

# Mg Alloy Text Mining Tool

A Python Application for Automated Data Extraction and Analysis

Dr. Yaa Takyiwaa Acquah  
Postdoctoral Scholar  
Department of Computer Science

12/12/2024



## Outline:

- Background and Motivation
- Objectives
- Research question
- Methodology diagram
- Text mining application
- Demonstration
- Key achievements
- Future directions



- Magnesium (Mg) alloys are widely researched for their lightweight, high-strength properties, making them critical in industries such as aerospace, automotive, and biomedical engineering.
- Research data on Mg alloys is often dispersed across scientific publications, stored in unstructured formats, and challenging to process manually.

## **Motivation**

- Manual extraction and cleaning of data from scientific articles are labor-intensive and prone to errors.
- The growing volume of scientific publications requires an automated and efficient solution for data extraction and analysis.



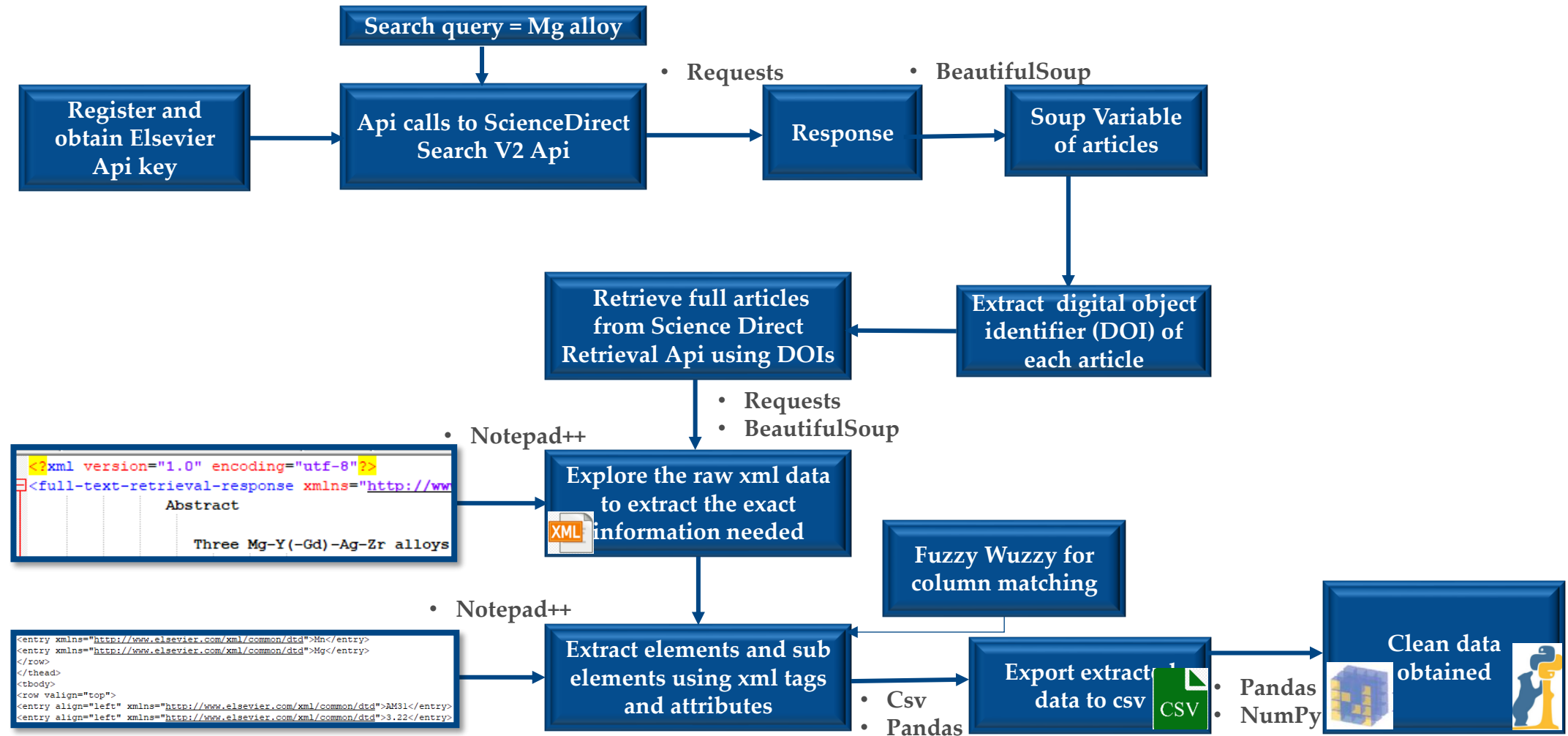
- The main objective of this research is to extract specific information intelligently from scientific articles.
- The specific information includes:
  - material compositions (wt%),
  - processing Methods and Parameters,
  - mechanical Properties and
  - Corrosion Properties.
- Consolidate the extracted information into clean, comprehensive datasets for further material science research and analysis.



