Area 3

Big Data Approaches to Knowledge Discovery

Dr. Etienne Gnimpieba, Carol Lushbough

Biomedical Engineering

University of South Dakota

Area 3 - Big Data Approaches to Knowledge Discovery

Goals

1. Development of a Biofilms Knowledge and Information Discovery System (Biofilm-KIDS) to assist researchers in screening for 2D materials that withstand corrosive effects of SRB biofilms.

2. Apply data mining and machine learning approaches to predict 2D material functions, biofilm phenotypes, and bio corrosion resistance in response to surface properties (Area 1).

3. Develop an infrastructure to assist in the correlation of gene expression patterns between specific plant pathways (that confer bacteria colonization) and specific microbial pathways (Area 2)

Overview

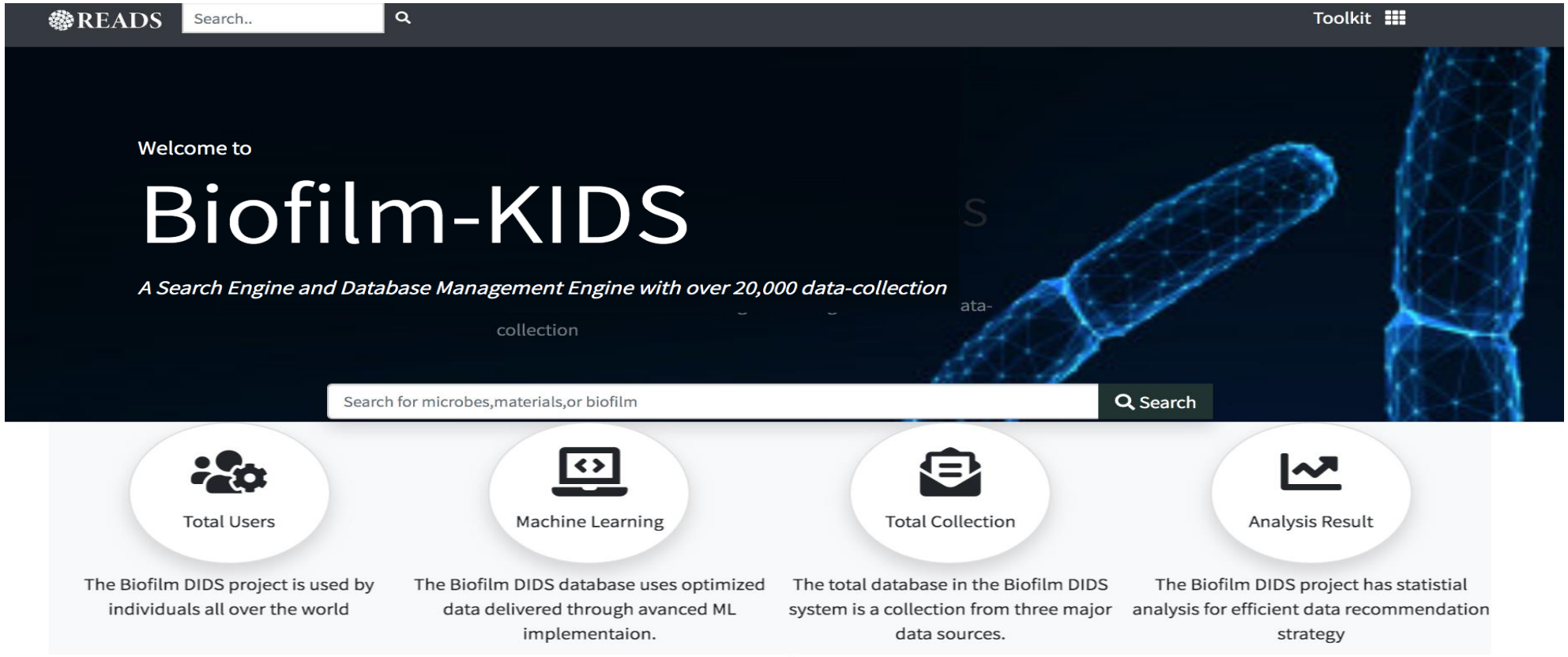
Question: What resources can be developed and leveraged to help address a questions related to 2D materials to biofilm phenotype investigations?

Assertion: At each level from 2D materials to biofilm phenotype investigations, applying predictive and systematic tools such as machine learning and systems biology to integrated preexisting and related project task data will play an important role in improving research outcomes.

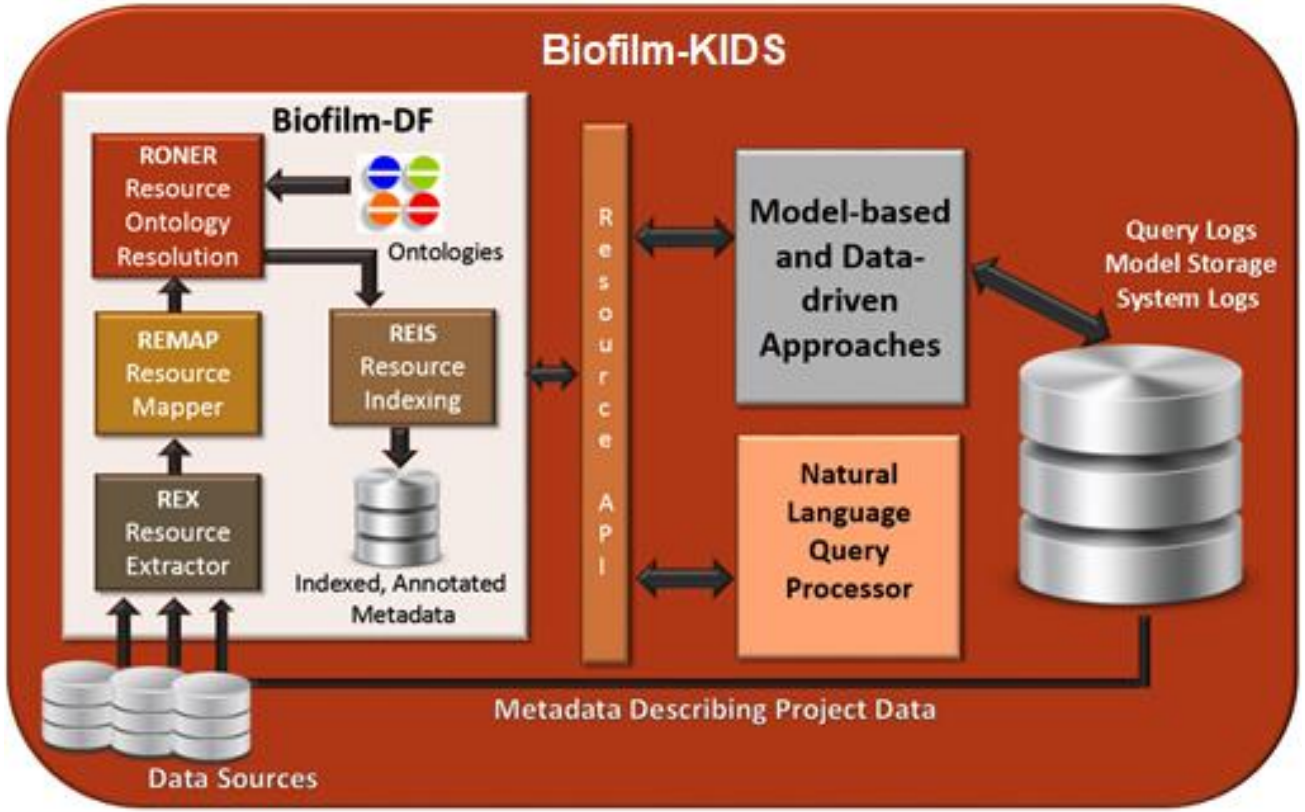
Approaches:

1. Use data mining to extract existing dataset from literature and bioscience databases,
2. Build a free text based search engine to assist user in their discovery,
3. Build a comprehensive, predictive, multilayer framework for knowledge discovery, and
4. Get the community involved to curate and enrich the collection.

Biofilm-KIDS Version 0.1: Biofilm Knowledge Information and Data Discovery System

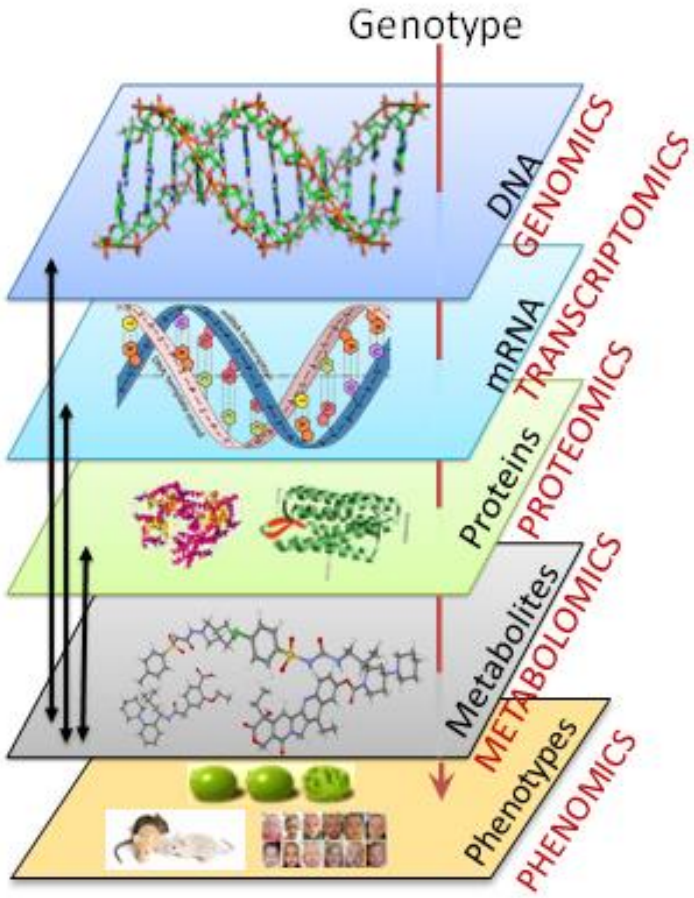


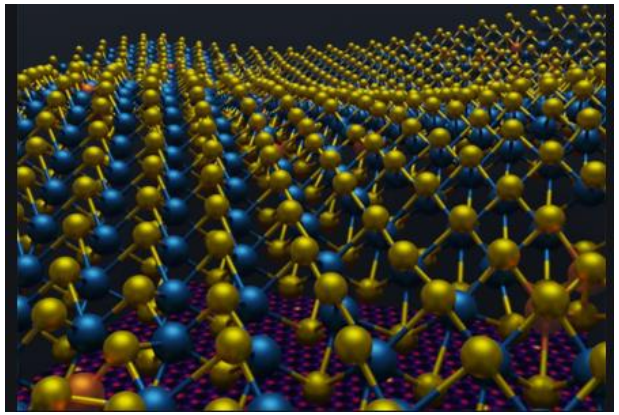
Task 1 - Biofilms Knowledge and Information Discovery System (Biofilm-KIDS) Architecture

