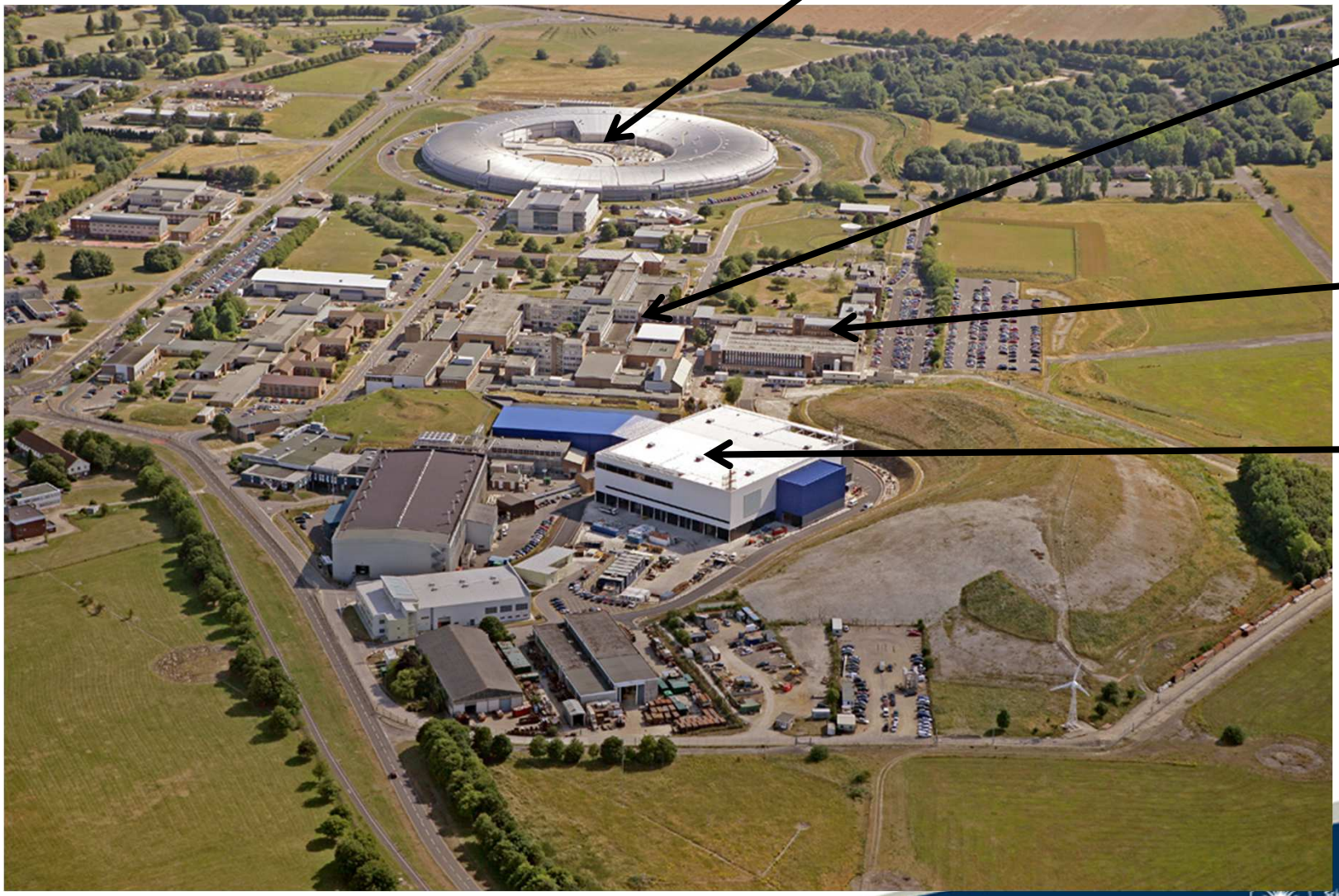
Managing risks in the preservation of research data with preservation networks

Esther Conway, Brian Matthews, David Giaretta, Simon Lambert, Nick Draper
Michael Wilson: Science and Technology Facilities Council

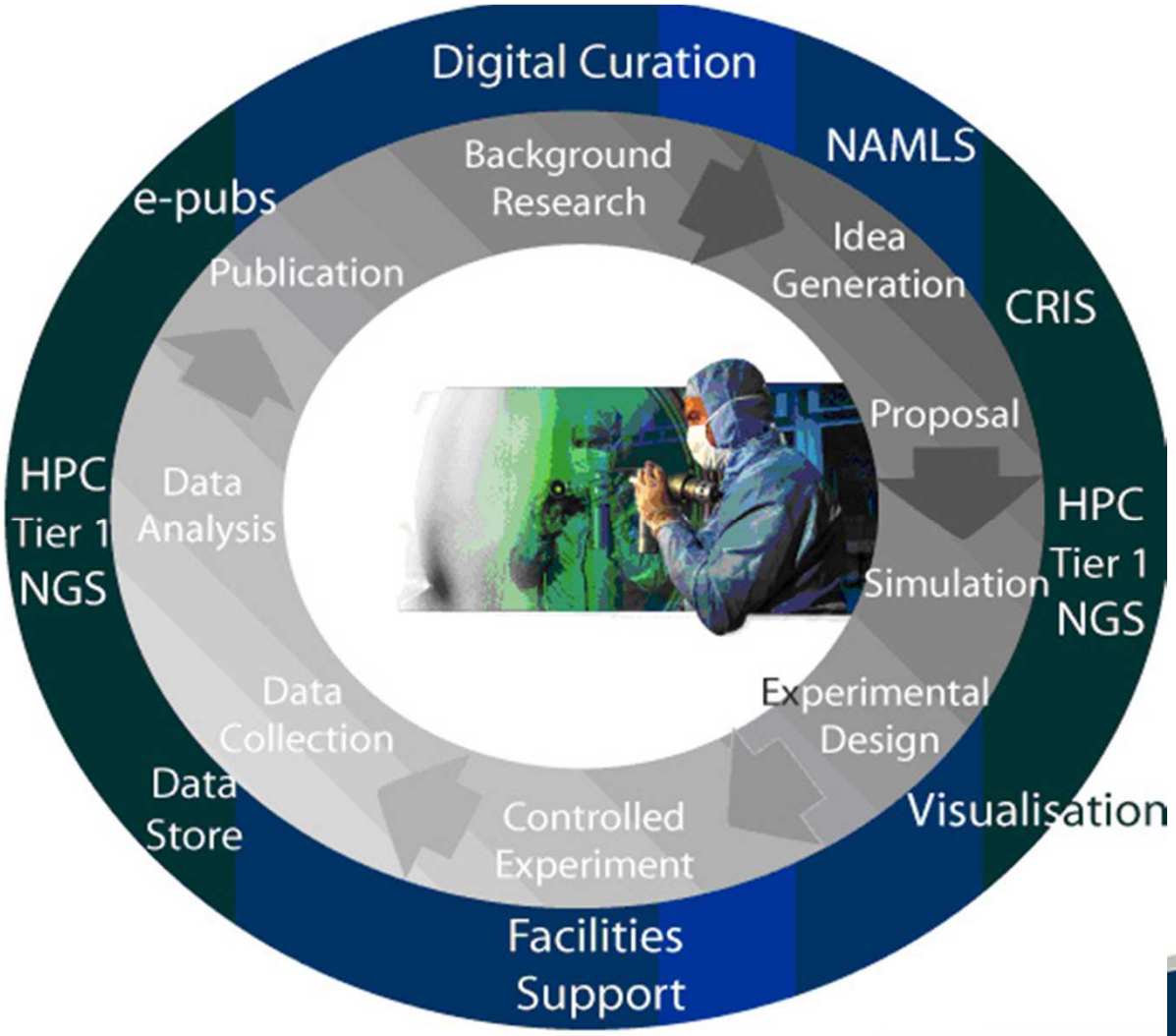
Nick Draper: Tessella

RAL



Science & Technology Facilities Council
e-Science

E-Science supports the science lifecycle



Science & Technology Facilities Council
e-Science

EU Projects

FP6

- CASPAR
- PARSE INSIGHT

FP7

- APARSEN
- SCAPE
- ENSURE
- SCIDIP

Objects

Objects are uniquely identified digital entities capable of an independent existence which possess the following attributes:

Information is a description of the key information contained by the digital object. This information should have been identified during preservation analysis as being the information required to satisfy the preservation objective for the designated user community.

Location information is the information required by the end user to physically locate and retrieve the object. Archival Information Packages (AIPs) may be logical in construction with key digital objects being distributed and managed within different information systems. This tends to be the case when data is in active use with resources moving in dynamic environment. Note the difference between being able to locate and retrieve information and having control of it.

Physical state describes the form of the digital object. It should contain sufficient information relating to the version, variant, instance or format.

A PNM terminates when a user requires no additional information or assistance to achieve the defined preservation objective, given that the accepted risks will not be imminently realised.

Relationship

A relationship captures how two objects are related to one another in order to fulfil the specified preservation objective whilst being utilized by a member of the designated user community. Relationships can possess the following attributes:

Function, in order to satisfy the preservation objective a digital object, will perform a specific function for example the delivery of textual information or the extraction and graphical visualisation of specific parameters.

Risks and Dependencies, most digital solutions will have inherent risks and a finite lifespan. Risks could include the interpretability of information, technical dependencies or loss designated community skills. Risks should be recorded against the appropriate object so they can be monitored and the implication of them being realised assessed.

Tolerance, not every function is critical for the fulfilment of the preservation objective, some objects are included as they enhance the quality of the solution or ease of use. Loss of this function is denoted in the model as a tolerance.

Quality assurance and testing, The ability of an object to perform the specified function may have been subjected to quality assurance and testing which may be recorded against the relationship.

