Metadata Models for Experimental Science Data Management

Brian Matthews

Facilities Programme Manager

Scientific Computing Department, STFC

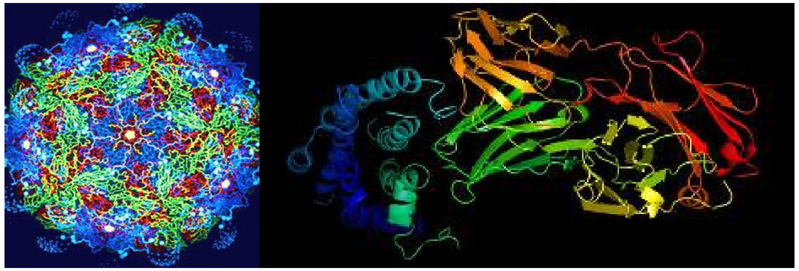
Co-Chair RDA Photon and Neutron Science Interest Group

Task lead, NA on metadata standardisation, NFFA-Europe

nffa.eu
nanoscience foundries & fine analysis

Science & Technology
Facilities Council

Large-Scale Analytic Facilities



Key challenges of the 21st century

- energy, global climate, health and security
- study matter at the scales
  - from single atoms (10^{-10} m) to living cells (10^{-6} m) to whole systems (10^{-3} - 1 m)

High resolution "microscopes" ⟶ intense beams of particles ⟶ Specialist sources

Requires large scale research infrastructures that are beyond the capability of any single university or research group

Diamond

ISIS

Photons (X-Rays) "see" electric charge – high atomic number nuclei

Neutrons "see" nucleons – including hydrogen atoms

Experimental Method

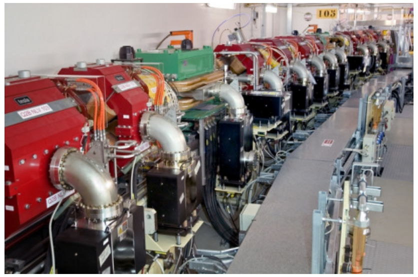
- Fundamental in science
  - *The* defining feature
- Experimental methodology
  - A Subject of study
  - Controlled environmental conditions
  - Vary chosen parameters
  - Measure and take data
  - Analysis to interpret data
  - Compare with hypothesis (model)
- Data alone is useless
  - With some simple descriptive metadata
- Need full-context of the experiment
  - Restartability
  - Validation
  - Reproducability

MONITI MELIORA
The right Hon.ble Francis Lo: Veru-lam, Viscount St Alban mortuus 9 Aprilis, Anno Dni. 1626. Annoq Aetat 66.

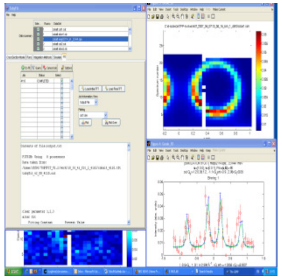
Science & Technology
Facilities Council

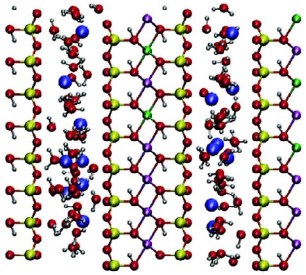
The science we do - Structure of materials

Visit facility on research campus


Place sample in beam

Diffraction pattern from sample


Fitting experimental data to model


Structure of cholesterol in crude oil

- A particular view on what an “experiment” is
  - Structural determination of materials
    - Possibly multiple runs, multiple techniques
  - Compared and contrast with computational models
  - Increasingly dynamics
  - May be used in a wider context
    - E.g. Drug candidates
- May differ from other views of experiments
  - Observations and measurements
  - Longitudinal studies
  - Etc
- But a “useful”subclass
  - And may be generalisable (?)