## Research Question

- How can we automate the extraction, processing, and analysis of large volumes of scientific data from unstructured and semi-structured sources, such as research articles, to enable efficient data-driven insights in material science, specifically for magnesium alloys?"

| Title  | Source (link)   | Coresponding Author                               | Nationality | Institute  | Material Composition |       |    |      |       |    |    |    |    |    | Mechanical Properties |           |       | Corrosion Properties |      |
|--|---|---|-------------|--|----------------------|-------|----|------|-------|----|----|----|----|----|-----------------------|-----------|-------|----------------------|------|
|  |   |   |             |  |                      |       |    |      |       |    |    |    |    |    | Compression           |           |       | Immersion            |      |
|  |   |   |             |  | Al                   | Ca    | Zn | Mn   | Sn    | Li | Cu | Pb | Na | Zr | YS (MPa)              | UCS (MPa) | E (%) | mg/mm <sup>2</sup> h | mm/y |
| Phase equilibria, thermodynamics and solidification microstructures of Mg–Sn–Ca alloys, Part 1: Experimental investigation and thermodynamic modeling of the ternary Mg–Sn–Ca system | <a href="https://www.sciencedirect.com/science/article/pii/S0966979507002087?via%3Dihub">https://www.sciencedirect.com/science/article/pii/S0966979507002087?via%3Dihub</a> | schmid-fetzer@tu-clausthal.de (R. Schmid-Fetzer). | Germany     | Institute of Metallurgy, Clausthal University of Technology, Robert-Koch-Str       |                      | 40.28 |    |      | 0.21  |    |    |    |    |    |                       |           |       |                      |      |
|  |   |   |             |  |                      | 22.91 |    |      | 33.93 |    |    |    |    |    |                       |           |       |                      |      |
|  |   |   |             |  |                      | 15.51 |    |      | 45.65 |    |    |    |    |    |                       |           |       |                      |      |
|  |   |   |             |  |                      | 0.04  |    |      | 66.68 |    |    |    |    |    |                       |           |       |                      |      |
|  |   |   |             |  |                      | 21.75 |    |      | 64.21 |    |    |    |    |    |                       |           |       |                      |      |
|  |   |   |             |  |                      | 21.04 |    |      | 70.35 |    |    |    |    |    |                       |           |       |                      |      |
|  |   |   |             |  |                      | 3.51  |    |      | 69.25 |    |    |    |    |    |                       |           |       |                      |      |
|  |   |   |             |  |                      | 31.23 |    |      | 61.59 |    |    |    |    |    |                       |           |       |                      |      |
| Recent research and developments on wrought magnesium alloys   | <a href="https://www.sciencedirect.com/science/article/pii/S2213956717300464">https://www.sciencedirect.com/science/article/pii/S2213956717300464</a>                       | yuanding.huang@hzg.de (Y. Huang).                 | Germany     | MagIC-Magnesium Innovation Centre, Helmholtz-Zentrum Geesthacht, Max-Planck-Str. 1 | 3.00                 | 3.00  |    | 0.30 |       |    |    |    |    |    | 285                   | 472       | 9.5   |                      |      |







Mg Alloy Text Mining Tool

## Elsevier Mg Alloy Text Mining Tool

Enter search query here

Enter Start page of such (additions of 25;Eg.1,26,51,...)