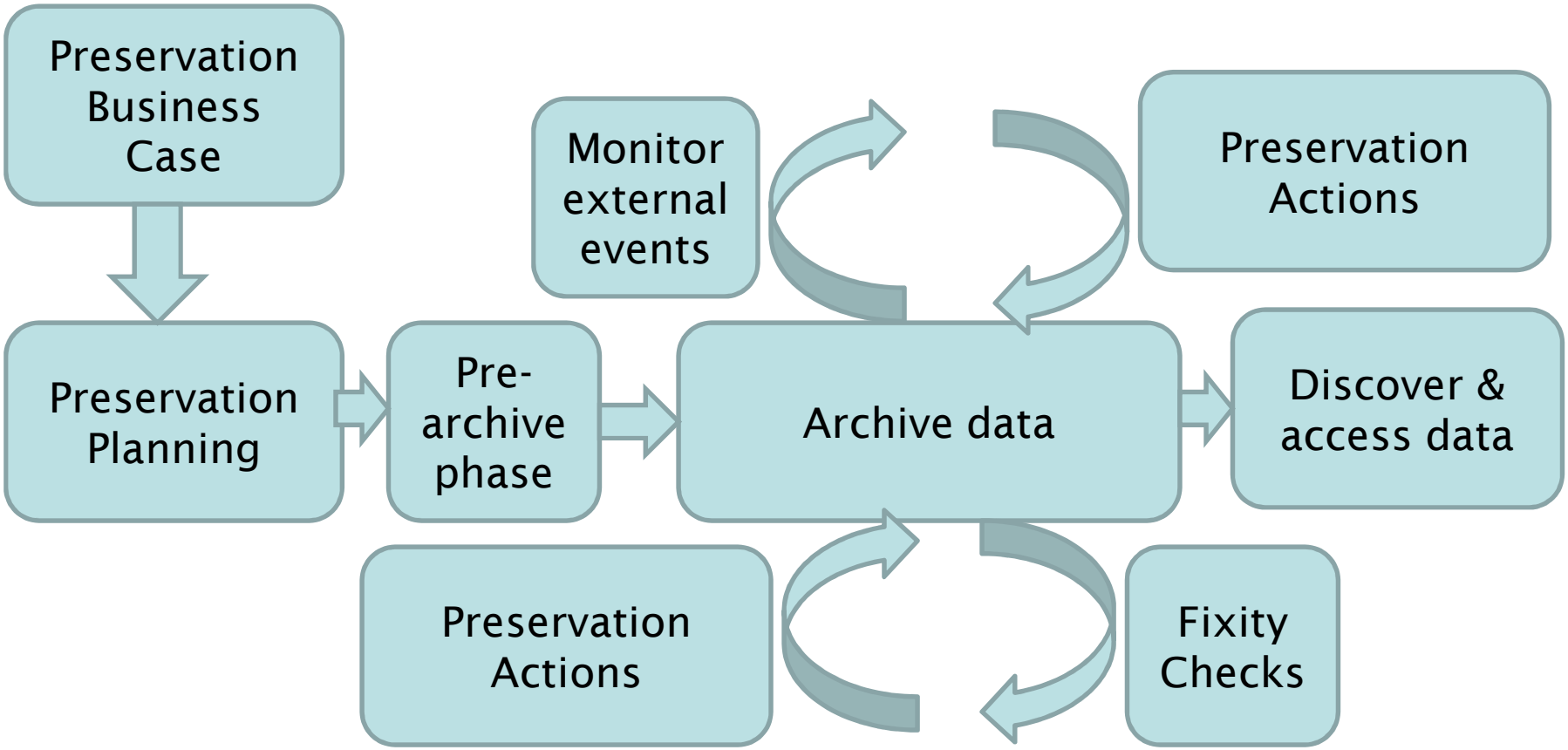
Cost and Value: Evaluate the cost and benefit of different quality solutions, enabling a business to choose the most cost effective solution.

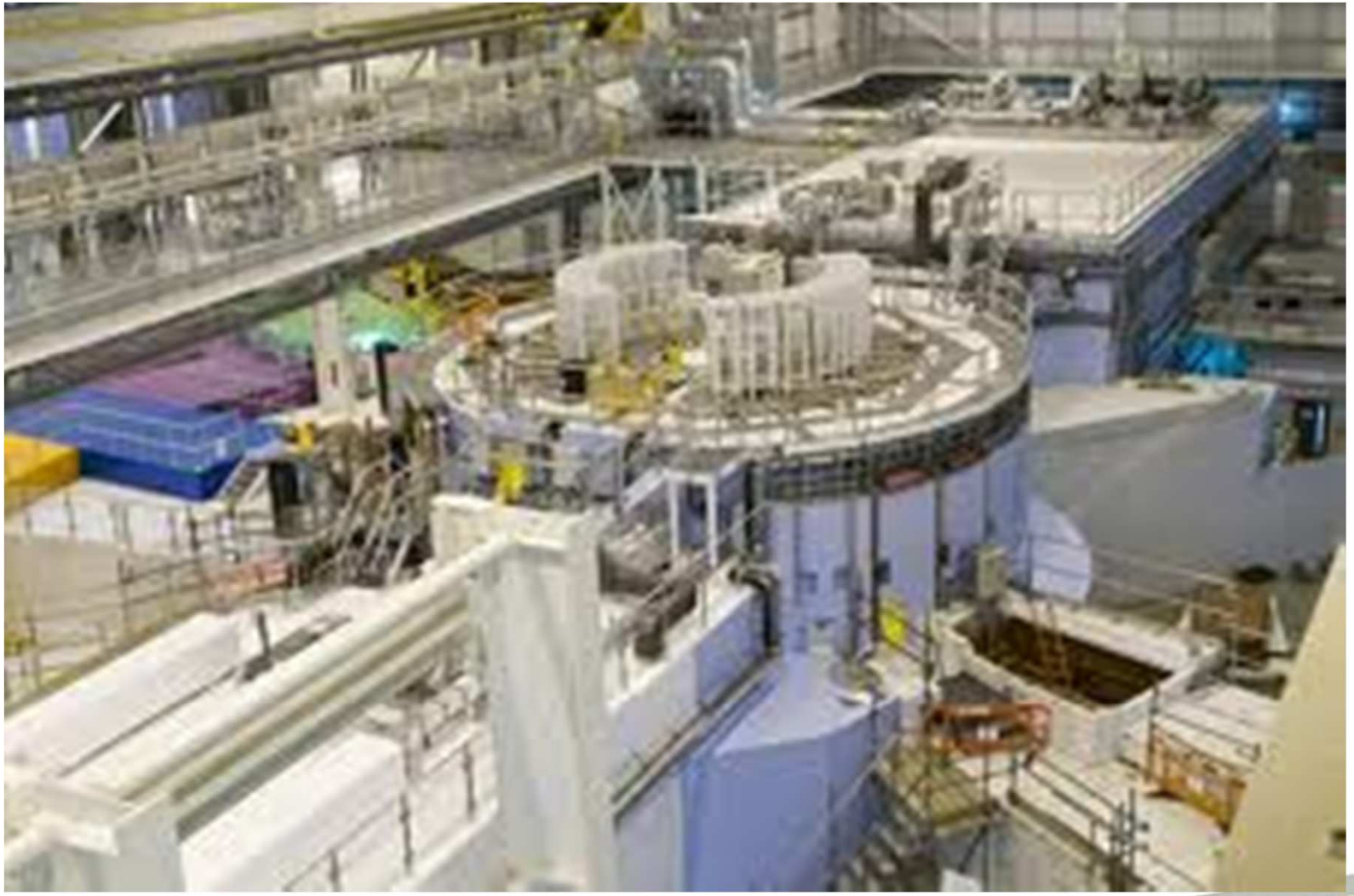
Preservation Lifecyle Management: Build on industry standard lifecycle management approaches to manage the preservation lifecycle, meet regulatory compliance, allow changes in the preservation approach to reflect environmental changes, address evolution of ontologies and manage the quality of digital objects over time.

Content-aware Long Term Data Protection: Provide data protection over long periods of time, addressing changes to personally identifiable information, new and evolving regulations, and manage user identities over the decades.

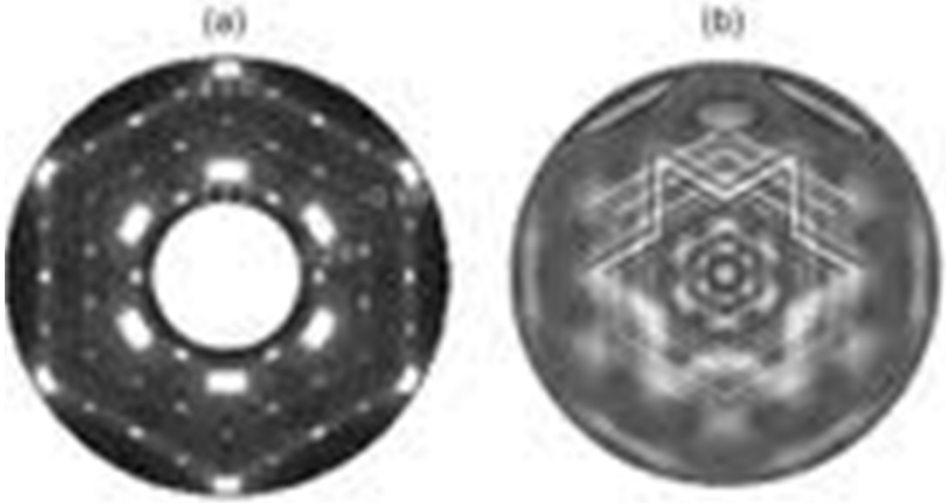
Utilize Emerging ICT: Evaluate the costs, risks and benefits and demonstrate how to use emerging, commonly available Information Technology to enable scalable solutions for digital preservation, in particular considering Cloud Storage and virtualization techniques.

