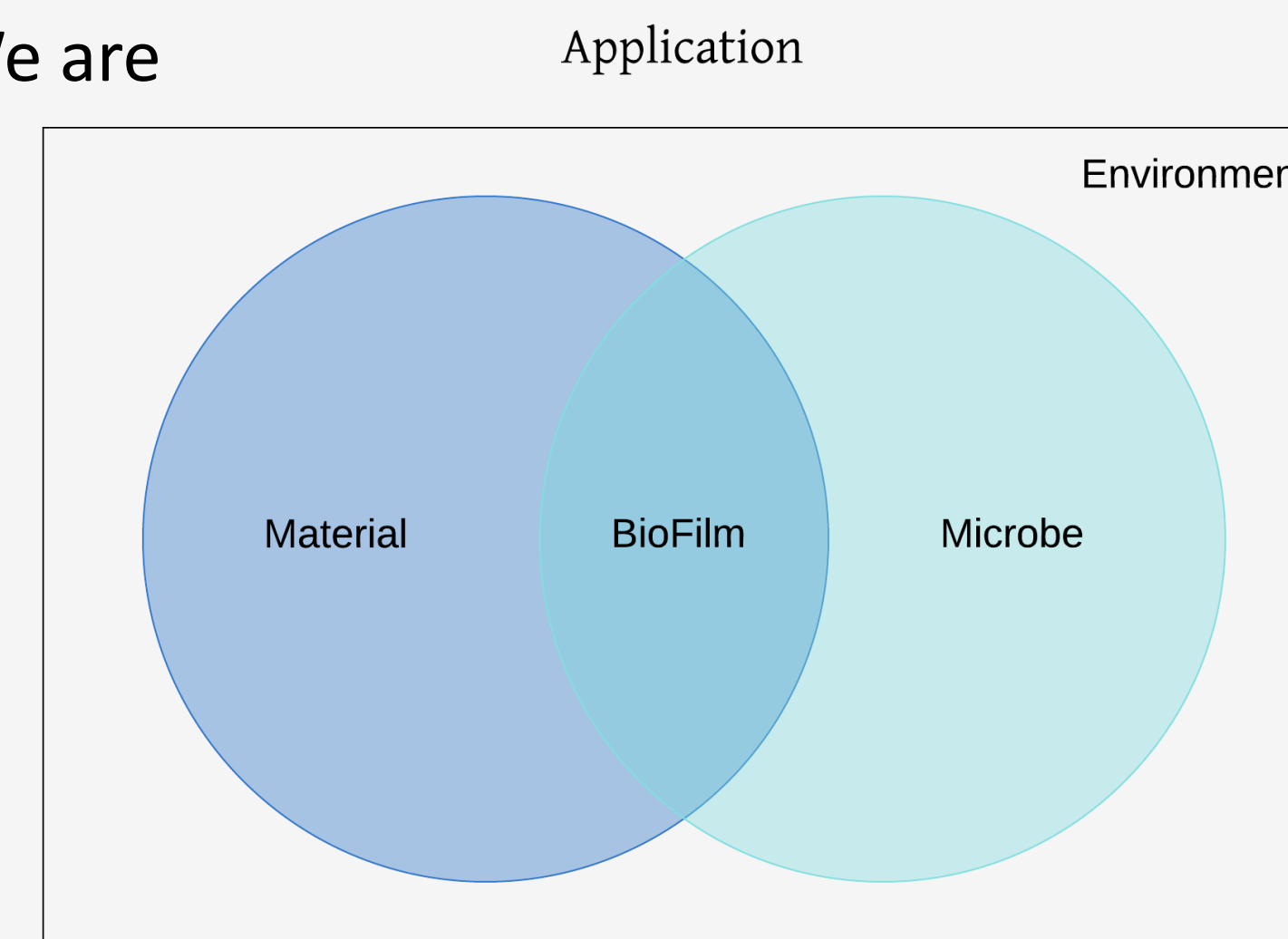


# Dataset Collection for the Biofilm Data and Information Discovery System (Biofilm-DIDS)

Subrat Subedi, Carol Lushbough, Etienne Z. Gnimpieba  
*Biomedical Engineering Program, University of South Dakota*

## Introduction

The Bioinformatics Tool Discovery System (Bio-TDS, *biotds.org*) was developed to assist researchers in retrieving the most applicable analytic tools by allowing them to formulate their questions as free text. We are working with SDSMT on a project that would leverage the Bio-TDS architecture in the development of a Biofilms Data and Information Discovery System (Biofilm-DIDS). Biofilm-DIDS is a Data-Driven Material-Biofilm Discovery Framework that collects and combines disparate big datasets pertinent to biology and material scientists, using artificial intelligence (machine learning approaches) and natural language processing. This system will be used to analyze and predict microbial responses and biofilm phenotypes impacted by nanosomic surficial properties. The primary objective of this project is to develop an extraction module to retrieve the most relevant dataset to populate our Biofilm-DIDS.



