LightMAT
Lightweight Materials
CONSORTIUM

LightMAT DataHUB

Chitra Sivaraman

Industry Kick-Off Meeting, Detroit
February 09, 2017

Energy Materials Network
U.S. Department of Energy

LightMAT Data Management: Capabilities

- Web Portal – https://lightmat.org/
- DataHUB
  - Data Archive and Storage
  - Data Collection
  - Data Sharing
  - Security
  - Discovery
- Acceleration Engine
  - Standards
  - Integration and Transformation
  - Publication→Science and Discovery

Mission and Vision

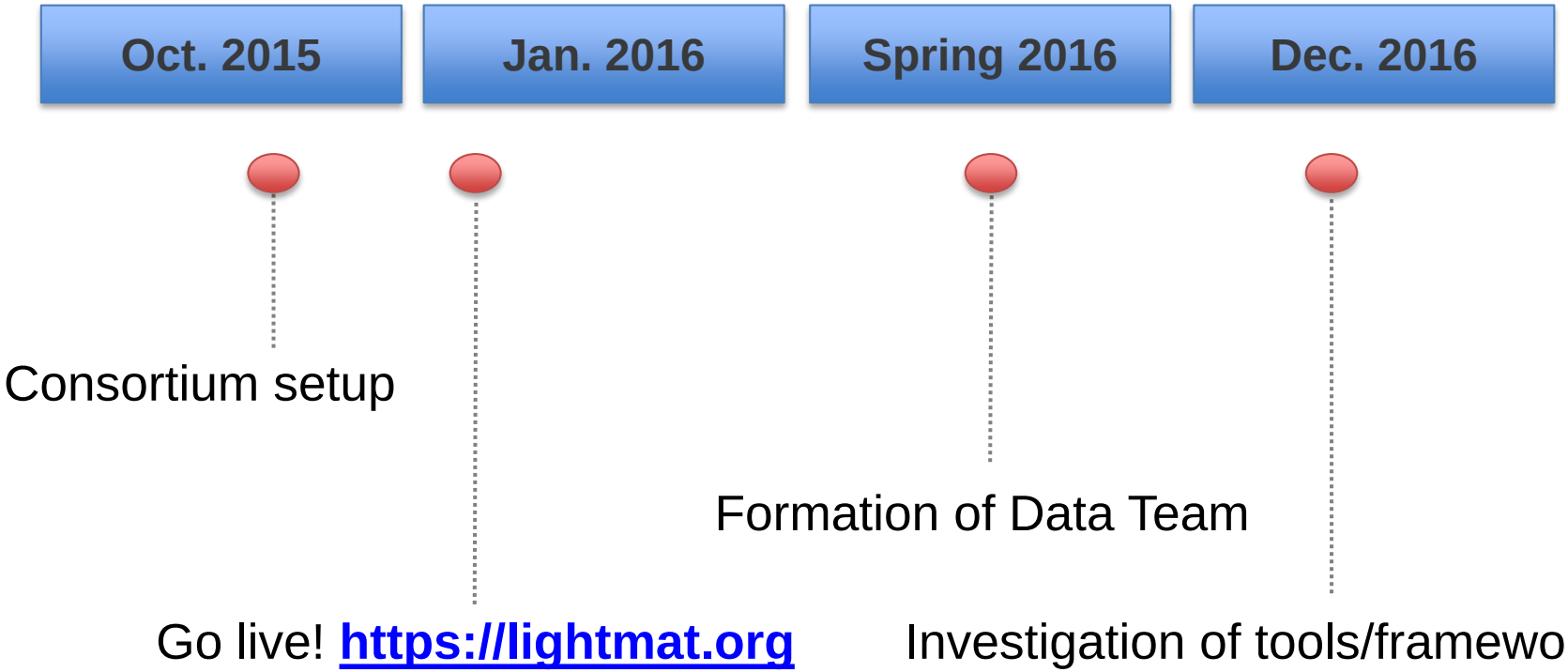
Mission

The LightMAT (LM) DataHUB's (DH) objective is to provide **secure, timely, easy and open** access to data produced by LightMAT consortia.