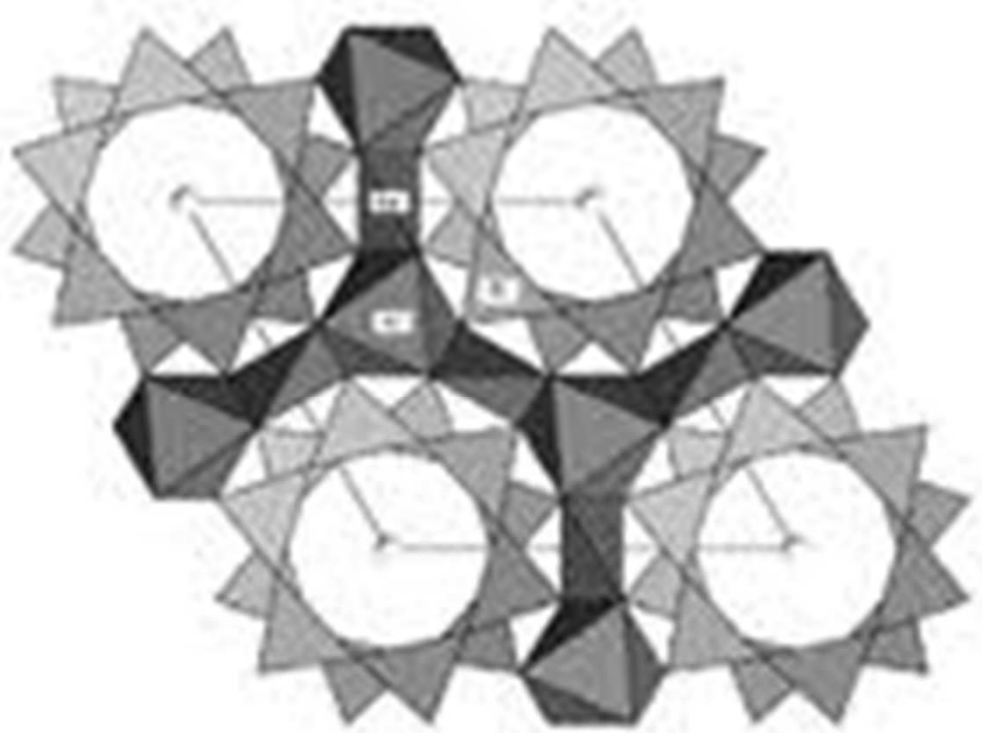
Data Preservation

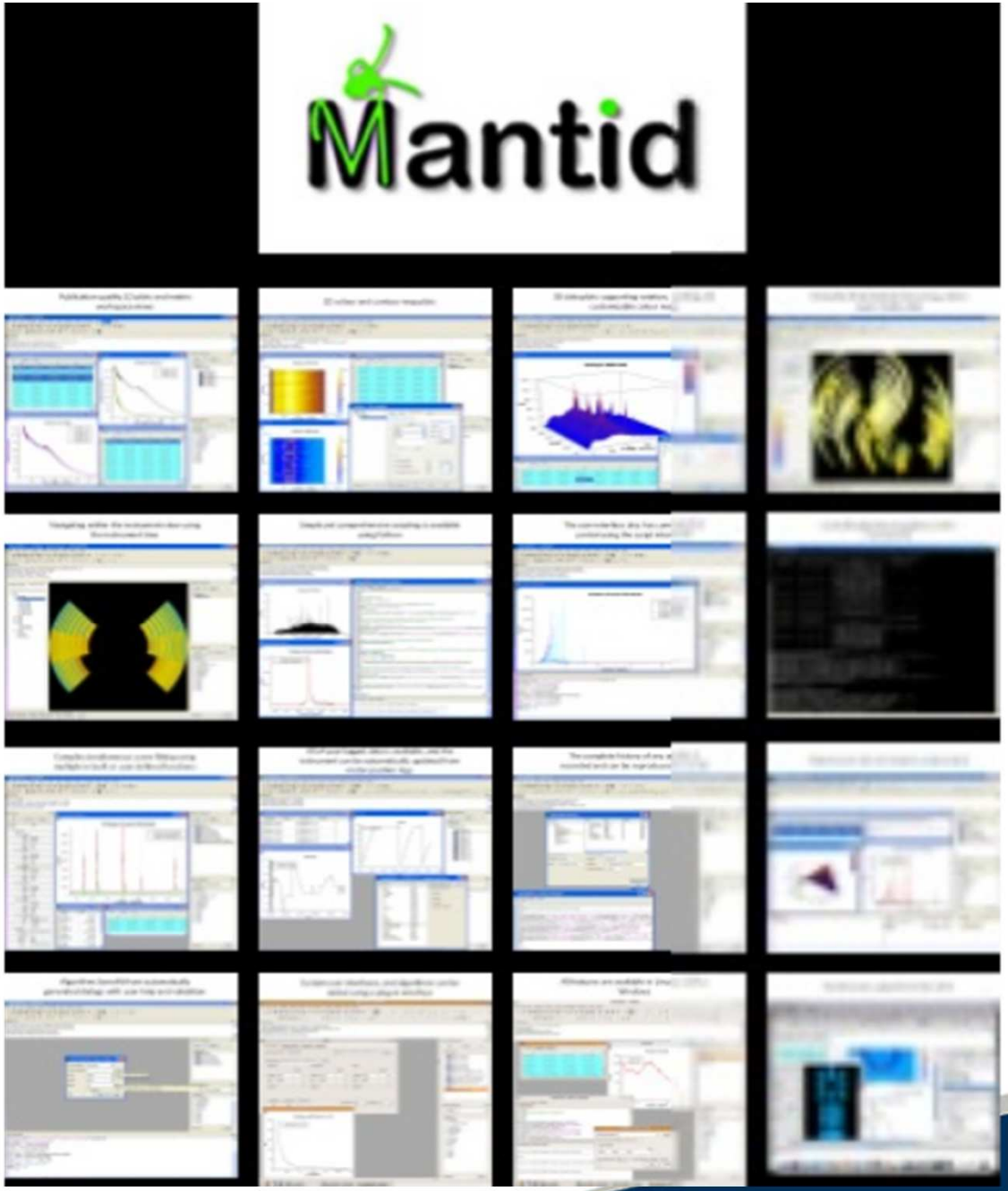




Science & Technology Facilities Council
e-Science


(a)
(b)


B
A
C

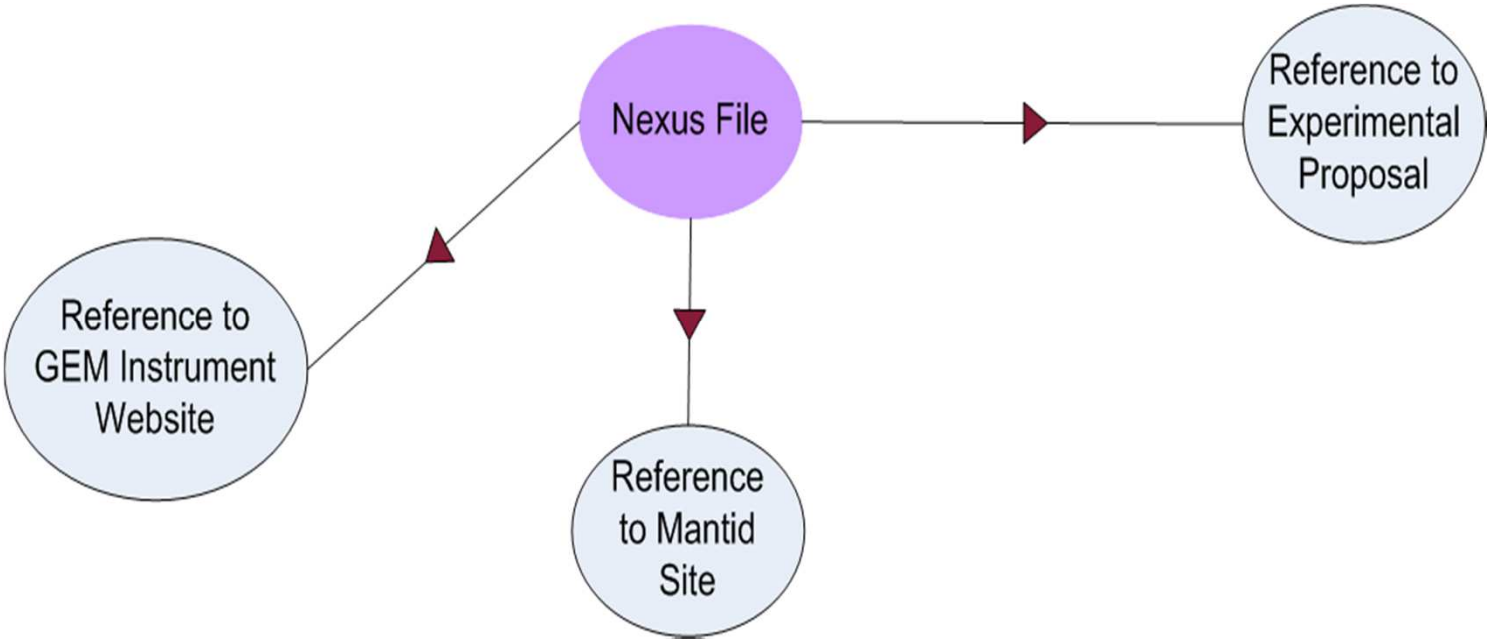


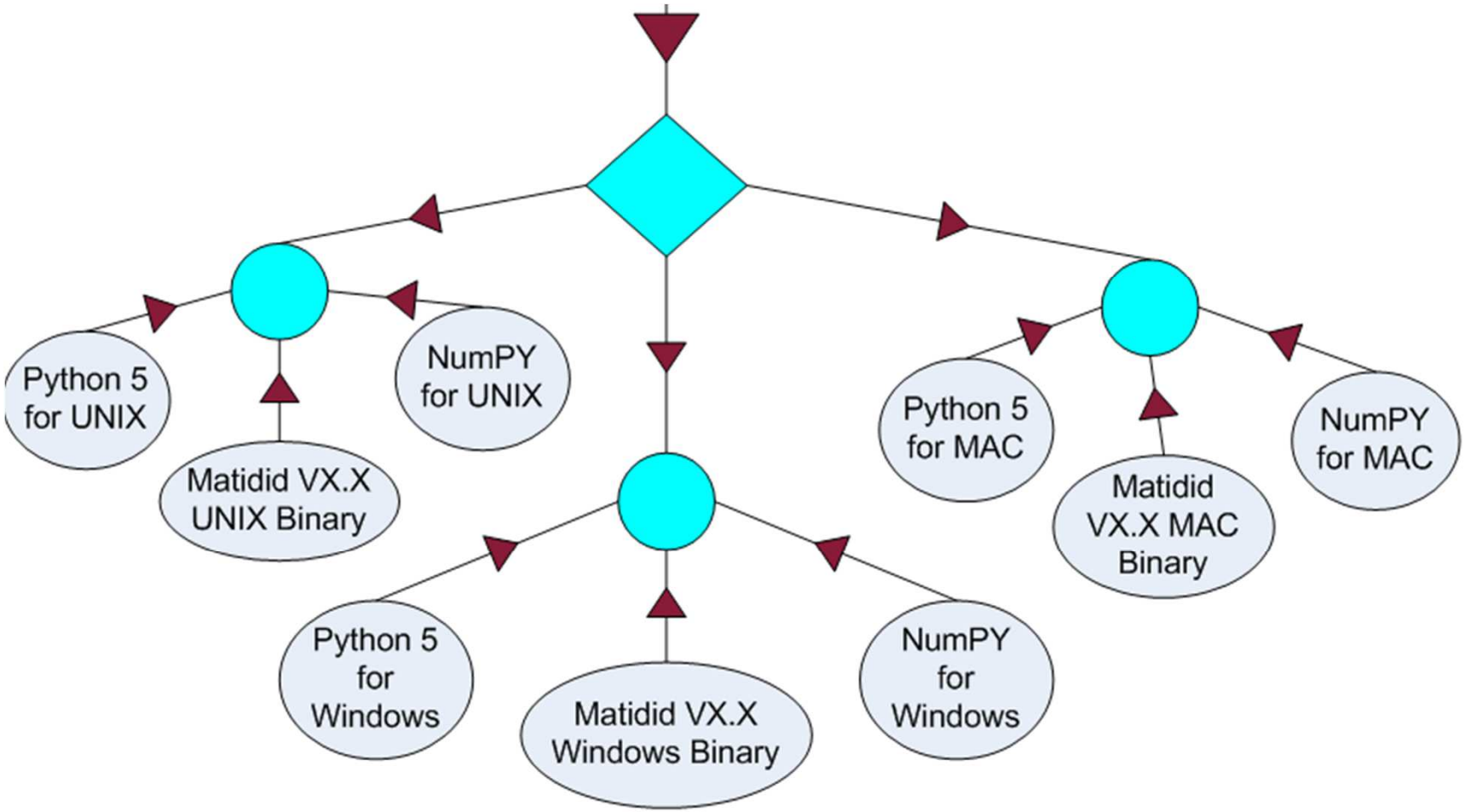
Science & Technology Facilities Council
e-Science