Science & Technology Facilities Council

Data Management Systems

ICAT Data Management Suite

Integrated data management pipelines

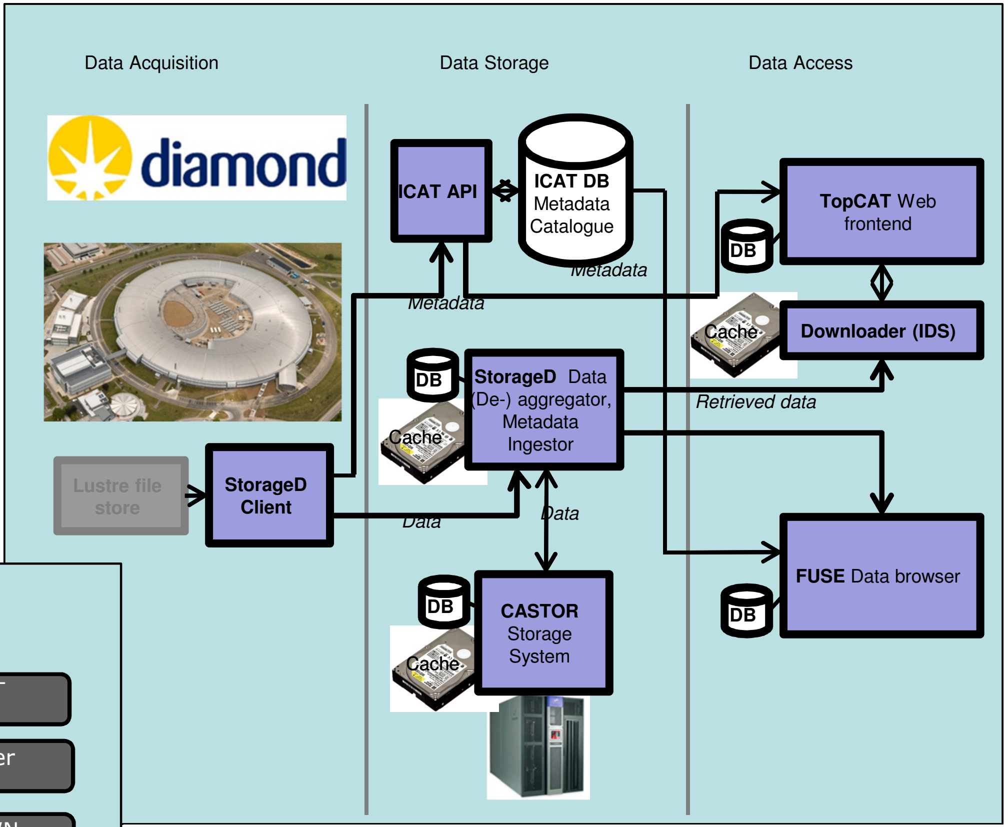
- From data acquisition to storage to publication

