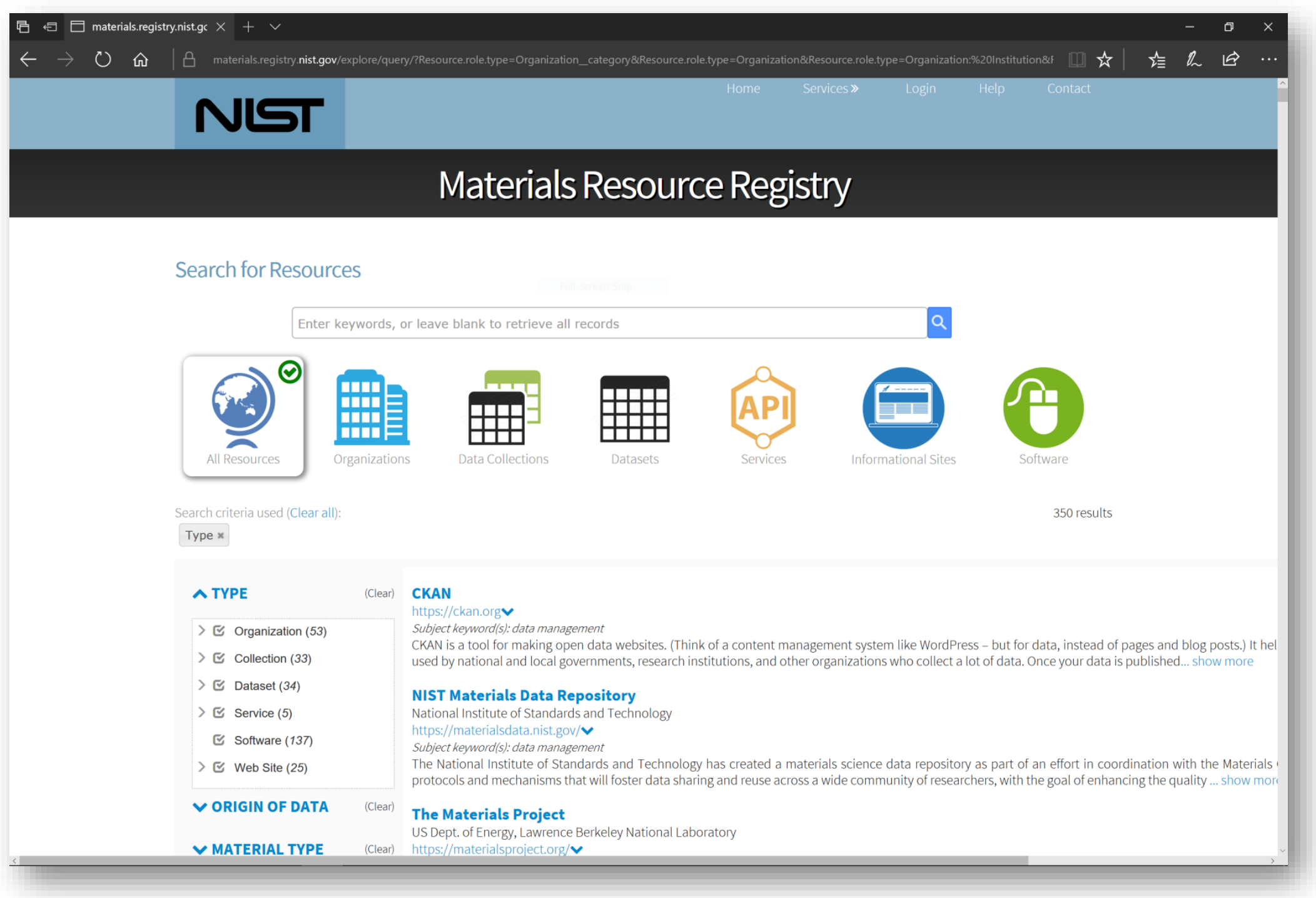
The Configurable Data Curation System (CDCS)