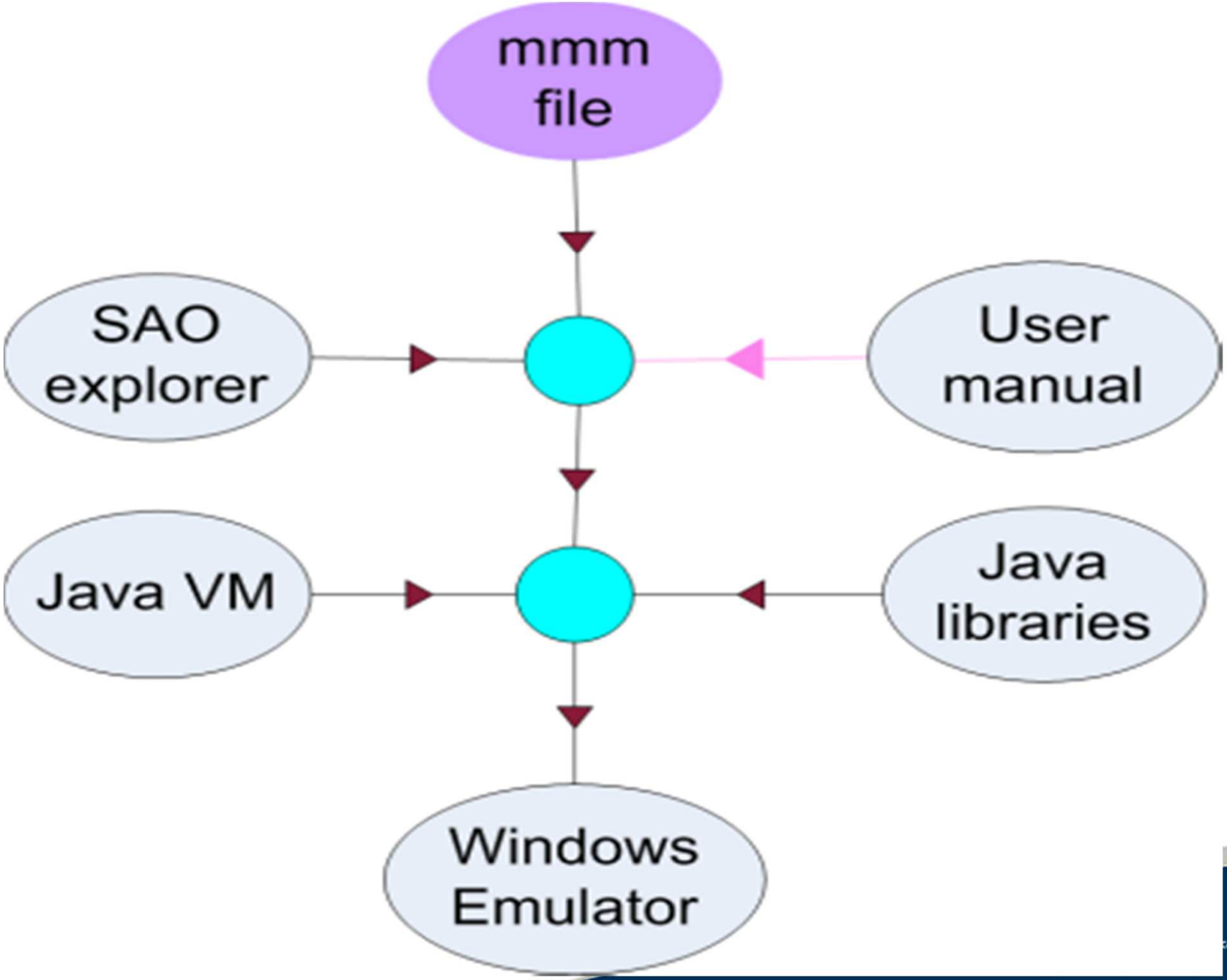
Preservation is all about control

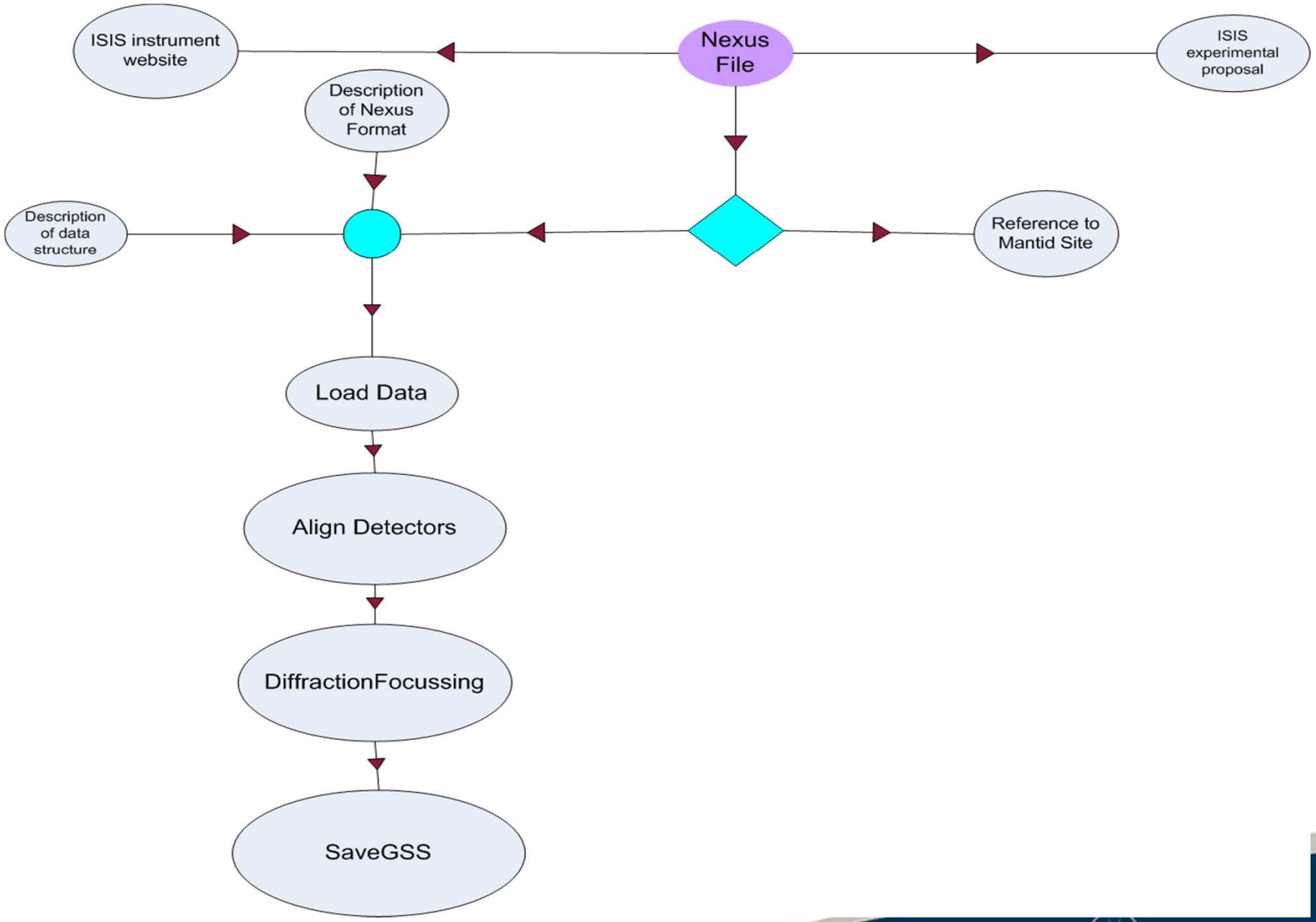
Preservation Action

- Risk Acceptance and Monitoring
- Software Capture
- Description
- Transformation


Nexus File
Reference to Experimental Proposal
Reference to GEM Instrument Website
Reference to Mantid Site


Python 5
for UNIX
Matidid VX.X
UNIX Binary
NumPY
for UNIX
Python 5
for
Windows
Matidid VX.X
Windows Binary
NumPY
for
Windows
Python 5
for MAC
Matidid
VX.X MAC
Binary
NumPY
for MAC

