CRISP WP 17
1 / 2

Proposed Metadata Catalogue Architecture Document

Work package 17 - IT & DM:
Metadata Management and Data Continuum

- Objectives:

  choose, implement data management and metadata mining services and establish an environment permitting a data continuum from raw data to publications across the participating Research Institutes (RIs): ILL, ESRF, SLHC and EuroFEL.

- Task plan:

1) Evaluate and adapt metadata catalogues according to the RIs requirements.
2) Deploy and integrate metadata catalogue
3) Prototype of data mining on metadata services.

Evaluate metadata catalogues: Use cases

- Identified a list of requirement based on ILL, ESRF and DASY use cases.
- Select a list of most suitable metadata catalogue system on the market.

- Match the requirements with features proposed by the metadata catalogues.

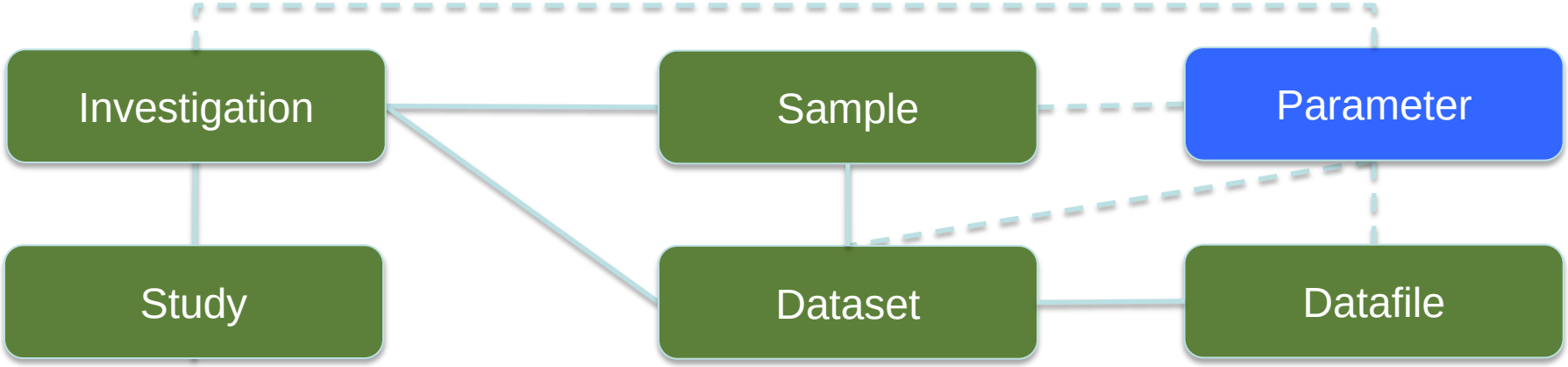
Evaluate metadata catalogues: Requirements

1) AAA
   1. Authentication
      Modular integration of different authentication systems.
   2. Authorization
      Customizable access control system.
   3. Accounting
      Granular logging information levels.

Evaluate metadata catalogues: Requirements

2) Metadata Model

Core Scientific Metadata Model (CSMD) already been developed at STFC



Evaluate metadata catalogues: Requirements

3) Searching method

   Fulfill user's search needs, being easy to use and to access (web).
   Provide data mining to Facilities and Scientific management about data use/access/search/modific.

4) Cross platform

5) Service API

   Stable set of API possibly programming language agnostic.

Evaluate metadata catalogues: Requirements

6) Sustainability
   1. Open source
   2. Project organization:
      Actively maintained, Release plan (documentation, update mechanism, backward comp.), Patch release process (security, bug fix)
   3. Cutting edge Technology
7) License
   Free of charge

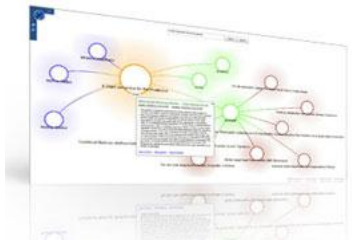
Evaluate metadata catalogues: Requirements

8) Data policy
   Dynamic authorization system.
9) Scalability & Performance
   ILL host ~2'000 experiment /year producing ~10'000 datasets. Other facilities possibly more…
10) Data ingestion
   Manually & automatic + possible harvest (OAI-PMH)
11) Security
   Protect intellectual property.

Evaluate metadata catalogues:

Metadata catalogue systems

1. ICAT
2. Dspace
3. Fedora
4. Ckan
5. Invenio
6. Tardis
7. ISPyB
8. iRODS
9. SRB-MCAT
10. MS. Zentity



Evaluate metadata catalogues: Selection result

- Different solutions have been explored, amongst them ICAT appears to be the only one that currently fits the Data Model requirements. This is the key element for a successful implementation in a reasonable time frame.

Evaluate metadata catalogues: ICAT

- Authentication plug-in
- Rule based authorization mechanism
- Flexible metadata model
- Search method: full-text, numerical and string search and SQL like query syntax
- Set of API (Java and Python)
- Database configurable (Oracle, Posgres and MySQL)
- Federated search via TopCAT
- Core Scientific Meta-Data Model (CSMD)

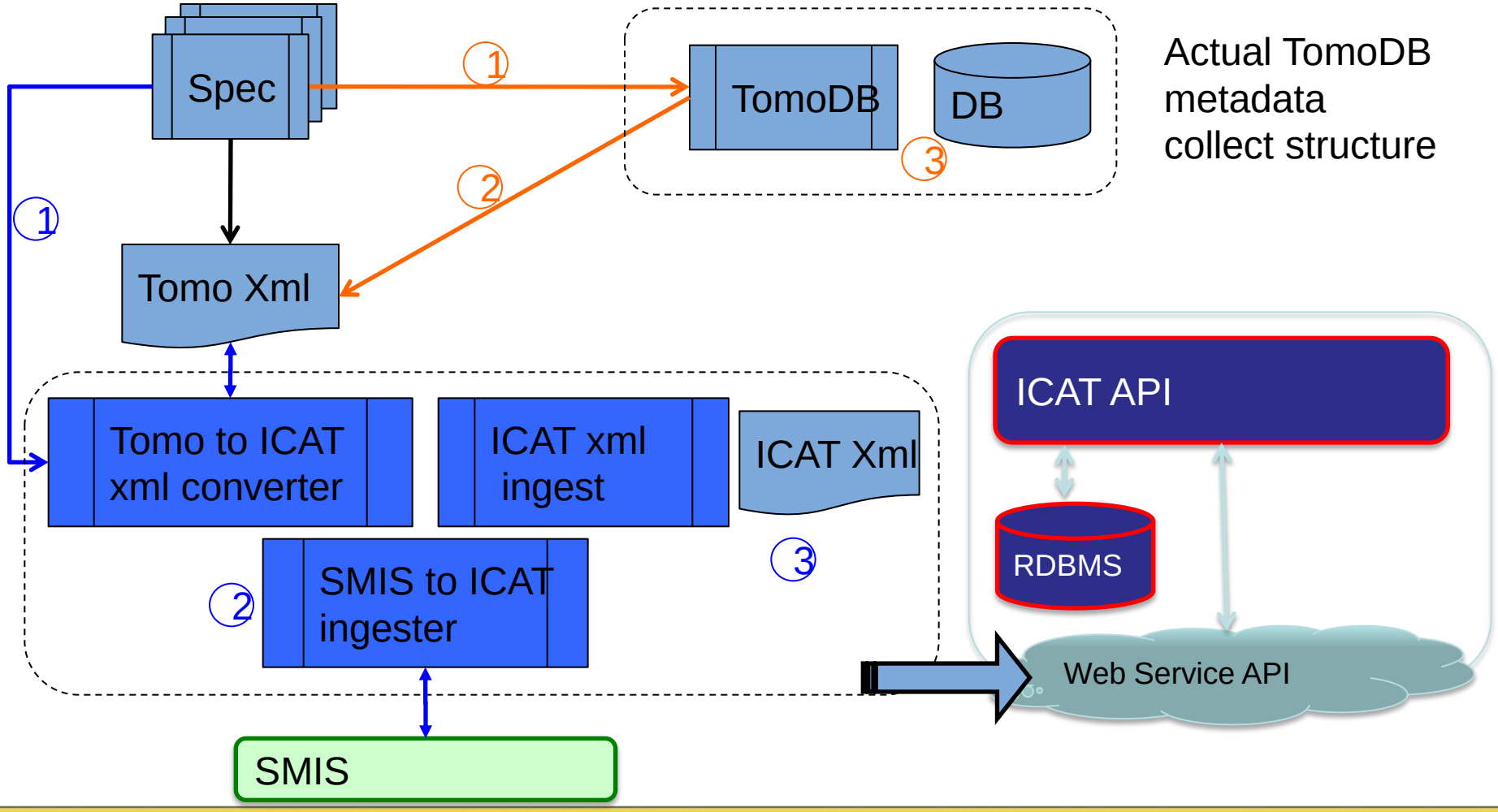
Evaluate metadata catalogues: ICAT

- Plug-in for DAWN/Mantid
- Licence: FreeBSD
- Web interface: TopCAT
- In use at 11+ RIs

