## Geoscience Data Management Infrastructure

### Data Providers
- Generate new data sets, derived products, models, software, and documents. All need to be cataloged and made available to co-researchers.

### Centralized Metadata Catalog
- Each data set is archived within Lab repositories. Structural metadata is documented with the data set.
  - An initial data management infrastructure (DMI) that includes one Centralized Data Catalog (CDC) and a Distributed Data Repository (DDR) has been created to manage the geophysical data sets.
  - The current catalog is hosted by LLNL with access granted to SNL, LANL, and PNNL participants.

### Distributed Data Repository
- The repository system is a collection of geographically distributed sub-repositories.
- Each Lab hosts the data set for which they are responsible.
- The repository must meet requirements described in the Technical Requirements.

### Process
- The flow chart below shows the process followed to define the DMI.

### Ongoing Work:
- Repository build will be a focus of work through FY22.
- A critical aspect of this work is managing a range of data sensitivities and the awareness of restrictions in sharing data between National Labs. Efficient data transfer mechanism needs to be defined accordingly.
- Discoverability is enhanced when data sets are metadata tagged using established schema e.g., DCAT. Data storage such as Object Store could provide a way to meet this need.

### Data Management Group
- Core DMG: Two members from each participating Lab. Others brought into the team as expertise needed.
- System Engineers to verify the system meets requirements.
- Data Coordinators responsible for helping data providers submit data to the catalog and repository system.

### Data Coordinators
- Two members from each participating Lab. Others brought in as expertise needed.

### Data Providers
- Per the Data Policy for the projects, each data set is hosted by one Host Lab. That data set is the primary, or authoritative, data set. The Host Lab takes on the responsibility of long-term curation of the data set. The Host Lab facilitates data transfer upon request from a researcher.

### Centralized Metadata Catalog
- Data sets and products generated within the project are recorded in a centralized catalog hosted by Lawrence Livermore National Lab.
- Each data record is assigned a unique ID in the catalog.
- Descriptive metadata is entered by a provider and coordinator.
- Once a data set is identified, participants can request it from the host Lab.
- Authorized access only.

### Distributed Repository System
- The repository system is a collection of geographically distributed sub-repositories.
- Each Lab hosts the data set for which they are responsible.
- The repository must meet requirements described in the Technical Requirements.
- Innovation in catalog platforms and data management architecture is rapid, the proposed architecture is modularized and flexible allowing emerging tools to be implemented providing responsiveness to research needs.

### Diagram

![Geoscience Data Management Infrastructure Diagram](dummy.png)

### Table: Technical Requirements

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Sets</td>
<td>archived within Lab repositories.</td>
</tr>
<tr>
<td>Structural Metadata</td>
<td>documented with the data set.</td>
</tr>
<tr>
<td>Centralized Data Catalog</td>
<td>hosted by LLNL, with access granted to SNL, LANL, and PNNL participants.</td>
</tr>
<tr>
<td>Distributed Data Repository</td>
<td>collects geographically distributed sub-repositories.</td>
</tr>
<tr>
<td>Data Transfer</td>
<td>facilitated by Host Lab.</td>
</tr>
<tr>
<td>Data Use</td>
<td>controlled by technical requirements.</td>
</tr>
<tr>
<td>Data Sensitivity</td>
<td>managed by the authorizing process.</td>
</tr>
<tr>
<td>Data Access</td>
<td>determined by the authorizing process.</td>
</tr>
</tbody>
</table>

### Key
- #A: Authorizing process
- #B: Technical requirements
- #C: User access

---

**Kathleen Hodgkinson¹, Rebecca Rodd², Jennifer Mendez³, Richard Stead⁴, Jonathan MacCarthy⁴, Rose Borden⁵, Amanda Price⁶, Erin McCann⁷, Ian Smith⁷, Michael Hofmockel⁸, Jose Falliner⁹**


---

**Notes**
- Data sets archived within Lab repositories.
- Structural metadata is documented with the data set.
- The current catalog is hosted by LLNL with access granted to SNL, LANL, and PNNL participants.
- Innovation in catalog platforms and data management architecture is rapid, the proposed architecture is modularized and flexible allowing emerging tools to be implemented providing responsiveness to research needs.

---

**Questions**
- How does the centralized metadata catalog facilitate the sharing of data?
- What is the role of the distributed repository system in managing geographically distributed sub-repositories?
- How are data transfer requirements determined and managed in this system?