dh2loop 1.0: an open-source ~~python~~Python library for automated processing and classification of geological logs

Ranee Joshi[1,2], Kavitha Madaiah[1,2], Mark Jessell[1,2], Mark Lindsay[1,2] and Guillaume Pirot[1,2]

[1]Centre of Exploration Targeting, School of Earth Sciences, University of Western Australia, 35 Stirling Highway, Crawley 6009 Western Australia
[2]Mineral Exploration Cooperative Research Centre (MinEx CRC), School of Earth Sciences, University of Western Australia, 35 Stirling Highway, Crawley 6009 Western Australia

Correspondence to: Ranee Joshi (ranee.joshi@research.uwa.edu.au)

Abstract. ~~Exploration and mining companies rely on geological drill core logs to target and obtain initial information on geology of the area to build models for prospectivity mapping or mine planning.~~ A huge amount of legacy drilling data is available in geological survey but cannot be used directly as it is compiled and recorded in an unstructured ~~textural~~textual form and using different formats depending on the database structure, company, logging geologist, investigation method, investigated materials and/or drilling campaign. It is subjective and plagued with uncertainty as it is likely to have been conducted by tens to hundreds of geologists, all of whom would have their own personal biases. ~~However, this is valuable information that adds value to geoscientific data for research and exploration, specifically in efficiently targeting sustainable new discoveries and providing better shallow subsurface constraints for 3D geological models.~~

dh2loop (https://github.com/Loop3D/dh2loop) is an open-source python library ~~that provides the functionality to extract and standardize~~for extracting and standardizing geologic drill hole data and export it into readily importable interval tables (collar, survey, lithology). In this contribution, we extract, process and classify lithological logs from the Geological Survey of Western Australia Mineral Exploration Reports Database in the Yalgoo-Singleton Greenstone Belt (YSGB) region. ~~For this study case, the extraction rate for collar, survey and lithology data is respectively 93%, 865 and 34%.~~ It also addresses the subjective nature and variability of nomenclature of lithological descriptions within and across different drilling campaigns by using thesauri and fuzzy string matching. For this study case, 86% of the extracted lithology data is successfully matched to lithologies in the thesauri. Since this process can be tedious, we attempted to test the string matching with the comments, which resulted to a matching rate of 16% (7,870 successfully matched records out of 47,823 records). The standardized lithological data is then classified into multi-level groupings that can be used to systematically upscale and downscale drill hole data inputs for multiscale 3D geological modelling. *dh2loop* formats legacy data bridging the gap between utilization and maximization of legacy drill hole data and drill hole analysis functionalities available in existing python libraries (*lasio*, *welly*, *striplog*).

Formatted: Font: Italic

Formatted: Space After: 8 pt, Line spacing: Multiple 1.08 li

Formatted: Font: Not Bold, Font color: Auto

1 Introduction

Drilling is ~~the~~a process of penetrating through the ground ~~and~~ that is capable of extracting information about rocks from various depths ~~beneath~~below the surface. This is useful for ~~confirming~~establishing the geology beneath ~~and/~~the surface. Drill core or cuttings can be collected thus providing samples for ~~chemical~~ description, interpretation and analysis. ~~As it penetrates the ground it forms drill holes from which drill core is collected.~~ The location of where drilling starts is referred to as the collar. As the drilling progresses, survey orientation measurements are taken to be able to convert the specific depths to exact coordinate locations of the drill core being retrieved. In a hard rock setting, geological drill core logging is the process whereby the recovered drill core sample is systematically studied to determine the lithology, mineralisation, structures, and alteration zones of a potential mineral deposit. It is usually performed by geologists who classify a rock unit into a code, based on one or multiple properties such as rock type, alteration intensity and mineralisation content. Exploration and mining companies rely on the diverse geoscientific information obtained by drill core logging techniques to target and to build models for prospectivity mapping or mine planning. This work focuses on lithological logs which is the component of a geological log that refers to the geological information on the dominant rock type in a specific downhole interval. Inevitably, lithological drill core logging is subjective and plagued with uncertainty~~, particularly~~ as ~~it is likely to have been conducted by ten to hundred~~all logging geologists~~, all of whom would~~ have their own personal biases ~~(Lark et al., 2014).~~(Lark et al., 2014). The information and level of detail contained in logs is highly dependent on the purpose of the study, this already makes geological logging subjective. This subjectivity is also influenced by the lack of a standards between projects and/or companies combined with the personal biases of the logging geologist. Furthermore, it can be difficult to recognize lithology with confidence and to establish subtle variations or boundaries in apparently homogeneous sequences.

With the advent of the digital age, semi-automated drill core logging techniques such as X-Ray Diffraction (XRD), X-Ray Fluorescence spectrometry (XRF) and Hyperspectral (HS) imaging have provided higher detail of data collection and ~~even detection of~~ other properties such as conductivity, volumetric magnetic susceptibility, density using gamma-ray attenuation, and chemical elements during logging (Zhou et al., 2003; Rothwell and Rack, 2006; Ross et al., 2013). This has prompted a shift towards using numerical data rather than depending on traditional geological drill core logging procedures ~~(Culshaw, 2005)~~(Culshaw, 2005). Multiple methods have been recently applied to geological drill core logging such as wavelet transform analysis or data mosaic (Arabjamaloei et al., 2011; Hill et al., 2020; Le Vaillant et al., 2017; Hill et al., 2015), artificial neural network model (Lindsay, 2019; Zhou et al., 2019; Emelyanova et al., 2017) and inversion ~~(Zhu et al., 2019)~~(Zhu et al., 2019). Relying solely on these semi-automatic methods comes with drawbacks as it excludes some of the subjective interpretations that cannot be replaced. The semi-automatic methods also are poor at describing textural characteristics (foliation, banding, grain size variation). Furthermore, a rich amount of legacy data ~~was~~is collected in the traditional drill core logging method and disregarding this information limits the dataset.

Legacy data are information collected, compiled and/or stored in the past into many different old or obsolete formats or systems, such as handwritten records, aperture cards, floppy disks, microfiche, transparencies, magnetic tapes and/or newspaper clippings making it difficult to access and/or process ~~(Smith et al., 2015).~~(Smith et al., 2015). Legacy digital data also suffer from lack of standardisation and inconsistency. In geoscience, these are currently scattered amongst unpublished company reports, departmental reports, publications, petrographic reports, printed plans and maps, aerial photographs, field notebooks, sample ticket books, drill core samples, tenement information and geospatial data providing a major impediment to their efficient use. This includes geological drill core logs that are the outcome of most expensive part of most mineral exploration campaigns: drilling. This is valuable information source and key assets that can be used to add value to geoscientific data for research and exploration; design mapping programs and research questions of interest; more efficiently target remapping and sustainable new discoveries; and provide customers with all existing information at the start of the remapping program. It should not be abandoned for it may have lower intrinsic quality than observations made with more modern equipment, its recovery and translation to a digital format is too tedious. ~~Elizabeth Griffin (2015)~~Griffin (2015) argues that there is no distinction in principle between legacy data and 'new' data, as all of it is data. The intention of recovering legacy data is to a) upcycle information with integration into modern datasets, b) use salvaged data for new scientific applications and c) allow reuse of that information into utility downstream applications ~~(Vearncombe et al., 2017).~~(Vearncombe et al., 2017). Furthermore, extracting information from legacy datasets is ~~high~~valuable and relatively low-risk as geoscientific insight is added to a project for little or no cost compared to those of drilling ~~(Vearncombe et al., 2016)~~(Vearncombe et al., 2016).

~~The primary challenge in dealing with geological legacy datasets is a large amount of important data, information and knowledge are recorded in an unstructured textural form, such as host rock, alteration types, geological setting, ore-controlled factors, geochemical and geophysical anomaly patterns, and location (Wang and Ma, 2019). Moreover, the geological drill core logging forms and formats vary depending on the company, logging geologist, investigation method, investigated materials and/or drilling campaign. Natural language processing (NLP) also known as computational linguistics has been used for information extraction, text classification and automatic text summarization. NLP relies on data-driven computation involving statistics, probability, machine learning and "deep learning" (Otter et al., 2020). NLP applications on legacy data have been demonstrated in the fields of taxonomy (Rivera-Quiroz and Miller, 2019), biomedicine (Liu et al., 2011) and legal services (Jallan et al., 2019). Qiu et al. (2020) proposed an ontology-based methodology to support automated classification of geological reports using word embeddings, geoscience dictionary matching and bidirectional long short-term memory model (Dic-Att-BiLSTM) that assists in identifying the difference in relevance from a report. Padarian and Fuentes (2019) also introduced the use of domain-specific word embeddings (GeoVec) which was used to automate and reduce subjectivity of geological mapping of drill hole descriptions (Fuentes et al., 2020).~~

~~Similarity matching has many applications in natural language processing as it is one of the best techniques for improving retrieval effectiveness (Park et al., 2005). The use of text similarity is beneficial for text categorization (Liu and Guo, 2005) and text summarization (Erkan and Radev, 2004;Lin and Hovy, 2003). Fuzzy string matching, also known as approximate string matching, is the process of finding strings that approximately match a given pattern (Cohen, 2011;Gonzalez et al., 2017). It has been used in language syntax checker, spell-checking, DNA analysis and detection, spam detection, sport and concert event ticket search (Higgins and Mehta, 2018), text re-use detection (Recasens et al., 2013) and clinical trials (Kumari et al., 2020).~~

The primary challenge in dealing with geological legacy datasets is that a large amount of important data, information and knowledge are recorded in an unstructured textural form, such as host rock, alteration types, geological setting, ore-controlled factors, geochemical and geophysical anomaly patterns, and location (Wang and Ma, 2019). To acknowledge the ambiguity in the context of "unstructured textual form", we define it in this paper as, "descriptive text that lacks a pre-defined format and/or metadata thus cannot be readily indexed and mapped into standard database fields". The geological drill core logging forms and formats also vary depending on the company, logging geologist, investigation method, investigated materials and/or drilling campaign. Natural language processing (NLP) also known as computational linguistics has been used for information extraction, text classification and automatic text summarization (Otter et al., 2020). NLP applications on legacy data have been demonstrated in the fields of taxonomy (Rivera-Quiroz and Miller, 2019), biomedicine (Liu et al., 2011) and legal services (Jallan et al., 2019). Qiu et al. (2020) proposed an ontology-based methodology to support automated classification of geological reports using word embeddings, geoscience dictionary matching and bidirectional long short-term memory model (Dic-Att-BiLSTM) that assists in identifying the difference in relevance from a report. Padarian and Fuentes (2019) also introduced the use of domain-specific word embeddings (GeoVec) which is used to automate and reduce subjectivity of geological mapping of drill hole descriptions (Fuentes et al., 2020).