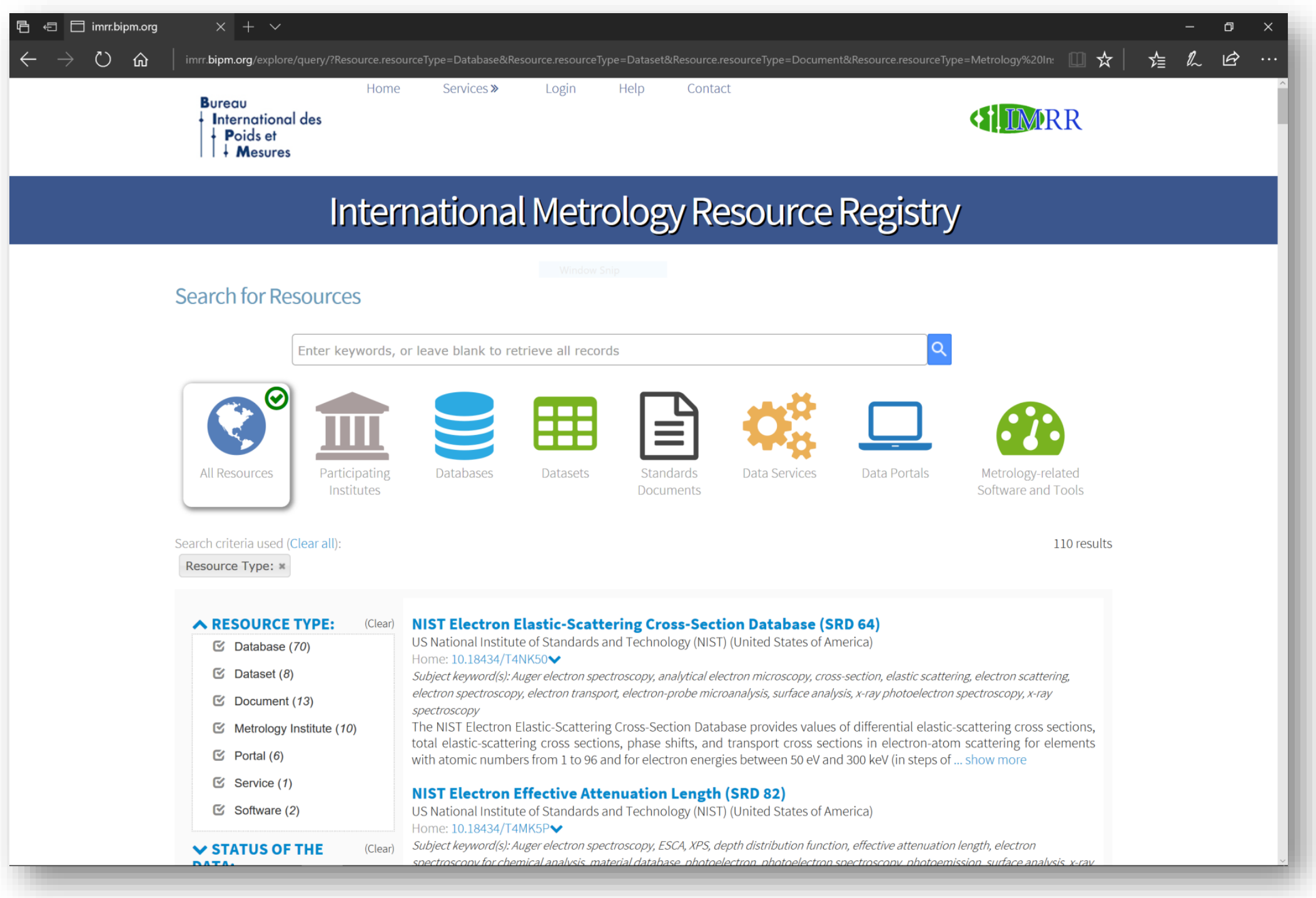
C.A. Becker, M.C. Brady, K.G. Brady, C.E. Campbell, A.L. Catel, P.J. Dessauw, A.A. Dima, G.R. Greene, R.J. Hanisch, B. Long, M.W. Newrock, R.L. Plante, P.F. Rigodiat, X. Schmitt, G. Sousa Amaral, Z.T. Trautt, J.A. Warren

