ElectroCat
Electrocatalysis Consortium

Building an ElectroCat Data Hub

IAN FOSTER **Ben Blaiszik**

Argonne National Laboratory & University of Chicago

Kristin Munch, NREL
John Perkins, NREL
Robert White, NREL

Network Requirements (for a EMN Consortium)

1. ***WORLD CLASS MATERIALS CAPABILITY NETWORK:*** Create and manage a **unique, accessible set of capabilities** within the DOE National Laboratory system
2. ***CLEAR POINT OF ENGAGEMENT:*** Provide a **single point-of-contact** and concierge to direct interested users (e.g. industry research teams) to the appropriate laboratory capabilities, and to **facilitate efficient access**.
3. ***DATA AND TOOL COLLABORATION FRAMEWORK:*** **Capture data, tools, and expertise** developed at each node such that they can be **shared and leveraged** throughout the EMN and **in future programs**. Establish data repositories and, where appropriate, distribute data to the scientific community and public. Accelerate learning and development through data analysis using advanced informatics tools.
4. ***STREAMLINED ACCESS:*** Facilitate **rapid completion of agreements** for external partners, and aggressively pursue approaches to reduce non-technical burden on organizations seeking to leverage the EMN for accelerated materials development and deployment.

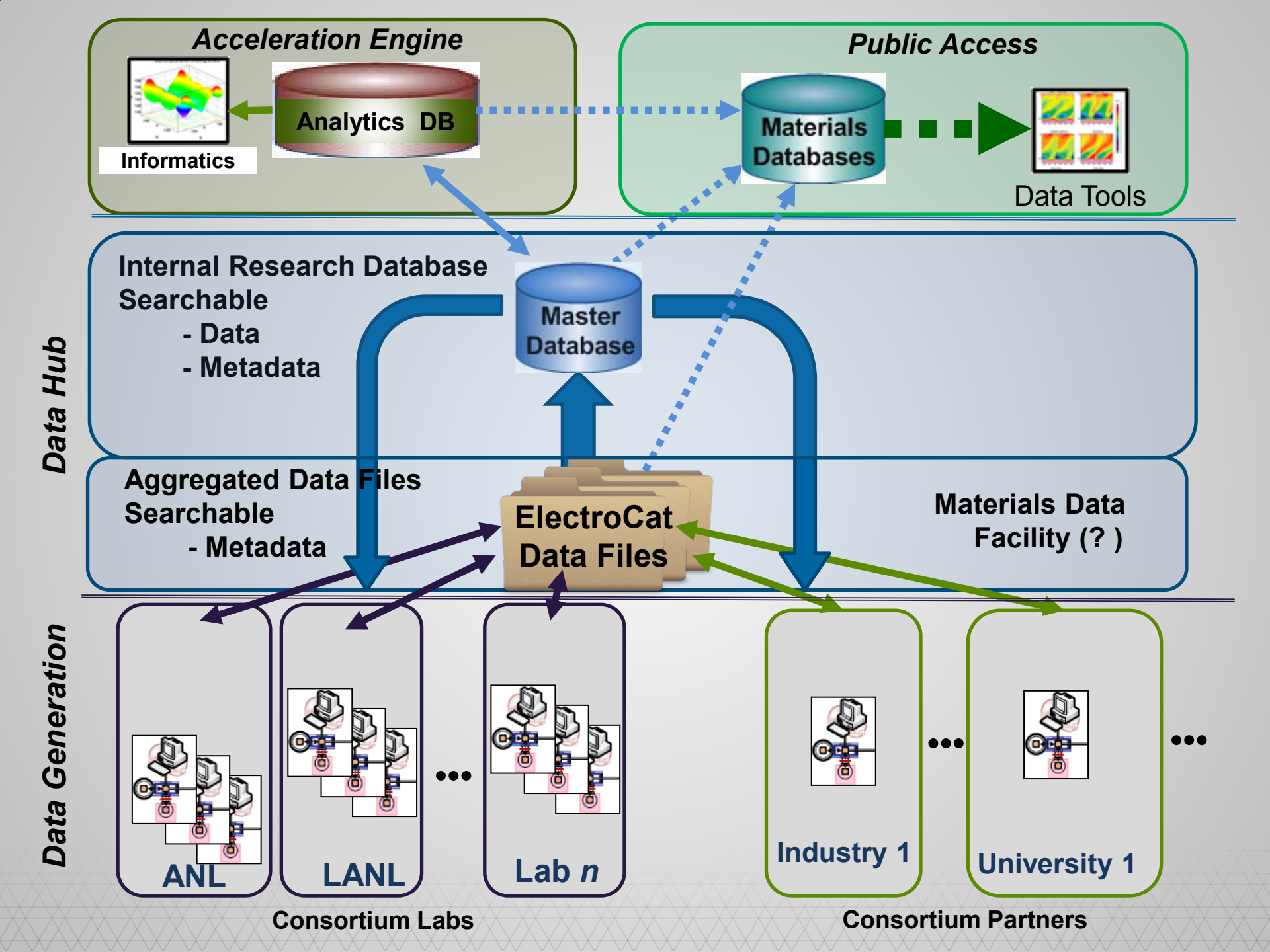
ELECTROCAT DATA MANAGEMENT

This [data] system will be comprised of the following three components:

1) ElectroCat *Portal*

2) ElectroCat *Data Hub*

3) ElectroCat *Acceleration Engine*


Acceleration Engine
Informatics
Analytics DB
Public Access
Materials Databases
Data Tools
Data Hub
Internal Research Database
Searchable
- Data
- Metadata
Master Database
Aggregated Data Files
Searchable
- Metadata
ElectroCat Data Files
Materials Data Facility (?)
Data Generation
ANL
LANL
Lab n
Industry 1
University 1
Consortium Labs
Consortium Partners

MORE ON THE *DATA HUB* ….

- "the Data Hub … will **house data** generated through use of ElectroCat facilities to make it publicly available"
- "also **develop protocols** that researchers must follow when submitting data to the Hub (data formats and information on test conditions…)
- "[it] will **provide centralized data/model storage** capability for various ElectroCat generated information …: 1) Codes/models; 2) Experimental/simulated data; and, 3) Journal publications/presentations tagged with the appropriate provenance"

DATA HUB (2)