Similarity matching has many applications in natural language processing as it is one of the best techniques for improving retrieval effectiveness (Park et al., 2005). The use of text similarity is beneficial for text categorization (Liu and Guo, 2005) and text summarization (Erkan and Radev, 2004; Lin and Hovy, 2003). It has been used to extract lithostratigraphic markers from drill lithology logs (Schetselaar and Lemieux, 2012). Fuzzy string matching, also known as approximate string matching, is the process of finding strings that approximately match a given pattern (Cohen, 2011; Gonzalez et al., 2017). It has been used in language syntax checker, spell-checking, DNA analysis and detection, spam detection, sport and concert event ticket search (Higgins and Mehta, 2018), text re-use detection (Recasens et al., 2013) and clinical trials (Kumari et al., 2020).

Most of the available python libraries available have been built to process extracted and standardized drill hole data. The most common of these are: *lasio* (https://lasio.readthedocs.io/en/latest/) which deals with reading and writing Log ASCII Standard (LAS) files, a drill hole format commonly used in the oil and gas industry, *welly* (https://github.com/agile-geoscience/welly) which deals with loading, processing, and analysis of drill holes and *striplog* (https://github.com/agile-geoscience/striplog) which digitizes, visualizes and archives stratigraphic and lithological data. ~~*Striplog* (Hall and Keppie, 2016)~~*Striplog* (Hall and Keppie, 2016) also parses natural language 'descriptions', converting them into structured data via an arbitrary lexicon which allows further querying and analysis on drill hole data. The main ~~limitations~~limitation of these existing libraries, with respect to legacy data in the mining sector is that they assume that the data is already standardized and pre-processed.

dh2loop provides the functionality to extract and standardize geologic drill hole data and export it into readily importable interval tables (collar, survey, lithology). It addresses the subjective nature and variability of nomenclature of lithological descriptions within and across different drilling campaigns by integrating published dictionaries, glossaries and/or thesauri that ~~were~~are built to improve resolution of poorly defined or highly subjective use of terminology and idiosyncratic logging methods. It is however important to highlight that verifying the accuracy and/or correctness of the geological logs being standardized is outside the scope of this tool, thus we assume logging has been conducted to the best of the geologist's ability.

Furthermore, it classifies lithological data into multi-level groupings that can be used to systematically upscale and downscale drill hole data inputs in multiscale 3D geological model. It also provides drill hole desurveying (computes the geometry of a drillhole in three-dimensional space) and log correlation functions so that the results can be plotted in 3D and analysed against each other. It also links the gap between utilization and maximization of legacy drill hole data and the drill hole analysis functionalities available in existing python libraries.

~~**2 Materials and Methods**~~

2 *dh2loop* Drillhole Data Extraction

2.1 Conventions and Terminologies

This paper involves multiple python libraries, database tables and fields. For clarity, the following conventions are used for this paper~~:~~ (Appendix A1):

1. Python libraries are written in italics: *dh2loop*
2. Python functions are written in italics followed by an open and close parenthesis: *token_set_ratio()*
3. Database tables are written in ~~Lucinda~~Lucida Console Italics: *dhgeology*

Formatted: Font color: Auto

Formatted: Space Before: 12 pt, After: 12 pt, Line spacing: single, Don't keep lines together

Formatted: Font color: Auto

Formatted: Font: Times New Roman

Formatted: Font: Times New Roman

4. Database table fields are written in ~~Lucinda~~Lucida Console: `CollarID`

5. Workflows are written in Century Gothic Bold: **Lithology Code workflow**

Formatted: Font: Times New Roman

Formatted: Font: Times New Roman

~~2.2 Dependencies~~

Formatted: Font color: Auto, Highlight

Formatted: Space Before: 12 pt, After: 12 pt, Line spacing: single, Don't keep lines together

~~*dh2loop* stands for drill hole data extracted into a 3D modelling input format, compatible with/for the Loop platform. It is a drill hole processing tool that integrates published dictionaries, glossaries and/or thesauri to and improve standardize highly subjective use of terminology and idiosyncratic logging methods and classify lithological logs. It primarily depends on a number of external open-source libraries:~~

1. ~~*fuzzywuzzy* (https://github.com/seatgeek/fuzzywuzzy) which uses fuzzy logic for string matching (Cohen, 2011)~~
2. ~~*pandas* (https://pandas.pydata.org/) for data analysis and manipulation (McKinney, 2011)~~
3. ~~*psycopg2* (https://pypi.org/project/psycopg2/), a PostgreSQL database adapter for python (Gregorio and Varrazzo, 2018)~~
4. ~~*numpy* (https://github.com/numpy/numpy)~~
5. ~~*nltk* (https://github.com/nltk/nltk), the Natural Language Toolkit is a suite of open source Python modules, data sets, and tutorials supporting research and development in Natural Language Processing (Loper and Bird, 2002).~~
6. ~~*pyproj* (https://github.com/pyproj4/pyproj), python interface to PROJ (cartographic projections and coordinate transformations library)~~

~~Code describing basic drill hole operations, such as desurveying (process of translating collar (location) and survey data (azimuth, dip, length) of drill holes into XYZ coordinates in order to define its 3D geometry of the non-vertical borehole), was heavily inspired from *pyGSLIB* drill hole module (Martínez-Vargas, 2016). *pyGSLIB* (https://github.com/opengeostat/pygslib) is an open-source python package to perform mineral resource estimations. The *pyGSLIB* drillhole module handles drill hole data, desurveying interval tables and other drill hole related processes. The module was re-written into python to be make it more compact with less dependencies and tailor it to the data extraction output.~~

~~2.3~~2.2 Data Source

Formatted: Font color: Auto

Formatted: Space Before: 12 pt, After: 12 pt, Line spacing: single, Don't keep lines together

The Geological Survey of Western Australia Mineral Exploration Reports Database contains open-file reports submitted as a compliance to the Sunset Clause, Regulation 96(4) of the Western Australia legislation Mining Regulations 1981. These reports contain valuable exploration information in hardcopy (1957-2000), hardcopy and digital format (2000-2007) and digital format (2000-present) ~~(Riganti et al., 2015)~~(Riganti et al., 2015). The minimum contents of a drilling report comprise a collar file which describe the geographic coordinates of the collar location. ~~(Fig. 1).~~ Additional files may be included, such as a survey file describing the depth, azimuth and ~~dip~~inclination measurements for the drilling path; assays; downhole geology and property surveys (e.g. downhole geochemistry, petrophysics) may also be available depending on the company's submission ~~(Riganti et al., 2015).~~(Riganti et al., 2015). The data in the drilling reports ~~were~~are extracted with spatial attribution and imported to a custom-designed relational database (also called the Mineral Drillhole Database) curated by the GSWA that allows easy retrieval and spatial querying. For simplicity, we will refer to this database as the WAMEX database in this text.

Formatted: Highlight

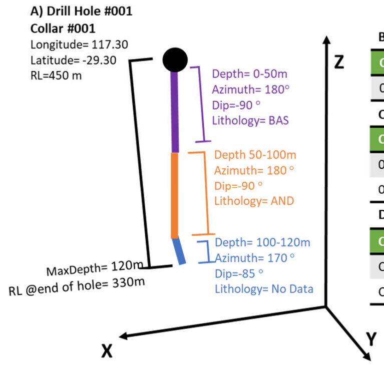
Formatted: Highlight

Formatted: Highlight

The WAMEX database contains more than 50 years' worth of mineral exploration drill hole data with more than 2.05 million drill holes, imported from over 1,514 companies. Each drill hole is identified by its surface coordinates and its unique ID (`CollarID`) in the *`collar`* table). The drill hole 3D geometry is described in the survey tables (*`dhsurvey`*,

Formatted: Font: Times New Roman

dhsurveyattr).



B) Collar Data

CollarID	HoleID	Longitude	Latitude	RL	MaxDepth
001	DDH_001	117.30	-29.30	450	120

C) Survey Data

CollarID	Depth	Azimuth	Dip
001	0	180	-90
001	100	170	-85

D) Lithology Data

CompanyID	CollarID	FromDepth	ToDepth	Attribute Value	Company_Litho
CompanyA	001	0	50	BAS	Hi-Mg basalt
CompanyA	001	50	100	AND	Andesite

The lithology along the drill hole is described as a function of depth in the lithology tables (*dhgeology* and *dhgeologyattr*). However, it is important to emphasize that the drill hole data is of variable quality and reliability and that no validation has been done. The necessary amendments and reformatting enabling to extract and utilize data from the WAMEX database are part of the functionalities provided by *dh2loop*.

Figure 1. Simplified example of a drill hole (1.A) and its corresponding interval tables collar (1.B), survey (1.C) and lithology (1.D). The black circle denotes the collar location of the drill hole which is obtained from a collar table (1.B). The purple line represents the first downhole interval taking its deviation data from the survey table (1.C) and the lithology information from the lithology table (1.D). The same applies for the second interval (orange line) and third interval (blue line). The orange line follows the same trajectory as the first interval as it uses the same entry in the survey table (1.C). The blue line has no lithology data as this information is not present in the lithology table (1.D). The `MaxDepth` denotes the total drill length (1.B).

Figure 2. Simplified WAMEX database schema showing the one-to-many relationship between the *collar* table and the *collarattr* table (red solid line). Each drill hole in the WAMEX database is identified by its surface coordinates and a unique ID (`CollarID`) in the *collar* table. The drill hole 3D geometry is described in the survey tables (*dhsurvey*, *dhsurveyattr*). The properties logged are described in pair of tables such as *dhgeology*, *dhgeologyattr*. The WAMEX schema is linked by a main table, *collar*, which assigns a primary key (`CollarID`) for each drill hole in Western Australia and corresponding geographic location (Fig. 2). *collar* has a one-to-many relationship with another table, *collarattr*, to store other attributes that describe each unique drill hole, such as maximum depth and elevation. The deviation of the drill hole is stored in a table, *dhsurvey*, with a primary key (`DHSurveyID`) that refers to each unique depth of a drill hole. This primary key has a many-to-one relationship with collar, as there are multiple depth measurements for each drill hole. Furthermore, *dhsurvey* also has a one-to-many relationship with table *dhsurveyattr*, which stores additional attribute information regarding survey, such as azimuth and dip readings. This similar relationship is maintained with interval tables, except that the primary key (e.g.`DHGeologyID`) is used to refer a unique downhole interval rather than a depth measurement. For lithological information, we refer to tables: *dhgeology* and *dhgeologyattr*. *dhgeologyattr* which contain information such as rock names and free text comments while *dhgeology* provides information to which hole and interval depth that data refers to. This information can be joined and extracted through SQL (Structured Query Language) queries. Other interval tables available in the database refer to typical logs and downhole surveys: alteration, events, geochemistry geophysics, geotech, hyperspectral, magnetic susceptibility, mineralogy, recovery, regolith, specific gravity, structure, veining, water level and weathering information.

