UNIVERSITY OF Southampton
School of Chemistry

Case Study 1

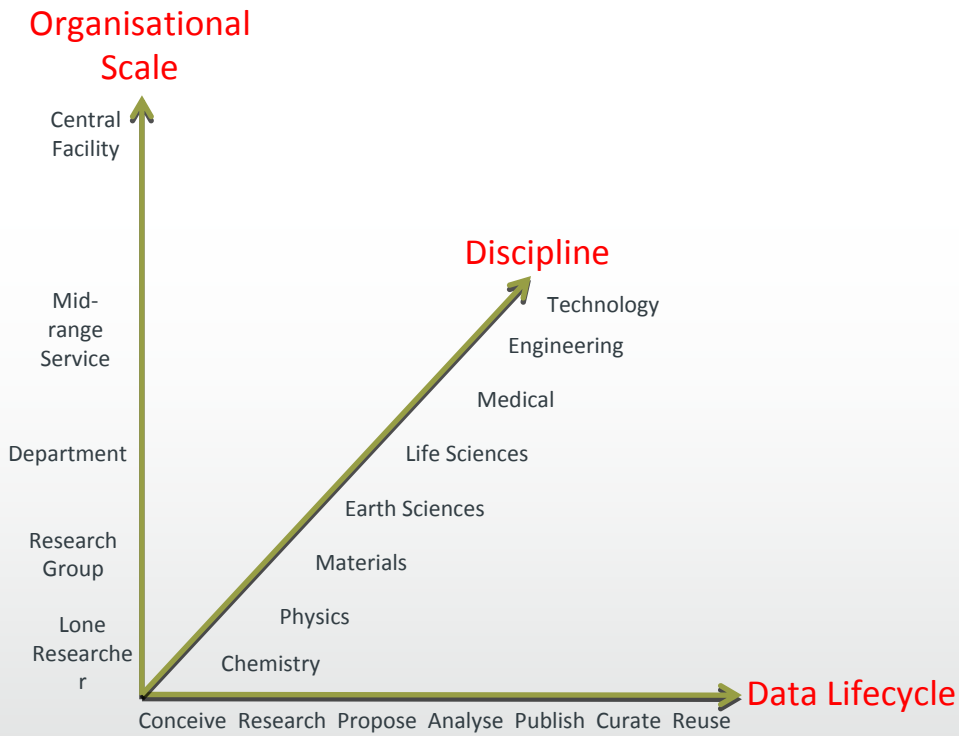
Probing Data Lifecycle and Inter-Institutional Issues.

Simon Coles (s.j.coles@soton.ac.uk)
5th May 2011

UNIVERSITY OF Southampton
School of Chemistry

Aims

- Increase efficiency and cope with scale across institutions
- Support whole data lifecycle management so that processes become seamless
- Map across organisational infrastructures
- Develop from established starting platforms



Challenges

- Researchers capture, manage, discuss & disseminate data in relative isolation with highly fragmented data infrastructures and poorly integrated software applications
- Conventional systems of publication lead to insufficient provenance of results and irreproducible experiments
- The processes for recognition lead to a lack of inclination and incentive to share or make all the supporting information for a study publicly available
- Low awareness of data curation and preservation issues leads to data loss and reduced productivity

UNIVERSITY OF Southampton
School of Chemistry

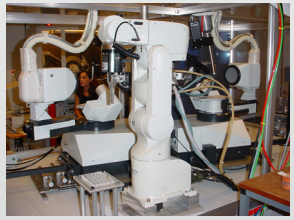
The Lone Researcher

- Considers research as ‘linear’ &/or ‘individual’ processes
- Considers data volumes to be manually manageable
- Relatively low-powered instrumentation
- Is less concerned with sharing
- Has little or no support infrastructure

University of Southampton
School of Chemistry

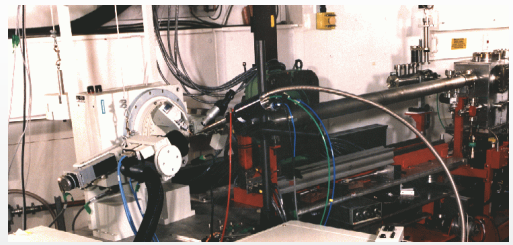
Mid-range Facility

- Service provision function / working as a team
- Technique focused
- High powered lab-based instrumentation
- Operating across institutions
- Moderate infrastructure
- Local data management solutions



Central Facilities

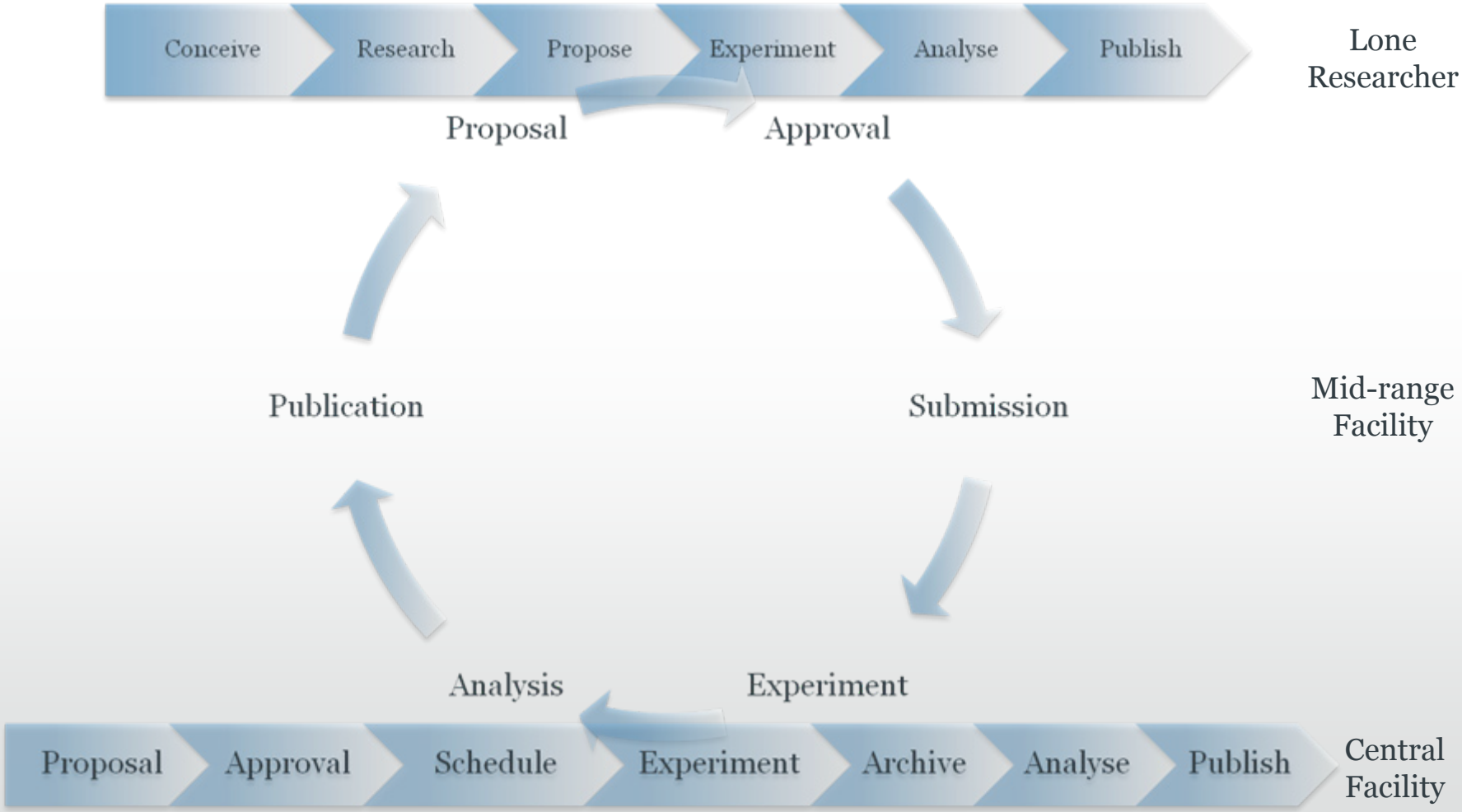
- Operate on behalf of multiple institutions (community)
- Have established and often bespoke & formulaic experimental processes
- Very powerful custom instrumentation
- Have large supporting infrastructure
- Users take data away (different mechanisms)
- Don’t often know what happens to data



