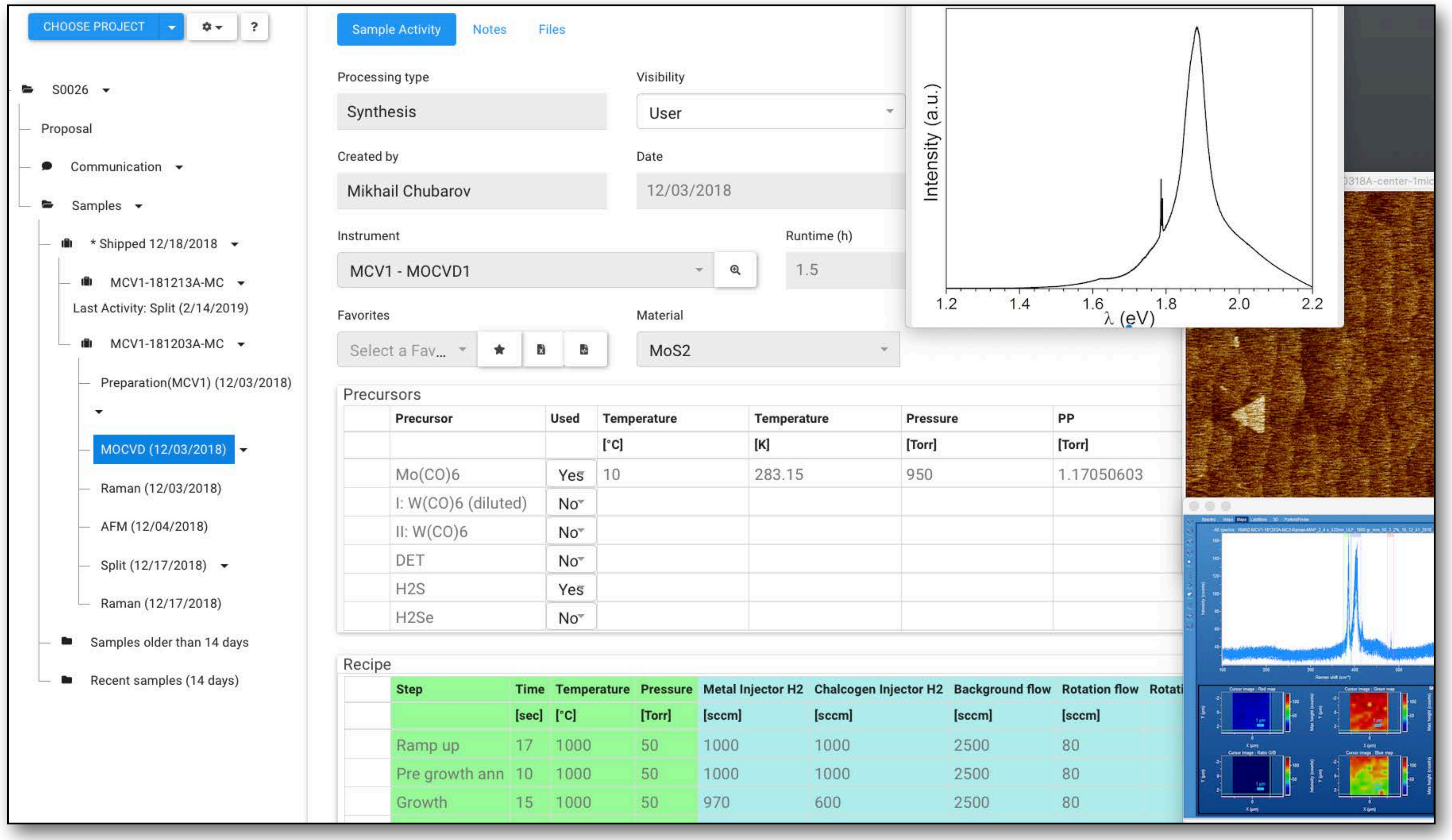
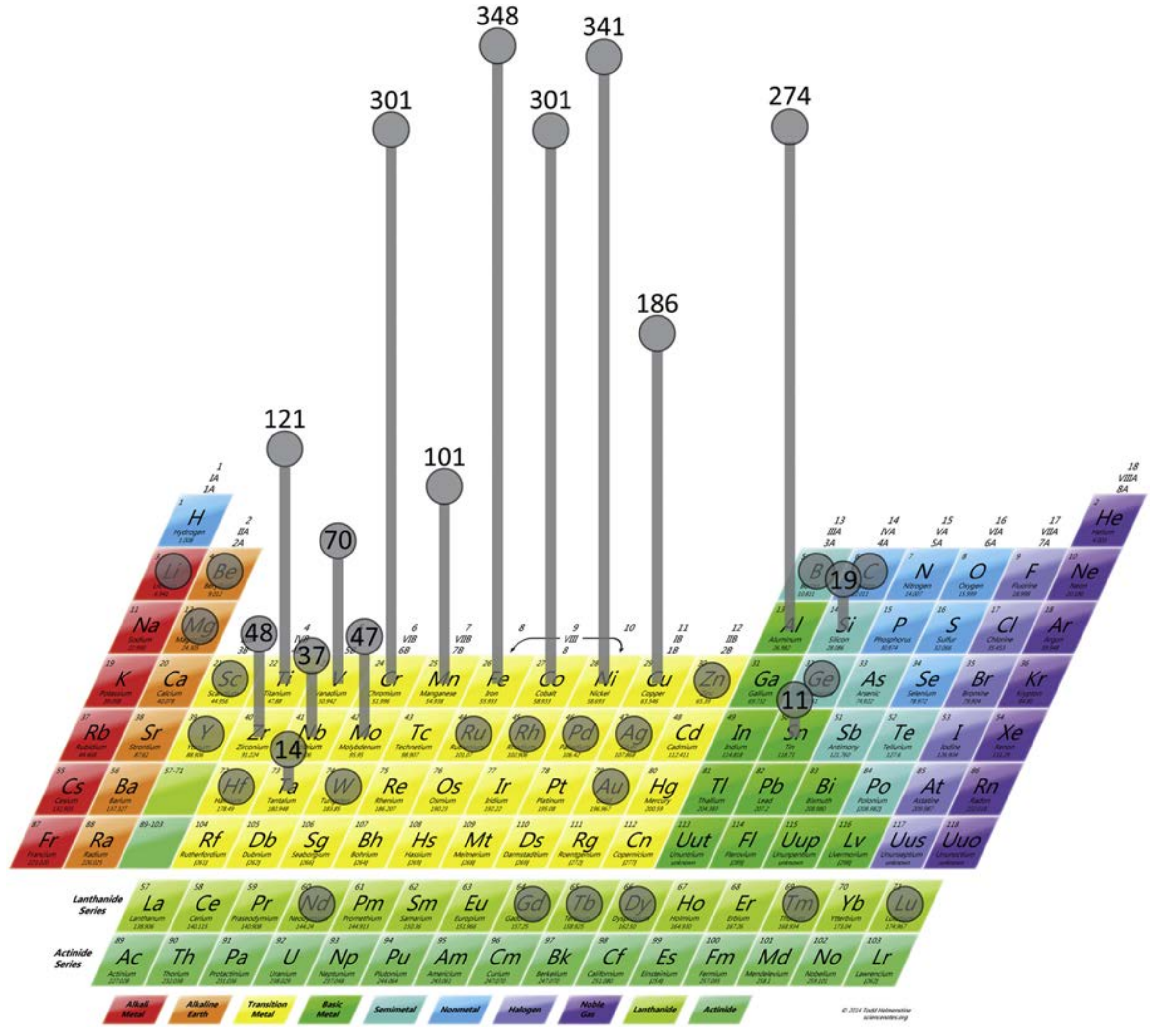
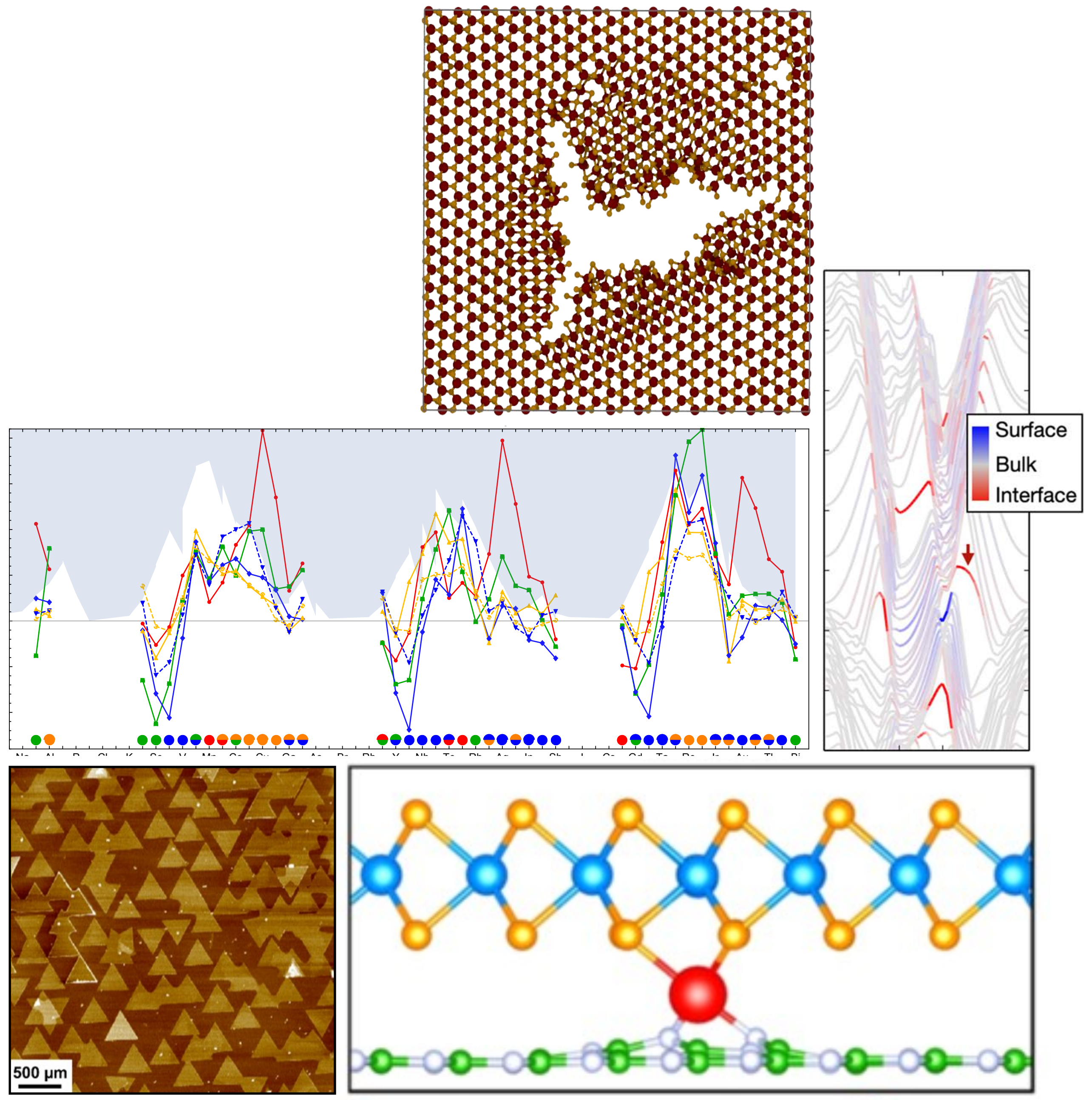
Data Practice and Plans in the Two-Dimensional Crystal Consortium





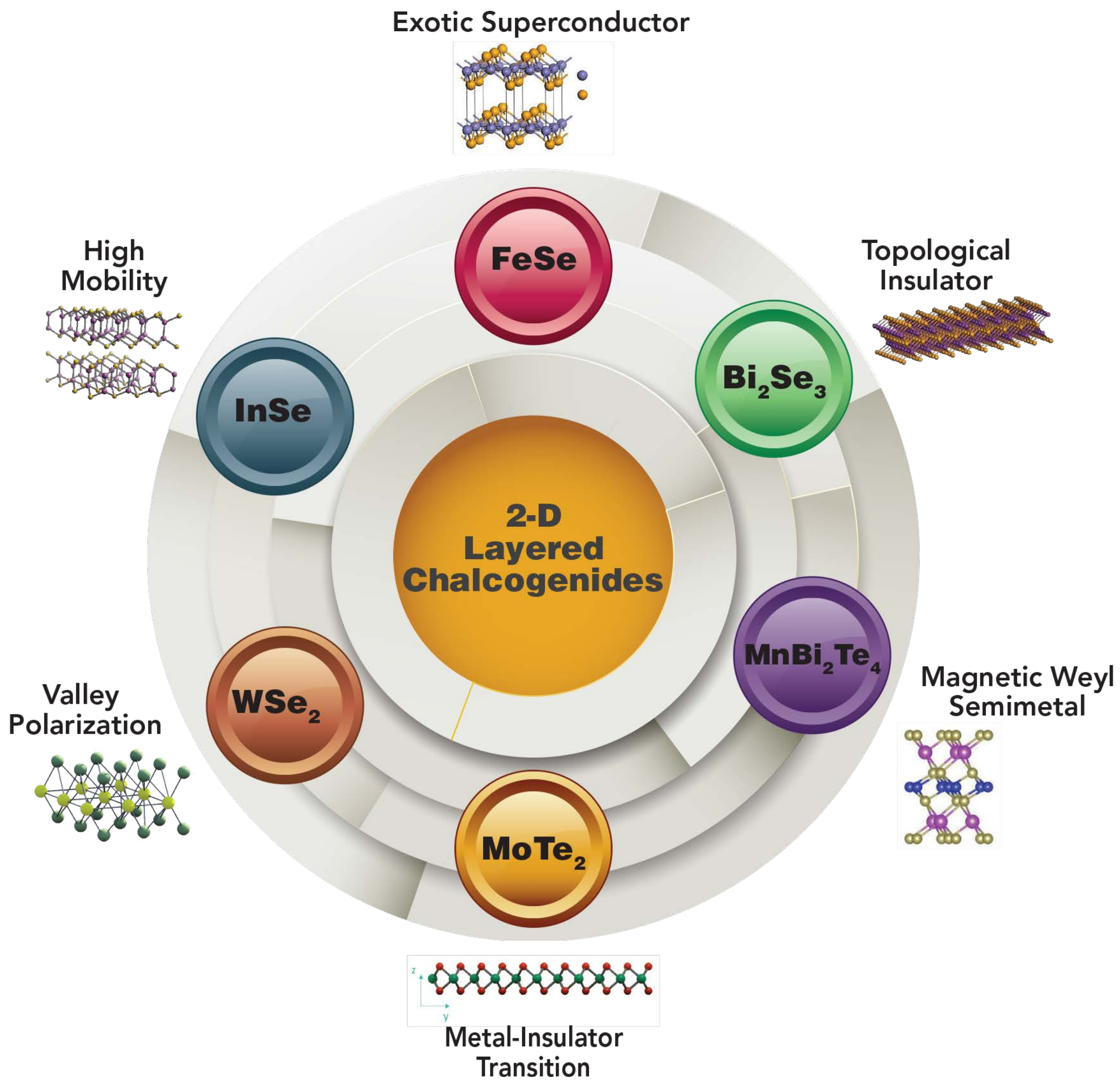


Capture and curation of **sample history as metadata** through **Li**fetime **S**ample **T**racking.

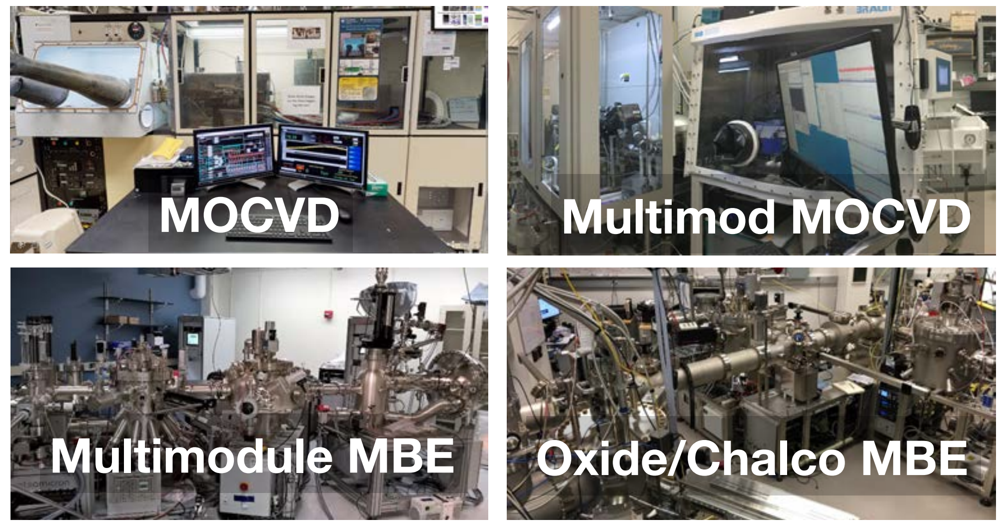
A developing program of discoverability, access, and interoperability/reuse

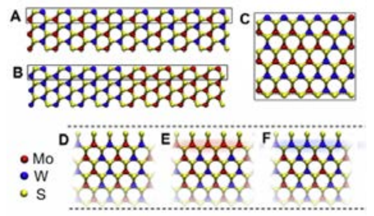
A National User Facility in 2D Materials Growth

2D Crystal Consortium
NSF Materials Innovation Platform

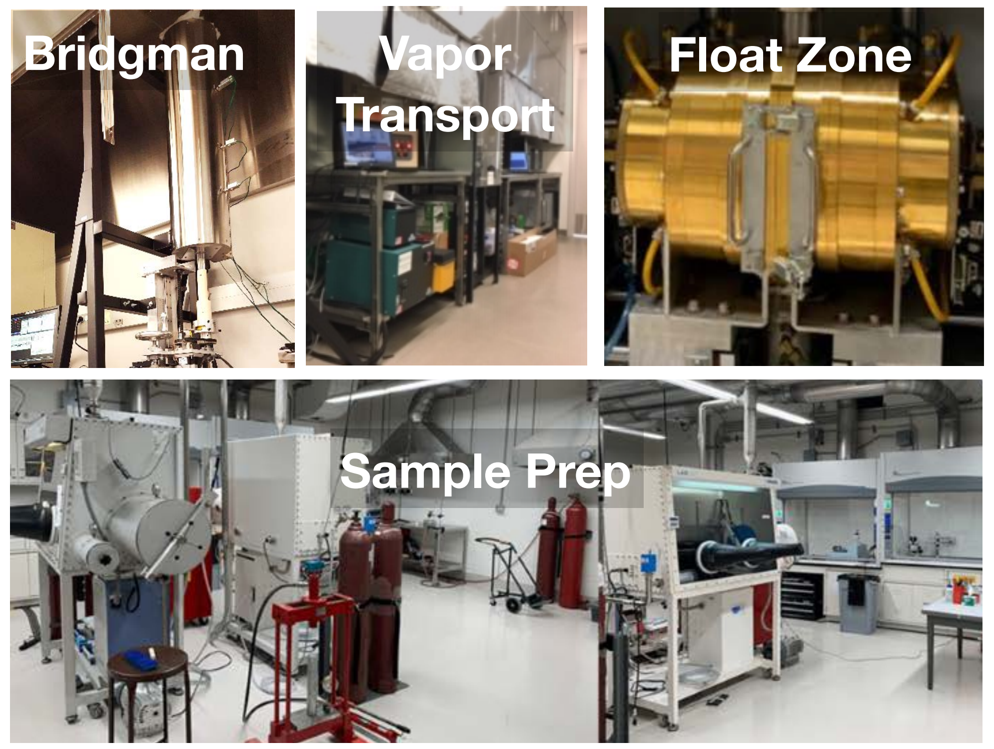


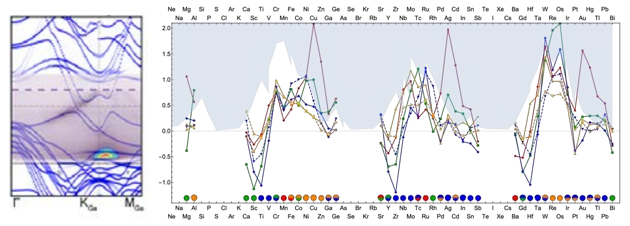
Thin Films and In Situ Characterization



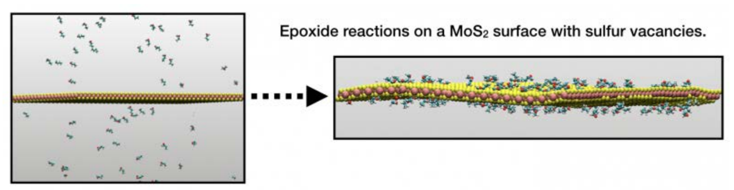


Bulk Crystal Growth

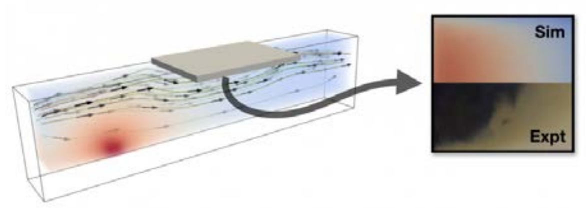




First Principles



Reactive Force Fields



Multiscale Modeling

NEW

Double Crucible Vertical Bridgman Furnace for Stoichiometry Control

High Pressure Thermal Evaporator for Confinement Heteroepitaxy of 2D Metals

Glovebox Cluster Tool for 2D/2D Hybrid Structures and Devices

2DCC Vision

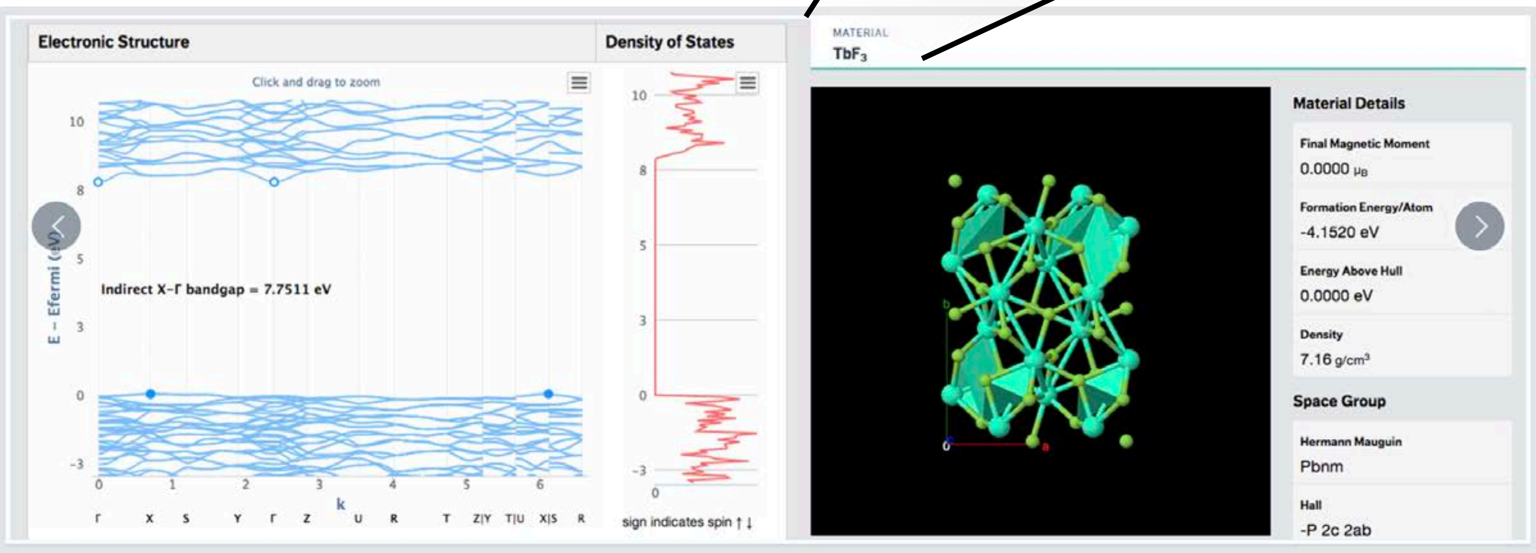
Advance discovery-driven research into the growth, properties and applications of 2D layered chalcogenide crystals and related 2D materials for next-generation technologies by innovating state-of-the-art synthesis, characterization and data/computational tools to foster a diverse scientific ecosystem of in-house experts and external users to drive international leadership by the US 2D materials research community.

Character of a 2D Materials Data Ecosystem

Astronomers (and biologists) pioneered community data platforms.

We share a single sky…

…there is only one “TbF_3” crystal in Materials Project…



TbF_3

while crystal structure databases key on “archetypal” and near-universal properties of materials.

Numerous **sample-unique properties** resulting from complex synthesis kinetics mean that **every sample of a 2D material is its own “sky”**.

Some use cases for experimental materials data

It depends very much on the type of data!
For example, two diametric poles:

X-ray crystal structure data:

- well-defined mathematically with a shared understanding
- relatively insensitive to sample imperfections, the effects of which are fairly well-understood
- reducing imperfections consistently makes “better” data
- thermodynamic equilibrium provides a unique and precisely defined ideal material state

Charge transport measurement of 2D heterostructures:

- complex system-specific interpretation without a shared understanding
- very sensitive to imperfections in the sample in ways not necessarily well-understood
- sometime reducing imperfections improves data; sometimes certain imperfections *are* the data
- heterostructure formation is often far from equilibrium and **every sample is unique**, with properties that depend very sensitively on **how it was made**.

https://americaexplained.wordpress.com/2010/12/17/

Every sample is an embodied history

Material synthesis is a usually combination of **intuition, experience, equipment** and **hard work.** Contributions from computational design are challenged by the need to handle extreme spans of time/length-scale with limited opportunity for experimental validation of kinetic intermediates.

Kinetics >> Thermodynamics: A comprehensive database of $[\text{kinetic intermediates} \times \text{substrate interactions} \times \text{defect geometries} \times (\text{key impurities}+1)]$ for *a single material system* could be comparable in size to an equilibrium crystal structure database across *the entire periodic table.*

A data infrastructure for diverse forms of materials **measurement** would ideally build on a data infrastructure for materials **growth**. **Sample growth history is *metadata* for sample measurement.**

Naive Theorist: "Every sample is identical”

Savvy Experimentalist: “Every sample is unique”

X-ray crystal structure data:

- extremely well-defined mathematically and universally understood
- relatively insensitive to sample imperfections, the effects of which are fairly well-understood
- reducing imperfections consistently makes “better” data
- thermodynamic equilibrium as a unique and precisely defined material state

Charge transport measurement of a 2D heterostructure:

- complex system-specific interpretation
- very sensitive to imperfections in the sample in ways not necessarily well-understood
- sometime reducing imperfections improves data; sometimes imperfections *are* the data
- heterostructure formation is often far from equilibrium and **every sample is unique**.

A materials data infrastructure that covers ground around this second pole should address the challenge of data acquisition, annotation, and curation for the entire history of the sample, ideally **from the precursor atoms or molecules to a fully grown, processed and measured sample**.

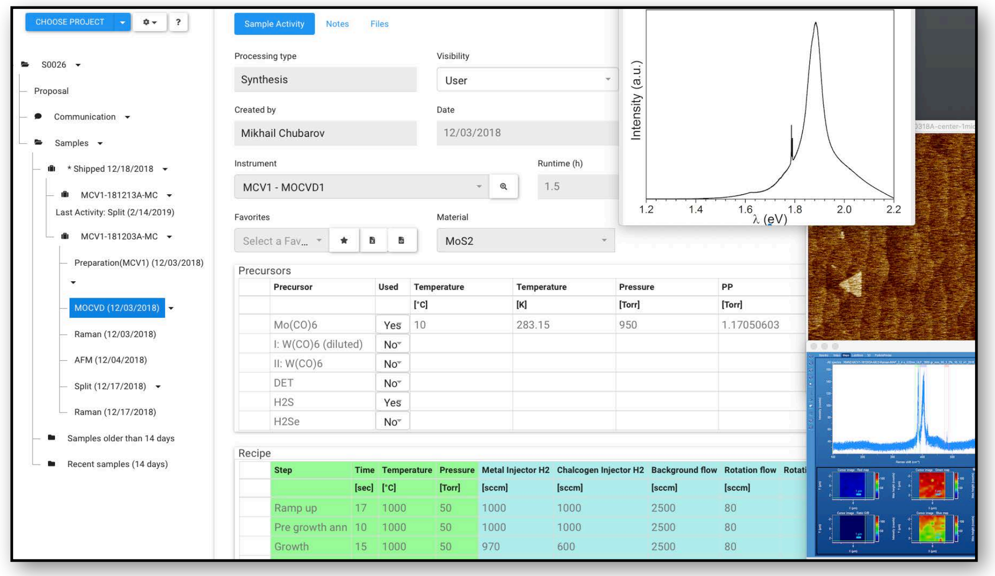
This history-defining metadata may well exceed the data in size and scope.

Lifetime Sample Tracking 1.0 captures and curates sample history

Tool development to date:

Konrad Hilse · Kevin Dressler

LiST



Year 1: 2DCC-MIP Platform focus on hardware ramp-up and tool development.
The need for and concept of a LiST tool begins to form.
NSF introduces the 10 Big Ideas, including Harnessing Data.

