

Populating and Refining an Ontology of Cellulose Materials with Terms from Scientific Publications

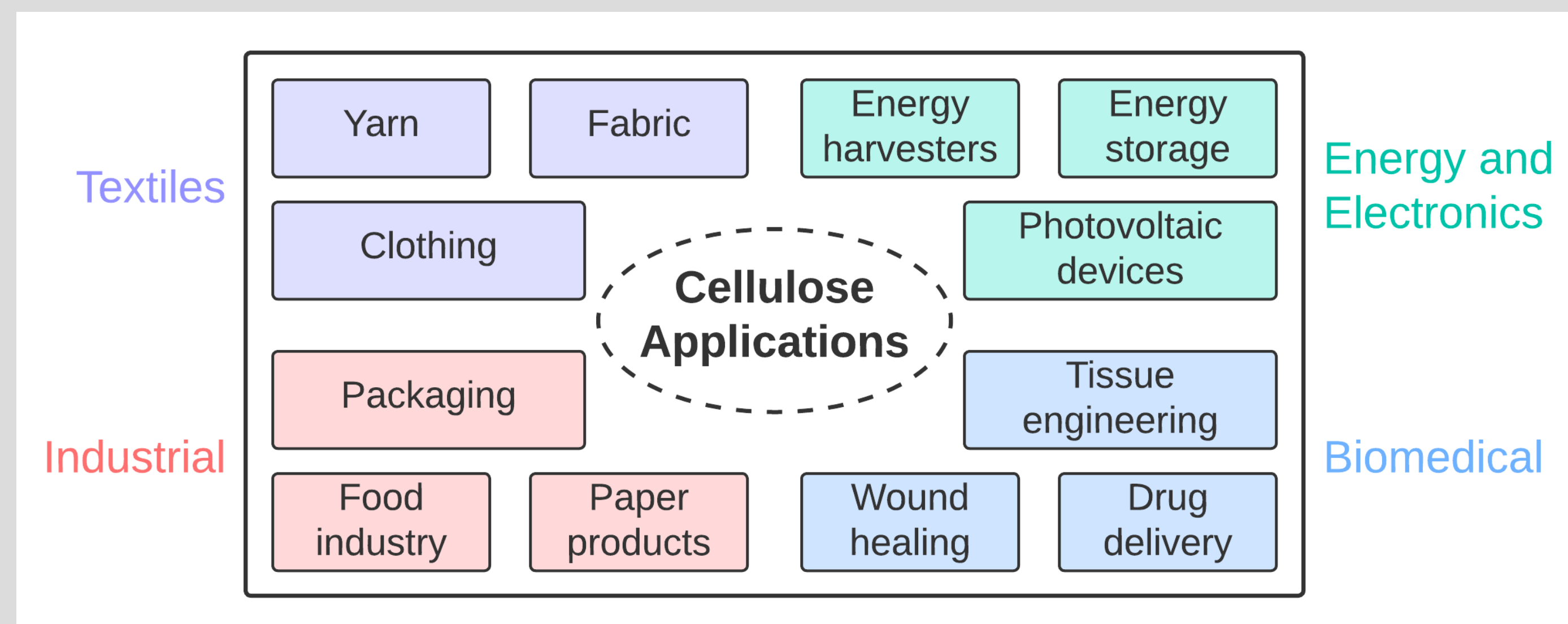
Umayer Reza, Torsten Hahmann

School of Computing and Information Science, University of Maine

{a.reza, torsten.hahmann}@maine.edu

Motivation

- Cellulose is widely utilized in various domain due to its sustainable, renewable, and biodegradable nature.
- The growth of cellulosic knowledge is directly proportional to the increasing number of scientific publications about cellulose.
- The extraction of knowledge from scientific publications is challenging due to the scattered and ambiguous nature of natural language text.



Research Objective

- **Information Extraction of Cellulosic Terms:** Design and develop an automated system to generate data for populating the cellulose ontology.
- **Classification of Cellulosic Terms:** Leverage cellulose ontology to improve performance of named entity classification.
- **Refinement of Cellulose Ontology:** Maintain the ontology up-to-date with more relevant and accurate concepts.

Research Question

- Under what conditions can pre-trained language models be incrementally re-trained for improved named entity recognition of terms in the cellulosic domain?

Evaluation Methods: Performance metrics (precision, recall, and f1-score), Benchmark dataset, Comparative analysis, and Expert feedback