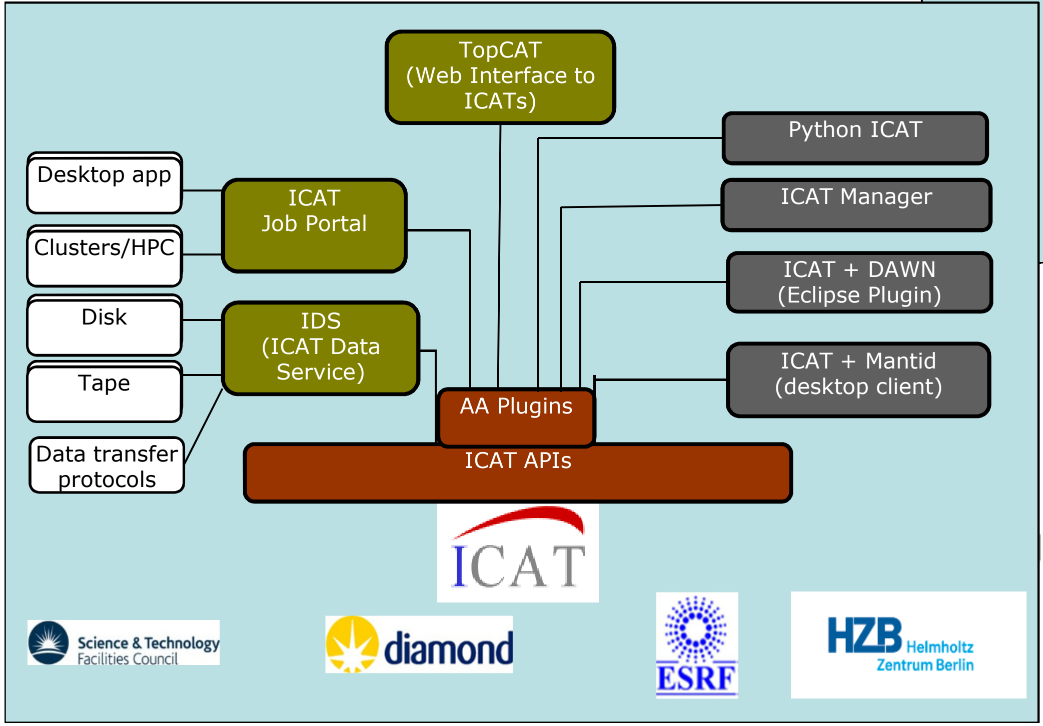
Metadata as Middleware

- A Catalogue of Experimental Data
- Automated metadata capture
- Integrated with the User Office and data acquisition system

Providing access to the user

- TopCat web front end
- Integrated into Analysis frameworks
  - Mantid for Neutrons, DAWN for X-Rays




TopCAT (Web Interface to ICATs)
Python ICAT
Desktop app
ICAT Job Portal
ICAT Manager
Clusters/HPC
ICAT + DAWN (Eclipse Plugin)
Disk
IDS (ICAT Data Service)
Tape
ICAT + Mantid (desktop client)
AA Plugins
Data transfer protocols
ICAT APIs
ICAT
Science & Technology Facilities Council
diamond
ESRF
HZB Helmholtz Zentrum Berlin

15 years effort to build data management systems

DLS Archive of

- 4.7PB, 1100 million files (Aoril 2016)
  - cf 2.2PB, 620m Jan 2015

ISIS Data Archive

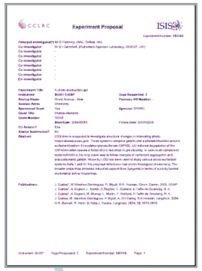
- ~50Tb
- Full experimental Metadata

ICAT Open Source Collaboration: www.icatproject.org

Facility Data Lifecycle

ICAT

Metadata Catalogue

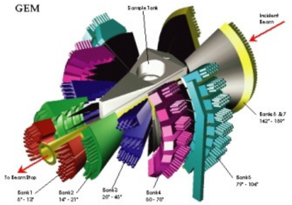


Experiment Proposal

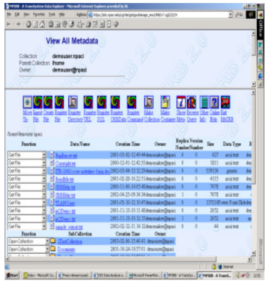
Proposal

Approval

Scheduling

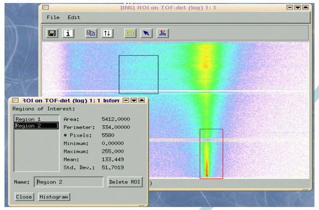


Experiment



View All Metadata

Data reduction


[IN6] ROI on TOF-det (log) 1:1
File
Edit
ROI on TOF-det (log) 1:1 Inform
Region 1
Region 2
Delete ROI
Close
Histogram

Regions of Interest:

Area: 5412.0000
Perimeter: 334.00000
Pixels: 5580
Minimum: 0.00000
Maximum: 255.000
Mean: 133.449
Std. Dev.: 51.7019

