Infrastructure for Integration in Structural Sciences

Scaling Up Scientific Data Management Infrastructure

Erica Yang, Manjula Patel, Brian Matthews

UK e-Sceince All Hands Meeting
Cardiff, Wales
13-16th September 2010

Infrastructure for
Integration in Structural Sciences

Objectives

- Identify requirements for a data-driven research infrastructure
  - Understand localised data management practices
  - Understand data management infrastructure in large centralised facilities
- Examine 3 complementary infrastructure axes:

  Scale and complexity: small laboratory to institutional Installations to large scale facilities e.g. DLS & ISIS, STFC

  Interdisciplinary issues: research across domain boundaries

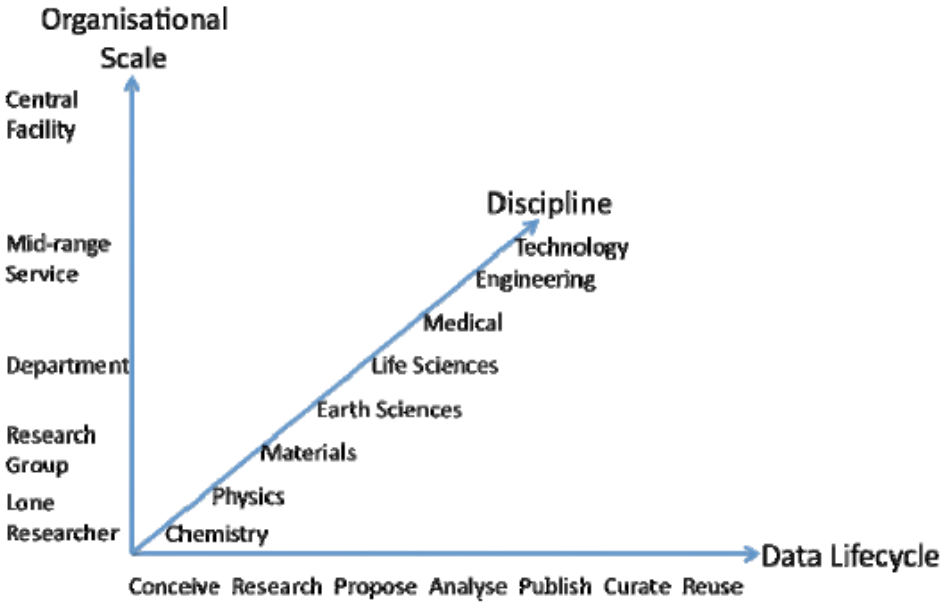
  Data lifecycle: data flows and data transformations over time

DLS & ISIS, STFC

EPSRC National Crystallography Service

University of Cambridge (Chemistry)

University of Cambridge (Earth Sciences)



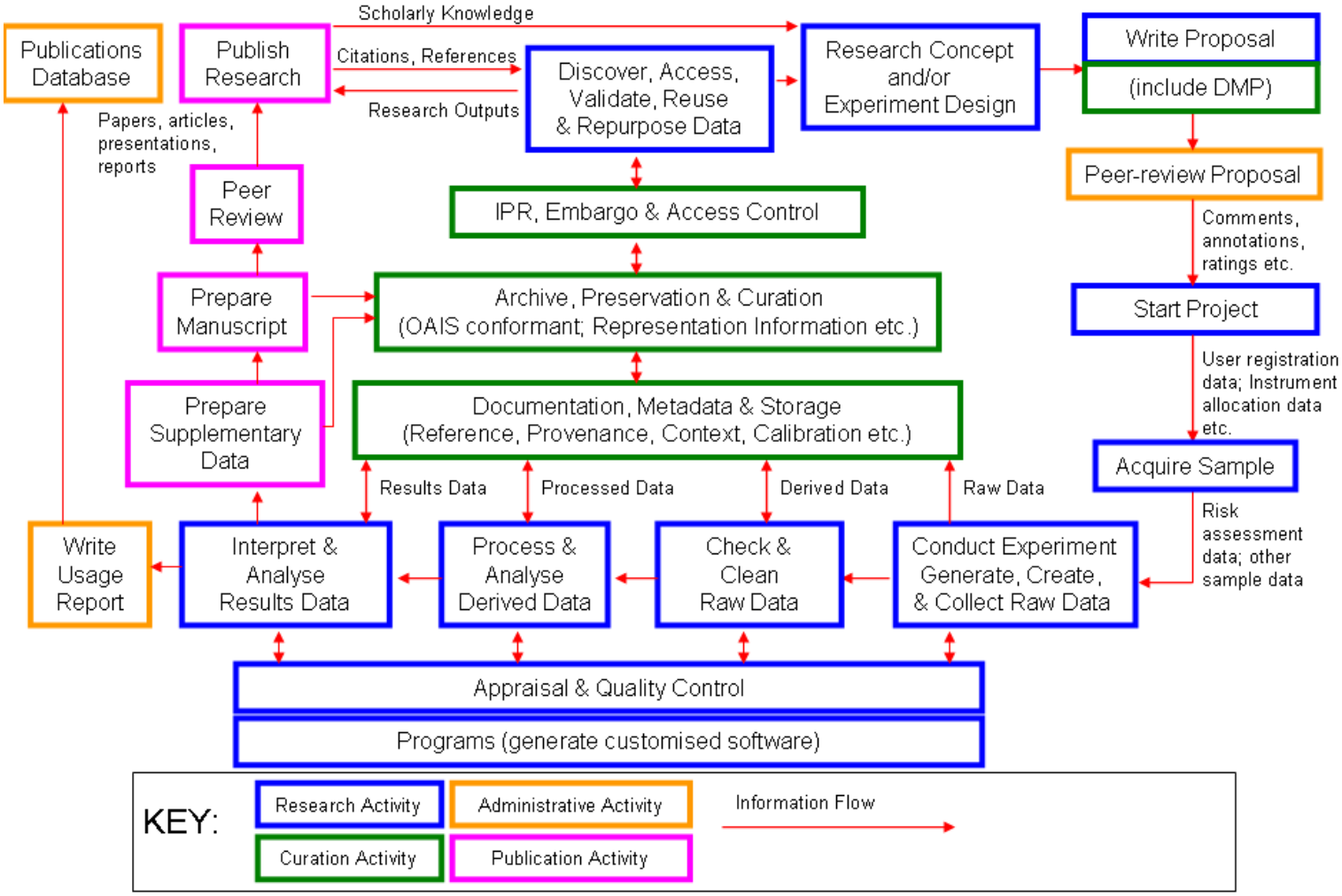
I2S2
Infrastructure for
Integration in Structural Sciences

