



Infrastructure for Integration in Structural Sciences

Manjula Patel
UKOLN, University of Bath

Data Infrastructure Challenges: working across
scale, disciplinary and institutional boundaries

University of Leicester

5th May 2011





Infrastructure for
Integration in Structural Sciences

Outline

- I2S2 Project overview & aims
- Project Team
- Research data & infrastructure
- Requirements analysis
- A Scientific Research Activity Lifecycle Model
- An integrated services approach
- Testing the I2S2 information model
- Cost-Benefits analysis



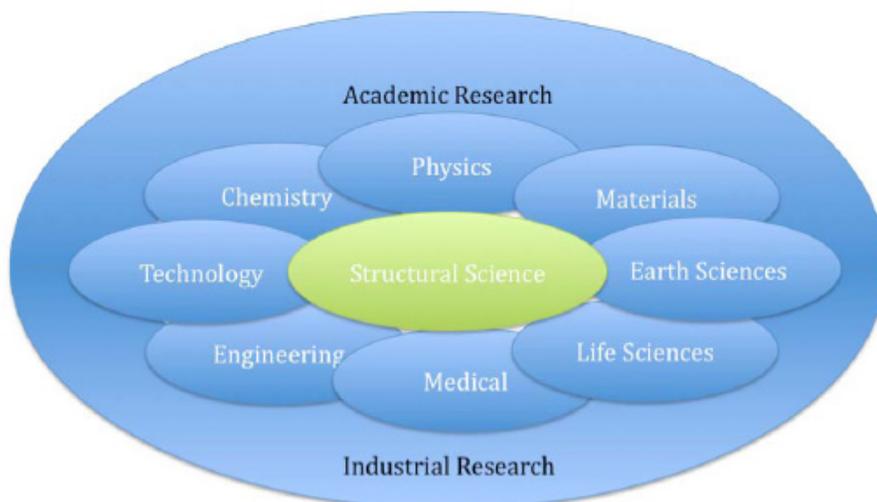
Diamond Light Source (DLS),
Science & Technology Facilities Council, UK



Infrastructure for
Integration in Structural Sciences

Overview & Aims

- Research Data Management Infrastructure strand of the JISC's Managing Research Data Programme (18 month project, Oct 2009 - March 2011)
- 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

Project Team

- Liz Lyon (Project Director, UKOLN (University of Bath) & Digital Curation Centre)
- Manjula Patel (Project Manager, UKOLN (University of Bath) & Digital Curation Centre)
- Sarah Hext (Financial Administrator, UKOLN (University of Bath))
- Simon Coles (EPSRC National Crystallography Centre, University of Southampton)
- Neil Beagrie (Charles Beagrie Ltd.)
- Brian Matthews (Science & Technology Facilities Council)
- Erica Yang (Science & Technology Facilities Council – now at Bodleian libraries, University of Oxford)
- Martin Dove (Earth Sciences, University of Cambridge)
- Peter Murray-Rust (Chemistry, University of Cambridge)

- Simon Hodson (JISC Managing Research Data Programme Manager)