Proposal → Approval → Schedule → Experiment → Archive → Analyse → Publish

UNIVERSITY OF Southampton
School of Chemistry

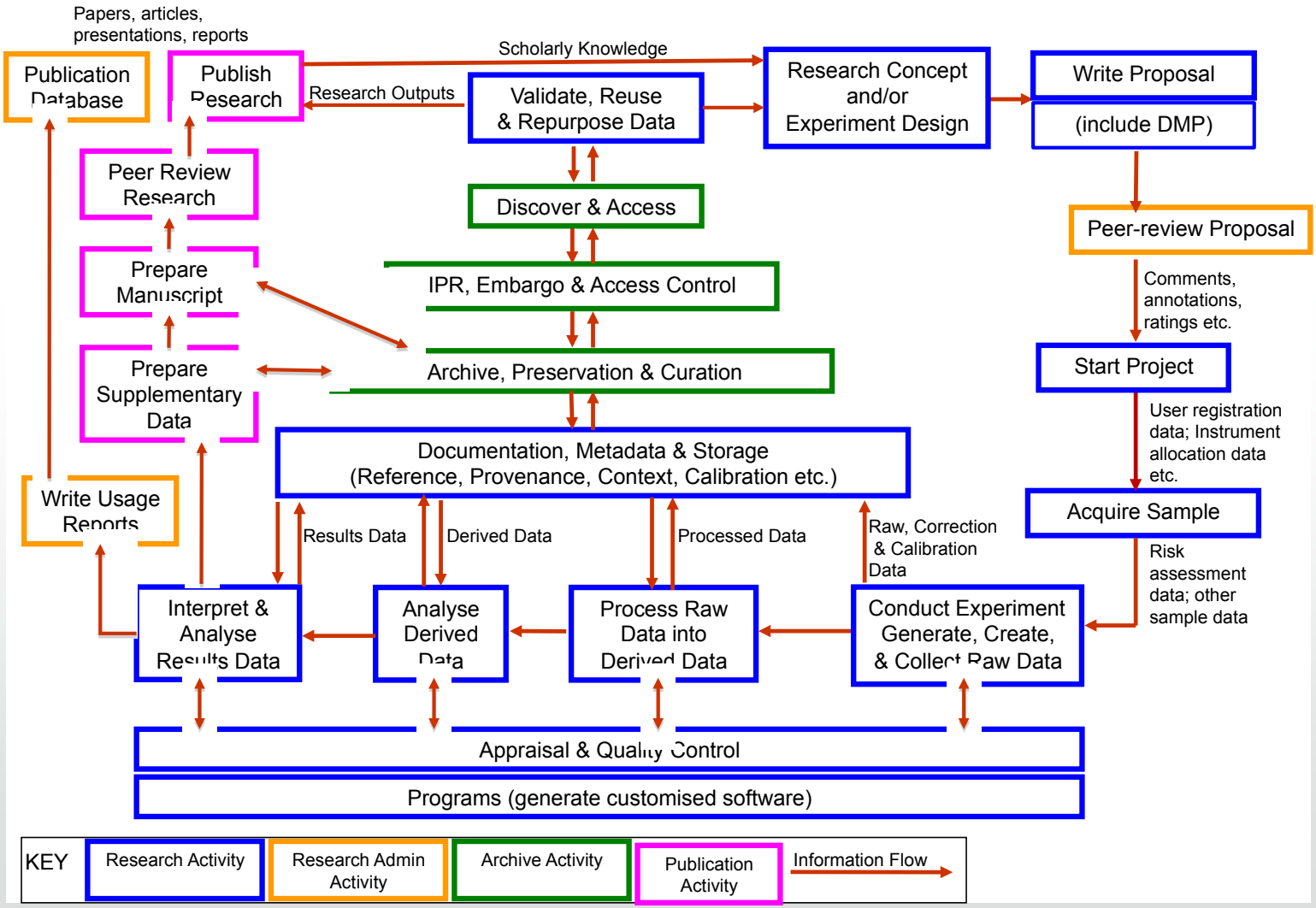
The Bigger Picture



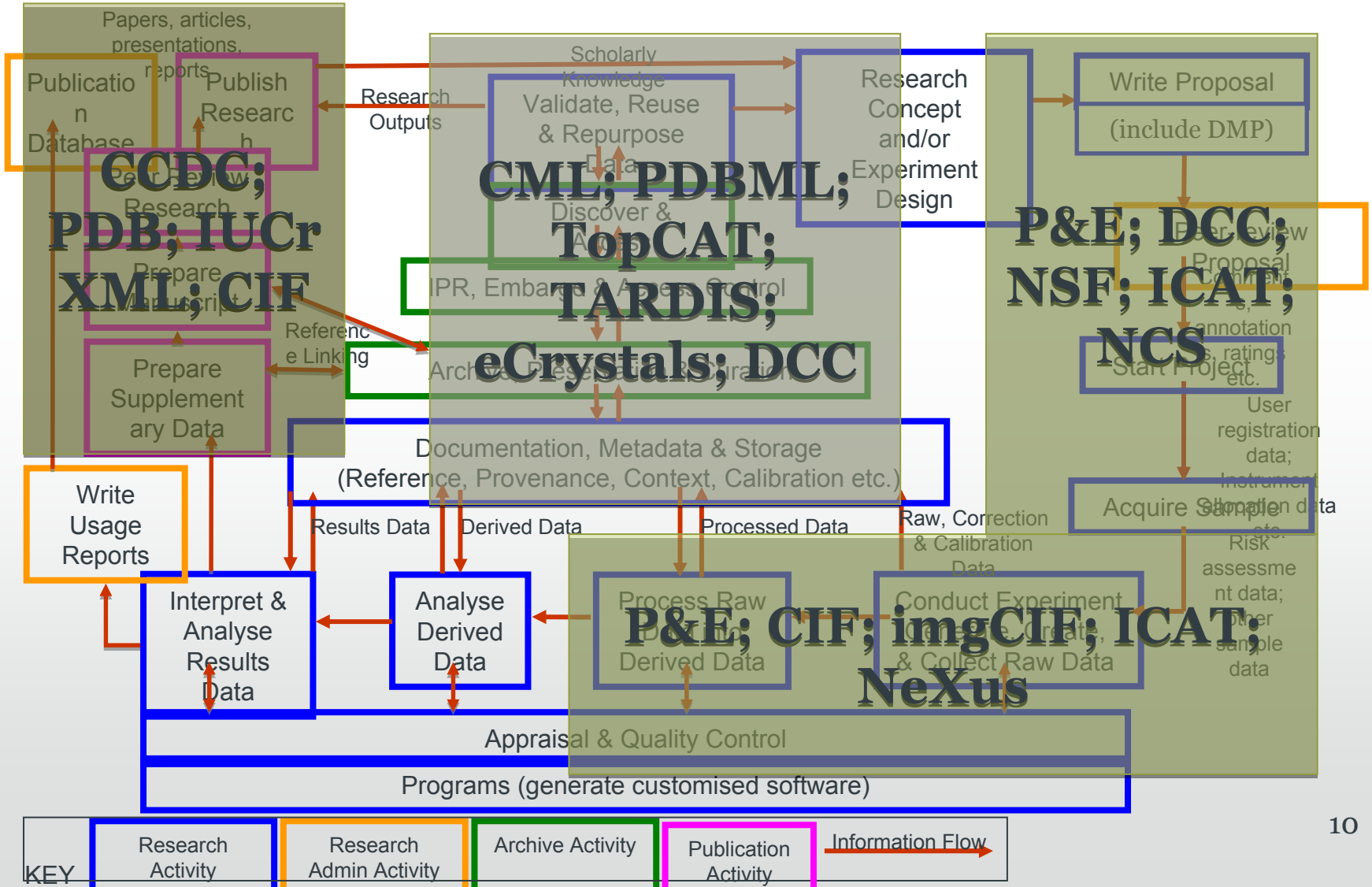
Our Approach

- Model the research processes
- Analyse established working practice
- Adopt and develop established systems
- Examine the interfaces between infrastructures
- Develop a harmonised generic model that fits all infrastructures