Formatted: Caption

Formatted: Font: Not Italic, Highlight

Formatted: Font: Not Italic, Highlight

Formatted: Font: Times New Roman, Highlight

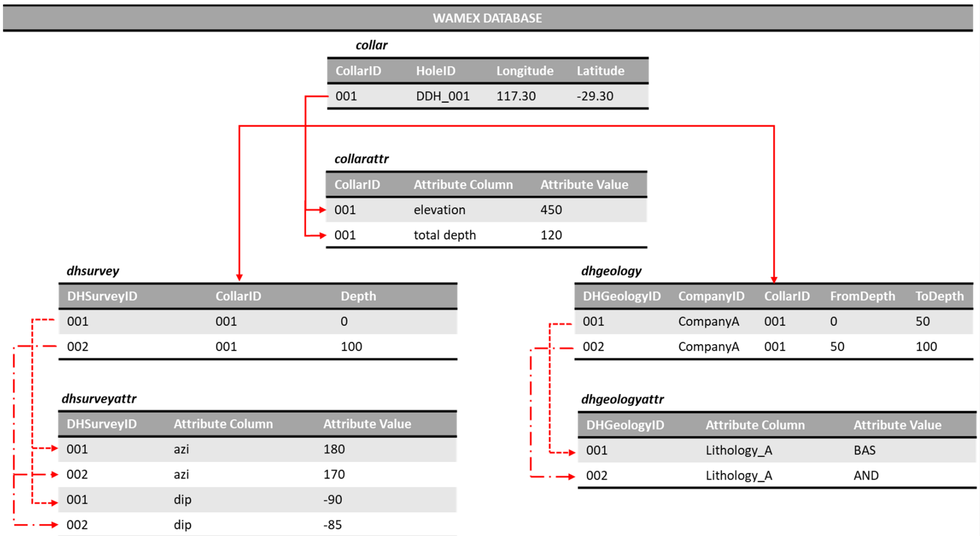


Figure 2. Simplified WAMEX database schema showing the one-to-many relationship between the *collar* table and the *collarattr* table (red solid line). It also shows the relationship between the collar table and the other interval tables such as *dhsurvey, dhsurveyattr, dhgeology, dhgeologyattr*. The example show the relationship between tables for the first (red dashed line) and second interval (red dashed-dot line).

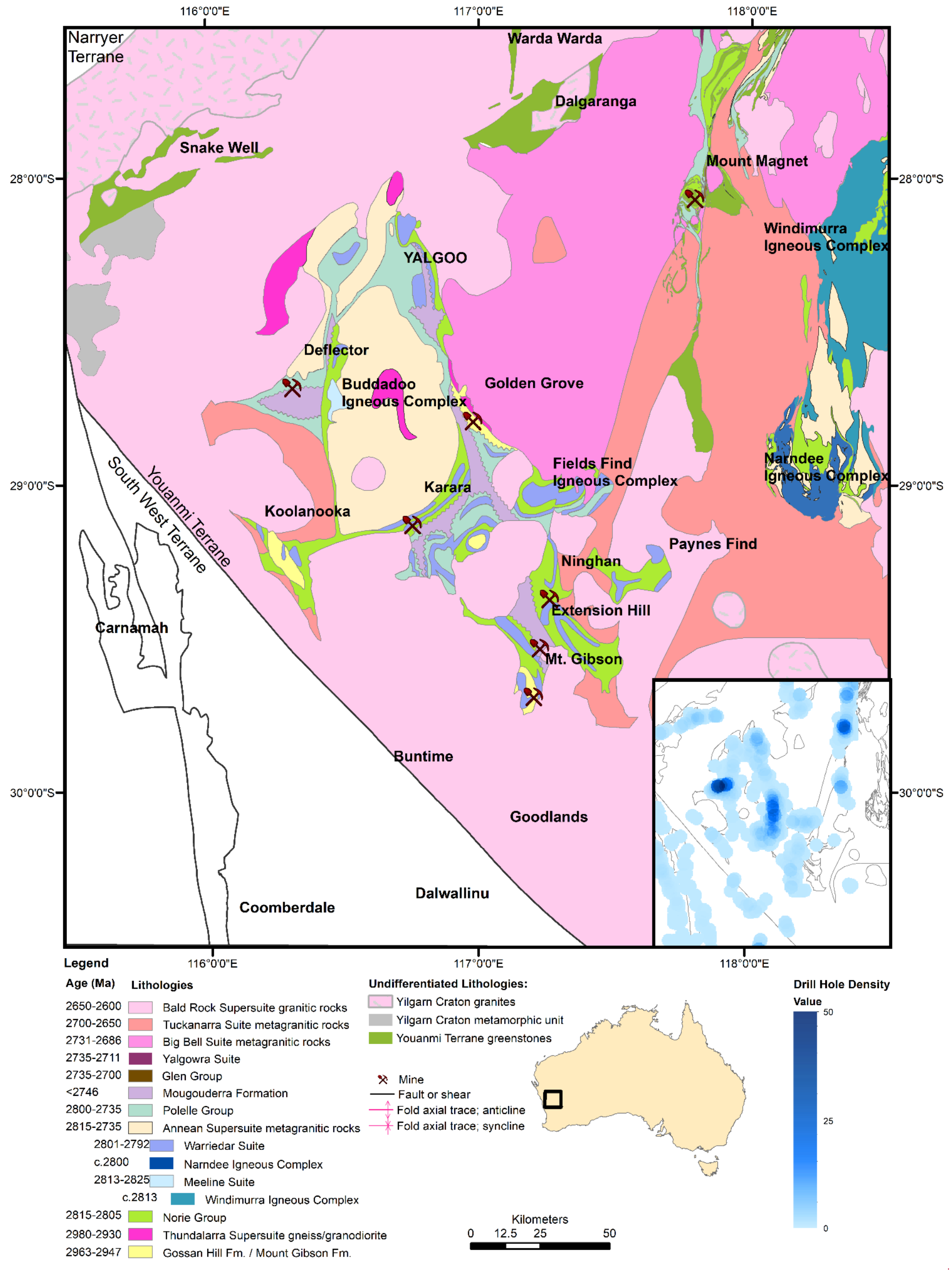
It contains more than 50 years worth of mineral exploration drill hole data with more than 2.05 million drill holes, imported from over 1,514 companies. Each drill hole is identified by its surface coordinates and its unique ID (CollarID). The drill hole 3D geometry is described in the survey tables (*dhsurvey, dhsurveyattr*). However, it is important to emphasize that the drill hole data is of variable quality and reliability and no validation has been done. The amendments and reformatting necessary to be able to extract and utilize data from the WAMEX database is part of the functionalities provided by *dh2loop*.

2.42.1 Study Area

In this paper, we demonstrate the application of *dh2loop* to data from the Yalgoo-Singleton greenstone belt (YSGB) (Fig. 2), a geologically complex, largely heterogeneous and highly mineralized arcuate granite-greenstone terrane, in the western Youanmi Terrane, Yilgarn Craton in Western Australia (Anand and Butt, 2010). The YSGB has good range of different lithologies in the area. Igneous rocks occur as extensive granitoid intrusions that occurred between 2700 and 2630 Ma (Myers, 1993), as well as ultramafic mafic volcanic rocks formed as extensive submarine lavas and local centres of felsic and mafic volcanic rocks. Some layered gabbroic sills intruding the greenstone are also observed. Sedimentary rocks formed in broad basins during tectonic and volcanic quiescence consist of mostly banded iron formation (BIF) and felsic volcanoclastic rocks. The greenstone belt is metamorphosed to greenschist facies (Barley et al., 2008). The area is also covered by deeply weathered regolith profiles which both conceals mineral deposits hosted by the underlying bedrock and contains signatures of mineralisation that are distal signatures of possible economically significant deposits (Cockbain, 2002). Furthermore, the YSGB is a major target for exploration as it has considerable resources of gold, nickel, bauxite, as well as lesser amounts of a wide range of other commodities (Cockbain, 2002). It hosts multiple mineral deposits ranging from volcanogenic massive sulphide (Golden Grove, Gossan Hill), orogenic gold (Mt. Magnet), banded iron formations (Mount Gibson, Karara, Extension Hill). The geological complexity and relevance to mineral exploration makes the YSGB a reasonable and sensible area to test the dh2loop thesauri, matching and upscaling.

Formatted: Font: Times New Roman

Formatted: Space Before: 12 pt, After: 12 pt, Line spacing: single, Don't keep lines together



~~**Figure 3.** The map shows the Yalgoo-Singleton greenstone belt highlighting the different mines and prospects in the area. The inset map shows the heterogeneous distribution and drill hole density from the legacy data available from the WAMEX database.~~

~~2.5~~2.3 Thesauri

Since most exploration companies have their own nomenclature and systems, which could also change between drilling campaigns, it is necessary to build thesauri: dictionaries that list equivalent and related nomenclature (or synonyms) for different attribute names and values. Synonyms include terminologies that share a similar intent, for example, RL (relative level) terms, whether elevation or relative level, as long as the words are recording a vertical height. These thesauri are stored as additional tables in the database. For example, if we are interested in the major lithology in a specific interval, this information can be tabulated as "Major Rock Type", "Lithology_A" or "Main_Geology_Unit" depending on the drill core logging system used. The resulting thesauri considers change in cases, abbreviations, addition of characters, typographical errors and a combination of these. Although listing out these terms is manual and tedious, it only needs to be done once and can be re-used and forms the basis for future text matching and as a training set to automate finding similar terms. This ~~was~~is preferred over selection based on regular expressions as when parsing these terms, there are complex patterns in the terms used and the inconsistencies in the way they are written that can be understood by a person with a geological background but not by a simple regular expression. The complexity of the regular expression required to catch all the terms of interest means an optimal expression is difficult, if not impossible, to define, and also tends to be computationally burdensome. *dh2loop* ~~1.0~~ provides several thesauri that can easily be updated (if needed) for the following attribute~~ss (Appendix B: Thesauri):~~ (Appendix A. A1). In order to extract the other attributes, we envisage developing other thesauri, following the same workflow.