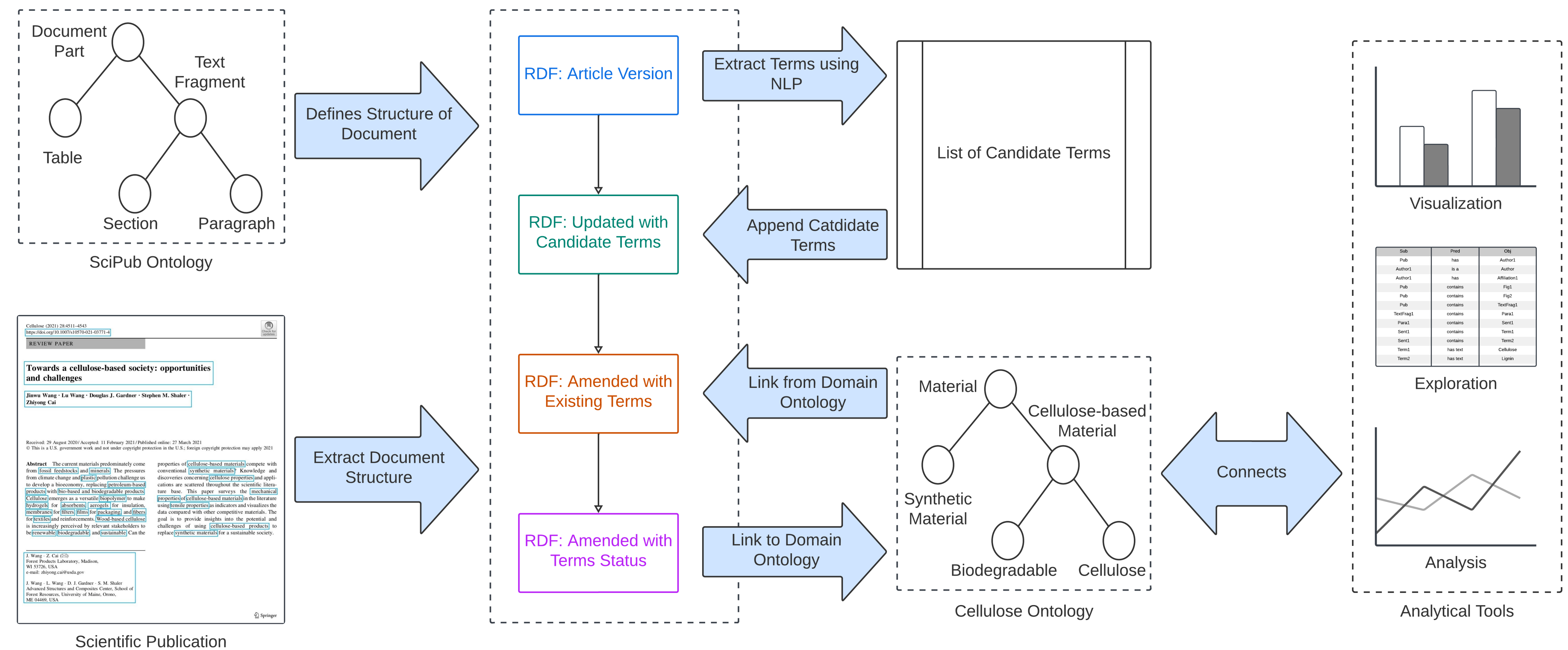
- How can cellulose-related terms that are identified by such NER approaches be categorized more effectively and precisely by leveraging a small hand-curated ontology of cellulose materials?

Evaluation Methods: Performance metrics (precision, recall, and f1-score), Comparative analysis, and Expert feedback

- What methods are suitable for determining whether a particular identified term refers to a concept that already exists in the ontology, or a new concept that requires amending the ontology?

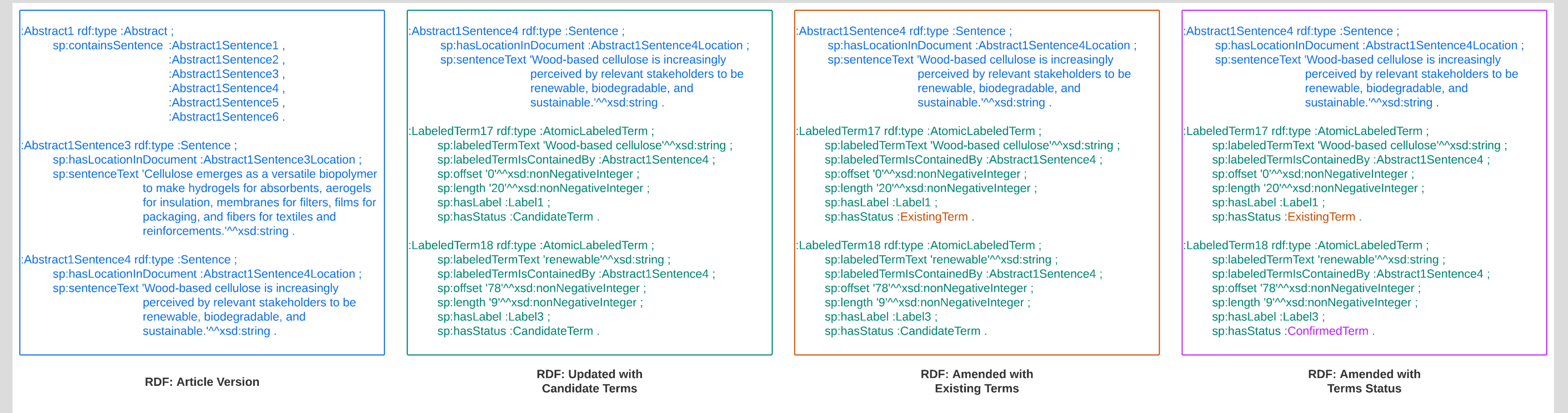
Evaluation Methods: Ontology consistency assessment, Semantic similarity matching, Term-Concept mapping, and Expert feedback

System Architecture



Data and Knowledge Workflow

- Produce RDF article version using SciPub ontology
- Extract list of terms using NLP algorithms
- Update RDF with extracted (candidate) terms
- Identify existing terms from domain ontology
- Amend RDF with existing terms
- Get the candidate terms status from domain expert
- Amend RDF with candidate terms status
- Refine ontology with confirmed candidate terms



Acknowledgement

- Thanks to Jinwu Wang, Douglas Gardner, and Stephen Shaler for fruitful discussions of this work.
- This research is supported in part by the U.S. Department of Agriculture: Forest Service, Project 20-JV-1111124-055, USDA Agricultural Research Service (ARS), Project 0204-41510-001-98S, National Institute of Food and Agriculture (NIFA), Award 2021-67022-34366