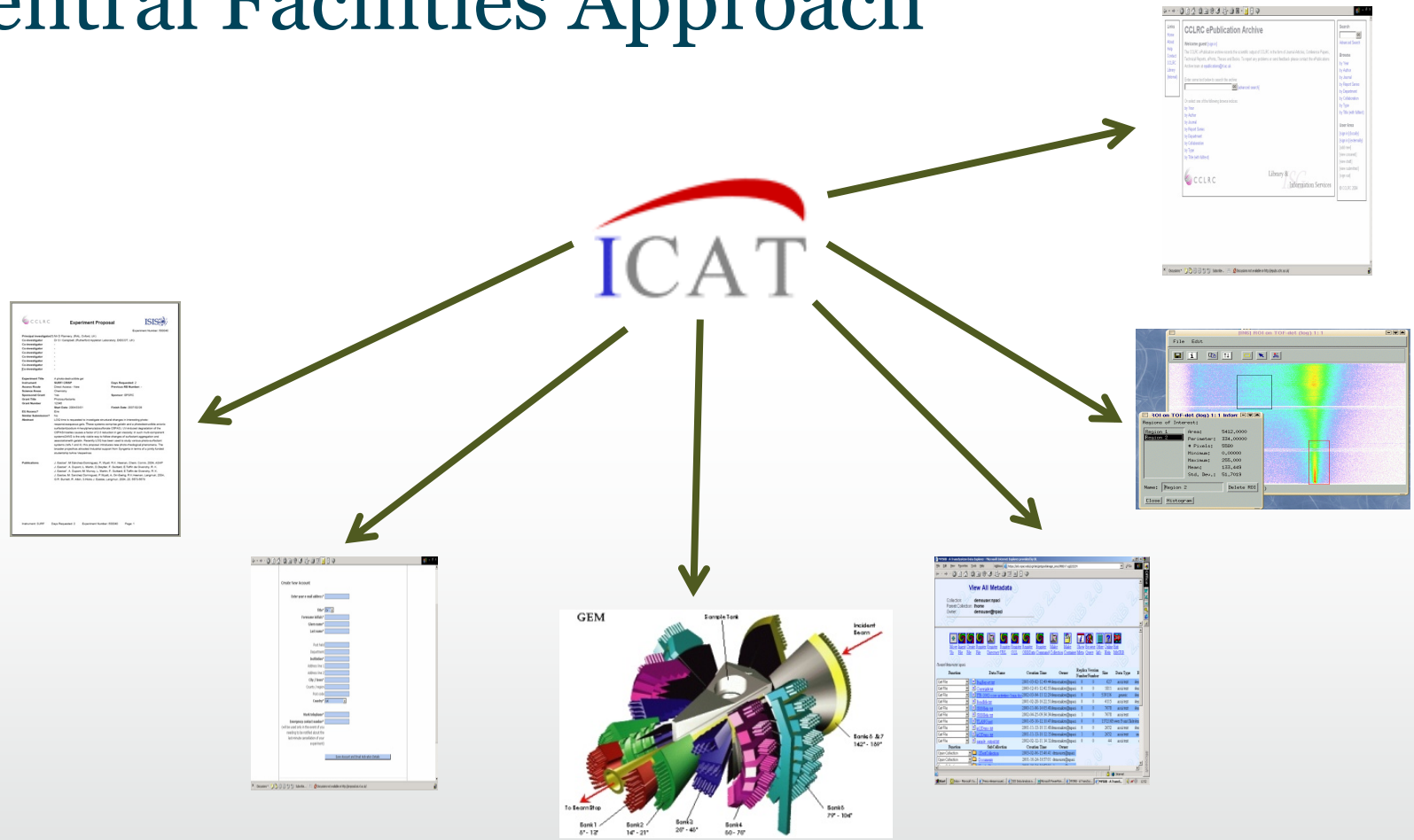
Idealised Research Activity Model



Established Approaches



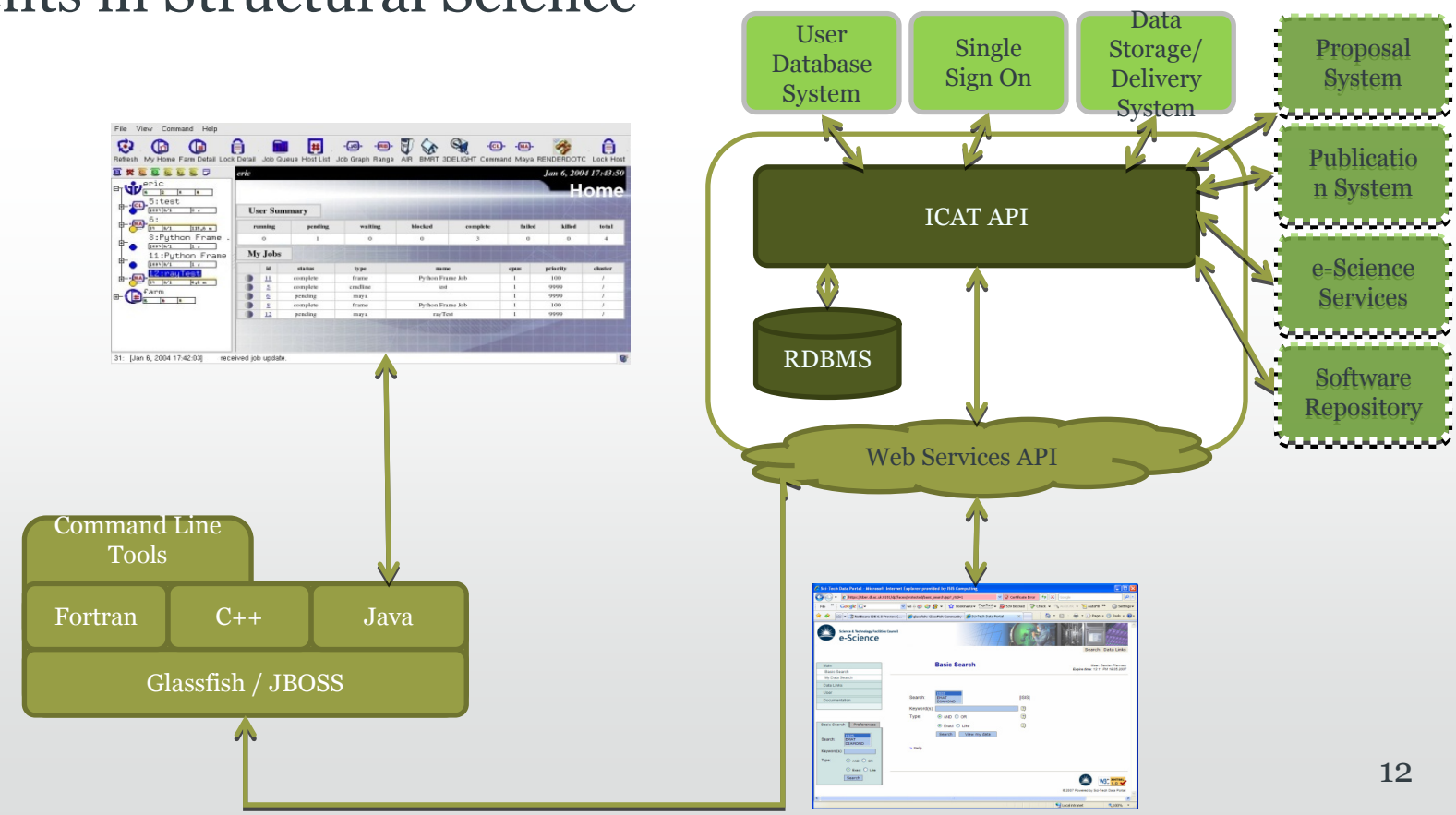
Central Facilities Approach



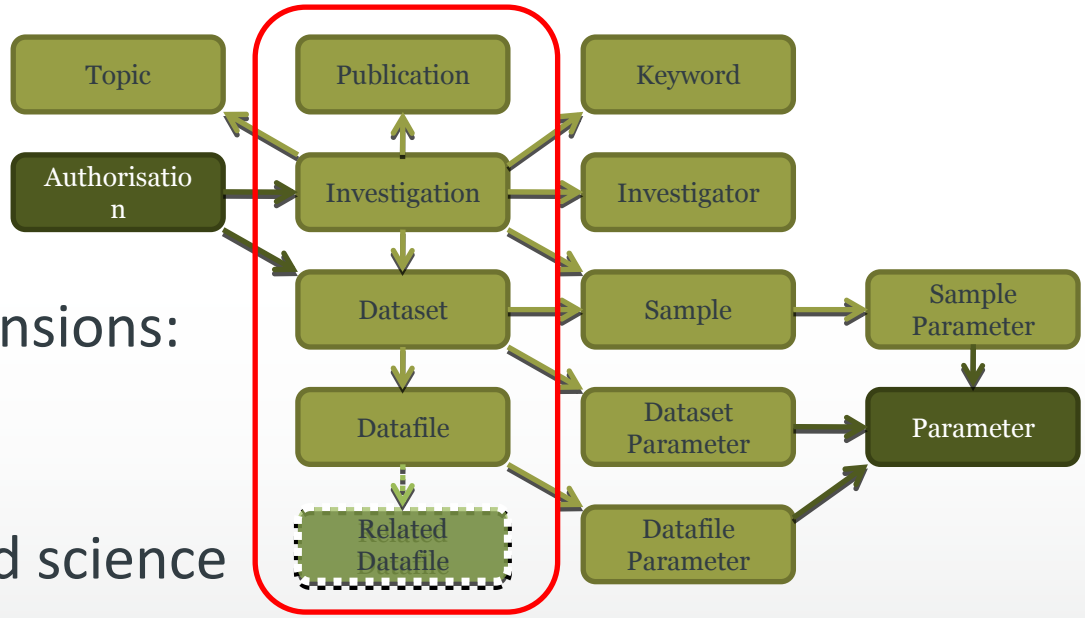
Proposal → Approval → Schedule → Experiment → Archive → Analyse → Publish

ICAT

The Core Scientific Metadata model forms the information model for ICAT & is designed to describe facilities-based experiments in Structural Science



CSMD as a Starting Point



- Forms the basis for extensions:
  - To derived data
  - To laboratory based science
  - To secondary analysis data
  - To preservation information
  - To publication data

UNIVERSITY OF Southampton
School of Chemistry

Planning and Enactment

Rationalisation of objects by way of the explicit statement of intent and action (Mark Borkum).