Evaluate metadata catalogues: ICAT

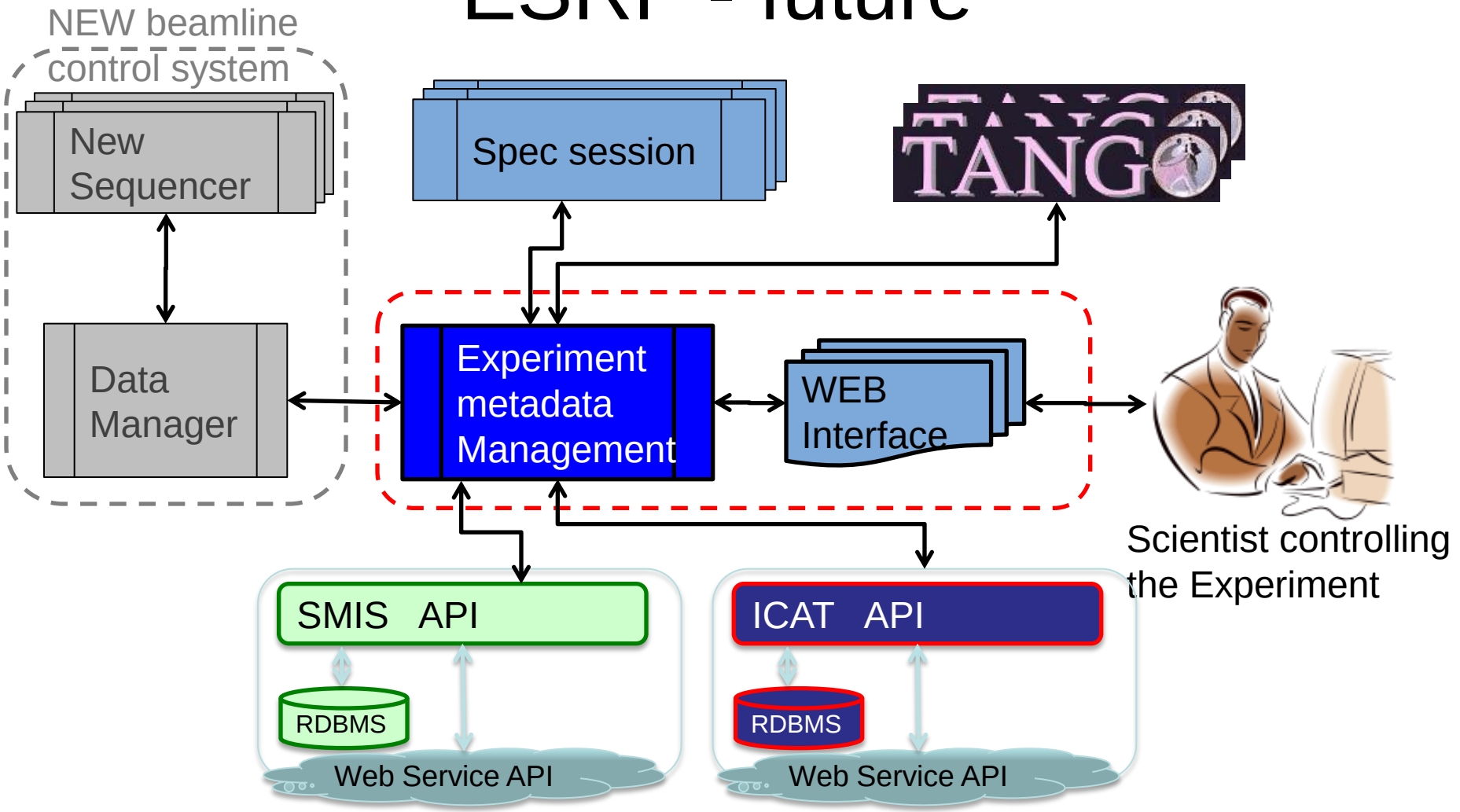
- Work-in-progress:
  - Improve web interface (TopCat)
  - Possibility to harvest (OAI-PMH)
  - Installation process
  - Synonym mechanism
  - Integration with Umbrella authentication

Deploy and integrate ICAT:
ESRF - Pilot



Actual TomoDB metadata collect structure

Deploy and integrate ICAT:
ESRF - future



Deploy and integrate ICAT:
ILL

- Data policy published in Dec 2011
- Implementation Oct 2012
- ICAT deployment Dec 2012
- Currently, ingestion of the Data since Nov 2012

Future work

- Complete the deployment (ingestion) at the participating facilities.
- Data mining
  - Collect uses cases from the different facilities
  - Currently all use cases are technically simple (no request for correlation for instance)
    - Work on the search engine (lucene)
    - Reporting


"That's all Folks!"