Constructing a Structured Corpus from Geoscience Literature: A Case Study using Western Australia Iron and Lithium Deposits

Villacorta Chambi, S. P.,[1*] Lindsay, M.,[2,3,4] Gessner, K.,[5] Klump, J.[2] and McFarlane, H.[2]
[1]IAEG Peruvian Group, Piura, Zipcode: 20301, Peru, E-mail: villacortasp@gmail.com
[2]Commonwealth Scientific and Industrial Research Organization - CSIRO, Mineral Resources, Kensington, Kensington, WA 6151, Australia
[3]School of Earth Sciences, The University of Western Australia, Crawley, WA 6009, Australia
[4]ARC ITTC Data Analytics for Resources and Environment, Crawley, WA 6009, Australia
E-mails: mark.lindsay@csiro.au; jens.klump@csiro.au; helen.mcfarlane@csiro.au
[5]Geological Survey of Western Australia, Australia, East Perth, WA 600, Australia
E-mail: klaus.GESSNER@demirs.wa.gov.au
**Corresponding Author*

DOI: https://doi.org/10.52939/ijg.v21i5.4159

Abstract
The field of geoscience is facing challenges when it comes to combining various data sets and formats to facilitate knowledge discovery. Academic literature often contains unstructured data, which require intricate processing for machine comprehension. Information Extraction can plays a vital role in effectively organizing these data by enhancing the functionality of artificial intelligence algorithms. Natural Language Processing (NLP) and Named Entity Recognition (NER), particularly using ontologies, have been pivotal in achieving semantic consistency within integrated datasets. Advances in geospatial artificial intelligence, propelled by deep learning and machine learning, have been crucial in converting vast amounts of spatial data into practical knowledge. However, the reliance on extensively labelled datasets imposes limitations due to the intensive nature of the process. This paper showcases a case study demonstrating the practical application of these methods to journal articles on mineral deposits in Western Australia. The results highlight the efficacy of our approach in transforming unstructured data into structured information, thereby contributing to the advancement of knowledge extraction in geoscience. This study not only underscores the importance of robust data processing strategies but also shows how advanced machine learning and deep learning techniques in geospatial artificial intelligence, enable more informed geological research and decision-making.

Keywords: Corpora, Geo-databases, NLP, Pre-processing, Unstructured Text

1. Introduction

Geoscientists encounter a major challenge when managing vast and diverse sources of data and knowledge products, such as seismic records, geochemical analyses, geological surveys and research publications. These data types range from structured databases like EarthChem and Geobiodiversity, which are formatted in accessible tables enhancing machine readability [1] to unstructured formats such as academic literature that necessitate sophisticated processing for computational analysis [2]. In addition to text documents, the spatial and temporal aspects of geology are commonly presented in maps or in constructed or parametric three dimensional (3D) models. 3D models, explicit and parametric, offer visual and analytical representations of geological phenomena but often fail to capture complex geological semantics and lack flexibility, making them less suitable for dynamic geological analysis [3] and [4]. Despite their advantages, these models often overlook the granular details of geological knowledge, such as topology and the logic of geological history, which are crucial for informed decision-making and hypothesis testing in research and exploration [5]. The heterogeneous formats of geoscience data make automated processing challenging specially when managing dynamic datasets across different platforms [6].

Moreover, the rapid increase in geoscientific research publications and the diversity of data formats such as reports, patents, and geological surveys present significant challenges for researchers. Inconsistencies in data presentation, particularly in scanned Portable Document Format (PDF) documents from legacy sources, complicate automated text extraction and hinder standardization efforts necessary for further analysis [7]. In response, the use of Artificial Intelligence (AI) particularly through Natural Language Processing (NLP) and Information Extraction (IE) techniques have begun to transform the handling of unstructured texts. These AI-driven methods have become essential for managing unstructured geoscientific texts. Early techniques, like gazetteers and rule-based systems, laid the groundwork for more advanced AI applications, such as Named Entity Recognition (NER), which is now widely used in complex domains like geosciences. The development of structured corpora ("corpus" in singular) large, annotated datasets has enabled more effective NER applications in geosciences, forming the backbone of machine learning (ML) and data-driven research [8]. Alongside this, Extract, Transform, Load (ETL) processes, supported by programming languages and Structured Query Language (SQL), have been instrumental in standardizing and transforming data for AI-powered analysis [9].

As part of this AI evolution, data and text mining techniques such as clustering, decision trees, and deep learning have become crucial tools for extracting patterns and organizing large datasets. Data mining focuses on structured datasets, while text mining, a branch of data mining, applies similar principles to unstructured textual data, especially valuable in processing vast amounts of scientific literature. These techniques are becoming increasingly important for mineral exploration research and real-time geological analysis [10] and [11]. AI-driven data and text mining techniques have the potential to uncover hidden patterns, relationships, and trends, ultimately supporting more informed and accurate geoscience research [12]. Knowledge Graphs (KGs) have gained attention for their ability to connect and integrate diverse data sources. This integration helps in organising complex geological relationships, ensuring data consistency across datasets [13]. Ontologies withing these KGs provide a semantic framework crucial for effective data integration and processing [14] and [15]. However, challenges remain in adapting general ontologies to specific domain needs, often requiring labor-intensive modifications [16] and [17].

With the rapid evolution of neural network and language models like Bidirectional Encoder Representations from Transformers (BERT), Bidirectional Gated Recurrent Unit (BiGRU), Attention-Conditional Random Fields (CRF) and Attention-weighted BERT- Bidirectional Long Short-Term Memory (BiLSTM), there is significant potential in various NER scenarios, enhancing accuracy in other complex domains [18] and [19].

Additionally, embedding techniques such as those from Embeddings from Language Models (ELMo), Extra-Long Network (XLNet), have further enhanced NER performance by capturing the context within domain-specific datasets [20] and [21]. Despite these advancements, automating the creation of high-quality corpora remains a challenge, as accurate preprocessing is crucial to ensure the precision of NER systems. Since 2023, large language models (LLMs), such as the Generative Pre-trained Transformer (GPT) have demonstrated the ability to generate annotated datasets automatically, reducing the manual effort involved in corpus creation. This advancement is particularly promising for geosciences, where the need for precise entity recognition and classification is critical in data analysis and research [22] and [23].

In this context, this paper aims to address three interconnected research objectives against these challenges: firstly, by reviewing state-of-the-art methods, showing the application in a study case and applying fine-tuned models over generic approaches, the goal is to enhance NER performance in the geosciences and lay the groundwork for further research on the topic; secondly, presents a methodology for creating domain-specific corpora from geoscience literature, specifically focused on mineral deposit terminology, that can serve for NER training and a benchmark for data processing techniques; and thirdly, to show the refinement of preprocessing methods to effectively handle the complex features of geoscience texts. These methods are shown in a pipeline for building geoscience corpora including tools and scripts to facilitate the adoption of the proposed pipeline by other geoscientists. The hypothesis is that tailored NER approaches supported by well-curated corpora and refined preprocessing will significantly improve the accuracy and efficiency of data processing in geosciences.

Moreover, this study explores the integration of data and text mining techniques such as clustering, decision trees, and deep learning to transform unstructured data into structured formats, enhancing the discovery of hidden patterns and relationships.

A case study illustrates the extraction and examination of corpora from PDF documents, showcasing the application of NER in geosciences. The structure of this manuscript is as follows: Section 2 provides an overview of advancements in NLP for geoscientific data analysis, emphasizing existing challenges. Section 3 outlines the methodology for converting PDF documents into structured corpora tailored for NER tasks. Section 4 presents a case study that demonstrates the preparation process of corpora from unstructured text related to mineral deposit papers along with the tools and techniques employed. Section 5 details the results, followed by Section 6, which discusses the findings, their implications, and the broader applicability of the workflow to other geological contexts. Finally, the conclusions are presented in Section 7 contributing to ongoing discussions on NLP applications in geoscience.

2. Advances in NLP for Geoscientific Data Analysis and Remaining Challenges

The Earth is continuously observed by an array of satellites, drones, and sensors, producing a wealth of global datasets of varying resolutions. These include high-frequency imagery from Sentinel-2 satellites and detailed revisits from Landsat-7 and -8 missions. However, traditional operational geophysical mapping often relies on single-date spectral data for classification, which fails to fully exploit the richness of these datasets, thus limiting the potential for enhanced classification outcomes [24]. Deep learning algorithms have significantly impacted remote sensing by improving the extraction of complex features from raw data, thereby enhancing image analysis and anomaly detection [25].

Integrating geospatial science with artificial intelligence has given rise to geospatial artificial intelligence (GeoAI), which employs algorithms that emulate human spatial cognition. This technology addresses complex spatially centred human-environmental system challenges in geological research, broadening application to include predictive interpolation methods, demographic data enrichment, and environmental event simulation [26]. Aligned with thise efforts, the National Aeronautics and Space Administration (NASA) of the USA, in collaboration with the International Business Machines Corporation (IBM), developed INDUS, a suite of LLMs tailored for Earth science, biology, physics, heliophysics, planetary sciences, and astrophysics. INDUS, trained in curated scientific corpora, includes an encoder model for natural language understanding, a contrastive-learning-based text embedding model for information retrieval, and distilled versions for resource-constrained applications. Additionally, scientific benchmark datasets on climate change and Earth Science were created. The modelcreators stated that INDUS models outperform general-purpose and existing domain-specific encoders [27].

Under the Deep-Time Digital Earth (DDE) project, the GeoDeepShovel initiative, exemplifies effective data structuring, transforming unstructured scientific literature into formats applied in geoscience research, utilizing neural networks and weak supervision learning [7]. Similarly, the Global Earth Observation System of Systems (GEOSS) contributes to global data sharing and environmental monitoring by integrating various Earth observation datasets, having evolved since its 2005 inception into a continually expanding and functional system [28]. NER technology is facilitating the identification of key geological concepts such as geological age and structure, challenged by the unique language and sentence structures of geoscience texts [29]. Advanced models like BERT-BiGRU-CRF have enriched the semantic representation of character vectors [4], surpassing the capabilities of static models like Word2Vec in extracting complex geological. The Convolutional Neural Networks (CNN)-BiLSTM-CRF model merged dynamic and character-level features offering a more nuanced analysis of geological data [30]. These efforts were developed to address the limitations of traditional NER methods, which heavily rely on feature engineering and often suffer from tag inconsistency, with newer models, that achieves high F1scores [31]. F1 is a metric which combines precision and recall and is commonly used to as indicative of NER model accuracy [32] and [33].

The effectiveness of NER applications heavily relies on the linguistic richness of training datasets. While extensively trained corpora such as OntoNotes significantly boost NLP tool performance by capturing linguistic nuances [34], languages with scarce annotated datasets continue to exhibit lower NER accuracy, showing a prevalent challenge within the NLP community [34] and [35]. Table 1 delineates the progress of NER research on geosciences, illustrating the increased adoption of sophisticated NLP tools and methodologies, to extract geological terminologies from geological articles and databases like the Wanfang database (http://www.wanfangdata.com.cn) [36]. These studies have utilized tools like CRF, Bi-LSTM and BERT which can handle complex syntax and semantics typical of geoscience texts. Early efforts focused on developing a NER system tailored for geological texts [37].

