3.18: Convergence Informatics (CI) Approach: Data Collection and Fusion

Quanpeng Yang[1], Sahil Changlani[2], Nahed Abu Zaid[2], Azizah Conerly[3], Brooke Mayers[4], Paul Westerhoff[5], Cranos Williams[3], Alexey V. Gulyuk[1], Rada Chirkova[2], Yaroslava G. Yinging[1]

STEPS Site Visit 2024

1 Materials Science and Engineering; 2 Computer Science, 3 Electrical and Computer Engineering, North Carolina State University, 4 Department of Civil, Construction and Environmental Engineering, Marquette University, 5 School of Sustainable Engineering, Arizona State University

Introduction

Convergence Informatics (CI) at STEPS

CI **integrates data and knowledge from all STEPS disciplines**, creating an informatics paradigm conducive to the diverse STEPS analytics needs.

Theme 1 Data: combinatorial materials synthesis, high throughput materials characterization, multiscale materials modeling and simulations

Theme 2 Data: testing of materials in real matrices, batch and continuous flow microcosms, microbiological experiments, mechanical and molecular modeling

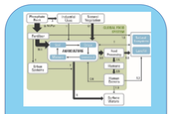
Theme 3 Data: Geospatial mapping, LCA and TEA, economic costs, ecotoxicological risks, social network assessments, stakeholders interview

Historical Data Mining: Published P -research data analyzed with web -scrapers and data analytics tools for trends, correlations, and developments

STEPS Knowledge Hub
(A Data-driven tool for phosphorus community)

Data Storage Knowledge Graph AI Tools

Users



Information Search
Data Analysis
Materials Design
P-flow Optimization

What the CI Team Does

- **Gathering of heterogeneous data** from STEPS themes and literature
- **Developing STEPS Knowledge Hub** for **phosphorus-related** research data **collection**, **storage**, **analysis**, and **utilization**

Methods

Gathering Publicly-available Phosphorus-related Data

wiley Google scholar Springer ACS Publications Web Scraper → Database (SQL)

- Using data mining tools (web-scraper with tailored functionality) to retrieve large amounts of published data and store in the SQL database

Creating Knowledge Graph and AI Tools

neo4j

Large language models

- Utilizing a set of data science tools and approaches to integrate data stored in several different formats into a relational graph database (**Neo4j**)
- Using **natural language processing (NLP)** and **large language models (LLMs)** to create AI tools for data analysis and applications

Developing Graphical User Interface (GUI) for Data Utilization

- Utilized dynamic web design approach to create GUI for interacting with STEPS Knowledge Hub

django node React HTML CSS

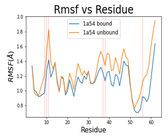
Results

Objective 1: *Data Mining*

- **Outcome:** Results of data mining stored in a database capable of further integration with STEPS Knowledge Hub
- Literature-based ontology is integrated with our pilot study ontology and incorporated in the data search tools

Objective 2: *Data Preprocessing and Knowledge Graph (KG)*

Source 1: MD Simulations



Example: Graphical information format from MD simulations of P-capturing proteins

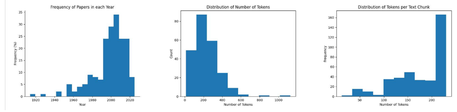
Source 2: Experimental Data

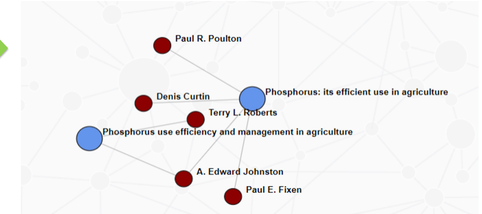
Experiment	Trial	Material	Solution	Solvent	Ion	Temperature	Absorbance	Original Concentration
Temperature	temp1.1	phosphate-binding proteins immobilized on iron oxide nanoparticles	solution-temp	clean water	Ortho-phosphate	10	0.234	1.12
Temperature	temp2.1	phosphate-binding proteins immobilized on iron oxide nanoparticles	solution-temp	clean water	Ortho-phosphate	20	0.209	1.12

Example: Spreadsheet (text and numerical) format of experimental data on P-capturing protein-based materials

Ontology
Natural Language Processing
Data Transformation

Data Statistics





- **Outcome:** Designed *KG* applies semantics to give context and relationships to data, providing a framework for data integration, unification, analytics and sharing

Objective 3: *AI Tools for Data Analysis and Utilization*

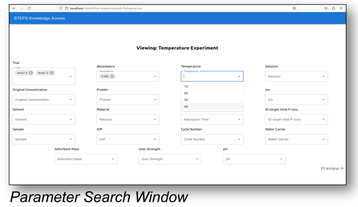
Chatbot

Paper Recommendation by AI

- **Outcome:** Developed various ML Tools for analysis for diverse datasets, including Chatbot, Papers Recommendation by AI, etc.

Objective 4: *Graphical User Interface*

STEPS Knowledge Hub Home Page



Parameter Search Window

- **Outcome:** We developed a GUI interface that parses the KG and allows users to retrieve the data from it in the most efficient way
- Dynamic web interface design allows for new experiments to be added and viewed with no changes in the code

Discussion

- The CI team developed a continuous workflow for collection, processing and sharing the data from various existing sources and new STEPS research projects. Literature-based ontology is integrated with our CI-ontology and increases the “portability” across the Themes.
- We designed databases that include simulation, experimental characterization data in various formats–available outcomes across the Themes 1, 2, and 3.
- We develop AI-powered tools allowing users to upload, preprocess, parse, search, retrieve various types of data.
- We utilized the materials informatics-based approach in which process-structure-properties-performance relations designed by analyzing large materials data sets with machine learning algorithms.

Value Added

The collaborative environment and funding provided by the Center were instrumental in fostering interdisciplinary collaboration, facilitating the implementation of key project components, and ensuring the overall success of the project.

Supports the STEPS Center's mission to achieve phosphorus sustainability through innovative, data-driven research and interdisciplinary collaboration.

Workflows developed here can be applicable to data integration in other proposals.

The module on Data imputation was incorporated into the MSE723 Materials Informatics class at NCSU.

Acknowledgements
This material is based upon work supported by the National Science Foundation CBET-2019435.

NC STATE UNIVERSITY Appalachian State University ASU CLEMSON JSNN Joint School of Nanoscience and Nanoengineering MARQUETTE UNIVERSITY RTI INTERNATIONAL UNIVERSITY OF ARKANSAS UF UNIVERSITY of FLORIDA UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN