FAIR IMPLEMENTATION FOR NI4OS-EUROPE SERVICE PROVIDERS

29 April 2022

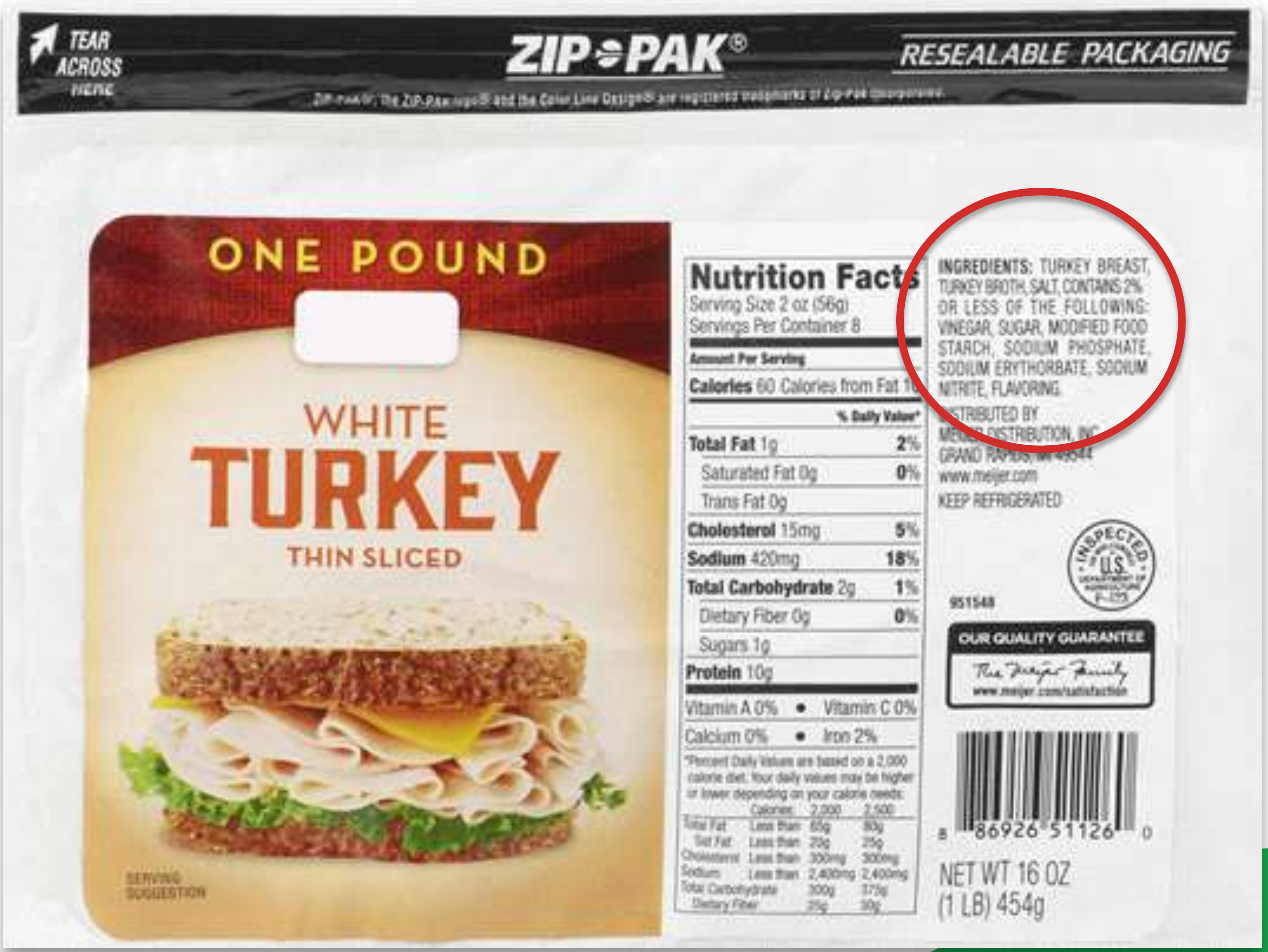
Dr. Zoe Cournia
Life Science Coordinator
Biomedical Research Foundation
Academy of Athens
Greece

Enhancing the food & cosmetics OpenAIRE Research Graph for consumer health

29 April 2022

Z Cournia, M Kounadis, A Chatzigoulas, D Papakonstantinou

Consumers don’t understand complex chemical labels


TEAR ACROSS HERE
ZIP-PAK®
RESEALABLE PACKAGING
ZIP-PAK®, The ZIP-Pak logo® and the Color Line Design® are registered trademarks of Zip-Pak Incorporated.
ONE POUND
WHITE
TURKEY
THIN SLICED
SERVING SUGGESTION
INSPECTED U.S. DEPARTMENT OF AGRICULTURE P-229
951548
OUR QUALITY GUARANTEE
The Meijer Family
www.meijer.com/satisfaction
8 86926 51126 0

Nutrition Facts
Serving Size 2 oz (56g)
Servings Per Container 8

Amount Per Serving	
Calories 60 Calories from Fat 10	
	% Daily Value*
Total Fat 1g	**2%**
Saturated Fat 0g	**0%**
Trans Fat 0g	
Cholesterol 15mg	**5%**
Sodium 420mg	**18%**
Total Carbohydrate 2g	**1%**
Dietary Fiber 0g	**0%**
Sugars 1g	
Protein 10g	

Vitamin A 0% • Vitamin C 0%
Calcium 0% • Iron 2%

*Percent Daily Values are based on a 2,000 calorie diet. Your daily values may be higher or lower depending on your calorie needs:

	Calories:	2,000	2,500
Total Fat	Less than	65g	80g
Sat Fat	Less than	20g	25g
Cholesterol	Less than	300mg	300mg
Sodium	Less than	2,400mg	2,400mg
Total Carbohydrate		300g	375g
Dietary Fiber		25g	30g

INGREDIENTS: TURKEY BREAST, TURKEY BROTH, SALT, CONTAINS 2% OR LESS OF THE FOLLOWING: VINEGAR, SUGAR, MODIFIED FOOD STARCH, SODIUM PHOSPHATE, SODIUM ERYTHORBATE, SODIUM NITRITE, FLAVORING.