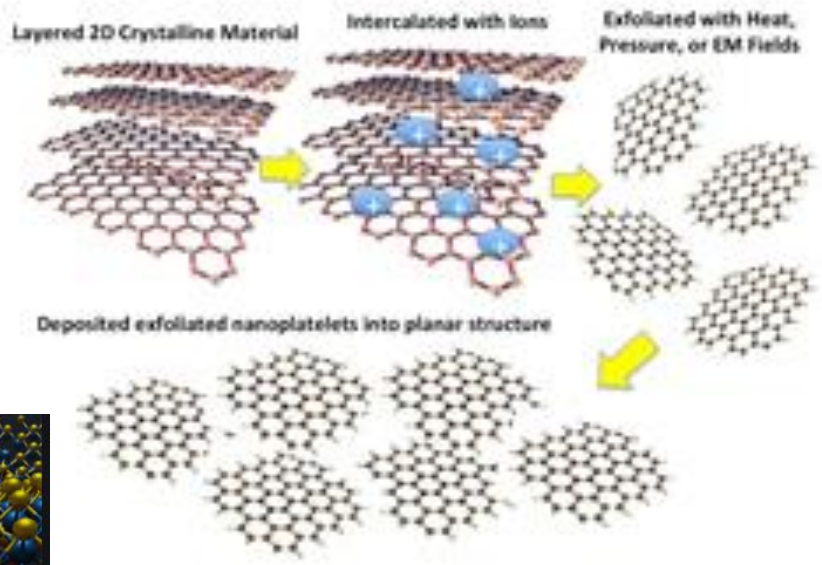


Data Sources

Biofilm-KIDS Data Sets



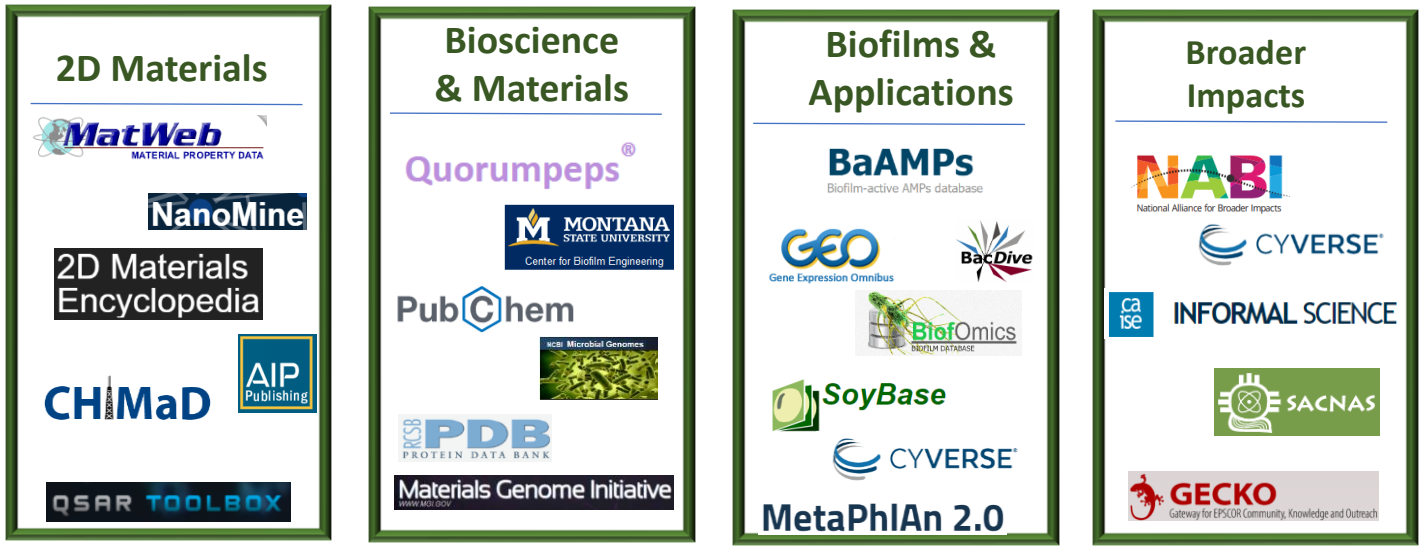




Biofilm-KIDS

Sample Use Case: Identify materials and the surficial properties that impact DA-G20 biofilm phenotypes

- Expand database collection.



Biofilm-KIDS Data Sets

The current collections that have been integrated into Biofilm-KIDS version 0.1 are:

- NCBI Microbe,
- BaAMPs Biofilm-active AMPs Database, and
- BacDive – Bacterial Diversity Meta-database.

NCBI Microbial Genomes

https://www.ncbi.nlm.nih.gov/genome/microbes/

- Microbial Genomes resource presents public data from prokaryotic genome sequencing projects.
- Prokaryotes are the earliest forms of life, appearing on earth 4 billion years ago.
- The Prokaryotes include the Archaea, which include inhabitants of some of the most extreme environments on the planet, and the Bacteria, which include both important pathogens and producers of fermented food, antibiotics, and vitamins.

BaAMPs Biofilm-active AMPs Database

http://www.baamps.it/

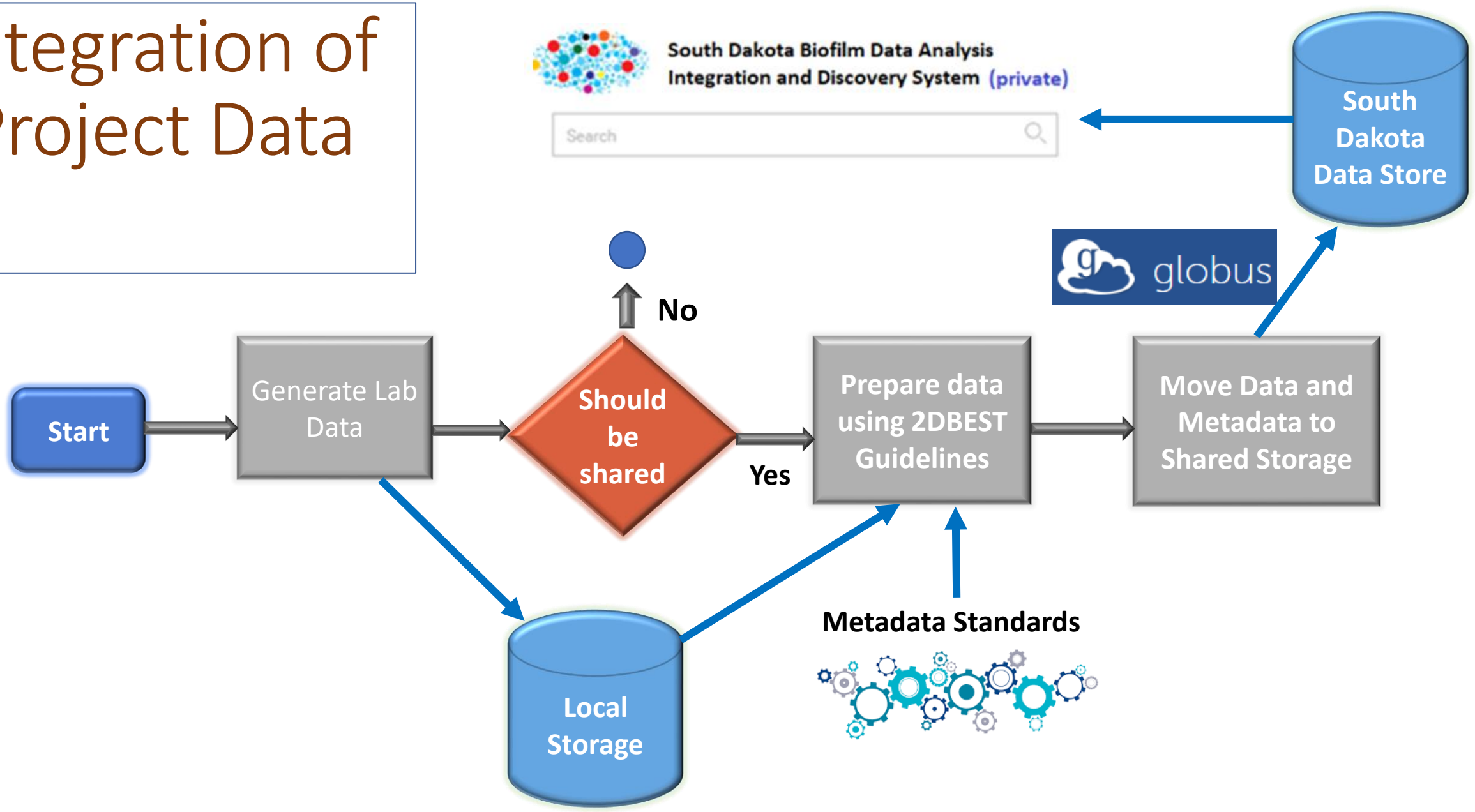
- Dedicated to antimicrobial peptides (AMPs) specifically tested against microbial biofilms.
- The aim of this project is to provide useful resources for the study of AMPs against biofilms to microbiologist, bioinformatics researcher and medical scientist working in this field in an open-access framework.

BacDive – Bacterial Diversity Metadatabase

https://bacdive.dsmz.de/about

- BacDive represents a collection of organism-linked information covering the multifarious aspects of bacterial and archaeal biodiversity.
- The content encloses information on taxonomy, morphology, physiology, sampling and environmental conditions as well as molecular biology.