1. Drill ~~hole collar elevation (Appendix B1):~~Hole Collar Elevation Thesaurus: 360 synonyms such as "elevation" and "relative level"
2. Drill ~~hole maximum depth (Appendix B2):~~Hole Maximum Depth Thesaurus:160 synonyms such as "end of hole", "final depth" and "total depth"
3. Drill ~~hole survey azimuth (Appendix B3):~~Hole Survey Azimuth Thesaurus: 142 synonyms
4. Drill ~~hole survey dip (Appendix B4):~~Hole Survey Inclination Thesaurus: 8 synonyms such as "~~inclination~~dip"
5. Drill ~~hole lithology (Appendix B5):~~Hole Lithology Thesaurus: 688 synonyms such as "geology", "Lithology_A", "Major_Geology_Unit" and "Major_Rock_Type"
6. Drill ~~hole comments (Appendix B6):~~Hole Comments Thesaurus: 434 synonyms such as "description"

The thesauri created specifically for further processing lithology and comments information are:

7. Drill ~~hole lithology codes thesaurus (Appendix B7,~~ Hole Lithology Codes Thesaurus

   ~~7.~~ It compiles the equivalent lithology for a given lithological code based on the reports submitted to GSWA. This thesaurus is identified by a company id and report number.~~discussed further in Sect. 2.5.1)~~

8. Clean-~~-~~up ~~dictionary (Appendix B8,~~ Dictionary

   It is a list of words and non-alphabetic characters that are used as descriptions in the geological logging syntax. This dictionary is used to remove these terms from the `Company_Litho` and/or `Comments` free text descriptions prior to the fuzzy string matching. The dictionary is composed of terms that describe age, location, structural forms, textures, amount/distribution, minerals, colors, symbols and common phrases, compiled from abbreviations in field and mine geological mapping (Chace, 1956) and the CGI-IUGS geoscience vocabularies accessible at http://geosciml.org/resource/def/voc/ (Simons et al., 2006; Richard et al., 2007; Raymond et al., 2012).

~~8. discussed further in Sect. 2.5.2)~~

Lithology ~~hierarchical thesaurus (Appendix B9,~~ Hierarchical Thesaurus~~discussed further in Sect. 2.5.3)~~

9.

~~In order to extract the other attributes we envisage developing other thesauri, following the same workflow.~~

~~**2.5.1 Drill hole lithology codes**~~

Formatted: Space Before: 12 pt, After: 12 pt, Line spacing: single, Don't keep lines together

Formatted: Font: Italic

Field Code Changed

Formatted: Indent: Left: 1.27 cm, No bullets or numbering

Formatted: Font: Not Italic

Formatted: Indent: Left: 1.27 cm, No bullets or numbering

Formatted: Font: Not Italic

Formatted: List Paragraph, Numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0.63 cm + Indent at: 1.27 cm

Formatted: List Paragraph

Formatted: Indent: Left: 1.27 cm, No bullets or numbering

~~This is a thesaurus compiling the equivalent lithology for a given lithological code based on the reports submitted to GSWA. This thesaurus is identified by a company id and report number. The current thesaurus covers 41 out of the 168 companies in the study area with a total of 352 entries (Appendix B7). It is important to note that the `Company_LithoCode` varies depending on the `CompanyID`. For example, "Company 551" refers to "Saprolite" as "CS" while Company "2551" uses CS to refer to "Cambrian Sediment". The opposite is true as well, a company may use "AMPH" to refer to "Amphibolite" while another company may use "MAA".~~

Formatted: List Paragraph

~~2.5.2 Clean-up Dictionary~~

Formatted: Indent: Left: 1.27 cm, No bullets or numbering

~~The clean-up dictionary is a list of words and non-alphabetic characters that are used as descriptions in the geological logging syntax. This dictionary is used to remove these terms from the lithology and/or comment free text descriptions prior to the fuzzy string matching. The dictionary is composed of 1662 records, most of which were compiled from abbreviations in field and mine geological mapping (Chace, 1956) and the CGI vocabularies: GeoSciML and EarthResourcesML (Simons et al., 2006;Richard et al., 2007;Raymond et al., 2012). 353 of these records are original to *dh2loop* and were added to accommodate the geological logging syntax in Western Australia. Added records include the following:~~

Formatted: List Paragraph

1. ~~Chronostratigraphic ages (Cambrian, Proterozoic)~~
2. ~~Location descriptors (above, below, between), relative time (ancient, older, youngest)~~
3. ~~Structural descriptors (anticlinal)~~
4. ~~Textures (rounded, angular, block)~~
5. ~~Mineralisation-related terminologies (absent, massive, disseminated)~~
6. ~~Minerals (bornite, cassiterite),~~
7. ~~Colors (brownish, cream)~~
8. ~~Adjectives and their root form (good, better, best, extremely, extreme, fragmental, fragments, fragment)~~
9. ~~Symbols (>, ?,\,@); and~~
10. ~~Common phrases (same as above, as per usual).~~

Formatted: Indent: Left: 1.27 cm, No bullets or numbering

~~2.5.3 Lithology Hierarchical Thesaurus~~

Formatted: Indent: Left: 1.27 cm, No bullets or numbering

~~The lithology hierarchical thesaurus~~It is a list of 757 rock names (`Detailed_Lithology`), their synonyms and a two-level upscale grouping (`Lithology_Subgroup` and `Lithology_Group`) (Fig 1). Each row in `Detailed_Lithology` refers to a rock name. Each rock name row lists the standardized terminology first, followed by its synonyms. The two corresponding columns for this row indicated the two-level upscale grouping. Many of the `Lithology_Subgroups` listed have parent-child relationships e.g. 'mafic_fine_grained_crystalline' is a child of 'mafic'. Parents in parent-child relationships are included in their children as catch-all groups to capture free text descriptions that do not include details that would be captured by only using the child terms alone. 169 of these rock names ~~were compiled from GeoSciML~~ are compiled from the CGI-IUGS Simple Lithology vocabulary available at: http://resource.geosciml.org/classifier/cgi/lithology (Simons et al., 2006; Richard et al., 2007; Raymond et al., 2012). The synonyms ~~were~~are obtained from mindat.org ~~(Ralph and Chau, 2014;Ralph, 2004). The hierarchical classification was~~(Ralph and Chau, 2014; Ralph, 2004). The hierarchical classification is inherited from both mindat.org ~~(Ralph and Chau, 2014;Ralph, 2004)~~(Ralph and Chau, 2014; Ralph, 2004) and the British Geological Survey (BGS) Classification Scheme (Gillespie and Styles, 1999; Robertson, 1999; Hallsworth and Knox, 1999; McMillan and Powell, 1999; Rosenbaum et al., 2003). It is important to use multiple libraries to be able to build an exhaustive thesauri as some libraries are limited by the nomenclature, level of interest and presence of the lithology or rock group in a geographic area. For example, the BGS classification did not have a comprehensive regolith

Formatted: List Paragraph

Formatted: Not Highlight

dictionary, but one could argue that this is because the regolith is not extensive and not known to be a host for mineralisation in the United Kingdom.. Thus, regolith has been classified using the regolith glossary (Eggleton, 2001)(Eggleton, 2001).

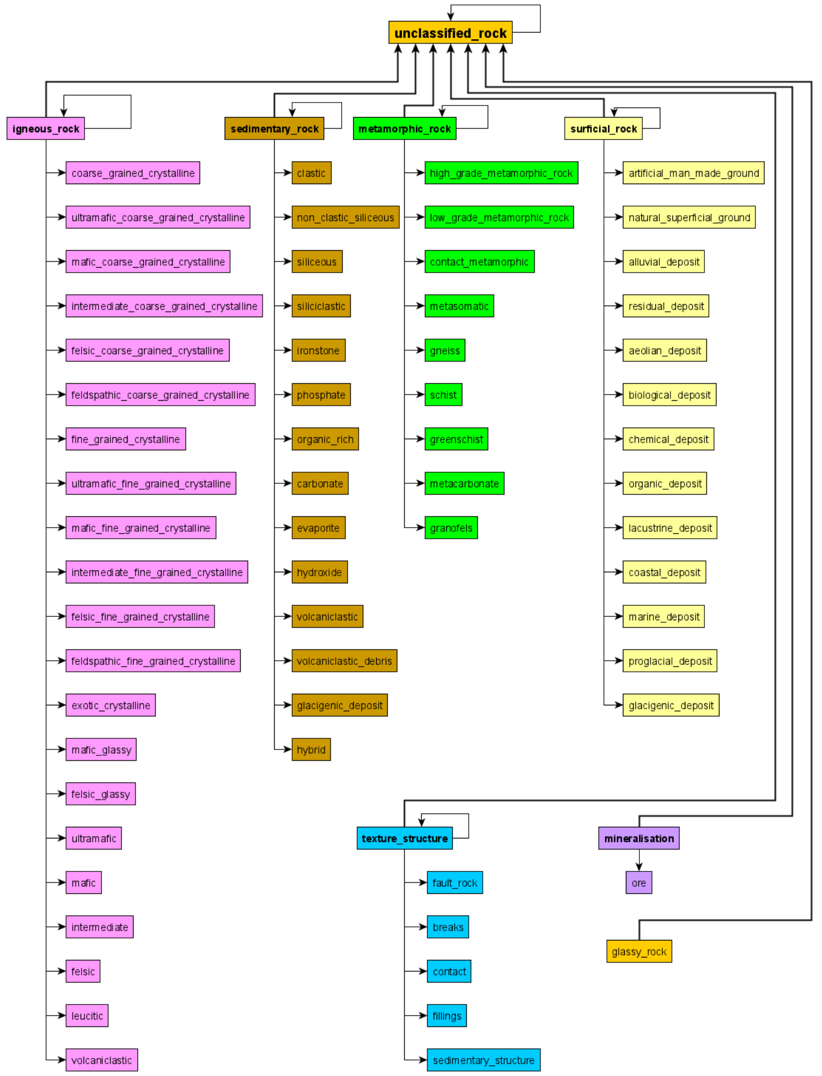


Figure 1. Lithology Hierarchical Thesaurus showing the 7 major Lithology_Groups: Igneous rocks (pink), Sedimentary rocks (light brown), Metamorphic rocks (green), Surficial Rocks (light yellow), Texture and Structure (blue), Mineralisation (purple) and Unclassified Rocks (dark yellow) and their corresponding Lithology_Subgrouups. The hierarchical classification consists of seven major lithology groups (Lithology_Group) (Fig. 4):

1) Igneous rocks. The igneous rocks are further classified to 12 lithology subgroups, considering grain size, composition and a combination of both.