Name: Region 2

Data analysis

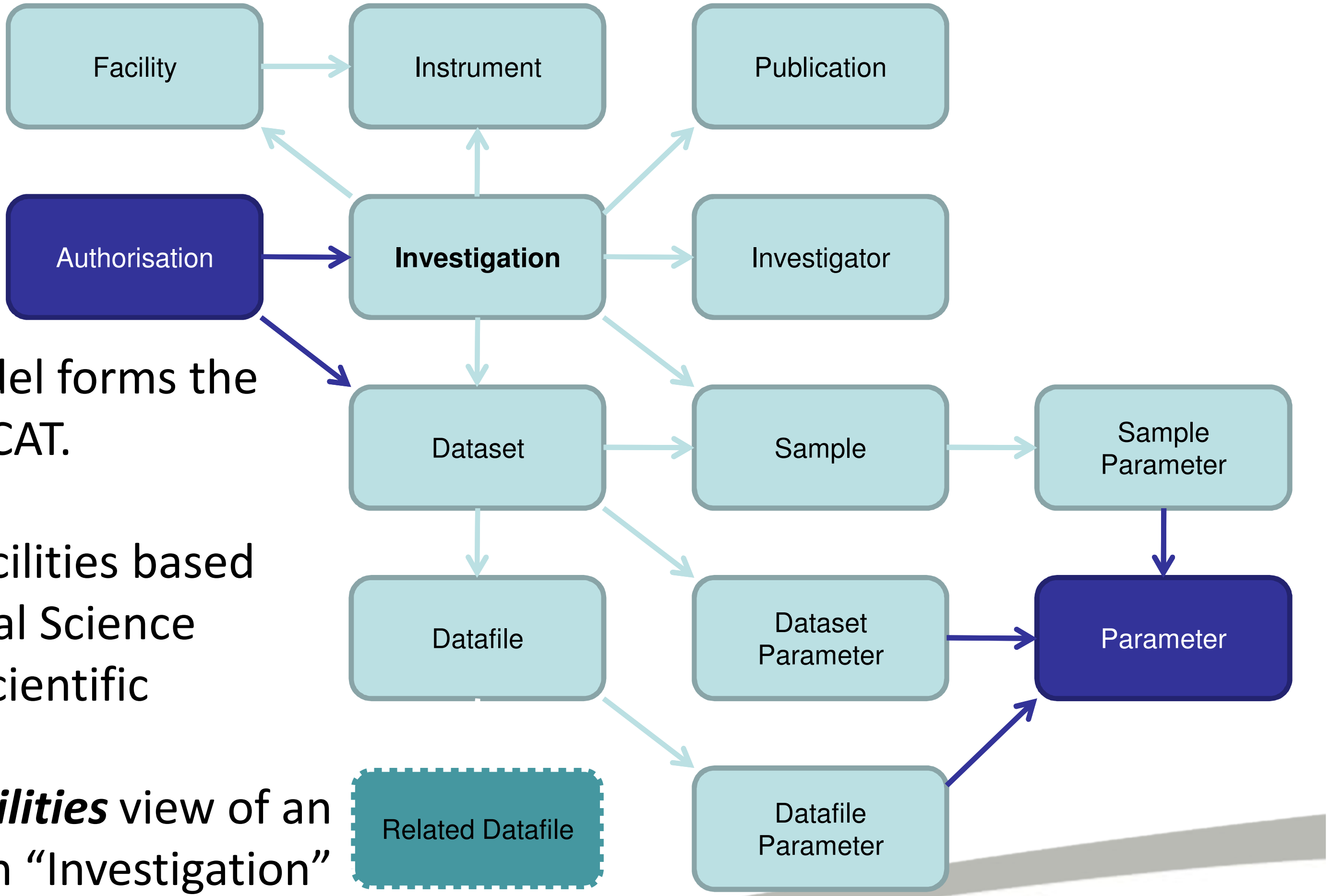
CCLRC ePublication Archive

Publication

ICAT

http://www.icatproject.org

Core Scientific Metadata Model (CSMD)



The Core Metadata model forms the information model for ICAT.

Designed to describe facilities based experiments in Structural Science throughout a facility's scientific workflow.

