

Plan Overview

A Data Management Plan created using DMPTool

DMP ID: <https://doi.org/10.48321/D1TH43>

Title: Use of machine learning tools to accelerate the development of bioplastic films based on seaweed.

Creator: Davor Ibarra - **ORCID:** [0000-0002-1972-6085](https://orcid.org/0000-0002-1972-6085)

Affiliation: University of Santiago Chile (usach.cl)

Principal Investigator: Maria Jose Galotto

Data Manager: Davor Ibarra

Funder: University of Santiago Chile (usach.cl)

Template: Digital Curation Centre

Project abstract:

The study of biodegradable polymers, from biological and sustainable resources, is becoming increasingly popular at the scientific level. The governmental policies of the most developed countries recognize the need to develop new materials that meet both the technical requirements of their application and the reduction of the environmental impact generated after their use. In this sense, biopolymers based on algal polysaccharides are widely studied in the development of new packaging materials, proposing the inclusion of various components to improve barrier, mechanical, antioxidant and antimicrobial properties. In this study, we propose the development of a methodology to extract formulations and material properties from the scientific literature to accelerate the development of seaweed-based bioplastic films. This, by means of Artificial Intelligence techniques, such as models of Language Large Memory (LLM) for information detection in scientific articles to identify groups of components that improve certain properties in a sparse and highly dispersed data matrix.

Start date: 04-03-2023

End date: 11-30-2023

Last modified: 01-22-2024

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customize it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

Use of machine learning tools to accelerate the development of bioplastic films based on seaweed.

Concentrations of ingredients and their mechanical properties of bioplastic films made from seaweed are extracted from the scientific literature. The data will be presented in table format, specifically in JSON extension. Processing for fabrication bioplastics films, combinations of ingredients with their respective properties reported are expected to be extracted and published for open access and use.

The extraction of the ingredients and properties reported in the scientific literature will be carried out using the following methodology:

1. **Data obtaining.** Obtaining metadata (Title, Authors, Year, Journal, Abstract, DOI) of publications in WOS and SCOPUS search engines.
2. **Data selection.** Through the review of Titles and Abstracts, we selected the articles that correspond to bioplastic packaging films and that report the mechanical properties.
3. **Workflow Develop.** Development of an application for the extraction of information by means of intelligent artificial intelligence systems. A virtual assistant is created with capabilities to extract information from PDF files, this requires both the connection of computer services and the configuration of the technology for the specific case proposed.
4. **Data extraction.** Of the selected articles and application developed, specific content is extracted by means of a set of detailed instructions. From the section on bioplastic manufacturing methodology and results is reviewed for the IA assistance, leaving out those that do not exceed the exclusion criteria, such as manufacturing method, ingredients reported in concentration format and mechanical properties reported in table format.
5. **Data processing.** Within the assisted extraction workflow, it is possible to configure the wizard to standardize the units of the extracted variables.

The results will be presented in their final version, i.e. without change control and in GitHub repository: "davor-ibarra/AI_extract_PDF", where the more technical development of the project is detailed at length.

Respect to the general project aspect, in the repository, the main folder will contain a readme (.md) file with the explanation all relevant points for the project.

In specific, the data extracted have the following associated metadata:

"metadata":

```
{  
  "name_file": "Content Task 1",  
  "type_doc": "Content Task 1",  
  "title": "Content Task 1",  
  "authors": "Content Task 1",  
  "date_doc": "Content Task 1",  
  "doi": "Content Task 1":  
}
```

In addition, a final report (.pdf) file will be included in the root folder with basic details to help find the data, who created or contributed to its creation, its title, date of creation, conditions under which it can be accessed, methodology used for its creation, data processing and workflow used.

not applicable.

The results will be of free access and use under the license **Attribution 4.0 International (CC BY 4.0)**. For their use, their origin must be duly mentioned and their authorship acknowledged. It is forbidden to use the data for scientific publications that are not in collaboration or prior agreement with the authors of these.

The research data will be published in the Institutional Data Repository: <http://datos.usach.cl> which has sufficient space and secure storage and transfer protection protocols. A backup of the research data will be available through a retrieval archive on Google cloud storage services (accessed via personal institutional email) and a secondary physical backup storage on the laptops of the project researchers. The backup storage will be available for at least 5 years and in case it is required, the principal investigator of the project should be contacted.

In any case, all the project results will be published in GitHub repository: "davor-ibarra/AI_extract_PDF"

The platform developed during the project will be completely available to anyone in the world through the repository mentioned above, the nature of this repository allows to create a space for collaborative development through a version control managed entirely by the creator of the repository, but does not allow the safeguarding of information or technology, since the goal is that the information is completely open.

In the case of research data, these will be published in the internal repository of the Universidad de Santiago de Chile, where there are robust access and authentication policies for access to this information.

The resulting research data will be useful as long as the predictions made are validated, other more efficient models exist for this type of analysis, or the information is outdated. If it is observed that the data do not provide complementary value for the manufacture of algae-based bioplastic films, the storage and safeguarding of this information should be reconsidered.

Respect to the application, the advance of technology today is fast and dynamic, so it will be valuable as long as there is no more robust free software that meets the same proposed objectives.

not applicable.

The software developed will be published in GitHub repository: https://github.com/davor-ibarra/AI_extract_PDF

The research data will be published in the Institutional Data Repository: <http://datos.usach.cl>.

For its use, its origin must be duly mentioned and its authorship must be acknowledged. It is forbidden to use the data for scientific publications that do not have the collaboration or prior agreement of the authors.

This is not necessary.

Data collection and software developed, metadata production, data quality, and data plan management is the collaborating researcher Davor Ibarra.

The supervision and responsible for the execution of the data plan, provision of backup data, and contact for authorization of data use for scientific publications is Principal Researcher Maria Jose Galotto.

Not applicable.

Planned Research Outputs

Dataset - "extraction_biopolymers_data.xlsx"

Summary table with the extractions made from scientific articles on the processing and manufacture of bioplastic films based on marine algae, with their respective concentrations and reported mechanical properties.

Software - "AI_extract_PDF"

Artificial intelligence-based tool for the extraction of open access scientific knowledge. It aims to facilitate research and data analysis in emerging fields, contributing to environmental sustainability and the advancement of open science.

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
extraction_biopolymers_data.xlsx	Dataset	Unspecified	Open	GitHub DATAVERSE USACH		Creative Commons Attribution 4.0 International	None specified	No	No
AI_extract_PDF	Software	Unspecified	Open	GitHub		Creative Commons Attribution 4.0 International	None specified	No	No