DISTRIBUTED BY
MEIJER DISTRIBUTION, INC
GRAND RAPIDS, MI 49544
www.meijer.com
KEEP REFRIGERATED

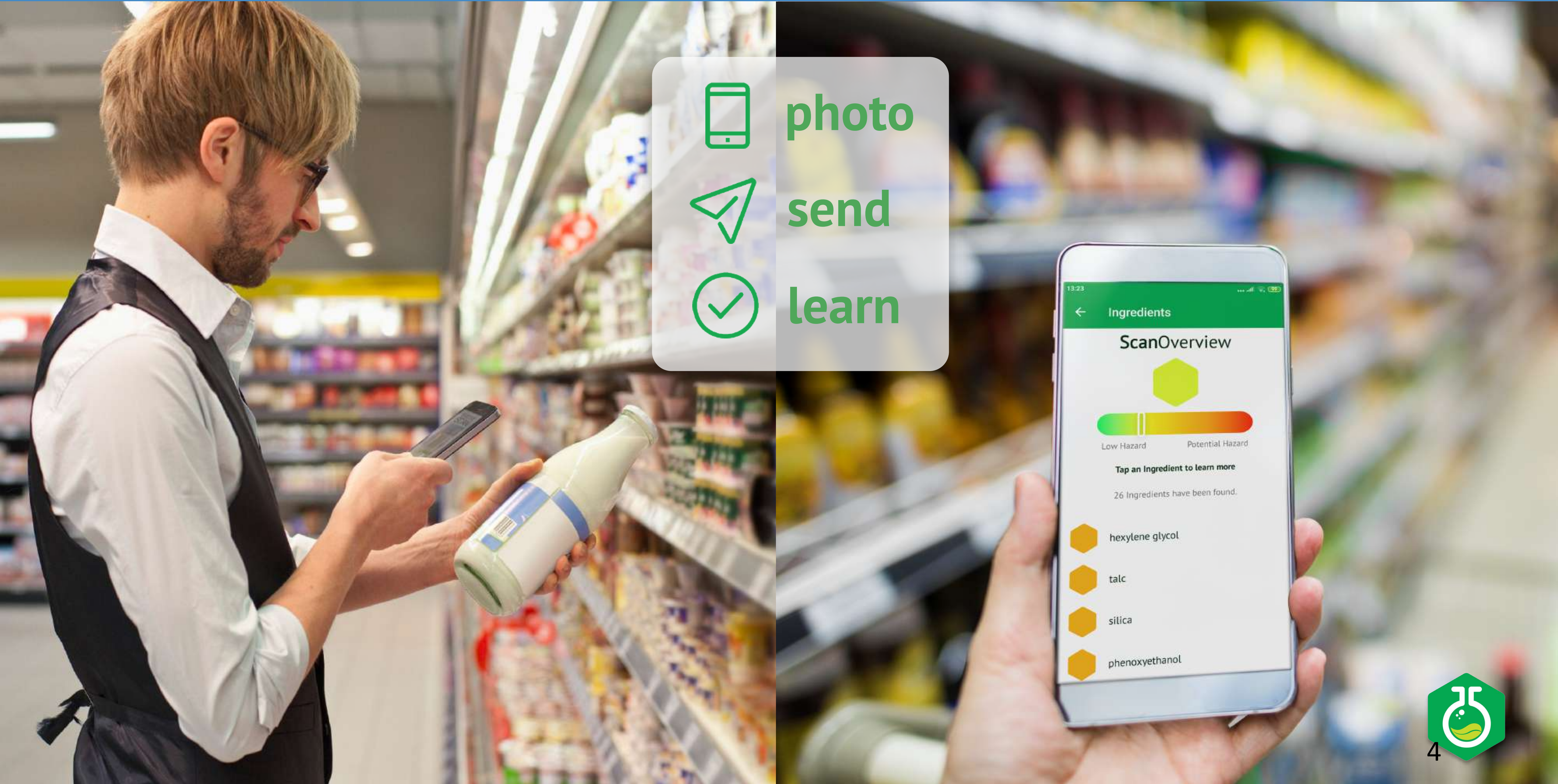
NET WT 16 OZ
(1 LB) 454g

World Health Organization has classified

Sodium Nitrite

found in processed meats as TYPE 1 CARCINOGEN

Ingredio is a tool that makes ingredients easy to understand



photo

send

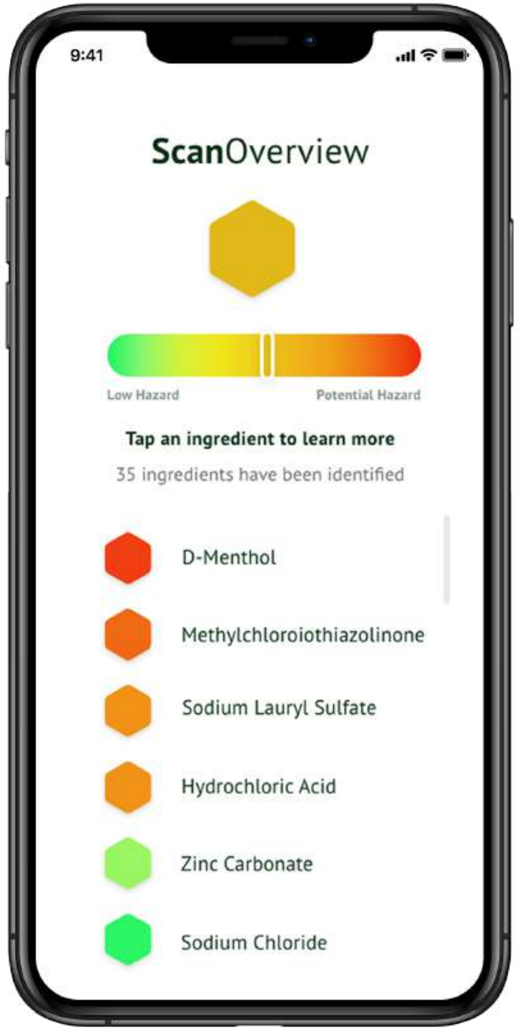
learn

Tap an Ingredient to learn more

26 Ingredients have been found.

Ingredio: Users can check if the product has ingredients with potential hazards for human health

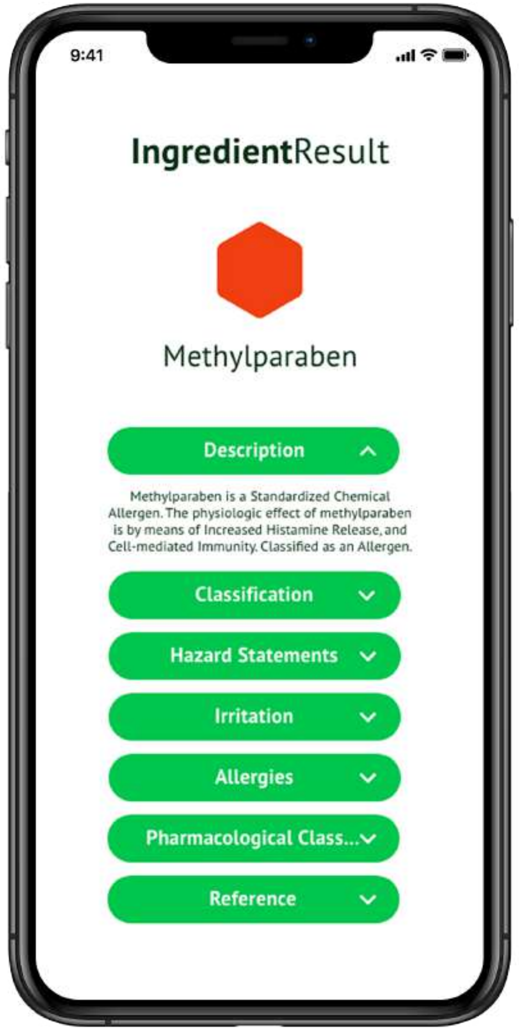
Unique scientific algorithms & scoring function



Tap an ingredient to learn more

35 ingredients have been identified

- D-Menthol
- Methylchloroiothiazolinone
- Sodium Lauryl Sulfate
- Hydrochloric Acid
- Zinc Carbonate
- Sodium Chloride



Methylparaben is a Standardized Chemical Allergen. The physiologic effect of methylparaben is by means of Increased Histamine Release, and Cell-mediated Immunity. Classified as an Allergen.