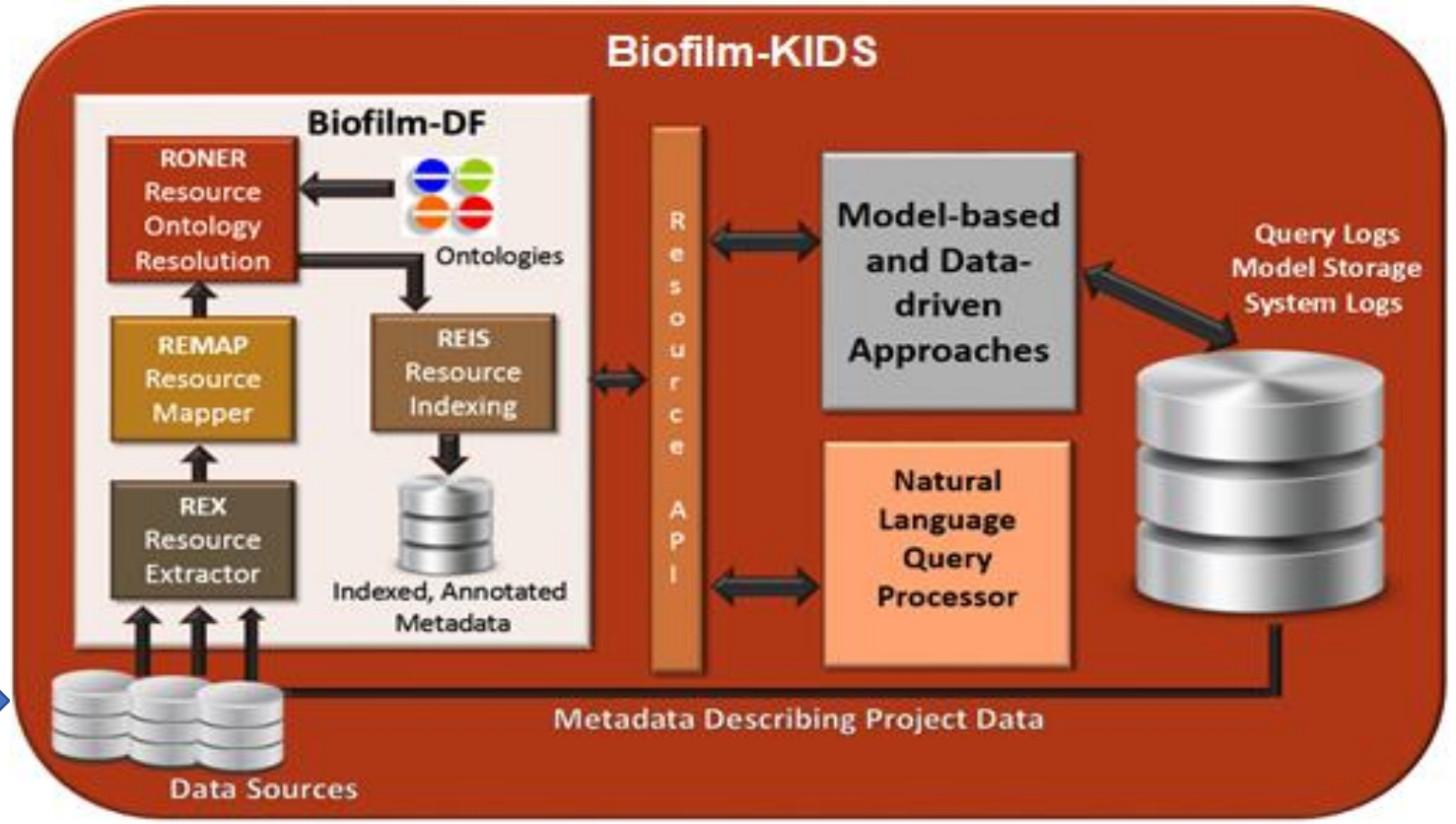
Integration of Project Data



Task 1 - Biofilms Knowledge and Information Discovery System (Biofilm-KIDS) Architecture

Data Sources

- Publications
- Project Data
- NCBI Microbe
- BaAMPs
- BacDive



Biofilm-KIDS Standards and Ontologies

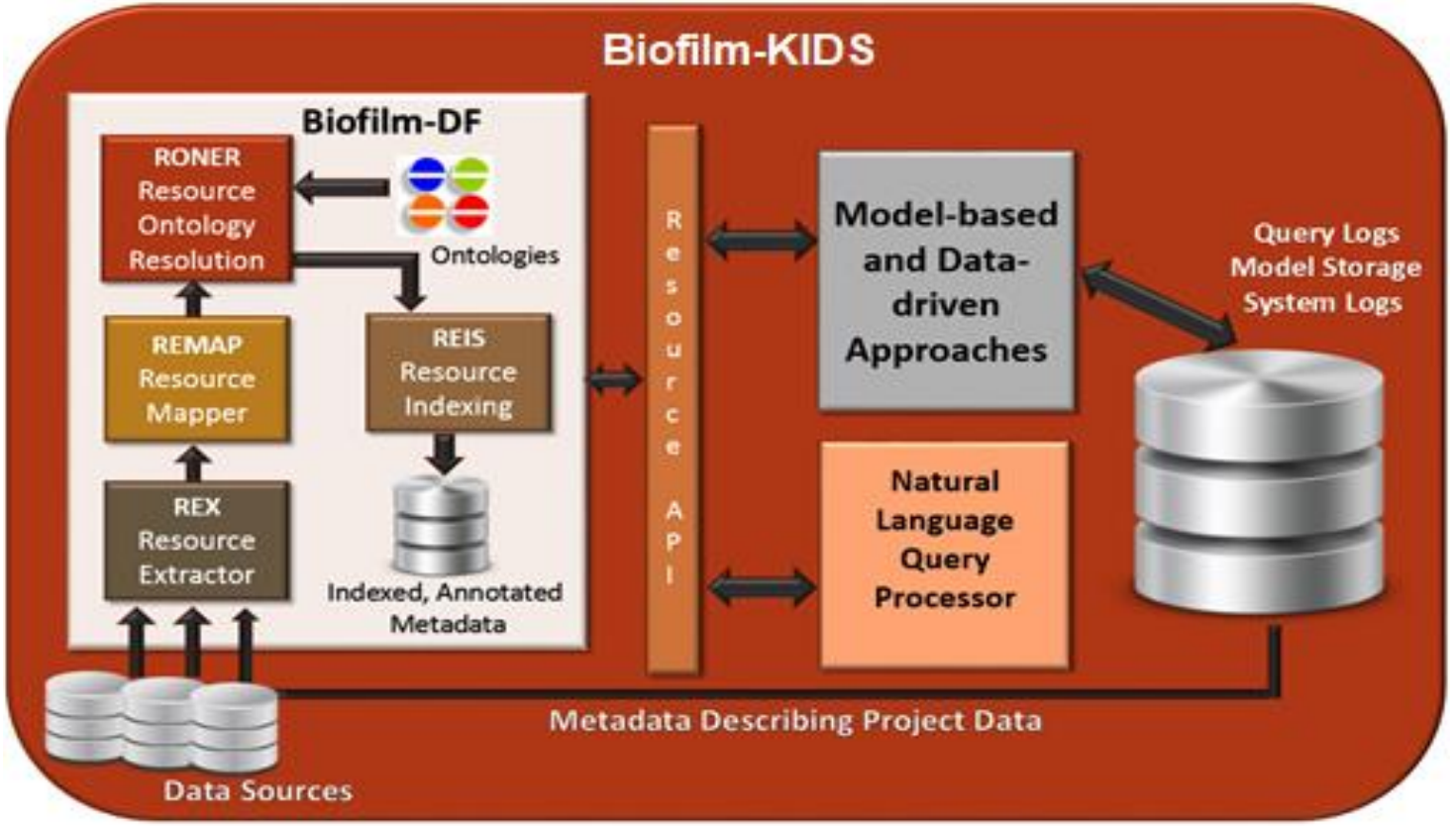
- We are following the research community's minimum information about biofilm experiment (MIABiE) standards for storing, sharing and publishing our biofilm experimental data.
- Other Minimum Information Standards will be used to standardize the description of other generated research data.
- Ontologies integrated in the system include Metagenome and Microbes Environmental Ontology and Word net from for English Dictionary (e.g. Python NLTK Corpus)