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

Registries

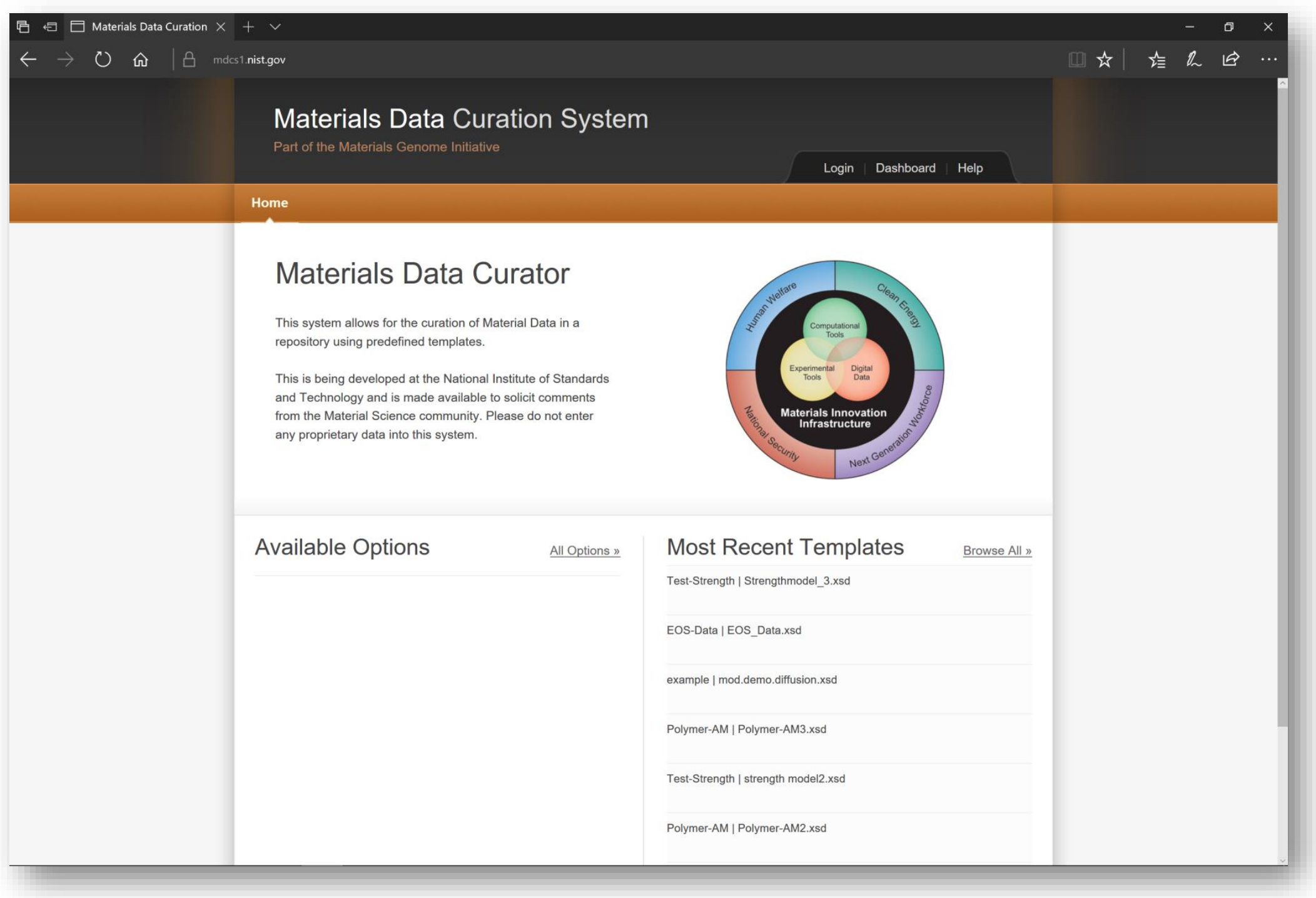
The CDCS empowers communities to deploy and operate specialized registries by enabling **Findability** of data and resources. Examples for materials and metrology are shown below.