Level of hazard per ingredient*

European Commission

CosIng COSMETIC INGREDIENTS & SUBSTANCES

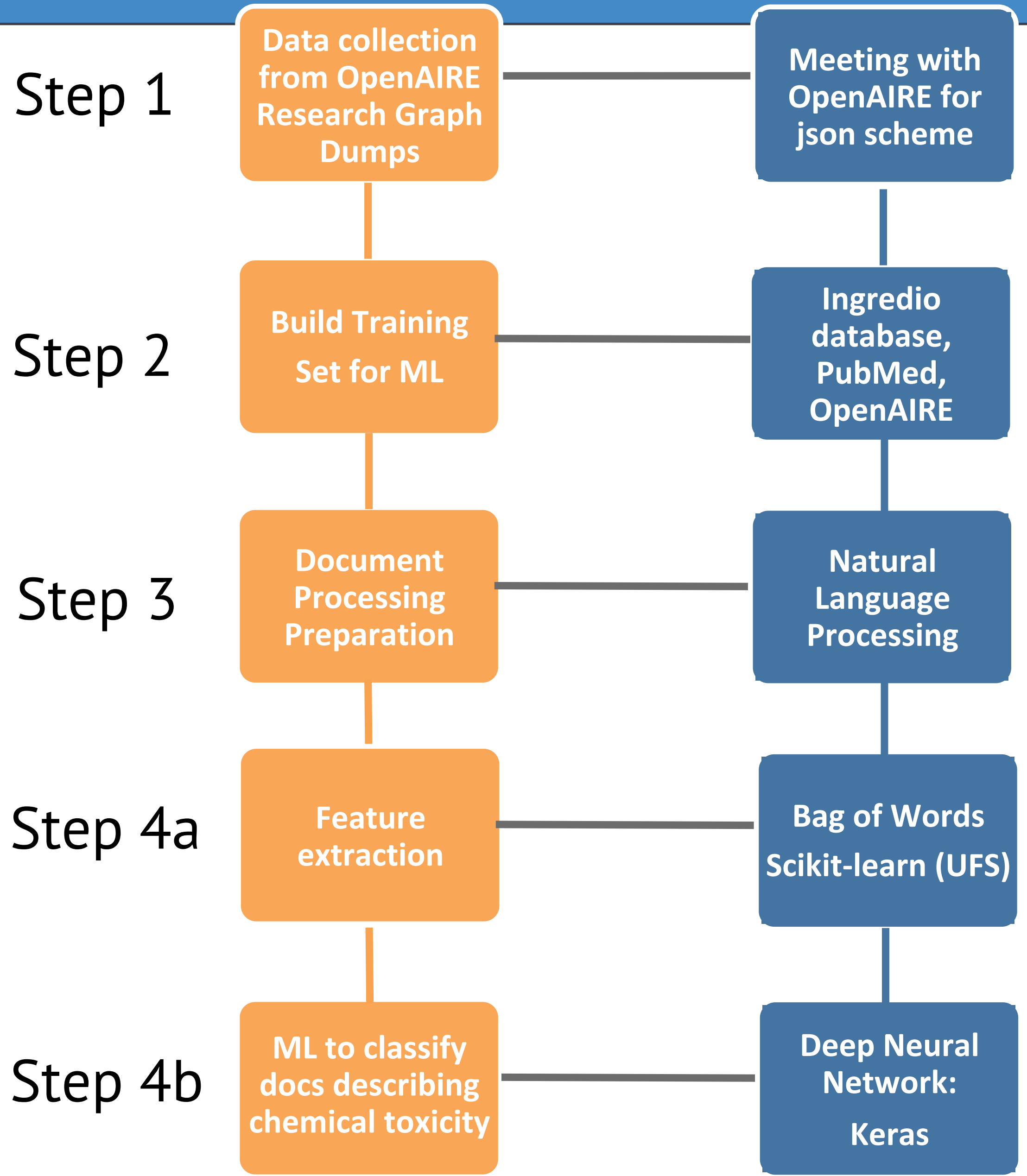
NATIONAL INSTITUTES OF HEALTH

PubChem

*** Information sourced from institutional databases**

OpenAIRE

Workflow



Step 1

Data collection from OpenAIRE Research Graph Dumps

Meeting with OpenAIRE for json scheme

Step 2

Build Training Set for ML

Ingredio database, PubMed, OpenAIRE

Step 3

Document Processing Preparation

Natural Language Processing

Step 4a

Feature extraction

Bag of Words Scikit-learn (UFS)

Step 4b

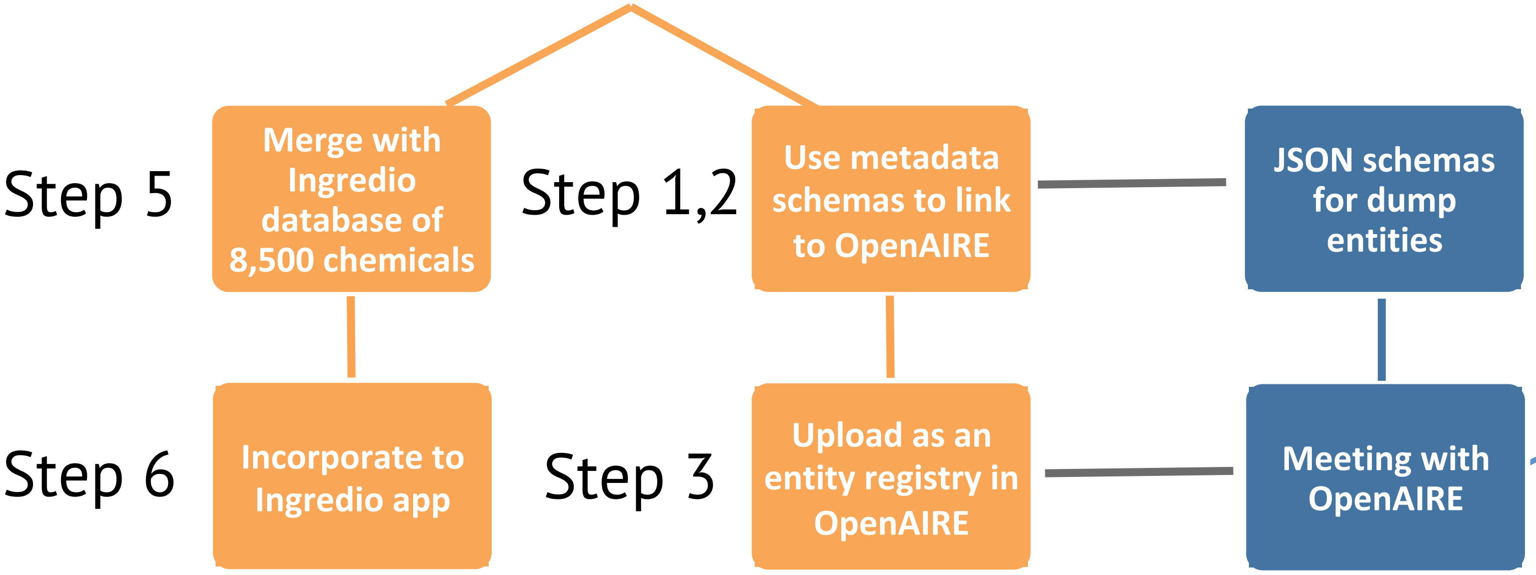
ML to classify docs describing chemical toxicity

Deep Neural Network: Keras

Feature 1

Develop text mining and Machine Learning algorithms to extract OpenAIRE data linking chemical ingredients of food &cosmetics to potential health hazards

Workflow



Step 5

Merge with Ingredio database of 8,500 chemicals

Step 1,2

Use metadata schemas to link to OpenAIRE

JSON schemas for dump entities

Step 6

Incorporate to Ingredio app

Step 3

Upload as an entity registry in OpenAIRE

Meeting with OpenAIRE

Feature 2

Support the curation of the data with appropriate metadata schemas for efficient integration of information in the OpenAIRE Research Graph

OpenAIRE Integration

Input/Output records:

Input (single JSON record in a single line):

```
{"id": "50|dedup_wf_001::00a9849ddb1168ffca2d599d316a7b19", "abstract": "Some abstract"}
{"id": "50|dedup_wf_001::03f197b3ba4270d7ca7d677f604f3e35", "abstract": "Different abstract"}
{"id": "50|dedup_wf_001::0262bfa18da0451870d7d9e62aea6ed1", "abstract": "Yet another abstract"}
```

and expected output (again, single JSON record in a single line):

```
{"id": "50|dedup_wf_001::00a9849ddb1168ffca2d599d316a7b19", "label": "some_mesh_class_label1", "confidenceLevel": 0.5}
{"id": "50|dedup_wf_001::03f197b3ba4270d7ca7d677f604f3e35", "label": "some_mesh_class_label1", "confidenceLevel": 0.9}
{"id": "50|dedup_wf_001::0262bfa18da0451870d7d9e62aea6ed1", "label": "some_mesh_class_label1", "confidenceLevel": 0.1}
```

Where confidenceLevel values are in the <0,1> range.

