I2S2

Infrastructure for Integration in Structural Sciences

Manjula Patel
JISC MRD Progress Meeting
Manchester Conference Centre
17-18th May 2010
http://www.ukoln.ac.uk/projects/I2S2/

I2S2

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

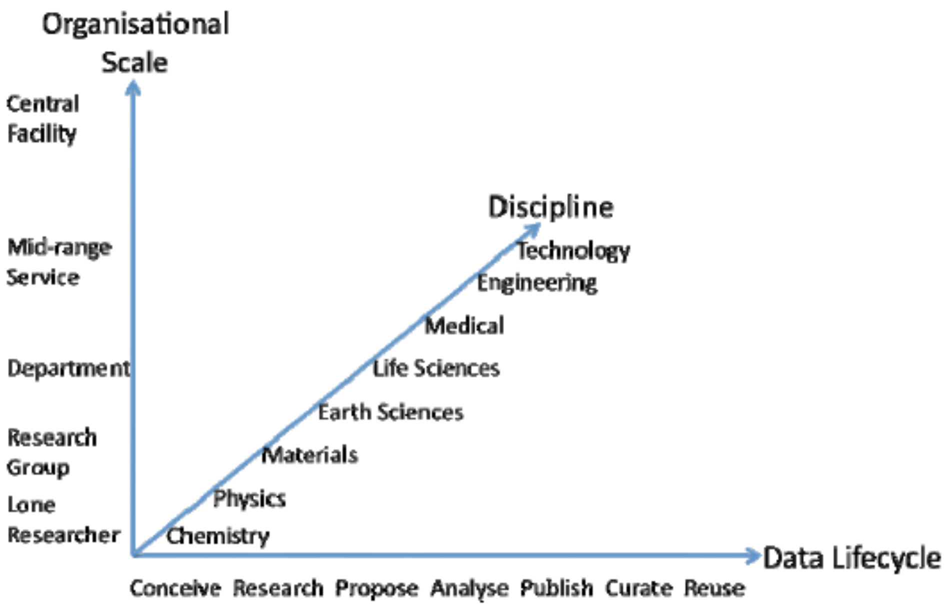
  Data lifecycle: data flows and data transformations over time

DLS & ISIS, STFC

EPSRC National Crystallography Service

University of Cambridge (Chemistry)

University of Cambridge (Earth Sciences)



Infrastructure for
Integration in Structural Sciences

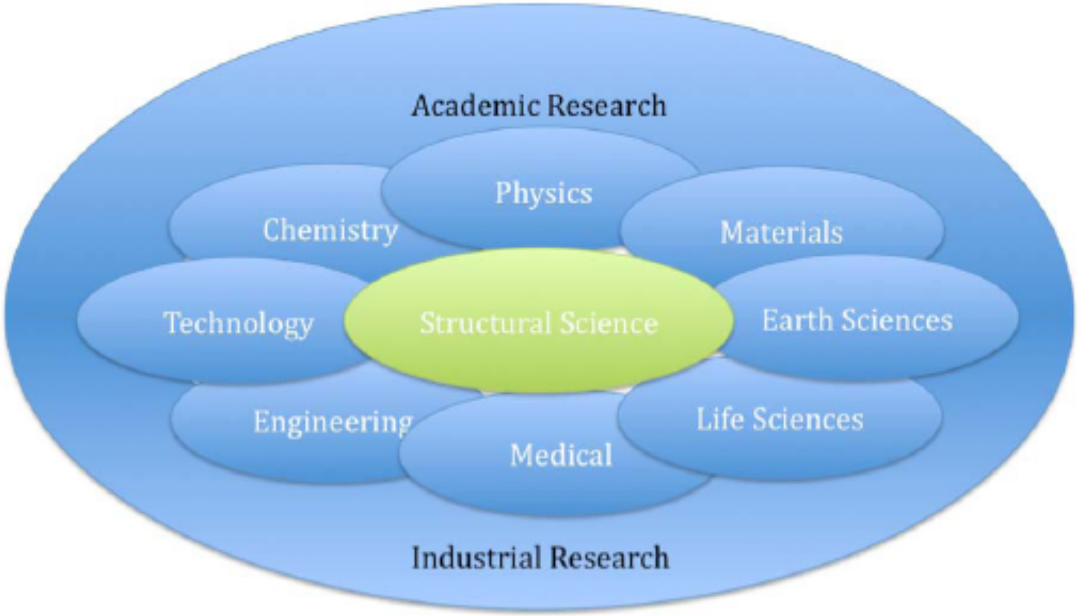
Research Infrastructure

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)

Progress Outline

- Requirements Gathering
- Use Case Studies & Pilot Implementations
- Integrated Information Model
- Cost/Benefits Analysis

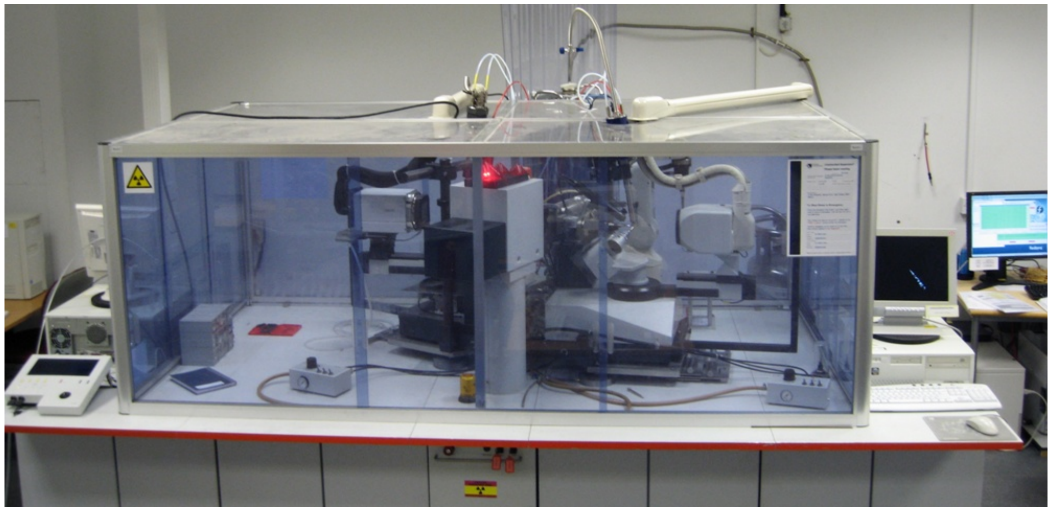


I2S2
Infrastructure for
Integration in Structural Sciences

Requirements Gathering

Methodology:

