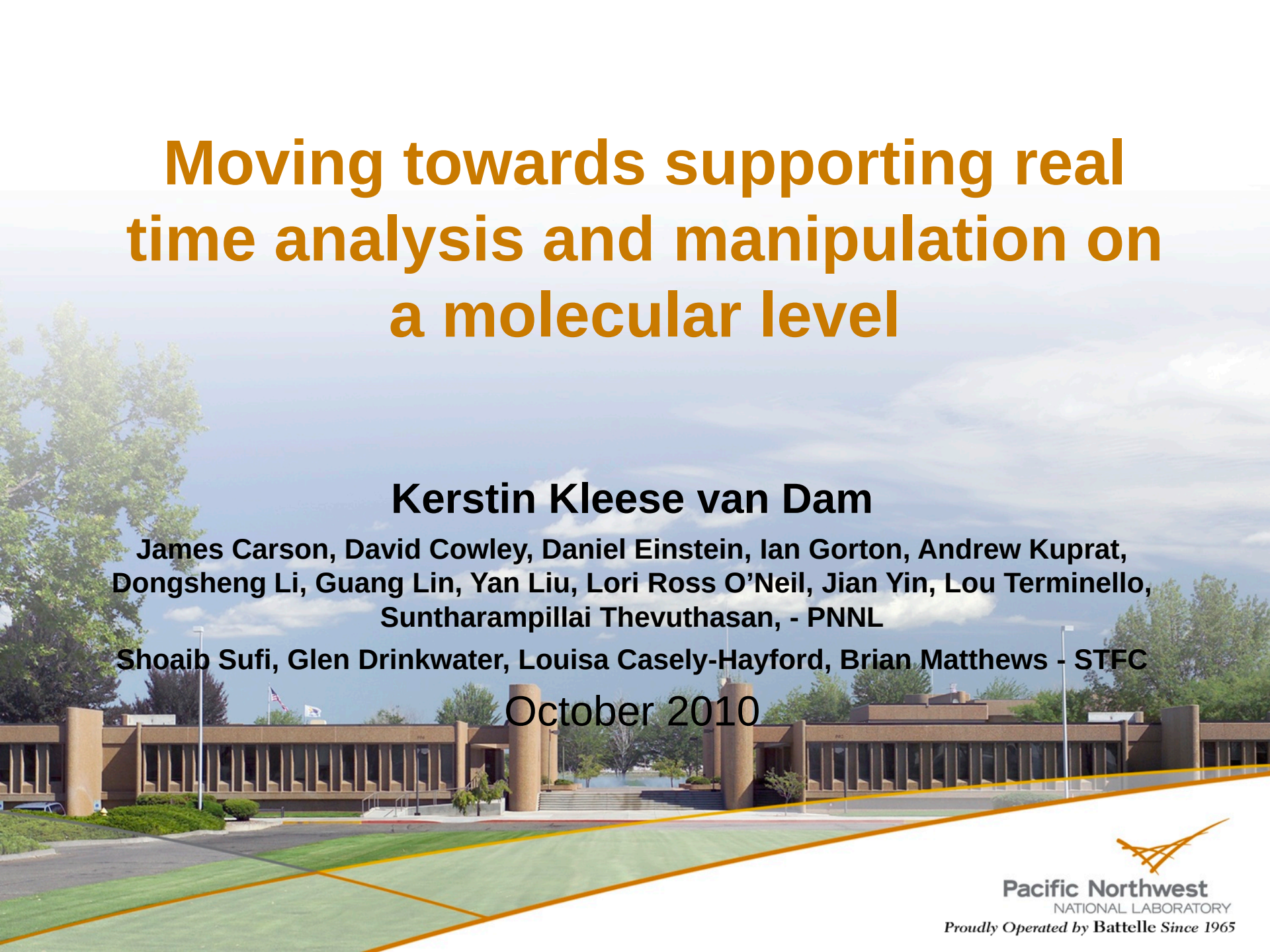
Moving towards supporting real time analysis and manipulation on a molecular level

Kerstin Kleese van Dam

James Carson, David Cowley, Daniel Einstein, Ian Gorton, Andrew Kuprat, Dongsheng Li, Guang Lin, Yan Liu, Lori Ross O'Neil, Jian Yin, Lou Terminello, Suntharampillai Thevuthasan, - PNNL

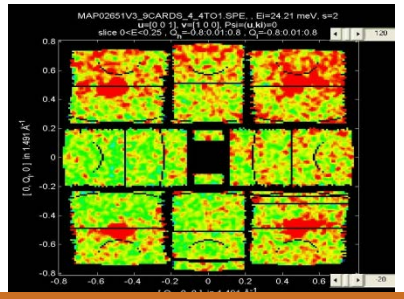
Shoaib Sufi, Glen Drinkwater, Louisa Casely-Hayford, Brian Matthews - STFC

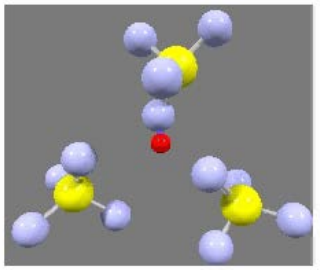
October 2010

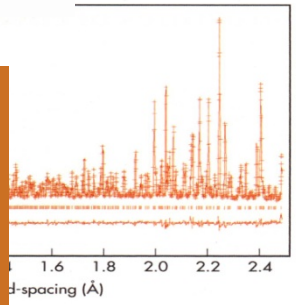


Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*

Science needs to answer complex Questions

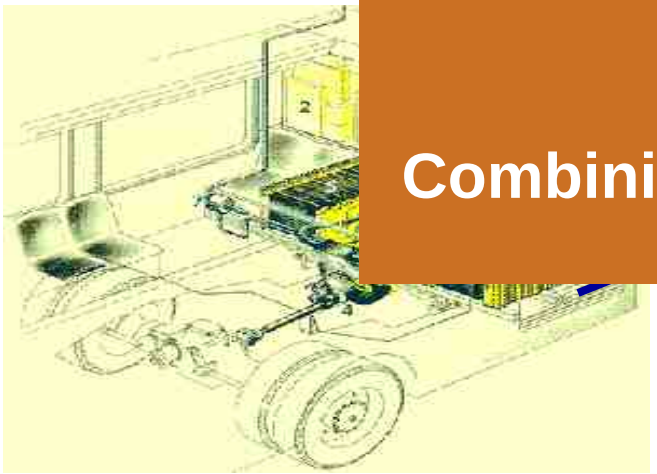


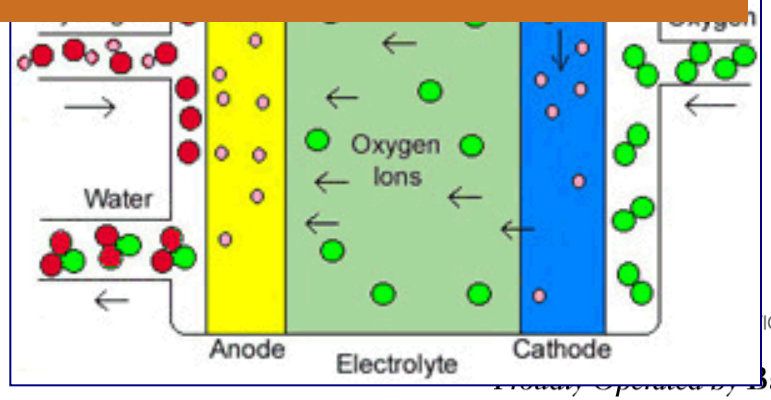




Making it Happen

Many People –
Combining complementary Expertise





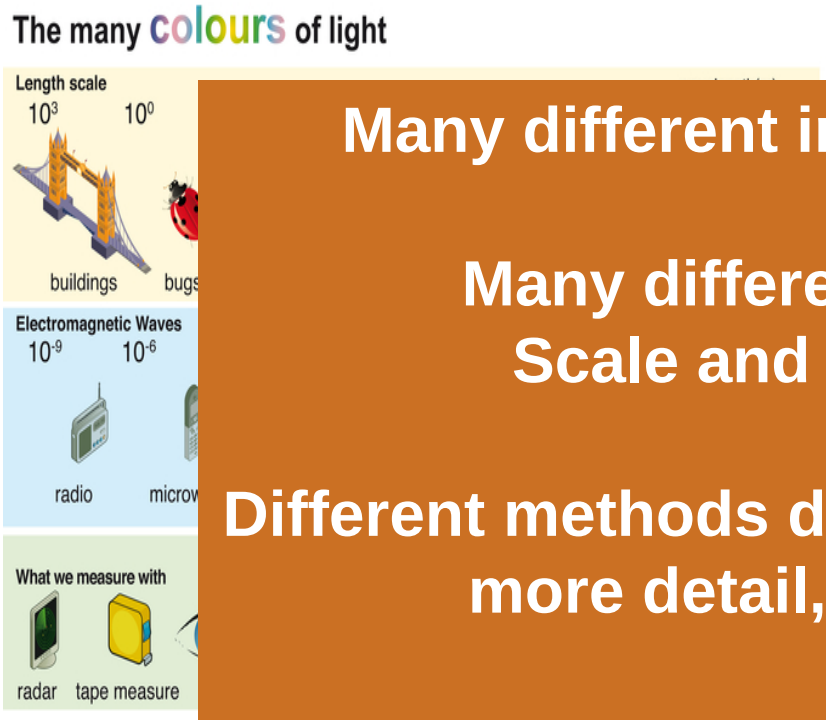
Dr. Robert McGreevy, ISIS, UK

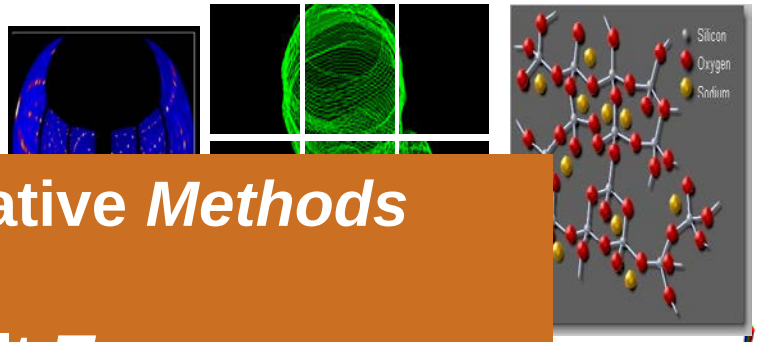
Northwest
IONAL LABORATORY
Battelle Since 1965