Execution scheme:

The script will be run relying on streaming approach (provided by oozie) so an inline equivalent of what is going to happen on each one of 32 data nodes would be just:

```
cat input.json | python classify.py
```

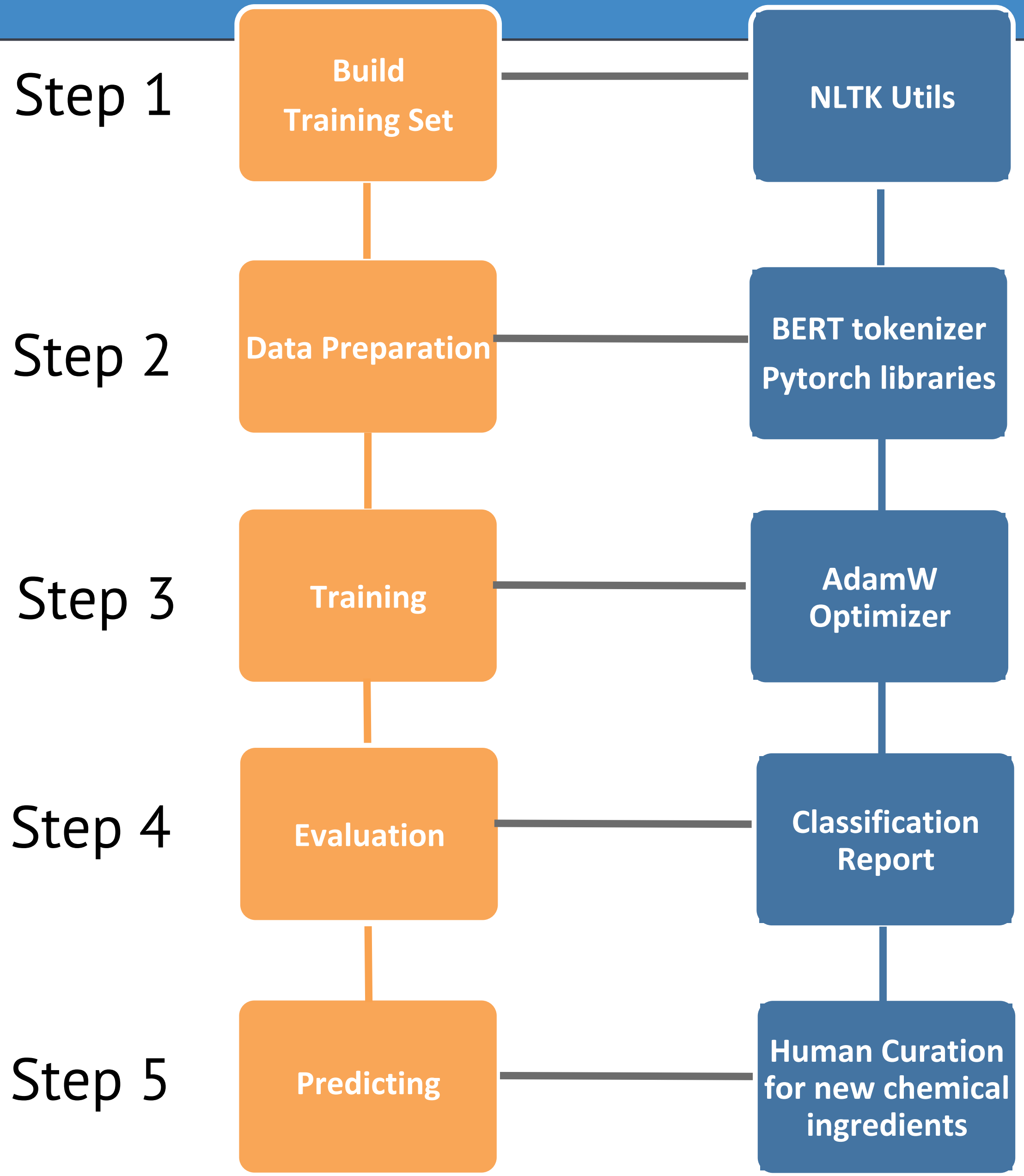
producing JSON records at stdout.

Mesh_label = "Toxic Actions"

Ensuring Interoperability with OpenAIRE

- ✓ Followed the OpenAIRE semantic layer with the OpenAIRE- CERIF XML format
- ✓ OpenAIRE semantic linkage and input from supervisor was used for all entities
- ✓ Constructed a dataframe with the articles containing useful information
- ✓ Transformed our dataset based on the “DataSource” core entity of the OpenAIRE data model
- ✓ We followed the OpenAIRE JSON schema for dumped entities
- ✓ 10 JSON files have been uploaded keeping the provenance of information (ready for d/l)

Workflow



Step 1	Build Training Set	NLTK Utils
Step 2	Data Preparation	BERT tokenizer Pytorch libraries
Step 3	Training	AdamW Optimizer
Step 4	Evaluation	Classification Report
Step 5	Predicting	Human Curation for new chemical ingredients

Feature 3

Identifying new chemical ingredients from the OpenAIRE data to enrich the OpenAIRE research graph & the Ingredio database