Formatted: Font: Not Italic

Formatted: Centered

Formatted: Centered, Line spacing: 1.5 lines

Formatted: Caption

Formatted: Font: Times New Roman

Formatted: Caption, No bullets or numbering

2) Sedimentary rocks. Sedimentary rocks sub-classified to 16 lithology subgroups based on genetic source and composition (carbonate, clastic, evaporate, hybrid, hydroxide, ironstone, non-clastic siliceous, organic-rich, phosphate, siliceous, siliciclastic, volcaniclastic, glacigenic).

3) Metamorphic rocks. Metamorphic rocks are subdivided into nine lithology subgroups based on the degree and type of metamorphism (metasomatic, contact, low-grade, schist, gneiss, high-grade, granofels, greenschist, metacarbonate).

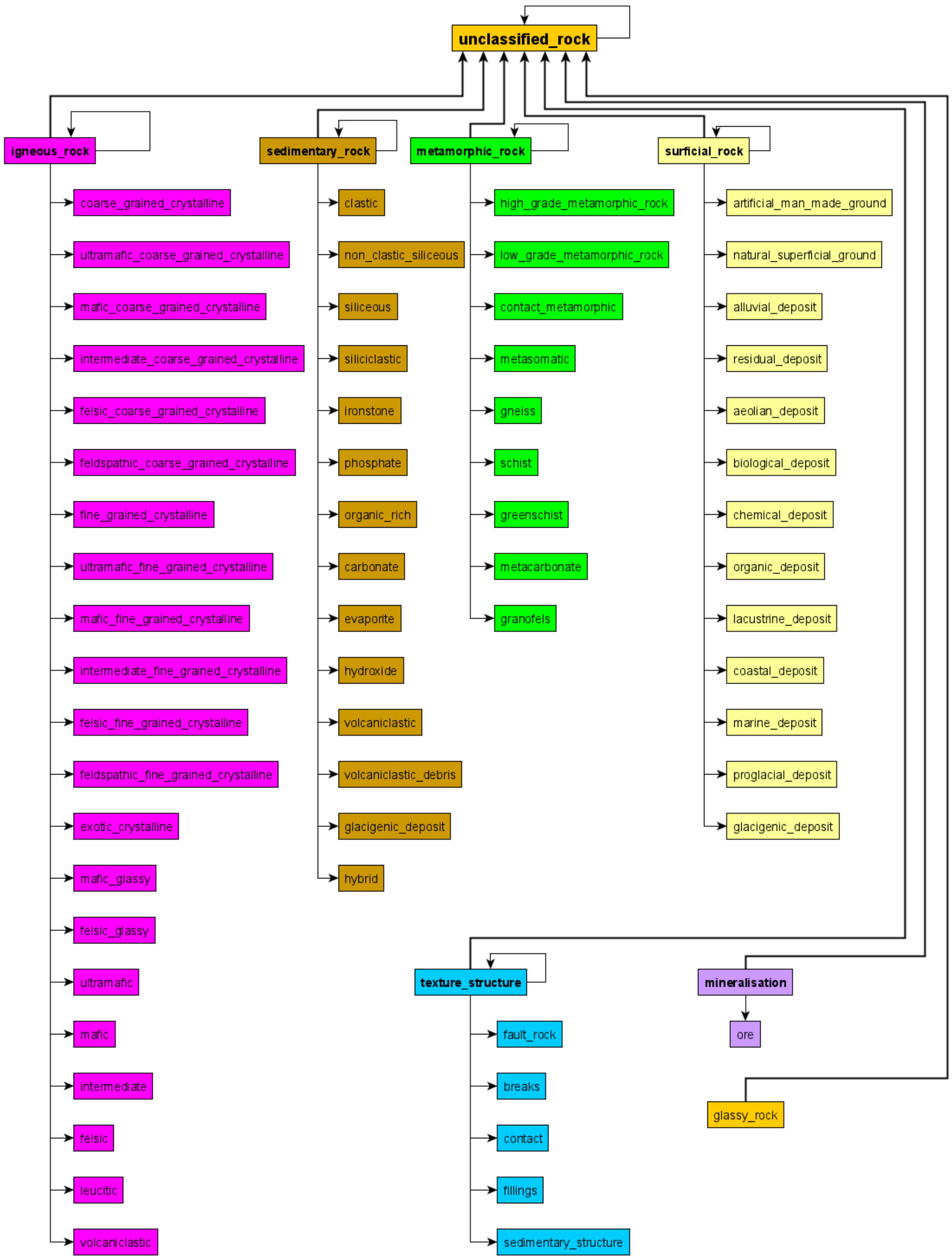
4) Surficial rocks. Surficial rocks are subdivided into 13 lithology subgroups based on the depositional environment and composition. The residual deposit `Lithology_Subgroup` includes the regolith detailed lithologies.

5) Mineralisation. Mineralisation is considered as a separate classification to be able to classify ore zones.

6) Structure and texture. Since structures and textures can sometimes be logged as lithologies in geological logging, they are classified separately. Structure and Texture is divided into five lithology subgroups: fault rock, breaks, contact, fillings and sedimentary structures. And

7) Unclassified. The final classification is a catch–all for unclassified rocks.

Formatted: Caption, Don't keep with next



~~Figure 4. Lithology hierarchical thesauri showing the 7 major `Lithology_Groups` and their corresponding `Lithology_Subgroups`: Igneous rocks (pink), Sedimentary rocks (light brown), Metamorphic rocks (green), Surficial Rocks (light yellow), Texture and Structure (blue), Mineralisation (purple) and Unclassified Rocks (dark yellow). Igneous rocks `Lithology_Subgroups` considers grain size, composition and a combination of both. Sedimentary rocks are sub-classified based on genetic source and composition. Metamorphic rocks are subdivided based on the degree and type of metamorphism. Surficial rocks are subdivided based on the depositional environment and composition. Mineralisation is considered as a separate~~

~~classification to be able to classify ore zones. Structure and texture addresses situations that structures are logged as lithologies in geological logging.~~ ~~The final classification is a catch–all for unclassified rocks.~~ ~~The final classification is a catch–all for unclassified rocks.~~

Formatted: Caption

2.44 Data Extraction

Currently, the *dh2loop* library extracts collar, survey and lithology information. The paper focus on the lithological extraction. Database structure and extraction results for collar and survey are available in Appendix B3 & B4 . The extraction~~It~~ uses a configuration file ~~(Appendix C1)~~ that allows the user to define the inputs, which are:

1. Region of interest (in WGS 1984 lat/long); and/or
2. List of drill hole ID codes codes, if known.
3. If reprojection is desired, the EPSG code of the projected coordinate system (e.g. EPSG:28350 for MGA Zone 50; http://epsg.io)
4. The connection credentials to the local copy of the WAMEX database

Input and output file directories/~~locations~~location

5.

~~**2.4.1 Collar Extraction**~~

~~With the minimum input of a region of interest, the *dh2loop* library exports a Comma-Separated Values file (CSV) listing the drill holes in the area with the following information (Fig. 5):~~

1. ~~`CollarID`: This is the primary key from the *collar* table. It is used to associate data in different tables with a single drill hole. The `CollarID` for a drill hole is identical in all tables in order for data to be associated with that drill hole.~~
2. ~~`HoleID`: This is the drill hole name, as the company would internally identify the drill hole.~~
3. ~~`Longitude` and `Latitude`: The geographical coordinates locating the collar of the drill hole. Both values are expressed in WGS 1984 lat/long (EPSG:2436).~~
4. ~~Relative level (`RL`): This refers to the Z coordinate of the collar location. This value is extracted by using the drill hole collar elevation thesaurus to filter the values referring to relative level (Fig. 5b). More than one value can be fetched due to duplicate company submissions or multiple elevation measurements, in which case the code retains the value with most decimal places assuming higher precision corresponds to better accuracy. If no elevation values are fetched from the database the entire record is skipped. Non-numeric values are also ignored.~~
5. ~~Maximum depth (`MaxDepth`): This refers to the maximum downhole length drilled for a drill hole, commonly referred as the end-of-hole. This value is extracted by using the drill hole collar maximum depth thesaurus (Fig. 5c). Due to duplicate company submissions, there can be more than one value fetched. Since there is no submission date information, the code takes the value with largest value assuming it is the latest submission.~~
6. ~~Calculated `X`, `Y` values of projected coordinates: These values are commonly calculated and used to be able to plot the drill hole in a metric system to be able to accurate display and measure distance within and between drill holes.~~

Formatted: Numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0.63 cm + Indent at: 1.27 cm

Formatted: Keep with next

The projection system used in the calculation is based on the input specified in the configuration file (Sect. 2.4, Appendix C1)

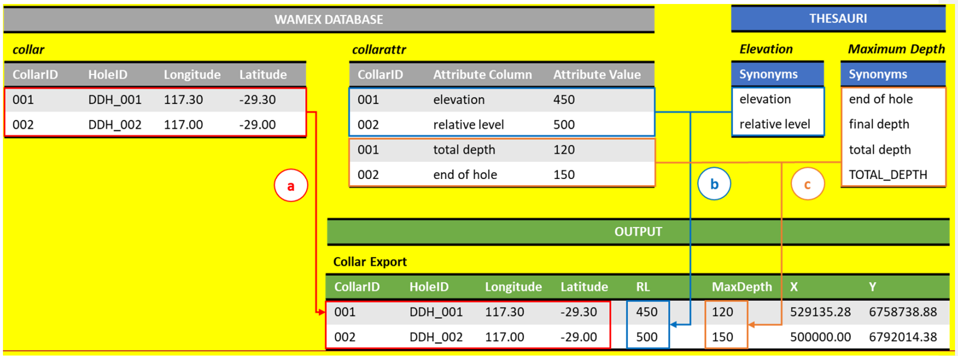


Figure 5. Collar extraction workflow showing the `CollarID`, `HoleID`, `Longitude` and `Latitude` information is fetched from the *`collar`* table (a, red), the corresponding RL and `MaxDepth` values are fetched from the *`collarattr`* table using the Elevation (b, blue) and Maximum Depth thesauri (c, orange).

