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

Data Collection

What data will you collect or create?

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.

How will the data be collected or created?

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.

Documentation and Metadata

What documentation and metadata will accompany the data?

Respect to the general project aspect, in the repository, the main folder will contain a readme (.md) file with the exlanation all relevants 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.

Ethics and Legal Compliance

How will you manage any ethical issues?

not applicable.

How will you manage copyright and Intellectual Property Rights (IP/IPR) issues?

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.

Storage and Backup

How will the data be stored and backed up during the research?

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 publish in GitHub repository: "davor-ibarra/AI_extract_PDF"

How will you manage access and security?

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.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

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.

What is the long-term preservation plan for the dataset?

not applicable.

Data Sharing

How will you share the data?

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.

Are any restrictions on data sharing required?

This is not necessary.

Responsibilities and Resources

Who will be responsible for data management?

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.

What resources will you require to deliver your plan?

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