Table 1: Overview of NER research in geoscience domain

Application / Goal	Corpora	Entities	Methods/tools	References
Developing a NER system specifically for geological texts	Geological reports and articles from India	Geographic entities, Mineral, Year, Organization, Measures, Person, Time, Fault, and Rock.	CRF	[37]
Improve map-based retrieval of scientific data by tagging figure and table captions with geological time expressions and geographic locations in scientific articles	80 geological articles including figure and table captions	Geological timescale expressions and geographic locations	Gazetteer-based methods using geological ontologies and GeoNames and a Bi-LSTM model	[38]
Extraction of geological named entities from unstructured Chinese geoscience reports	Geological reports in Chinese language	Geological history, geological structure, rock, stratum	Bi-LSTM	[31]
Extraction of named entities related to geological hazards to enhance knowledge graph construction for geological hazards.	Wanfang database targeting geological hazards literature	Geological hazard-related entities like location names, methods, and data	BiGRU-CRF, BiGRU and CRF	[36]
Automatic labelling of geological entities from mineral exploration reports in Western Australia	Western Australian mineral exploration reports	Rock, mineral, timescale, stratigraphy, location, ore deposit	BiLSTM, CNNs, CRF	[40]
Developing a NER model for geological texts	Geological reports and manually annotated geological data	Strata, rocks, minerals, geological processes, and geological time	few-shot learning, pre-trained BERT (GeoBERT), Bi-LSTM, CRF	[41]
Improve the extraction of geological named entities from unstructured texts	Geological reports in Chinese language	geological ages, stratigraphic units, rock types, minerals, and geographic locations.	BERT, BiGRU and CRF	[4]
Mining geological literature for building ontologies and annotation schemas specifically for porphyry copper deposits	Scholarly articles on porphyry copper deposits	Mineral deposit, Location, Geological setting, Igneous rock, and others	Bi-LSTM and CRF	[42]
Creating a platform for integrating multimodal data from geoscience literature, including text, images, and tables	40,000 documents from geoscience literature	Names, lithologies, geological ages, geographic locations, geoscience features	Unified Information Extraction for text and You Only Look Once, Version 3 for image and table detection.	[38]
Developing a suite of LLMs tailored for Earth science, biology, physics, heliophysics, planetary sciences, and astrophysics	300 million samples from internet sources like Wikipedia, StackExchange, scientific data from PubMed, PMC, Arxiv, and S2ORC, ADS data	Domain-specific entities across multiple science disciplines	Contrastive learning for sentence embedding, BERT, mean pooling, knowledge distillation for smaller models, Retro-Masked Autoencoder style pretraining, and supervised fine-tuning with annotated datasets (NQ, SQuAD, and SPECTER pairs).	[27]

Later, this approach expanded the scope of entity recognition to improve data retrieval and visualization [38], bridging the gap between raw text data and usable information for geoscientists. Moreover, integrating multimodal data from geoscience literature, underscore a shift towards more holistic data processing [39], which emphasizes the integration of various data types to enrich scientific research, akin to the approaches in DDE and GEOSS.

Additional research have been directed towards building knowledge graphs that classify entities related to mineral resources [40] which represents advancements in mining exploration research. This work often requires manual verification of entity classifications, which can limit the scope of analysis. Despite notable advancements in NLP applied to geosciences, challenges persist, particularly in integrating geo-semantic features effectively and the need for extensive human intervention. These limitations present significant opportunities for advancing research and technology, particularly through the development of specialized embedding models and automated tools focused on geoscience-related entities [43]. Even though the application of NLP in the geoscience domain is limited, it has been employed across various fields for the analysis of complex tasks such as the classification of geo-databases and the extraction of insights from data on natural phenomena [44]. In these instances, using lexicon-based methods [45] and [46] and strategies relying on manually created pattern-matching rules.

The specific requirements of geosciences require the development of dedicated ontologies such as those created for the geologic time scale [47], geological modelling [5], structural geology [48], and geotechnics [49]. Unlike general domain language, the absence of tailored ontologies/schemas in the geoscience sector can hinder the process of automatic information extraction. Geological language is characterized by high ambiguity, extensive character lengths, a plethora of distant words, and varied combinations of terms relating to geological named entities, rendering these models dependent on context [39]. As a result, the formal characterization of geological concepts presents difficulties [44][47] and [50]. This underscores the importance of expert input in selecting appropriate names and verbs, and highlights the challenges associated with identifying, verifying, and swiftly updating new knowledge within KGs [14]. Examples of term usage in both general and geoscientific texts are illustrated in Table 2.

An additional limitation is the limited online publication of such geoscience domain ontologies. Further challenges include the restricted reusability, linking, and interoperability of existing geoscience vocabularies and a scarcity of qualified professionals skilled in executing data annotations. Creating high-quality, structured corpora is essential for facilitating a wide range of ML applications, including predictive modelling, classification tasks, and anomaly detection within geoscience data, underscoring its significance in specialised research domains.

3. Methodology

The following methodology explains how to convert PDF documents into structured corpora for NER tasks. The pipeline, as shown in Figure 1, consists of four main steps: data collection, preprocessing, training, and extraction.

3.1 Workflow (Pipeline)

The pipeline described below outlines the general steps for gathering, preprocessing, training, and extracting data using NLP techniques.

Table 2: Examples of use of terms in common domain and their use in geoscientific domain [51]

Word	Common	Geology
Dating	Romance	Finding the age of a rock
Formation [49]	A natural process that causes something to form (for instance: "the formation of a crystal").	A fundamental unit of stratigraphy
Bomb	An explosive weapon	A partly molten material from a volcanic vent which solidifies in flight or after landing.
Deposit	Money in the bank	An accumulation of terrestrial materials
Hamersley	Geographic locality, a town	Geological grouping of rock formations

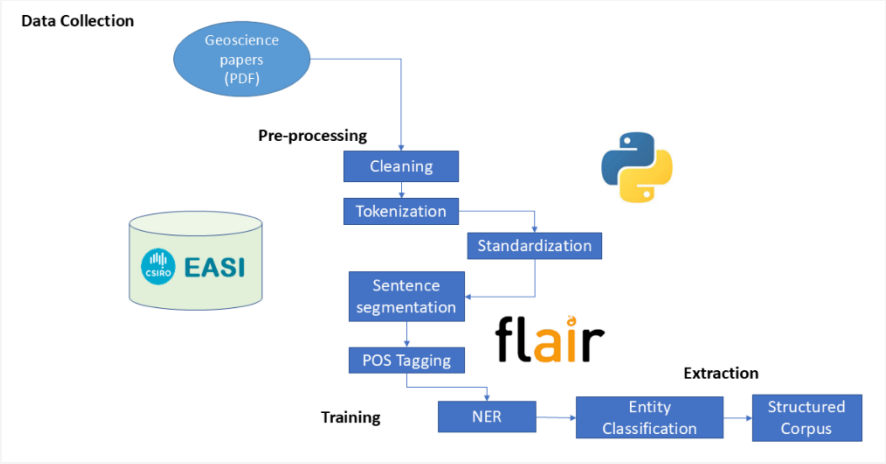

Figure 1: Workflow diagram of a pipeline using EASI, Python, and Flair for converting unstructured geoscience PDFs into structured corpora for NLP tasks

3.1.1 Data collection

Data relevant to the topic of interest can be gathered from online repositories such as Google Scholar, Scopus, Academia.edu, and ResearchGate. These platforms provide access to a vast array of scholarly articles, ensuring a comprehensive dataset for analysis.

3.1.2 Preprocessing

The preprocessing stage involves text cleaning to remove extraneous elements like headers, footers, and publication dates. This stage ensures that the text is in a uniform format, facilitating accurate and efficient analysis. This process includes tokenization, stemming, sentence segmentation and part-of-speech (POS) tagging.

Tokenization involves breaking down the text into individual units, such as words or phrases. This task is crucial for preparing text for deeper NLP tasks, as it allows for the analysis of text at a granular level. A common approach is using the Natural Language Toolkit (NLTK) in Python language, which provides robust methods for tokenizing text [52] and is employed for text structuring, setting the foundation for subsequent processing steps. After tokenization, it is required a standardization of textual data by converting all characters to lowercase, correcting misspellings, removing punctuation and converting synonyms to a standardized form. Regular expressions (RegEx) can be used in this section to handle multi-word terms effectively. This step helps reduce text data's complexity and improves NLP models' performance [51]. Techniques such as lemmatization or stemming are often employed to reduce words to their base or root form. These steps help to normalize the text, reducing the complexity of variations of the same word and improving the performance of NLP models [53].

Sentence segmentation is the process of dividing the text into individual sentences which is essential for understanding the context in which every term is used. Clear sentence segmentation boosts NLP model performance by establishing essential boundaries for context [53]. POS tagging involves assigning parts of speech (e.g., noun, verb, adjective) to each token. This tagging is essential for understanding the grammatical structure of sentences and for subsequent syntactic parsing. The Stanford NLP Group provides tools that accurately perform POS tagging [54]. POS tagging is particularly useful in complex domain texts to differentiate between similar terms used in different contexts [55].

3.1.3 Training and entity classification

At this stage, NER tools are utilized to apply a schema (labelled dataset, often called "domain dictionary") which is designed to training the NER model making it can identify and categorize domain-specific terminology. Once trained, the model can annotate new texts classifying entity labels. Such schema is crucial in deriving insights within a specific domain. It is a labelling framework that categorizes each entity under types pertinent to the domain under study, such as "OzRock" for mineral resources [40]. Verification is crucial to ensure the accuracy and reliability of the model's annotations. This process involves systematically addressing any discrepancies that may arise from automated processing. The model's efficacy is assessed using standard metrics such as precision, recall, and F1-score, which help confirm that the model meets the required standards for scientific research.

3.1.4 Corpora extraction

Post-training, the models from the classification phase are applied to extract and categorize entities within the texts. The annotated text is organized into a structured corpus, which simplifies data access and manipulation and supports advanced analytical methods. The extraction and structuring of the corpus require advanced NLP tools and frameworks. Text extraction from PDFs can be performed using tools like PDF Plumber [56], while for text processing other packages like the NLTK. State-of-the-art NLP library, Flair [57], SpaCy [58], and Prodigy [59] are recognized tools to apply in the entity recognition phase.

SpaCy: Known for its fast and efficient NER capabilities, especially in the general domain, SpaCy also facilitates easy integration with other NLP tasks like tokenization and POS tagging. While it allows for training on domain-specific data, it lacks the advanced contextual embedding capabilities of Flair and necessitates substantial expertise for effective customization.

Prodigy: This tool is popular for its user-friendly annotation interface, which is instrumental in creating tailored datasets through its active learning framework. Prodigy allows data annotation and integrates smoothly with SpaCy for rapid model development and iteration. The primary drawback is its licensing cost, which may be prohibitive for non-commercial use, and the complex setup for using the packages in external scripts.

Flair: this tool provides state-of-the-art contextual string embeddings that encapsulate both semantic meaning and contextual usage of words. It supports multiple languages and allows for fine-tuning domain-specific datasets, which is particularly beneficial for incorporating geoscience terminology. Yet, it requires significant computational resources particularly when training bespoke models.

4. Case Study: Creating Corpora from Geoscience Papers on Iron and Lithium Deposits in Australia

This section shows how the previously explained pipeline was applied to create a corpus of academic papers on mineral deposits.

4.1 Datasets Selection and Collection

The geology of Western Australia, characterized by its distinct tectonic history and diverse mineral resources, provides an excellent test bedfor economic and academic studies. Papers describing banded iron ore were selected: banded iron ore deposits due to their substantial economic and environmental impacts strongly influencing the global steel markets [60]. In addition, we explored lithium deposits, which have garnered increasing attention due to their classification as critical minerals vital for various modern renewable energy technologies [61].