m.patel@ukoln.ac.uk

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



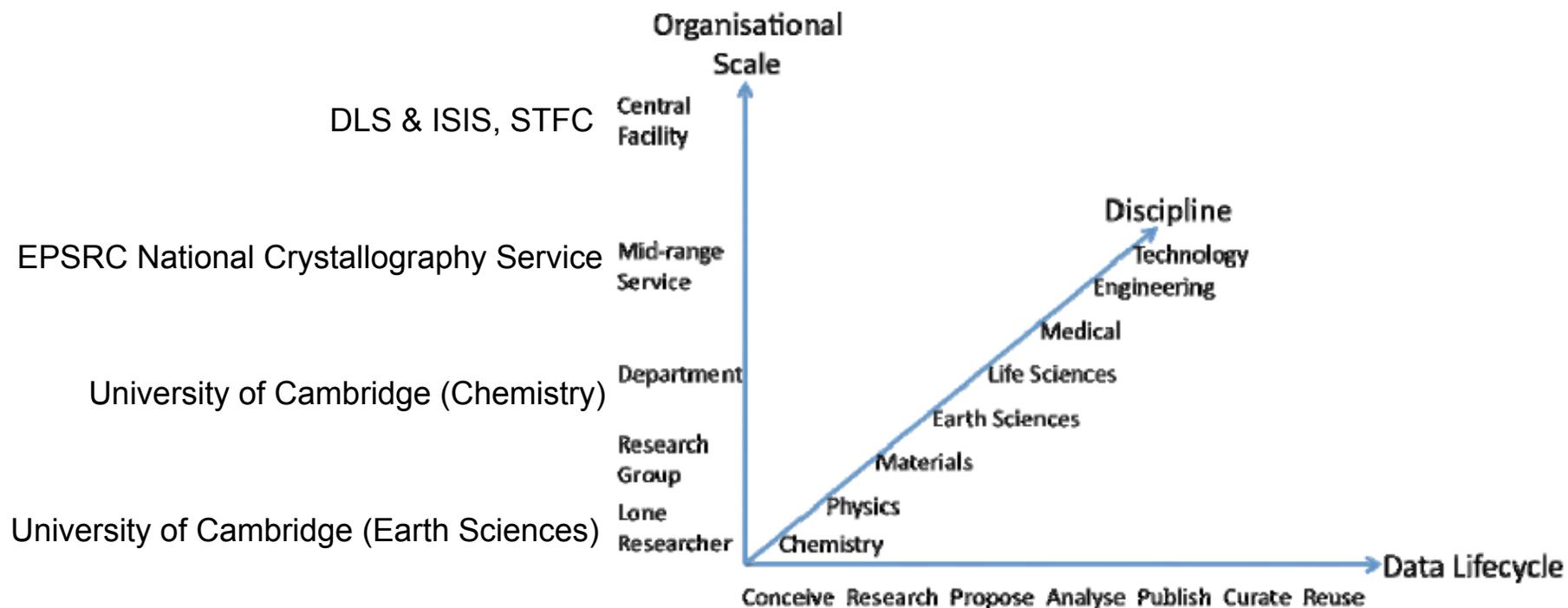
Infrastructure for
Integration in Structural Sciences

Overall Methodology

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





Infrastructure for
Integration in Structural Sciences

Research Data & Infrastructure

- **Research Data** includes (all information relating to a physical 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



Infrastructure for
Integration in Structural Sciences

Earth Sciences, Cambridge

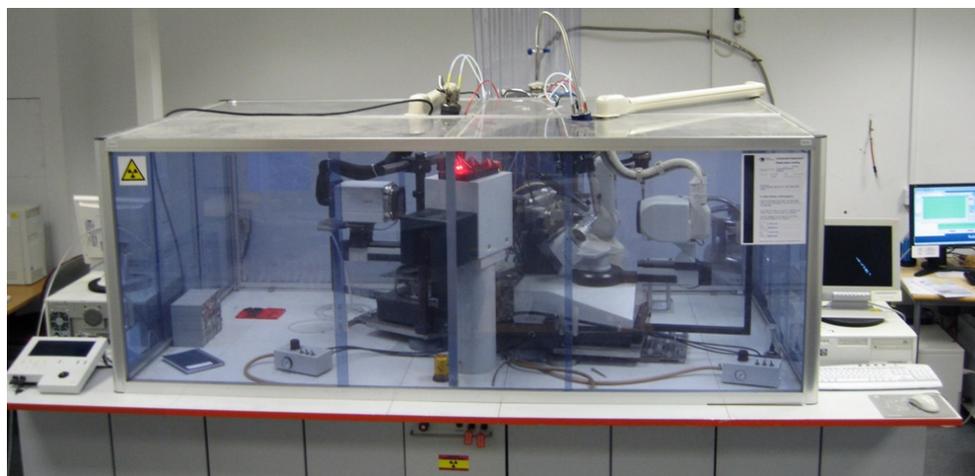
- Construct large scale atomic models of matter that best match experimental data; using Reverse Monte-Carlo Simulation techniques
- Experiment and data collection conducted at ISIS Neutron Source (GEM)
- **Little or no shared infrastructure**
 - Data sharing with colleagues via email, ftp, memory stick 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 **non-intrusive**