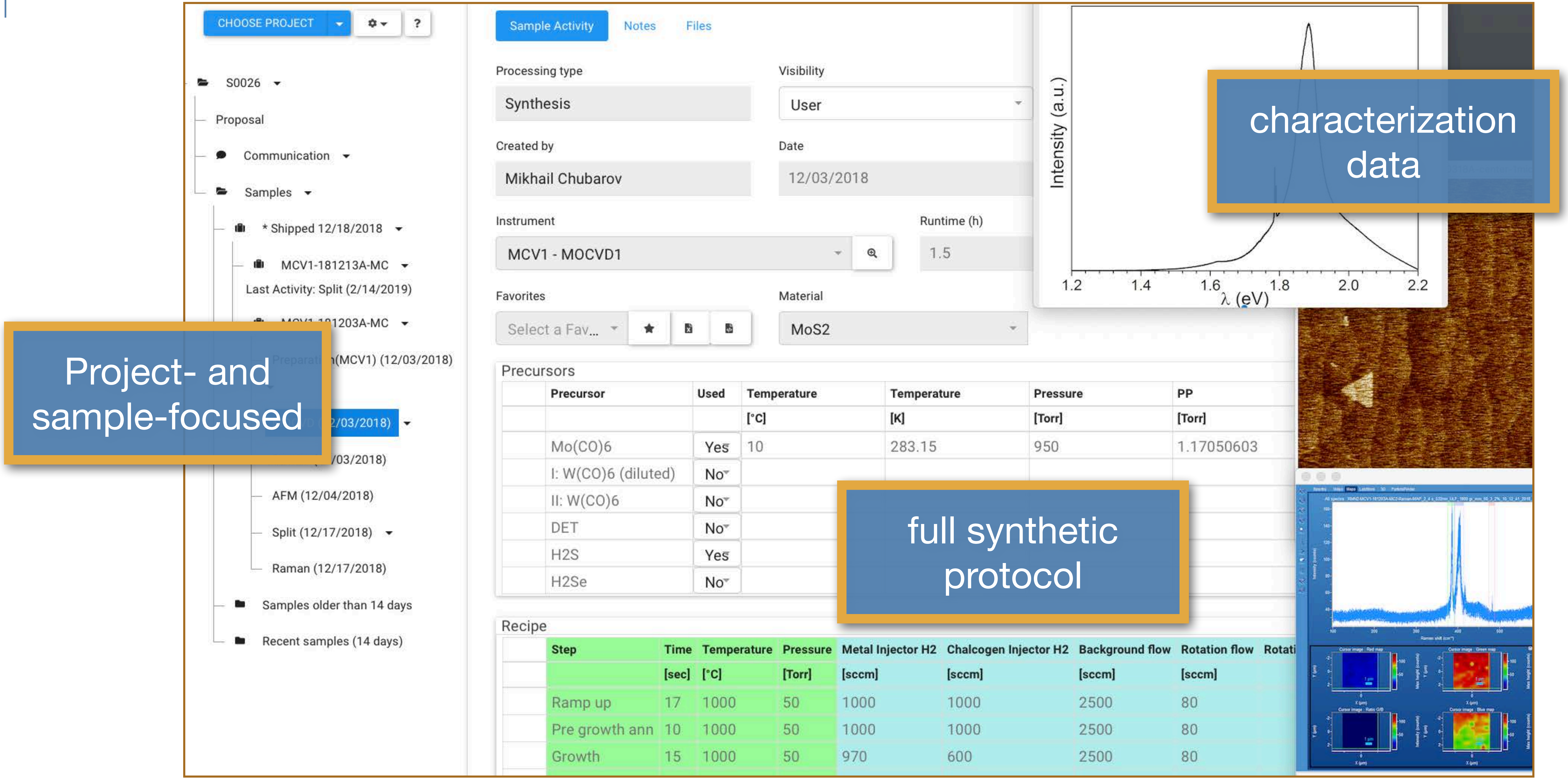
Year 2: Initial conception of a comprehensive, facility-wide sample-tracking tool.
Budget reconfiguration to enable recruitment of an IT staff member.
Initial design of database schema, access modes, and user interface.

Year 3: Implementation of “LiST 0.5” with drag-and-drop data import.
Iterative feedback and optimization of LiST workflows, working with technical staff across 2DCC.

Years 4,5: Automated data import across all synthesis tools, compatible with diverse staff workflows.
Import of current and historical synthesis data for >5000 2DCC samples.
Import of characterization data on these samples.

Year 6: MetaData Explorer provides a public view of LiST: Findable
Sample sets attached to User Data Projects makes LiST data accessible for data-oriented 2DCC users.

LiST 1.0

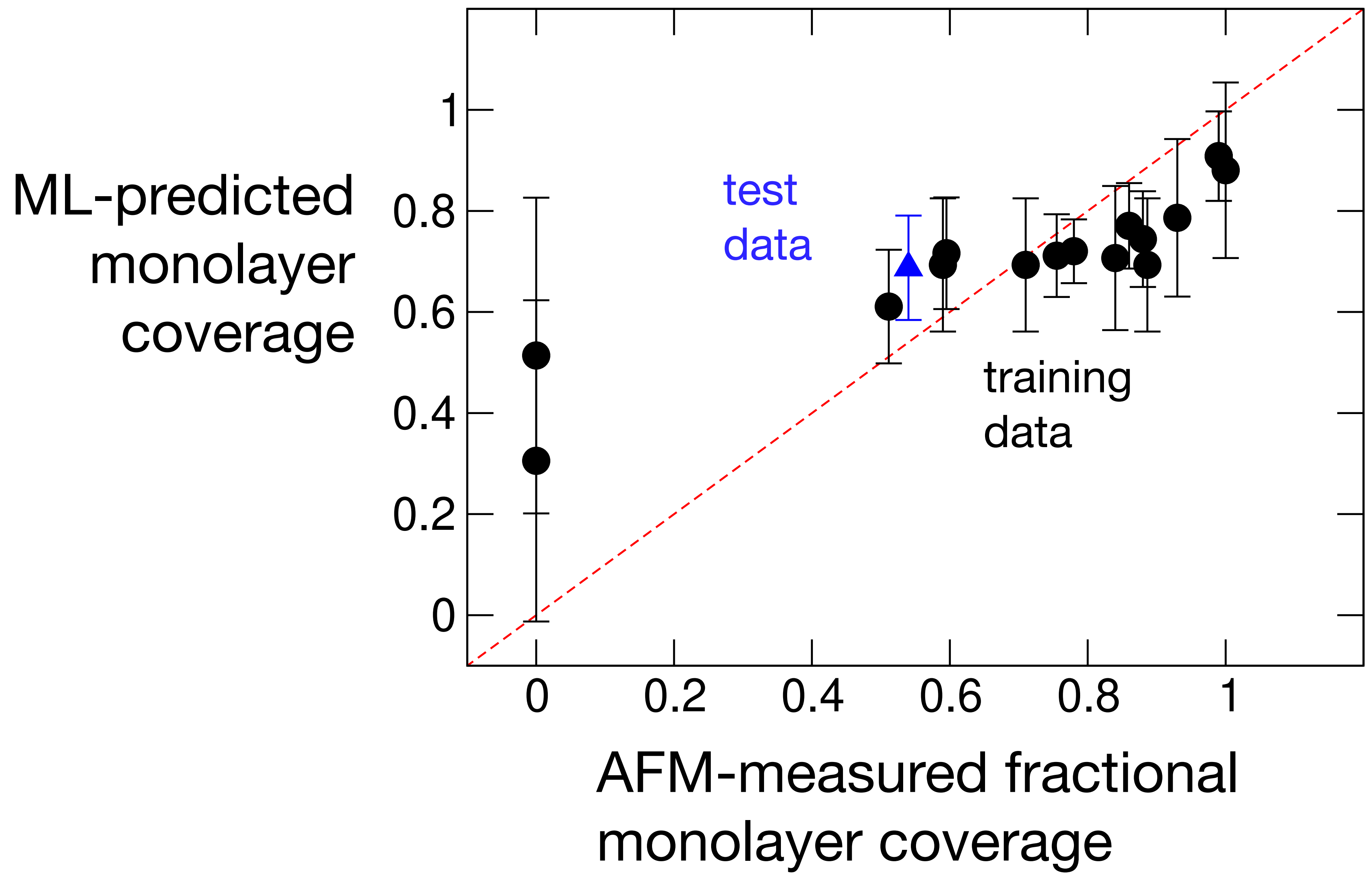


LiST 1.0 is a platform for materials synthesis data capture and curation, compatible with staff workflows and extendable to new data from diverse sources. It is populated with all synthesis data since platform inception, with ongoing import of characterization data.

How LiST is used currently:

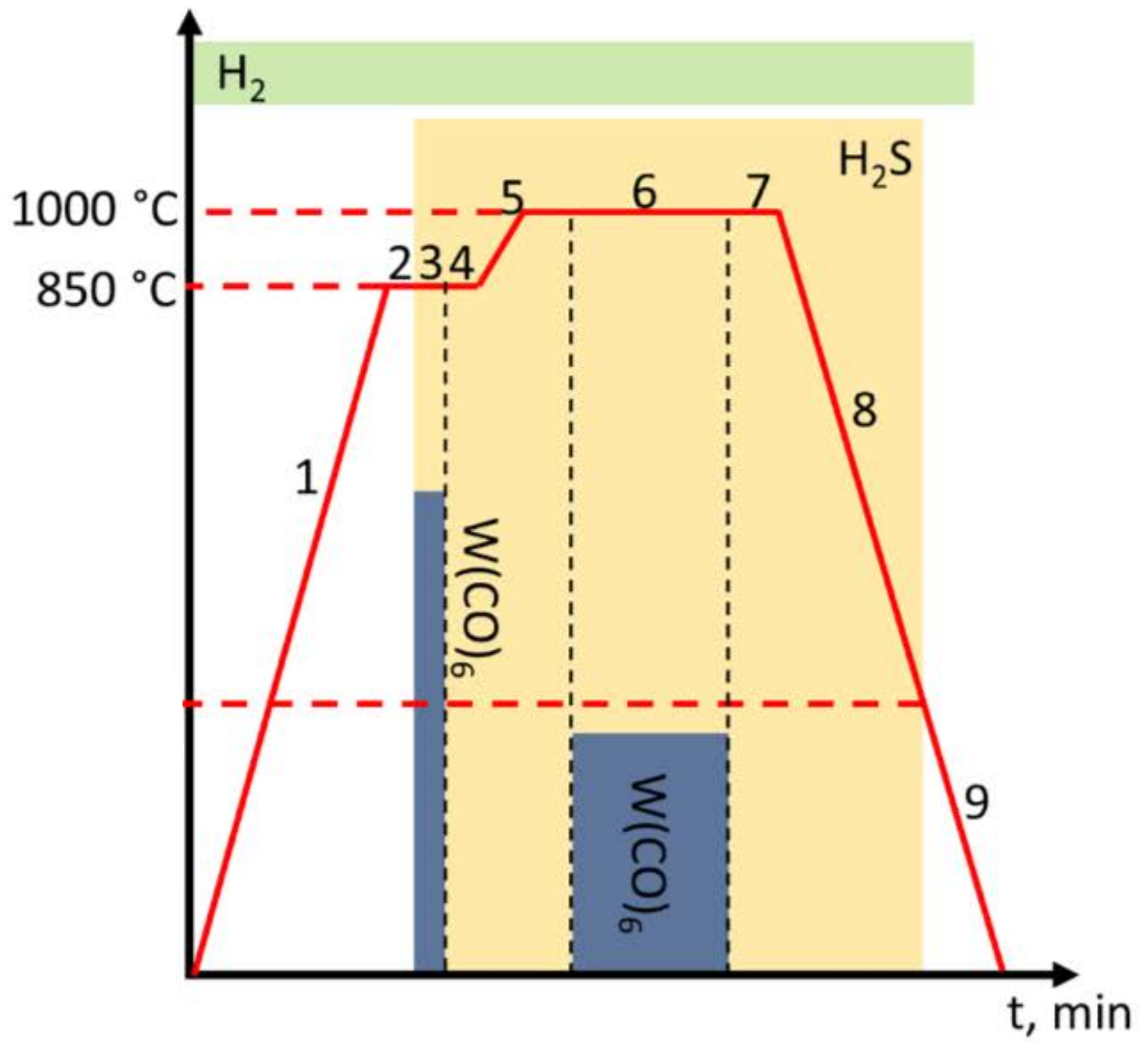
- **In-house staff and visitors** import all new growth runs, using automated data import tools.
- **Users** access data transitioning to direct access via LiST.
- User Data Requests for platform-wide data link together samples from many existing projects into a distinct **User Data Project**, accessed via LiST.