- Desk Study
- Data management planning tools
- Immersive Studies
- Gap Analysis



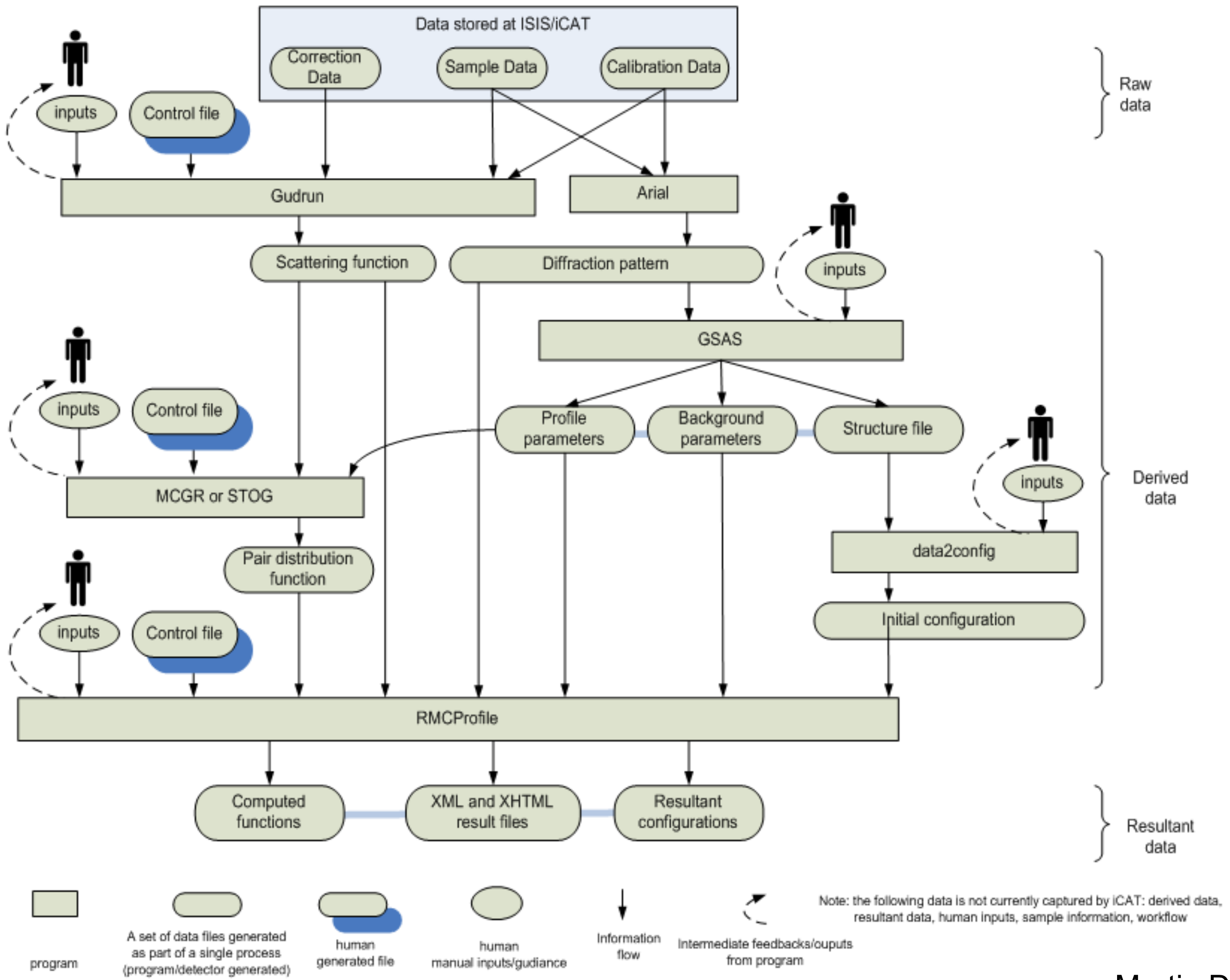
Infrastructure for Integration in Structural Sciences

Mini Immersive Studies

- Focusing on interface between local laboratories and large-scale facilities:
  - Visit Simon Coles @ NCS, 17th Nov 2009
  - Visit Martin Dove @ Cambridge Earth Sciences, 24th Nov 2009
  - Visit Martin Dove @ ISIS, 7th & 14th Dec 2009 (excluding ISIS User Office)
  - Visit Simon Coles @ DLS, 15th Jan 2010 (including DLS User Office)
  - Visit Peter Murray-Rust @ Cambridge Chemistry, 4th Mar 2010
- Critical to developing an effective data management infrastructure is a thorough understanding of
  - data themselves
  - workflows and processes involved in generating and processing data
  - file formats in use
  - inter-relationships between processing software and data files
- Processes and workflows in each scientific laboratory differ considerably