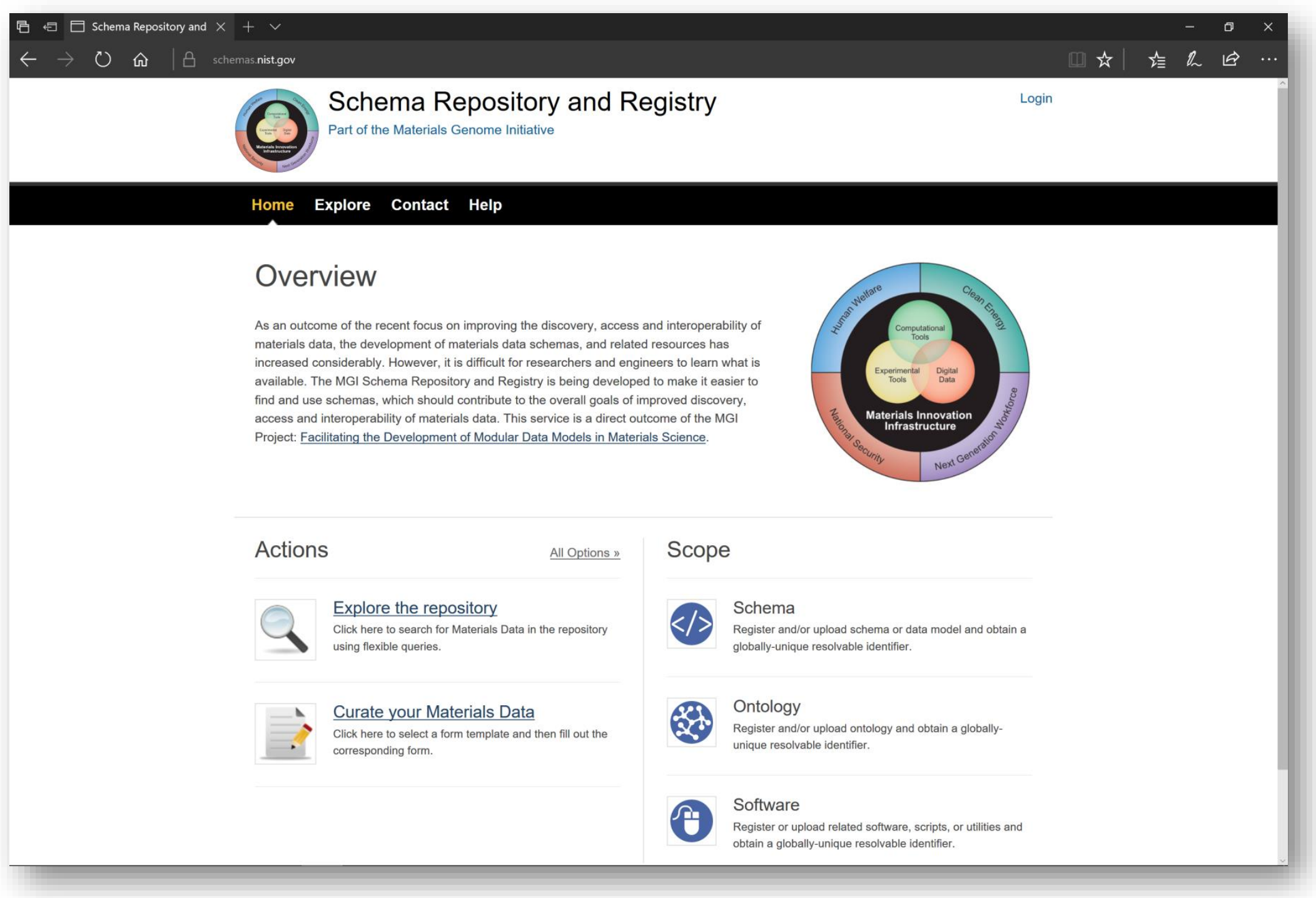
Data Ingest

The CDCS enables **Acquisition** and **Storage** of data in **Interoperable** formats early in the data lifecycle. For example, the Materials Data Curation System[2] (MDCS) is shown below. The MDCS API enables automated curation.



Schemas

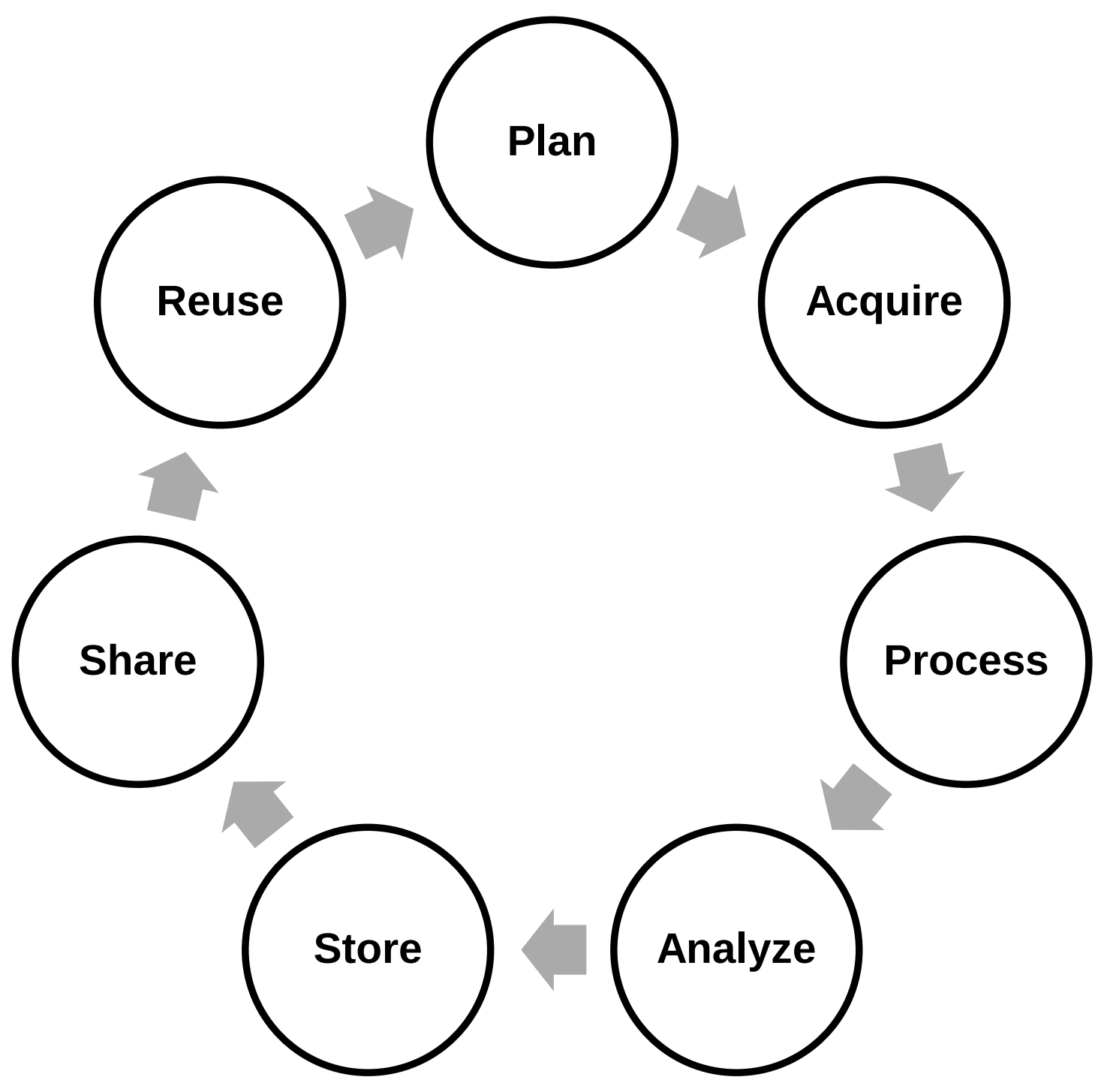
The CDCS enables **Interoperability** of data and metadata via XML technologies such as XML Schemas. In this way, it can support existing XML-based data standards, such as MatML, ThermoML, and MTConnect, which have a stable following in scientific and engineering communities. The CDCS also support data reuse through creation and exchange of modular, community-based schemas, supporting diverse users who leverage shared concepts. The MGI is developing a registry and repository for supporting this kind of exchange which will enable the **Findability**, **Accessibility**, and **Reuse** of these resources.