Infrastructure for
Integration in Structural Sciences

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



Infrastructure for
Integration in Structural Sciences

EPSRC NCS, Southampton

- **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



Infrastructure for
Integration in Structural Sciences

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**

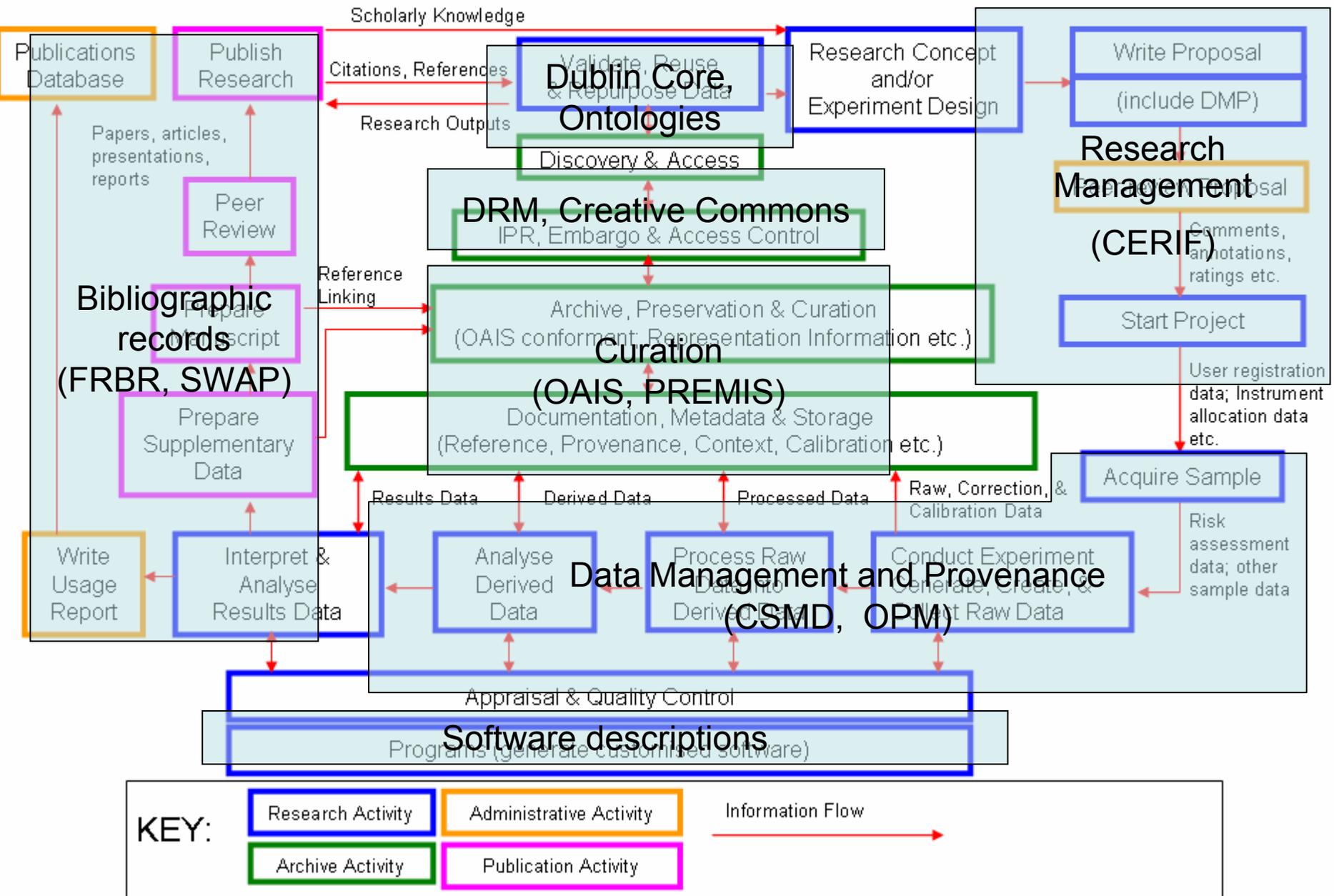


Infrastructure for
Integration in Structural Sciences

Generalised Issues

- Basic requirement for robust **data storage and backup** facilities to sophisticated needs such as **structuring and linking** together of data
- **Management of intermediate, derived and results data a major issue**
- **Contextual information is not routinely captured**
- Processing pipeline is dependent on a **suite of software**
- The actual **workflow or processing pipeline is not routinely recorded**
- Need for 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**
 - Unique persistent identifiers
 - Standardised Experiment Risk Assessment forms
- Solutions should be as **non-intrusive** as possible