Infrastructure for
Integration in Structural Sciences

Earth Sciences: typical workflow



Martin Dove & Erica Yang

Infrastructure for Integration in Structural Sciences

Earth Sciences: some requirements …

- 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
- Need department or research group level data storage and management infrastructure to capture, manage and maintain:
  - Metadata and contextual information (including provenance);
  - Control files and parameters;
  - Processing software;
  - Workflow for a particular analysis;
  - Derived and results data;
  - Links between all the datasets relating to a specific experiment or analysis
- Any changes should be embedded into scientist's workflow and be non-intrusive

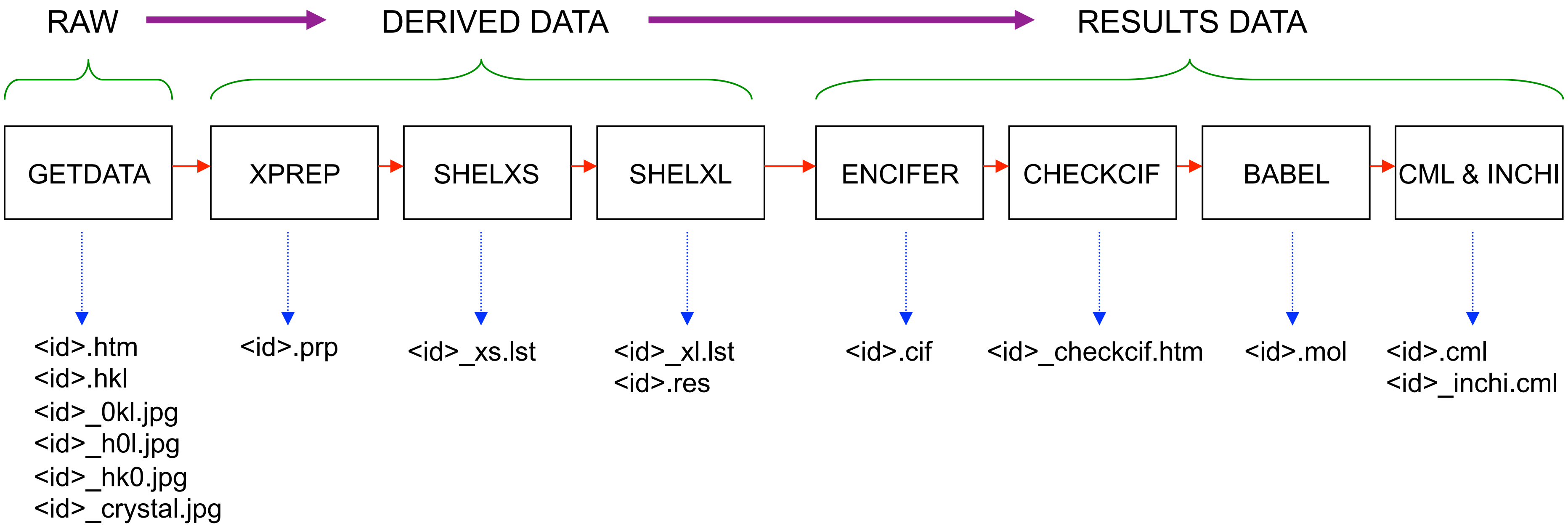
Chemistry: some requirements …

Infrastructure for Integration in Structural Sciences

- Implementation and enhancement of a repository for crystallography data underway (CLARION Project)
  - will require additional effort to convert into a robust service level infrastructure
- 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

Infrastructure for
Integration in Structural Sciences

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
- Report: generate Crystal Structure Report
- CML, INChI

EPSRC NCS: some requirements …

- Service function implies an obligation to:
  - Retain experiment data
  - Maintain administrative and safety data
  - Transfer data to end-researcher
- eCrystals repository (may need further development)
  - Metadata application profile
  - Public and private parts (embargo system)
  - Digital Object Identifier, InChi
- 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 (researcher assigned; researcher institution assigned, NCS assigned, DLS assigned)
- Administrative functions require streamlining between NCS and DLS
  - e.g. standardisation of ERA forms, identifiers