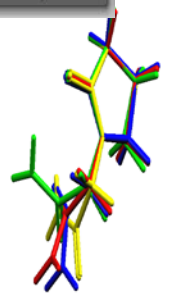
Investigative Methods

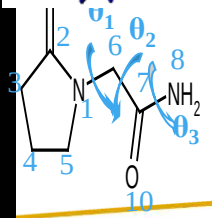
Result Representation

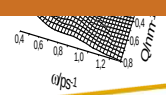
Methods











Many different investigative *Methods*

Many different result *Types* –
Scale and Representation

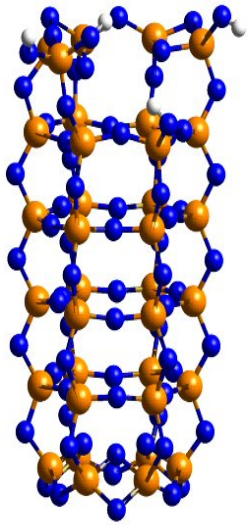
Different methods deliver different *Insights* –
more detail, new information

DIAMOND Light Source UK, 2010

Motivation

- **Enabling improved analysis through more complete information, facilitated by allowing users to get rapid access to current and past data, related projects, publications etc.**

- **Enabling rational design and synthesis of new chemical, biological, and materials systems through integrated molecular scale imaging technologies, real time analysis and manipulation.**



Capturing Data and Making it Accessible

Science & Technology Facilities Council

Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

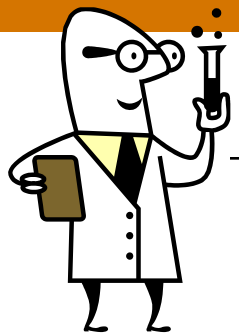
UK Science and Technology Facilities Council (STFC)

STFC (formerly CCLRC) facilitates the access to large scale experimental and computational facilities for the UK research community, both through subscriptions to international institutions and by operating a range of world class facilities e.g.:

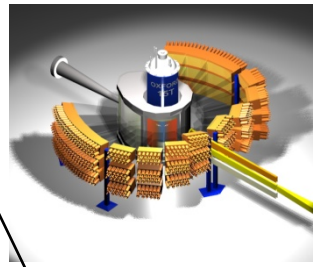
- **ISIS Neutron and Muon Facility**
- **Daresbury Laboratory Synchrotron**
- **DIAMOND Light Source**
- **Central Laser Facilities**

Science & Technology Facilities Council

Integrated e-Infrastructure

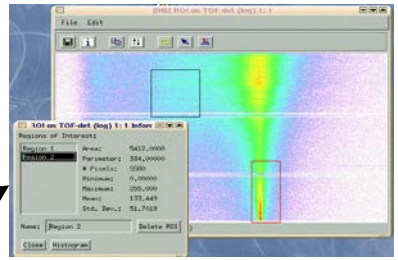


Experiment



Data Acquisition System

Data Analysis



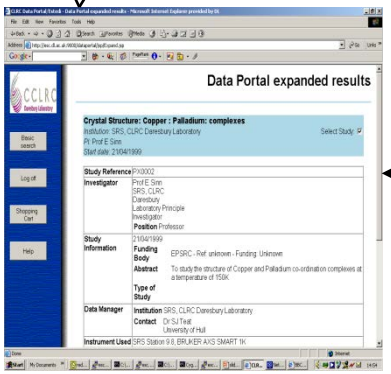
Proposal

All Data and Metadata Capture is automated.

Publication

Proposal System

Information



Metadata Catalogue

Secure Storage

E-Pubs

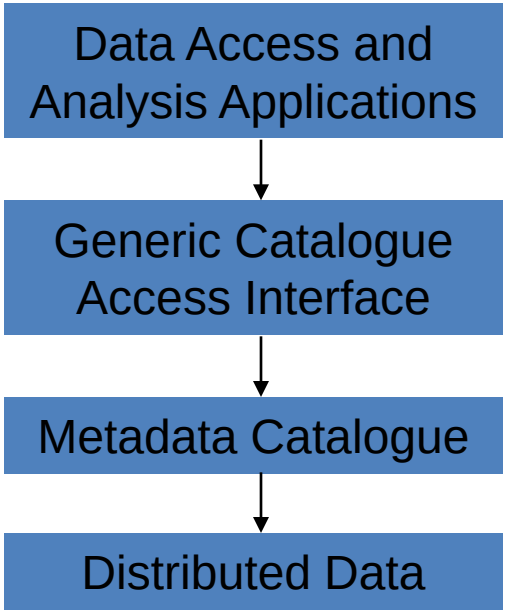
Science & Technology Facilities Council

Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

ICAT Software Suite

- **The ICAT software suite centrally catalogues all experiment related information and extracts key results.**
- **Where ever possible information is gathered automatically trough integration with existing IT systems such as proposal systems or data acquisition.**
- **The catalogue and the data it references are accessible via a well defined API for easy embedding into any applications.**



Integrated Infrastructure Architecture

Online Proposal System

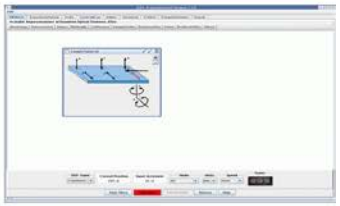
User Office System incl.:
User Database
Scheduling
Health and Safety
Proposal Management