Task 1 - Biofilms Knowledge and Information Discovery System (Biofilm-KIDS) Architecture

Ontologies & Standards

MIABiE

Metagenome Microbes

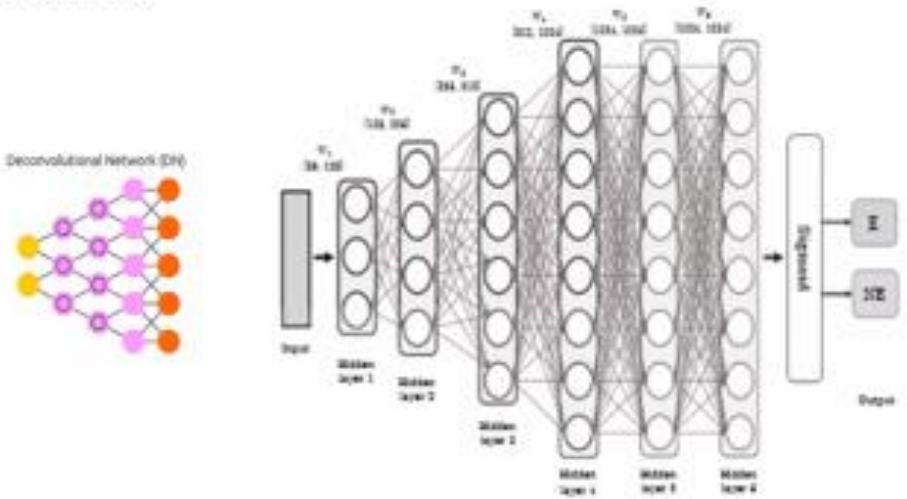


SGS 1.0 – Smart Gene Set Predictor (Python package)

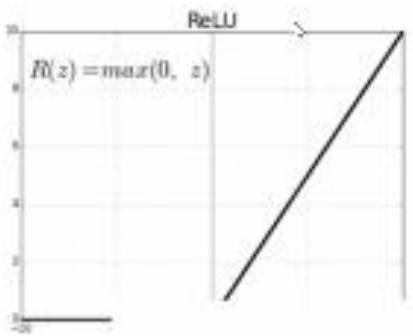
Input: Microbe genomics and systems biology features (variables) dataset

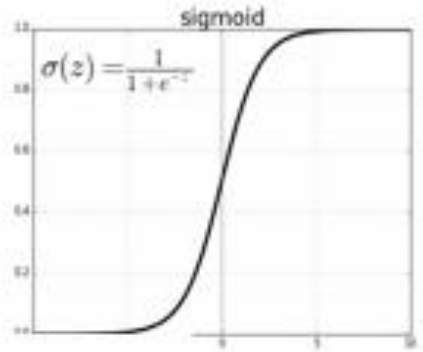
Output: Essential gene list involved in microbe phenotype (e.g. survival) in a given environment

2.Model Arch: deconvolutional network with an additional sigmoid layer for output.