Infrastructure for
Integration in Structural Sciences

DLS & ISIS: some observations …

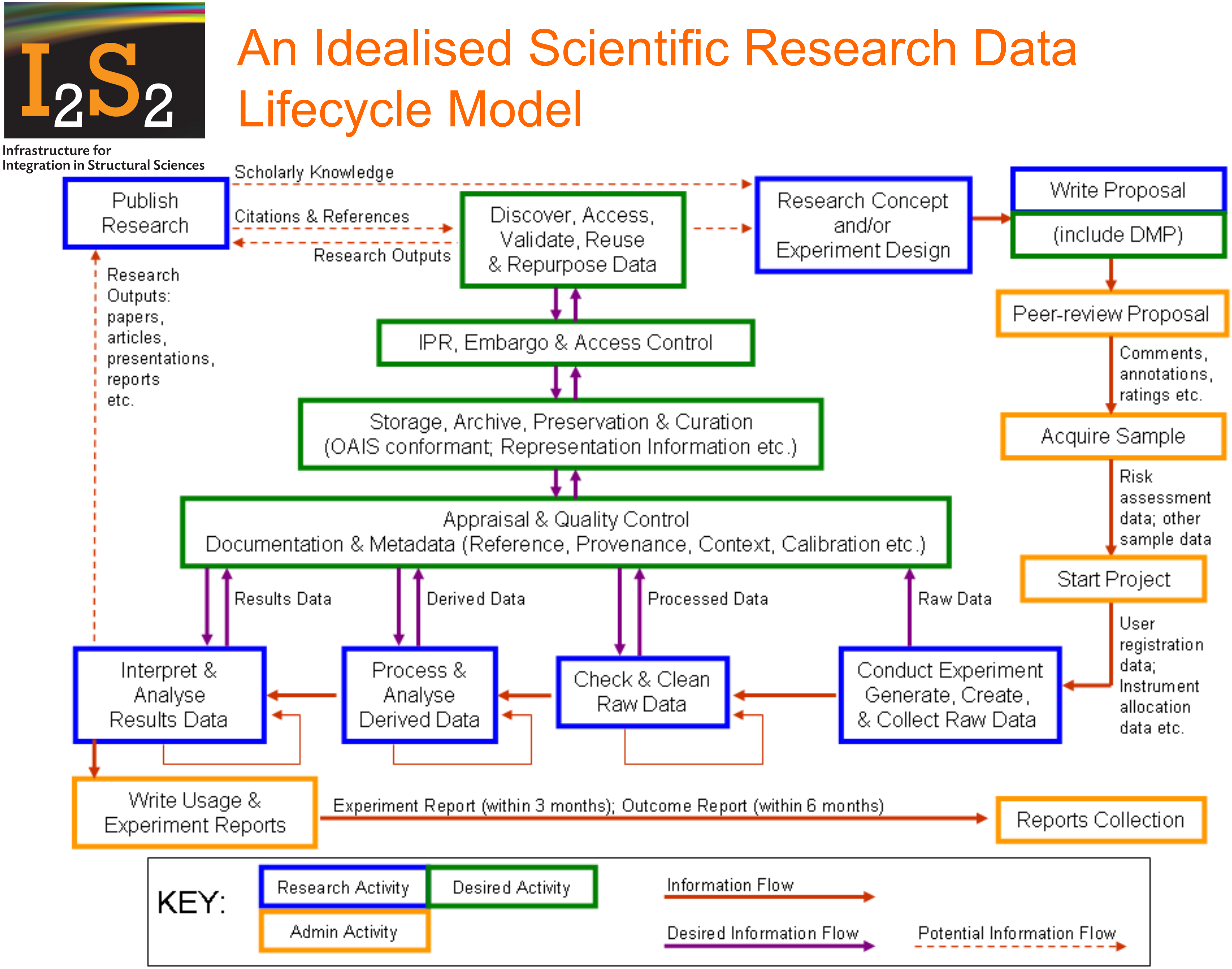
- Service function implies an obligation to retain raw data
- Efficiencies and benefits to be gained by working across organisational boundaries through an integrated approach
- Simplification of inter-organisational communications and tracking, referencing and citation of datasets
  - Standardised ERA forms
  - Unique persistent identifiers (Experiment/Sample identifiers currently based on beam line number)
- Core Scientific Metadata Model (CSMD) needs to be extended
  - For additional info e.g. costs; preservation
  - For use beyond STFC
  - Storage or management of derived and results data