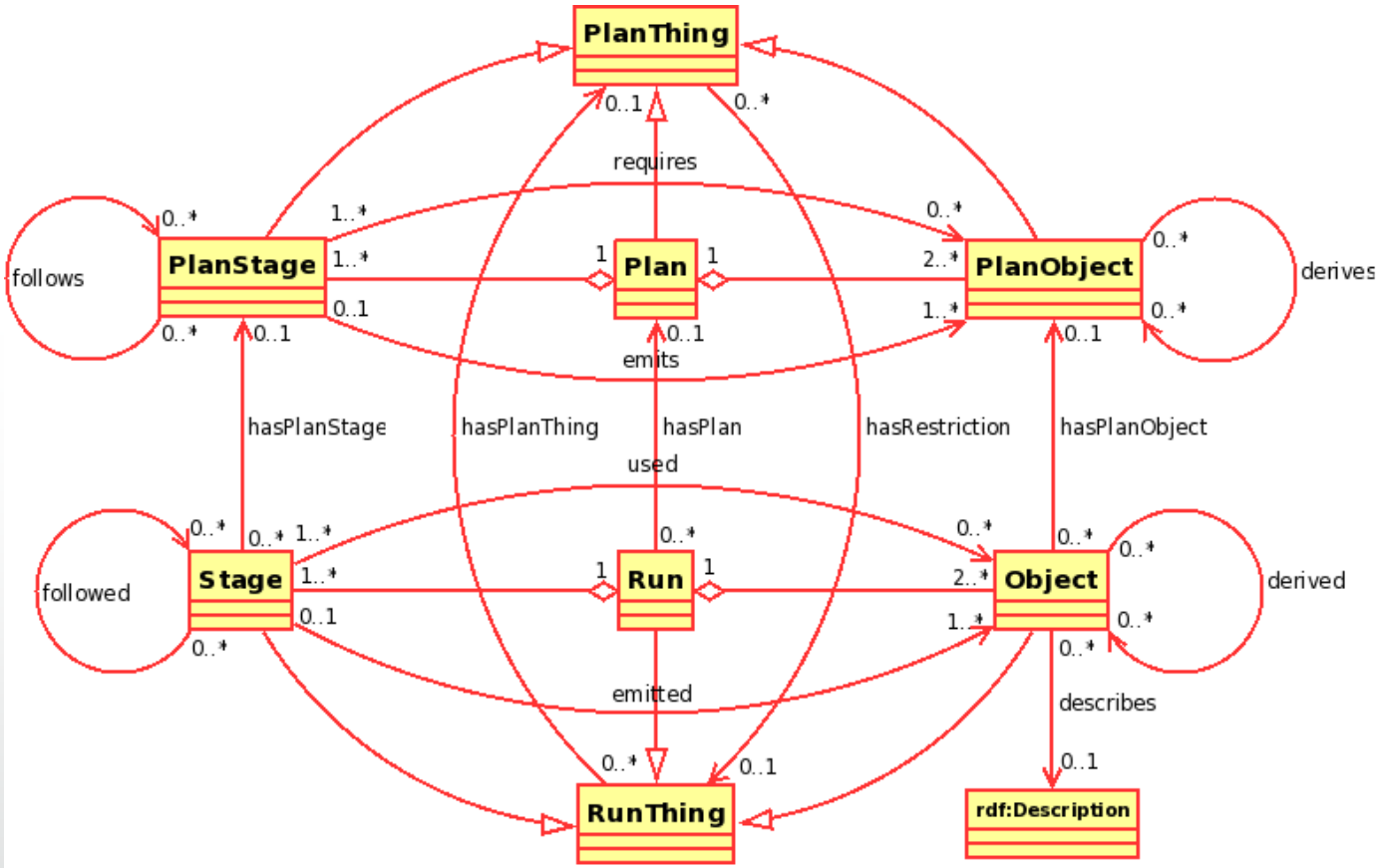
"In theory, there is no difference between theory and practice. But, in practice, there is." Unknown (possibly Yogi Berra)

Introduction

- Q: What defines a Semantic Web application?
- A: 3x concepts:
  1. Identity
     – *Who*
  2. Vocabulary
     – *What*
  3. Provenance
     – *Where, When, Why and How*

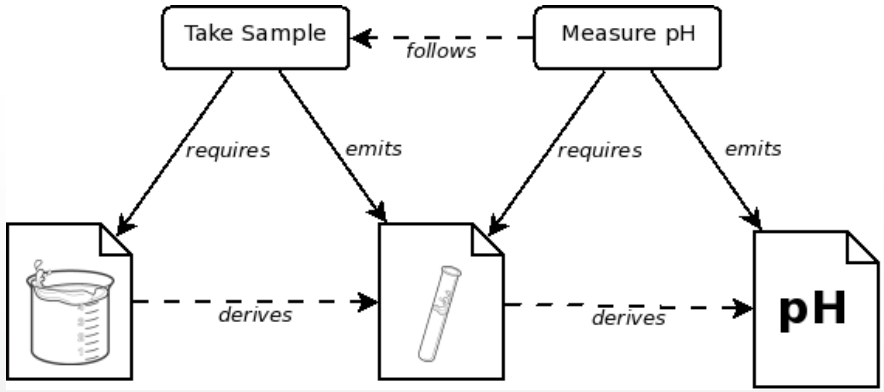
http://memory-alpha.org/wiki/File:Sickbay_laboratory.jpg

Planning and Enactment Ontology – UML



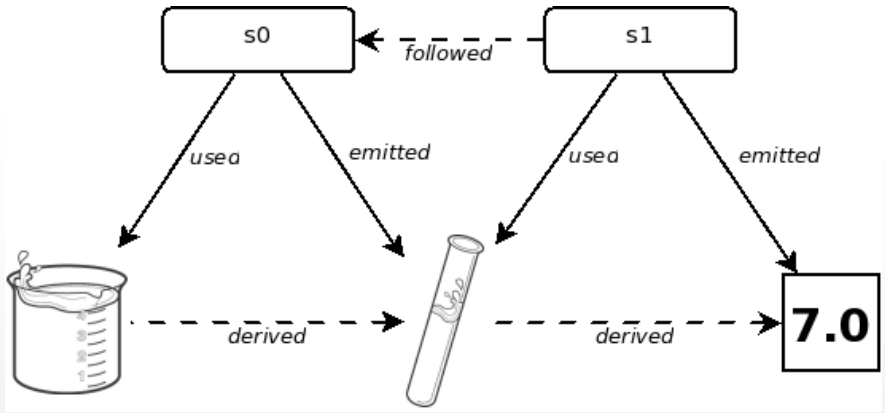
Planning

- Prospective provenance
- Describes a method that *will be enacted* (in the future)
- Three entity types:
  - Plan
  - Plan Stage
  - Plan Object



Enactment

- Retrospective provenance
- Describes a method that *was enacted*
- Three entity types:
  - Run
  - Stage
  - Object