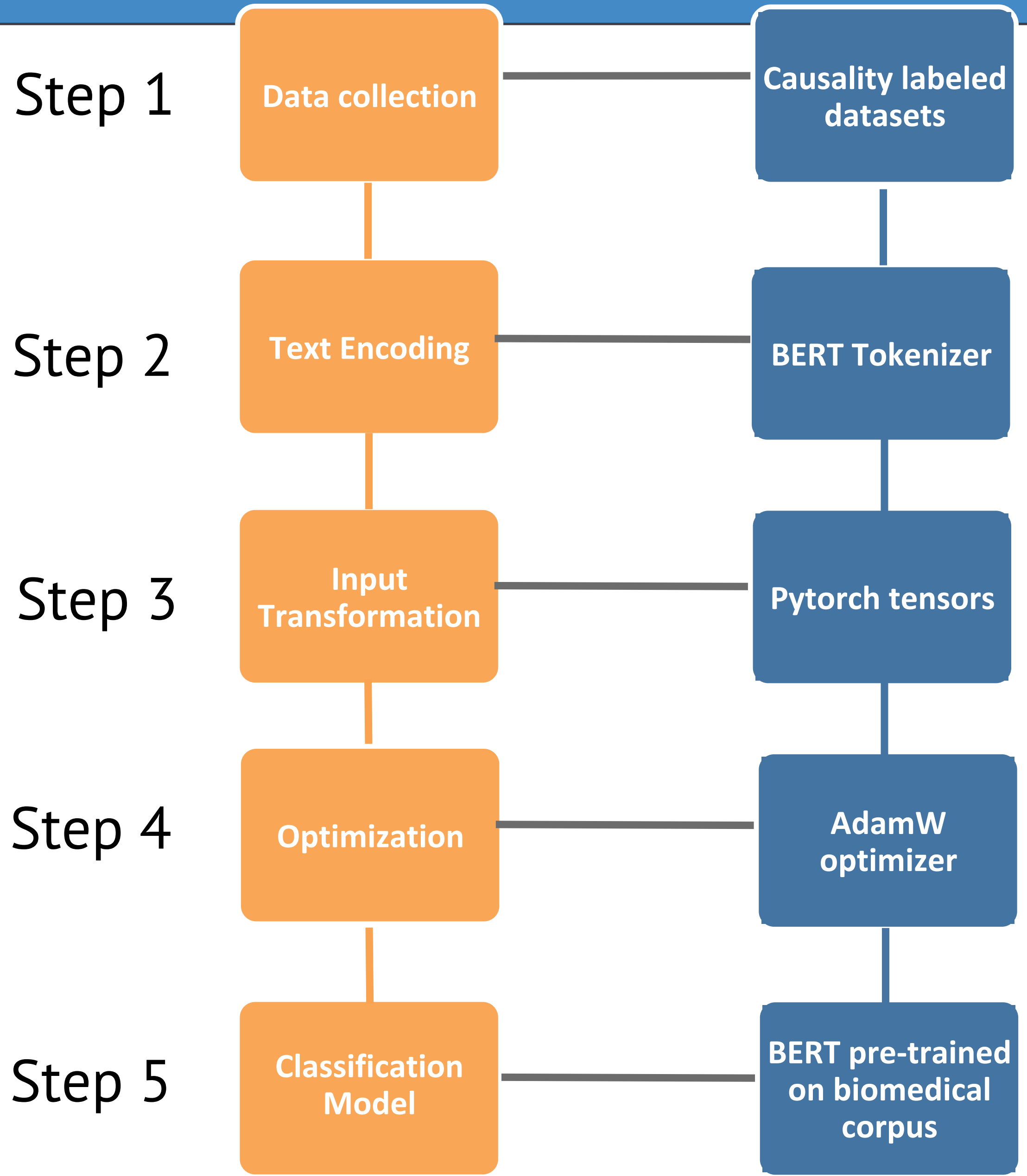
Example of identifying new chemical ingredients from biomedical text

Text:
Genetic disruption of thioredoxin reductase 1 protects against acetaminophen (APAP) toxicity. To determine the role of the thioredoxin system on xenobiotic metabolism we challenged wildtype and txnrd1liver-null mice with acetaminophen. Adult male wildtype and txnrd1 liver-null mice (C57BL6/J) were treated with either saline (PBS) or 100mg/kg APAP. Liver RNA was harvested eight hours after challenge and processed for microarray analysis. Comparison of 2 treatment conditions in 2 genotypes, biological replicates in triplicate.

Other examples of candidate compounds found from predicting articles classified in phase 2:

{'lumbar', 'propylene', 'anabolic', 'nicotine', 'xylan', 'thalidomide', 'acetaminophen', 'carvone'…}

Ingredio Main Features



Step 1 — Data collection — Causality labeled datasets

Step 2 — Text Encoding — BERT Tokenizer

Step 3 — Input Transformation — Pytorch tensors

Step 4 — Optimization — AdamW optimizer

Step 5 — Classification Model — BERT pre-trained on biomedical corpus

Feature 4

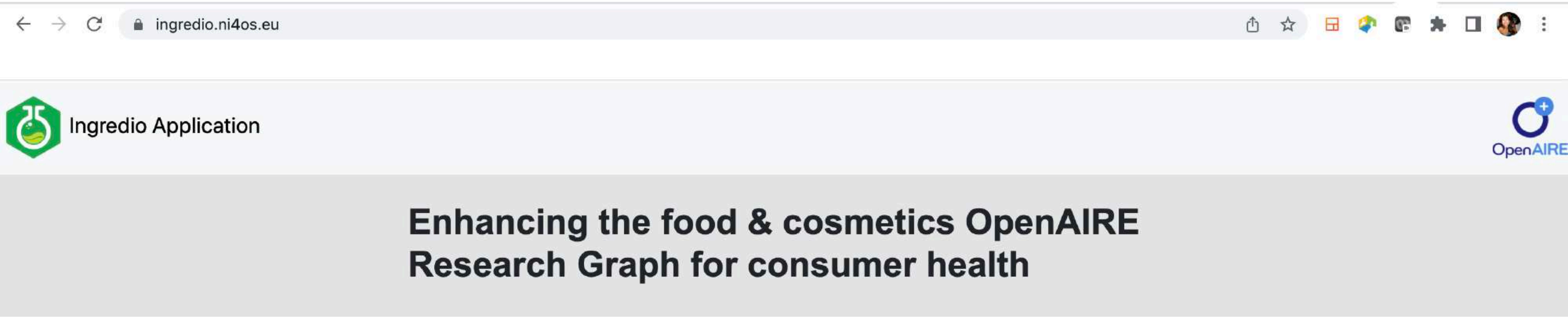
Understanding the relationship of chemical ingredients with hazards using the provided information –
Building causal relationships

Example of finding causal relationships

	Precision	Recall	F1-score
Non-causal	0.90	0.88	0.89
Causal	0.91	0.92	0.91
Accuracy			0.90
Macro avg	0.90	0.90	0.90
Weighted avg	0.90	0.90	0.90

[PubChemID: 6579, Adverse Effect: 'neurotoxicity’, Sentence: '*acrylamide (acr) is known to induce neurotoxicity in humans and occupational exposure to acr has an effect on human health.*’, Probability: 0.99]

Onboarding Ingredio to EOSC



ingredio.ni4os.eu

Ingredio Application

OpenAIRE

Enhancing the food & cosmetics OpenAIRE Research Graph for consumer health

- ✓ A dedicated server was provided by NI4OS-Europe (BAS - Bulgaria) with one NVIDIA V100 on 24.1.2021
- ✓ A web-server was developed and uploaded in https://ingredio.ni4os.eu/
- ✓ Web-server has been onboarded in NI4OS

Ingredio Main Features

Feature 1
Develop text mining and Machine Learning algorithms to extract OpenAIRE data linking chemical ingredients of food &cosmetics to potential health hazards

Feature 2
Support the curation of OpenAIRE data with appropriate metadata schemas for efficient integration of information in the OpenAIRE Research Graph

Feature 3
Identifying new chemical ingredients from OpenAIRE data to enrich the OpenAIRE research graph & the Ingredio database