Vision

The DH will **collect, store, catalog, process, preserve, and disseminate** all significant LM data (codes, models, experimental, simulated data and journal publications) with state-of-the-art technology while conforming to or helping define industry data standards and leveraging existing tools.

LightMAT DataHUB: Timeline



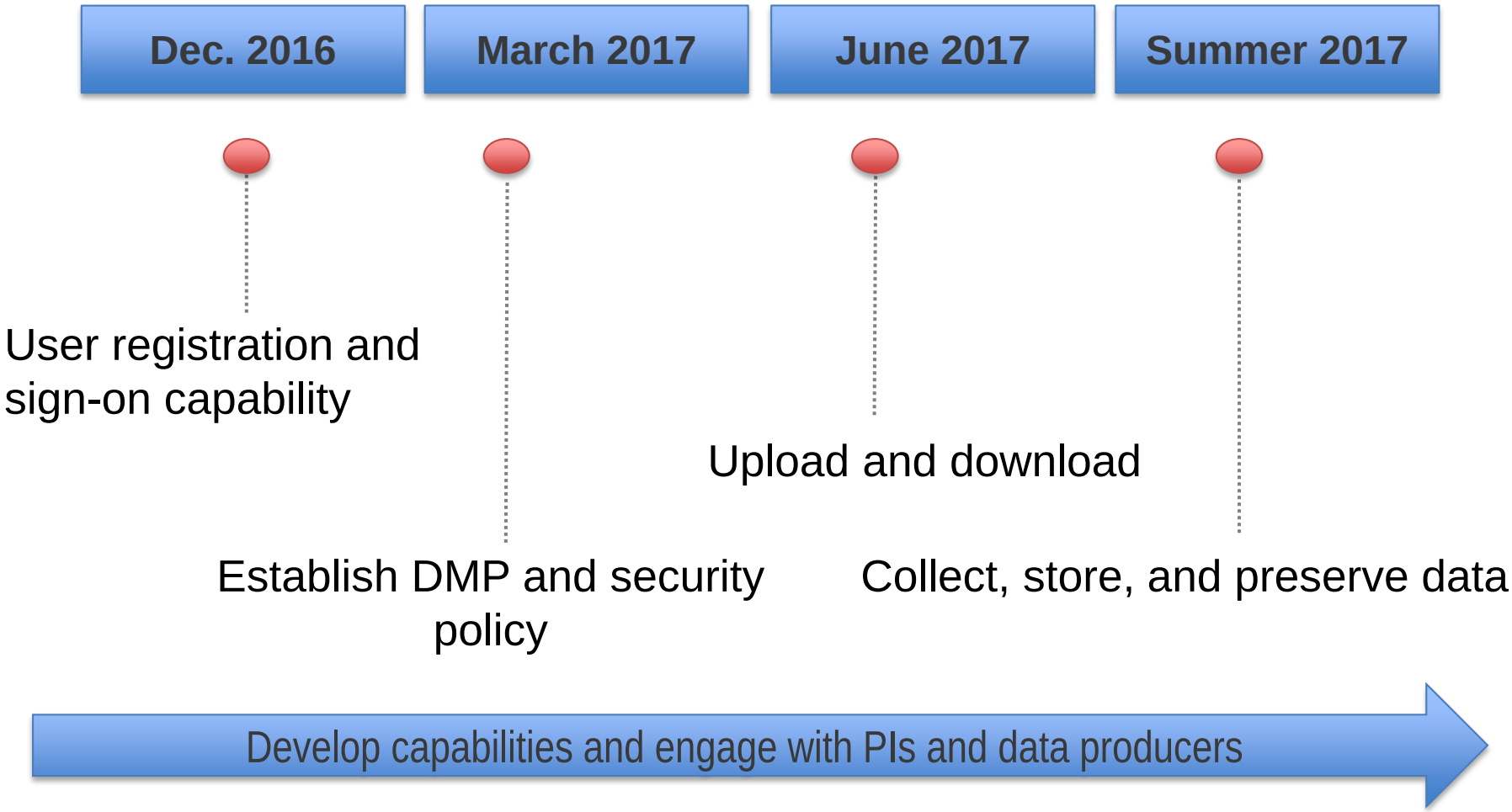
LightMAT DataHUB Team

Name	Organization
Jeff Florando	LLNL
Richard Karnesky	SNL
Matt Macduff	PNNL
Kristin Munch	NREL
Amit Shyam	ORNL
Dilip Singh	ANL