- Uses a ***Facilities*** view of an experiment – an "Investigation" (proposal)

For use within the repository for organising data

http://purl.org/net/CSMD
http://icatproject.org/CSMD/

nffa.eu
nanoscience foundries & fine analysis

An open access resource for experimental & theoretical nanoscience

Information and Data Management Repository Platform for nanoscience

- An integrated platform
  - ➢ covering the full research cycle by the users.
  - ➢ automatic acquisition of key metadata
  - ➢ a data repository for future data access

.

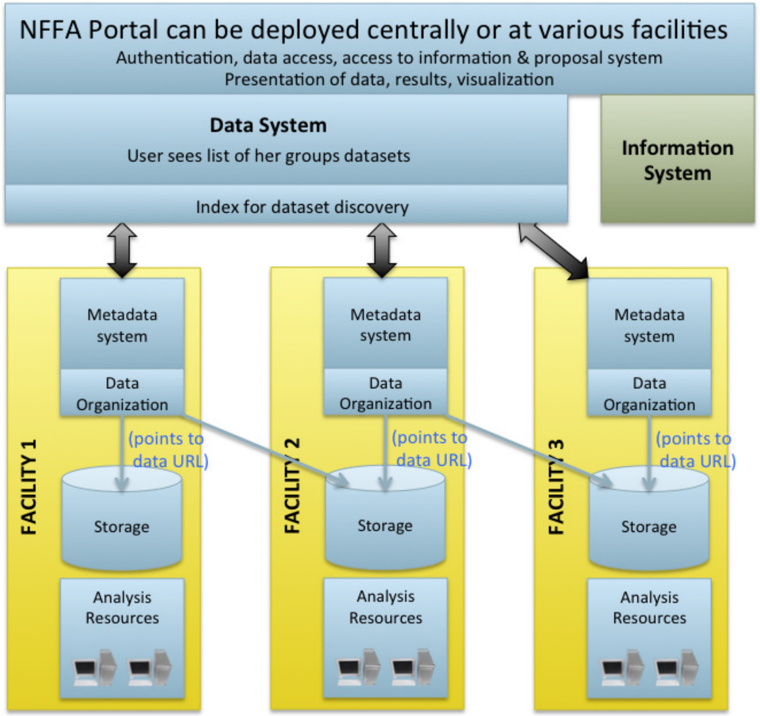
Defining metadata standards *for data sharing in nanoscience*

➢ To represent data from nanoscience experiment and theoretical analysis.