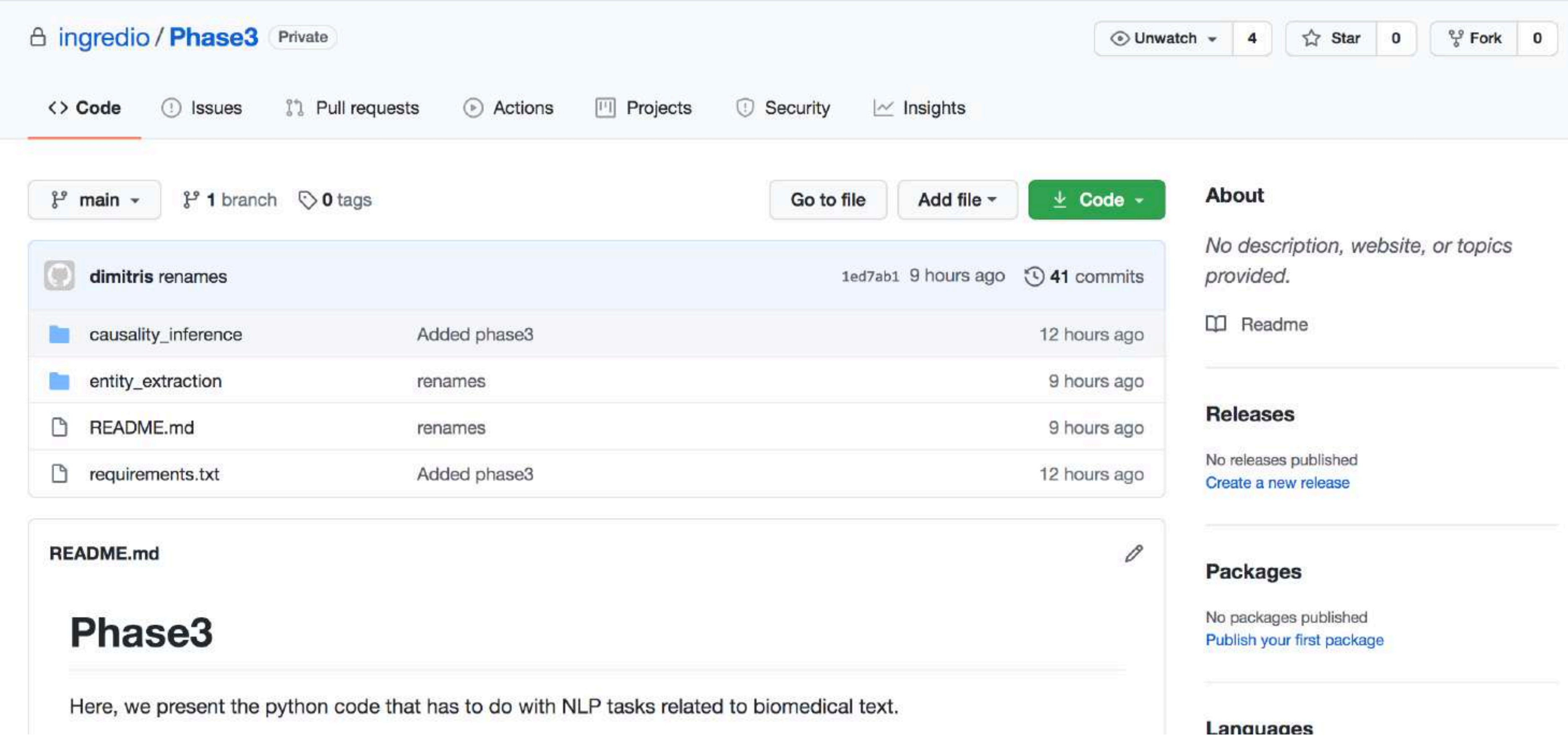
Feature 4
Identify causal relationships of chemical ingredients with hazards using the collected information

Ingredio FAIR practices: Findability and Accessibility

```
Phase3
 |── causality_inference
 |── dataset
 |── input
 |    |    └── bert_base_uncased
 |    |         ├── config.json
 |    |         └── vocab.txt
 |    └── src
 |         ├── config.py
 |         ├── dataset.py
 |         ├── engine.py
 |         ├── model.py
 |         ├── predict.py
 |         └── train.py
 |── entity_extraction
 |    ├── datasets
 |    └── src
 |         ├── config_task1.py
 |         ├── dataset_task1.py
 |         ├── find_compounds_task1.py
 |         ├── model_task1.py
 |         ├── predict_task1.py
 |         └── train_task1.py
 |── README.md
 └── requirements.txt

8 directories, 16 files
```

Ingredio code and data are publicly available at
https://ingredio.ni4os.eu/
https://github.com/ingredio/NLP_utils



ingredio / **Phase3** Private

Unwatch 4 | Star 0 | Fork 0

Code | Issues | Pull requests | Actions | Projects | Security | Insights

main | 1 branch | 0 tags | Go to file | Add file | Code

dimitris renames — 1ed7ab1 9 hours ago — 41 commits

causality_inference	Added phase3	12 hours ago
entity_extraction	renames	9 hours ago
README.md	renames	9 hours ago
requirements.txt	Added phase3	12 hours ago

README.md

Phase3

Here, we present the python code that has to do with NLP tasks related to biomedical text.

About

No description, website, or topics provided.

Readme

Releases

No releases published
Create a new release

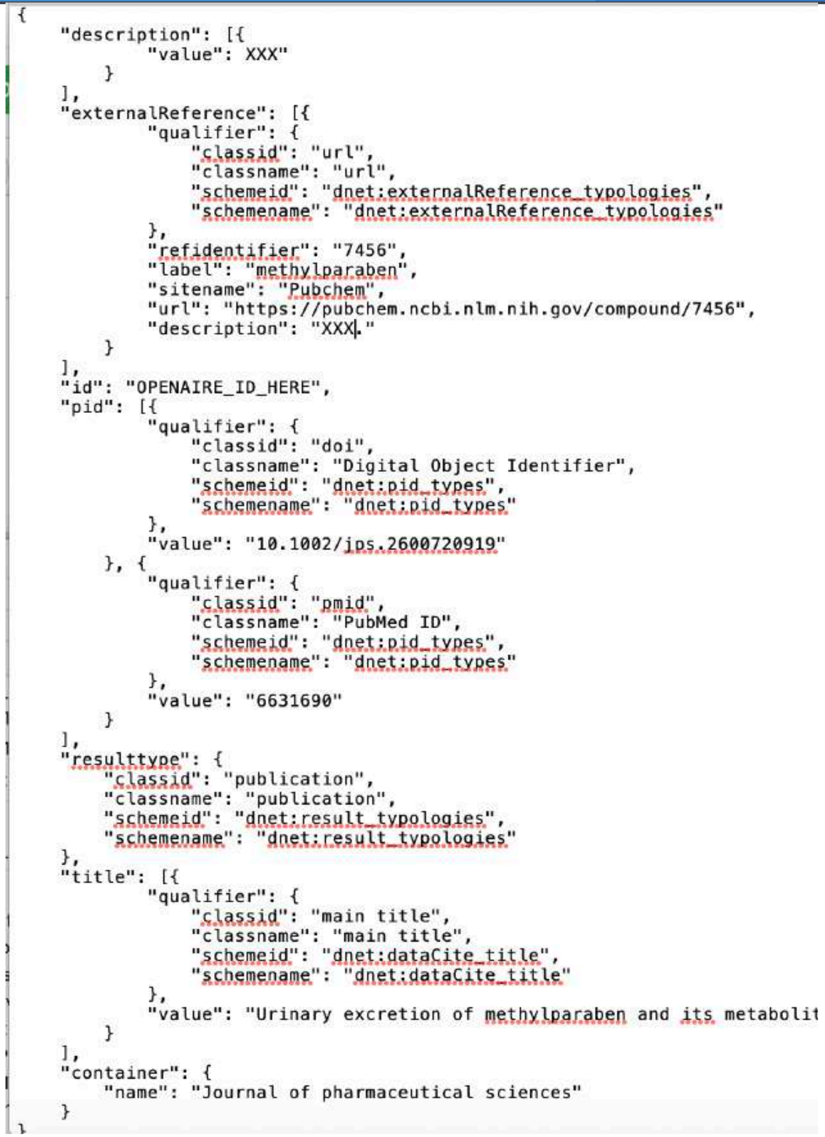
Packages

No packages published
Publish your first package

Languages

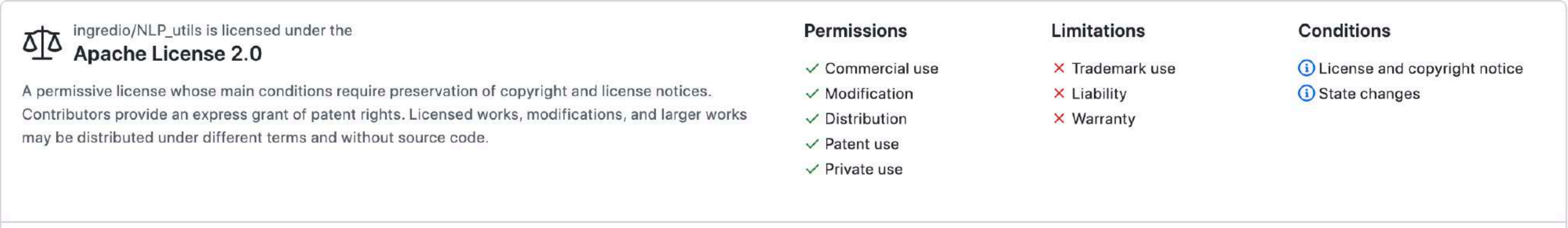
Ingredio FAIR practices - Interoperability