Single Sign On Account Creation and Management

DataPortal

Metadata Catalogue

Data Acquisition System



Storage Management System

Data Analysis and Visualisation Interfaces

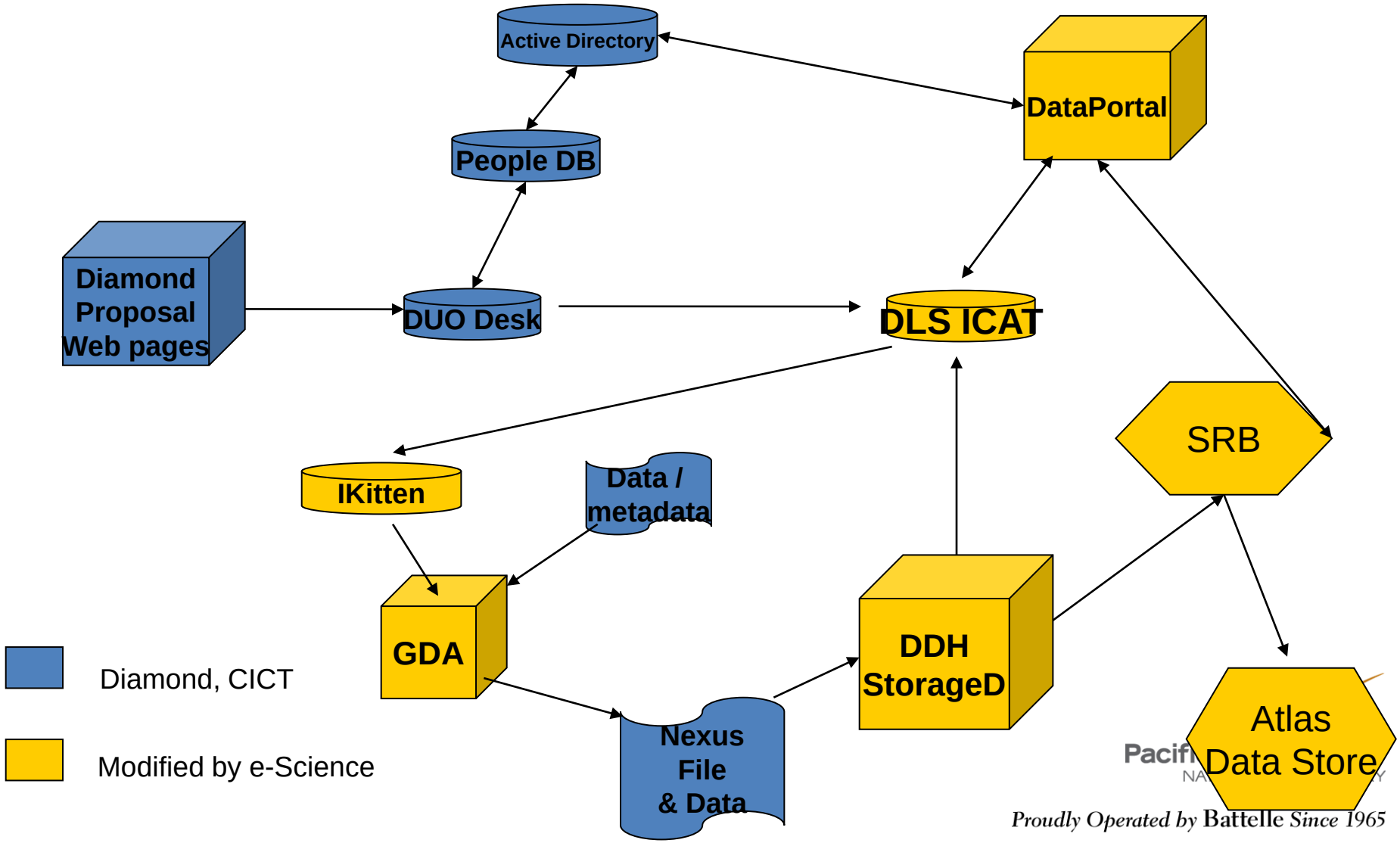
Annotation

Data Analysis Simulation
Code Repository

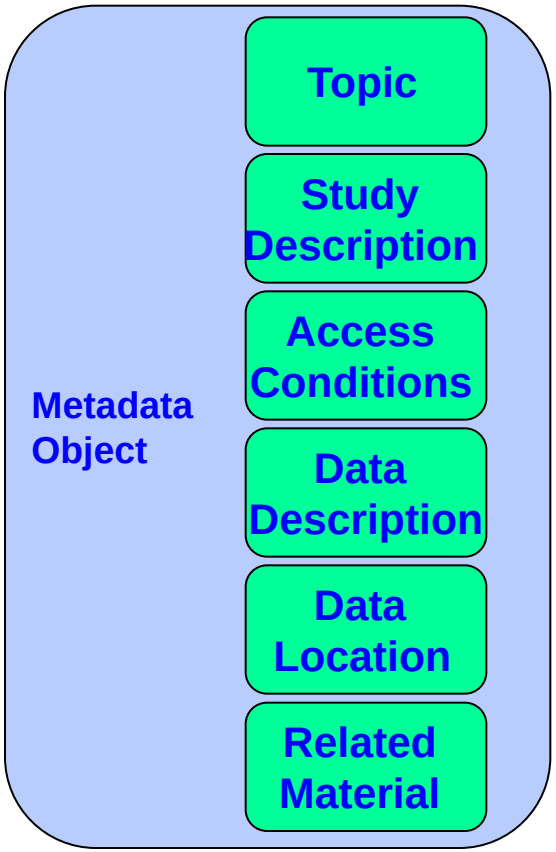
XML Output

CCLRC Integrated e-Infrastructure for Facilities

DIAMOND Implementation



Metadata Model



Keywords **providing a index on what the study is about.**

Provenance **about what the study is, who did it and when.**

Conditions of use **providing information on who and how the data can be accessed.**

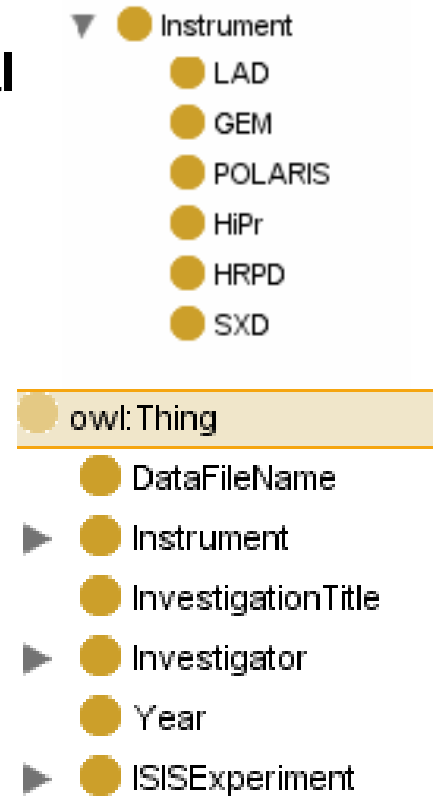
Detailed description **of the organisation of the data into datasets and files.**

Locations **providing a navigational to where the data on the study can be found.**

References **into the literature and community providing context about the study.**

Ontology Support

- **1,700,000 distinctive keywords ISIS ICAT**
- **These keywords are used to index experimental studies**
- **The creation of ontology's at ISIS aids the mapping of familiar terms in one domain as well as related concepts in different domains.**
- **Facilitates searching of data by category and grouping of data into keywords across studies. Faster results and enabling of cross facility search.**



Science & Technology Facilities Council

Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

Science & Technology Facilities Council
e-Science

Search Results

My Data Search Results

User: kk44
Expire time: 18:09 PM 04.08.2008

- Search
- User Preferences
- Log out

Keyword | Advanced | ISIS

Keyword(s):

Auto Case

Search Reset

Investigations

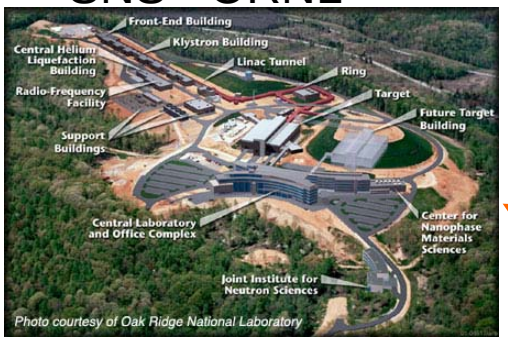
#		Rb Number	Title	Type	Instrument	Investigator	Run Range	Year
1	⊞	720378	Authentication of a bronze figure from the Florence's National Museum of Archaeology	experiment	ENGINX			
2	⊞	14995	Bronze Geth=1MeV Bipar+TFA	experiment	EVS	GG,CA,MT,EP,AP,RS - UNIMIB, UNITOV	10661-10662	2004
3	⊞	720582	The manufacturing of Middle and Late Bronze Age Ceremonial Weapons .	experiment	ENGINX			

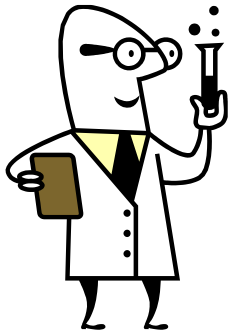
3 Investigations found, displaying 3, from 1 to 3. Page 1 / 1

> Help

Infrastructure – Access to Multiple Facilities

SNS - ORNL





CSL - Canada

Data Portal

ISIS – TS1 + 2

DLS

SRS + ERLP

CLF

Science & Technology Facilities Council

Pacific Northwest NATIONAL LABORATORY

Operated by Battelle *Since 1965*

What was achieved by 2008

Agreed common metadata and data formats and single sign on allow scientist to have rapid access at any stage of their work to:

- **Present and past projects at STFC.**
- **Publications**
- **Raw data from DLS, ISIS and CLF.**
- **Advanced Analysis Software and High Performance and High Throughput Computing**
- **Advanced Visualisation**

Created a Knowledge Repository:

- **A 20 year back catalogue of ISIS raw data is available**
- **All future data collected at STFC Facilities and DLS will be curated and made available for reuse now and in the future**

Initial Hurdles

- Existing catalogues were not as metadata rich as expected – existing metadata poor in quality
- No Provenance Information linked to the stored data (who, what when and why)
- Data stored insecurely partly on decaying media
- Data was not externally accessible – off-line, unorganized etc.
- ‘Not invented here’ syndrome
- Reliability

Lessons Learned

- **Address Early on:**
  - Resolve institutional and/or personal issues in dealing with mistrust and the use of the data collected
  - Developing agreements, policies and procedures for data

Don’t loose sight of the final goal – Create an environment to enable better science!

- **Design:**
  - Metadata is the key enabling technology
  - Automation and reliability of processes are vital
  - Interfaces should be as familiar as possible
  - Close integration into existing scientific processes
  - Step wise progression to take users ‘along’

MyEMSL 2010

Pacific Northwest
NATIONAL LABORATORY

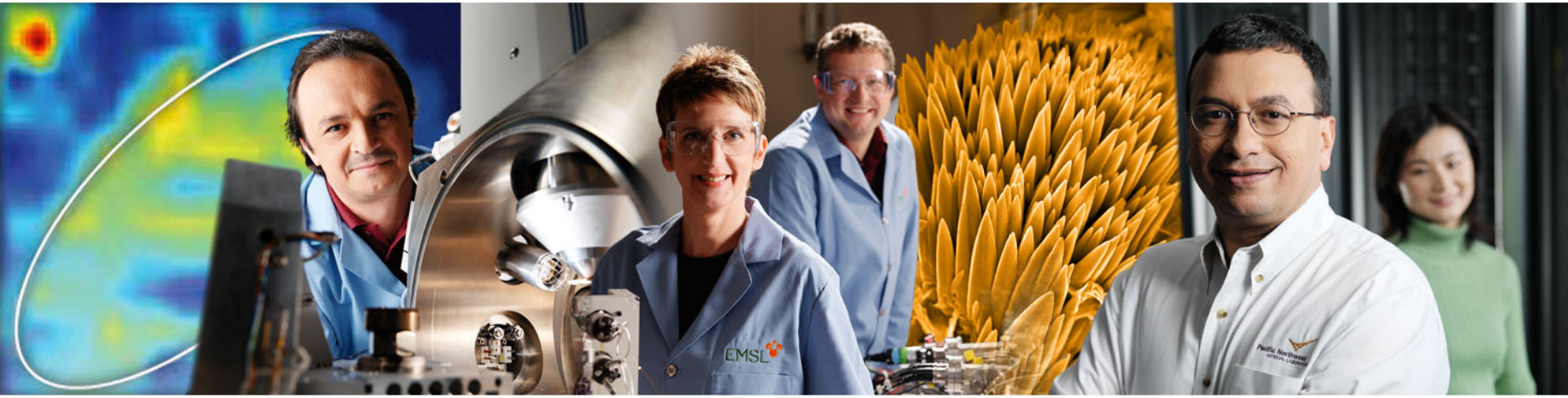
Proudly Operated by Battelle *Since 1965*