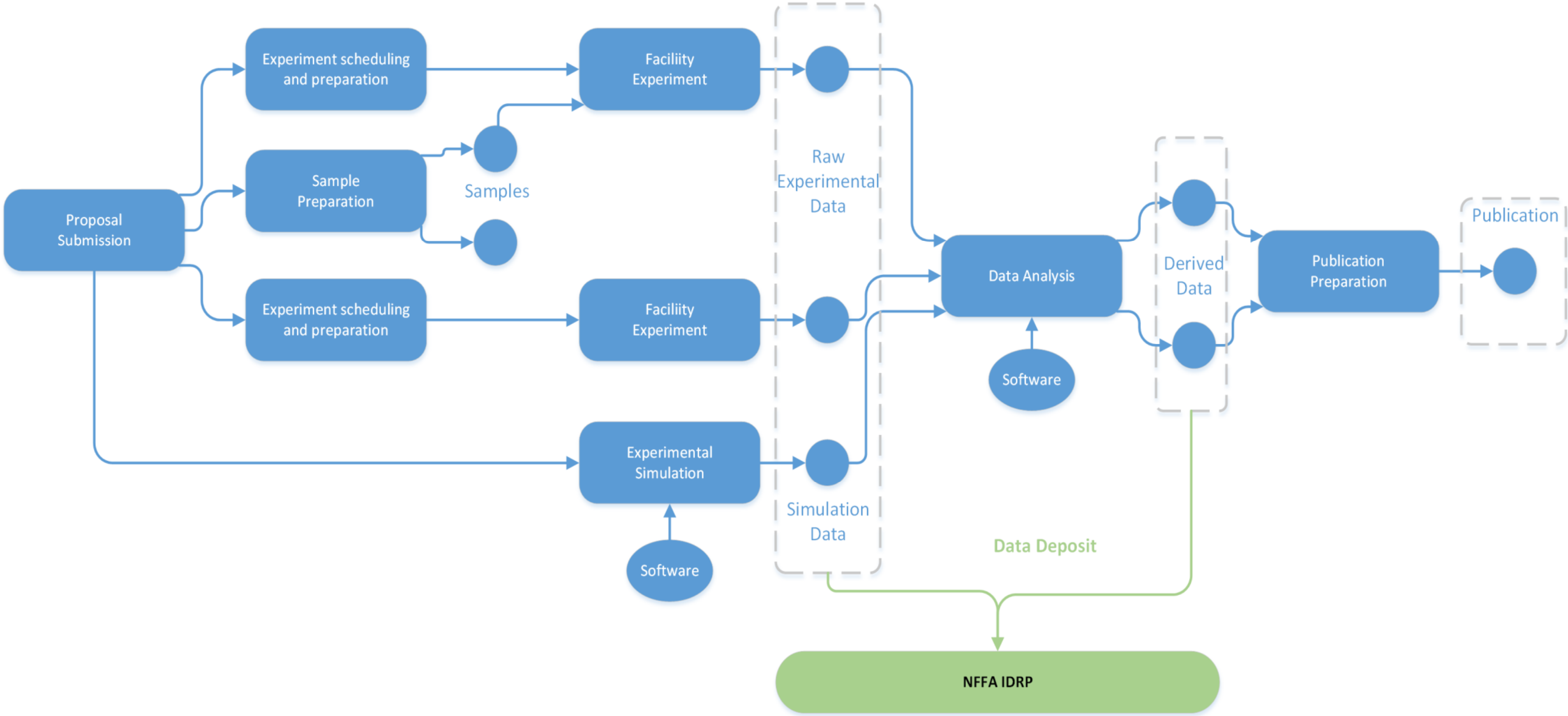
➢ Use currently available standards e.g. from PaNData project.

➢STFC, CNR-IOM, ESRF, KIT, FORTH

➢Materials IG - and International Materials Resource Registries WG


NFFA Portal can be deployed centrally or at various facilities
Authentication, data access, access to information & proposal system
Presentation of data, results, visualization
Data System
User sees list of her groups datasets
Index for dataset discovery
Information System
FACILITY 1
FACILITY 2
FACILITY 3
Metadata system
Data Organization
(points to data URL)
Storage
Analysis Resources

Metadata for Nanomaterials Data



- Workflow for Nano-structured Science
- Metadata focussed around the Project
  - A user centred view
- NFFA Deliverable 11.2: Draft Metadata Standard
  - 29th February 2016