The selection of twenty scholarly articles to create corpora on iron and lithium mineral deposits as a demonstration of the proposed pipeline's flexibility and scalability. The chosen number is not a determining factor in the effectiveness of the methodology, as the pipeline is flexible and scalable to any corpus size, whether it consists of 7, 20, or 50 documents. We used the public open-access repository Google Scholar (https://scholar.google.com/) for its extensive collection of academic articles, and the Commonwealth Scientific and Industrial Research Organization (CSIRO) Libraries (https://alumni.csiro.au/csiro-libraries/) allowing efficient collection of relevant documents in PDF format. For the iron ore deposits corpus, keywords like "banded iron formation" and "Western Australia" were used to ensure relevance to the geological context. In the search for lithium, we used keywords like "pegmatite lithium" and "Western Australia"; however, the search was broadened to include 'Australia' due to a limited number of relevant papers found under the initial filter. This strategy guaranteed that the chosen papers were closely aligned with the topics of interest.

To facilitate labelling and categorization of geological entities withing the selected corpora and assess the effectiveness of an NER model to identify and classify such entities, the OzRock Evaluation Set was adopted as schema [40] (Figure 2). OzRock developed from a corpus of hundreds of mineral exploration documents, categorize geological entities into six types: stratigraphic units (STRAT), ores and deposits (ORE_DEPOSIT), minerals (MINERAL), rocks (ROCK), the geological timescale (TIMESCALE), and locations (LOCATION). This Evaluation Set contains: 83838 sentences and 3278 entities and its schema was directly retrieved from its publicly available GitHub repository (https://github.com/majiga/OzROCK) without additional annotation since it was already annotated by domain experts. Table 3 shows an overview of this schema in JSON format, detailing the entity types and providing examples of geological entities.

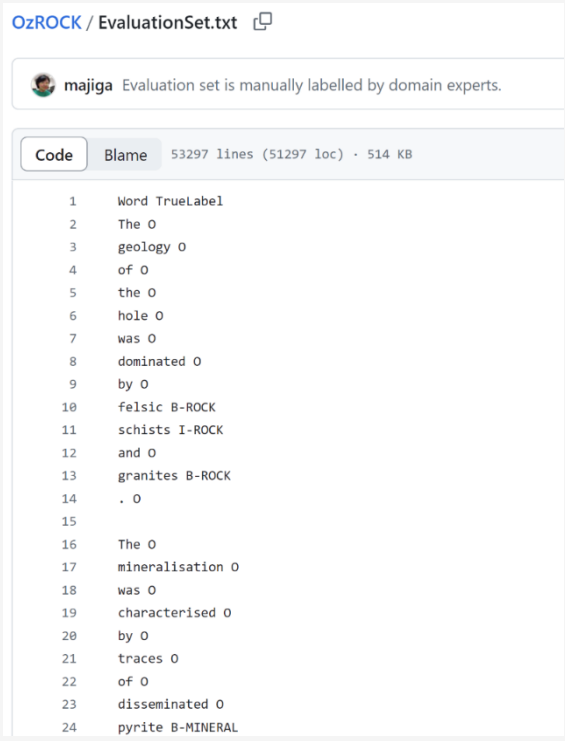

Figure 2: OzRock Evaluation set used as a schema for training a geoscience-specific NER model, enhancing recognition of geological classes in iron deposit papers. Extract taken form: *https://github.com/majiga/OzROCK/blob/master/EvaluationSet.txt.*

Table 3: Description of OzRock entity types [40]

Label (class)	Description	Example	Sentence example
MINERAL	Mineral	Copper, fire opal, goethite, gold, Iceland spar, magnesite, iron, natural salt, silica	Supergene martite-goethite mineralization is hosted within the Marra Mamba Iron Formation.
ROCK	Lithology	Conglomerate sandstone, felsic volcanic rock, migmatite, volcaniclastic sedimentary rock	The upper Bee Gorge Member, comprises intercalated shales, cherts, volcaniclastics, and turbiditic dolomites.
ORE_DEPOSIT	Ore types	Channel iron deposit, iron ore, nickel ore, silver ore	The ore is denser with lower porosity and higher goethite content compared with primary ore.
TIMESCALE	Geological Time	Archean, Lower Proterozoic, Paleoproterozoic, Triassic, Upper Cretaceous	It is not clear why a distinct BIF facies, developed on the seafloor in the Archean to Paleoproterozoic, should now show such an intimate spatial connection.
STRAT	Stratigraphy	Angas Hills Formation, Bingy Bingy Basalt Member, Marra Mamba Iron Formation	Mineralization is hosted within the Marra Mamba Iron Formation.
LOCATION	Geographical location	Kalgoorlie terrane, Kimberley craton, Perth, Pilbara, Pilbara craton, Western Australia	These operations have been the mainstay of the Western Australian iron ore industry.

4.2 Code Development and Refinement

A modified workflow was adopted after the initial pipeline execution uncovered multiple preprocessing and text extraction errors that significantly impacted readability and NER accuracy. Initially Python libraries such as PDFPlumber and NLTK (Natural Language Tool Kit) were used to extract and process the text from the selected PDF papers. However, during testing, we encountered several issues that affected the tokenization and sentence segmentation -specifically, concatenated words and inconsistent sentence breaks. These issues hindered the effective processing and tagging of entities by Flair NER model. To resolve these preprocessing challenges, several refinements were implemented in the Enhanced version of the Workflow:

PDF Text Extraction Library Change: The original Workflow utilized the PDFplumber library for text extraction, which, while effective in many scenarios, occasionally failed to accurately segment text into discrete words, particularly in cases involving concatenated words. To mitigate this issue, the fitz module from PyMuPDF was introduced as an alternative. PyMuPDF is recognized for its robust handling of diverse PDF content structures, enhancing the likelihood of retrieving cleaner text outputs.

Enhanced Regex Adjustments: The original Workflow employs a series of refined Regex operations designed to correct specific text formatting issues more conservatively. This corresponds with splitting improperly merged camelCase words, where a lowercase letter is directly followed by an uppercase letter, thus improving tokenization accuracy; ensuring spaces after punctuation marks when followed directly by letters, addressing common errors in punctuation handling in extracted texts; separating digits from adjacent letters, which is particularly relevant in geoscientific texts where numerical data and text often coalesce without adequate spacing; and correcting instances where a letter follows a period without a space, a frequent occurrence in abbreviations and acronyms in scientific literature.

Integration Into the Workflow: After text extraction and preprocessing, the cleaned text undergoes sentence tokenization using the NLTK library, which then feeds into the Flair-based NER model. Each sentence is processed individually by the SequenceTagger from Flair, which predicts entity spans based on the model trained on a similarly preprocessed corpus. The annotations include not only the textual spans of detected entities but also their corresponding entity types, contributing to a rich, structured dataset ready for further analytical tasks. The modified script is included in Appendix 1. The implemented changes in the workflow are summarized in Table 4.

4.3 Tools, Experimental Setup and Reproducibility

The experimental set up for creating corpora was developed using Python 3.8 as the programming language using libraries within the NLP framework, such as PyMuPDF for PDF text extraction, NLTK for text processing, and Flair for tagging and classification. The experiments were conducted on Windows 11 operating system and CSIRO platform EASI Hub instance equipped with a Tesla V100 GPU was used as a computational resource. To promote reproducibility, the scripts used for processing the datasets and to train the NER models are provided in the appendices. These scripts contain detailed documentation and usage instructions. The next section of the paper outlines the steps to processing our two corpora and train the NER Flair framework using the OzRock schema and its annotated dataset.

Table 4: Differences in functionality and tools used among the original and modified work-flow to extract corpora

Aspect	Original Script	Modified Script
Library for PDF extraction	pdfplumber	fitz (PyMuPDF)
Text correction function	N/A	Regex to correct concatenated words
Sentences extraction method	nltk.sent_tokenize	nltk.sent_tokenize
Text extraction function	pdfplumber	fitz
Entity recognition model	Flair SequenceTagger	Flair SequenceTagger
Entity count calculation	Counts entities in loop after model prediction	Entity counts directly calculated after model prediction

4.4 Using the Pipeline (Workflow)

This section outlines the two key phases of the pipeline implementation demonstrating how the data was pre-processed and annotated, followed by the model training and evaluation steps.

4.4.1 Preprocessing and annotation

The initial phase of this process involved processing the collected PDF documents (Iron and Lithium deposits) which were processed using PDFPlumber to extract raw text. Yet, one of the challenges when processing PDFs is that this format makes recognition of text 'line by line' which can fragment sentences hindering sentence-based analysis. To avoid that issue, we applied sentence tokenization to properly convert the raw text into complete sentences after extraction. Following text extraction, text cleaning was performed to remove non-essential elements like headers, footers, and publication dates; and then converting the text to lowercase and eliminating irrelevant punctuation. RegEx facilitated the management of multi-word terms, while text structuring and normalization were executed using the NLTK, preparing the textual data for deploying the NER models.

Subsequently, to facilitate NER tasks, the text was transformed into the format created by the Special Interest Group on Natural Language Learning during the Conference on Natural Language Learning (CoNLL) using the pdf_to_conll function and PDFplumber. The CoNLL format is recognized for being efficient when handling sequence labeling problems in NLP tasks [62] and provides a standardized dataset and evaluation framework for comparing NER systems. The BIO (Beginning, Inside, Outside) tagging scheme, which is commonly used in NER to label sequences of tokens [63] was used in our case study to run on Flair, facilitating the capture of relevant geological entities. Here is a breakdown of the BIO tags applied to our dataset:

B- (Beginning): Indicates the first token of a multi-token entity. Example (Figure 2): In the name "felsic schists", "felsic" is tagged as "B-ROCK", indicating it is the beginning of a ROCK entity.

I- (Inside): Indicates subsequent tokens of a multi-token entity. Example (Figure 2): " schists " in "felsic schists", is tagged as "I-ROCK", showing it continues the ROCK entity started by "felsic".

O- (Outside): Indicates tokens that are not part of any entity. Example (Figure 2): Common words like "by" or "and" that do not belong to any specific entity are tagged as "O".

The OzRock annotated dataset was split adopting a 70/15/15 ratio between training, testing and validation, adhering to recommended practices in ML tasks [64]. This approach is supported by NER previous efforts where ratios like 70-80% for training, 10-15% for testing and 10-15% for validation are commonly used to balance model development and evaluation effectively [65][66] and [67]. This division helped maintaining the balance between various geological entity types in the dataset, even when found class imbalances. These splits were used to train the final model (obtaining the best model.pt) which was subsequently used in the next steps of the pipeline.

The preparation of textual data culminated when structuring a corpus using Flair NER function: ColumnCorpus, employed to tokenize the text, for further processing. Flair embeddings framework ''forward and backward', was used to improve the model's contextual understanding of geological entities. The corpus thus comprised two columns: the textual content and its corresponding named entity tag.

4.4.2 Model training and assessment

For the training stage we utilized the ModelTrainer function from Flair, executing multiple epochs while fine-tuning hyperparameters such as learning rate and batch size to optimize performance. Upon applying the Flair trained model for entity extraction and classification, each corpus was subject to manual verification to verify any inaccuracies introduced by the algorithms reading the PDFs. The model's performance across different entity classes in the Iron Deposits and Lithium Deposits corpora was assessed using precision, recall, and F1-score metrics (Tables 5 and 6). The results, including the entity distributions across the structured corpora, were compared, which highlights the differences between the original and Enhanced Workflow. By comparing the performance of the Flair NER model on these two corpora, we were able to examine how the model handles the terminology, structure, and complexity inherent in each dataset. This analysis provided insights into the model's strengths and areas for improvement, especially in handling specific geological entities. Figure 3 illustrates an example of the structured corpus derived from one of the academic papers used in this study.