- … provide **keyword-based searching** capability and **publish curated metadata** to commercial search engines
- … **virtual linkages** to other relevant materials databases
- a **security system** with flexibility to accommodate access and use by different classes of projects and data – from highly sensitive proprietary, to embargoed, to publicly available data

OBSERVATIONS

- An exciting and appropriate vision for data-driven, reproducible science & engineering
- Challenging to realize in its entirety due to variety of data types and usage modalities
- Knowledge of how people are going to use the DataHub is still developing

DATA STRATEGY

We propose a “tiered approach” to data management, evaluating & leveraging existing resources where applicable, and incrementally developing capabilities where needed:

1) **Basic consortium collaboration:** Data submission, metadata, basic search, continued metadata development, data sharing
2) **Consortium materials database:** Database of relevant materials properties and data
3) **Public access:** Release data to public
4) **Advanced analysis:** Organize and structure data for data mining and informatics approaches

A PROPOSED APPROACH

The Materials Data Facility (MDF) could provide several of the data components for ElectroCat:

- **Year 1: Basic consortium collaboration**
  - Explore Argonne/UChicago/NIST Globus-based **Materials Data Facility** as configurable publication pipeline and data repository (details below)
  - Evaluate **ElectroCat-specific use cases** to determine MDF-Globus applicability and any need for custom features
  - Leverage **NREL expertise** to support MDF configurations: specific data types, formats, ingestion processes, metadata design, accessibility
- **Out-years**
  - Develop policies and configure MDF for public data release
  - Develop API into MDF, enabling programmatic data accessibility
  - Leverage MDF connectivity to express MDF data into analysis-ready Consortium Materials Database

The Materials Data Facility

Ben Blaiszik (blaiszik@uchicago.edu),
Kyle Chard, Rachana Ananthakrishnan
Michael Ondrejcek, Kenton McHenry
PIs: Ian Foster (foster@uchicago.edu), Steven Tuecke, John Towns

materialsdatafacility.org
globus.org

Materials Genome Initiative

NIST National Institute of Standards and Technology U.S. Department of Commerce
CHiMaD Center for Hierarchical Materials Design
THE UNIVERSITY OF CHICAGO
ILLINOIS UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
Argonne NATIONAL LABORATORY

What is MDF?