Submit

RETRIEVE DOI

Retrieve Tables

Retrieve combined article info

Click RETRIEVE DOI BOTTON to continue

25 Article Titles, Links and Authors extraction sucessful

check for results in the csv file

Using DOIs to retrieve full articles; Click Retrieve Tables button.

Table extraction sucessful. check the folder articletables for details.

Table extraction sucessful. check the folder articletables for details.

ws (C:) > Users > NaYa > PycharmProjects > GraphicUserInterfaceAndBatchFile > articletables

| Name  | Date modified       | Type        | Size |
|---|---------------------|-------------|------|
| <input checked="" type="checkbox"/> S2352-4928(24)02936-2 | 12/12/2024 10:27 AM | File folder |      |
| <input type="checkbox"/> S2238-7854(24)02532-8            | 12/12/2024 10:27 AM | File folder |      |
| <input type="checkbox"/> S0921-5093(24)01206-1            | 12/12/2024 10:27 AM | File folder |      |
| <input type="checkbox"/> S0925-8388(24)03738-1            | 12/12/2024 10:26 AM | File folder |      |
| <input type="checkbox"/> S0167-577X(24)01785-3            | 12/12/2024 10:26 AM | File folder |      |
| <input type="checkbox"/> S0925-8388(24)04426-8            | 12/12/2024 10:26 AM | File folder |      |
| <input type="checkbox"/> S1005-0302(24)00971-X            | 12/12/2024 10:26 AM | File folder |      |
| <input type="checkbox"/> S0022-3697(24)00642-5            | 12/12/2024 10:25 AM | File folder |      |
| <input type="checkbox"/> S0925-8388(24)03198-0            | 12/12/2024 10:25 AM | File folder |      |
| <input type="checkbox"/> S2352-4928(24)03247-1            | 12/12/2024 10:25 AM | File folder |      |
| <input type="checkbox"/> S2213-9567(24)00378-5            | 12/12/2024 10:24 AM | File folder |      |
| <input type="checkbox"/> S2238-7854(24)02624-3            | 12/12/2024 10:24 AM | File folder |      |
| <input type="checkbox"/> S0921-5093(24)01580-6            | 12/12/2024 10:24 AM | File folder |      |
| <input type="checkbox"/> S2352-4928(24)02874-5            | 12/12/2024 10:23 AM | File folder |      |
| <input type="checkbox"/> S2213-9567(24)00342-6            | 12/12/2024 10:23 AM | File folder |      |

ws (C:) > Users > NaYa > PycharmProjects > GraphicUserInterfaceAndBatchFile > artbatch