Core vocabulary for Entities

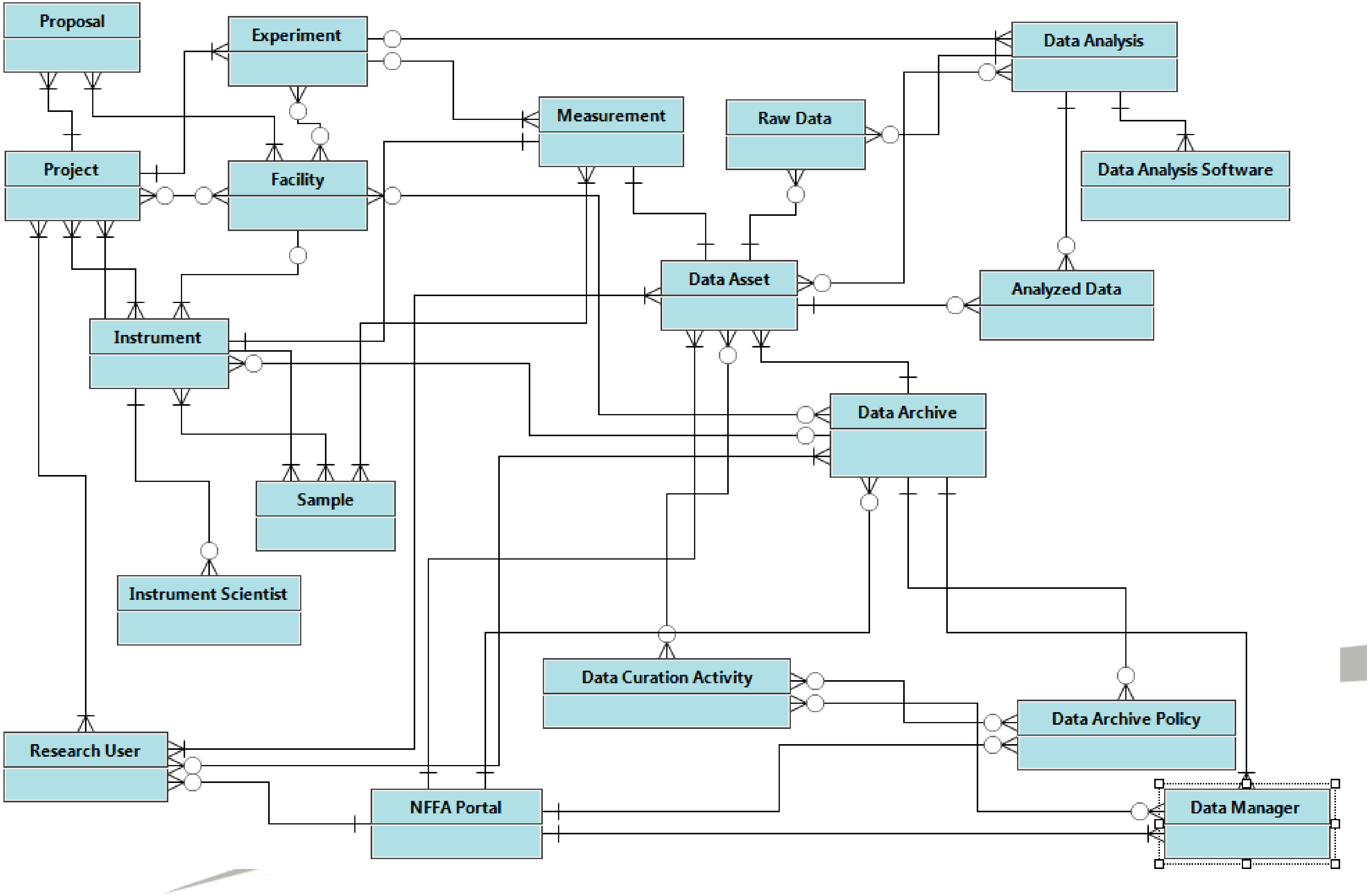
Experiment Concepts

- Research User
- Instrument Scientist. .
- Project
- Proposal
- Facility
- Instrument
- Experiment
- Measurement
- Sample

Data Concepts

- Raw Data
- Analyzed Data
- Data Asset
- Data Analysis
- Data Analysis Software
- Data Archive
- Data Policy
- Data Manager
- Data Curation Activity