- ❑ Metadata schemas used in Ingredio follows OpenAIRE schemas:
- ❑ Exported and appended to a file each publication in JSON format
- ❑ Mapping directly as an OpenAIRE internal Oaf model, specifically for publications
- ❑ Outputs classified documents in the following schema:
- ❑ {"id": "some form of id", "label": "relevant", "confidenceLevel": 0.52}
- ❑ Vocabulary used: Medical Subject Headings (MeSH)



```
{
    "description": [{
            "value": XXX"
        }
    ],
    "externalReference": [{
            "qualifier": {
                "classid": "url",
                "classname": "url",
                "schemeid": "dnet:externalReference_typologies",
                "schemename": "dnet:externalReference_typologies"
            },
            "refidentifier": "7456",
            "label": "methylparaben",
            "sitename": "Pubchem",
            "url": "https://pubchem.ncbi.nlm.nih.gov/compound/7456",
            "description": "XXX."
        }
    ],
    "id": "OPENAIRE_ID_HERE",
    "pid": [{
            "qualifier": {
                "classid": "doi",
                "classname": "Digital Object Identifier",
                "schemeid": "dnet:pid_types",
                "schemename": "dnet:pid_types"
            },
            "value": "10.1002/jps.2600720919"
        }, {
            "qualifier": {
                "classid": "pmid",
                "classname": "PubMed ID",
                "schemeid": "dnet:pid_types",
                "schemename": "dnet:pid_types"
            },
            "value": "6631690"
        }
    ],
    "resulttype": {
        "classid": "publication",
        "classname": "publication",
        "schemeid": "dnet:result_typologies",
        "schemename": "dnet:result_typologies"
    },
    "title": [{
            "qualifier": {
                "classid": "main title",
                "classname": "main title",
                "schemeid": "dnet:dataCite_title",
                "schemename": "dnet:dataCite_title"
            },
            "value": "Urinary excretion of methylparaben and its metabolit
        }
    ],
    "container": {
        "name": "Journal of pharmaceutical sciences"
    }
}
```

Ingredio FAIR practices – Reusability: License



ingredio/NLP_utils is licensed under the

Apache License 2.0

A permissive license whose main conditions require preservation of copyright and license notices. Contributors provide an express grant of patent rights. Licensed works, modifications, and larger works may be distributed under different terms and without source code.

Permissions

- ✓ Commercial use
- ✓ Modification
- ✓ Distribution
- ✓ Patent use
- ✓ Private use

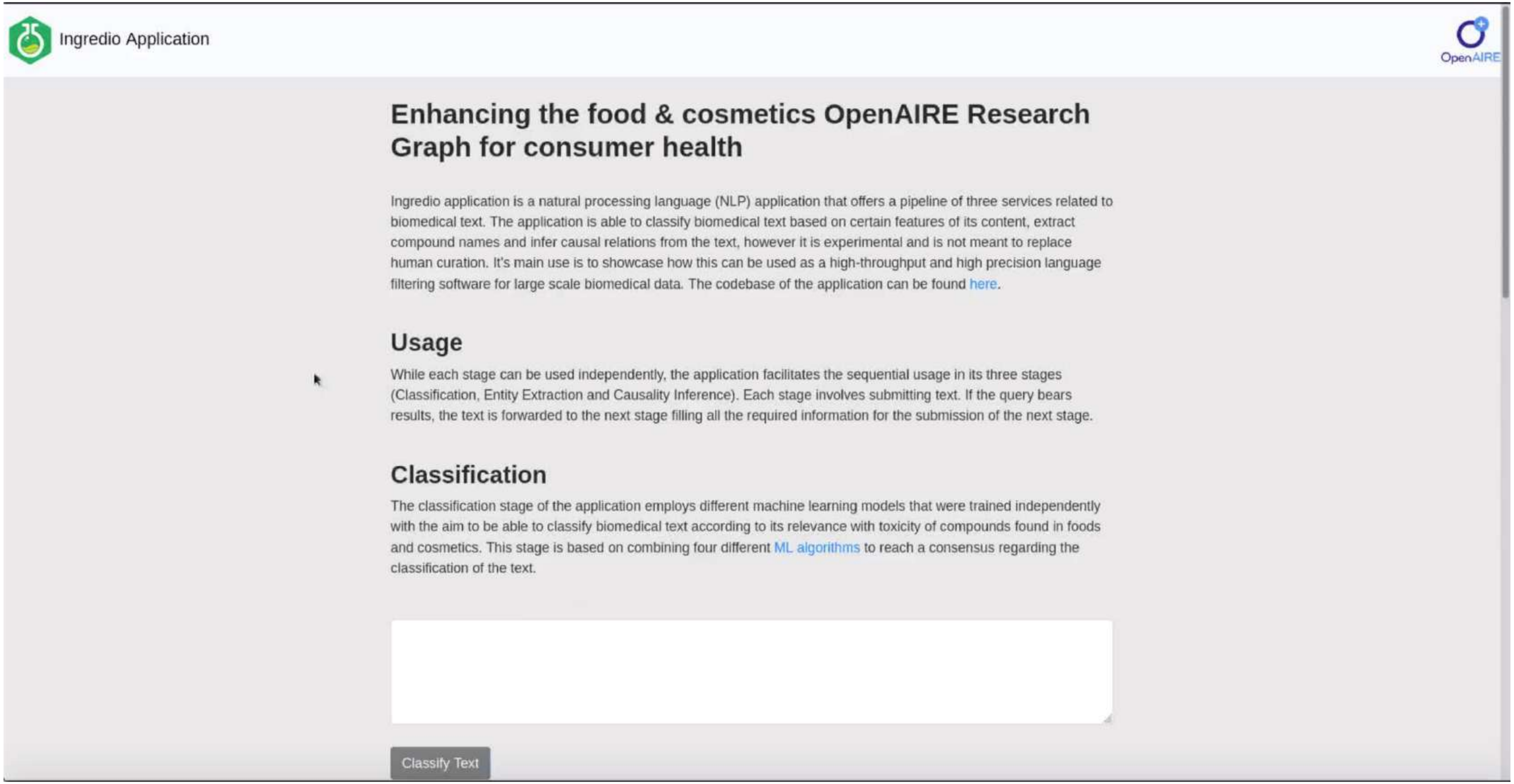
Limitations

- × Trademark use
- × Liability
- × Warranty

Conditions

- ⓘ License and copyright notice
- ⓘ State changes

Ingredio FAIR practices – Reusability: Video Example



Enhancing the food & cosmetics OpenAIRE Research Graph for consumer health

Ingredio application is a natural processing language (NLP) application that offers a pipeline of three services related to biomedical text. The application is able to classify biomedical text based on certain features of its content, extract compound names and infer causal relations from the text, however it is experimental and is not meant to replace human curation. It's main use is to showcase how this can be used as a high-throughput and high precision language filtering software for large scale biomedical data. The codebase of the application can be found here.