Gap Analysis

Infrastructure for Integration in Structural Sciences

- 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
- Effective validation, reuse and repurposing of data requires
  - Trust and a thorough understanding of the data
  - Transparent contextual information detailing how the data were generated, processed, analysed and managed
- Based on idealised scientific research data lifecycle and case studies:
  - NCS & DLS
  - Earth Sciences & ISIS

An Idealised Scientific Research Data Lifecycle Model



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

Requirements: implementation

Infrastructure for Integration in Structural Sciences

- Relevant Technologies
  - Persistent Identifiers (URIs, DOIs etc.)
  - Metadata schema (PREMIS, XML, CML, RDF?)
  - Controlled vocabularies (ontologies?)
  - Integrated information model (structured, linked data?)
  - Extensions to CSMD & ICAT
  - Interoperability and exchange (OAI-PMH, file formats)
  - Data packaging (OAI-ORE)
  - OAIS Representation Information?

- Cultural Issues: responsibilities at different roles and levels of scale (research student, research supervisor, research laboratory, department, institution, service facility, large scale facility)
  - Best practice guidelines
  - Use of Standards
  - Advocacy
  - Training

Infrastructure for
Integration in Structural Sciences

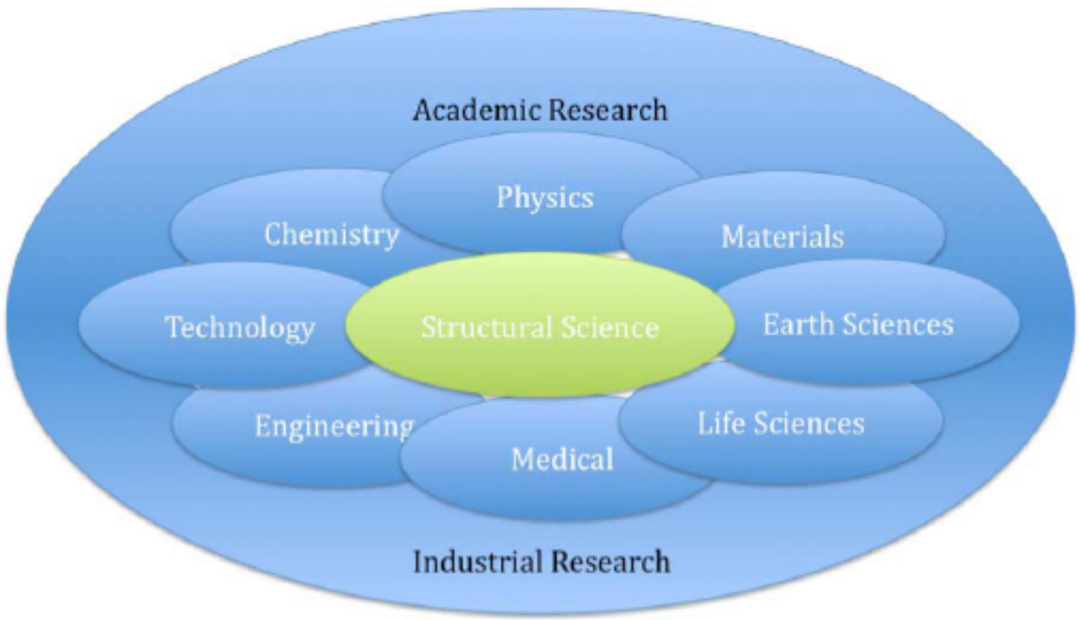
Requirements Summary

- 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:
  - Efficient exchange and reuse of data across disciplinary boundaries
  - Aggregation and/or cross-searching of related datasets
  - Data mining to identify patterns or trends