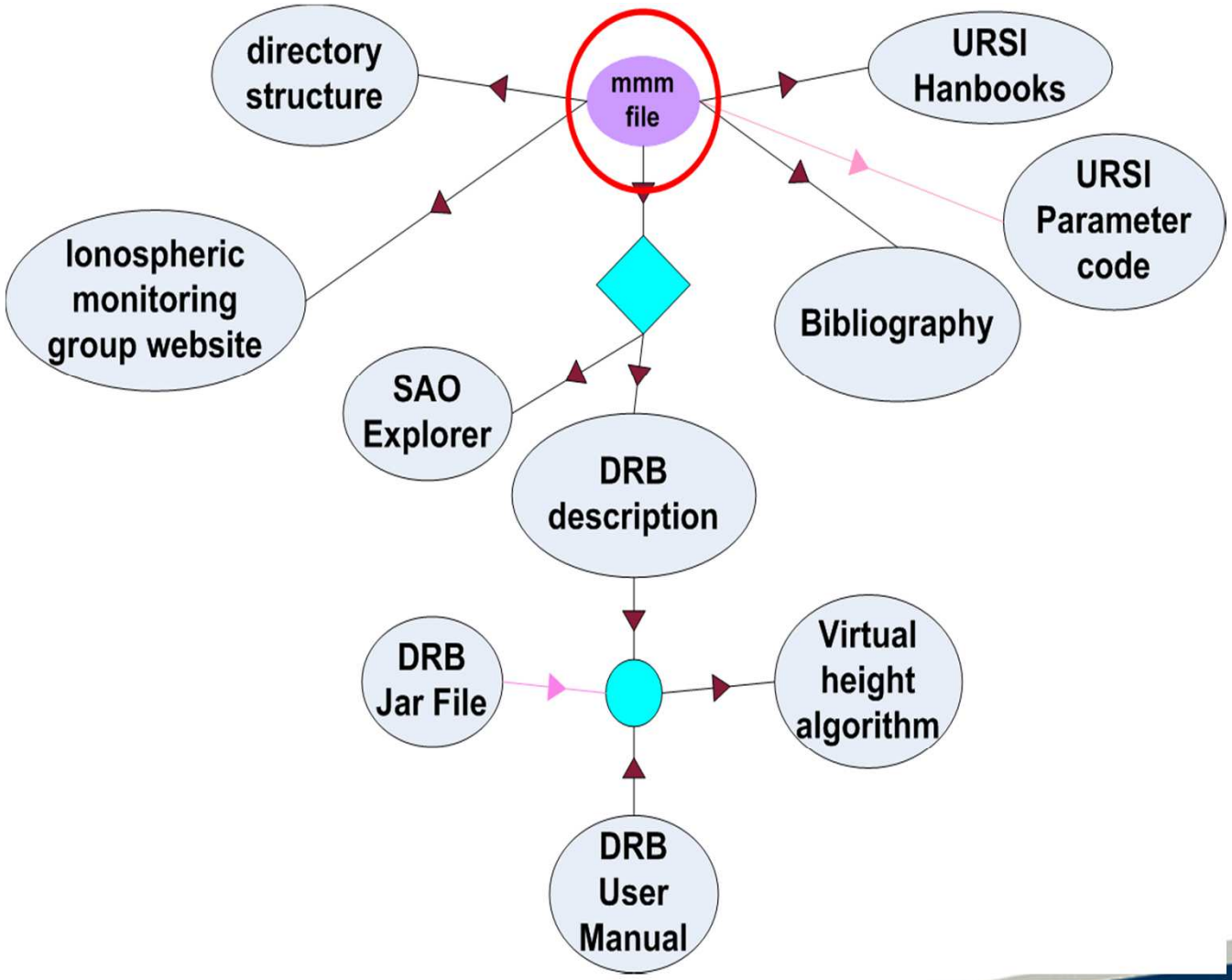

mmm
file
SAO
explorer
User
manual
Java VM
Java
libraries
Windows
Emulator

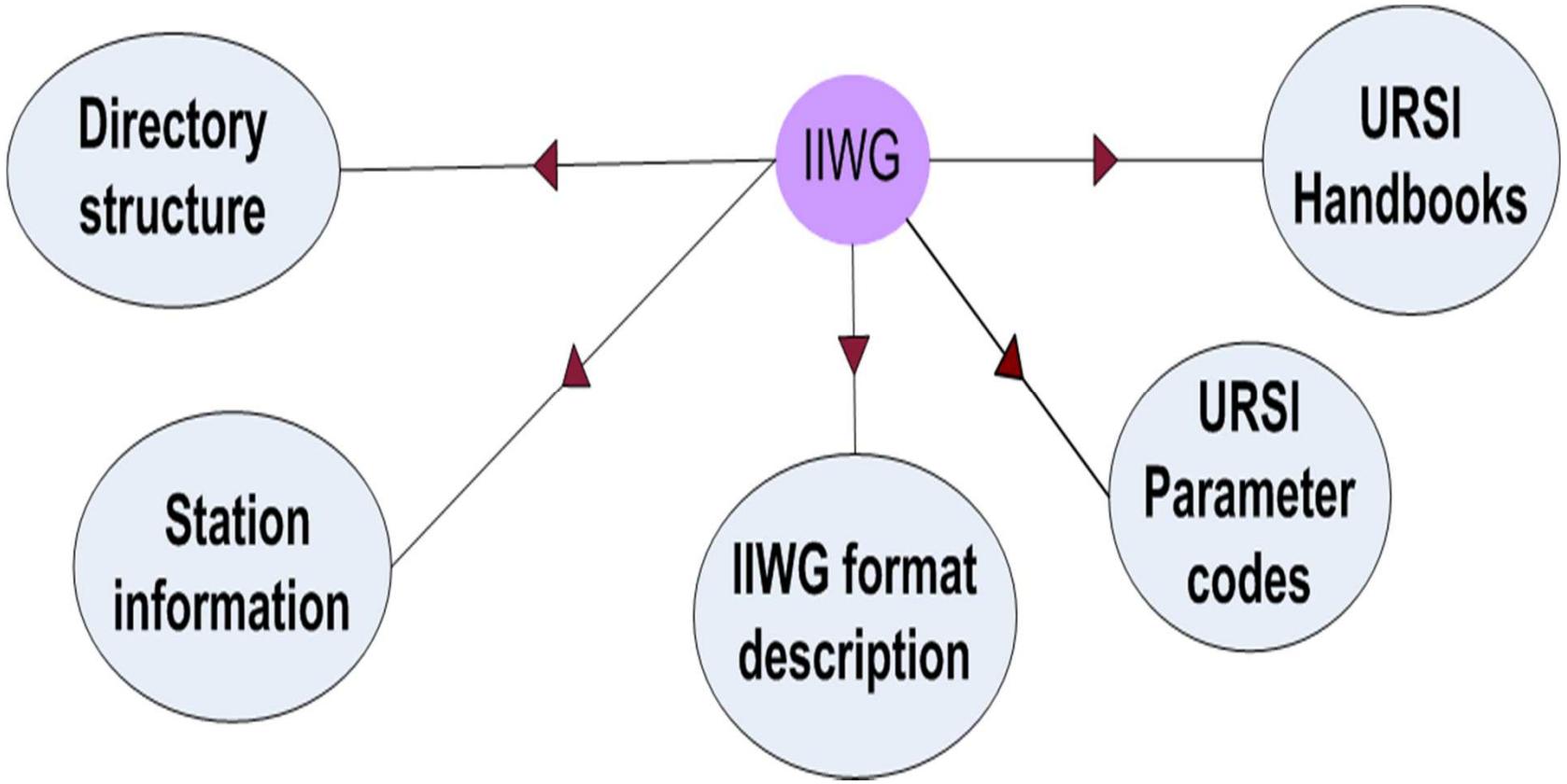

ISIS instrument website
Nexus File
ISIS experimental proposal
Description of Nexus Format
Description of data structure
Reference to Mantid Site
Load Data
Align Detectors
DiffractionFocussing
SaveGSS

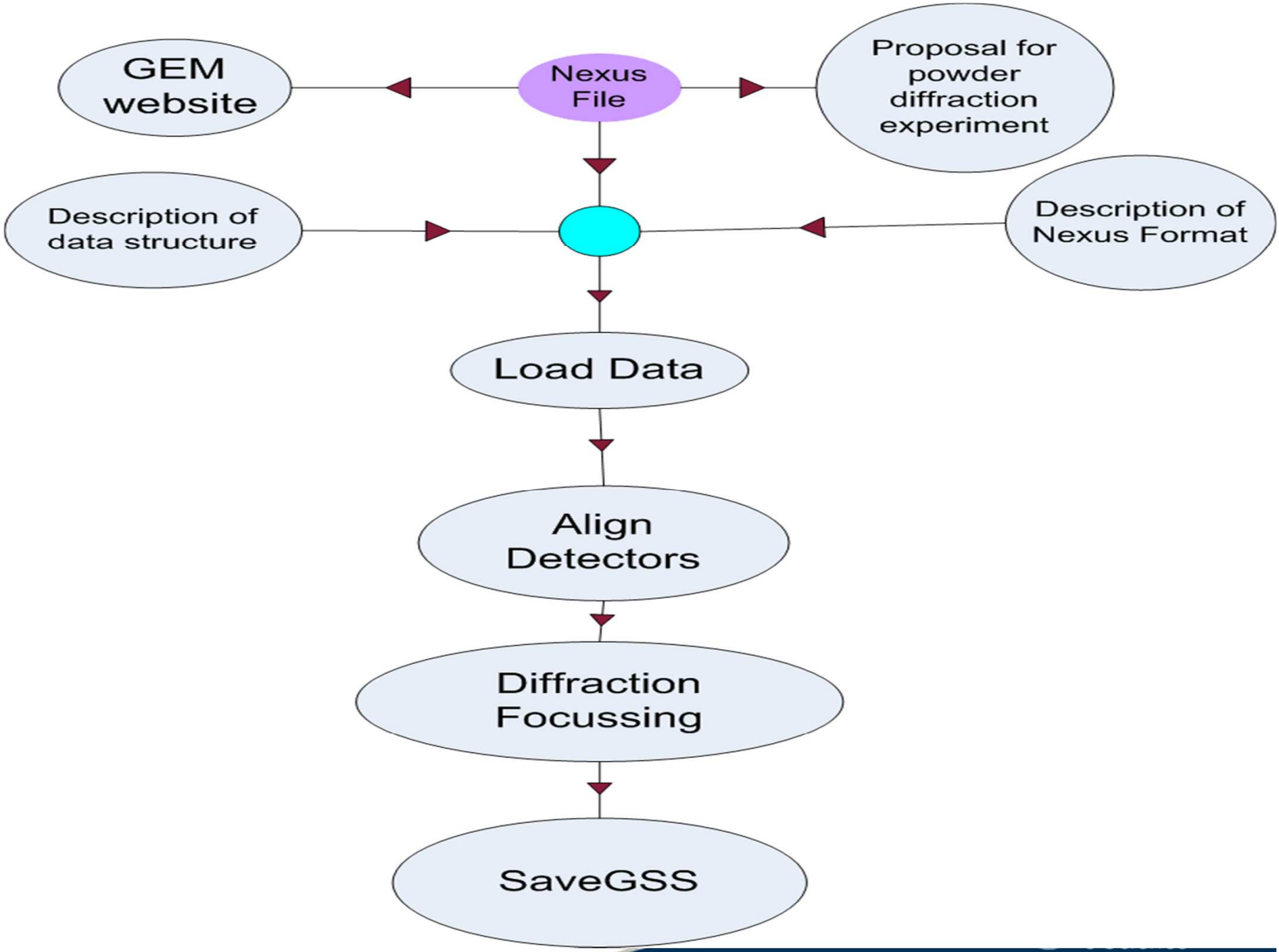
Technology Agnostic but Expensive !!!!!