Usage

While each stage can be used independently, the application facilitates the sequential usage in its three stages (Classification, Entity Extraction and Causality Inference). Each stage involves submitting text. If the query bears results, the text is forwarded to the next stage filling all the required information for the submission of the next stage.

Classification

The classification stage of the application employs different machine learning models that were trained independently with the aim to be able to classify biomedical text according to its relevance with toxicity of compounds found in foods and cosmetics. This stage is based on combining four different ML algorithms to reach a consensus regarding the classification of the text.

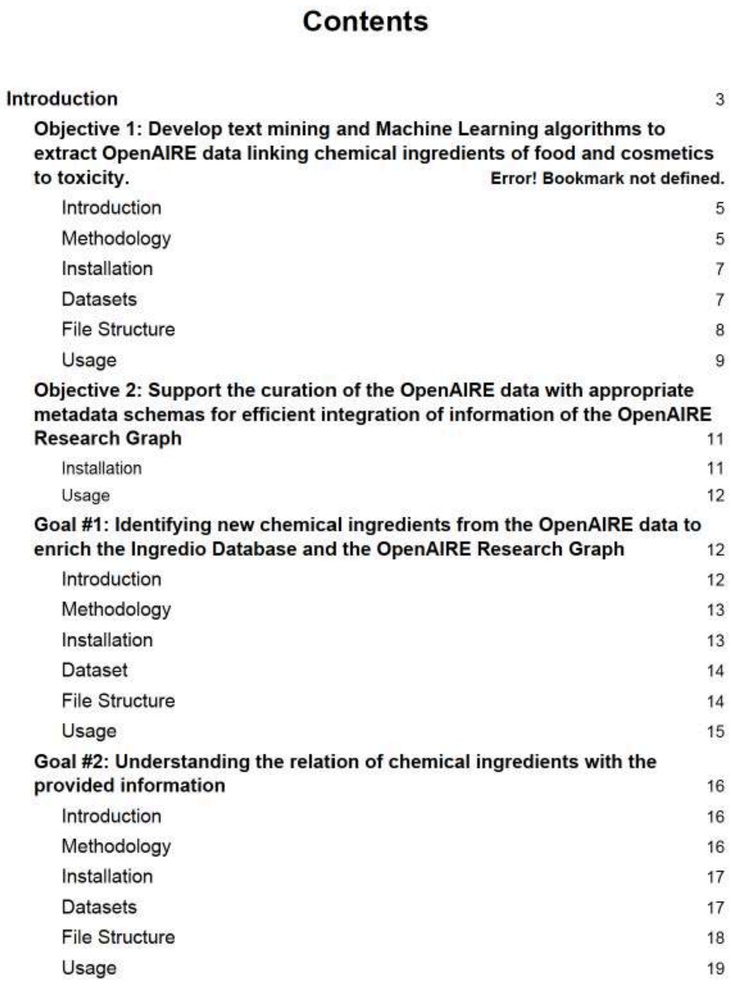
Ingredio FAIR practices – Reusability: User Manual

Enhancing the food & cosmetics OpenAIRE Research Graph for consumer health

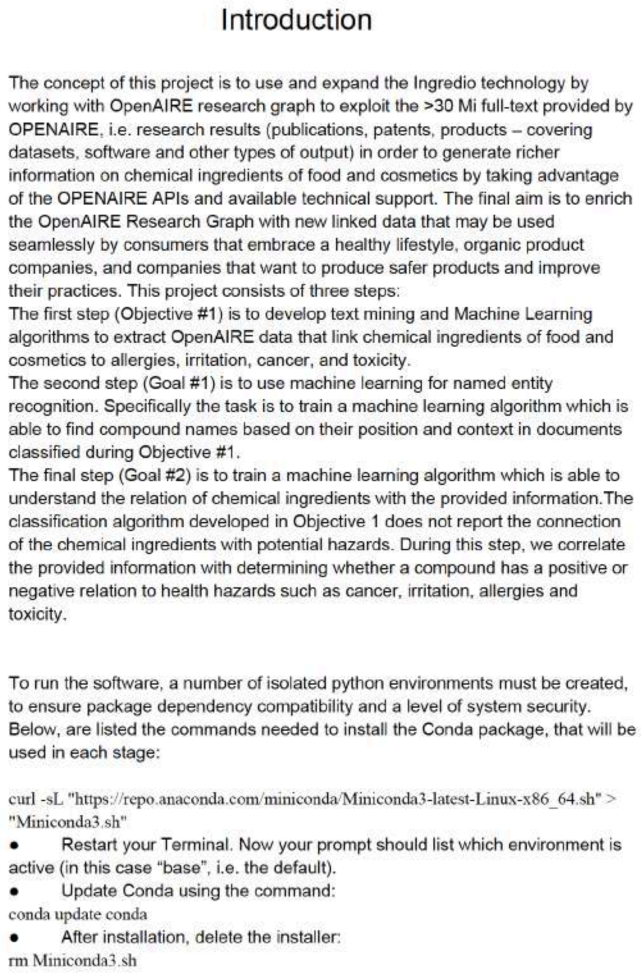
USER MANUAL

Z Cournia, M Kounadis, A Chatzigoulas, D Papakonstantinou

Powered by DENTICA LTD



Contents



Introduction

The concept of this project is to use and expand the Ingredio technology by working with OpenAIRE research graph to exploit the >30 Mi full-text provided by OPENAIRE, i.e. research results (publications, patents, products – covering datasets, software and other types of output) in order to generate richer information on chemical ingredients of food and cosmetics by taking advantage of the OPENAIRE APIs and available technical support. The final aim is to enrich the OpenAIRE Research Graph with new linked data that may be used seamlessly by consumers that embrace a healthy lifestyle, organic product companies, and companies that want to produce safer products and improve their practices. This project consists of three steps:

The first step (Objective #1) is to develop text mining and Machine Learning algorithms to extract OpenAIRE data that link chemical ingredients of food and cosmetics to allergies, irritation, cancer, and toxicity.

The second step (Goal #1) is to use machine learning for named entity recognition. Specifically the task is to train a machine learning algorithm which is able to find compound names based on their position and context in documents classified during Objective #1.

The final step (Goal #2) is to train a machine learning algorithm which is able to understand the relation of chemical ingredients with the provided information.The classification algorithm developed in Objective 1 does not report the connection of the chemical ingredients with potential hazards. During this step, we correlate the provided information with determining whether a compound has a positive or negative relation to health hazards such as cancer, irritation, allergies and toxicity.

To run the software, a number of isolated python environments must be created, to ensure package dependency compatibility and a level of system security. Below, are listed the commands needed to install the Conda package, that will be used in each stage:

```
curl -sL "https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh" > "Miniconda3.sh"
```

- Restart your Terminal. Now your prompt should list which environment is active (in this case “base”, i.e. the default).
- Update Conda using the command:

```
conda update conda
```

- After installation, delete the installer:

```
rm Miniconda3.sh
```

Ingredio FAIR practices – Reusability: Training Material

- Information in NI4OS-Europe Agora: https://catalogue.ni4os.eu/?_=/resources/7cc4118c-e637-4463-a841-92831e897368

- Access: https://ingredio.ni4os.eu/

- Training Material: https://training.ni4os.eu/mod/scorm/view.php?id=1182

Acknowledgements

Team Ingredio

M Kounadis
A Chatzigoulas
D Papakonstantinou
D Trovas

Team OpenAIRE

H Dimitropoulos
M Horst
Y Foufoulas
A Manocci
A Bardi
N Berikou

Team NI4OS-Europe

D Vudragovic
A Mishev
S Spasov
E Atanassov
M Durchova
K Koumantaros

Thanks for your attention!