MOCVD-grown WS_2

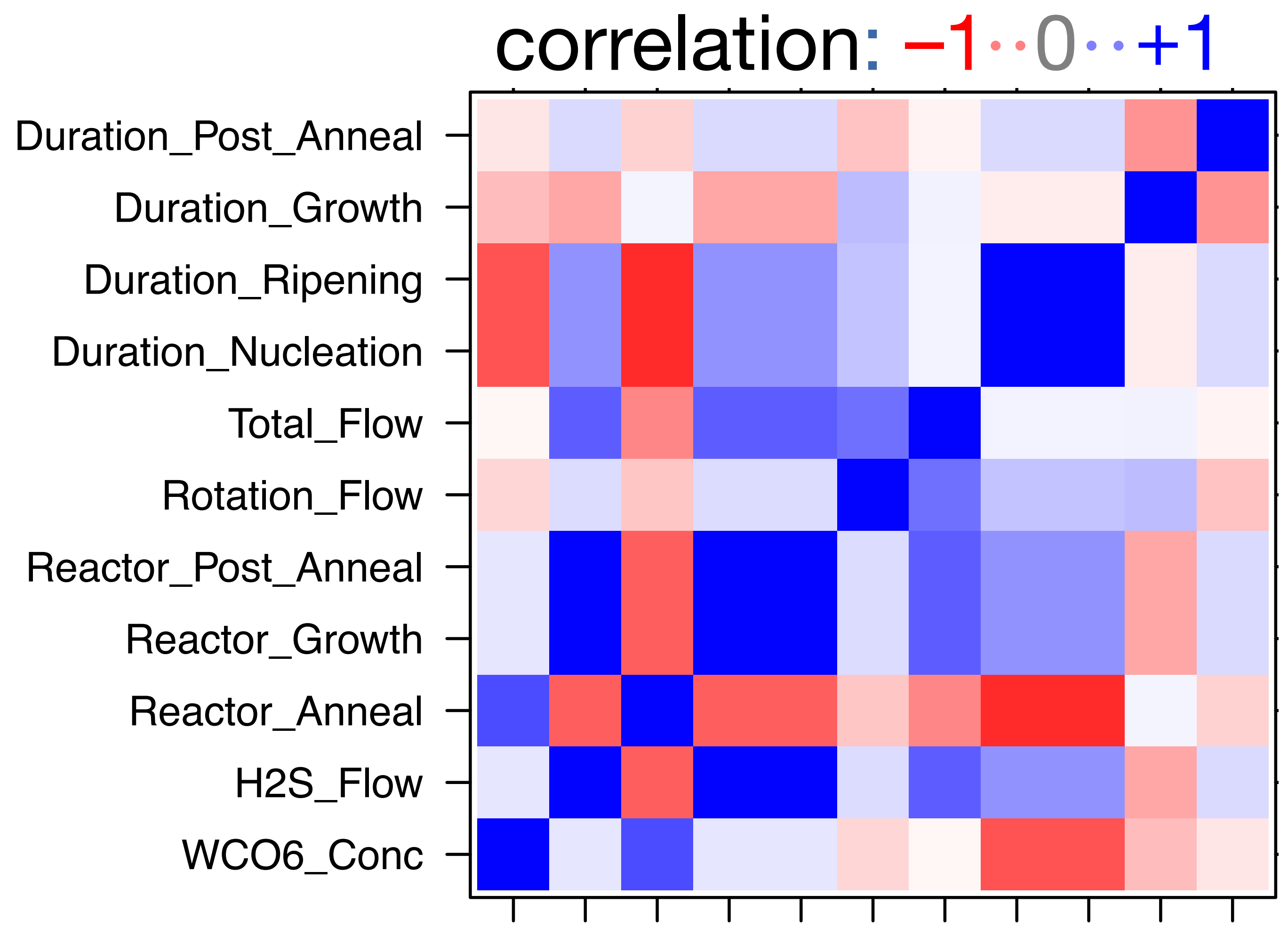


Next Steps:

- API-based access.
- Better alignment of data formats for characterization data with community standards
- Integration with associated simulation methods and results.
- 3rd-party data?
- **Findable: MetaData Explorer**



17 WS_2 growth runs **on a single tool**



<u>User Data Request S0073:</u>
P. Balachandran & P. Reinke @ UVA
T. Choudhury & J. Redwing @ 2DCC

MetaData Explorer

- Publicly accessible interface to filter LiST sample listings by material, characterization technique, synthesis tool, etc.
- Can guide User Data Requests by revealing the amount and character of data that is available for analysis.
- The elemental filters will also act as filters for the availability of reactive force fields covering different sets of compositions.
- **F**indability of LiST Data.
- **A**ccessibility through User Data Projects with associated sample sets, logging in via Shibboleth. (Future: API access)
- **I**nteroperability and **R**eusability are facilitated by the metadata that LiST surfaces through its user interface. (Planned: standardize machine-readable metadata formats for synthesis, characterization, simulation).

MetaData Explorer

Publications | Availabe Samples | Show Projects | Search materials web

2DCC-MIP

Growth Method: Any

Characterization Technique: Any

Materials: ✖ SHOW PERIODIC TABLE | SEARCH

 # Samples with Characterization Data

# Samples	Material	Growth Method	Contact	AFM	ARPES	Auger	Ellipsometry	ICP-AES	Optical Profilometry	Raman	SEM	SQUID	ST-FMR	STM	TEM	Transport	XPS	XRD
153	$ZrTe_2$	MBE	Nitin Samarth	48	91	0	0	0	0	1	0	0	0	7	0	18	1	23
101	$CrTe_2$	MBE	Nitin Samarth	8	79	0	0	0	0	0	0	5	0	2	0	2	0	4
92	FeTe	MBE	Cui-zu Chang	0	49	0	0	0	0	0	0	0	0	0	0	0	0	0
225	$(Bi,Sb)_2Te_3$	MBE	Nitin Samarth	90	26	0	0	0	0	1	0	1	0	0	0	55	5	31
60	$Cr\text{-}ZrTe_2$	MBE	Nitin Samarth	2	24	0	0	0	0	0	0	1	0	2	0	22	0	16
30	Se, Te, $NbSe_2$	MBE	Cui-zu Chang	0	24	0	0	0	0	0	0	0	0	0	0	8	0	0
37	$NbSe_2$	MBE	Cui-zu Chang	0	19	0	0	0	0	0	0	0	0	0	0	10	0	0
137	Bi_2Se_3	MBE	Nitin Samarth	52	16	0	0	0	0	0	0	0	0	0	0	6	0	31
16	$CrTe_2$, $ZrTe_2$	MBE	Nitin Samarth	0	12	0	0	0	0	0	0	0	0	2	0	0	0	0
11	Se, $NbSe_2$	MBE	Cui-zu Chang	1	10	0	0	0	0	0	0	0	0	0	0	0	0	0

Showing 1 to 10 of 457 entries

Previous 1 2 3 4 5 ... 46 Next

How are we doing on FAIR? (mostly *in re* LiST)

Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

- F1. (Meta)data are assigned a globally unique and persistent identifier (locally unique with a scheme to globalize)
- F2. Data are described with rich metadata, defined by R1 below
- F3. Metadata clearly and explicitly include the identifier of the data they describe (via UI)
- F4. (Meta)data are registered or indexed in a searchable resource (indexed internally, not externally)

Accessible

Once the user finds the required data, she/he/they need to know how can they be accessed, possibly including authentication and authorization.

- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol (via UI, not API yet)
- A1.1 The protocol is open, free, and universally implementable (via Shibboleth, requires user data project)
- A1.2 The protocol allows for an authentication and authorisation procedure, where necessary (via Shibboleth)
- A2. Metadata are accessible, even when the data are no longer available (via MetaData Explorer)

Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation (List 2.0)
- I2. (Meta)data use vocabularies that follow FAIR principles (sporadic, a target for List 2.0)
- I3. (Meta)data include qualified references to other (meta)data (not all qualified, a target for List 2.0)

Reusable

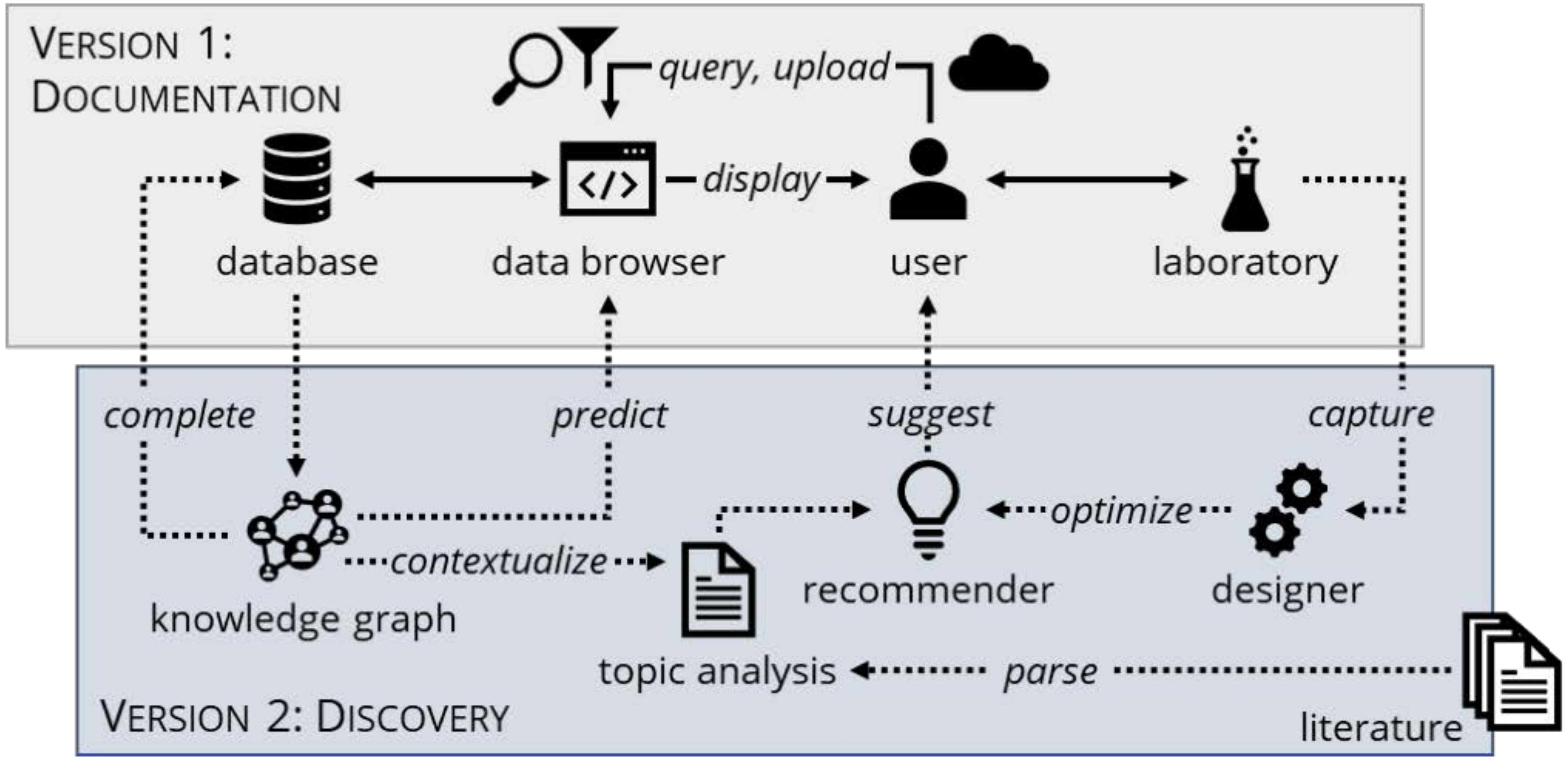
The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