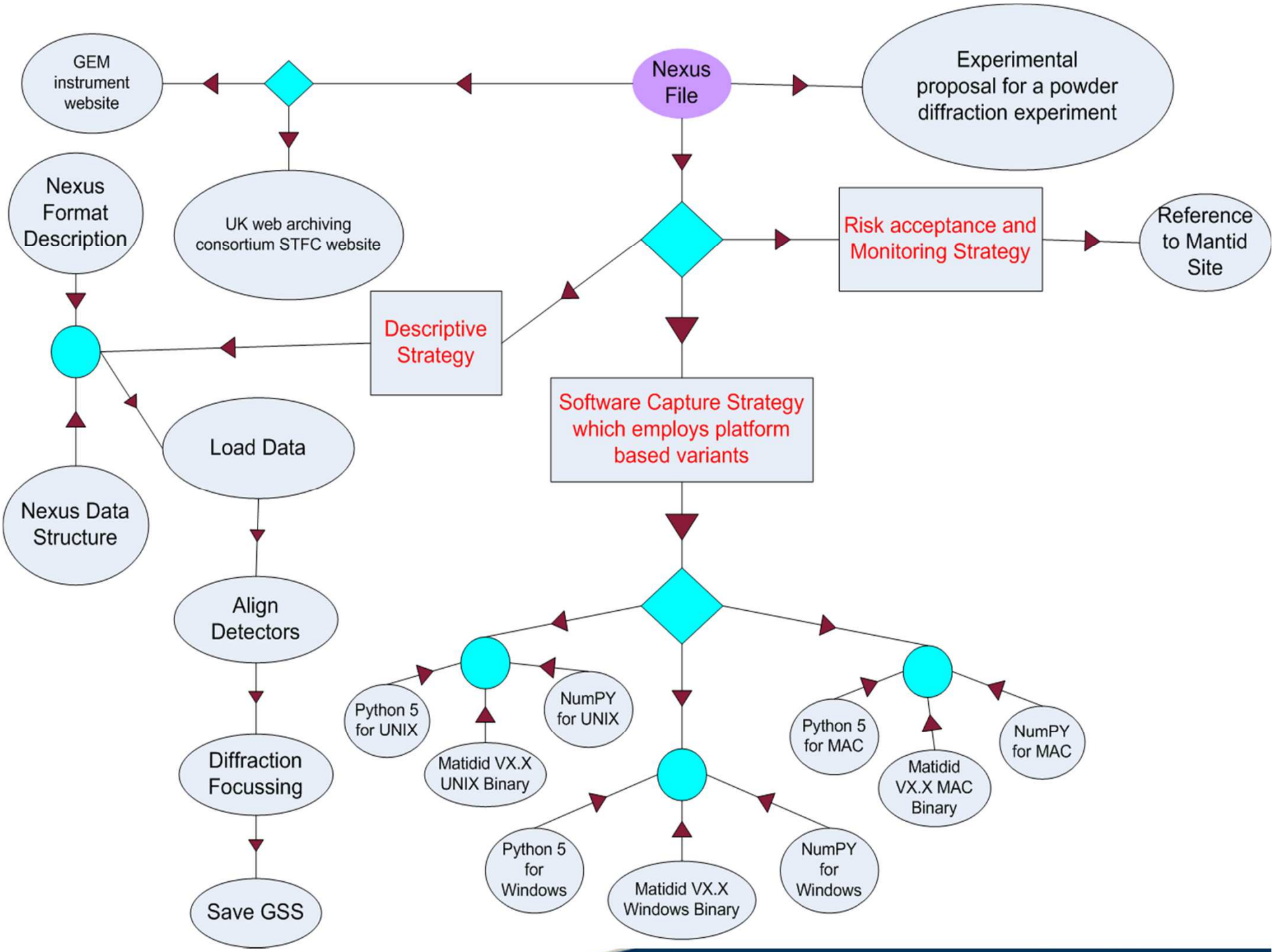

MST
NetCDF
Cartesian
Conference
proceedings
Word
CEDA

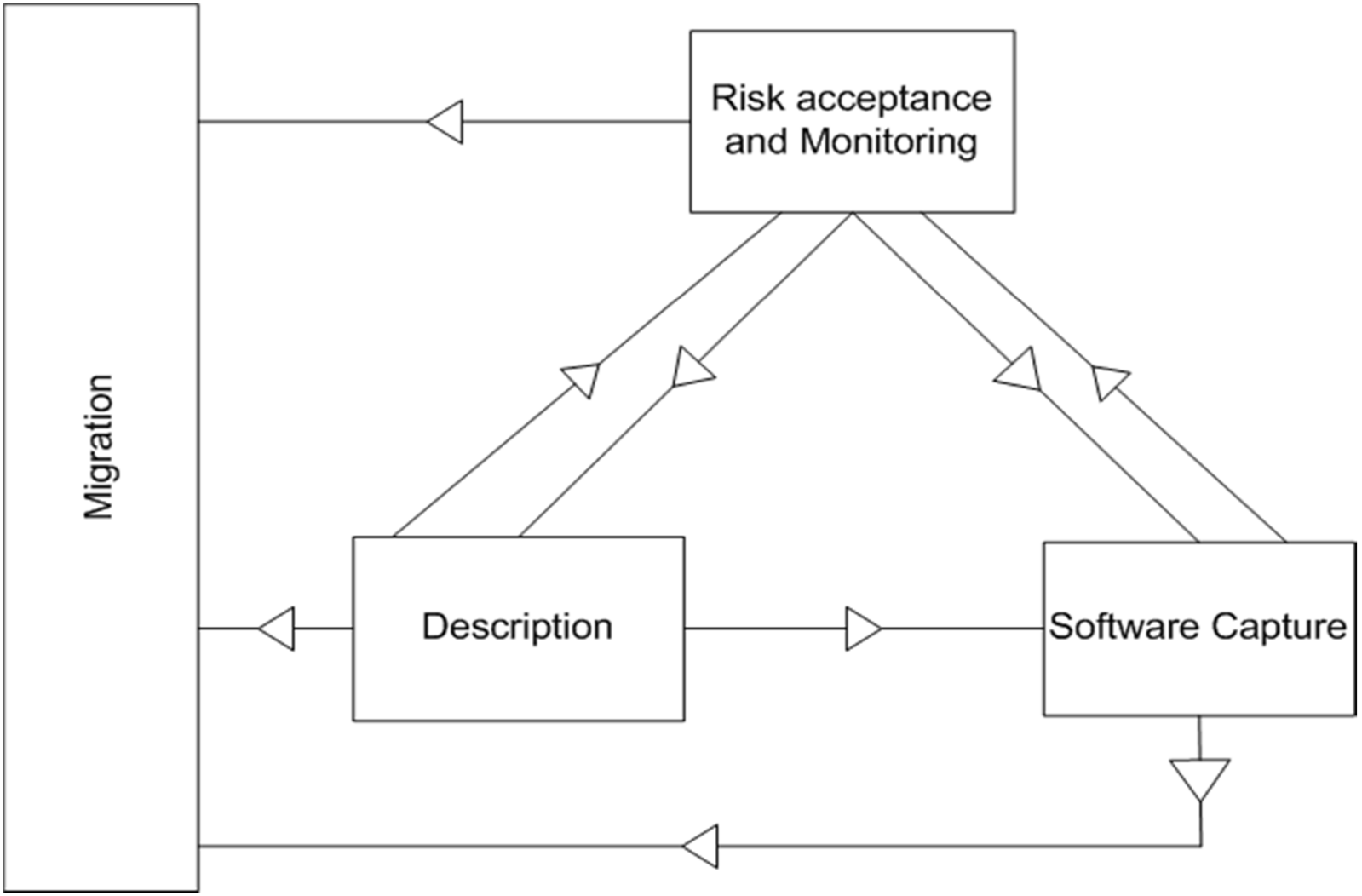

MST
NetCDF
Cartesian
Conference
proceedings
pdf
CEDA
repository


directory structure
mmm file
URSI Hanbooks
URSI Parameter code
Ionospheric monitoring group website
Bibliography
SAO Explorer
DRB description
DRB Jar File
Virtual height algorithm
DRB User Manual

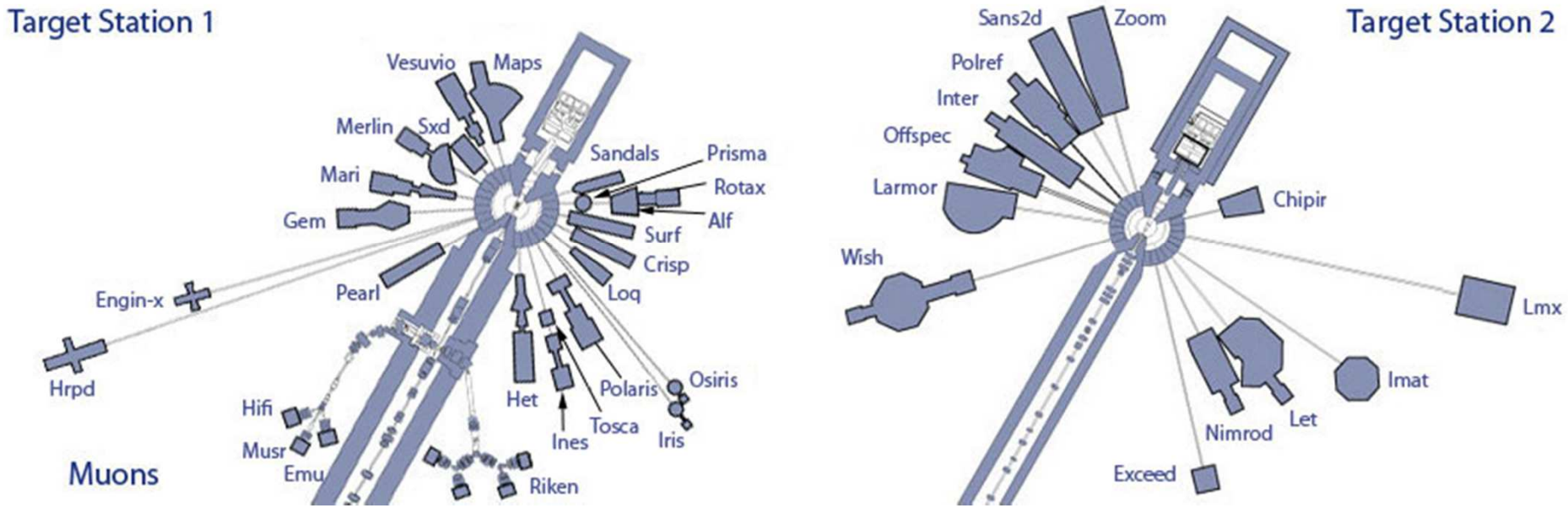

Directory structure
IIWG
URSI Handbooks
Station information
IIWG format description
URSI Parameter codes


GEM website
Nexus File
Proposal for powder diffraction experiment
Description of data structure
Description of Nexus Format
Load Data
Align Detectors
Diffraction Focussing
SaveGSS


GEM instrument website
Nexus File
Experimental proposal for a powder diffraction experiment
Nexus Format Description
UK web archiving consortium STFC website
Risk acceptance and Monitoring Strategy
Reference to Mantid Site
Descriptive Strategy
Software Capture Strategy which employs platform based variants
Load Data
Nexus Data Structure
Align Detectors
Diffraction Focussing
Save GSS
Python 5 for UNIX
Matidid VX.X UNIX Binary
NumPY for UNIX
Python 5 for MAC
Matidid VX.X MAC Binary
NumPY for MAC
Python 5 for Windows
Matidid VX.X Windows Binary
NumPY for Windows

