

EMSL Use Case

The Environmental Molecular Science Laboratory (EMSL) is a national user facility that provides world-class fundamental research capabilities for scientific discovery and the development of innovative solutions to the nation's environmental challenges and energy production. EMSL's distinctive focus on integrating computational and experimental capabilities as well as collaborating among disciplines yields a strong, synergistic scientific environment. Bringing together experts and state-of-the-art instruments critical to their research under one roof, EMSL has helped thousands of researchers use a multidisciplinary, collaborative approach to solve some of the most important national challenges in energy, environmental sciences, and human health. These challenges cover a wide range of research, including synthesis, characterization, theory and modeling, dynamical properties and environmental testing. EMSL houses a collection of over 100 state-of-the-art capabilities and is operated by PNNL for the DOE's Office of Biological and Environmental Research.

EMSL overall Data and Storage Needs:

Time	Daily Data Size	LAN Transfer	WAN Transfer
Today	6.5 TB/day	5 TB/day	200 GB/Months
2-5 years	20-40 TB /day	40 TB/day	600 TB/months
5+ years	100 TB/day	200 TB/day	3 PB/months

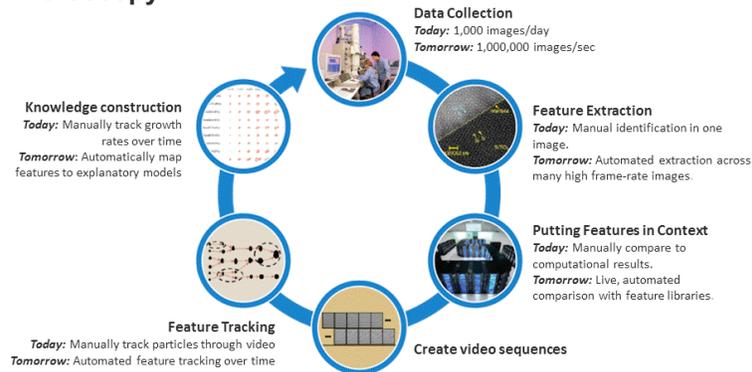
Specific Data Challenges:

Data Volume and Rate Challenge - Example: EMSL houses a range of state of the art Transmission Electron Microscopes (TEM), a fast growing imaging method, with currently ~600 instruments worldwide, increasing by ~50/year. TEM's usually produce a series of 2kx2k or 4kx4k images, at increasingly higher rates due to new detector technologies.

Time	Instrument	Data Rates	Burst	Volume
Today	Conventional	100-1000 images/day	single	
1-2 years	Environmental	1000 images/sec	10 min	11-13 TB/day
2-3	Dynamic	1,000,000 images/sec	10 sec	20 TB /burst

At present the full analysis and interpretation of one experiment (using only a limited number of images) can take up to 6 months. The aim is however to gain a deeper understanding of physical, chemical and biological processes, by analyzing all of the generated images. Furthermore science would greatly benefit from real time data processing capabilities, to enhance the quality of data taking and enabling the interaction with the sample based on real time results. Please find below a graphic of a typical TEM analysis workflow.

Exemplary Challenges in Transmission Electron Microscopy



Example Transmission Electron Microscope – The Analysis of one experiment (selected images only) takes today 6 months, in the future all images should be analyzed as rapid as possible with an increase from 1000 images /day to 1,000,000 images/sec

12

TEM is only one example of the complex, high data rate analysis workflows required by increasingly data intensive instrumentation in scientific user facilities.

Multi Modal Challenge - EMSL experiments typically can involve multiple instruments (multi-modal) that can consist of a number of experimental techniques and computational simulation. Depending on the science being studied this will require new techniques and tools for data assimilation and integration. Data assimilation combines a number of data sources for comparison including numerical simulations and observational data, using statistical methods and applied mathematics techniques. Data integration collects disparate data sets for [meta-analysis](#) (methods for contrasting and combining results from different studies, etc.). This type of data integration is especially challenging for many scientific disciplines where there are many different data types produced in these fields. In addition some of those data and analytical tools might be found at other sites, both openly accessible facilities and private user collections.

Cost Challenge - An important concern for user facilities is the cost associated with maintenance and support of software stacks. Potential solutions might be 1) coordinating with the user facilities to generate open source efforts with the larger community (i.e. Semantic Physical Sciences, NWChem, etc.), and 2) identify reusable and generally applicable software components (i.e., analysis tools, etc.).

Cross Facilities Collaboration - In addition, new data plans, data sharing and data policies will be coming sometime in the near future and facilities need to coordinate as much as possible to have a more consistent data plans, sharing and policy to enable data sharing. It is conceivable that an experiment could be done at a beamline (i.e. APS) and the same sample be taken to a user facility at a different laboratory and if the data plans, etc. are not consistent this could cause issues for the user to obtain all of their data. Similarly the users need to be able to easily access and analyze their data across the different facilities used and do so jointly with their collaborators.