| Name   | Date modified       | Type                 | Size |
|--|---------------------|----------------------|------|
| <input type="checkbox"/> article-1             | 12/12/2024 10:22 AM | Microsoft Excel C... | 1 KB |
| <input checked="" type="checkbox"/> article-2  | 12/12/2024 10:22 AM | Microsoft Excel C... | 2 KB |
| <input checked="" type="checkbox"/> article-3  | 12/12/2024 10:22 AM | Microsoft Excel C... | 1 KB |
| <input checked="" type="checkbox"/> article-4  | 12/12/2024 10:23 AM | Microsoft Excel C... | 1 KB |
| <input checked="" type="checkbox"/> article-5  | 12/12/2024 10:23 AM | Microsoft Excel C... | 1 KB |
| <input checked="" type="checkbox"/> article-6  | 12/12/2024 10:23 AM | Microsoft Excel C... | 1 KB |
| <input checked="" type="checkbox"/> article-7  | 12/12/2024 10:23 AM | Microsoft Excel C... | 1 KB |
| <input checked="" type="checkbox"/> article-9  | 12/12/2024 10:24 AM | Microsoft Excel C... | 1 KB |
| <input checked="" type="checkbox"/> article-10 | 12/12/2024 10:24 AM | Microsoft Excel C... | 1 KB |
| <input checked="" type="checkbox"/> article-11 | 12/12/2024 10:24 AM | Microsoft Excel C... | 1 KB |
| <input checked="" type="checkbox"/> article-12 | 12/12/2024 10:25 AM | Microsoft Excel C... | 1 KB |
| <input checked="" type="checkbox"/> article-13 | 12/12/2024 10:25 AM | Microsoft Excel C... | 1 KB |
| <input checked="" type="checkbox"/> article-14 | 12/12/2024 10:25 AM | Microsoft Excel C... | 1 KB |
| <input checked="" type="checkbox"/> article-16 | 12/12/2024 10:26 AM | Microsoft Excel C... | 1 KB |
| <input checked="" type="checkbox"/> article-18 | 12/12/2024 10:26 AM | Microsoft Excel C... | 1 KB |
| <input checked="" type="checkbox"/> article-19 | 12/12/2024 10:26 AM | Microsoft Excel C... | 1 KB |
| <input checked="" type="checkbox"/> article-20 | 12/12/2024 10:26 AM | Microsoft Excel C... | 1 KB |
| <input checked="" type="checkbox"/> article-21 | 12/12/2024 10:27 AM | Microsoft Excel C... | 1 KB |
| <input checked="" type="checkbox"/> article-24 | 12/12/2024 10:27 AM | Microsoft Excel C... | 1 KB |
| <input checked="" type="checkbox"/> article-25 | 12/12/2024 10:27 AM | Microsoft Excel C... | 1 KB |

- **Digital Object Identifier (DOI)**, is a string of numbers, letters and symbols used to uniquely identify an article or document.
- The **Publisher Item Identifier (PII)** is a unique identifier used by a number of scientific journal publishers to identify documents.