A New Fluid-Flow Model for the Genesis of Banded Iron Formation-
Martite-Goethite Mineralization, with Special Reference to the Nort
South Flank Deposits of the Hamersley Province, Western Austra

Caroline Perring,[1,†] Matt Crowe,[2] and Jon Hronsky[3,4]

[1]BHP Iron Ore, 125 St. Georges Terrace, Perth, Western Australia 6000, Australia

[2]Cassini Resources, 16 Ord Street, West Perth, Western Australia 6005, Australia

[3]Western Mining Services, Suite 26/17 Prowse Street, West Perth, Western Australia 6005, Australi

[4]Centre for Exploration Targeting, School of Earth Science, University of Western Australia, 35 Stirling H
Crawley, Western Australia 6009, Australia

Abstract

The North and South Flank deposits are located on the flanks of the Weeli Wolli anticline at Mining Are
the central Hamersley Province. Supergene martite-goethite mineralization is hosted within the Marra M
Iron Formation and is developed over a strike length of more than 60 km. This multibillion metric ton re
has been drilled out on a 150- × 50- to 50- × 50-m grid, thus providing us with an unprecedented data
analysis. This study synthesizes the drill hole data and presents a physical process model that can acco
the observed distribution of mineralization.

A fluid and mass flux model is proposed which envisages a three-stage process: (1) leaching of Fe
banded iron formation (BIF) in the vadose zone by reduced, acidic, meteoric-derived fluids; (2) penet
of an Fe-rich supergene-fluid plume, driven by gravity and focused by bedding-parallel permeability in
body of ambient alkaline groundwater, effecting nonredox, mimetic replacement of magnetite by hemati
of the gangue minerals (carbonate, silicate, and chert) by goethite coupled with the release of silica into th
phase; and (3) a change from silica leaching to silica deposition on the downdip margins of the system befo
ore-fluid plume is eventually diluted and becomes indistinguishable from the surrounding body of ground

Despite the undoubted secondary role played by structurally enhanced permeability, the primary c
on ore-fluid hydrology is gravity-driven flow along bedding planes. This central observation explains
observed feature of the three-dimensional distribution of martite-goethite mineralization, and the inh
structural architecture simply provides the context for this process to play out. This type of control is
means obvious–the ingress of meteoric fluids during later lateritic weathering of the mineralization do
show this control and produces broadly subhorizontal, bedding-discordant zones of overprinting.

The fundamental control exerted on the distribution of martite-goethite mineralization by bedding

```
2
3  Sentence[147]: "3, pp.627–659 A New Fluid-Flow Model for the Genesis of Banded Iron Formation-Hosted Mart
   Deposits of the Hamersley Province, Western Australia Caroline Perring,1,† Matt Crowe,2 and Jon Hronsky3,
   2Cassini Resources, 16 Ord Street, West Perth, Western Australia 6005, Australia 3Western Mining Services
   Exploration Targeting, School of Earth Science, University of Western Australia, 35 Stirling Highway, Cra
   located on the flanks of the Weeli Wolli anticline at Mining Area C in the central Hamersley Province." -
   Province"/LOCATION, "Western Australia"/LOCATION, "Caroline"/LOCATION, "Iron Ore"/ORE_DEPOSIT, "Perth"/LO
   Australia"/LOCATION, "Australia"/LOCATION, "Perth"/LOCATION, "Western Australia"/LOCATION, "Australia"/LO
   "Australia"/LOCATION, "Weeli Wolli"/LOCATION, "Hamersley Province"/LOCATION]
4
5  Sentence[24]: "Supergene martite-goethite mineralization is hosted within the Marra Mamba Iron Formation
   goethite"/MINERAL, "Marra Mamba Iron Formation"/STRAT]
6
7  Sentence[34]: "This multibillion metric ton resource has been drilled out on a 150- × 50- to 50- × 50-m g
8
9  Sentence[23]: "This study synthesizes the drill hole data and presents a physical process model that can
10
11 Sentence[136]: "A fluid and mass flux model is proposed which envisages a three-stage process: (1) leach
   meteoric-derived fluids; (2) penetration of an Fe-rich supergene-fluid plume, driven by gravity and focus
   effecting nonredox, mimetic replacement of magnetite by hematite and of the gangue minerals (carbonate, s
   and (3) a change from silica leaching to silica deposition on the downdip margins of the system before th
   surrounding body of groundwater." → ["banded iron formation"/ROCK, "magnetite"/MINERAL, "hematite"/MINERA
   "silica"/MINERAL, "silica"/MINERAL, "silica"/MINERAL]
12
13 Sentence[24]: "Despite the undoubted secondary role played by structurally enhanced permeability, the pri
14
15 Sentence[31]: "This central observation explains every observed feature of the three-dimensional distribu
   simply provides the context for this process to play out."
16
17 Sentence[37]: "This type of control is by no means obvious–the ingress of meteoric fluids during later la
   subhorizontal, bedding-discordant zones of overprinting."
18
19 Sentence[42]: "The fundamental control exerted on the distribution of martite-goethite mineralization by
   plume created its own porosity via the relevant ore-forming reactions, and that these were in turn contro
20
```

Figure 3: Structured corpus derived from a paper on Iron deposits, showing text extraction from an academic paper on martite-goethite mineralization, with preprocessing output displayed

Table 5: Performance metrics for NER using Flair on the iron deposits dataset

Class	Precision	Recall	F1-Score	Support
MINERAL	0.788	0.866	0.825	807
ROCK	0.717	0.694	0.706	1030
ORE_DEPOSIT	0.882	0.641	0.742	128
TIMESCALE	0.816	0.912	0.861	68
STRAT	0.824	0.642	0.722	212
LOCATION	0.661	0.434	0.524	477
Micro avg	0.751	0.698	0.724	2722
Macro avg	0.781	0.698	0.730	2722
Weighted avg	0.747	0.698	0.716	2722

Table 6: Performance metrics for NER using Flair on the lithium deposits dataset

Class	Precision	Recall	F1-Score	Support
MINERAL	0.782	0.889	0.832	807
ROCK	0.709	0.688	0.699	1030
ORE_DEPOSIT	0.874	0.648	0.744	128
TIMESCALE	0.863	0.927	0.894	68
STRAT	0.870	0.566	0.686	212
LOCATION	0.580	0.486	0.529	477
Micro avg	0.734	0.707	0.720	2722
Macro avg	0.778	0.701	0.731	2722
Weighted avg	0.7321	0.7069	0.7144	2722

5. Results

This section presents the results from the NER performance evaluation using the Flair framework applied to the OzRock schema. we assess the performance of the model on two geoscience corpora (Iron Deposits and Lithium Deposits). Initially, we detail results of enhancements applied to the workflow (Figures 4 and 5), followed by the assessment of the NER performance metrics (Tables 5 and 6).

5.1 Corpus Construction and Workflow Adjustments

During the developing of the described processes, several challenges emerged, especially when handling PDF documents and text tokenization. As anticipated, tools like PDFPlumber failed to accurately segment the raw text of the papers into sentences on several occasions resulting in errors like missing spaces, misplaced formatting characters, incorrect punctuation, and inconsistent formatting and embedded URLs referring to the journal paper's online version. For instance, concatenated words like "greenstonesequenceconsistsofultramaficandmaficla vasandflowswithinterbeddedsedimentary" (line 9, sentence {12} in Figure 4A) were identified, hindering the readability and accuracy of text parsing. To address these issues, we introduced in our script PyMuPDF (fitz) as a more robust alternative. This change proved effective in better handling various PDF content structures resulting in a cleaner and more accurately structured corpus. Additionally enhanced RegEx operations were applied to correct formatting issues more conservatively. These adjustments in the original workflow (appendix 2) addressed problems such as improper camelCase word splitting, incorrect punctuation, and errors in abbreviations, which are particularly relevant in geoscientific texts where numerical data often coalesce with words (e.g., "100kmnorthwest"). The enhanced workflow (appendix 1) allowed to obtain a better-structured corpora, significantly improving tokenization and the accurate labeling of entities. Figure 5 illustrates the increase in recognized entities across both corpora (Iron and Lithium) after implementing these adjustments.

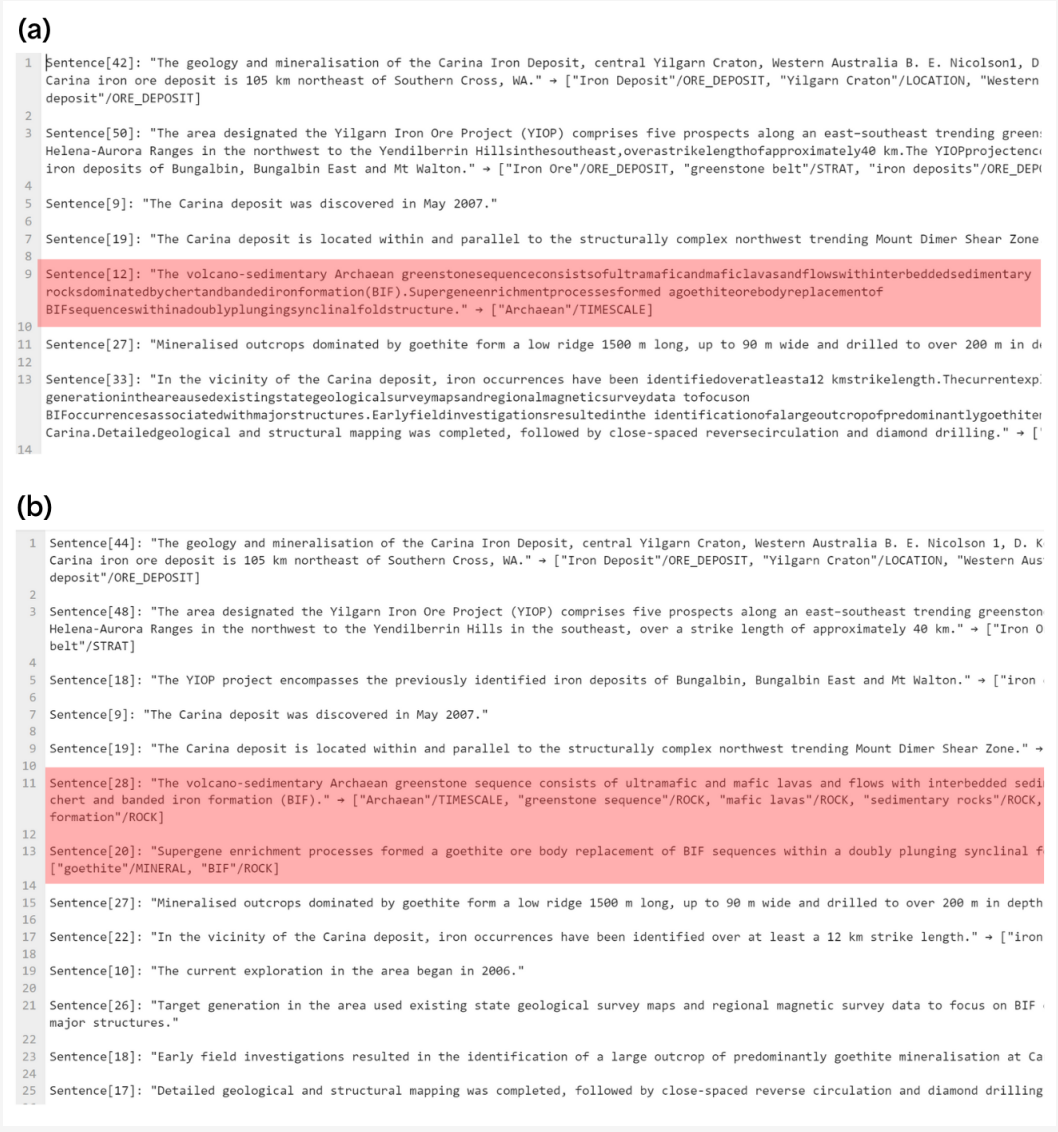

Figure 4: (a): Corpus on iron deposits, highlighting areas where spaces were missed. (b): Sentences and Token Recognition improved in the updated Iron Corpus. Note the correlation between line 9 in A and lines 11 to 13 in B, where the previously identified errors have been rectified

5.1.1 NER performance

The performance of the Flair NER model is assessed using standard NLP metrics such as precision, recall, and F1-score metrics. In the geoscience domain, acceptable F1-scores for geological NER tasks typically range from 0.70 to 0.90 [61], with higher values suggesting better accuracy in recognizing geological entities. This section analyzes the model's overall performance focusing on the macro and micro F1-scores and presents a class-based analysis of the entities recognized across the Iron and Lithium Deposits corpora.