3. Function / Algo

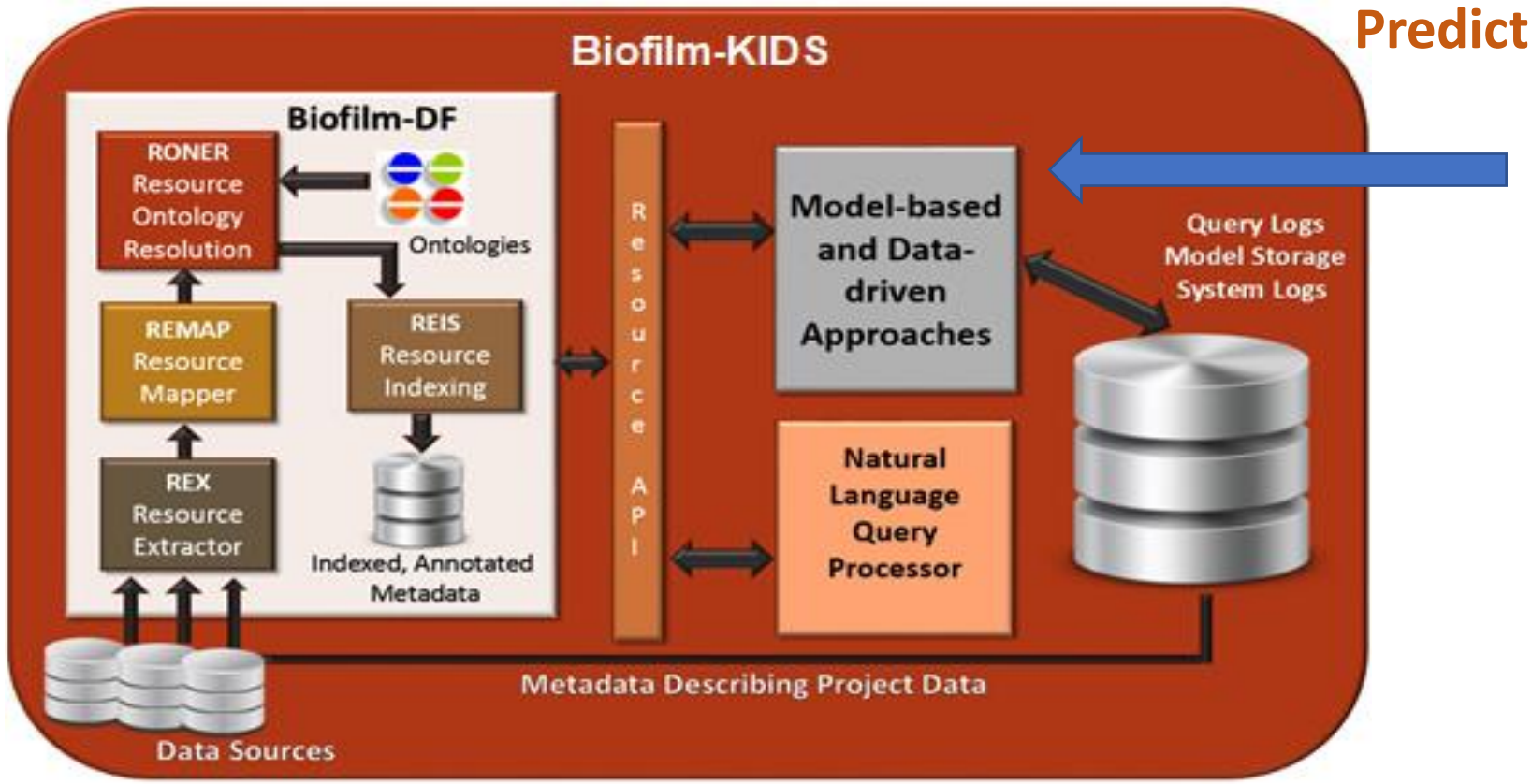




$$L_{\text{cross-entropy}}(\hat{y}, y) = -\sum_i y_i \log(\hat{y}_i)$$

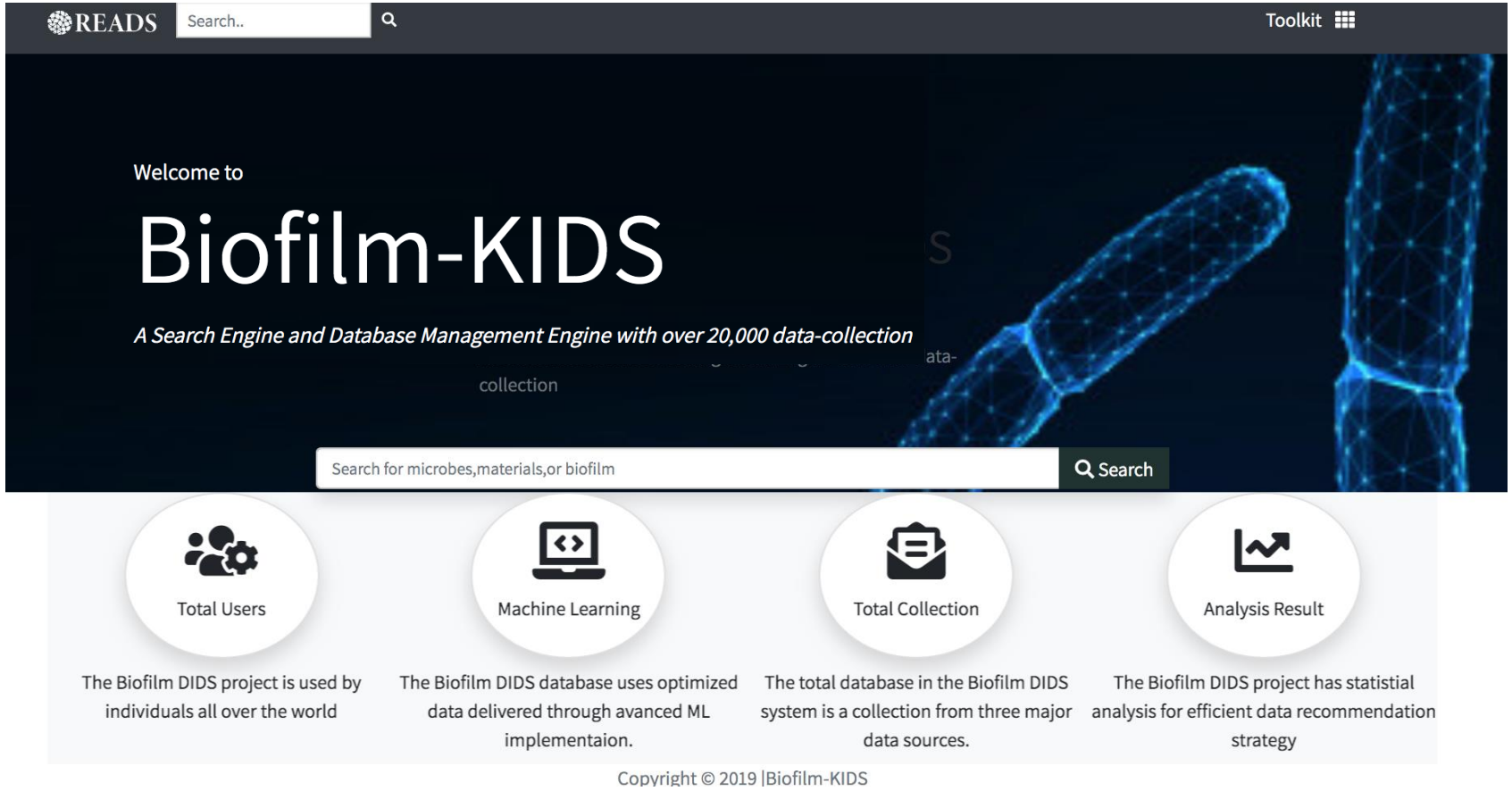
AUC: ~ 0.80, loss is calculated via cross-entropy at the bottom and backpropagation/optimization is handled by gradient descent.

Task 1 - Biofilms Knowledge and Information Discovery System (Biofilm-KIDS) Architecture