- R1. (Meta)data are richly described with a plurality of accurate and relevant attributes (this is the purpose of LiST and DMR-IDB)
- R1.1. (Meta)data are released with a clear and accessible data usage license (requires sustained stakeholder engagement)
- R1.2. (Meta)data are associated with detailed provenance (could be more detailed)
- R1.3. (Meta)data meet domain-relevant community standards (needs work)

Orange: work-in-progress Green: implemented Parentheses: (caveat or context)

https://www.go-fair.org/fair-principles/

LiST 2.0: From Data Store to Knowledge Graph



1.0: Documentation

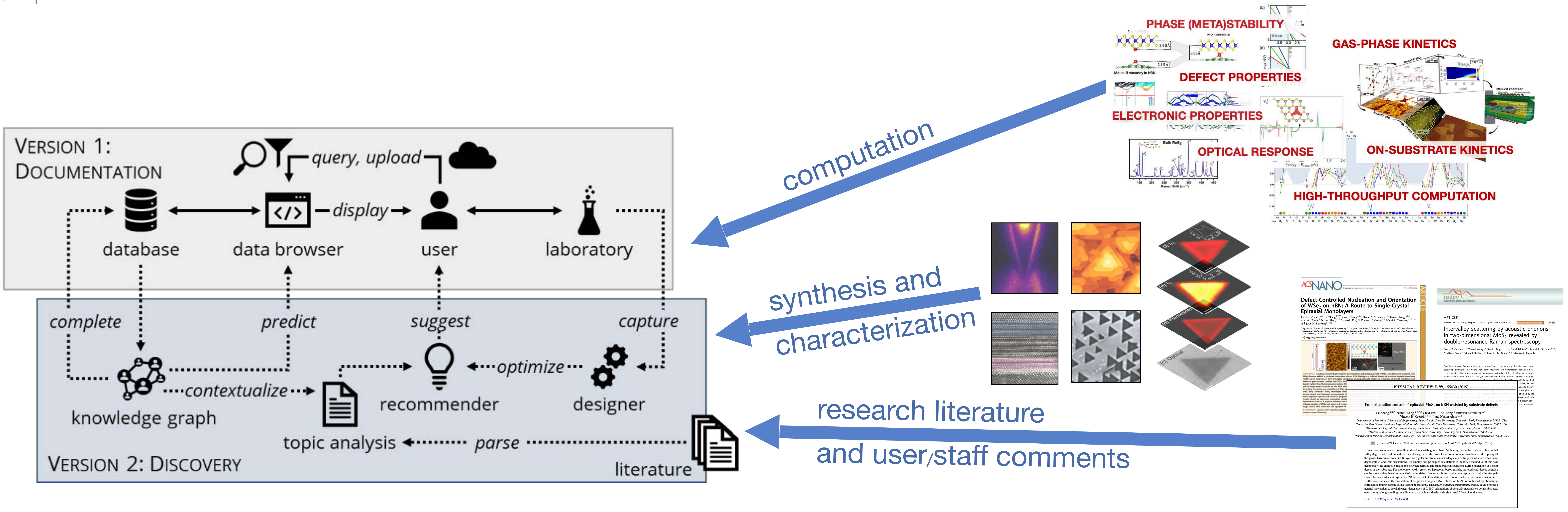
⬇

2.0: Discovery

LiST 1.0 is a **traditional database** whose consistent schema are straightforward to query, synthesize, and maintain, but it can only map explicit relationships between data.

List 2.0's **knowledge graph** will contain both a data store and an ontology describing the relationships between entities and will be able to infer relationships not explicitly provided.

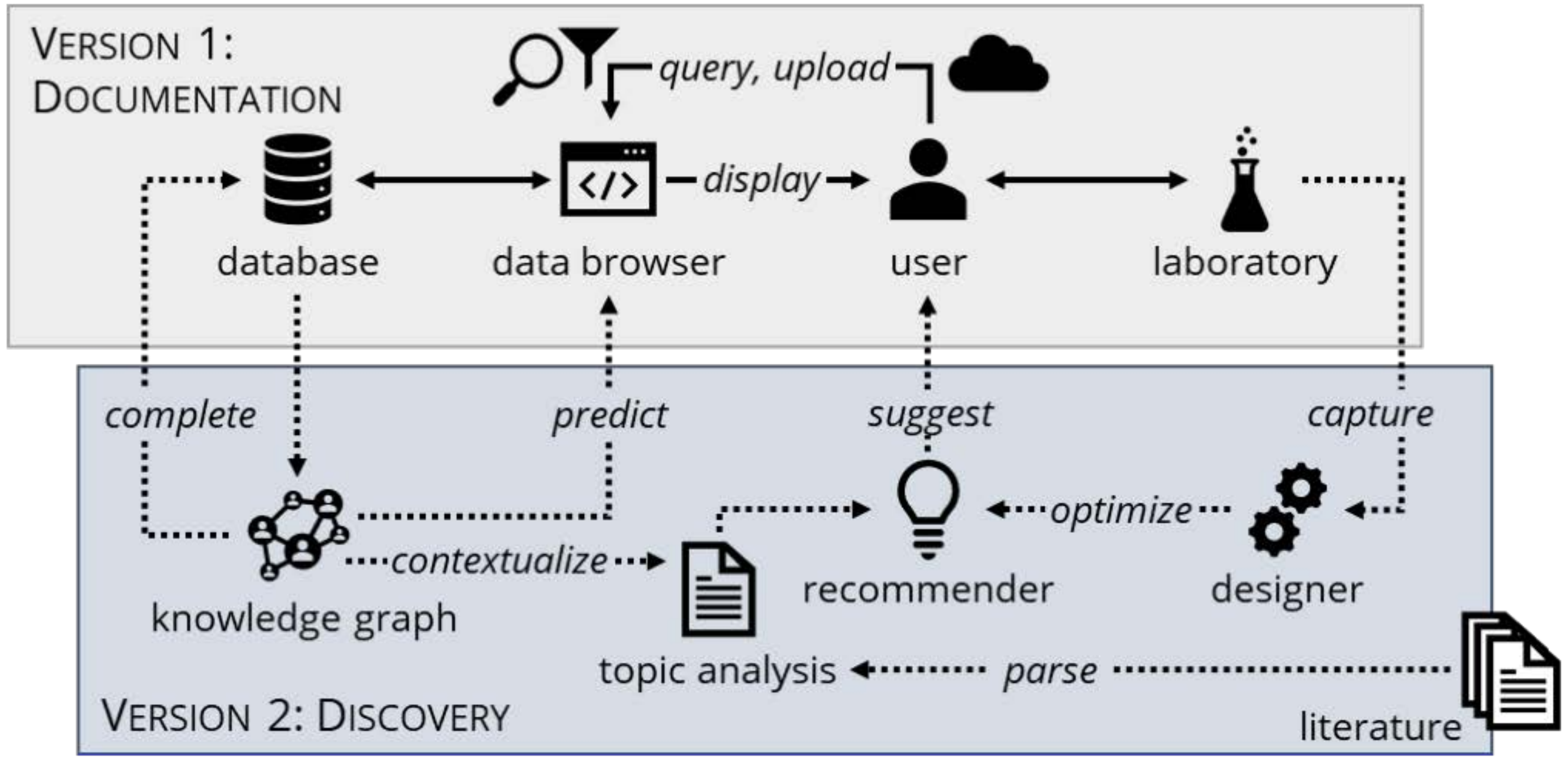
Integrating diverse data into the Knowledge Graph