Chitra Sivaraman	PNNL
Xin Sun	PNNL
Shrikant Nagpure	INL

- Planning Process
  - Transparent process with experts from eight laboratories.
  - Bi-monthly data team calls.
  - LightMAT Workshop at PNNL held on April 05, 2016.
  - Provide guidance on what tools to host on the portal.

Platforms Reviewed
Granta, Citrine, DSpace, Materials Common, CKAN, DKAN, HUBzero, Socrata, DataVerse, FedoraCommons

LightMAT DataHUB: Near-term Action Items



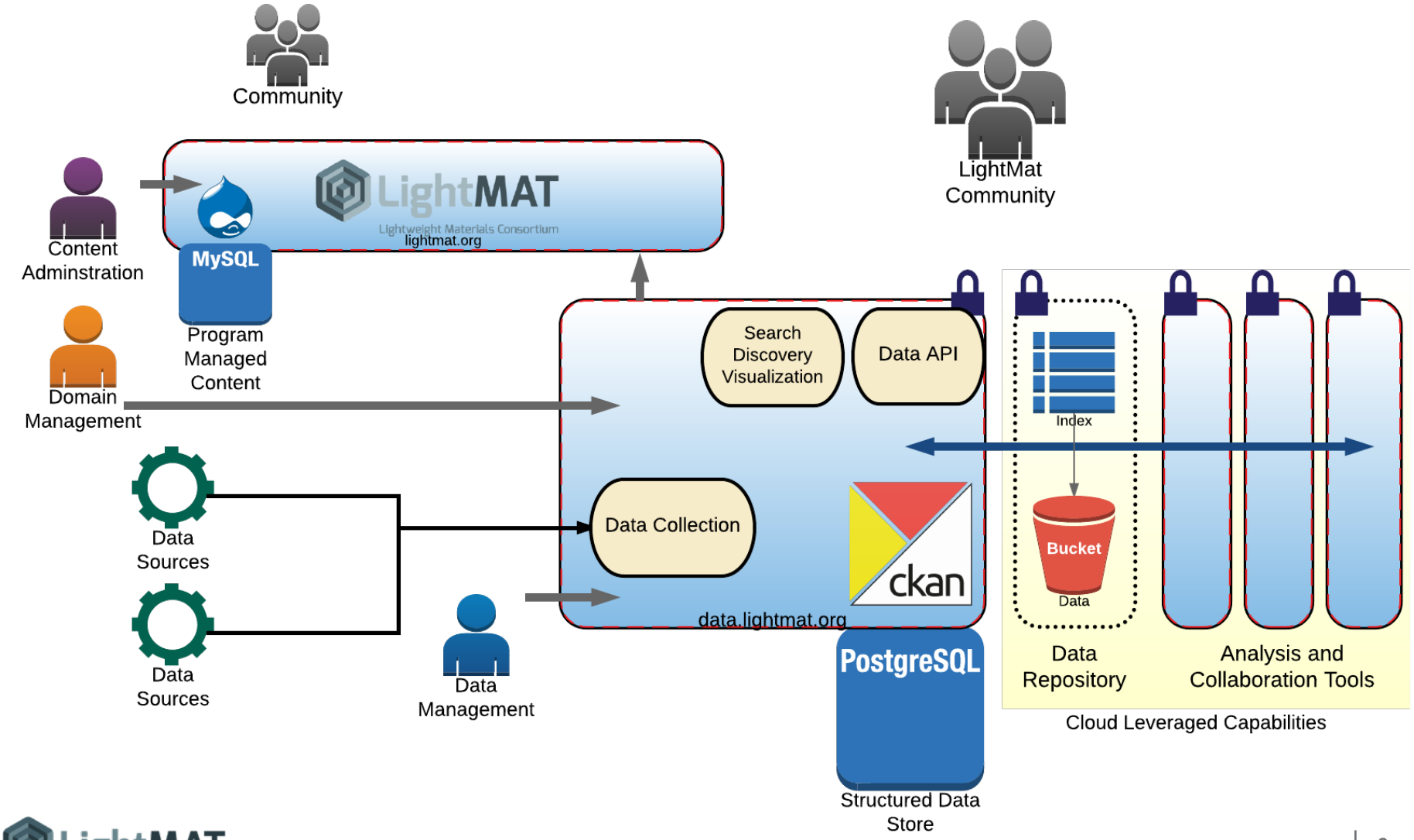
Key Considerations: Making the DataHUB Effective