2.4.2 Survey Extraction

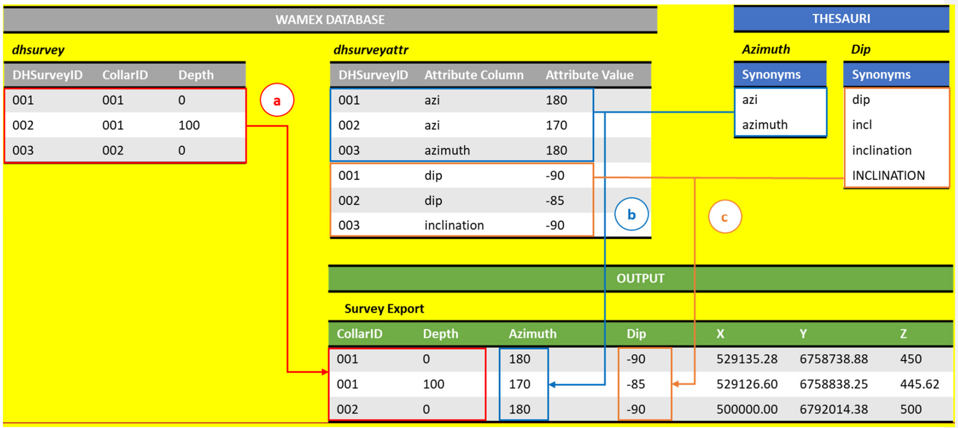
With the same inputs defined in the configuration file, the *dh2loop* library outputs a survey CSV file containing the following information (Fig. 6):

1. `CollarID`: The primary key to link the survey information to the collar file.
2. `Depth`: This is from the survey table and refers to the downhole depth where the survey measurement was taken. All non-numeric values were ignored. If the depth value fetched is negative, the absolute value is taken as it may have been used to denote direction of drilling. This assumption was made as some drill holes have negative depth information that is technically not possible to have a negative length. This was done by some companies to denote that the depth measure was going upwards (usually for underground probing drill holes) rather than downhole. To put all the drill holes in the same cartesian plane, this correction was done.

3. `Azimuth`: It is the trend direction indicated by an angle between 0-360 degrees from the north going clockwise. This is extracted using the drill hole azimuth thesaurus (Fig. 6b). The code fetches values between 0-360 degrees, thus ignoring non-numeric value and values greater than 360. Values between -360 to 0 are assumed to be counter-clockwise from the north. If there is no survey information for a drill hole present in collar, the azimuth value is set to 0.

4. `Dip`: It is the inclination angle perpendicular to the azimuth indicated by an angle between -90 to 90. It is measured from the horizontal plane, thus a positive value would describe a drill hole deviating to the surface or points upward while a negative value would be a drill hole pointing downwards. This is extracted using the drill hole dip thesaurus (Fig. 6c).
5. Calculated `X`, `Y`, `Z` values: These values are the project location of the survey measurement. This is calculated using the minimum curvature (also called spherical arc) algorithm. The minimum curvature algorithm (Amorin, 2009) desurveys downhole distances as distances along a circular arc. The algorithm matches the survey at two consecutive measurements exactly and the curvature is constant between these two measurements. The direction remains continuous meaning there are no sharp changes in direction. The code for the minimum curvature was based on the *pyGSLIB* drill hole module.

Formatted: Normal, No bullets or numbering



~~**Figure 6. Survey extraction workflow showing the `DHSurveyID`, `CollarID` and `Depth` information is fetched from the *`dhsurvey`* table (a, red), the corresponding Azimuth and Dip values are fetched from the *`dhsurveyattr`* table using the Azimuth (b, blue) and Dip thesauri (c, orange).**~~

~~**2.4.3 Lithology Extraction**~~

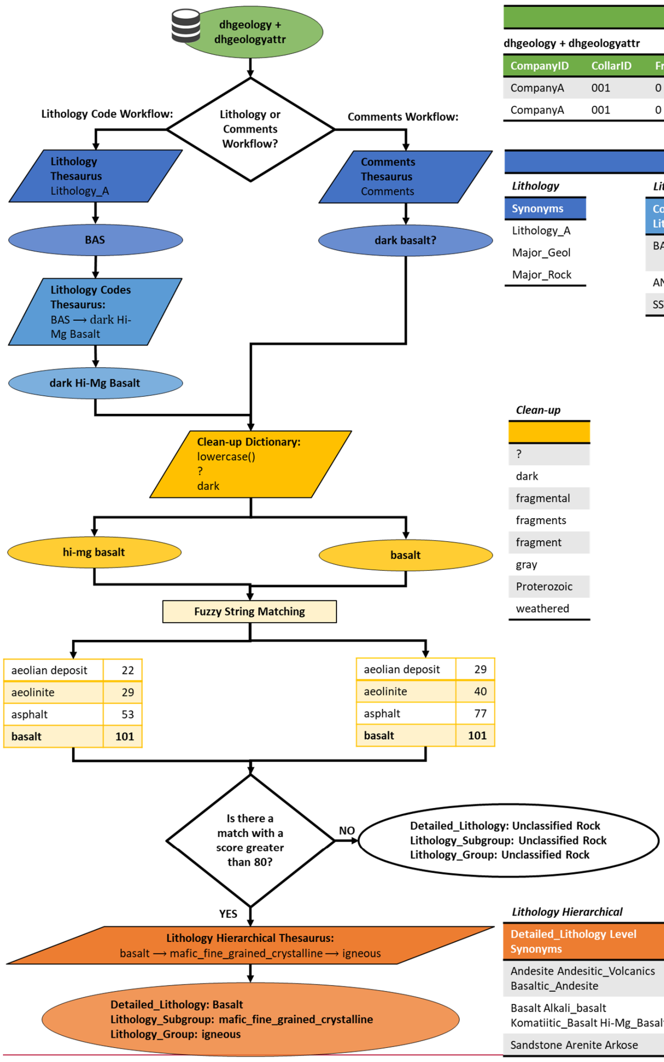
The lithology extraction ~~outputs~~is divided into two workflows: **Lithology Code Workflow** and **Comments Workflow**. Both workflows output a lithology CSV file containing the following information (Fig. ~~7):~~2):

1. `CompanyID`: The primary key to link the lithology code to the ~~drill hole lithology codes thesaurus~~Drill Hole Lithology Codes Thesaurus and decode the lithologies.
2. `CollarID`: The primary key to link the lithology information to the collar file.
3. `FromDepth` and `ToDepth`:~~ The start/from and end/to downhole depth values.~~ If the `ToDepth` is null, we assume `ToDepth` to be equal to `FromDepth` + 0.01. If the `FromDepth` is larger than `ToDepth`, the `FromDepth` and `ToDepth` values are switched.
4. `Detailed_Lithology`: This value is the lithology matched through fuzzy string matching. The string that serves as input to the fuzzy string matching may either be the `Company_Litho` (decoded lithology from `Company_LithoCode`) or from the `Comments` (free text descriptions).
   4.1. Decoding Lithological Codes
      4.1.1. `Company_LithoCode`: This fetches the lithology codes that are typically three-letter codes using the ~~drill hole lithology thesaurus~~Drill Hole Lithology Thesaurus.
      4.1.2. `Company_Litho`: The `Company_Litho` is fetched by matching the `CompanyID` and `Company_LithoCode` to the ~~drill hole lithology codes thesaurus~~Drill Hole Lithology Codes Thesaurus.
   4.2. `Comments`: This fetches the free text descriptions using the ~~drill hole comments thesaurus~~Drill Hole Comments Thesaurus.
5. `Lithology_Subgroup` and `Lithology_Group`: Upscales the lithological information to more generic rock groups. For example, `Detailed_Lithology`: "basalt" is upscaled to `Lithology_Subgroup`:"mafic_fine-grained crystalline" and further upscaled to `Lithology_Group`:"igneous rock".
6. Calculated `X`, `Y`, `Z` for the start, mid and endpoint also using the minimum curvature algorithm. The desurveying code ~~was~~is heavily based on the *pyGSLIB* drill hole module.

Formatted: Font: Lucida Console

Formatted: Font: Lucida Console

Formatted: Font: Lucida Console



OUTPUTS

dhgeology + dhgeologyattr

CompanyID	CollarID	FromDepth	ToDepth	Attribute Column	Attribute Value
CompanyA	001	0	50	Lithology_A	BAS
CompanyA	001	0	50	Comments	dark basalt?

THESAURI

Lithology

Synonyms
Lithology_A
Major_Geol
Major_Rock

Lithology Codes

Company_ LithoCode	Company_ Litho
BAS	dark Hi-Mg basalt
AND	andesite
SST	sandstone

Comments

Synonyms
Comment
Comments
DESC
description

Clean-up

?
dark
fragmental
fragments
fragment
gray
Proterozoic
weathered

Lithology Hierarchical

Detailed_Lithology Level Synonyms	Lithology_Subgroup Level	Lithology_Group Level
Andesite Andesitic_Volcanics Basaltic_Andesite	intermediate_fine_grained_crystalline	igneous
Basalt Alkali_basalt Komatiitic_Basalt Hi-Mg_Basalt	mafic_fine_grained_crystalline	igneous
Sandstone Arenite Arkose	clastic	sedimentary

