Artificial Intelligence And Open Science In Research Software Engineering

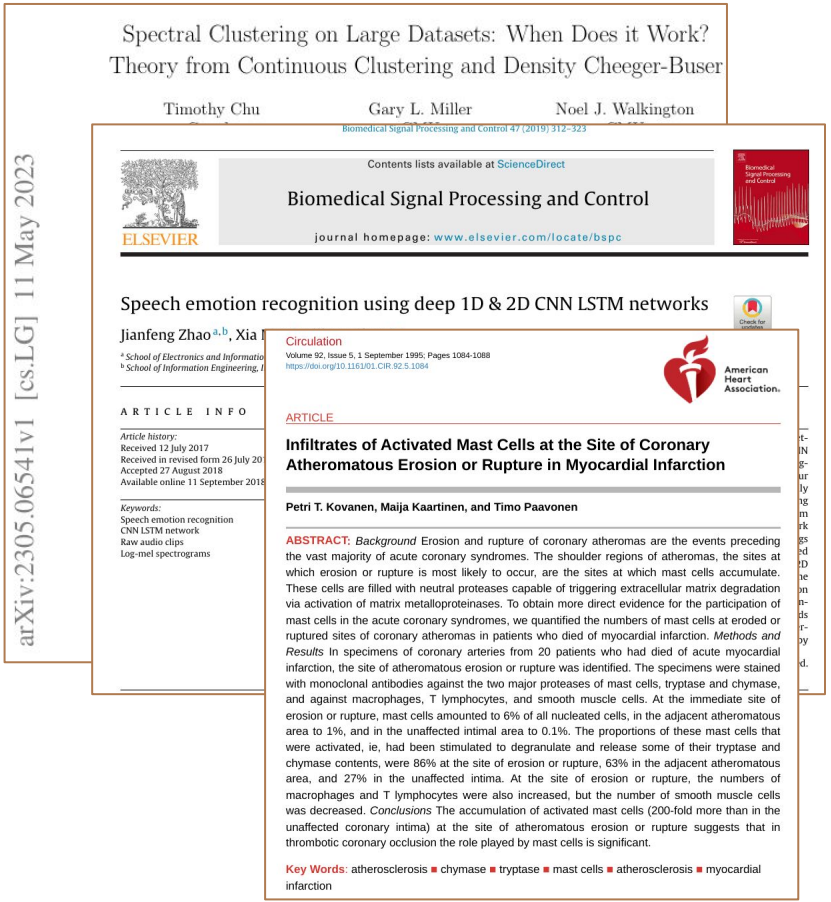
Deliverable 2

Sebastian Bayona
Alejandro Ayuso
Daniel Cabrera
Alejandro Morán

https://github.com/dan415/openscience-grupo

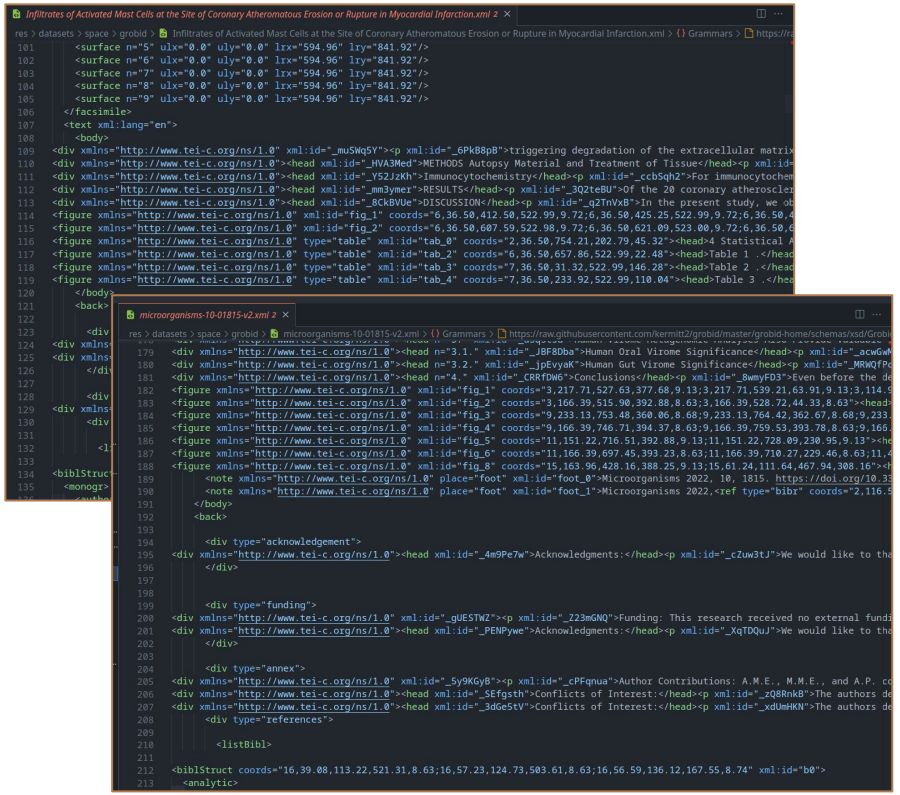
Papers selection

There are papers from 2 distinct research fields, and the main reason for selecting them is their complete metadata: Acknowledgments, authors, affiliations, etc.

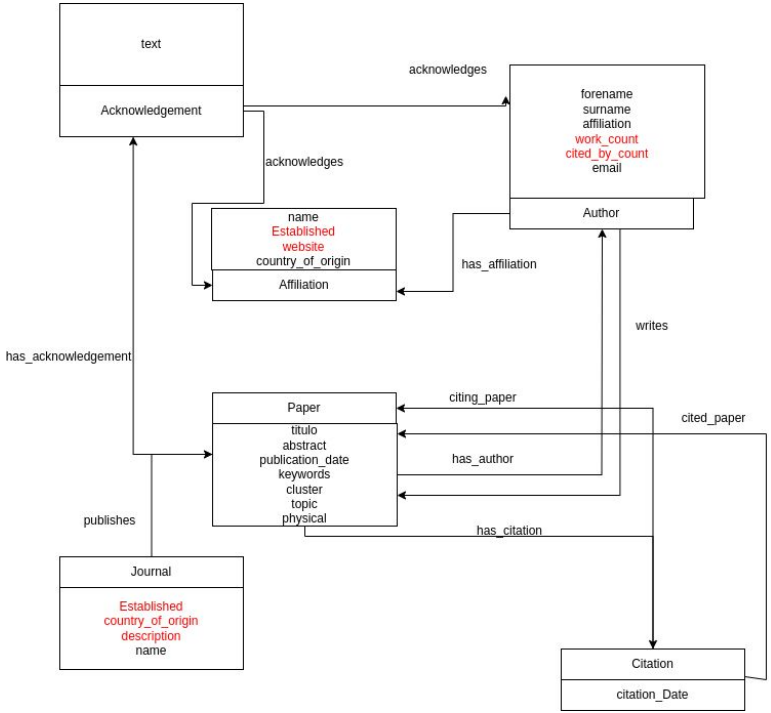


Leveraging GROBID for Knowledge Graph Construction: Extracting Relevant Information from Selected Papers

In the realm of technical and scientific publications, extracting structured information from raw documents like PDFs is crucial for various research and knowledge management tasks. GROBID, a powerful machine learning library, enables the extraction, parsing, and re-structuring of PDFs into structured XML/TEI encoded documents. This presentation focuses on utilizing GROBID to extract relevant information from selected papers and leveraging this data to construct a knowledge graph. By employing GROBID's functionalities.