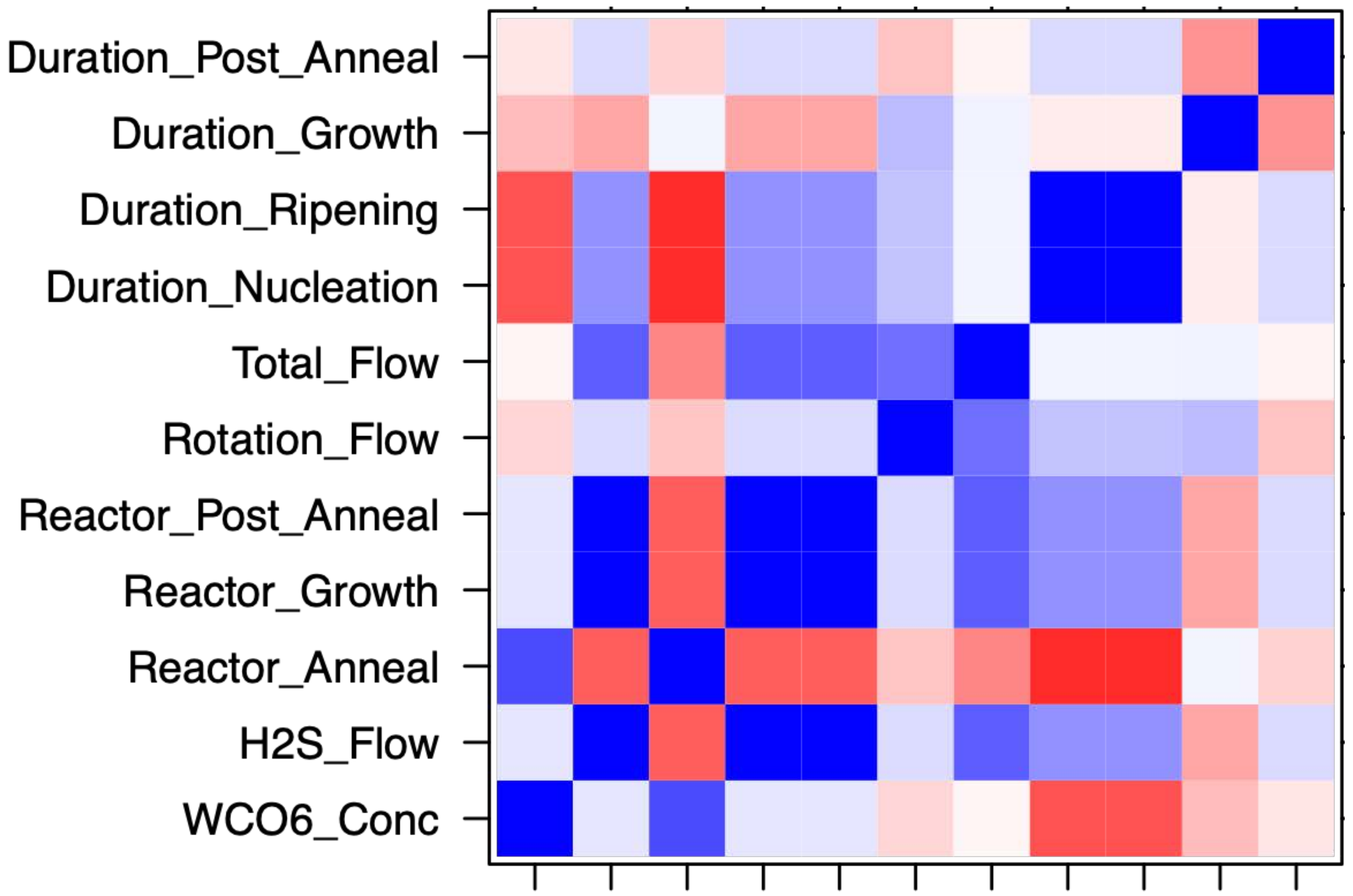


Experiments by different researchers using different conditions or facilities will be leveraged to build common models of underlying mechanisms, augmented by kinetics-focused computations, characterization data, and natural language sources such as researcher comments and published literature, with quantified uncertainties.

LiST 2.0 aims to enable the research community to pursue new modes of materials discovery.

A platform to reveal hidden relationships in the data





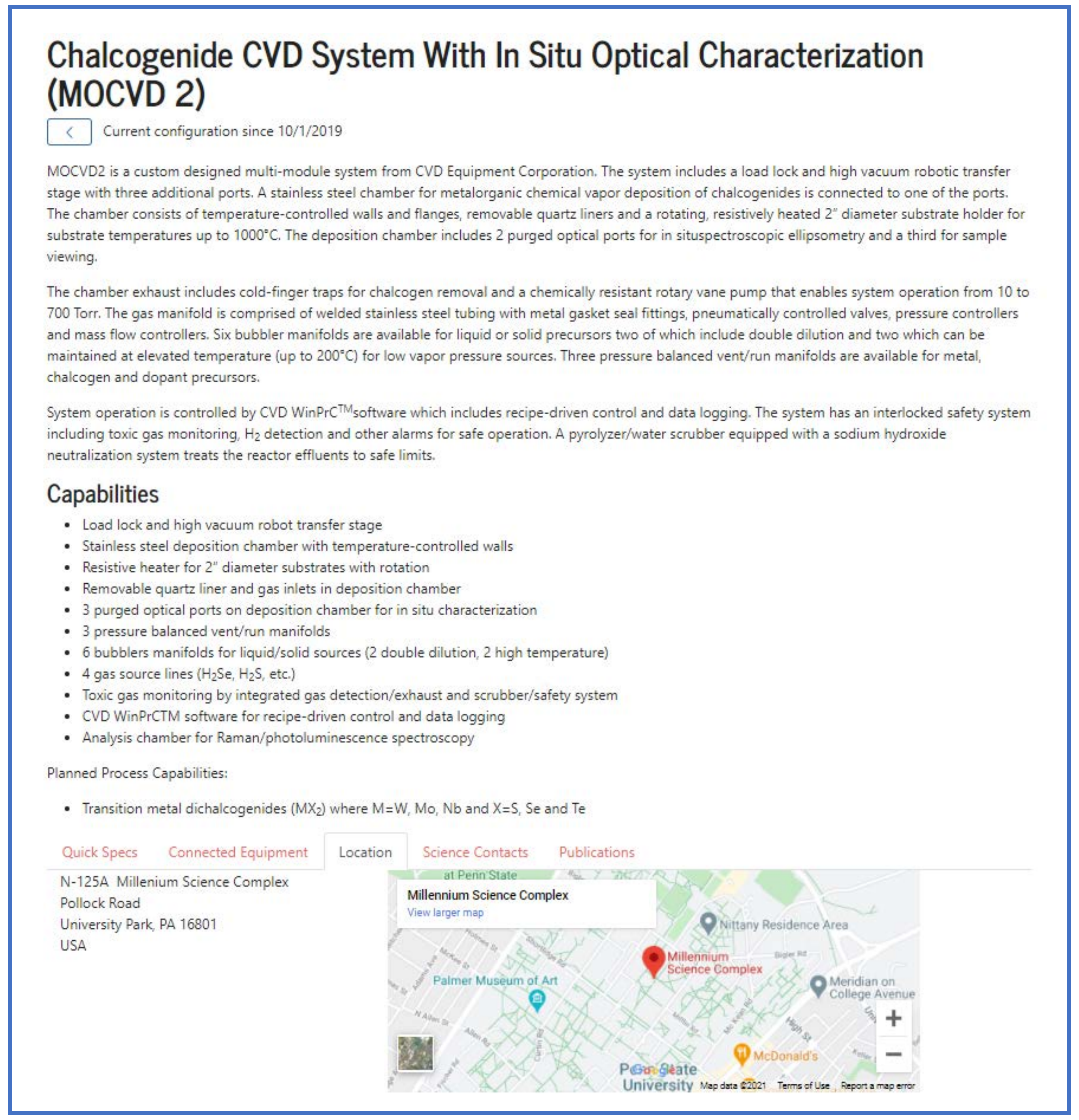
User project S0073

The LiST 2.0 ontology will describe relationships between samples, tools, people, characterization data, simulations, etc. in context-aware, graph-structured data with machine-learning pipelines that can infer new facts and uncover otherwise hidden relationships:

- Recommend optimizations of growth conditions (what helps?)
- Accelerate diagnosis of faulty protocols or equipment (what changed?)
- Suggest new experiments (most informative, most promising) or new collaborations
- Facilitate similar growth outcomes in different growth tools across different sites
- Enable useful data aggregation across different growth tools in different groups (↔multiscale modeling)

Data Infrastructure for Materials **Instrumentation**

- A national database of **DMR-supported instruments.** (Placeholder name "DMR-IDB").
- Semantic web modules for local content management systems will **embed machine-readable metadata on instruments** into facility and investigator webpages.
- This will **integrate into existing workflows:** updating an instrument's institutional webpage will automatically update the central server.
- A central server will "spider" these pages, similar to how a search engine gathers information, to **populate a central database** where it is searchable by location, capability, access, per-publication, etc.
- A **unique ID for each instrument** allows e.g. publication acknowledgements that can be tracked and analyzed by PIs, NSF, and 3rd-party researchers.



Chalcogenide CVD System With In Situ Optical Characterization (MOCVD 2)

Current configuration since 10/1/2019

MOCVD2 is a custom designed multi-module system from CVD Equipment Corporation. The system includes a load lock and high vacuum robotic transfer stage with three additional ports. A stainless steel chamber for metalorganic chemical vapor deposition of chalcogenides is connected to one of the ports. The chamber consists of temperature-controlled walls and flanges, removable quartz liners and a rotating, resistively heated 2" diameter substrate holder for substrate temperatures up to 1000°C. The deposition chamber includes 2 purged optical ports for in situspectroscopic ellipsometry and a third for sample viewing.

The chamber exhaust includes cold-finger traps for chalcogen removal and a chemically resistant rotary vane pump that enables system operation from 10 to 700 Torr. The gas manifold is comprised of welded stainless steel tubing with metal gasket seal fittings, pneumatically controlled valves, pressure controllers and mass flow controllers. Six bubbler manifolds are available for liquid or solid precursors two of which include double dilution and two which can be maintained at elevated temperature (up to 200°C) for low vapor pressure sources. Three pressure balanced vent/run manifolds are available for metal, chalcogen and dopant precursors.

System operation is controlled by CVD WinPrC™software which includes recipe-driven control and data logging. The system has an interlocked safety system including toxic gas monitoring, H_2 detection and other alarms for safe operation. A pyrolyzer/water scrubber equipped with a sodium hydroxide neutralization system treats the reactor effluents to safe limits.

