin Core, Ontologies
- DRM, Creative Commons
- IPR, Embargo & Access Control

- **Keywords**
- Bibliographic records (FRBR, SWAP)
- Data Management and Provenance (CSMD, OPM)
- Curation (OAIS, PREMIS)
- Dublin Core, Ontologies
- DRM, Creative Commons
- Research Concept and/or Experiment Design
- Write Proposal (include DMP)
- Research Management (CERIF)
- Start Project
- Acquire Sample
- Collect Raw Data
- Process & Analyse Derived Data
- Conduct Experiment (Reference, Provenance, Context, Calibration etc.)
- Check & Analyse Raw Data
- Archive, Preservation & Curation (OAIS conformant; Representation Information etc.)
- Documentation, Metadata & Storage (Reference, Provenance, Context, Calibration etc.)
- Dublin Core, Ontologies
- DRM, Creative Commons
- IPR, Embargo & Access Control
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
CSMD-Core

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

- 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
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
Testing the I2S2-IM

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)
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
Project Team

- 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.)

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