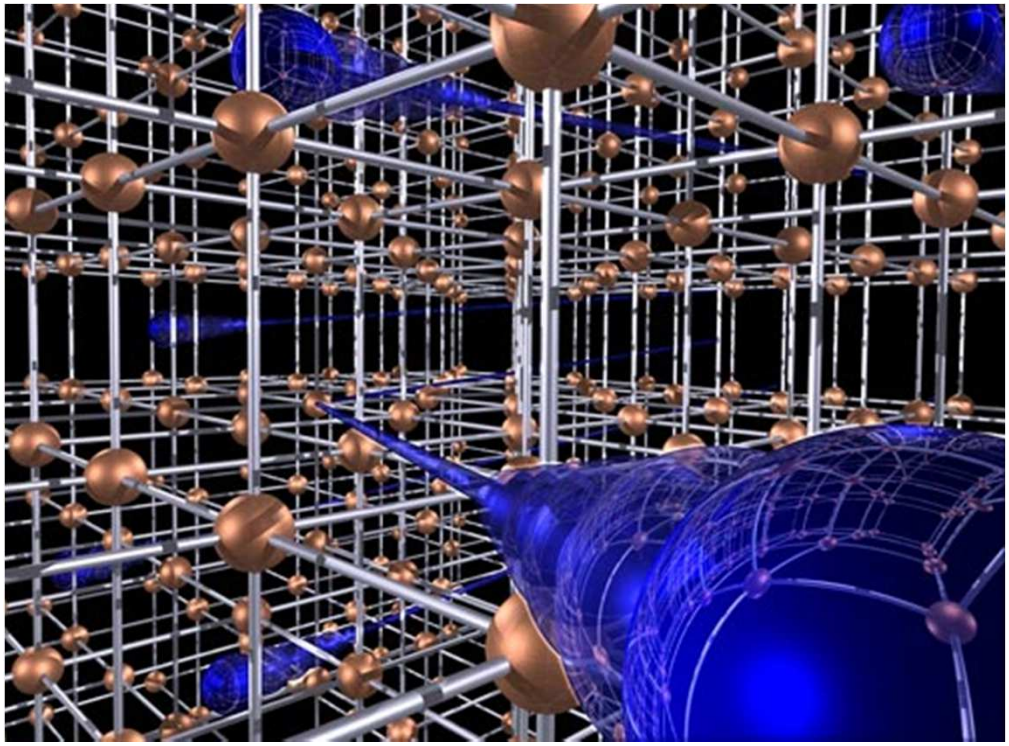

Migration
Risk acceptance and Monitoring
Description
Software Capture

ISIS Instruments

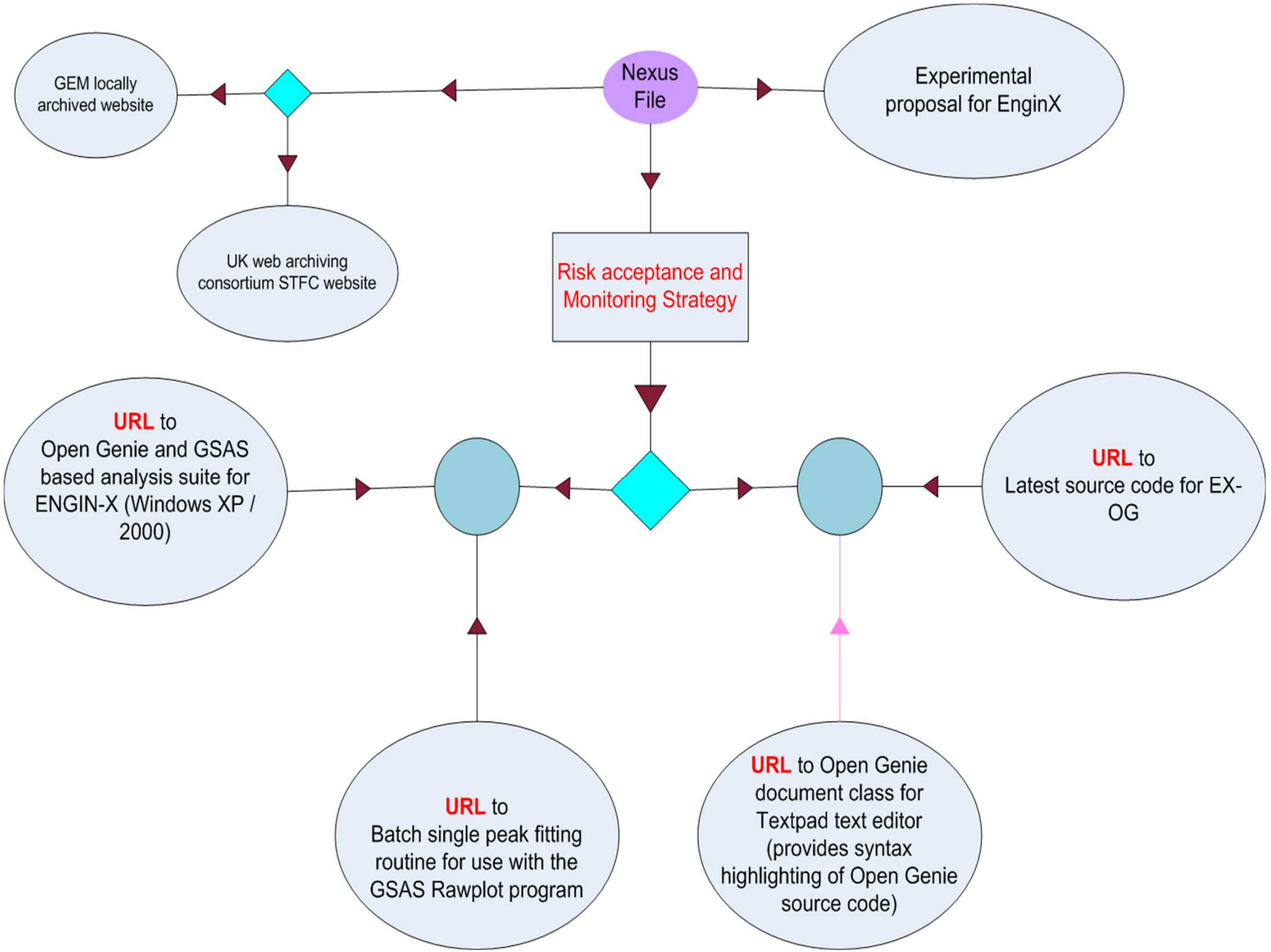


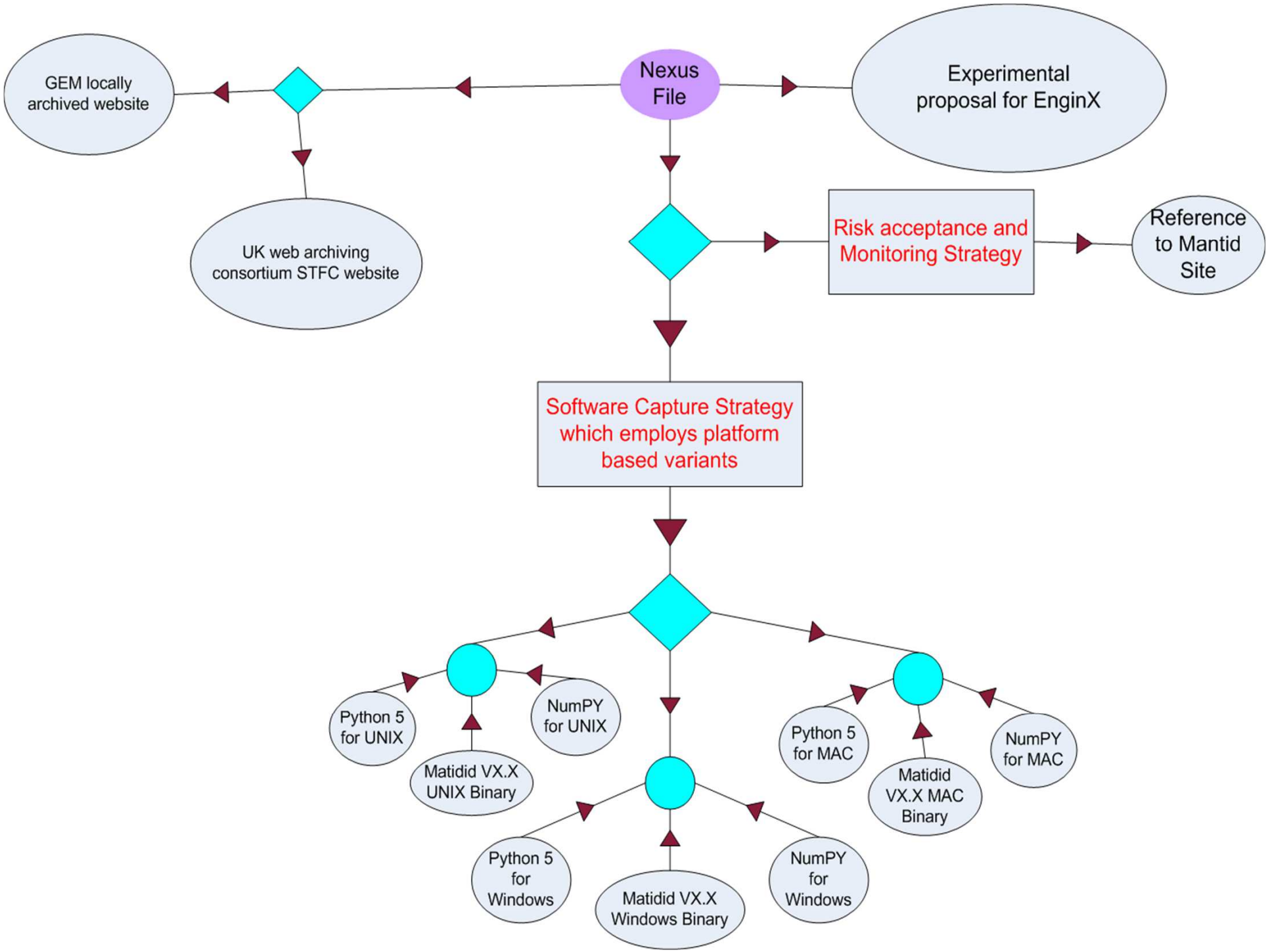
ISIS techniques

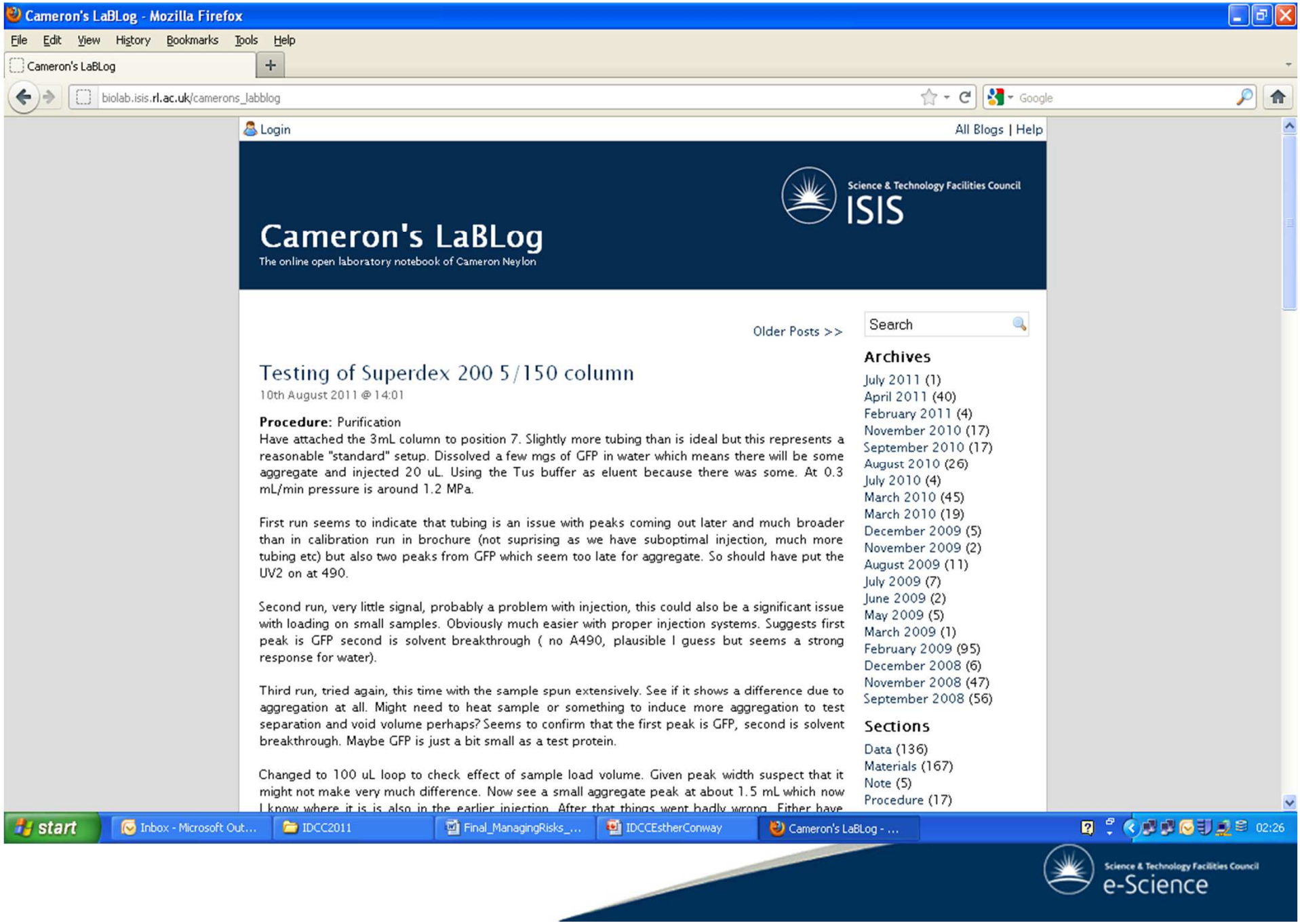
Muon spectroscopy
Neutron diffraction
Neutron spectroscopy
Reflectometry
Small Angle scattering

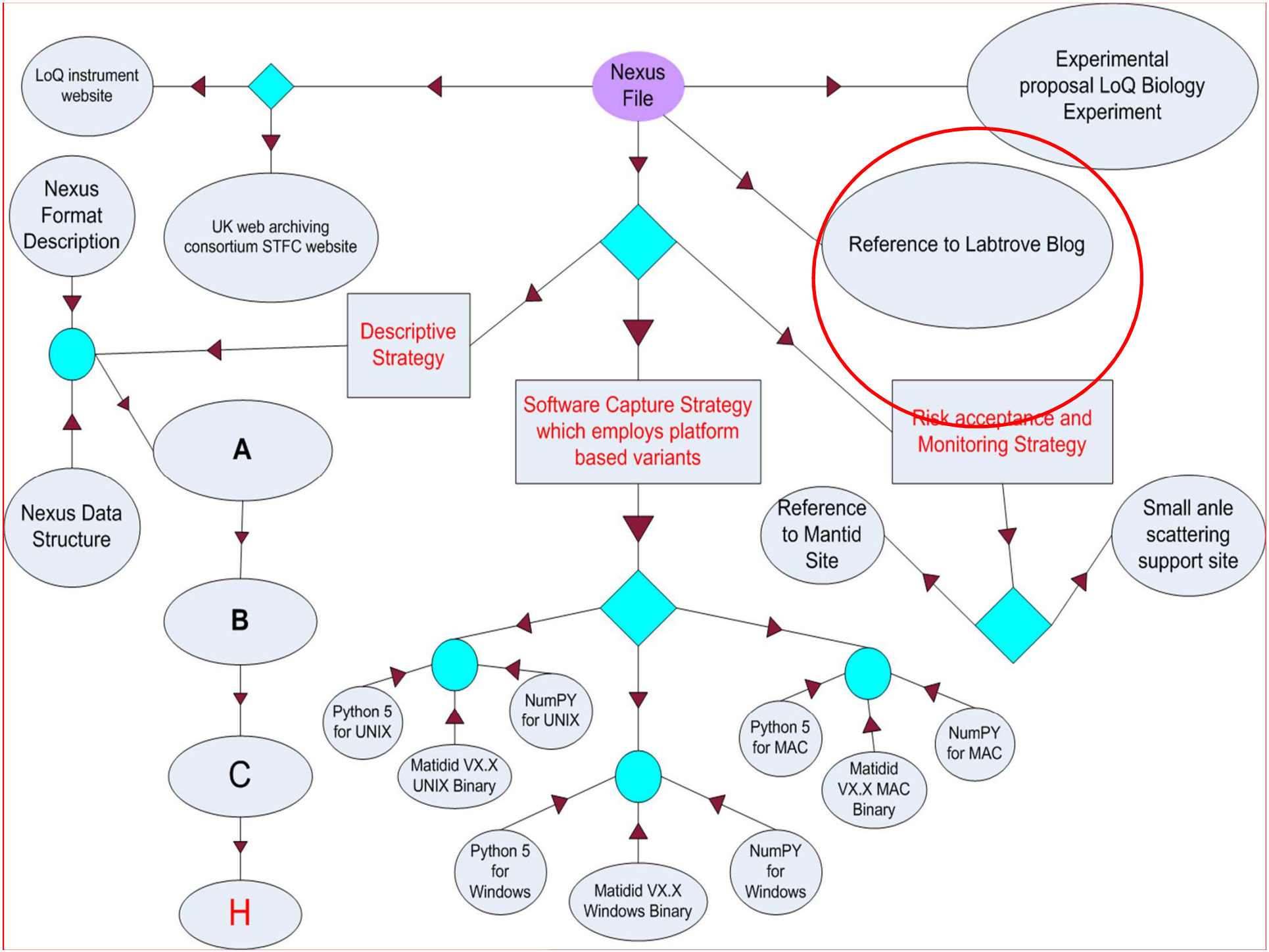


Finding the structure and properties of chemicals
To make new materials, drugs etc..


GEM locally archived website
Nexus File
Experimental proposal for EnginX
UK web archiving consortium STFC website
Risk acceptance and Monitoring Strategy
URL to Open Genie and GSAS based analysis suite for ENGIN-X (Windows XP / 2000)
URL to Latest source code for EX-OG
URL to Batch single peak fitting routine for use with the GSAS Rawplot program
URL to Open Genie document class for Textpad text editor (provides syntax highlighting of Open Genie source code)


GEM locally archived website
Nexus File
Experimental proposal for EnginX
UK web archiving consortium STFC website
Risk acceptance and Monitoring Strategy
Reference to Mantid Site
Software Capture Strategy which employs platform based variants
Python 5 for UNIX
NumPY for UNIX
Matidid VX.X UNIX Binary
Python 5 for MAC
NumPY for MAC
Matidid VX.X MAC Binary
Python 5 for Windows
Matidid VX.X Windows Binary
NumPY for Windows


Cameron's LaBLog - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Cameron's LaBLog
biolab.isis.rl.ac.uk/camerons_labblog
Google
Login
All Blogs | Help
Science & Technology Facilities Council
ISIS
Cameron's LaBLog
The online open laboratory notebook of Cameron Neylon
Older Posts >>
Search
Testing of Superdex 200 5/150 column
10th August 2011 @ 14:01
Procedure: Purification
Have attached the 3mL column to position 7. Slightly more tubing than is ideal but this represents a reasonable "standard" setup. Dissolved a few mgs of GFP in water which means there will be some aggregate and injected 20 uL. Using the Tus buffer as eluent because there was some. At 0.3 mL/min pressure is around 1.2 MPa.
First run seems to indicate that tubing is an issue with peaks coming out later and much broader than in calibration run in brochure (not suprising as we have suboptimal injection, much more tubing etc) but also two peaks from GFP which seem too late for aggregate. So should have put the UV2 on at 490.