Vision

The Configurable Data Curation System (CDCS) aims to support and enable:

- An Effective Research Data Lifecycle
- FAIR Data Principles[1]
  (Findable, Accessible, Interoperable, and Reusable)
- Modular System Design



Visit:
https://cdcs.nist.gov/

Findable

- (meta)data are assigned a globally unique and persistent identifier
- data are described with rich metadata
- metadata clearly and explicitly include the identifier of the data it describes
- (meta)data are registered or indexed in a searchable resource

Accessible

- (meta)data are retrievable by their identifier using a standardized communications protocol
  - the protocol is open, free, and universally implementable
  - the protocol allows for an authentication and authorization procedure, where necessary
- metadata are accessible, even when the data are no longer available

Interoperable

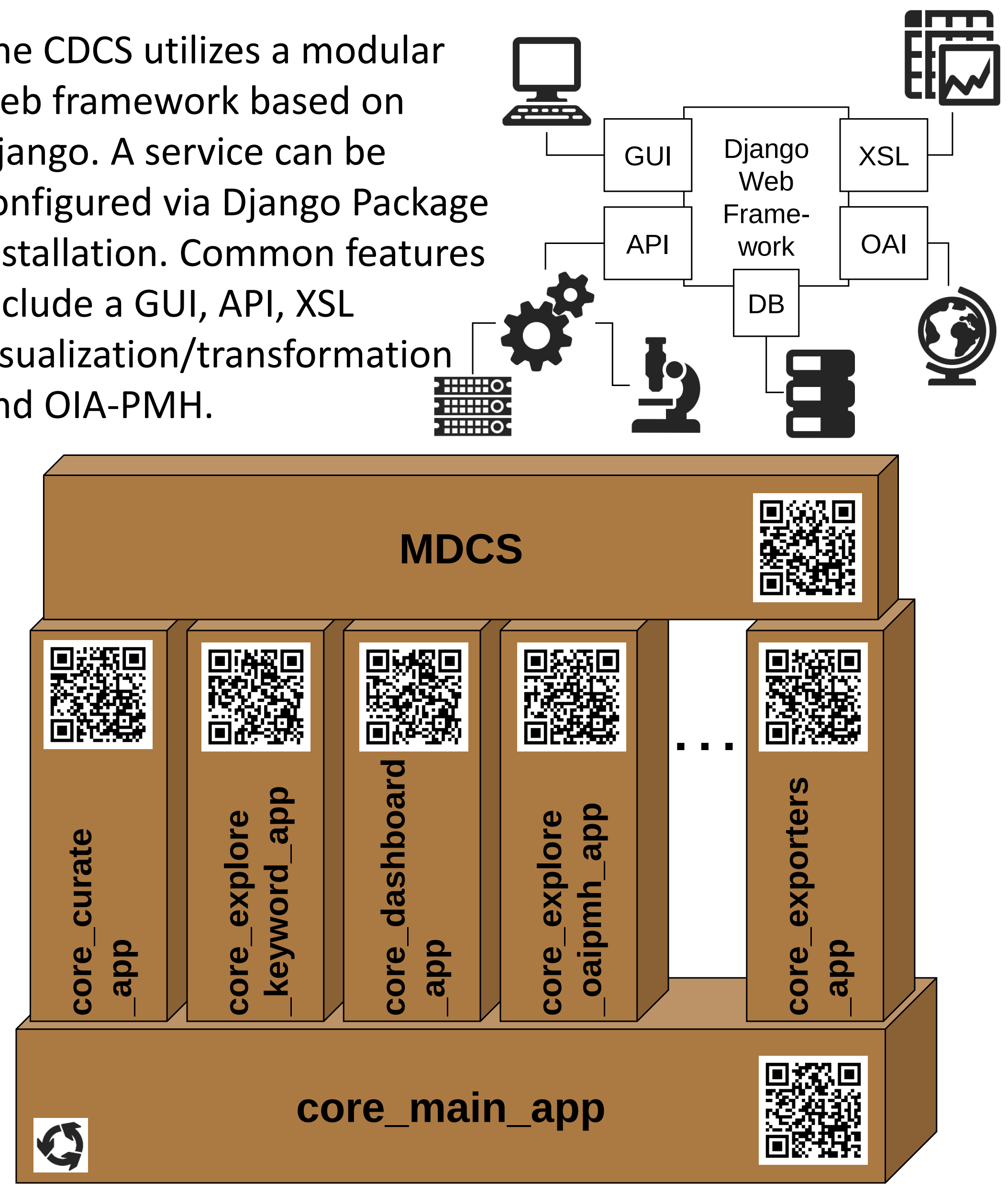
- (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- (meta)data use vocabularies that follow FAIR principles
- (meta)data include qualified references to other (meta)data