EMSL

EMSL

Environmental Molecular Sciences Laboratory:

A national user facility integrating experimental and computational resources for discovery and technological innovation



www.emsl.pnl.gov

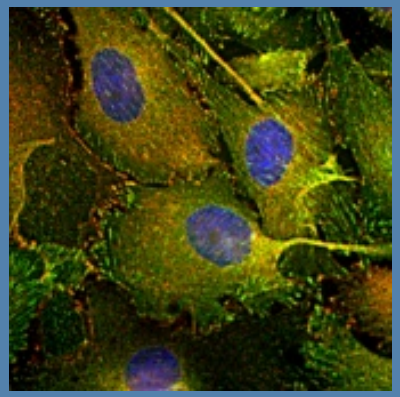
Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

U.S. DEPARTMENT OF ENERGY

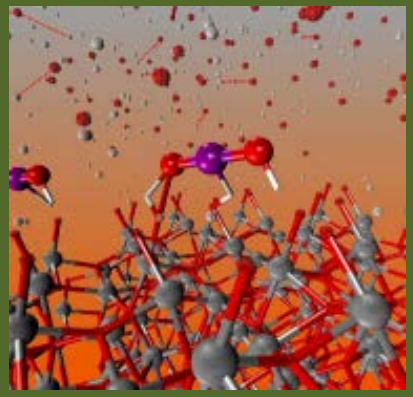
The user program is focused on three science themes

Biological Interactions and Dynamics



Understanding and optimizing the response of biological systems to their environment.

Geochemistry/Biogeochemistry & Subsurface Science



Unraveling molecular-level phenomena to determine their impact on contaminant migration and transformation.

Science of Interfacial Phenomena

Developing & verifying predictive models for interfacial processes and advancing understanding of structure-function relationships in complex systems.

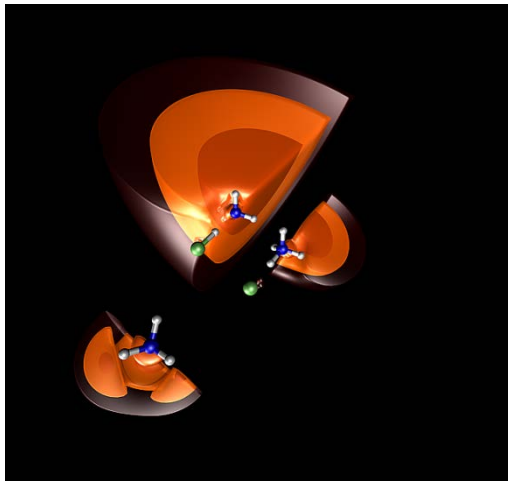
Environmental Molecular Sciences Laboratory: A National Scientific User Facility

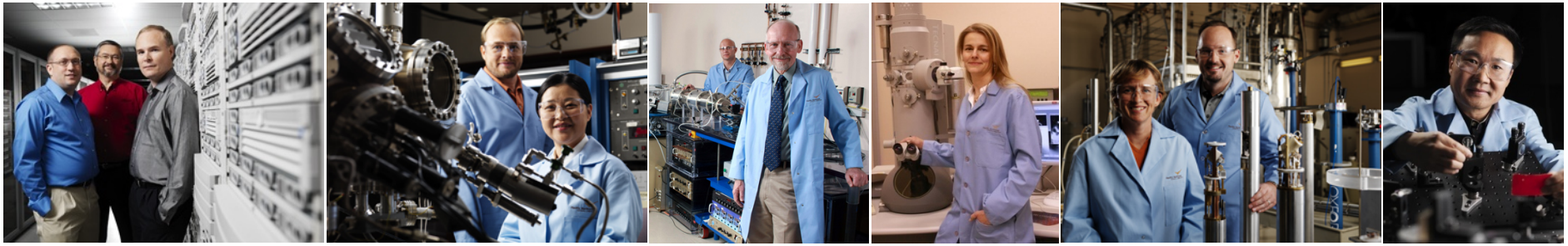
EMSL

Capabilities

EMSL houses an unparalleled collection of state-of-the-art capabilities that are used to address scientific challenges:

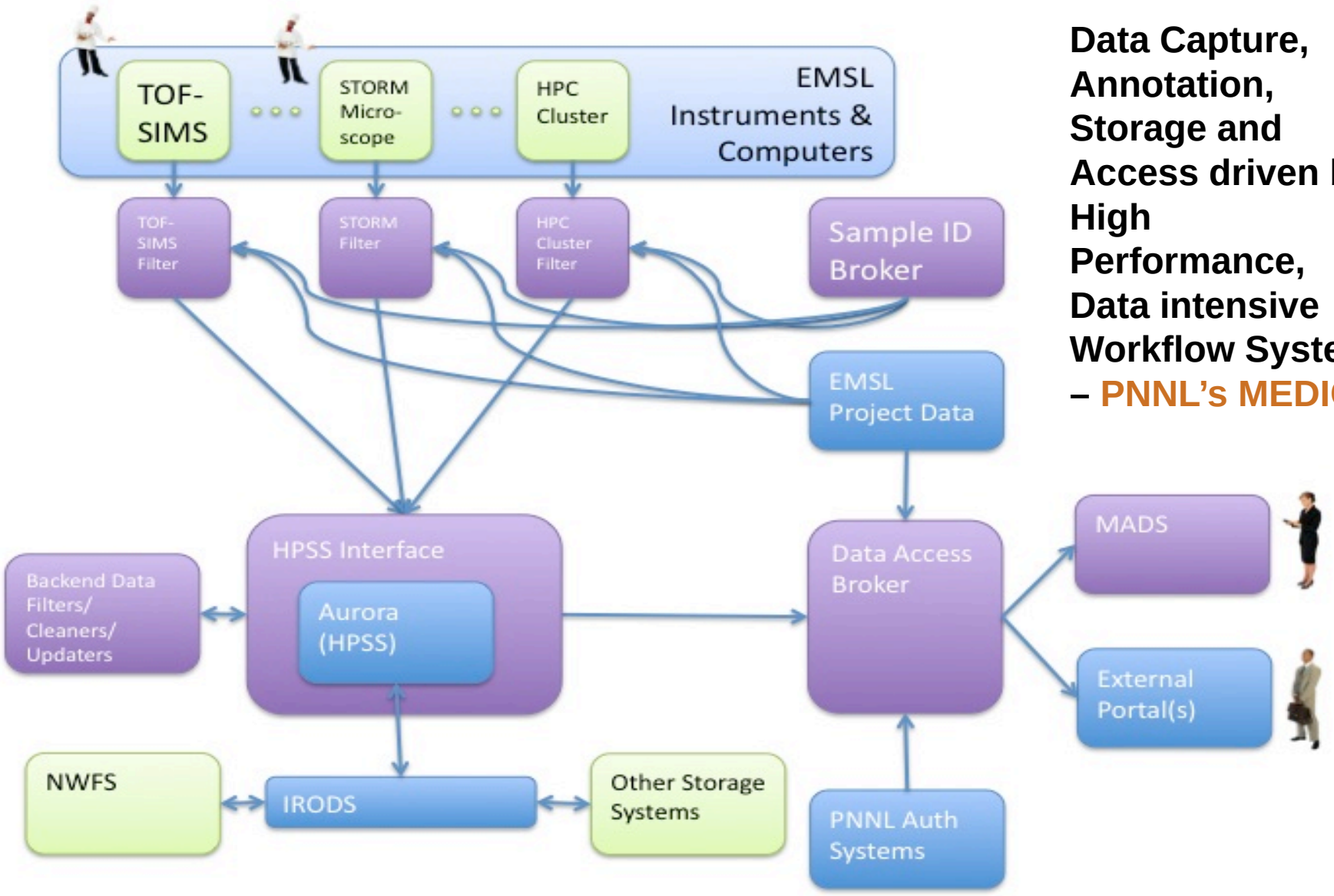
- Molecular Science Computing
- Deposition and Microfabrication
- Kinetics and Reactions
- Mass Spectrometry
- Microscopy
- NMR and EPR
- Spectroscopy and Diffraction
- Subsurface Flow and Transport