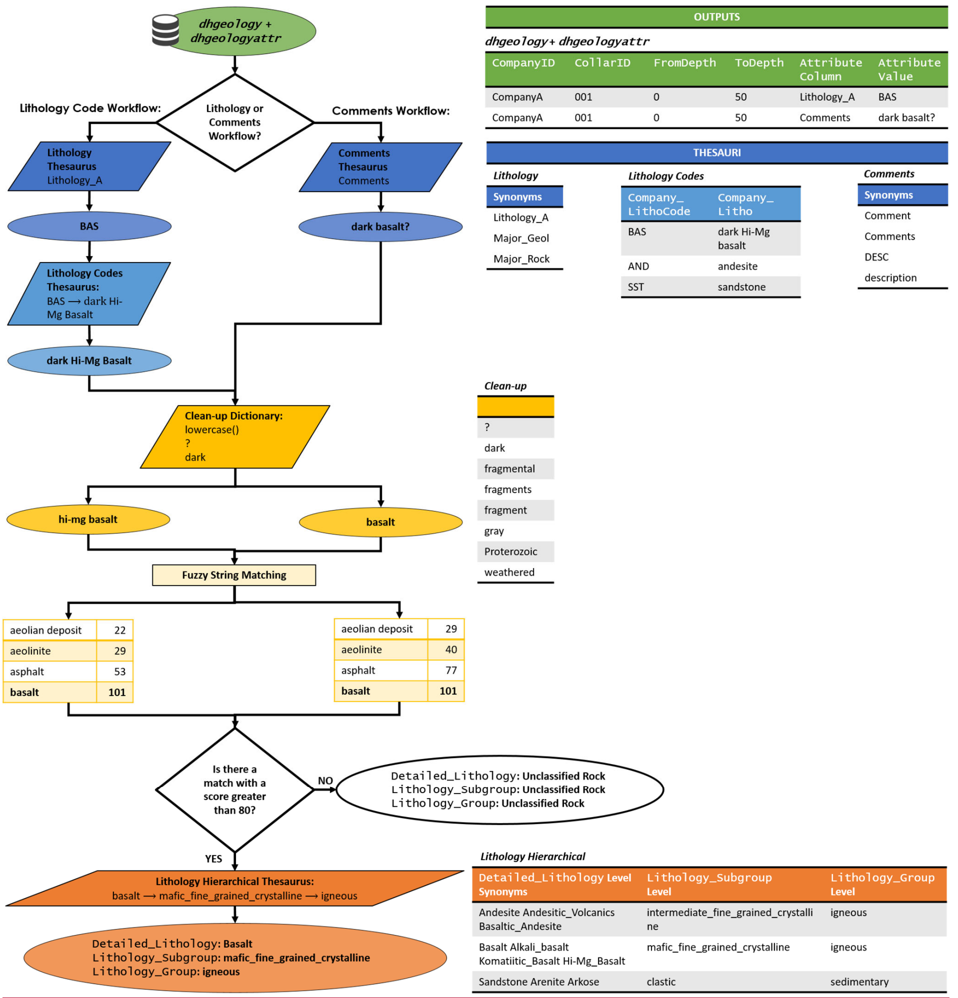


Figure 7.2. Lithology extraction is done through the Lithology Code workflow and Comments workflowsworkflow. The values are fetched from the *dhgeology* and *dhgeologyattr* table (green) using either the Drill Hole Lithology Thesaurus (blue) and Drill Hole Lithology CodeCodes Thesaurus (light blue) thesauri or the Drill Hole Comments Thesaurus (blue) thesaurus.). The string fetched is then cleaned prior to the fuzzy string matching using the CleanupClean-up Dictionary (dark yellow). The result is then matched against the Detailed_Lithology level of the Lithological Hierarchical Thesaurus. If there is a match with a score greater or equal to 80, the match is taken and matched with the rest of the columns in the Lithology Hierarchical thesaurus. If not, it is labelled as unclassified rock.

Formatted: Font: Century Gothic

Formatted: Font: Century Gothic

Once the `Company_Litho` (decoded lithology from `Company_LithoCode`~~)~~ or from the `Comments` (free text descriptions) have been extracted from the database, the lithology strings ~~were~~are pre-processed such that:

a) The strings ~~were~~are converted to lowercase form.
b) The string inside parenthesis, brackets and braces ~~were~~are removed, as these ~~were~~are found to reduce the accuracy of the matching.
c) The string ~~followed~~preceded by key phrases such as "with", "possibly", "similar to" ~~were~~are removed.
d) If any of the words listed in the ~~clean~~Clean-up ~~dictionary were~~Dictionary are present in the string, these words ~~were~~are removed.
e) Lemmatization, the removal of the inflections at the end of the words in order the "lemma" or root of the words, ~~was~~is applied to all nouns ~~(Müller et al., 2015).~~(Müller et al., 2015).
f) All words with non-alphabetic characters and tokens with less than three characters ~~were removed.~~are removed. This include two-letter words such as "to", "in", "at".
~~g) Stopwords, a set of words frequently used in language which are irrelevant for text mining purposes (Wilbur and Sirotkin, 1992), were removed. Examples on stopwords are: as the, is, at, which, and on.~~

g) Stopwords, a set of words frequently used in language which are irrelevant for text mining purposes (Wilbur and Sirotkin, 1992), are removed. Examples on stopwords are: as the, is, at, which, and on.

This is followed by fuzzy string matching, ~~an algorithm which~~a technique that finds the string that matches a pattern approximately. Fuzzy string matching is typically divided into two sub-problems: 1) finding approximate substring matches inside a given string, and 2) finding dictionary strings that match the pattern approximately. Fuzzy string matching uses the Levenshtein Distance to calculate the differences between sequences and patterns ~~(Okuda et al., 1976;Cohen, 2011).~~(Okuda et al., 1976; Cohen, 2011). The Levenshtein distance measures the minimum number of single-character edits (insertion, deletion, substitution) necessary to convert a given string into an exact match with the dictionary string ~~(Levenshtein, 1965)~~(Levenshtein, 1965).

We ~~utilized~~utilize *fuzzywuzzy* (https://github.com/seatgeek/fuzzywuzzy) for this. *fuzzywuzzy* provides two methods to calculate a similarity score between two strings: *ratio()* or *partial_ratio()*. It also provides two functions to pre-process the strings: *token_sort()* and *token_set()*. In this work, we used the *token_set_ratio()* scorer to do fuzzy string matching to classify the ~~database lithology description~~`Company_Litho` or `Comments` entries into one of the ~~lithology thesaurus~~Lithology Hierarchical Thesaurus entries (Table 1~~Table 1~~). *token_set()* pre-processes the strings by: 1) splitting the string on white-spaces (tokenization), 2) turning to lowercase and 3) removing punctuations, non-alpha non-numeric characters and unicode symbols. It tokenizes both strings (given string and dictionary string), splits the tokens into: intersection and remainder, then sort and compare the strings. The sorted intersection component refers to the similar tokens between the two strings. Since the sorted intersection component (similar tokens between two strings) of token*_set()*, will result in an exact match, the score will tend to increase when: 1) the sorted intersection makes up a larger percentage of the full string, and 2) the remainder component are more similar. The *ratio()* method then computes the standard Levenshtein distance between two strings. *token_set_ratio()* ~~was~~is found to be effective in addressing harmless misspelling and duplicated words but sensitive enough to calculate lower scores for longer strings (3-10 word labels), inconsistent word order and missing or extra words. *partial_ratio()* which takes the "best partial" of two strings or the best matching on the shorter substring ~~was~~is not preferred as it does not address the difference and order in substring construction. *token_sort()* ~~was~~is not preferred as it alphabetically sorts the tokens that ignores word order and does not weight intersection tokens which does not address the behavior of the strings in the logs.

Table 1. Examples of fuzzy string matching output using different combinations of the *fuzzywuzzy* functions. ~~The table demonstrates the corresponding effect of these functions to the given string. *Token_set_ratio ()* works best on geological free text descriptions as it weights the intersection tokens, honors substring construction and word order and ignores misspelling, extra and duplicated words. *Partial_ratio ()* ignores substring construction and order and is more sensitive to misspellings. *Token_sort_ratio ()* also ignores substring order and does not recognize duplicate and extra words~~The ticks and crosses indicated beside the score indicates the preferred (ticks) result between the methods clustered together.

fuzzywuzzy Function	Given String	Dictionary String	Score		Remarks
ratio ()	diorite	granodiorite rock	58	✓	*partial_ratio ()* ignores substring construction
partial_ratio ()	diorite	granodiorite rock	100	✗	
ratio ()	granodoirit rcok	granodiorite rock	85	✓	*ratio ()* mitigates misspelling
partial_ratio ()	granodoirit rcok	granodiorite rock	81	✗	
ratio ()	rock felsic granodiorite	granodiorite rock	59	✓	*partial_ratio ()* ignores substring order
partial_ratio ()	rock felsic granodiorite	granodiorite rock	83	✗	
token_set_ratio ()	rock felsic granodiorite	granodiorite rock	83	✓	*token_sort_ratio ()* ignores substring order
token_sort_ratio ()	rock felsic granodiorite	granodiorite rock	100	✗	
token_set_ratio ()	intermediate granodiorite rock	granodiorite rock	100	✓	*token_set_ratio ()* weights intersection tokens
token_sort_ratio ()	intermediate granodiorite rock	granodiorite rock	72	✗	
token_set_ratio ()	gray granodiorite granodiorite	granodiorite rock	83	✓	*token_set_ratio ()* ignores extra and duplicate words
token_sort_ratio ()	gray granodiorite granodiorite	granodiorite rock	64	✗	
token_set_ratio ()	**gray granodiorite granodiorite rckso**	**granodiorite rock**	**83**	✓	***token_set_ratio ()* weights intersection tokens, addresses substring construction and word order, ignores misspelling, extra and duplicate words**
partial_token_set_ratio ()	gray granodiorite granodiorite rckso	granodiorite rock	100	✗	

~~The code~~*dh2loop* calculates the t*oken_set_ratio()* between the `Company_Litho` or `Comments` (given string) and the entries in the ~~lithology hierarchical thesaurus~~Lithology Hierarchical Thesaurus (dictionary string). The tendency ~~of geologists when describing rocks~~ is to enumerate the descriptors before the rock name. For example, if the lithology in the logged interval is "basalt", the free text description could be something like "Dark gray to dark reddish brown, with olivine phenocrysts, largely altered andesitic basalt". After processing the string, it will be left with "andesitic basalt". To avoid, misclassifying the rock to "andesite", a bonus score is also added to add weight to the last word (in this case, "basalt~~") (Appendix C2 Pseudocode).~~"). Furthermore, the reader may worry that "basaltic andesite" will be simplified and classified into "andesite". Since "basaltic andesite" is an established volcanic rock name, it will remain as "basaltic andesite". For the pair between `Company_Litho` or `Comments` and the entries in the ~~lithology hierarchical thesaurus~~Lithology Hierarchical Thesaurus with the highest score, the first synonym is stored as `Detailed_Lithology`. If the score is less than 80, it is classified as "unclassified rock". The cut-off value ~~of 80~~ is user-defined~~,~~ and ~~in this case~~can be chosen based on the performance of the matching on ~~a~~the subset of ~~1,548 unique lithology codes (Fig. 8) from a subset of the YSGB dataset in~~ the ~~Golden Grove area. The matching performance may vary depending on the dataset being extracted. It is advised to test in a subset and adjust these cut-off score depending on these results~~desired region. If the performance is significantly lower, this indicates that the thesauri used in *dh2loop* may not be suitable to your area. The user may opt to update these thesauri to suit their needs. Once matched on `Detailed_Lithology`, the corresponding `Lithology_Subgroup` and `Lithology_Group` classifications are also fetched.