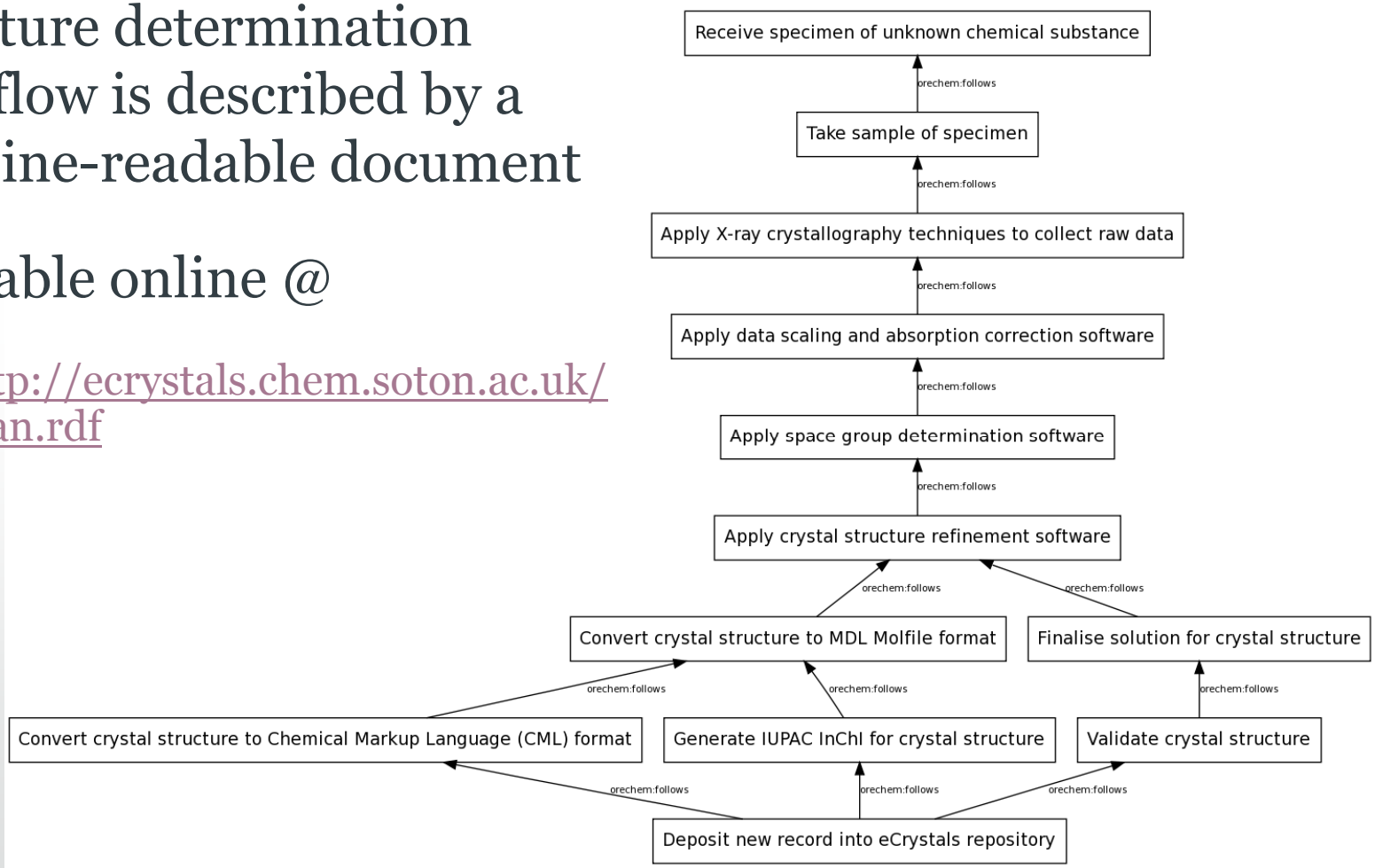


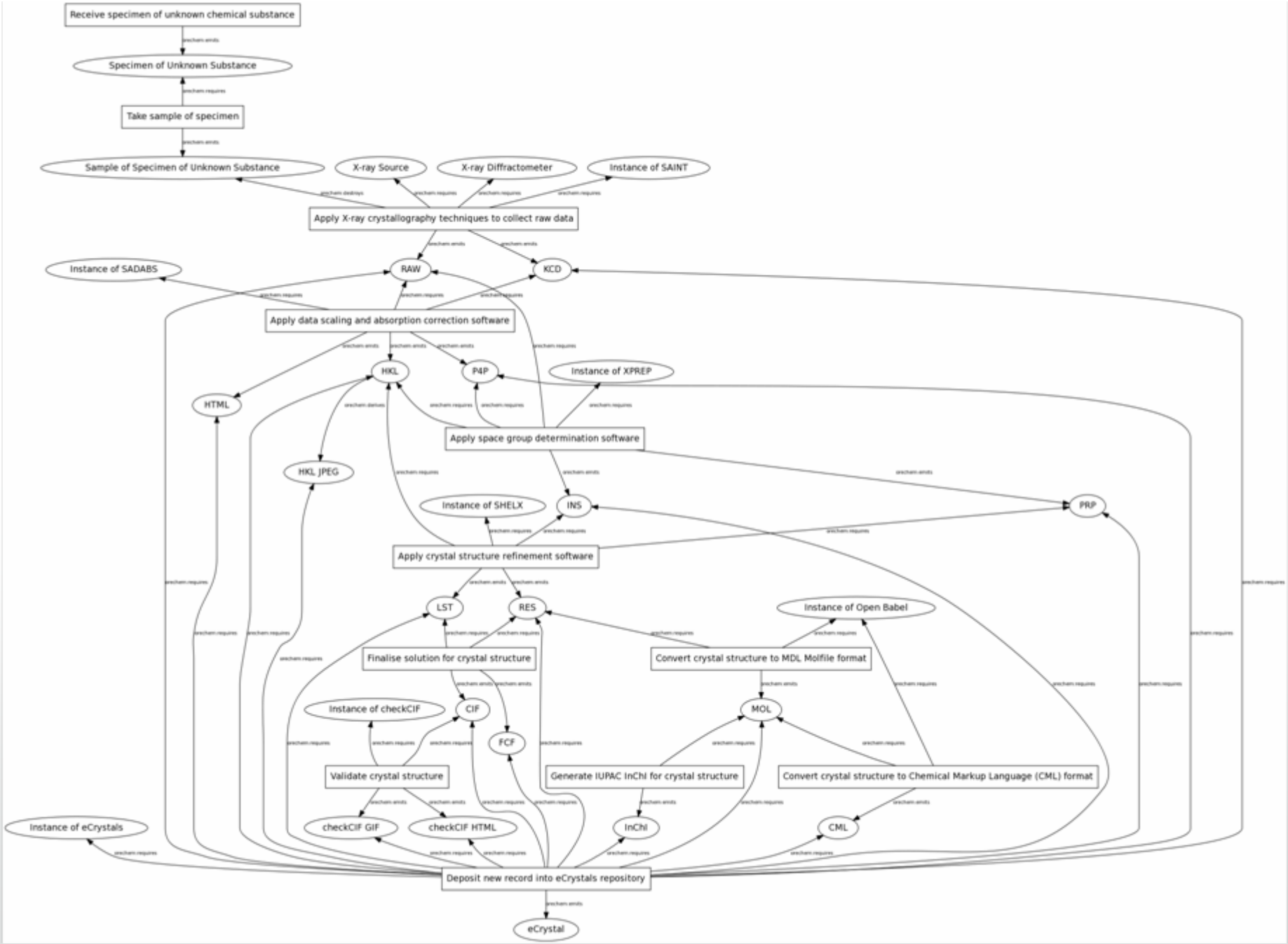
Satisfaction

- Given an enactment that has a plan, iff for all entities in the plan, there exists a corresponding entity in the enactment, then the enactment has "satisfied" the plan.
  - Satisfying the plan is equivalent to "being able to finish" the enactment.
  - This is distinct from "finishing" the enactment, e.g., the agent may repeat an action.
  - Each iteration creates a separate chain of events.