Second run, very little signal, probably a problem with injection, this could also be a significant issue with loading on small samples. Obviously much easier with proper injection systems. Suggests first peak is GFP second is solvent breakthrough (no A490, plausible I guess but seems a strong response for water).
Third run, tried again, this time with the sample spun extensively. See if it shows a difference due to aggregation at all. Might need to heat sample or something to induce more aggregation to test separation and void volume perhaps? Seems to confirm that the first peak is GFP, second is solvent breakthrough. Maybe GFP is just a bit small as a test protein.
Changed to 100 uL loop to check effect of sample load volume. Given peak width suspect that it might not make very much difference. Now see a small aggregate peak at about 1.5 mL which now I know where it is is also in the earlier injection. After that things went badly wrong. Either have
Archives
July 2011 (1)
April 2011 (40)
February 2011 (4)
November 2010 (17)
September 2010 (17)
August 2010 (26)
July 2010 (4)
March 2010 (45)
March 2010 (19)
December 2009 (5)
November 2009 (2)
August 2009 (11)
July 2009 (7)
June 2009 (2)
May 2009 (5)
March 2009 (1)
February 2009 (95)
December 2008 (6)
November 2008 (47)
September 2008 (56)
Sections
Data (136)
Materials (167)
Note (5)
Procedure (17)
start
Inbox - Microsoft Out...
IDCC2011
Final_ManagingRisks_...
IDCCEstherConway
Cameron's LaBLog - ...
02:26
Science & Technology Facilities Council
e-Science


LoQ instrument website
Nexus File
Experimental proposal LoQ Biology Experiment
Nexus Format Description
UK web archiving consortium STFC website
Reference to Labtrove Blog
Descriptive Strategy
Software Capture Strategy which employs platform based variants
Risk acceptance and Monitoring Strategy
A
Nexus Data Structure
Reference to Mantid Site
Small anle scattering support site
B
C
H
Python 5 for UNIX
Matidid VX.X UNIX Binary
NumPY for UNIX
Python 5 for MAC
Matidid VX.X MAC Binary
NumPY for MAC
Python 5 for Windows
Matidid VX.X Windows Binary
NumPY for Windows

Permits Automated Assessments

- Cost
- Value
- Risk
- Benefit

Added Benefits

- ***Assess exposure to risk and longevity of solutions***
- ***Impact analysis based on risk realisation***
- ***Tailoring solutions that are responsive to evolving environments***
- ***Tailoring solutions that are responsive to evolving environments***
- ***Scalability through characterising data collections***

Interesting Questions For Us?

- Risk Profiling
- Benefits and Performance
- Cost and Re-use of Solutions
- Quality and Trust

Questions From You?

esther.conway@stfc.ac.uk