Infrastructure for
Integration in Structural Sciences

Work in Progress

- Requirements Gathering
- Use case studies & Pilot Implementations
- Integrated Information Model
- Cost/Benefits Analysis



I2S2
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

Progress:

- Details of use cases presented at I2S2 Models workshop in February
- Identification of issues in the use cases
- Examination of workflows and processes based on the idealised lifecycle model
- Development of data lifecycles for each use case

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

Pilot Implementation 2

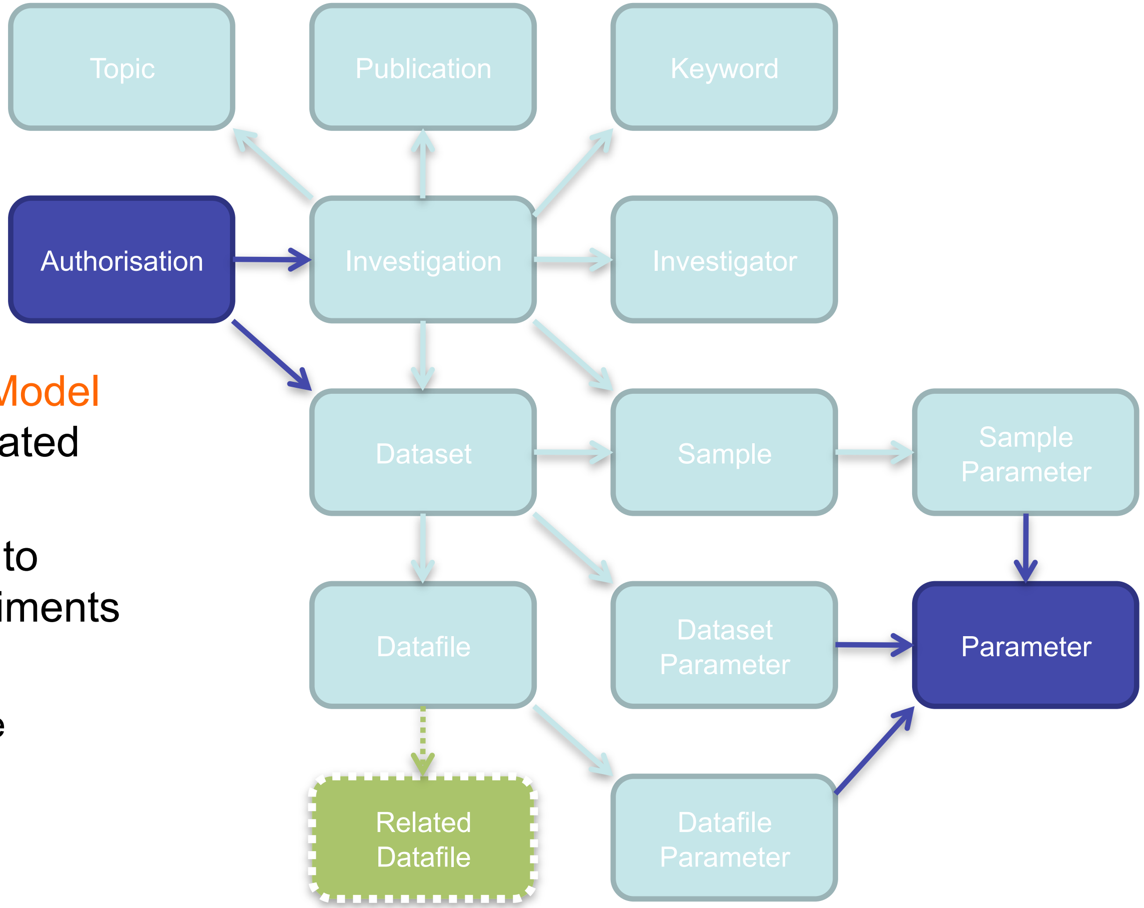
Infrastructure for
Integration in Structural Sciences