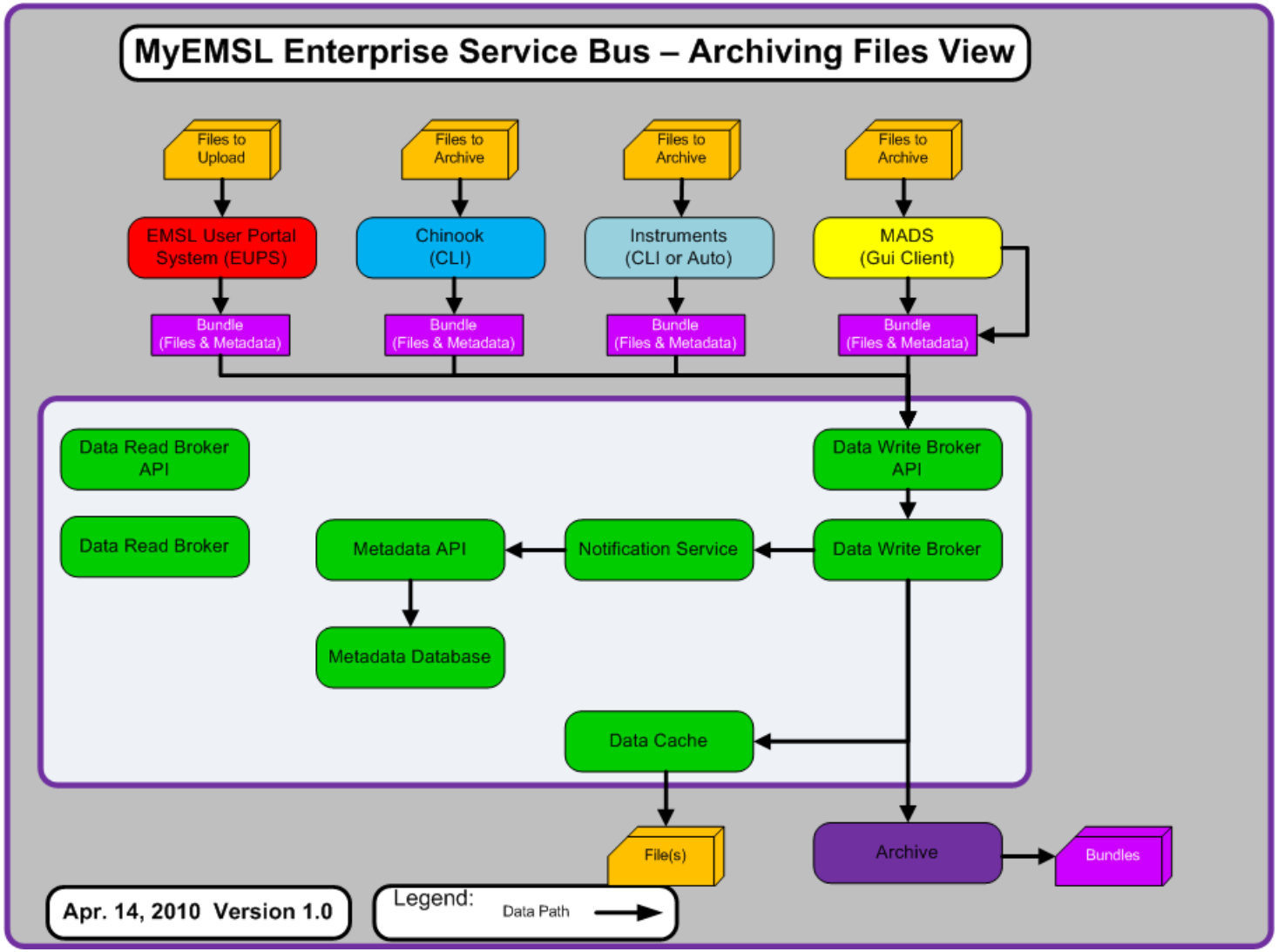
Pacific Northwest NATIONAL LABORATORY
U.S. DEPARTMENT OF ENERGY
Office of Science U.S. DEPARTMENT OF ENERGY

MyEMSL Architecture



Data Capture, Annotation, Storage and Access driven by High Performance, Data intensive Workflow System – PNNL's MEDICI

MyEMSL Data Flow



Pacific Northwest NATIONAL LABORATORY

Next Generation Proteomics Pipeline

Data Intensive Computing Research Enables Next Generation Proteomics Analysis Pipeline

Researching highly complex biological systems – such as microbial communities responsible for carbon uptake in our oceans – requires the development of complex experiments and the generation and evaluation of huge volumes of data.

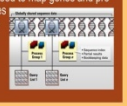
PNNL's *Next Generation Proteomics* technologies provide the capability to address this scale of problem, which in itself presents significant computational challenges. Our data intensive computing capabilities are advancing the scientific discovery process.

Data Genomes/ Metagenomes

Biology Samples

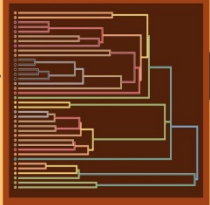
ScalaBLAST

ScalaBLAST is high performance software used by investigators to rapidly analyze large amounts of biological sequence data where detailed understanding of the molecular makeup of organisms or communities is crucial. ScalaBLAST provides scientists the information they need to map genes and proteins across species to help predict the specific functions and capabilities of the organisms.



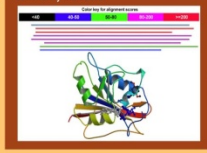
Develop Sequences

Comparative sequence analyses are used to mine spectra, explore biological activity and identify species.



Protein Families

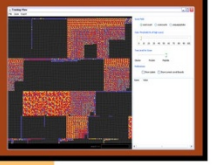
Protein sequences are organized by similarity, or "families," using advanced clustering algorithms which partition the metaproteome and metagenomes, and which allows for easier interpretation of the data. Grouping proteins by cellular function helps biologists address questions about what the system is doing. Similarities in sequences also indicate an evolutionary relationship among microbial community members.



Visualization/Interaction

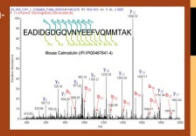
Interactive Visual Content Analysis of Real-Time Streams provides easily understandable visualizations of large datasets to systems biologists. This tool allows the user to visually navigate and explore data streams in manners not previously possible, making the search for relevant protein function much more dynamic and effective. Key features include:

- Data is incorporated into a visualization in real time.
- Easy-to-understand visualization organizes data points and are customizable to the user's set preferences.



Peptide Identification

Thousands of mass spectra are generated for a single sample. These spectra are automatically and routinely matched to sequences to identify proteins. Programs developed for this process are parallelized to take advantage of high-performance computing for the purpose of speeding up the sample analysis.



Mass Tag DB's

MS/MS

MS

SMART INSTRUMENT CONTROL

Ion Mobility Spectrometry (IMS) Time of Flight

Hybrid Algorithms for Networked Systems Analysis enables more effective Ion Mobility Mass Spectrometry measurements. Dynamic multiplexing schemes (such as Hadamard Transforms) addresses the inherent undersampling limitation of IMS-TOF measurements. These schemes require real-time demultiplexing of streaming data.

- Hybrid architectures allow real-time processing of streaming data thus supporting diagnostic application as well as intelligent control schemes.

Field Programmable Gate Array (FPGA) Acquisition

IMS-TOF signal detection requires multidimensional signal averaging of high speed streaming data. The current IMS-TOF detection systems do not utilize vendors proprietary detection hardware due to the data intensive challenges presented in the architecture. DICI capability developments have enabled the IMS-TOF team to form a collaboration with the vendor for an FPGA based solution to process the high speed streaming data. This research will result in IMS-TOF improvements in:

- Sensitivity
- Data density
- Resolution and mass measurement accuracy.

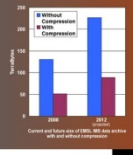
The Smart Instrument Control Technology facilitates intelligent data gathering in tandem mass spectrometry experiments by analyzing acquired spectra on-the-fly and performing pattern matching against a database of known signatures. Using this technology, researchers can reduce the amount of redundancy in their data collection procedures while also maximizing the information contained within the biological samples.

Mass Tag DB

Data Compression

This technology is a compression algorithm that reduces data size while significantly increasing the speed for storing and extracting data. Key features include:

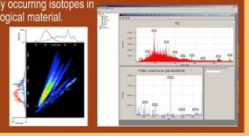
- Data storage size reduction by a factor of 3
- Compression and storage of 240 MB of data in less than one second
- Extract individual segments of uncompressed data from a compressed file



Archive

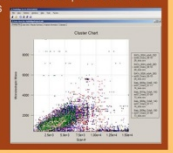
Feature Detection

Mass spectrometry features are extracted from raw instrument datasets and deisotoped - a process that involves the detection, deconvolution, and calculation of a monoisotopic mass from clusters of peaks – representing different combinations of naturally occurring isotopes in the biological material.



Feature Identification

Mass spectrometry features are aligned from multiple datasets and compared to mass tag databases to make quantitative comparisons of proteins from several samples at once. Additionally, protein dynamics can be determined in studies involving seasonal fluctuations in community populations, protein turnover, and cellular pathways.