Ontology


text
Acknowledgement
acknowledges
forename
surname
affiliation
work_count
cited_by_count
email
Author
acknowledges
name
Established
website
country_of_origin
Affiliation
has_affiliation
writes
has_acknowledgement
Paper
titulo
abstract
publication_date
keywords
cluster
topic
physical
citing_paper
cited_paper
has_author
publishes
has_citation
Journal
Established
country_of_origin
description
name
Citation
citation_Date

Wikidata → Affiliation (Established, Website)

Wikidata → Journal (Established, Country_of_origin, Description)

OpenAlex → Author (Cited_by_count, Works_count)

Research Knowledge Graph

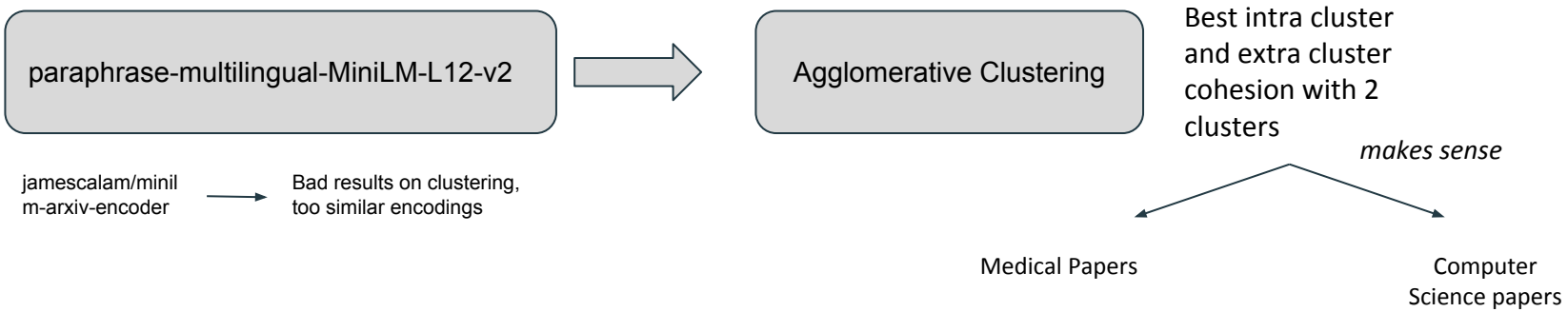
1. Get classes and properties from the xml already processed
2. Paper Space for a more useful representation
3. Obtain Knowledge Graph from Paper ~~Space~~ → JSON-LD

XML processed by Grobid

Representation to operate with

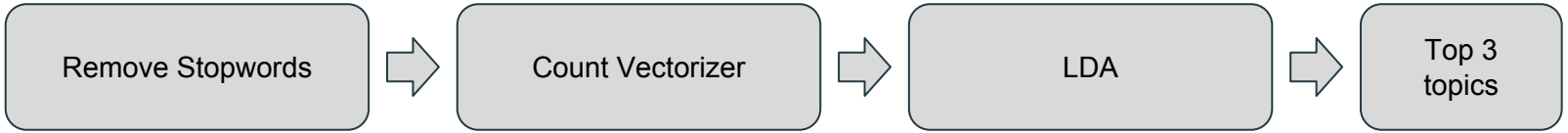
RDFParser -> JSON-LD

Clustering



"Predicted Cluster added as property"

Topic Modeling



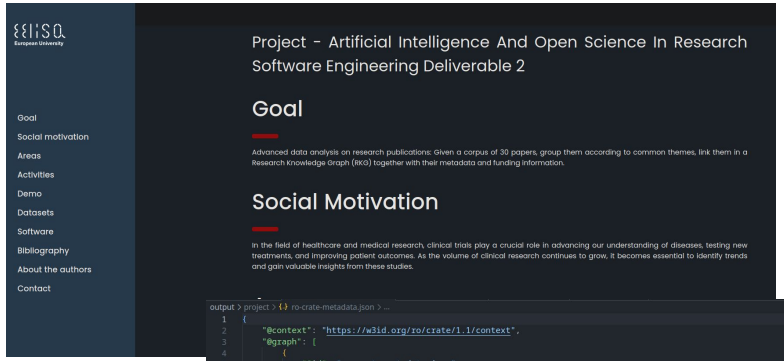
“Topics added as property”