Plan for eCrystals Record

- Structure determination workflow is described by a machine-readable document
- Available online @
  - http://ecrystals.chem.soton.ac.uk/plan.rdf




Receive specimen of unknown chemical substance
Specimen of Unknown Substance
Take sample of specimen
Sample of Specimen of Unknown Substance
X-ray Source
X-ray Diffractometer
Instance of SAINT
Apply X-ray crystallography techniques to collect raw data
Instance of SADABS
RAW
KCD
Apply data scaling and absorption correction software
HKL
P4P
Instance of XPREP
HTML
HKL JPEG
Apply space group determination software
Instance of SHELX
INS
PRP
Apply crystal structure refinement software
LST
RES
Instance of Open Babel
Finalise solution for crystal structure
Convert crystal structure to MDL Molfile format
Instance of checkCIF
CIF
MOL
FCF
Validate crystal structure
Generate IUPAC InChI for crystal structure
Convert crystal structure to Chemical Markup Language (CML) format
Instance of eCrystals
checkCIF GIF
checkCIF HTML
InChI
CML
Deposit new record into eCrystals repository
eCrystal

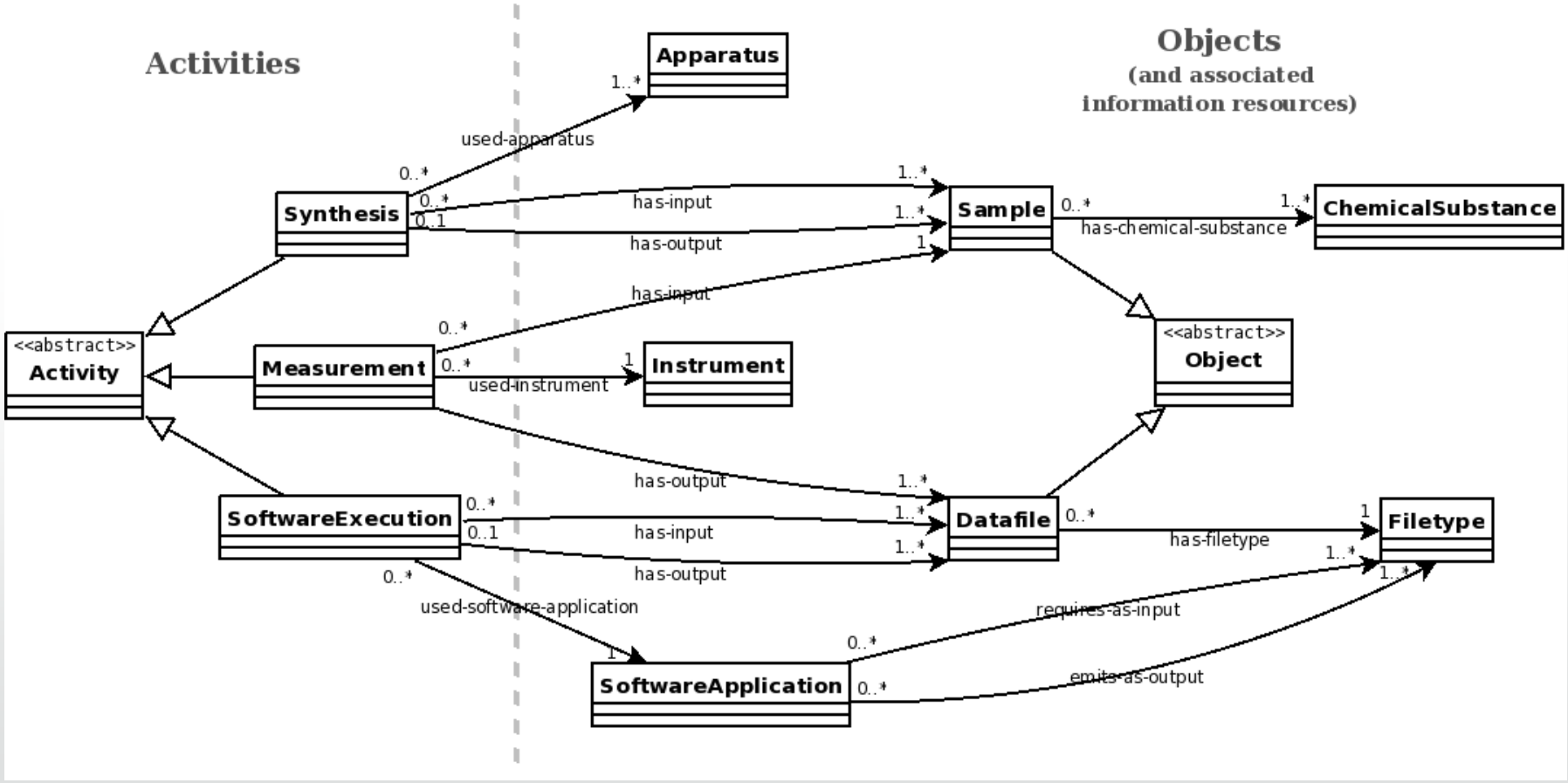
UNIVERSITY OF Southampton
School of Chemistry