5.1.2 Macro/Micro F1 scores

Slight variations in F1-scores due to class imbalances between the Iron and Lithium corpora, as detailed below, were observed:

Iron Deposits: The dataset exhibits a micro F1-score of 0.7238, which reflects the model's effectiveness at handling the overall dataset, where the influence of larger classes is more pronounced due to their higher number of samples. The macro F1-score of 0.7299 indicates consistent performance across various classes regardless of their frequency. The accuracy score of 0.5837 indicates a moderate overall success rate in classifying and identifying entities.

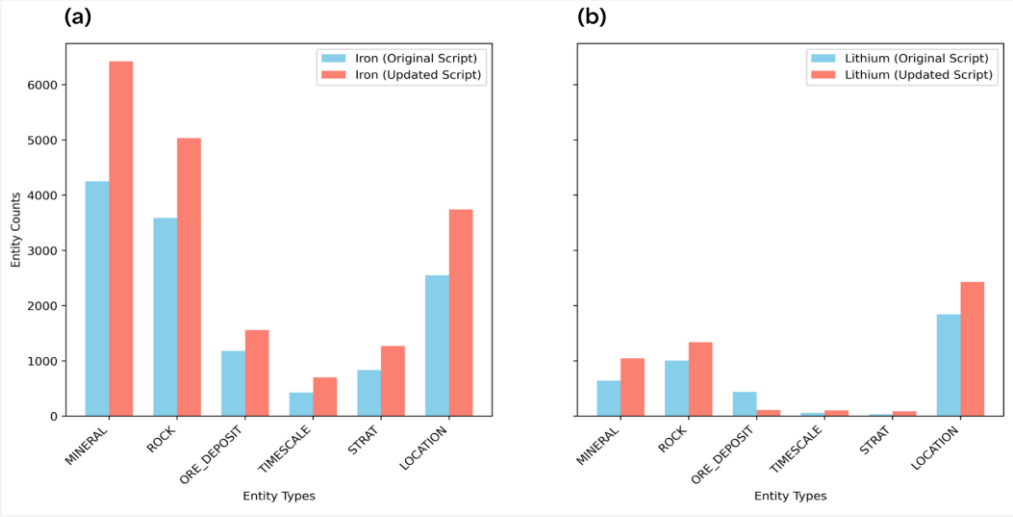

Figure 5: Number of entities by type in the obtained and improved corpora considering the original and updated scripts for both deposits Iron (a) and (b) Lithium dataset

Lithium Deposits: the micro F1-score stands at 0.7199. This score marginally diverges from the performance metrics associated with the Iron Deposits dataset, thereby illustrating the model's capability to manage entities across extensive datasets. Furthermore, the macro F1-score is recorded at 0.7305, subtly implying an enhanced ability of the model to address class imbalances in this specific dataset. Notwithstanding, the model's overall accuracy on this dataset is F1-score = 0.579, indicating a decreased level of model performance.
Tables 5 and 6 provide a detailed breakdown of the performance metrics for each corpus.

5.1.3 Class analysis

Differences in class-specific scores highlighted how geological terms are represented and utilized in each corpus, influenced by the prevalence of terminology and the clarity of contextual usage. The performance of each entity type using F1 score is shown in Table 5 (Iron deposits) and Table 6 (Lithium deposits).

Below some observation for each class:

- *MINERAL:* High F1-scores (0.825 in Iron, 0.832 in Lithium) suggest the model is adept at identifying common mineral-related entities, largely due to clear and frequent annotations in the OzRock dataset.
- *ROCK:* While the performance for this class is strong, the slightly lower F1-scores (0.706 in Iron, 0.699 in Lithium) reflect the complexity of identifying rock-related entities, which can be confused with other geological terms.
- *STRAT:* Although improvements were observed, recall values were lower, indicating that the model struggled to identify all stratigraphic terms, which tend to be more contextually dependent. F1-scores were 0.722 (Iron) and 0.686 (Lithium).
- *TIMESCALE:* This class performed exceptionally well, with F1-scores of 0.861 (Iron) and 0.894 (Lithium), suggesting that the model is proficient at recognizing geological time-related terms.
- *ORE_DEPOSIT:* The regular F1 scores (0.742 in Iron and 0.744) show that the model accurately identified ore deposit entities but likely missed some valid instances, especially in the Lithium dataset.
- *LOCATION:* Despite increased entity counts, LOCATION exhibited the lowest F1-scores (0.524 in Iron, 0.529 in Lithium). This reinforces the need for additional refinements in the dataset, particularly in dealing with the variability of geographical expressions.

5.1.4 Quality of corpora

The overall improvement on quality of each corpus (Iron and Lithium deposits) using the improved workflow are shown in Figure 5. This quality is quantified by the increase in the number of recognized entities across different categories:

- *ROCK:* The updated workflow resulted in a significant increase in the number of entities recognized in both datasets. For the Iron Deposits dataset, the entity count for ROCK rose by 40.2%, while in the Lithium Deposits dataset, the increase was 32.9%

- *MINERAL:* The enhanced text extraction incorporated in the updated workflow led to a substantial increase in recognized entities for the MINERAL class. In the Iron Deposits corpus, the number of MINERAL entities increased by 51 %, and in the Lithium Deposits corpus, the increase was 62.5%.
- *STRAT:* There was a noticeable improvement in the recognition of STRAT entities, particularly in the Lithium Deposits dataset, where the percentage increase is 161.8%, while for the iron dataset this value is of 51.7%.
- *TIMESCALE:* The TIMESCALE class showed a marked increase in both datasets, particularly for Iron Deposits, where the number of recognized entities grew by 64.6%. In the Lithium Deposits corpus, the improvement was notable as well, with an increase of 71.7%.
- *ORE_DEPOSIT:* While the updated script led to an increase in recognized entities for ORE_DEPOSIT in the Iron Deposits dataset (32.1%), the opposite effect was observed in the Lithium Deposits dataset, where the number of entities decreased by 74.4%.
- *LOCATION:* The number of recognized LOCATION entities increased in both datasets. In the Iron Deposits corpus, the count grew by 46.8%, and in the Lithium Deposits corpus, the increase was 31.9%.

6. Discussion

Effective data preprocessing is a cornerstone of NLP applications where, transforming unstructured textual data into a machine-readable corpus is essential. This involves processes such as text cleaning and text mining to ensure clarity and coherence, which directly impacts the usability of data for computational analysis [52] and [68].

6.1 Addressing Preprocessing Challenges

Errors identified in the original workflow prompted a series of adjustments, as detailed in Table 4, which compares differences in functionality and tools used between the original and modified scripts for corpus extraction. The transition from pdfplumber to PyMuPDF (fitz) enhanced the quality of text extraction, resulting in cleaner and more accurately segmented corpora. The addition of Regex operations resolved specific formatting issues, such as improperly merged camelCase words, ensuring spaces followed punctuation and separated digits from text. These enhancements were particularly useful in geoscientific texts where numerical data often merges with words (e.g., "100kmnorthwest"). The updated script (Appendix 2) contributed to better tokenization and more accurate entity recognition in the structured corpora.

6.2 Impact of Preprocessing Adjustments on Entity Recognition

Flair NER model demonstrated consistent performance across the Iron Deposits and Lithium mineral deposits corpora with noticeable improvements resulting from the preprocessing adjustments. The impact of these adjustments is illustrated in Figure 4 which presents the entity counts before and after workflow modifications. For example, the MINERAL class in both corpora shows an increase of more than 50% in recognized entities after applying the modified workflow. This increase highlights the importance of robust text extraction and tokenization in improving NER accuracy, especially in complex domains like geoscience. These results support findings that emphasize that geoscience text data must be annotated specifically to be effectively mined [69]. Standardized annotations are essential, as different researchers often use varying terms for the same geological features.

6.3 Error Analysis and Limitations of the NER Model

While NER performance improvements were noted across multiple classes, the analysis revealed challenges in handling specific geological terms, particularly context-dependent entities, ambiguous terminology, and domain-specific variations. These challenges are critical in geological NLP applications, where terms usually have different meanings depending on their surrounding context, regional naming conventions, or specific subfields within geology. The following subsections outline areas where the model underperformed highlighting class imbalances, complex terminology, potential biases in entity annotation, and discuss future directions for improvement.

6.3.1 Impact of class imbalances and specialized terminology

- The micro and macro F1-scores from the Iron and Lithium Deposits datasets reveal class imbalance effects, with frequent entity types (e.g., MINERAL, ROCK) achieving higher recognition rates, while less common entities (e.g., STRAT, TIMESCALE) exhibited lower performance.
- In the Iron Deposits dataset, the macro F1-score suggests overall robustness, with high recognition accuracy for dominant classes. However, less frequent entity types had lower F1-scores, indicating a need for additional training data or improved annotation strategies for these specialized terms.

- In the Lithium Deposits dataset, the slightly lower micro F1-score suggests challenges in handling entity diversity across a smaller corpus. The macro F1-score indicates that the model compensated for class imbalances, yet lower accuracy values reinforce the need for more refined annotation techniques, particularly for rare or highly specialized geological terms.

These results are in line with research emphasizing the need to annotate NER models for geology to handle low-frequency entities effectively [67] and [69]. Additionally, geographical entities (LOCATION class) recorded the lowest F1-scores across both datasets, reinforcing prior observations that geographical terms in geological literature are highly inconsistent. The model frequently misclassified geological formations (e.g., Pilbara Craton, Greenbushes Pegmatite) as administrative locations, underscoring the need for enhanced entity disambiguation methods.

6.3.2 Evaluation of entity recognition performance

The updated preprocessing workflow resulted in notable improvements across multiple entity types, as demonstrated in Figure 5.

- ROCK: Recognition improved notably, especially in the Iron Deposits dataset, aligning with higher precision and F1-scores, demonstrating that text extraction refinements enhanced entity identification.
- MINERAL: The updated script significantly improved recognition, particularly in the Iron Deposits corpus, where more consistent annotations in the OzRock dataset likely facilitated higher model accuracy.
- STRAT: Recognition increased modestly, with lower F1-scores, suggesting that stratigraphic terms may require annotation refinements to improve model accuracy.
- TIMESCALE: The most improved class, with substantial recognition increases, particularly in Iron Deposits, aligning with global geological time standards [67].
- ORE_DEPOSIT: Performance varied by dataset increased in Iron Deposits but declined in Lithium Deposits, likely due to differences in how deposits are described in literature.

LOCATION: Despite increased entity recognition, F1-scores remained low, reflecting challenges in handling geographical expressions, abbreviations, and regional naming variations as previously documented by previous studies [70].

6.3.3 Challenges in Recognizing Context-Dependent and Ambiguous Terms

During the analysis we confirmed that geological texts frequently contain terms with multiple interpretations, making NER in this domain particularly challenging. Specific examples include:

- *"Banded Iron Formation" (BIF):* While a recognized rock unit, the term "formation" may also refer to a general stratigraphic category, causing misclassification issues.
- *"Lithium-bearing minerals":* A general descriptor for lithium-enriched materials, whereas spodumene and lepidolite are specific mineral species requiring distinct categorization.
- *"100 km northwest":* Commonly appears as a spatial descriptor but is often embedded in geological feature descriptions, leading to parsing difficulties.