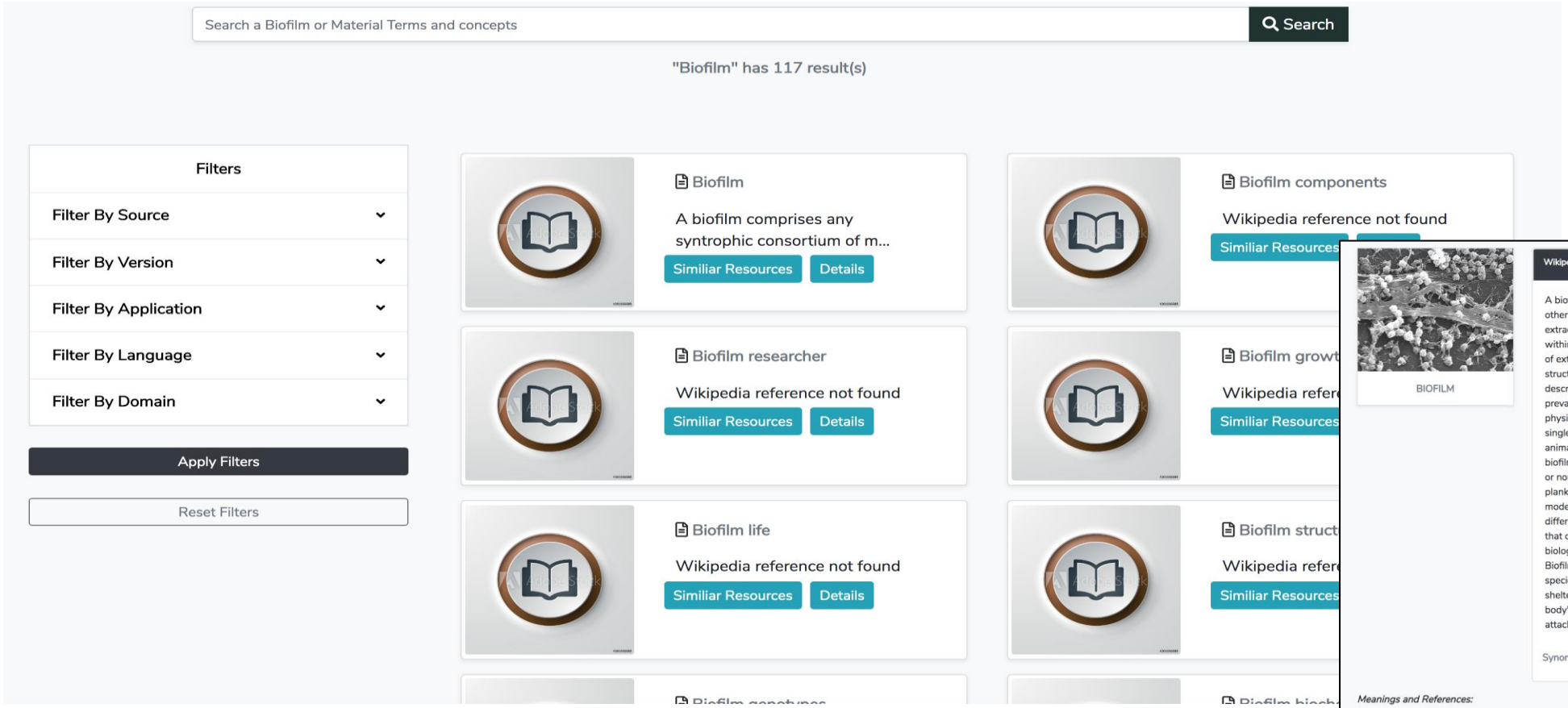


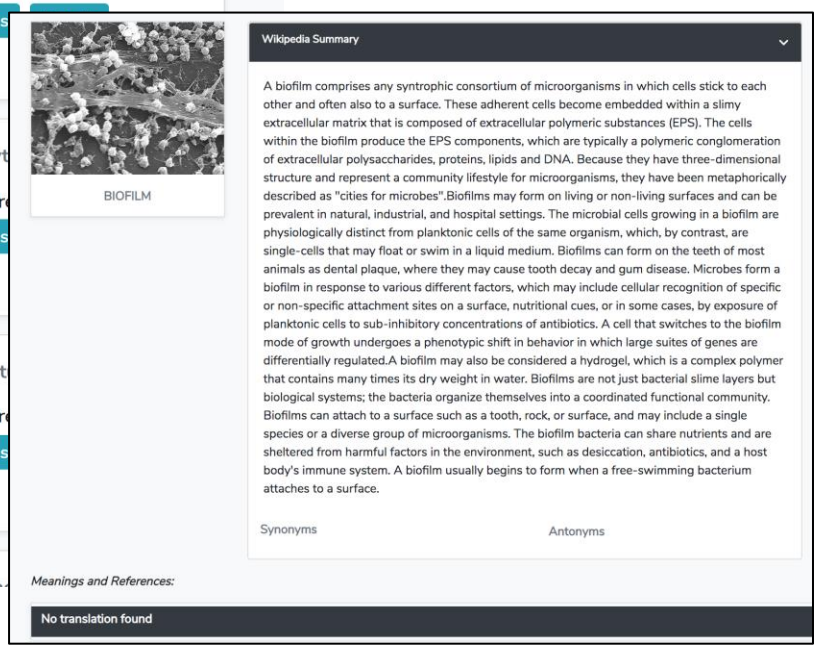
Biofilm Materials Search Engine

- Apply Natural Language Processing and domain ontology annotations to enrich the collection



Biofilm Materials Search Engine





Biofilm-KIDS Year 2 Goals

- Apply datamining and machine learning techniques to update collection using semi-automatic curation system leveraging identified data repositories and published review papers.
- Integrate project experimental data hosted by South Dakota Data Store into Biofilm-KIDS by leveraging Globus API
- Publish initial results.

Data Management and Sharing Overview

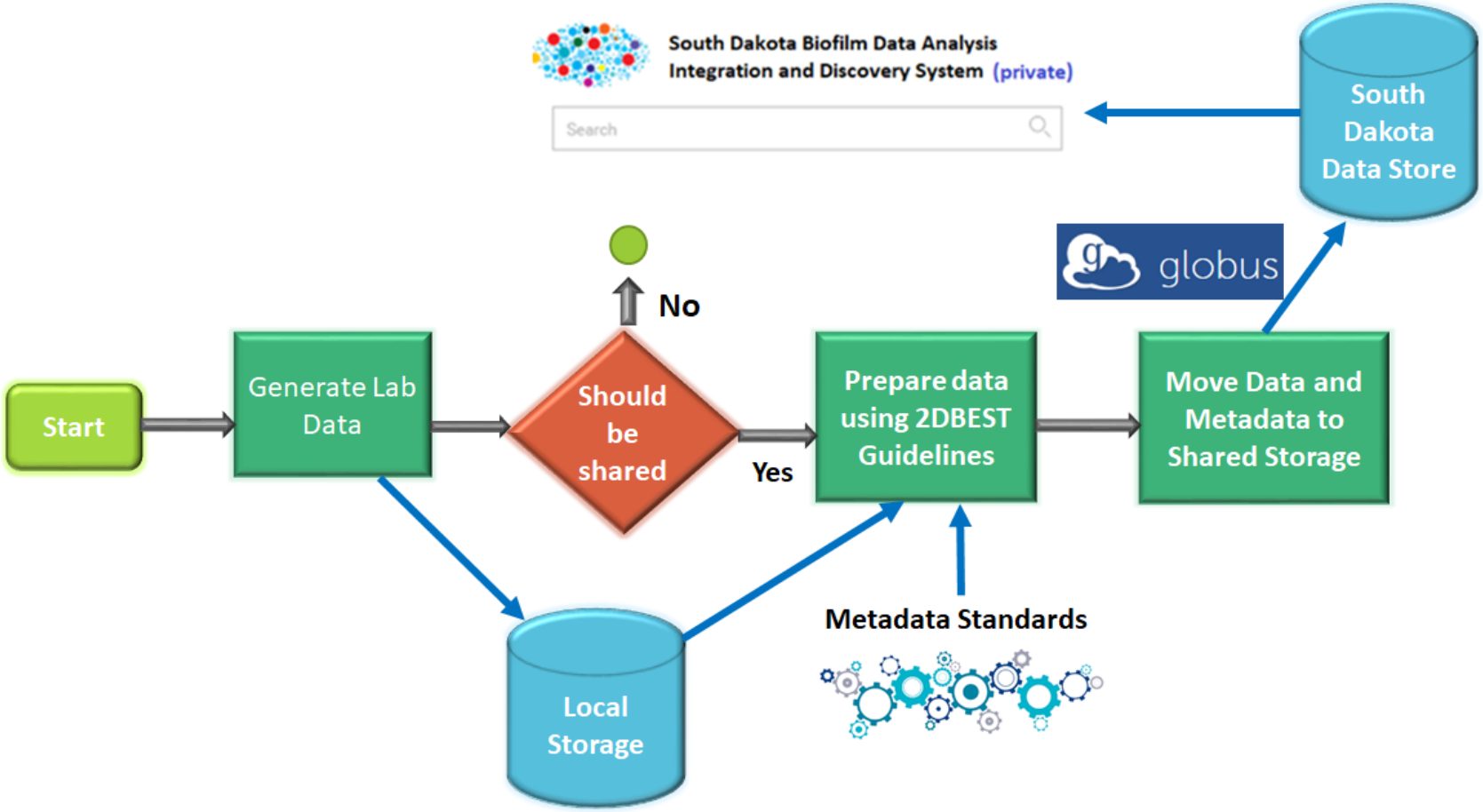
http://sd2dbest.org/

- One of the challenges of the South Dakota Biofilm Science and Engineering Center is the integration of large-scale, diverse datasets and analytic tools into a comprehensive framework to help provide a basis for a system level understanding of 2D materials – Biofilm interactions.
- To facilitate sharing of data among participants and with the broader scientific community, we are adopting open-source data formats.
- Importantly, we are employing the FAIR Data Principles in order to ensure that the data generated through this project is findable, accessible, interoperable and reusable.