Relations between Entities



Not just us of course: Chemical Process Description

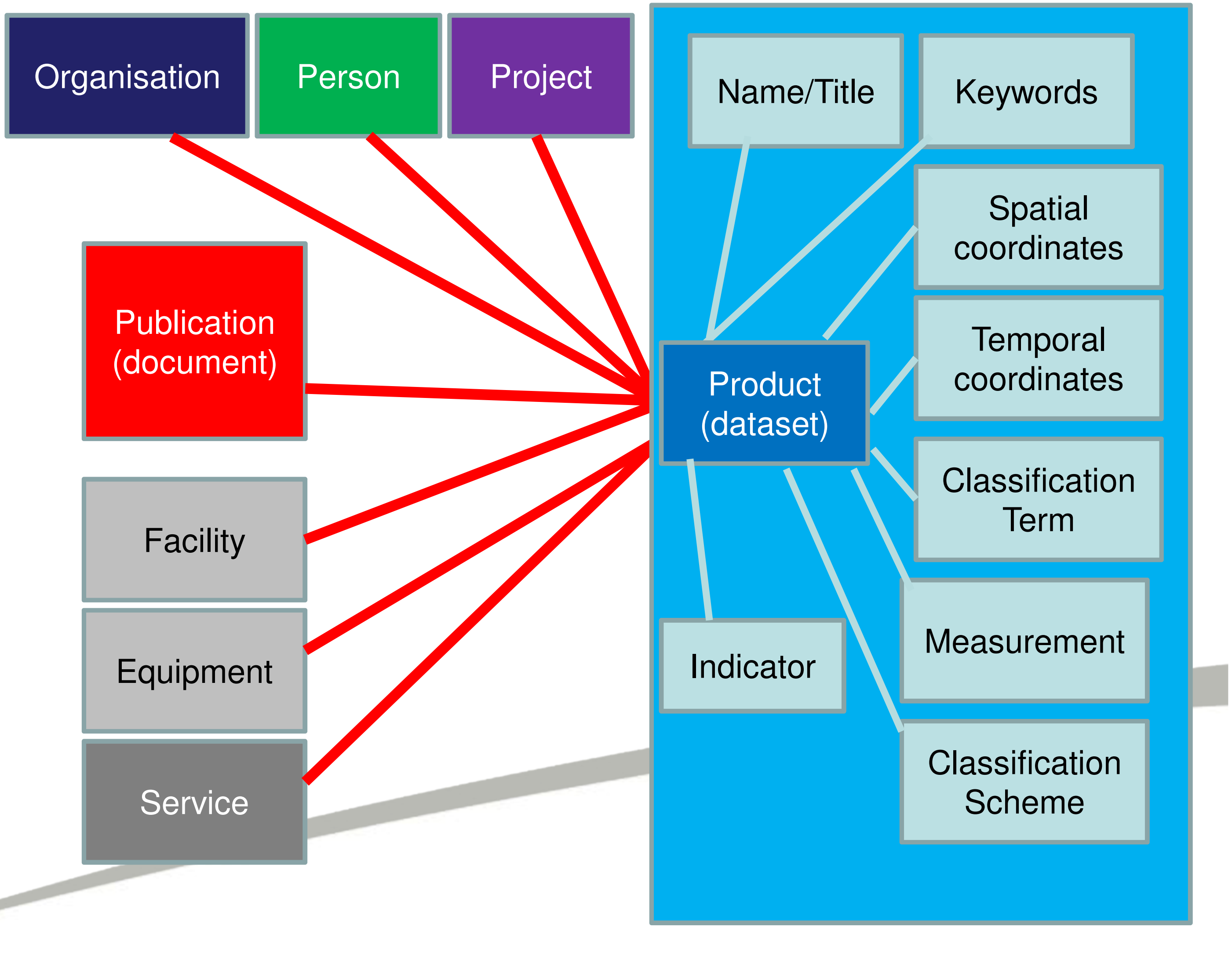
- Experimental process
- Measurement parameters
- Sample description/ preparation
- Observation/outcome description
- Data analysis
- Reaction transformation
- Equipment/apparatus
- Laboratory/environmental parameters

- Metadata used in data models (e.g.,oreChem)
- XML standards (e.g., AnIML, S88)
- Methods ontologies (e.g., ChMO)
- Analytical terminology (e.g,IUPAC Orange Book)
- Incident analysis (e.g., BowTie)

RDA Metadata IG Common Concepts

- RDA Metadata IG
- Defining common concepts
  - Straw man
- Need to agree good definitions
  - Take into account models of experimental science

Keith Jeffrey



For RDA

- FAIR: Interoperability, Reusability
  - Entities in a Core Metadata Vocabulary
    - Agreed definitions
  - Nature of Relationship between entities
  - Base Attributes for all Entities
- Based on models for research processes
  - General enough to be in common
  - Specific enough to be useful
- Role of Pids
  - Pids for everything!
- Relationships to other metadata
  - Provenance, Preservation …