Reusable

- meta(data) are richly described with a plurality of accurate and relevant attributes
  - (meta)data are released with a clear and accessible data usage license
  - (meta)data are associated with detailed provenance
  - (meta)data meet domain-relevant community standards

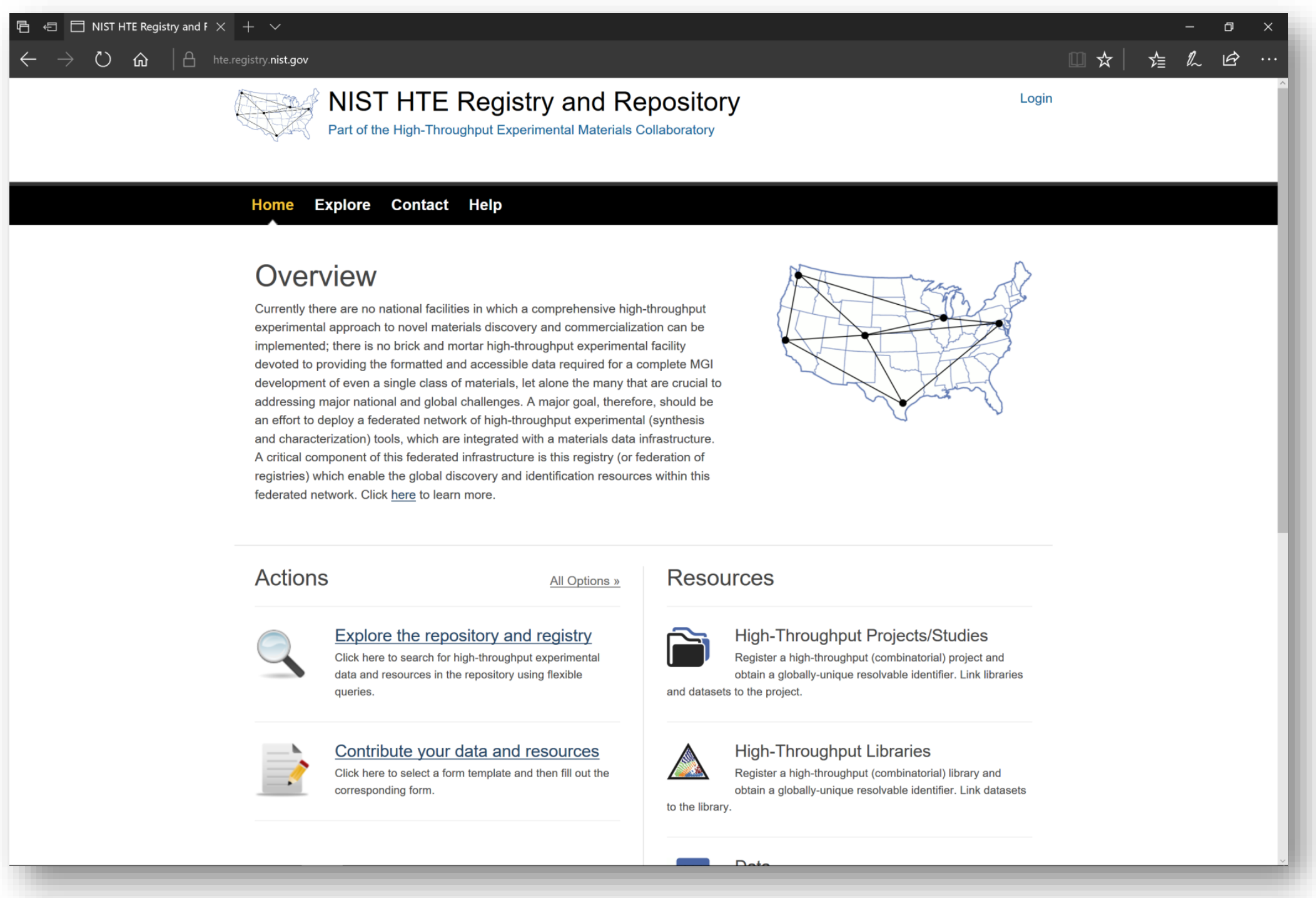
Modular System Design

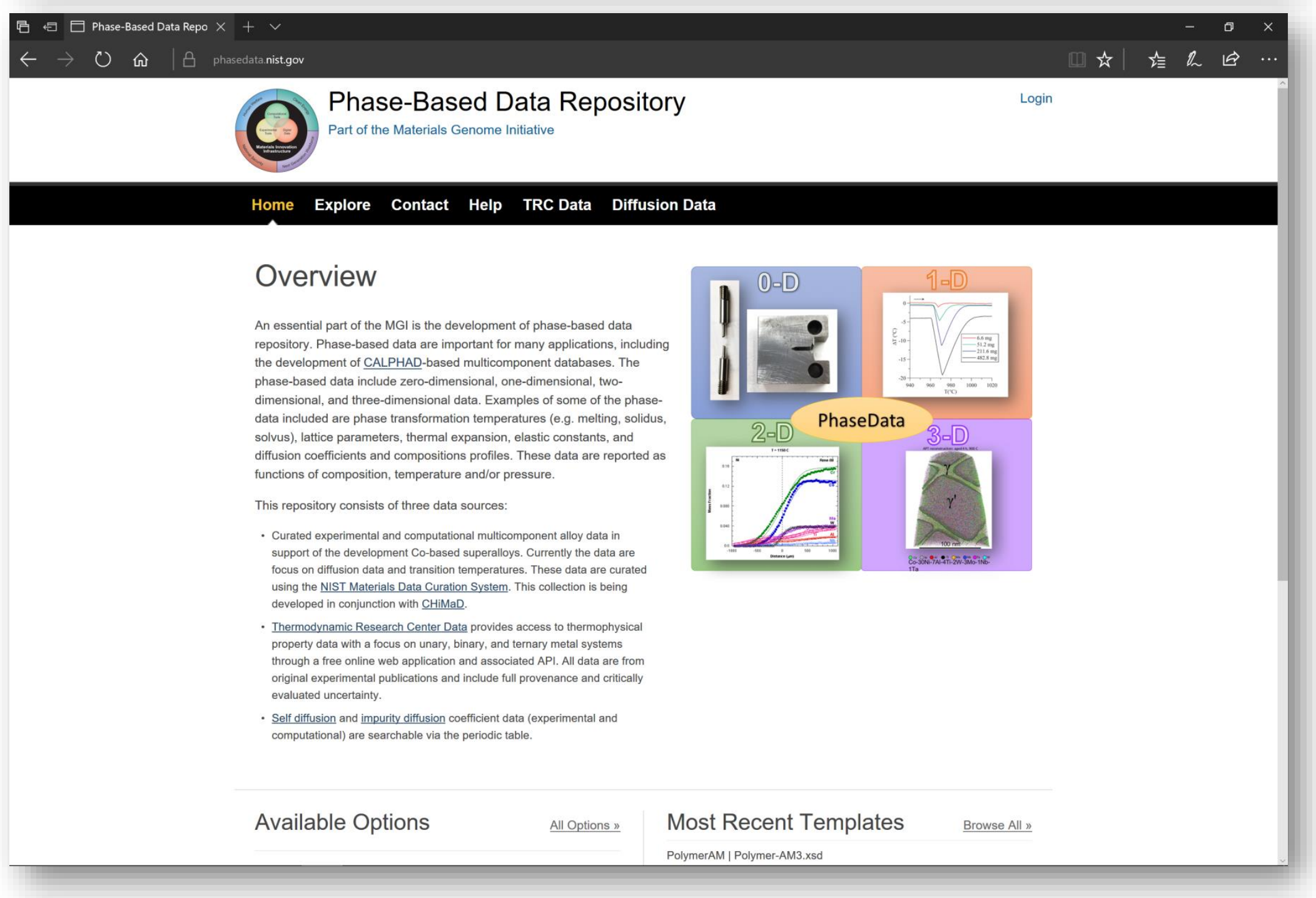
The CDCS utilizes a modular web framework based on Django. A service can be configured via Django Package installation. Common features include a GUI, API, XSL visualization/transformation and OIA-PMH.