These results indicate that Flair model's BiLSTM-CRF architecture relies on contextual embeddings and its performance decrease when encountering rare, domain-specific terms not well-represented in training data.

6.3.4 Limitations of the Schema and Potential Biases

This study applied the OzRock Evaluation Set as a reference schema for NER training and evaluation. While OzRock provides a structured taxonomy for geological terms, certain limitations and biases must be acknowledged. OzRock, while comprehensive, does not fully capture all geological subfields, leading to potential entity recognition inconsistencies. Examples include:

- Geological formations named after regions may be misclassified as locations (e.g., Hamersley Group interpreted as a geographical entity rather than a geological formation).
- Ambiguous stratigraphic units can cause annotation inconsistencies, especially when terminologies vary between studies.
- Granularity limitations in mineral classification may hinder recognition of less common lithium-bearing minerals, beyond those widely documented in existing datasets.

Similar challenges have been observed in previous studies when it comes to achieving high annotation consistency in domain-specific corpora [33].

It is important to clarify that this study did not perform independent expert validation of OzRock annotations, as the focus was on demonstrating corpus construction methodologies, not modifying existing geoscientific schemas. As detailed in Section 4.1 (Dataset Selection and Collection), the OzRock schema was sourced from a public GitHub repository and was already annotated by domain experts. Future work could incorporate additional expert-reviewed annotations to enhance model reliability across diverse geological subfields.

6.4 Computational Efficiency

In addition to improved NER, the updated pipeline significantly improved computational efficiency. By introducing PyMuPDF for text extraction, the preprocessing process became more efficient. The cleaning steps enhanced data quality by removing extraneous elements such as formatting artifacts and embedded URLs. This approach, supported using an optimization of RegEx for more accurate tokenization, ensured the prepared text was better prepared to computational analysis. The use of GPU cluster has significantly improved the time of the processes. For instance, it took around 35 minutes to process the 20 papers on iron deposits (averaging 15 pages per paper), using a GPU in contrasts with the 3.5 hours required when using a CPU. This significant reduction in processing time highlights the advantages of utilizing high-performance computing resources especially in a data-intensive field such as geoscience in line with previous efforts [67].

The choice of Flair over other tools for preprocessing tasks like SpaCy or Prodigy, was based on its robust performance in handling domain-specific NER tasks, particularly in English and due to its focus on advanced neural network architectures. Flair utilizes advanced neural network architectures with contextual embeddings using a BiLSTM-CRF model, which is effective for specialized tasks [71]. Further support for this choice is provided by experiments conducted in other academic contexts, where Flair embedding models and medium-sized corpora were employed, obtaining optimal results [72].

6.5 Potential Applications and Adaptability

The creation of structured corpora extends beyond academic exercises, directly impacting practical applications in geoscience including mineral exploration, geological modeling, and KGs construction. By providing a semantic foundation for improved data representation and analysis, structured corpora enable more accurate geological modeling and decision-making. This facilitates informed strategies in mineral exploration, where precise data is crucial for locating and evaluating potential mining sites [5][40] and [42]. Furthermore, well-constructed corpora can significantly advance knowledge discovery and management across different geological contexts by facilitating the development of geoscience KGs [14] and [68]. To further improve the utility and accuracy of structured corpora in geoscience, future research should incorporate advanced NLP technologies used in BERT, GPT, and ELMo. The adaptation of specialized models like GeoBERT and SciBERT is anticipated to refine entity recognition, particularly in handling complex and context-dependent geological entities. These enhancements will enable more sophisticated applications, including dynamic geological mapping and real-time data analysis, which are pivotal in high-stakes industries such as oil and gas exploration and environmental monitoring [67].

Moreover, maintaining high-quality annotated corpora remains a challenge, especially when integrating data from diverse sources such as scientific papers, technical reports, and exploration datasets. Overcoming data alignment and consistency challenges, especially from various sources, is crucial for maintaining data quality ensuring that structured corpora accurately reflect geological knowledge. The development of high-precision retrieval systems [73] and standardized annotation protocols [74] is becoming increasingly important in addressing the challenges mentioned before. Furthermore, integrating semantic reasoning into geological modelling will further reduce exploration risks and support decision-making in mineral resource exploration. Further assessments of the NER model's performance through comparisons with publicly available geological datasets are essential. These evaluations will establish benchmarks for the adaptability of NER models to domain-specific corpora in geosciences, thereby advancing AI-driven geoscientific text mining, enhancing exploration workflows, and improving knowledge discovery in mineral systems.

6.6 Future Work

Future research should focus on computational efficiency, evaluating the pipeline’s performance as it scales to larger and more complex datasets. This future investigation will clarify the computational demands and help optimize the pipeline for efficiency, particularly using high-performance computing resources. Additionally, exploring the integration of multimodal data will enhance the contextual richness and applicability of the corpora in real-world scenarios.

As a continuation of this research, the authors are exploring the integration of feedback mechanisms and case studies to provide insights into the practical applicability and areas for enhancement of the developed methodologies [74]. Such approaches would help validate the relevance and effectiveness of the NER models and and domain-specific schemas in diverse geoscience applications, establishing benchmarks for their adaptability.

This study does not imply that the constructed corpora are superior to existing external datasets. Instead, it aims to demonstrate the potential of an NLP-driven workflow to enhance corpus quality through structured preprocessing techniques, facilitate the effective application of NER in geoscience, and provide a reproducible methodology for corpus creation from geoscientific literature. The framework established here enables researchers to develop flexible, domain-specific corpora, scalable across various geosciences disciplines.

7. Conclusions

The research highlights the importance of preprocessing and quality control techniques in creating accurate NLP datasets for complex fields like geosciences. Managing data quality through meticulous preprocessing, including cleaning and removing noise (irrelevant characters and others), is crucial for improving the accuracy and reliability of information extraction. It is demonstrated that tailored strategies are necessary to extract valuable information efficiently, and identifying specific issues during text extraction is essential for robust preprocessing. For the last, expert review and validation are crucial as it is needed to validate a structured corpus that accurately represents the content of geoscientific papers, highlighting the collaborative interplay between domain experts and computational methods.

The observed performance variations across different classes of the OzRock schema when running Flair underscore the need for domain-specific model training and dataset-specific fine-tuning to address underperformance in critical categories. The analysis indicates that while the Flair NER model performs well on frequently used and well-labelled entities, it encounters difficulties with more specialized or context-dependent terms like and further refinements in annotation strategies and the dataset structure could improve performance in these areas. Enhanced recognition of certain categories in both datasets (Iron and Lithium deposits papers) illustrates the model's capacity to effectively capture and classify terms integral to understanding geological processes and resource characterization. A detailed error analysis for each class could demonstrate specific challenges and guide targeted improvements in data annotation or NER model refinement. This focused approach could further optimize the performance of NLP applications in geoscience, ensuring that the insights derived from automated text analysis are accurate and actionable.

The findings suggest that refining the NER model and enhancing the preprocessing pipeline to identify better and classify geological terms proved crucial in refining the quality of data extracted from scholarly articles in PDF format. By refining text preprocessing methodologies, the input fed into the NER model aligns more closely with conventional written language, thereby reducing inaccuracies and increasing the dependability of entity identification. This research also emphasizes advancements in NER capabilities within the geoscience domain and the role of sophisticated NLP tools like Flair in improving entity recognition and classification. Future research might explore the application of geological datasets in different formats or the integration of advanced NLP model architectures, such as those incorporating LLMs, to potentially rectify the precision and recall disparities observed across different classes. This includes integrating more advanced machine learning models to handle the complexity of geoscience data, aiming to improve the precision and efficiency of data processing and analysis in geosciences.

Acknowledgements

The authors gratefully acknowledge funding from the Commonwealth Scientific and Industrial Research Organisation (CSIRO) through the ResearchPlus Science Leader program. We also thank our CSIRO colleagues for their contributions to this research, with special thanks to Sarvnaz Karimi and John Hille for reviewing the entire manuscript and providing valuable feedback. Klaus Gessner publishes with permission from the Executive Director, Geological Survey of Western Australia. Moreover, we appreciate the constructive feedback provided by the anonymous reviewers, which helped enhance the clarity and overall quality of the manuscript.

References

[1] Ambika, P., (2020). Machine Learning and Deep Learning Algorithms on the Industrial Internet of Things (IIoT). *Advances in Computers*, Vol. 117(1), 321-338. https://doi.org/10.1016/bs.adcom.2019.10.007.

[2] Chen, L., Wang, L., Miao, J., Gao, H., Zhang, Y., Yao, Y., Bai, M., Mei, L. and He, J., (2020). Review of the Application of Big Data and Artificial Intelligence in Geology. *Journal of Physics:* Conference Series Vol. 1684(1). https://doi.org/10.1088/1742-6596/1684/1/012007.

[3] Lindsay, M. D., Aillères, L., Jessell, M. W., de Kemp, E. A., and Betts, P. G. (2012). Locating and Quantifying Geological Uncertainty in Three-dimensional Models: Analysis of the Gippsland Basin, Southeastern Australia. *Tectonophysics*, Vol. 546(1), 10-27. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0040195112002004. [Accessed: Jan. 06, 2025]

[4] Lv, X., Xie, Z., Xu, D., Jin, X., Ma, K., Tao, L., Qiu, Q. and Pan, Y., (2022). Chinese Named Entity Recognition in the Geoscience Domain Based on BERT. *Earth and Space Science*, Vol. 9(3), https://doi.org/10.1029/2021EA002166.

[5] Ailleres, L., Jessell, M., de Kemp, E., Caumon, G., Wellmann, F., Grose, L., Armit, R., Lindsay, M., Giraud, J., Brodaric, B., Harrison, M. T. and Courrioux, G. (2019). Loop-Enabling 3D Stochastic Geological Modelling. *ASEG Extended Abstracts*, Vol. 2019(1), 1-3, https://doi.org/10.1080/22020586.2019.12072955.

[6] Breunig, M., Bradley, P. E., Jahn, M., Kuper, P., Mazroob, N., Rösch, N., Al-Doori, M., Stefanakis, E. and Jadidi, M., (2020). Geospatial Data Management Research: Progress and Future Directions. ISPRS *International Journal of Geo-Information*, Vol. 9(2). https://doi.org/10.3390/ijgi9020095.

[7] Zhang, S., Xu, H., Jia, Y., Wen, Y., Wang, D., Fu, L., Wang, X. and Zhou, C., (2023). GeoDeepShovel: A Platform for Building Scientific Database from Geoscience Literature with AI Assistance. *Geoscience Data Journal*, Vol. 10(4), 519-537. https://doi.org/10.1002/gdj3.186.

[8] Biber, D., Conrad, S. and Reppen, R., (1998). *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge University Press, Accessed: Jan. 06, 2025. [Online]. Available: https://academic.oup.com/dsh/article-abstract/14/2/305/936240. [Accessed: Feb. 02, 2025]

[9] Yulianto, A. A., (2019). Extract Transform load (ETL) Process in Distributed Database Academic Data Warehouse. APTIKOM *Journal on Computer Science and Information Technologies*, Vol. 4(2), 61-68. https://doi.org/10.11591/APTIKOM.J.CSIT.36

[10] Pinto da Costa, J. F. and Cabral, M., (2022). Statistical Methods with Applications in Data Mining: A Review of the Most Recent Works. *Mathematics*, Vol. 10(6). https://doi.org/10.3390/math10060993.

[11] Jianping, C., Jie, X., Qiao, H. U., Wei, Y., Zili, L., Bin, H. U. and Wei, W., (2016). Quantitative Geoscience and Geological Big Data Development: A Review. *Acta Geologica Sinica-English Edition,* Vol. 90(4), 1490-1515. https://doi.org/10.1111/1755-6724.12782.