We are developing production services to make it more simple for materials datasets and resources to be ...

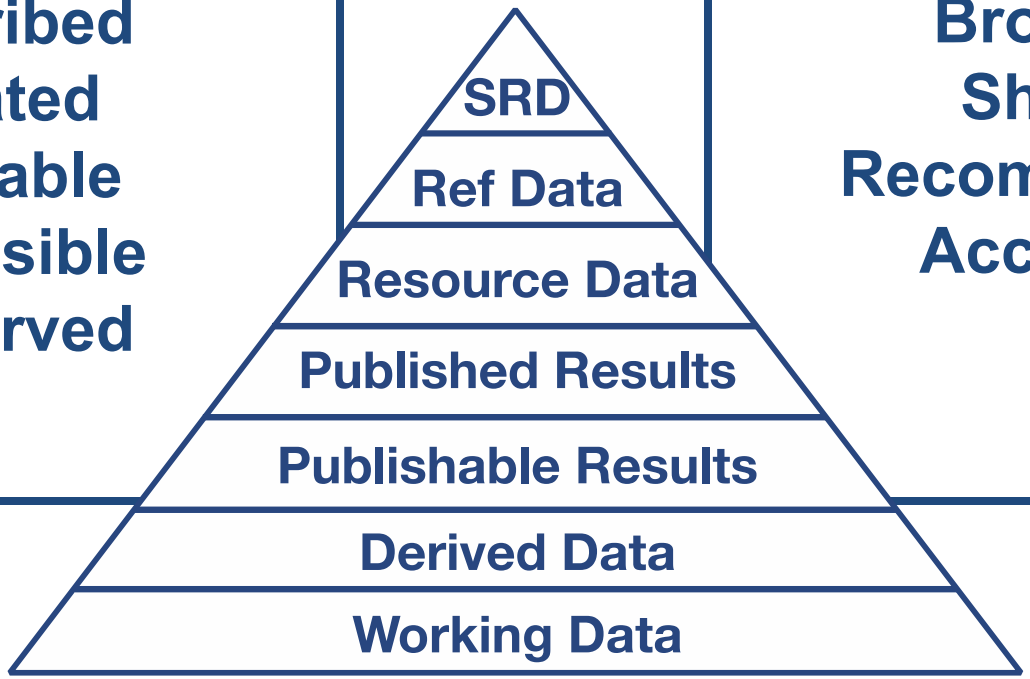
Published

Identified
Described
Curated
Verifiable
Accessible
Preserved

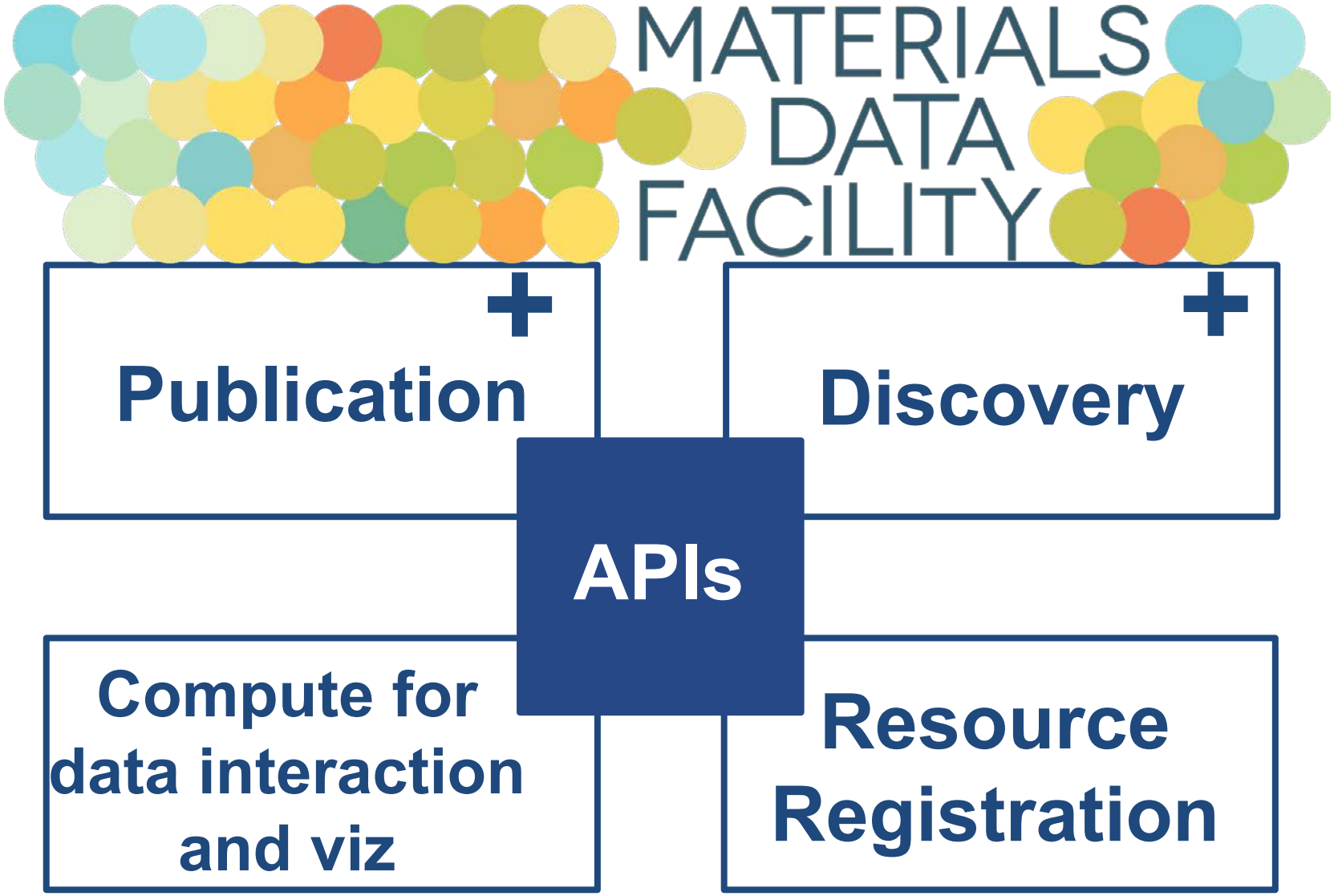
and

Discovered

Searched
Browsed
Shared
Recommended
Accessed



Data Service Infrastructure



Built on Globus Services

Identity management

- create and manage a unique identity linked to external identities for authentication

User groups

- Manage user group creation and administration flows
- Share data with user groups

Data publication

Data transfer

- High-performance data transfer from a web browser
- Optimize transfer settings and verify transfer integrity
- Add your laptop to the Globus cloud with Globus Connect Personal

Data sharing

- Share directly from your storage device (laptop or cluster)
- File and directory-level ACLs

MDF Data Publication

- Leverages Globus production services for Auth, sharing, groups, transfer
- Can handle small or large data (e.g. TB-sized datasets)
- Features customizable and shareable metadata schemas
- Allows for distributed data and metadata storage
- Capable of minting unique identifiers for datasets (e.g. DOI, Handle)
- Includes curation workflow tools (e.g. approval steps before data is made public)
- Has 100 TB allocation that could be used immediately without IT overhead (for public-facing data)
- Includes Web-UI and a planned API
- Will be interfaced with multiple national materials efforts over time

MDF Discovery Service

- Full text search across all metadata fields, even custom metadata fields, is available now
- Search by filenames, typed searches, range queries, and file contents for specific file types coming soon
- Goal: Intuitive search (e.g. Google-style) with support for more complex range queries and faceting (e.g. Amazon-style)

Future...

Spotlight for all data you have access to regardless of location

Integration with the Community is Key