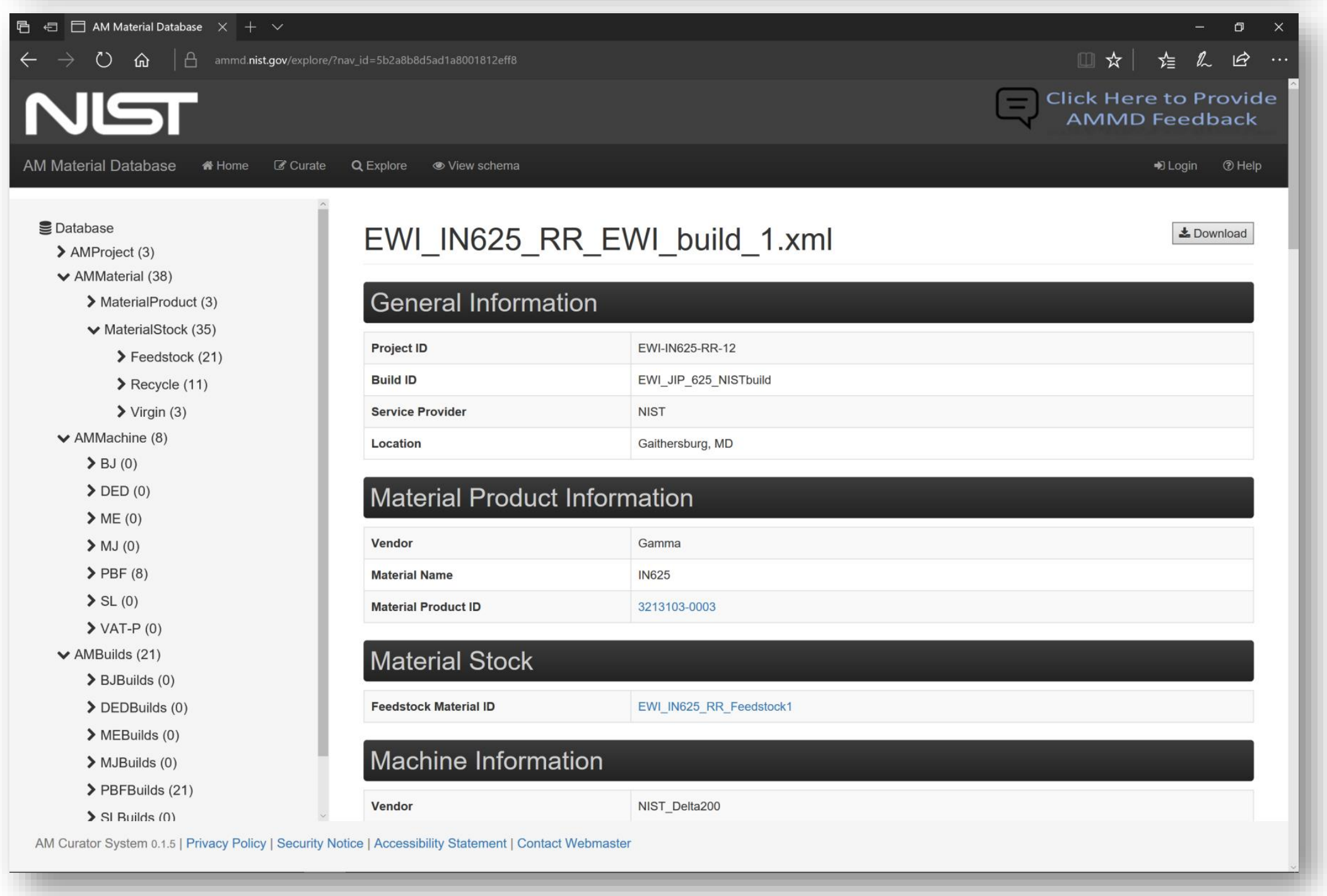


Repositories

The CDCS empowers communities to deploy and operate specialized repositories, which enables **Accessibility** of data and metadata. Data and metadata are stored in **Interoperable** formats and can be retrieved by API. Examples of three materials repositories are shown below.


NIST HTE Registry and Repository
Part of the High-Throughput Experimental Materials Collaboratory
Overview
Actions
Resources


Phase-Based Data Repository
Part of the Materials Genome Initiative
Overview
PhaseData
Available Options
Most Recent Templates



Engagement

The CDCS represents a platform through which NIST and related communities have begun to mutually engage in discussion, development, and problem-solving. Driven by FAIR data principles, this has given rise to a number of activities through NIST, including:

- **Community standards:** To increase the availability and quality of community standards, NIST hosted workshops focused on community data model development.
- **Interoperability:** To increase integration among data platforms (for materials science and beyond), NIST hosted a hackathon focused on such integrations.
- **Community development:** To support the development of FAIR data communities (in materials science and elsewhere), NIST began an annual convention for growing and nurturing the CDCS community of users.

	(Optional) pre-convention Tutorials	Annual Convention	(Optional) post-convention Code Sprints
Duration	1-2 Days	2-3 Days	1-2 Days
Organizers	Stakeholders	NIST Development Team	Stakeholders
Format	Parallel Track	Single Track	Parallel Track
Objective	Education	Stakeholder Networking & Annual Roadmapping Exercise	Software Development Planning & Implementation

References

[1] Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

[2] Dima, A., Bhaskarla, S., Becker, C., Brady, M., Campbell, C., Dessauw, P., et al. (2016). Informatics Infrastructure for the Materials Genome Initiative. JOM, 68(8), 2053–2064. https://doi.org/10.1007/s11837-016-2000-4