[12] Salloum, S. A., Al-Emran, M., Monem, A. A. and Shaalan, K., (2018). Using Text Mining Techniques for Extracting Information from Research Articles. *Intelligent Natural Language Processing: Trends and Applications*, Vol. 2018(1), 373-397. https://doi.org/10.1007/978-3-319-67056-0_18.

[13] Wang, R., Wang, M., Liu, J., Cochez, M. and Decker, S., (2020). Structured Query Construction Via Knowledge Graph Embedding. *Knowledge and Information Systems*, Vol. 62(1), 1819-1846. https://doi.org/10.1007/s10115-019-01401-x.

[14] Zhou, C., Wang, H., Wang, C., Hou, Z., Zheng, Z., Shen, S., Cheng, Q., Feng, Z., Wang, X., Lv, H., Fan, J., Hu, X., Hou, M. and Zhu, Y., (2021). Geoscience Knowledge Graph in the Big Data Era. *Science China Earth Sciences*, Vol. 64(7),1105-1114. https://doi.org/10.1007/s11430-020-9750-4.

[15] Portisch, J., Hladik, M. and Paulheim, H., (2024). Background Knowledge in Ontology Matching: A Survey. *Semantic Web*, Vol. 15(6), 2639-2693. https://doi.org/10.3233/SW-223085.

[16] Chi, N. W., Jin, Y. H. and Hsieh, S. H., (2019). Developing Base Domain Ontology from a Reference Collection to Aid Information Retrieval. *Automation in Construction*, Vol. 100(1), 180-189. https://doi.org/10.1016/j.autcon.2019.01.001.

[17] Zha, Y., Ke, Y., Hu, X. and Xiong, C., (2024). Ontology Attention Layer for Medical Named Entity Recognition. *Applied Sciences*, Vol. 14(1). https://doi.org/10.3390/app14010421.

[18] Liang, H., Zhou, Y., Wang, Y., Xu, X., Wei, Y. and Chen, Y., (2022). Named Entity Recognition of Diseases and Pests with Small Samples Based on Space Mapping. *2022 Euro-Asia Conference on Frontiers of Computer Science and Information Technology (FCSIT)* Vol. 2022(1), 64-72. https://doi.org/10.1109/FCSIT57414.2022.00025.

[19] Zhou, H., Ning, S., Liu, Z., Lang, C., Liu, Z. and Lei, B., (2020). Knowledge-Enhanced Biomedical Named Entity Recognition and Normalization: Application to Proteins and Genes. *BMC Bioinformatics*, Vol. 21(1), 1-15. https://doi.org/10.1186/s12859-020-3375-3.

[20] Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S. R., Das, D. and Pavlick, E., (2019). What do you Learn from Context? Probing for Sentence Structure in Contextualized Word Representations. arXiv preprint *arXiv*. https://doi.org/10.48550/arXiv.1905.06316.

[21] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R. and Le, Q. V., (2019). Xlnet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in Neural Information Processing Systems*, 1-11. https://proceedings.neurips.cc/paper_files/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf.

[22] Lawley, C. J., Raimondo, S., Chen, T., Brin, L., Zakharov, A., Kur, D., Hui, J., Newton, G., Burgoyne, S. L., and Marquis, G., (2022). Geoscience Language Models and their Intrinsic Evaluation. *Applied Computing and Geosciences*, Vol. 14(1). https://doi.org/10.1016/j.acags.2022.100084.

[23] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G. and Lample, G., (2023). Llama: Open and Efficient Foundation Language Models. *arXiv preprint arXiv*. https://doi.org/10.48550/arXiv.2302.13971.

[24] Gómez, C., White, J. C. and Wulder, M. A., (2016). Optical Remotely Sensed Time Series Data for Land Cover Classification: A Review. *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 116(1), 55-7. https://doi.org/10.1016/j.isprsjprs.2016.03.008.

[25] Cresson, R., (2018). A Framework for Remote Sensing Images Processing Using Deep Learning Techniques. *IEEE Geoscience and Remote Sensing Letters*, Vol. 16(1), 25-29. https://doi.org/10.1109/LGRS.2018.2867949.

[26] Gao, S., (2021). *Geospatial Artificial Intelligence (GeoAI)*, New York: Oxford University Press.

[27] Bhattacharjee, B., Trivedi, A., Muraoka, M., Ramasubramanian, M., Udagawa, T., Gurung, I., Pantha, N., Zhang, R., Dandala, B., Ramachandran, R., Maskey, M., Bugbee, K., Little, M., Fancher, E., Gerasimov, I., Mehrabian, A., Sanders, L., Costes, S., Blanco-Cuaresma, S., Lockhart, K., Allen, T., Grezes, F., Ansdell, M., Accomazzi, A., El-Kurdi, Y., Wertheimer, D., Pfitzmann, B., Berrospi, C., Dolfi, M., Teixeira de Lima, R., Vagenas, P., Mukkavilli, S. K., Staar, P., Vahidinia, S., McGranaghan, R. and Lee, T. (2024). INDUS: Effective and Efficient Language Models for Scientific Applications. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Vol. 2021(1), 98-112. https://doi.org/10.48550/arXiv.2405.10725.

[28] Craglia, M., Hradec, J., Nativi, S. and Santoro, M., (2017). Exploring the Depths of the Global Earth Observation System of Systems. *Big Earth Data*, Vol. 1(2), 21-46. https://doi.org/10.1080/20964471.2017.1401284.

[29] Qiu, Q., Xie, Z., Wu, L. and Tao, L., (2020). Dictionary-Based Automated Information Extraction from Geological Documents Using a Deep Learning Algorithm. *Earth and Space Science*, Vol. 7(3). https://doi.org/10.1029/2019EA000993.

[30] Chu, D., Wan, B., Li, H., Dong, S., Fu, J., Liu, Y., Huang, K. and Liu, H., (2022). A Machine Learning Approach to Extracting Spatial Information from Geological Texts in Chinese. *International Journal of Geographical Information Science*, Vol. 36(11), 2169-2193. https://doi.org/10.1080/13658816.2022.2087224.

[31] Qiu, Q., Xie, Z., Wu, L., Tao, L. and Li, W., (2019). BiLSTM-CRF for Geological Named Entity Recognition from the Geoscience Literature. *Earth Science Informatics*, Vol. 12, 565-579. https://doi.org/10.1007/s12145-019-00390-3.

[32] Hand, D. J., (2012). Assessing the Performance of Classification Methods. *International Statistical Review*, Vol. 80(3), 400-414. https://doi.org/10.1111/j.1751-5823.2012.00183.x.

[33] Li, J., Wei, Q., Ghiasvand, O., Chen, M., Lobanov, V., Weng, C. and Xu, H., (2021). Study of Pre-Trained Language Models for Named Entity Recognition in Clinical trial Eligibility Criteria from Multiple Corpora. *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, Vol. 2021(1), 511-512. https://doi.org/10.1109/ICHI52183.2021.00095.

[34] Weischedel, R., Hovy, E., Marcus, M., Palmer, M., Belvin, R., Ramshaw, Pradhan, S., Ramshaw, L. and Xue, N., (2011). Ontonotes: A Large Training Corpus for Enhanced Processing. Joseph Olive, Caitlin Christianson, and John McCary, editors, *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. https://www.cs.cmu.edu/~hovy/papers/09OntoNotes-GALEbook.pdf.

[35] Kyle, K., (2021). Natural Language Processing for Learner Corpus Research. *International Journal of Learner Corpus Research*, Vol. 7(1), 1-16. https://doi.org/10.1075/ijlcr.7.1?locatt=mode:legacy.

[36] Fan, R., Wang, L., Yan, J., Song, W., Zhu, Y. and Chen, X., (2019). Deep Learning-Based Named Entity Recognition and Knowledge Graph Construction for Geological Hazards. *ISPRS International Journal of Geo-Information*, Vol. 9(1). https://doi.org/10.3390/ijgi9010015.

[37] Sobhana, N., Mitra, P. and Ghosh, S. K., (2010). Conditional Random Field Based Named Entity Recognition in Geological Text[J]. *International Journal of Computer Applications*, Vol. 1(3), 143–147. https://doi.org/10.5120/72-166.

[38] Leveling, J., (2015). Tagging of Temporal Expressions and Geological Features in Scientific Articles. *Proceedings of the 9th Workshop on Geographic Information Retrieval* Vol. 2015(1), 1-10. https://doi.org/10.1145/2837689.2837701.

[39] Guo, Z., Wang, C., Zhou, J., Zheng, G., Wang, X. and Zhou, C., (2024). GeoKnowledgeFusion: A Platform for Multimodal Data Compilation from Geoscience. *Literature. Remote Sensing*, Vol. 16(9). https://doi.org/10.3390/rs16091484.

[40] Enkhsaikhan, M., (2021). *Geological Knowledge Graph Construction from Mineral Exploration Text*. Doctoral Dissertation. UWA (University of Western Australia).

[41] Liu, H., Qiu, Q., Wu, L., Li, W., Wang, B. and Zhou, Y., (2022). Few-Shot Learning for Name Entity Recognition in Geological Text Based on GeoBERT. *Earth Science Informatics*, Vol. 15(2), 979-991. https://doi.org/10.1007/s12145-022-00775-x.

[42] Wang, C., Chen, J. and Li, Y., (2022). Named Entity Annotation Schema for Geological Literature Mining in the Domain of Porphyry Copper Deposits. *AGU Fall Meeting Abstracts*, Vol. 2022(1). https://doi.org/10.1016/j.oregeorev.2022.105243.

[43] Ma, K., Zheng, S., Tian, M., Qiu, Q., Tan, Y., Hu, X., Li, H. Y. and Xie, Z., (2023). CnGeoPLM: Contextual Knowledge Selection and Embedding with Pretrained Language Representation Model for the Geoscience Domain. *Earth Science Informatics*, Vol. 16(4), 3629-3646. https://doi.org/10.1007/s12145-023-01112-6.

[44] Maskey, M., Ramachandran, R. and Miller, J., (2017). Deep Learning for Phenomena-Based Classification of Earth Science Images. *Journal of Applied Remote Sensing*, Vol. 11(4). https://doi.org/10.1117/1.JRS.11.042608.

[45] Rouhou, A. C., Dhiaf, M., Kessentini, Y. and Salem, S. B., (2021). Transformer-Based Approach for Joint Handwriting and Named Entity Recognition in Historical Document[J]. *Pattern Recognition Letters*. https://doi.org/10.1016/j.patrec.2021.11.010

[46] Santoso, J., Setiawan, E. I., Purwanto, C. N., Yuniarno, E. M., Hariadi, M. and Purnomo, M. H., (2021). Named entity Recognition for Extracting Concept in Ontology Building on Indonesian Language Using End-To-End Bidirectional Long Short-Term Memory[J]. *Expert Systems with Applications*, Vol. 176. https://doi.org/10.1016/j.eswa.2021.114856.

[47] Ma, X. and Fox, P., (2013) Recent Progress on Geologic Time Ontologies and Considerations for Future Works. *Earth Science Informatics*, Vol. 6(1), 31-46. https://doi.org/10.1007/s12145-013-0110-x.

[48] Zhong, J., Aydina, A. and McGuinness, D. L., (2009). Ontology of Fractures. *Journal of Structural Geology*, Vol. 31(3), 251-259. https://doi.org/10.1016/j.jsg.2009.01.008.

[49] Babaie, H. A. and Davarpanah, A., (2018) Semantic Modelling of Plastic Deformation of Polycrystalline Rock. *Computer Geoscience*, Vol. 111, 213–222. https://doi.org/10.1016/j.cageo.2017.11.002.

[50] Tessarollo, A. and Rademaker, A., (2020). Inclusion of Lithological Terms (Rocks and Minerals) in The Open Wordnet for English. *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)* Vol. 2020(1), 33-38. https://aclanthology.org/2020.mmw-1.6.pdf.