| Titles                    | DOI           | Pii        | Link       | Authors   | Al   | Ca | Zn   | Mn   | Sn | Cu | Zr   | Y    | Gd   | Mg    | Si | Bi   | As-cast                   | YS (MPa) | UTS (MPa) |
|---------------------------|---------------|------------|------------|---|------|----|------|------|----|----|------|------|------|-------|----|------|---------------------------|----------|-----------|
| Rationalizing Al-Mg, Al-I | 10.1016/j.mt  | S235249282 | https://ap | Junsheng WangJiaqiang HanRuifeng DouDongxu ChenHoubing Huang      |      |    |      |      |    |    |      |      |      |       |    |      |                           |          |           |
| Effect of Al on microstr  | 10.1016/j.pns | S100200711 | https://ap | Shida MaAitao TangPeng PengGe                                     | 3.22 |    |      | 1.22 |    |    |      |      |      | 95.6  |    |      |                           | 266      | 25        |
|                           |               |            |            |   | 5.8  |    |      | 0.92 |    |    |      |      |      | 93.28 |    |      |                           | 280      | 27        |
|                           |               |            |            |   | 9.07 |    |      | 0.88 |    |    |      |      |      | 90.05 |    |      |                           | 307      | 30        |
|                           |               |            |            |   |      |    |      |      |    |    |      |      |      |       |    |      |                           |          |           |
| Tuning texture and prec   | 10.1016/j.ma  | S104458032 | https://ap | Yu ZhangWei RongYujuan WuLiming Peng                              |      |    |      |      |    |    |      |      |      |       |    |      |                           |          |           |
| A study of microstructu   | 10.1016/j.ms  | S092150932 | https://ap | Dan WangPenghuai FuLiming PengYingxin WangWenjiang Ding           |      |    |      |      |    |    | 0.4  |      | 9.49 |       |    |      | Eutectic phase area ratio |          |           |
|                           |               |            |            |   |      |    |      |      |    |    | 0.35 | 0.96 | 8.86 |       |    |      | 86 μm                     | 222      |           |
|                           |               |            |            |   |      |    |      |      |    |    |      |      |      |       |    |      | 78 μm                     | 233      |           |
|                           |               |            |            |   |      |    |      |      |    |    |      |      |      |       |    |      |                           |          |           |
| Effects of Si content an  | 10.1016/j.ma  | S025405842 | https://ap | Tian GuoShusen WuXiong ZhouShulin LüLanc                          |      |    | 0    |      |    |    |      |      |      |       |    | 0.91 |                           |          |           |
|                           |               |            |            |   |      |    | 0    |      |    |    |      |      |      |       |    | 2.13 |                           |          |           |
|                           |               |            |            |   |      |    | 0    |      |    |    |      |      |      |       |    | 2.92 |                           |          |           |
|                           |               |            |            |   |      |    | 0    |      |    |    |      |      |      |       |    | 4.14 |                           |          |           |
|                           |               |            |            |   |      |    | 0.18 |      |    |    |      |      |      |       |    | 4.17 |                           |          |           |
|                           |               |            |            |   |      |    | 0.44 |      |    |    |      |      |      |       |    | 3.96 |                           |          |           |
|                           |               |            |            |   |      |    | 0.62 |      |    |    |      |      |      |       |    | 3.89 |                           |          |           |
|                           |               |            |            |   |      |    | 0.77 |      |    |    |      |      |      |       |    | 3.95 |                           |          |           |
|                           |               |            |            |   |      |    | 1.05 |      |    |    |      |      |      |       |    | 4.13 |                           |          |           |
| Achieving ultra-high stre | 10.1016/j.jms | S100503022 | https://ap | Yu ZhangWei RongYujuan WuLiming Peng                              |      |    |      |      |    |    |      |      |      |       |    |      |                           |          |           |
| Assessment of atomic r    | 10.1016/j.jms | S100503022 | https://ap | Yuhui ZhangYuling LiuShuhong LiuHai-Lin ChenYong Du               |      |    |      |      |    |    |      |      |      |       |    |      |                           |          |           |
| New Mg-Ca-Zn amorph       | 10.1016/j.mtl | S258915292 | https://ap | Sudeep PaulParthiban RamasamyMitun DasDurbadal MandalSupriya Bera |      |    |      |      |    |    |      |      |      |       |    |      |                           |          |           |
| Fabrication of high-stre  | 10.1016/j.ma  | S104458032 | https://ap | Qingchen DengYujuan WuYuanhang LuoNing SuLiming Peng              |      |    |      |      |    |    |      |      |      |       |    |      |                           |          |           |
| Investigation of the allo | 10.1016/j.jm  | S221395672 | https://ap | Alireza MaldarLeyun WangGaoming ZhuXiaoqin Zeng                   |      |    |      |      |    |    |      |      |      |       |    |      |                           | 90       | 18        |
|                           |               |            |            |   |      |    |      |      |    |    |      |      |      |       |    |      |                           | 83       | 21        |
|                           |               |            |            |   |      |    |      |      |    |    |      |      |      |       |    |      |                           | 148      | 22        |

- Above picture shows Titles, doi's, pii's, link, Authors and Tables of the 25 articles exported to csv with the





**Automation:**

Eliminates the need for manual data extraction by integrating with Elsevier's API and automating metadata and table retrieval.

**Scalability:**

Capable of processing large volumes of articles and combining information into a single, analyzable dataset.

**User-Friendliness:**

Provides a graphical user interface (GUI) that simplifies the data retrieval and processing pipeline for non-technical users.



Mg Alloy Text Mining Tool

## Elsevier Mg Alloy Text Mining Tool

Enter search query here

Enter Start page of such (additions of 25;Eg.1,26,51,...)

Submit

RETRIEVE DOI

Retrieve Tables

Retrieve combined article info





- Expand to Other Domains:
  - Adapt the tool for use with other material types or scientific fields.
- Enhanced NLP Integration:
  - Incorporate more sophisticated natural language processing techniques for semantic understanding of content beyond tables.
  - Intelligent column matching.
- Visualization:
  - Add data visualization features to generate insights directly within the tool.