To Conclude

Bringing the models together

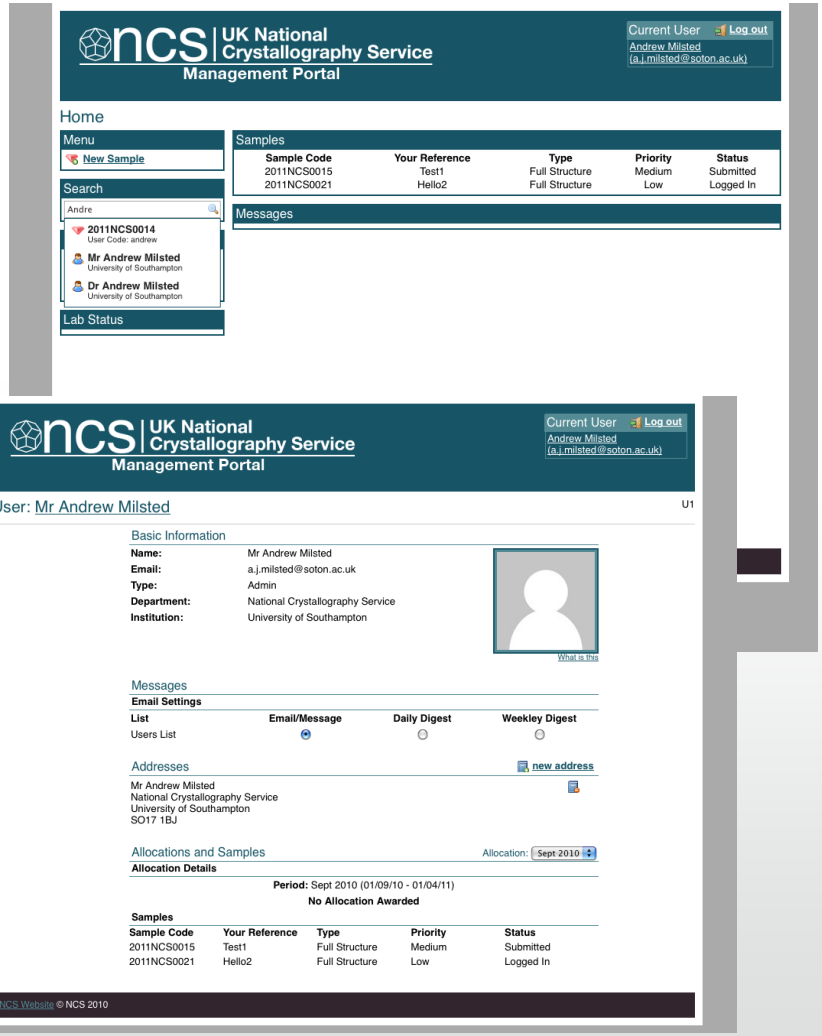
I2S2 Information Model (Core)

- Entities in the CSMD planned and enacted



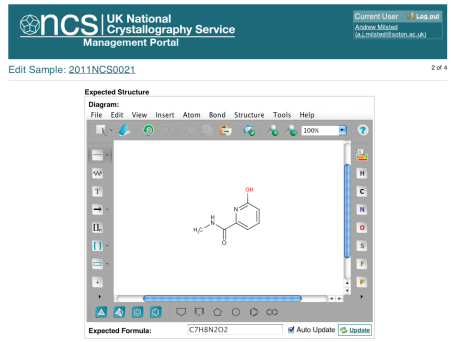
University of Southampton
School of Chemistry

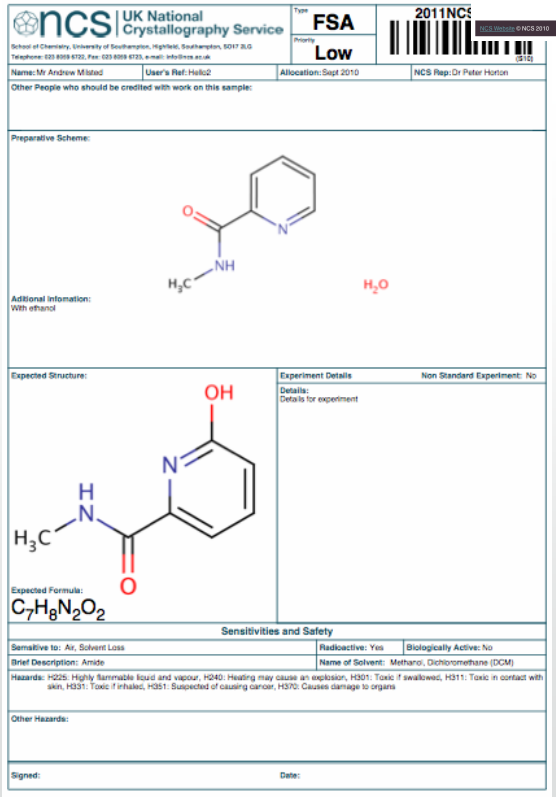
NCS Implementation



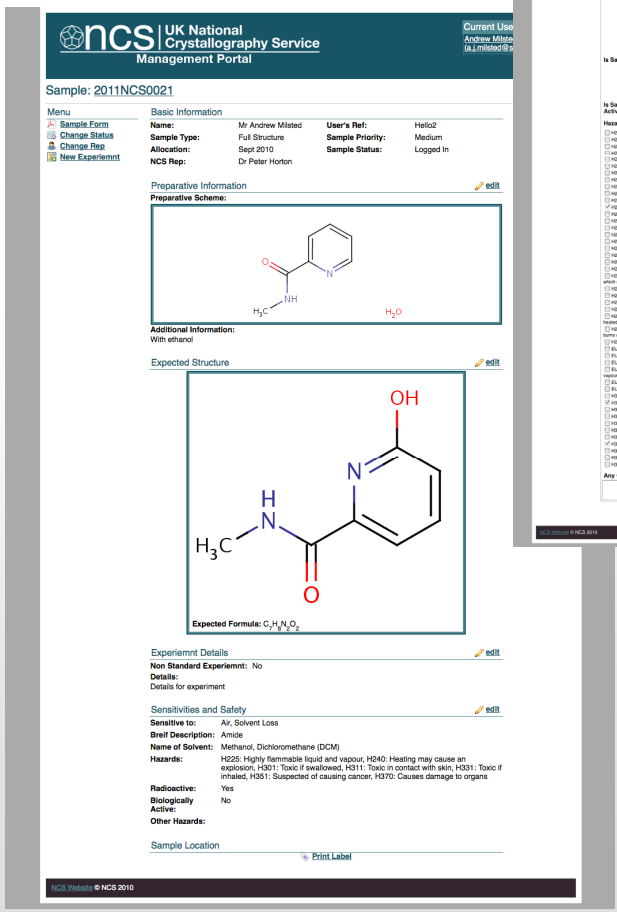
UNIVERSITY OF Southampton
School of Chemistry

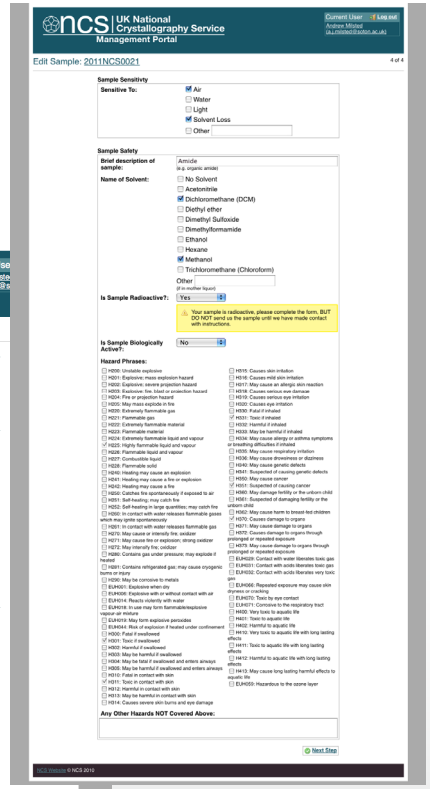
NCS Implementation





NCS Implementation





UNIVERSITY OF Southampton
School of Chemistry

Thanks!

- Project partners
  - UKOLN
  - STFC
  - Cambridge
  - DCC
  - Neil Beagrie
- Developers: Andrew Milsted & Mark Borkum
- oreChem project team & Microsoft
- JISC