MeDICI

Middleware for Data Intensive Computing (MeDICi) is a software toolkit for creating scientific pipelines that can pull data from multiple instruments at a time and manipulate, format and organize the data for researchers to effectively view its information. Features are:

- Allowing scientists to visually create and modify a data processing pipeline to match their analysis needs
- Supporting multiple programming languages, communication protocols, hardware platforms
- Making it possible to capture and integrate high-throughput sensor data in real-time

PNNL Chemical Imaging Initiative

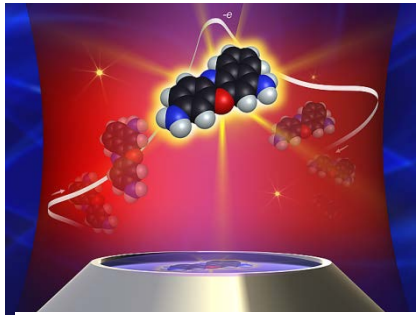
Pacific Northwest
NATIONAL LABORATORY

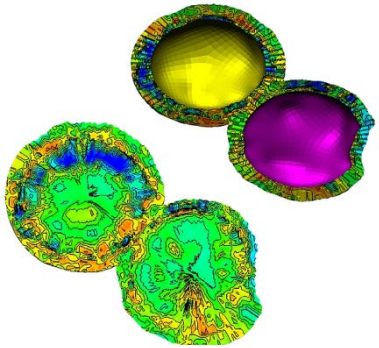
Proudly Operated by Battelle *Since 1965*

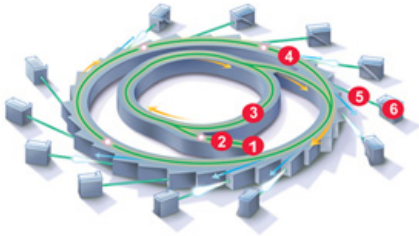
Real-World Manipulation on a Molecular Level

- **Purpose: This initiative will deliver a suite of unique tools with nanometer scale resolution and element specificity that will allow researchers to go from model system observation to real-world manipulation on a molecular level.**

- **Approach: We will build signature, *in-situ* capabilities in**
  - **Light source based x-ray and VUV probes coupled with laboratory based imaging capabilities for 3D tomographic, structural, and element specific interrogation at the molecular level**
  - **Coupled optical, electron, ion, mass, and scanned probe microscopies to understand chemical and biological transformations and mechanisms**
  - **Integrative hardware and software applications for image reconstruction, feature extraction and information integration**



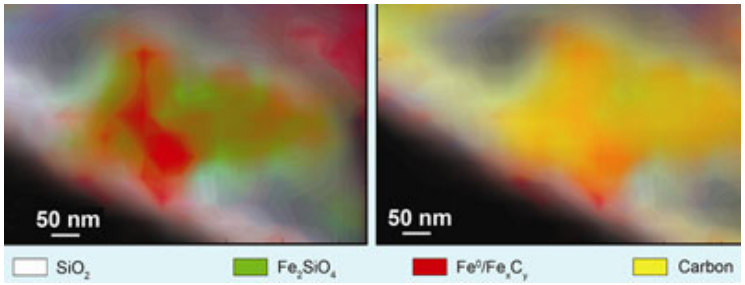




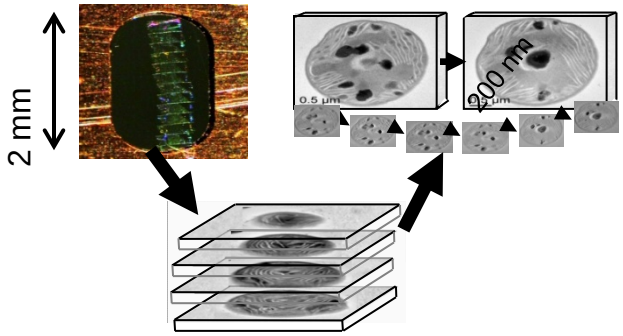
Pacific Northwest
NATIONAL LABORATORY

What will allow us to go from model system observation to real world manipulation of *in situ* interfaces on a molecular level?

- **Direct visualization of chemical, material, and biological transformations are essential to achieve a confident level of control over complex systems**
  - **e.g: the scanning tunneling microscope transformed surface science because of its visualization power**
- **Most of our ability to control matter today is through inference and interpretation of spectroscopic and structural data and modeling**
- **<u>Direct Observation:</u> With molecular scale imaging tools and concomitant data handling methods we can achieve the level of control enabling rational design and synthesis of new chemical, biological, and materials systems.**



Chemical map of a FT catalyst – proto molecular movie



3D EM tomography of Cyanothece Cell

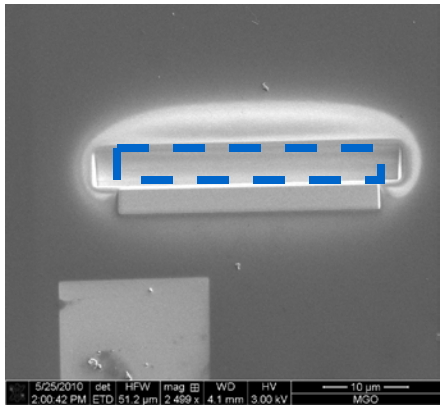
Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

Integrated Sample Preparation and Analysis Platform for Comprehensive 3-D Chemical Imaging

Challenges:

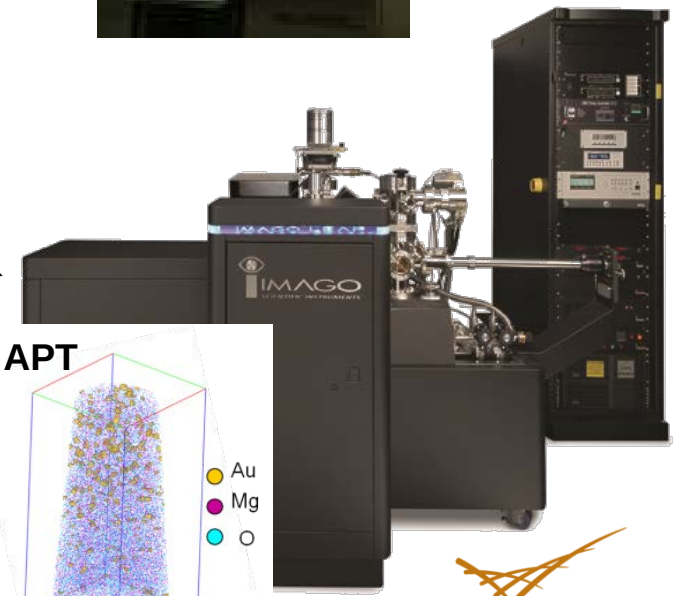
- **Determining 3-D positions and chemical identity of individual atoms in any materials system, analysis across different imaging techniques**



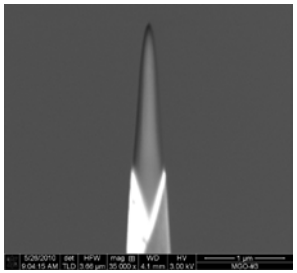
DB-FIB

TEM

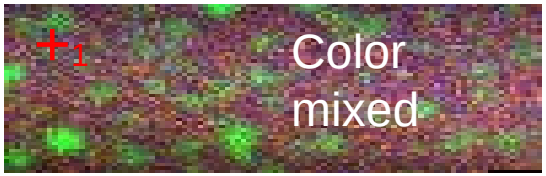
APT



DB-FIB



STEM



APT

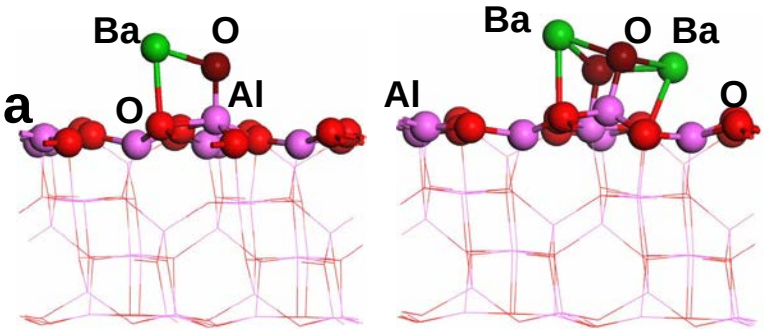
Mg: Orange
O: Blue
Au: Green

Pacific Northwest
NATIONAL LABORATORY

Atomic and Electronic Structure of Catalysts by Combination of In-situ and Ex-situ Imaging

Challenges:

- **Oxide nanoclusters supported by gamma alumina**
  - **Physical structure and chemical state of elements**
  - **Electronic structure**
  - **Influence of these in catalytic properties**

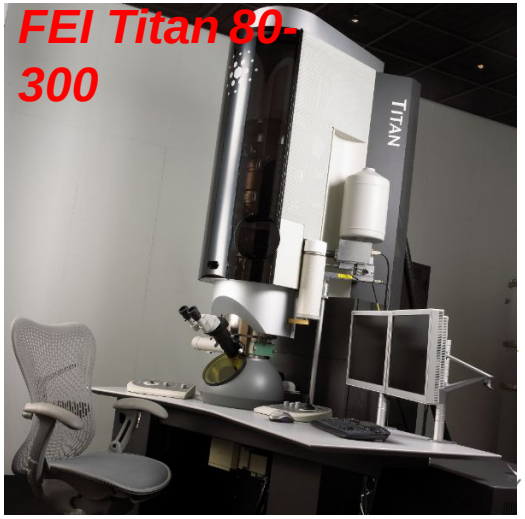


BaO monomer and dimer on a dehydroxylated γ -Al_2O_3(100) surface

Kwak et al, Journal of Catalysis 261, 17-22, 2009

Approach:

- **Class of TMO (e.g., WO_3, MO_3, V_2O_5) and TM (e.g., Pt, Cu) supported by g-Al_2O_3 substrate are important**
- **Combination of in-situ and ex-situ aberration corrected TEM, STEM-HAADF imaging, EDS and EELS spectroscopy**
- **Integration of high energy resolution EELS measurements with light source XAS**
- **Integration of experimental data with DFT calculations**



Understanding Chemical and Structural Changes in Battery Materials

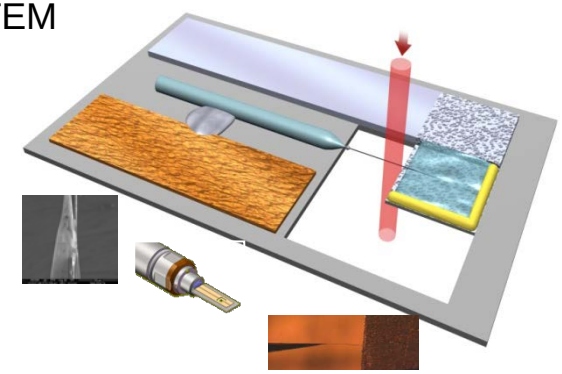
Challenges:

- **Discovering new materials with high capacity**
  - **- Less volume expansion**
  - **- Enhance charging and discharging rate**
  - **- Less irreversible microstructure formations**

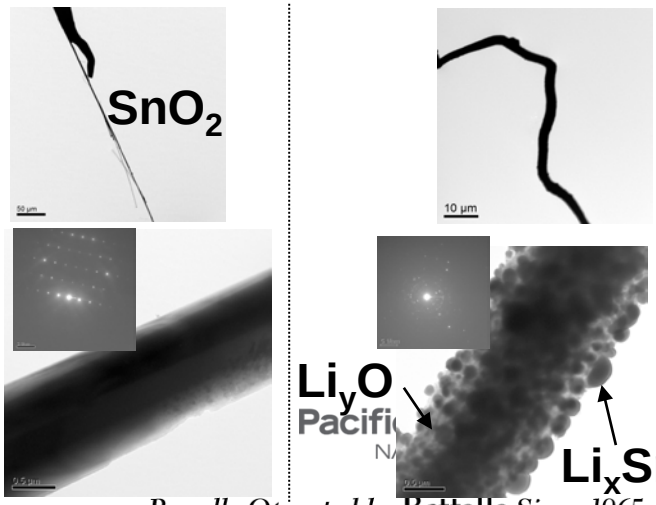
Approach:

- **Combination of nanomaterials and in-situ structural and electrochemical measurements**
- **- Combination of TEM, STEM and APT**
- **- Integration of high energy resolution EELS measurements with light source XAS techniques – chemical state charge transfer information**
- **- Integration of experimental data with calculations (DFT and MD)**

Basic conceptual design of battery using a single nanowire for in-situ TEM



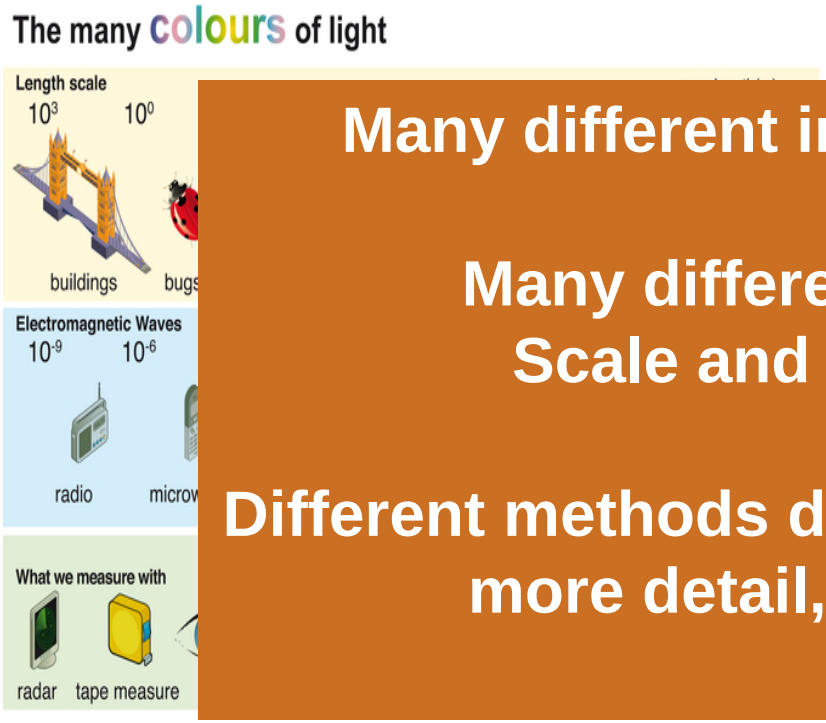
Microstructural evolution of SnO_2 anode upon initial charging

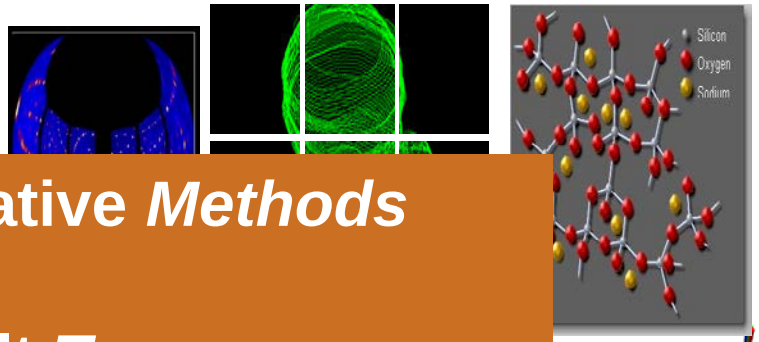


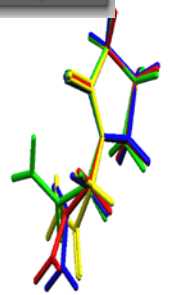
Investigative Methods

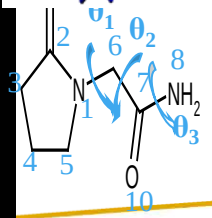
Result Representation

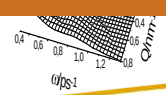
Methods











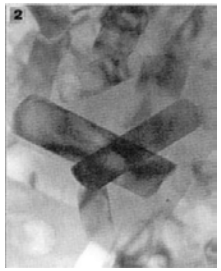
Many different investigative *Methods*

Many different result *Types* –
Scale and Representation

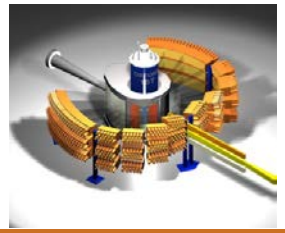
Different methods deliver different *Insights* –
more detail, new information

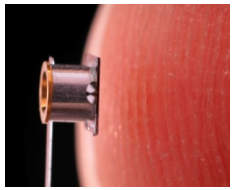
DIAMOND Light Source UK, 2010

Context sensitive, flexible Framework



Methods
- **Describe**
- **Map**
- **Classify**

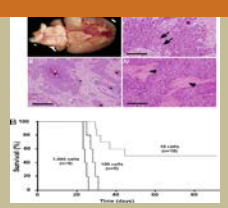


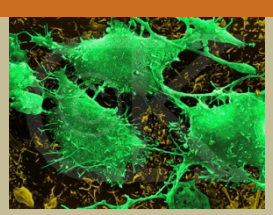


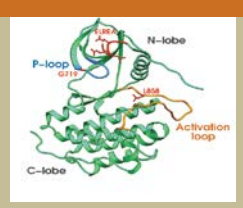
Different Components

We need a flexible framework that allows scientists to integrate, compare and contrast results from different investigative methods – and presents the results in a context that they are familiar with

Framework
Relating
Comparing
Combining
Synthesizing



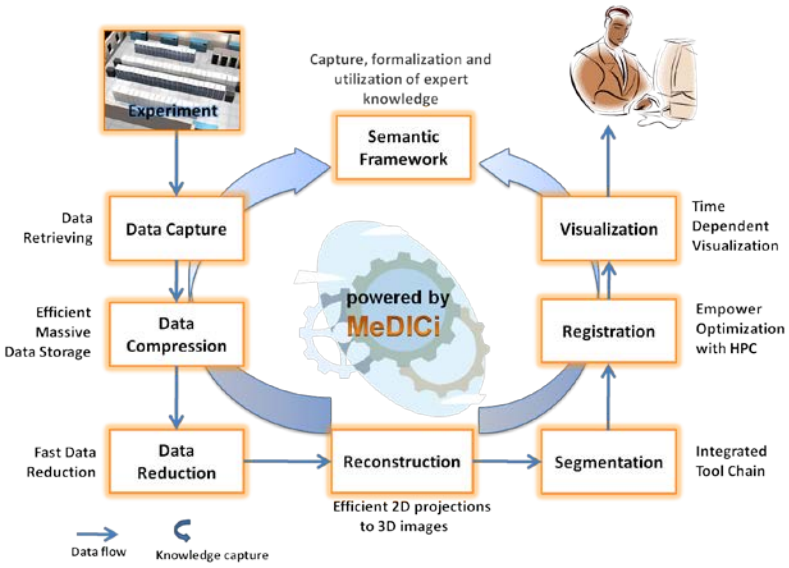




Different View Points – Shared Knowledge

Develop synergistic integration of multiple imaging, characterization, and simulation techniques