[51] Villacorta, S. P. and Lindsay, M., (2023). Exploring the Importance of Preprocessing Operations in Geoscience Knowledge Graphs through the Application of a Machine Learning Approach. *Proceedings of the 26th World Mining Congress*, Vol. 2023(1), 177–188. http://hdl.handle.net/102.100.100/487308?index=1.

[52] Bird, S., Klein, E. and Loper, E., (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.* O'Reilly Media, Inc.

[53] Ahmed, N., Amin, R., Aldabbas, H., Saeed, M., Bilal, M. and Song, H., (2024). A Novel Approach for Sentiment Analysis of a Low Resource Language Using Deep Learning Models. *ACM Transactions on Asian and Low-Resource Language Information Processing.* https://doi.org/10.1145/3696789.

[53] Frohmann, M., Sterner, I., Vulić, I., Minixhofer, B. and Schedl, M., (2024). Segment Any Text: A Universal Approach for Robust, Efficient and Adaptable Sentence Segmentation. *arXiv preprint arXiv.* https://doi.org/10.48550/arXiv.2406.16678.

[54] Qi, P., Zhang, Y., Zhang, Y., Bolton, J. and Manning, C. D., (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Vol. 2020(1), 101–108. https://doi.org/10.48550/arXiv.2003.07082.

[55] Hu, Z., Ma, X., Liu, Z., Hovy, E. and Xing, E., (2016). Harnessing Deep Neural Networks with Logic Rules. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, Vol. 4, 2410–2420. https://doi.org/10.48550/arXiv.1603.06318.

[56] Singer-Vine, J., (2024). 'pdfplumber', 2024, 0.11.0. Accessed: Jan. 07, 2025. [Online]. Available: https://pypi.org/project/pdfplumber/

[57] Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S. and Vollgraf, R., (2019). FLAIR: An Easy-to-use Framework for State-of-the-art NLP. *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics (demonstrations)*, 54–59. https://doi.org/10.18653/v1/N19-4010.

[58] Honnibal, M., Van Landeghem, S. and Boy, A., (2020). *'spaCy: Industrial-strength Natural Language Processing in Python*. https://doi.org/10.5281/zenodo.1212303.

[59] Montani, I. and Honnibal, M., (2017). *Prodigy: A New Tool for Radically Efficient Machine Teaching*. Available: https://prodi.gy/.

[60] Angerer, T., Duuring, P., Hagemann, S. G., Thorne, W. and McCuaig, T. C., (2015). A Mineral System Approach to Iron Ore in Archaean and Palaeoproterozoic BIF of Western Australia. *Geological Society, London, Special Publications*, Vol. 393(1), 81–115. https://doi.org/10.1144/SP393.11.

[61] Greim, P., Solomon, A. A. and Breyer, C., (2020). Assessment of Lithium Criticality in the Global Energy Transition and Addressing Policy Gaps in Transportation. *Nature Communications*, Vol. 11(1). https://doi.org/10.1038/s41467-020-18402-y.

[62] Sang, E. F. and De Meulder, F., (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 142–147. https://aclanthology.org/W03-0419.

[63] Ramshaw, L. A. and Marcus, M. P., (1995). Text Chunking Using Transformation-Based Learning. *ACL Third Workshop on Very Large Corpora*, 82-94. https://aclanthology.org/W95-0107.pdf.

[64] Nguyen, Q. H., Ly, H. B., Ho, L. S., Al-Ansari, N., Le, H. V., Tran, V. Q., Prakash, I. and Pham, B. T., (2021). Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil. *Mathematical Problems in Engineering*, Vol. 2021(1). https://doi.org/10.1155/2021/4832864

[65] Cheirmpos, G., Tabatabaei, S. A., Kanoulas, E. and Tsatsaronis, G., (2023). Benchmarking Named Entity Recognition Approaches for Extracting Research Infrastructure Information from Text. *International Conference on Machine Learning, Optimization, and Data Science* Vol. 2023(1), 131-141. https://doi.org/10.1007/978-3-031-53969-5_11.

[66] Ning, D., Xu, G., Chen, Y., Wang, X., Han, X., Xie, P., Zheng, H, and Liu, Z., (2021). Few-Nerd: A Few-Shot Named Entity Recognition Dataset. *arXiv preprint arXiv:2105.07464.* https://doi.org/10.48550/arXiv.2105.07464.

[67] Qiu, Q., Tian, M., Huang, Z., Xie, Z., Ma, K., Tao, L. and Xu, D., (2024). Chinese Engineering Geological Named Entity Recognition by Fusing Multi-Features and Data Enhancement Using Deep Learning. *Expert Systems with Applications*, Vol. 238(1). https://doi.org/10.1016/j.eswa.2023.121925.

[68] Ryen, V., Soylu, A. and Roman, D., (2022). Building Semantic Knowledge Graphs from (semi-) Structured Data: A Review. *Future Internet*, Vol. 14(5). https://doi.org/10.3390/fi14050129.

[69] Qiu, Q., Tian, M., Xie, Z., Tan, Y., Ma, K., Wang, Q., Pan, S. and Tao, L., (2023). Extracting Named Entity Using Entity Labeling in Geological Text Using Deep Learning Approach. *Journal of Earth Science*, Vol. 34(5), 1406-1417. https://doi.org/10.1007/s12583-022-1789-8.

[70] Tao, L., Xie, Z., Xu, D., Ma, K., Qiu, Q., Pan, S. and Huang, B., (2022). Geographic Named Entity Recognition by Employing Natural Language Processing and an Improved BERT Model. *ISPRS International Journal of Geo-Information*, Vol. 11(12). https://doi.org/10.3390/ijgi11120598.

[71] Vychegzhanin, S. and Kotelnikov, E., (2019). Comparison of Named Entity Recognition Tools Applied to News Articles. *2019 Ivannikov Ispras Open Conference (ISPRAS)* Vol. 2019(1). 72-77. https://doi.org/10.1109/ISPRAS47671.2019.00017.

[72] Smirnova, N. and Mayr, P., (2023). Embedding Models for Supervised Automatic Extraction and Classification of Named Entities in Scientific Acknowledgements. *Scientometrics*, Vol. 2023(1), 1-25. https://doi.org/10.1007/s11192-023-04806-2.

[73] Gayathri, R., Gobinath, T., Muthumari, A. and Swathi, R. S. V., (2023). Enhanced AI Based Feature Extraction Technique in Multimedia Image Retrieval. *ICTACT Journal on Image & Video Processing*, Vol. 13(4). https://doi.org/10.21917/ijivp.2023.0429.

[74] Yadav, V. and Bethard, S., (2019). A Survey on Recent Advances in Named Entity Recognition from Deep Learning Models. arXiv preprint *arXiv:1910.11470*. https://doi.org/10.48550/arXiv.1910.11470.

[75] Villacorta Chambi, S. P., Lindsay, M., Klump, J., Gessner, K., Gray, E. and McFarlane, H., (2025). Assessing named entity recognition by using geoscience domain schemas: the case of mineral systems. *Frontiers in Earth Science*, Vol.13-2025. https://doi.org/10.3389/feart.2025.1530004 .

Appendix 1 Enhanced Workflow

```
import os
import fitz  # PyMuPDF
import nltk
import glob
import re
from flair.data import Sentence
from flair.models import SequenceTagger

MODEL_DIR = "./ozrock/model"
PDF_DIR = "./FOLDER CONTAINING PDFs"

def extract_pdf_texts(pdf_dir: str) -> list:
    pdf_paths = glob.glob(os.path.join(pdf_dir, "*.pdf"))
    texts = []
    for pdf_path in pdf_paths:
        doc = fitz.open(pdf_path)
        raw_text = ""
        for page in doc:
            raw_text += page.get_text("text")
        corrected_text = correct_concatenated_words(raw_text)
        texts.append(corrected_text)
        doc.close()
    return texts

def correct_concatenated_words(text: str) -> str:
    text = re.sub(r'([a-z])([A-Z])', r'\1 \2', text)  # Split camelCase
    text = re.sub(r'([?.!,:;])([A-Za-z])', r'\1 \2', text)  # Ensure space after punctuation if followed by letter
    text = re.sub(r'(\d)([a-zA-Z])', r'\1 \2', text)  # Digit followed by letter
    text = re.sub(r'([a-zA-Z])(\d)', r'\1 \2', text)  # Letter followed by digit
    text = re.sub(r'([a-zA-Z])\.([a-zA-Z])', r'\1. \2', text)  # Letter dot letter without space
    return text

def annotate_pdf_texts(texts: list, model: SequenceTagger):
    annotations = []
    sentence_count = 0
    entity_count = 0
    for text in texts:
        sentences = nltk.sent_tokenize(text)
        sentence_count += len(sentences)
        for sentence in sentences:
            sent_obj = Sentence(sentence)
            model.predict(sent_obj)
            annotations.append(sent_obj.to_tagged_string())
            entity_count += len(sent_obj.get_spans('ner'))
    return annotations, sentence_count, entity_count

if __name__ == "__main__":
    trained_model = SequenceTagger.load(os.path.join(MODEL_DIR, 'best-model.pt'))
    pdf_texts = extract_pdf_texts(PDF_DIR)
    annotated_texts, total_sentences, total_entities = annotate_pdf_texts(pdf_texts, trained_model)

    output_file = "./NAME OF THE FILE.txt"
    with open(output_file, 'w') as f:
        for annotation in annotated_texts:
            f.write(annotation + "\n\n")
        f.write(f"Total sentences: {total_sentences}\n")
        f.write(f"Total recognized entities: {total_entities}\n")
        f.write(f"Total texts: {len(pdf_texts)}\n")

    print(f"Annotations, entity counts, and sentence counts saved to {output_file}")
```

Appendix 2 Original Workflow

```
import os
import nltk
import pdfplumber
import glob
from flair.data import Sentence
from flair.models import SequenceTagger

MODEL_DIR = "./ozrock/model"
PDF_DIR = "./FOLDER CONTAINING PDFs"

def extract_pdf_texts(pdf_dir: str) -> list:
    pdf_paths = glob.glob(os.path.join(pdf_dir, "*.pdf"))
    texts = []
    for pdf_path in pdf_paths:
        with pdfplumber.open(pdf_path) as pdf:
            text = "\n".join([page.extract_text() for page in pdf.pages])
            texts.append(text)
    return texts

def annotate_pdf_texts(texts: list, model: SequenceTagger):
    annotations = []
    sentence_count = 0
    entity_count = 0
    for text in texts:
        sentences = nltk.sent_tokenize(text)
        sentence_count += len(sentences)
        for sentence in sentences:
            sent_obj = Sentence(sentence)
            model.predict(sent_obj)
            annotations.append(sent_obj.to_tagged_string())
            # Count entities in the tagged sentence
            for entity in sent_obj.get_spans('ner'):
                entity_count += 1
    return annotations, sentence_count, entity_count

if __name__ == "__main__":
    trained_model = SequenceTagger.load(os.path.join(MODEL_DIR, 'best-model.pt'))
    pdf_texts = extract_pdf_texts(PDF_DIR)
    annotated_texts, total_sentences, total_entities = annotate_pdf_texts(pdf_texts, trained_model)

    # Output the annotated texts and total counts
    output_file = "./NAME OF THE FILE.txt"
    with open(output_file, 'w') as f:
        for annotation in annotated_texts:
            f.write(annotation + "\n\n")
        f.write(f"Total sentences: {total_sentences}\n")
        f.write(f"Total recognized entities: {total_entities}\n")
        f.write(f"Total texts: {len(pdf_texts)}\n")
    print(f"Annotations, entity counts, and sentence counts saved to {output_file}")
```