Generalised Requirements

- Basic requirement for data storage and backup facilities to sophisticated needs such as structuring and linking together of data
- 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

I2S2
Infrastructure for
Integration in Structural Sciences

An idealised scientific research data lifecycle model



Infrastructure for
Integration in Structural Sciences

Use Case Studies

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 (at STFC) and subsequent modelling of structures based on raw data
- Identification of infrastructural components and workflow modelling
- Aim to explore role of XML for data representation to support easier sharing of information content of derived data

Infrastructure for
Integration in Structural Sciences

Pilot Implementation 1

Scale and Complexity based on Use Case 1

- Involving: Cambridge Chemistry, NCS and DLS
- Centred around structural science support for the bench chemist
- Scenario
  - Cambridge organic synthesis PhD student generates new compound and crystallises.

  CLARION ELN
  - Student submits sample to local crystallographic service

  LOCAL SUBMISSION PROCESS (PAPER FORMS?)
  - Exploratory experiment performed – limited results obtained (unit cell and partial data collection)

  LOCAL LABORATORY INSTRUMENT AND DATA WORKUP SYSTEMS. ARCHIVAL
  - Decision to refer to NCS – undergo application / submission process

  ONLINE APPLICATION & SUBMISSION
  - Receipt by NCS – data collection performed

  ALERTING SERVICE, LOCAL DATA ACQUISITION & WORKUP, ONLINE AVAILABILITY & ARCHIVAL
  - Data not sufficient quality for publishable result – refer to DLS

  REFERRAL SYSTEM
  - Application, scheduling and receipt by DLS

  PROPOSAL, EXPERIMENTAL RISK ASSESSMENT, TRANSPORTATION
  - Beamtime – data collected

  LOCAL DATA COLLECTION, AVAILABILITY & ARCHIVAL
  - Result worked up, NCS status change, results conveyed to User, sample returned to NCS and then User.

  LOCAL DATA WORKUP, ONLINE ALERTING & AVAILABILITY, ARCHIVAL

Infrastructure for
Integration in Structural Sciences

Pilot Implementation 2

Interdisciplinary issues based on Use Case 2

- Involving: Cambridge Earth Sciences and ISIS
- Explore the use of XML for data representation at all stages in the workflow, particularly to ensure proper data interoperability
- Examine the possibility for automatic metadata collection at each stage
- Assess whether approach may be duplicated for other work processes
- Evaluate whether it is possible to make available all the derived data
- Ensure that innovations lead to changes that are as non-intrusive as possible for the researcher.
- Scenario
  - A powder diffraction experiment on the GEM diffractometer (ISIS facility) to measure "total scattering“
  - Analysis carried out using tools developed in collaboration between Cambridge and ISIS
  - Raw data sets, calibration and background correction data are collected and archived at ISIS
  - A series of complex processing workflows generate a derived dataset with potentially important new publishable information on the crystal structure
  - Transform CML files into XHTML representations that capture and display all key information
  - Investigate automation for simulation and/or computational analysis of data