- Engage with USAMP and project PIs
  - What are the data sources?
  - Who are the data consumers?
  - Why do they need the data (e.g., use case scenarios)?
  - In what format/interval is the data useful to the consumer?
  - What are the existing data formats/standards?
  - Open versus Embargoed versus Proprietary data?

Data Management Plan (DMP)

- Meet DOE's requirement that all projects within LightMAT have a DMP
  - Long-term archival preservation of 10 TB of data per project
  - Web-based search and discovery capabilities for user access
  - Digital Object Identifiers (DOIs)
  - Data is classified as **Public**, **Embargoed**, or **Restricted**
  - Continuous, bulk, or disk-based data submissions
  - Metadata repository of searchable data attributes
  - Host information on materials, properties, research data (raw, processed, derived, and reduced), graphs, tables, models, algorithms, etc.
  - Data formats will include most common formats such as ASCII, HDF, PNG, JPEG, PDF, NetCDF, etc.

DataHUB: Architecture



Data on the Web – Best Practices

Requirement	Description
Machine readable format	Data Should be available in a machine readable format that is adequate for its intended or potential use
Standardized format	When possible data should be available in a standardized format. Through standardization interoperability is also expected.
Multiple representations	Data should be available in multiple representations
Metadata Documentation	Local metadata and community vocabulary metadata should be clearly defined
Community Vocabularies	If available community vocabularies should be used to help facilitate interoperability between local metadata definitions.
Persistent identifiers	An identifier for a particular resource should be resolvable and associated for the foreseeable future with a single resource or with information about why the resource is no longer available.
Data Quality Metrics	Data should be associated with a set of standardized, objective quality metrics
Data Provenance Availability	Data provenance information should be available. Provenance data i s a type of metadata, so all metadata requirements also apply here.
Data Security	Data with security requirements should adhere to the owner's data governance policies
Data Citable	It should be possible to cite data with digital object identifier.
Data Archival	All designated data should be archived
Data Catalog	Data should be registered in catalog
Catalog Search	Data should be searchable in a catalog
Data Usage	Data usage metrics should be tracked

Capabilities within DataHUB

Capability	Normal	Advanced
Archival	Yes	Yes
Uploads	Automated/Ad-hoc	Real-time
Cloud Storage	10 TB	10 TB+
Meta-data Search	Datasets/Projects	File level
Download	Web request/manual	Instant real-time access
Access limitations	Public/Embargo/Restricted	Public/Embargo/Restricted
User Support	Yes – not real time	Yes
Meta-data support	Dataset naming and facet definition	File format standards, versions, relationships, quality flags.
Operational Support	Internal/Email	Monitoring

Key Questions

- Should the DataHUB be **processing** the data?
- Should we enable **visualization** of datasets?
- Should we host tools, codes, model runs, or simulation results?
- Do we need to create metrics?
- Are there existing data standards?
- Should we enable anonymization, quality checking, integration, validation as base capability for projects?
- Should we leverage other machine learning, natural language processing or deep learning frameworks?


Comments?
Questions?