An Idealised Scientific Research Activity Lifecycle Model

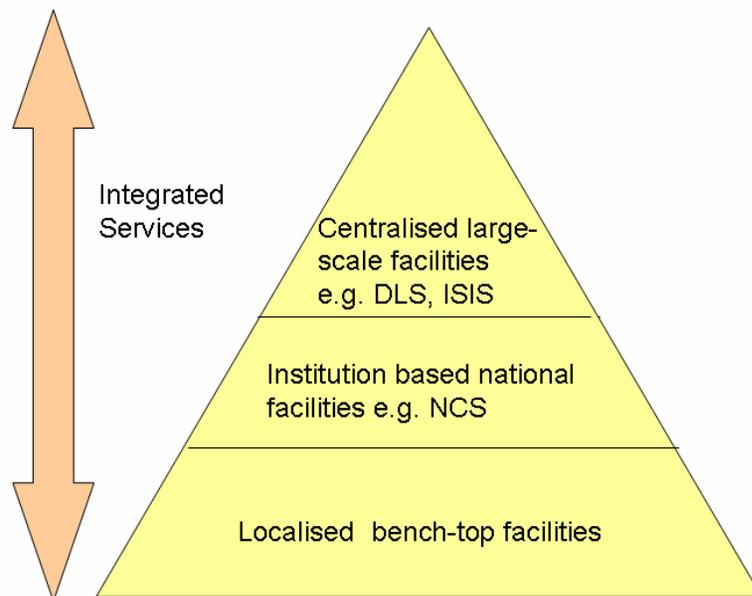




Infrastructure for
Integration in Structural Sciences

An Integrated Service Approach

- Individual researcher, group, department, institution, facilities all **working within their own frameworks**
- Merit in adopting an **integrated framework** which caters for all scales of scientific research
- Researchers need to be able to manage their data **across institutional and domain boundaries in a seamless manner**





Infrastructure for
Integration in Structural Sciences

I2S2 Integrated Information Model

- Core Scientific Metadata Model (CSMD)
 - Designed to describe **facilities based experiments** in Structural Sciences
 - Forms a **basis for extension to**: Laboratory based science; Derived data; Secondary analysis data; Preservation information; Publication data
 - Aim to cater for the **scientist's research lifecycle** as well as facilities data
- 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-IM = CSMD-Core + oreChem Model
 - Underpins **distributed data management**
 - Effective **inter-disciplinary data sharing**
- 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” for long-term storage**
 - Enables capture of **provenance** information



Infrastructure for
Integration in Structural Sciences

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: Disciplinary issues

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



Infrastructure for
Integration in Structural Sciences

Cost-Benefits Analysis

- A **cost-benefit analysis** using the Keeping Research Data Safe (KRDS) model
- **Extending the KRDS Model**
 - Focus has been on extensions and elaboration of activities in the research phase (KRDS “pre-Archive” stage)
- **Metrics and assigning costs**
 - Identification of activities in research activity lifecycle model that will represent significant cost savings or benefits
 - Identification of 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 etc.)



Infrastructure for
Integration in Structural Sciences

Conclusions

- Considerable variation in data management requirements across differing scales of science
- I2S2 Integrated framework aims to:
 - Support the scientific research activity lifecycle model
 - Capture processes and provenance information
 - Streamline flow of metadata, administrative information and experiment data across organisations
 - Interoperate with and complement existing models and frameworks
- Cost-benefits analysis to assess impact of interventions



Infrastructure for
Integration in Structural Sciences

Acknowledgements

- Liz Lyon (Project Director, UKOLN (University of Bath) & Digital Curation Centre)
- Sarah Hext (Financial Administrator, UKOLN (University of Bath))
- Simon Coles (EPSRC National Crystallography Centre, University of Southampton)
- Neil Beagrie (Charles Beagrie Ltd.)
- Brian Matthews (Science & Technology Facilities Council)
- Erica Yang (Science & Technology Facilities Council – now at Bodleian libraries, University of Oxford)
- Martin Dove (Earth Sciences, University of Cambridge)
- Peter Murray-Rust (Chemistry, University of Cambridge)

- Simon Hodson (JISC Managing Research Data Programme Manager)

m.patel@ukoln.ac.uk

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