Data Management and Sharing Overview

- The 2D-BEST team has developed a workflow to facilitate data management, sharing, and publishing.
- Once it is determined that experimental data should be shared with the project team, researchers follow the requisite protocol to move the data, along with its standard description, to shared storage.
- This project is leveraging the University of South Dakota Data Store (SDDS) and Globus in this data management plan.
- The SDDS provides high-reliability, high-availability, network-accessible storage for South Dakota researchers and collaborators.
- Globus is a not-profit service for secure, reliable research data management.

Data Management and Sharing Overview



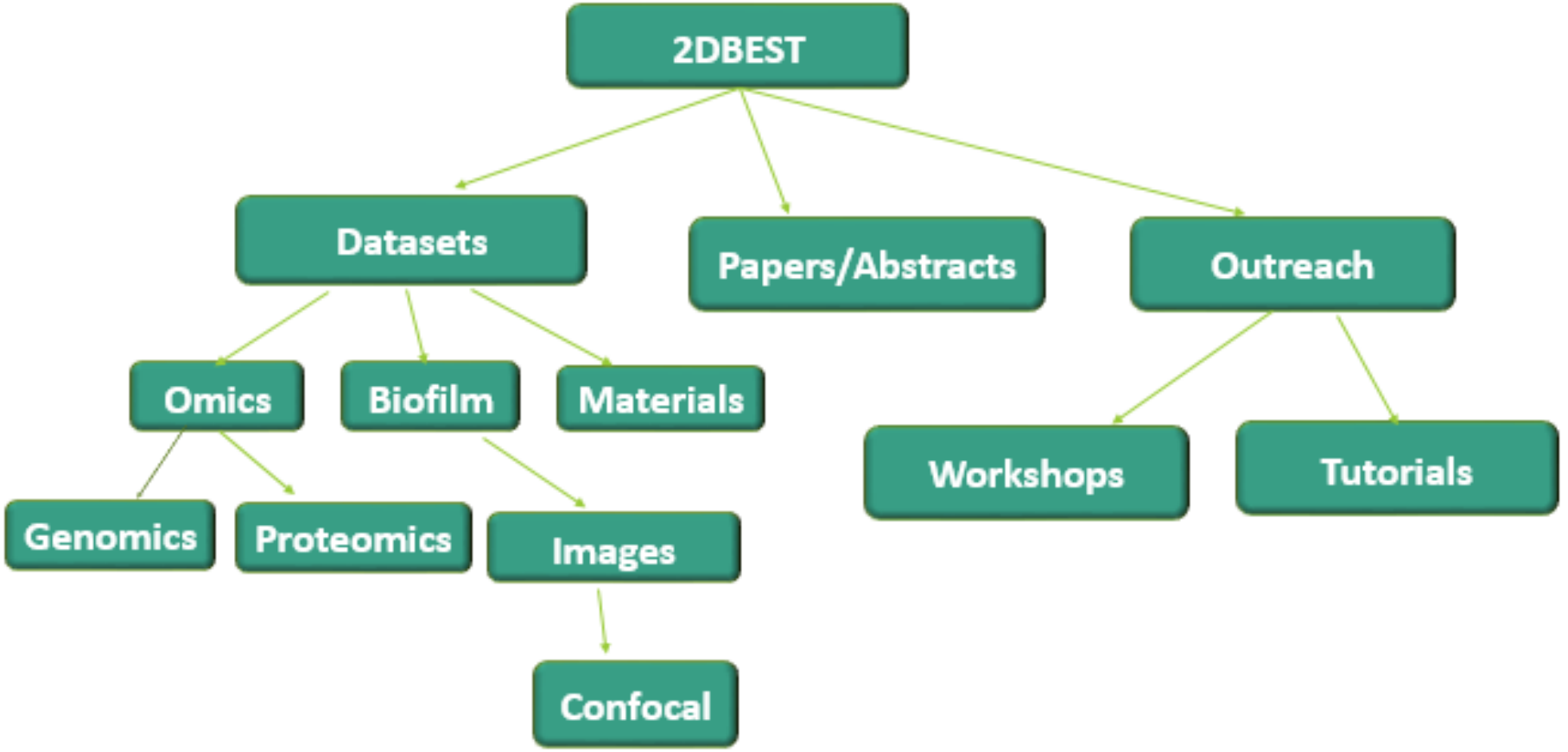
2D-BEST Data Sharing Steps

- Describe Data Types and Sources
- Register with Shared Data Store
- Move Data to Shared Storage
- Identify or Create a User Group
- Share Data with Collaborators

Guide to Writing Metadata Description

- All data collections moved to shared storage will be associated with a “readme” metadata description text file.
- We’ve adopted the guidelines outlined by the Comprehensive Data Management Planning Cornell University

2D-BEST Data Collection Examples



2D-BEST Data Collection Support

Move Data to Shared Storage

Organizing the 2D-BEST data will be important in order to provide easy accessibility to collaborators. Figure 2 offers an example of how that dataset collections could be categorized. After you have determined what data you will share and have completed the readme file that describes the data, the next step is to transfer the data to shared collection using the Globus platform.

To transfer data to shared storage:

- Identify where in the data collection organization you will place your data. The best way to approach this task is to contact one of the project data administrators. Below is a list of the 2D-BEST Data Administrators

Institution	Contact	Email
South Dakota School of Mines & Technology	Shankarachary Ragi	Shankarachary.Ragi@sdsmt.edu
South Dakota State University	Sen Subramanian	Senthil.Subramanian@sdstate.edu
University of Nebraska, Omaha	Parvathi Chundi	pchundi@unomaha.edu
University of South Dakota	Carol Lushbough	Carol.Lushbough@usd.edu
University of South Dakota	Etienne Gnimpieba	EtienneGnimpieba.usd.edu

Data Management
Next Steps Publishing

- We are still in the early days of publishing data.
- Our plan will be to leverage USD's Research, Engage, Design (RED) services to facilitate project data publishing.
- RED is a service of the University of South Dakota University Libraries that promotes and shares the scholarship, creative works, and data created by South Dakota faculty, students, and institutional partners .

Data Management Next Steps
Publishing