- **Phase 1: In collaboration with Thrust Areas 1+2 establish basic data analysis and handling framework**
- **Phase 2: Develop integrative analysis methods across different chemical imaging technologies with support of the wider community**
- **Phase 3: Evolve data handling and analysis methods to meet the real time and high data volume requirements of the new chemical imaging capabilities**

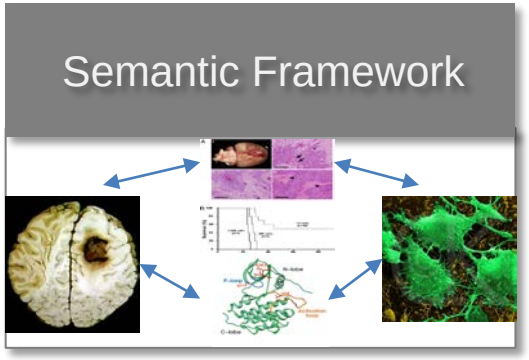


Pacific Northwest NATIONAL LABORATORY

Key Challenges

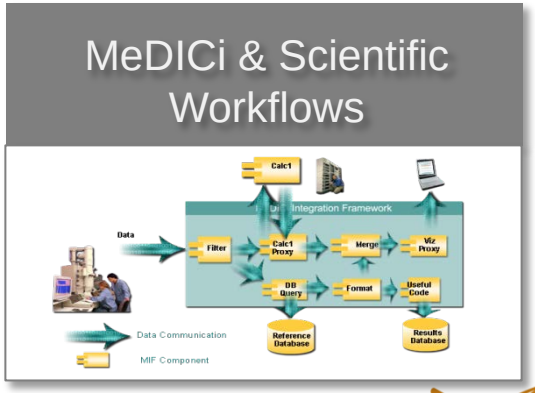
Challenge	Today	Tomorrow	Technologies
High Data Rate	5PB / Year LHC	3.5 PB / Day XFEL	Storage / Movement/ Analysis
Real Time Analysis	After Experiment -18 hours for 3D TEM reconstruction	Experimental Steering through real time analysis during experiment	Real Time, Parallel, Data Intensive Computing HW + SW
Integration across different Imaging Technologies	One off integration of 2 techniques	100's of combinations possible	Conceptual Relation, real time integration
Integrating across scales	On related scales only	Nano to Macro	Conceptual Relation
Geographical distribution of experimental sources to be integrated	None	10's – flexibly combined	Distributed Computing Paradigm

Proposed Core Framework Development (1)



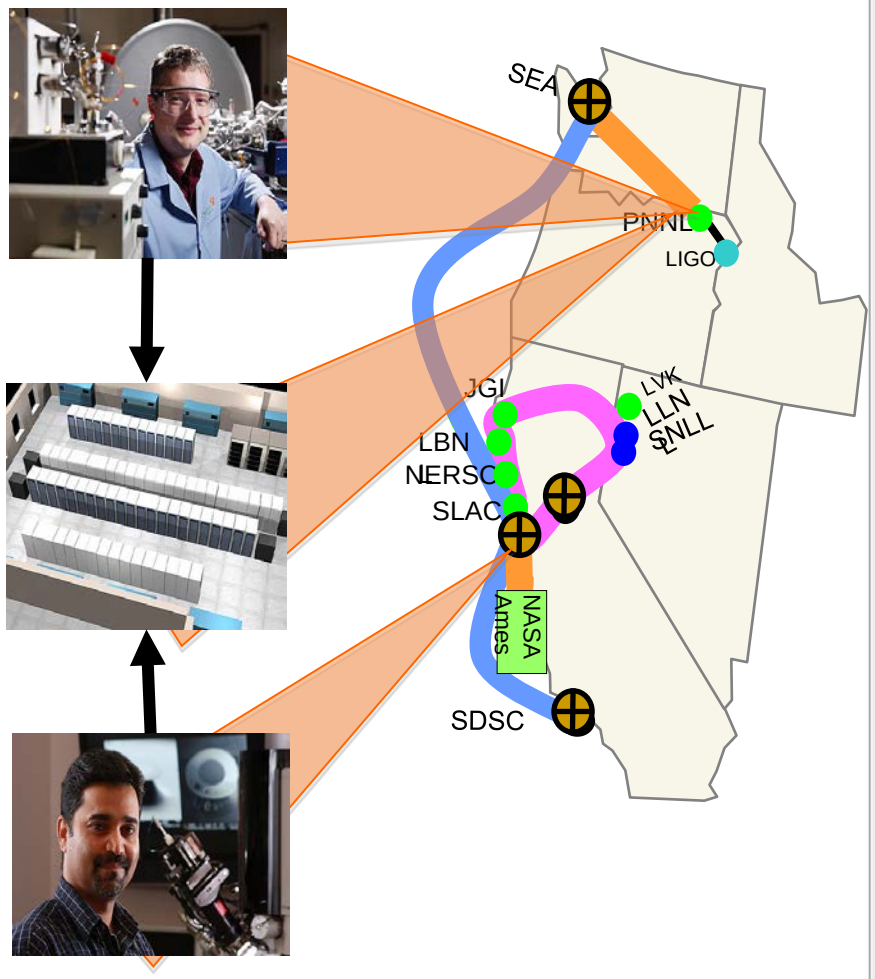
- **Formal characterizations of the methods, instruments, samples, analysis processes and associated data products.**
- **Formalized topology of the methods, their contribution and constraints.**

- **Flexible creation of data intensive workflows.**
- **Managing complex and intensive data exchange as well as rapid integration of data sets spanning different spatial and temporal scales**



Proposed Core Framework Development

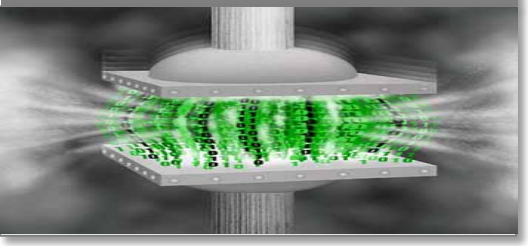
Data Capture and Distributed Computing Empowered by MeDICi



- **Leveraging MyEMSL framework providing workflow, data capture, metadata capture, a central data repository, and tools for data discovery.**
- **High volume data transfers.**
- **Taking analysis to the data via distributed computing.**

Proposed Example Components (1)

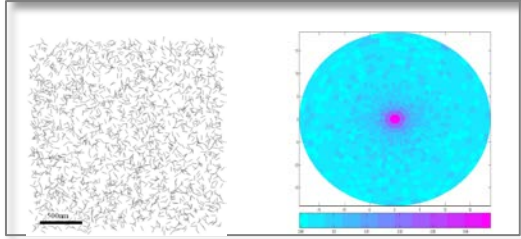
Data Compression



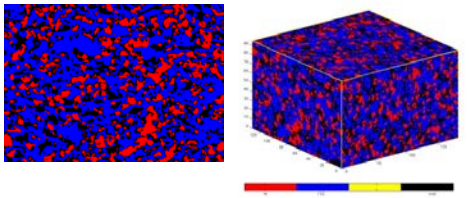
- Appropriate lossless and lossy compression algorithms
- High compression ratio with low computational overhead

- Reduce noise and smooth data
- Reconstructions will contain the most significant information, are feature-accentuated

Data Reduction



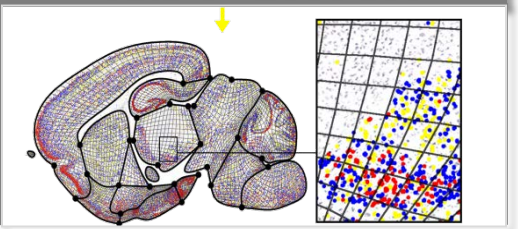
Reconstruction



- Accurate re-construction of high volume data in real time
- Combine correlation functions with parallelized filtered back projection

Proposed Example Components(2)

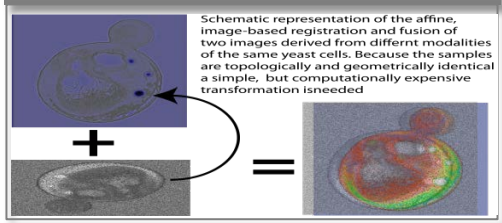
Segmentation & Feature Association



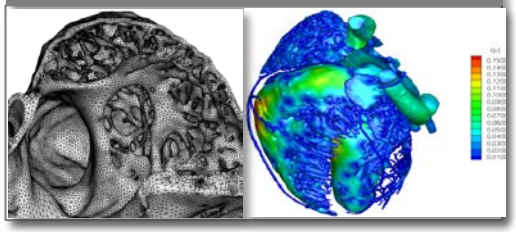
- **Scalable differential operators and primitives that can be combined at run time for application-specific chemical signature and feature recognition**

- **Enabling accurate localized comparisons between experimental datasets from different chemical imaging techniques at high resolutions**

Registration



Visualization & Analysis



- **Real time, remote, in-situ, high data volume 3+4D visualization**

Questions ?