Entity Recognition

Experiment as a Research Object: using ya2ro

project.yaml > [ ] participants > {} 2 > web

```yaml
# Mandatory field
type: "project"

title: "Project - Artificial Intelligence And Open Science In Research Software Engineering Deliverable 2"

goal: "Advanced data analysis on research publications: Given a corpus of 30 papers, group them according to commor

social_motivation: "In the field of healthcare and medical research, clinical trials play a crucial role in advanci

#sketch: "images/sketch_ya2ro.jpg"

areas:
  - "Area 1: Use information on clinical trials associated with the publications to identify trends in clinical res

activities:
  - "Subtask 1: Papers selection. There are papers from 2 distinct research fields, and the main reason for selecti
  - "Subtask 2: Leveraging GROBID for Knowledge Graph Construction: Extracting Relevant Information from Selected P
  - "Subtask 3: Ontology. Diagram of the ontology that was defined for the project"
  - "Subtask 4: Transform the classes and properties from the xml already processed (by GROBID) to the \"Paper Spac
  - "Subtask 5: Enriching with wikidata (affiliation, website and Journal) and OpenAlex (author's count of citation
  - "Subtack 6: Obtain Knowledge Graph from Paper Space -> JSON-LD"
  - "Subtask 7: Clustering"
  - "Subtask 8: Topic Modeling"
  - "Subtask 9: Entity Recognition"

demo:
  - name: "Git Repository"
    link: https://github.com/dan415/openscience-grupo
    description: "Find the docs and demos in the github repository."

datasets:
  - link: https://drive.google.com/drive/folders/1ACPAXCJiZzueSEnA3p1kjZoMWzqGY4qG?usp=share_link
    name: "Dataset 1"
    description: "List of selected papers"

software:
  - link: https://github.com/dan415/openscience-grupo
    name: "openscience-grupo"
    description: "Main code repository"
    license: "MIT-License"
```



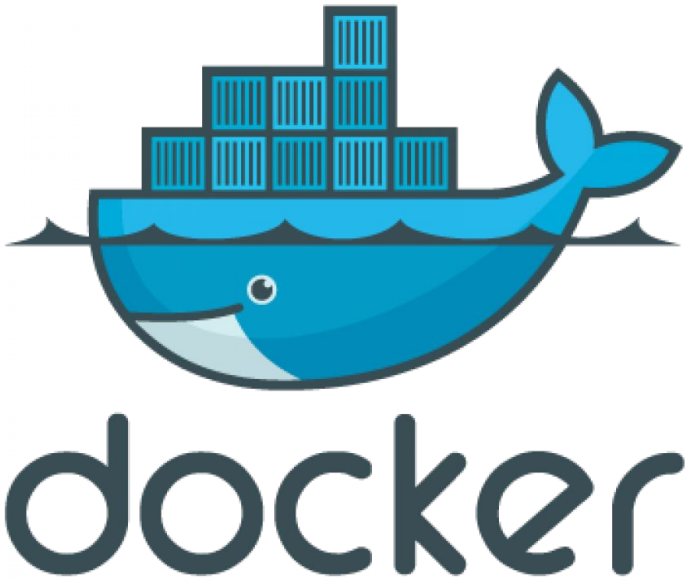
output > project > {} ro-crate-metadata.json > ...

```json
{
    "@context": "https://w3id.org/ro/crate/1.1/context",
    "@graph": [
        {
            "@id": "ro-crate-metadata.json",
            "@type": "CreativeWork",
            "about": {
                "@id": "./"
            },
            "conformsTo": {
                "@id": "https://w3id.org/ro/crate/1.1"
            },
            "description": "RO-Crate Metadata File Descriptor (this file)"
        },
        {
            "@id": "./",
            "@type": "Dataset",
            "DataType": "Project",
            "author": [
                {
                    "@id": "#daniel_cabrera"
                },
                {
                    "@id": "#alejandro_ayuso"
                },
                {
                    "@id": "#alejandro_mor\u00e1n"
                },
                {
                    "@id": "#sebastian_bayona"
                }
            ],
            "description": "Advanced data analysis on research publications: Given a corpus of 30 papers, group t
            "hasPart": [
                {
                    "@id": "#openscience-grupo"
                },
                {
```

ro-crate-metadata.json

Best practices

- Description
- Registries
- Releases
- Requeriments
- Installation
- Usage
- Tests



Demo