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

Integrated Information Model

Infrastructure for
Integration in Structural Sciences



- The Core Scientific Metadata Model (CSMD) is basis of I2S2 integrated information model
- CSMD was designed at STFC to describe facilities based experiments
- Forms a basis for extension:
  - To laboratory based science
  - To secondary analysis data
  - To preservation information
  - To publication data
- Covers the scientist's research lifecycle as well as the facilities

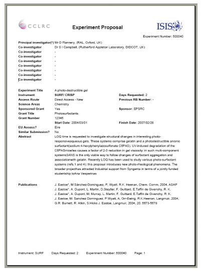
ICAT: A toolkit to catalogue and link facilities data

I2S2
Infrastructure for Integration in Structural Sciences

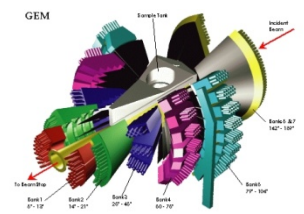
What is ICAT ?

ICAT is a database (with a well defined API) that provides a uniform interface to experimental data and a mechanism to link all aspects of research from proposal through to publication.

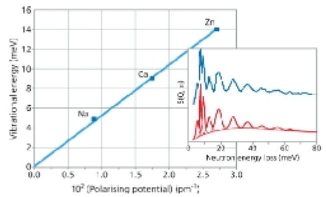
- Access data anywhere via the web
- Annotate your data
- Search for data in a meaningful way e.g. taxonomy, Sample, temperature, pressure etc
- Share data with colleagues
- Access data via your own programs (C++, Fortran, Java etc.) via the ICAT API
- Identify potential collaborations
- Utilise integrated e-Science High-Performance Computing and Visualisation resources
- Link to data from your publications
- Etc.



Example ISIS Proposal



GEM – High intensity, high resolution neutron diffractometer



H2-(zeolite) vibrational frequencies vs polarising potential of cations

B-lactoglobulin protein interfacial structure

ICAT

Proposals

Once awarded beamtime at ISIS, an entry will be created in ICAT that describes your proposed experiment.

Experiment

Data collected from your experiment will be indexed by ICAT (with additional experimental conditions) and made available to your experimental team

Analysed Data

You will have the capability to upload any desired analysed data and associate it with your experiments.

Publication

Using ICAT you will also be able to associate publications to your experiment and even reference data from your publications.

Developed at STFC e-Science for use in ISIS and DLS facilities

http://code.google.com/p/icatproject

ICAT
Science & Technology Facilities Council

Infrastructure for
Integration in Structural Sciences

Cost/Benefits Analysis

- A before and after cost-benefit analysis using the Keeping Research Data Safe (KRDS2) model
- Extending the KRDS Model
  - Early version presented at RDMI Programme workshop, Manchester, 12th March 2010
  - Initial focus has been on extensions and elaboration of activities in the research (KRDS “pre-Archive”) phase
  - Current work is elaborating the publication of research as an addition to the model
- Metrics and assigning costs
  - Identification of activities in idealised data lifecycle model that will represent significant cost savings or benefits at NCS
  - Work to identify non-cost benefits and possible metrics to associate with individual research projects

I2S2

Infrastructure for
Integration in Structural Sciences

Questions & Discussion

Further Information:

- I2S2 “Models Workshop” Presentations, Feb 2010:
  http://www.ukoln.ac.uk/projects/I2S2/events/modelling-workshop-2010-feb/
- Idealised scientific research data lifecycle model
  http://blogs.ukoln.ac.uk/I2S2/