## Material and Methods

For accurate searching of these biofilm related data sets, it is important to have meaningful annotations describing the microbes, their phenotypes and their behavior with various surfaces. Considering the large set of information to be analyzed, the Biofilm-DIDS platform is comprised of four core modules including:

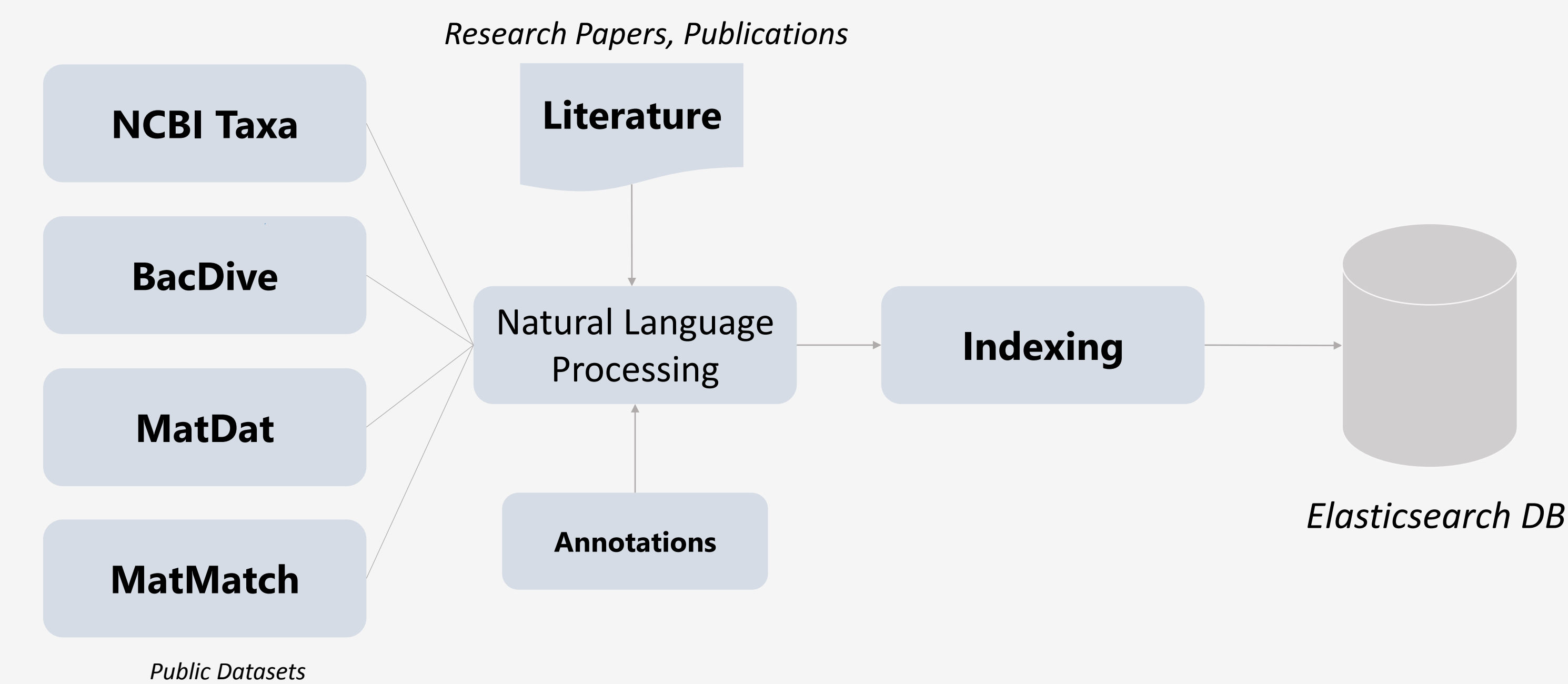
**Biofilm Data Fusion**

**Modeling and Data-driven Approaches**

**Natural Language Query Processor**

**Query Log Module**

The data fusion module retrieves, curates, annotates and indexes metadata from the public, disparate data sources, publications and research papers related to 2D materials, transcriptomics, proteomics, metabolomics, methylome, phenotypic information. The indexes are then integrated with the project's experimental data sets. Biofilm-DIDS stores the reference collections and other data needed to validate the biofilm hypotheses generated as a query result and return the biofilm phenotypes as function of 2D material properties.



## Results

- Extracted and annotated supporting data required for microbes and material identification in JSON format.

|                  |                |
|------------------|----------------|
| <b>NCBI Taxa</b> | 451721 entries |
| <b>BaAMPs</b>    | 221 entries    |
| <b>BacDive</b>   | 80584 entries  |

- Identified Dictionary and Patterns required for Text Mining and Natural Language Processing in order to extract data from Research Paper and Publications.

## Next Steps

- Develop Python Modules to perform paper-mining from relevant academic papers and research journals.
- Create a Front-End Interface as represented in the sample.
- Expand Microbial and Material data source collection.

*Supported by National Science Foundation/EPSCoR Award No. IIA-1355423, by the state of South Dakota's Governor's Office of Economic Development as a South Dakota Research Innovation Center (SDRIC), and with financial/match commitment from all participating institutions.*

*Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.*

## Sample Microbes

| Materials   | Sulfate-Reducing Desulfovibrio alaskensis G20 | Bacterioids sp. ECP-C1  |  |
|-------------|-----------------------------------------------|-------------------------|--|
| Copper (Cu) | Encourages Corrosion                          | Information Unavailable |  |
| SL-hBN-Cu   | Supresses Corrosion                           | Information Unavailable |  |
|             |                                               |                         |  |

Legend: ■ BioFilm Formation Unlikely ■ Possible BioFilm Formation ■ Data not available

*Representation of expected outcome*