Capabilities

- Load lock and high vacuum robot transfer stage
- Stainless steel deposition chamber with temperature-controlled walls
- Resistive heater for 2" diameter substrates with rotation
- Removable quartz liner and gas inlets in deposition chamber
- 3 purged optical ports on deposition chamber for in situ characterization
- 3 pressure balanced vent/run manifolds
- 6 bubblers manifolds for liquid/solid sources (2 double dilution, 2 high temperature)
- 4 gas source lines (H_2Se, H_2S, etc.)
- Toxic gas monitoring by integrated gas detection/exhaust and scrubber/safety system
- CVD WinPrCTM software for recipe-driven control and data logging
- Analysis chamber for Raman/photoluminescence spectroscopy

Planned Process Capabilities:

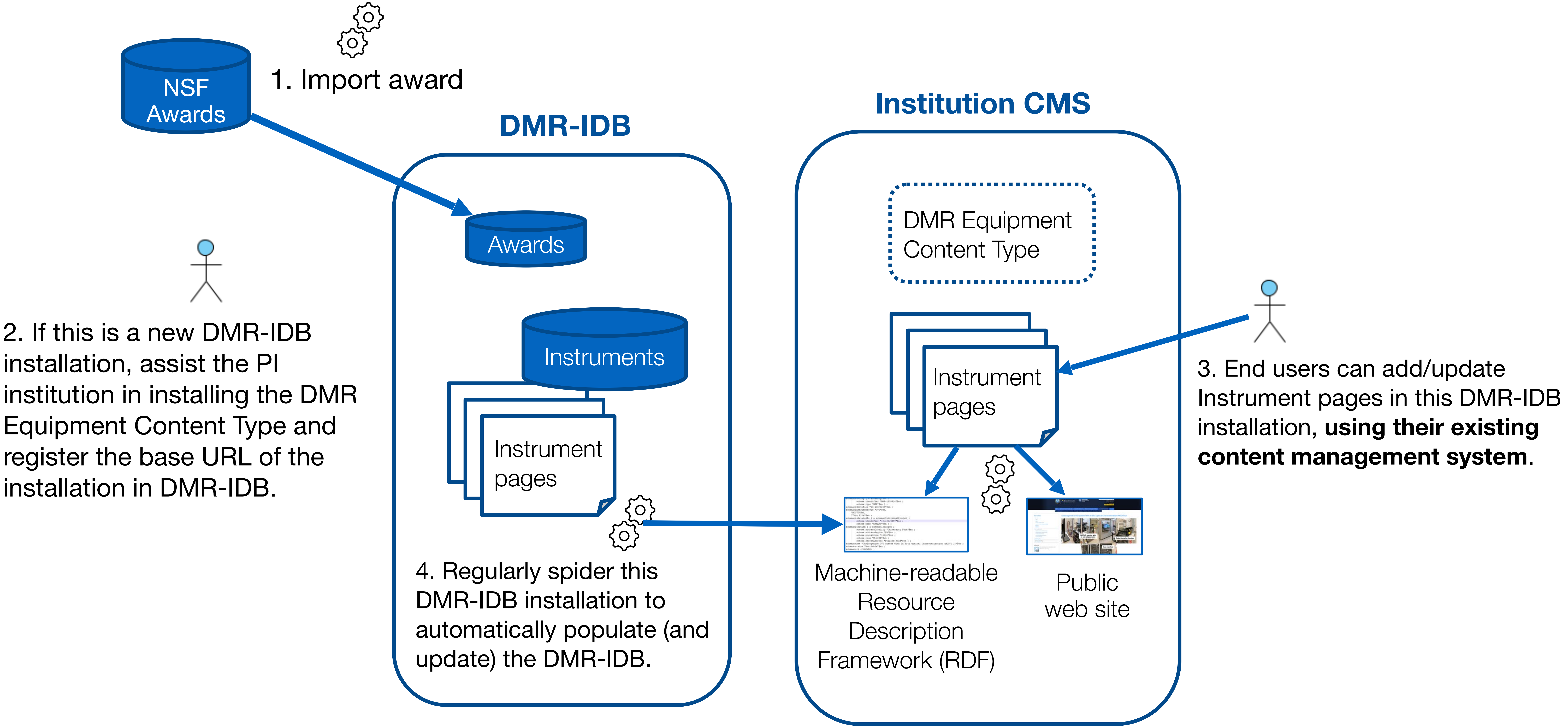
- Transition metal dichalcogenides (MX_2) where M=W, Mo, Nb and X=S, Se and Te

Quick Specs | Connected Equipment | Location | Science Contacts | Publications

N-125A Millenium Science Complex
Pollock Road
University Park, PA 16801
USA

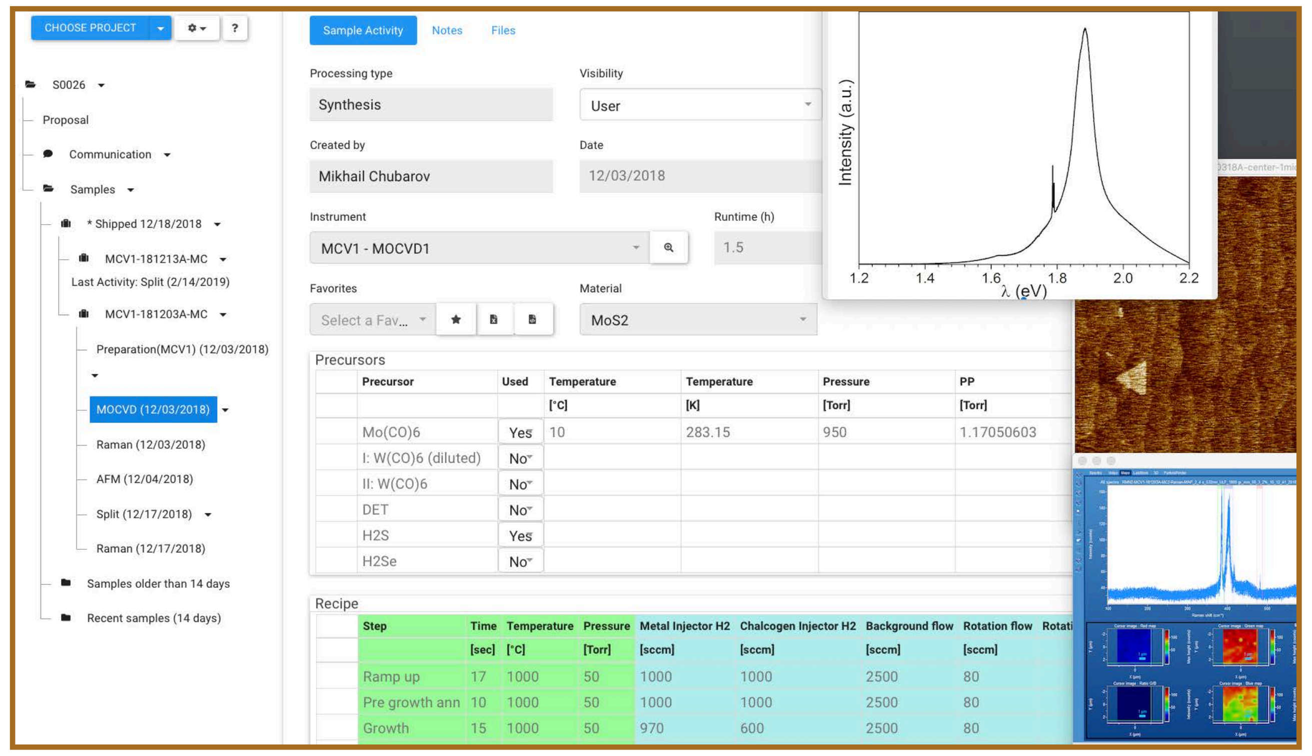
Initial demo complete. Currently supported by a 12-month supplemental funding request.

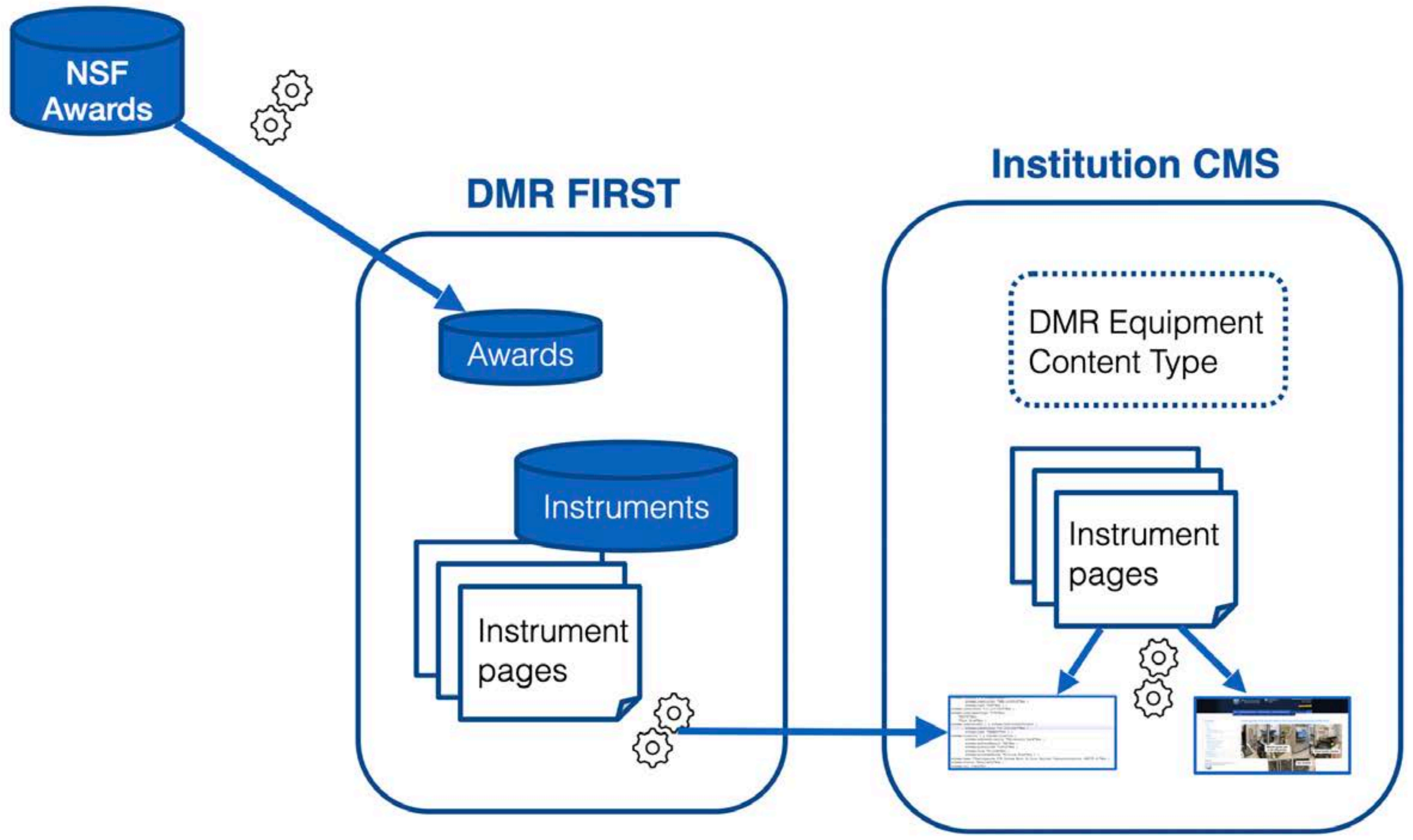
Integrate as Seamlessly as Possible into existing Facility Workflows



(For instruments not integrated into institutional facilities, the central server hosts the instrument data directly, providing the instrument PI with a web app for updates and serving the page content into their institution's web setting)

Data Infrastructure for Materials **Growth**





Data Infrastructure for Materials **Instrumentation**

2DCC-MIP