Formatted: Not Highlight

Field Code Changed

Formatted: Font: Italic

Formatted: Not Highlight

Formatted Table

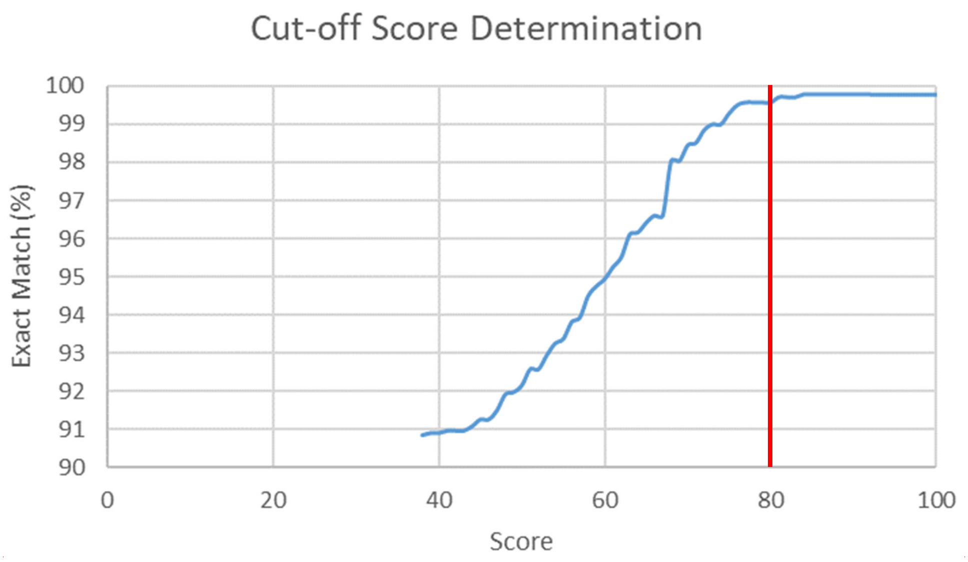


Figure 8. 2.5 Fuzzy String Matching Assessment

The user-defined cut-off score of 80 was chosen based on the results of the testing different cut-offs on a smaller dataset within the YSGB area. As seen in this figure, the number of exact matches plateau at a score of 80. This relationship may vary depending on the datasets available in the area. Thus, this cut-off value is user-defined and is best to test the matching performance on a subset in the user's area.

3 Data Extraction Results

3.1 Collar

Extraction of the collar data for YSGB resulted in a collar file with 68,729 drill holes (Table 2). This information was extracted from the *collar* table with 73,881 drill holes with 769,981 rows of information from *collarattr*. It includes the location of the collar both in geographic and projected coordinated systems, relative level (`RL`) and maximum depth (`MaxDepth`). A total of 136,100 records for `RL` were retrieved from the database, 1,526 of which were disregarded: 846 records for having an `RL` value greater than 10,000 meters and 680 non-numeric records. These discarded values were retrieved from the attribute column "RL_Local". In spite of it being an isolated issue for "RL_Local", the attribute column was retained as it is retrieved sensible values for other companies. The discarded values were limited to dataobjective is to compare the `Detailed_Lithology` classification results obtained from two companies (4085, 4670) for `RL` attribute columns "TD" and "DEPTH". A total of 58,706 records for `MaxDepth` were retrieved from the database: 58,642 of which were extracted as is, while 64 entries were disregarded for having a value of -999. The discarded values come from 8 companies. Null values are assigned to disregarded and absent `RL` or `MaxDepth` values. The "clean" collar export file contains at least either a value for `RL` or `MaxDepth`. The reasoning behind keeping records with at least one of the two field is there are other ways to extract for `RL` or `MaxDepth` from the database. `RL` values can be extracted from digital terrain models and `MaxDepth` values can be taken for the largest `ToDepth` values from the other tables.

Formatted: Font: Times New Roman

Formatted: Font: Times New Roman

Formatted: Font: Times New Roman

Formatted: Font: Times New Roman

Formatted: Font: Times New Roman

3.2 Survey

For the survey extraction, the *dhsurvey* table contained 146,713 survey depth intervals (from 45,708 drill holes) with corresponding 850,507 entries of supplementary survey information in *dhsurveyattr* (Table 3). Survey extraction in YSGB resulted in 126,669 survey depth information across 45,708 drill holes with azimuth (-52.5 to 359) and dip measurements (0-90) for each depth interval. A total of 517,592 records for `Azimuth` were retrieved from the database. 77 `Azimuth` values greater than 360 were retrieved and thus disregarded. 152 values were non-numeric values and were also disregarded.independent workflows: 1) These discarded values involved 228 holes across 10 companies. A value of 0 was assigned to missing `Azimuth` values. A total of 118,223 records for `Dip` were fetched from the database, 118,138 of which were extracted as is, while 95 entries were disregarded for having a value greater than 90. A values of -90 was assigned as the default for `Dip`. The discarded values correspond to 94 drill holes across 5 companies.

3.3 Lithology: Lithology Code **Workflow** **and 2)** Comments **workflow**

Lithology extraction is divided into two workflows. For the Lithology Code workflow, the extraction starts with filtering the *dhgeology* and *dhgeologyattr* table by the location extents and the Lithology thesaurus. The *dhgeology* table contained 47,062 drill holes across 115 companies with 797,975 lithology depth intervals with corresponding 820,612 entries of lithology information in *dhgeologyattr*. These records were matched with the entries from the Lithology Code thesaurus resulting to 273,684 matched records. The `FromDepth` and `ToDepth` for these records were then validated. 74 records had equal `FromDepth` and `ToDepth` values. 654 had values for `FromDepth` but null values for `ToDepth`. For both cases, `ToDepth` was calculated as `FromDepth`+0.01. The Lithology Code workflow resulted to 273,684 intervals across 12,793 drill holes wherein 235,606 records were successfully matched in the fuzzy string matching.

The **Comments workflow** extracts the records from the *dhgeology* and *dhgeologyattr* table as well, but this time using the Comments Thesaurus. For YSGB, the database has 262,567 records across 22,766 drill holes with comments. Since the comments are extracted here to compare their results from fuzzy string matching, only those records that matched in the Lithology Code workflow were retained. This resulted to 47,823 records, however, only 7,870 records were successfully matched on `Comments`. The dataset for the fuzzy string matching assessment (Sect. 5) consists only of the unique records matched on both workflows (3,074 records). It was visually checked from the records that the Lithology Code `Detailed_Lithology` results were sound classifications of the `Company_Lithology`. This was done to make sure that these results could be considered as the "true value" in the fuzzy string matching assessment (Sect. 5).

Formatted: Font: Century Gothic

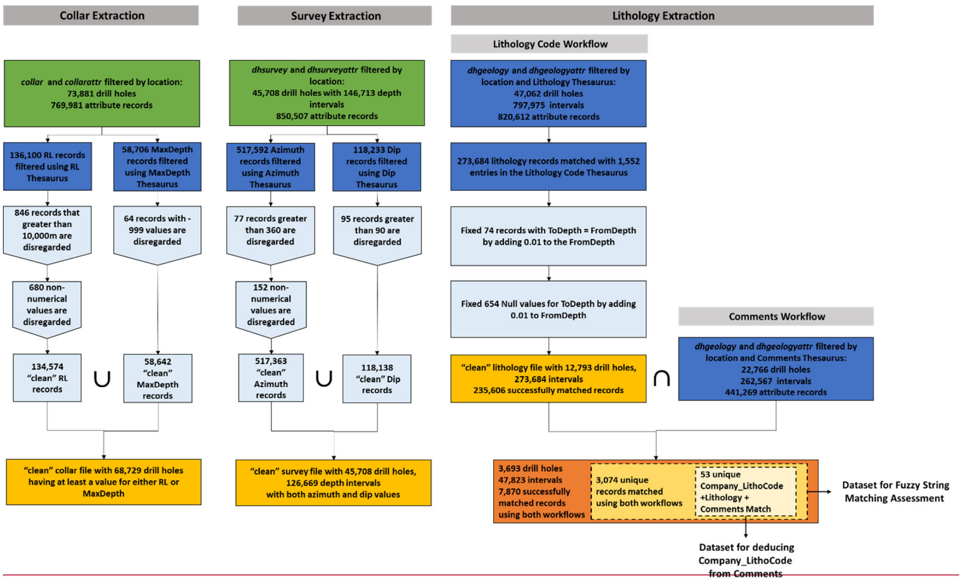
Formatted: Font: Century Gothic

Formatted: Font: Century Gothic

Formatted: Caption

Formatted: Font: Times New Roman, Not Italic

Formatted: Font: Century Gothic, Bold



~~Figure 9. Extraction of the collar, survey and lithology data for the YSGB. The collar extraction resulted in a collar file with 68,729 drill holes from the `collar` table with 73,881 drill holes with 769,981 rows of information from `collarattr`. A total of 136,100 records for RL were retrieved from the database, 1,526 of which were disregarded: 846 records for having an RL value greater than 10,000 meters and 680 non-numeric records. A total of 58,706 records for MaxDepth were retrieved from the database: 58,642 of which were extracted as is, while 64 entries were disregarded for having a value of -999. The "clean" collar export file contains at least either a value for RL or MaxDepth. Survey extraction in YSGB resulted in 126,669 survey depth information across 45,708 drill holes. The `dhsurvey` table contained 146,713 survey depth intervals (from 45,708 drill holes) with corresponding 850,507 entries of supplementary survey information in `dhsurveyattr`. 77 Azimuth values greater than 360 and 152 values were non-numeric values. Lithology extraction is divided into two workflows. For the Lithology Code workflow, the extraction starts with filtering the `dhgeology` and `dhgeologyattr` table by the location extents and the Lithology thesaurus. The `dhgeology` table contained 47,062 drill holes across 115 companies with 797,975 lithology depth intervals with corresponding 820,612 entries of lithology information in `dhgeologyattr`. These records were matched with the entries from the Lithology Code thesaurus resulting to 273,684 matched records. The FromDepth and ToDepth for these records were then validated. 74 records had equal FromDepth and ToDepth values. 654 had values for FromDepth but null values for ToDepth. For both cases, ToDepth was calculated as FromDepth+0.01. The Lithology Code workflow resulted to 273,684 intervals across 12,793 drill holes wherein 235,606 records were successfully matched in the fuzzy string matching. The Comments workflow extracts the records from the `dhgeology` and `dhgeologyattr` table as well, but this time using the Comments Thesaurus (262,567 records across 22,766 drill holes with comments). 47,823 records were present in both workflows, 7,870 records of which were successfully matched. The 3,074 unique entries from this was used as the dataset for the fuzzy string matching assessment (Sect. 5).~~

Formatted: Font: Lucida Console

Formatted: Font: Lucida Console

Formatted: Font: Lucida Console

Formatted: Font: Century Gothic

Formatted: Font: Lucida Console, Italic

Formatted: Font: Lucida Console

Formatted: Font: Lucida Console

Formatted: Font: Lucida Console

Formatted: Font: Lucida Console

~~4 Unique Lithology Code Results~~

Workflow. Using the Company_LithoCode, Company_Litho, **Lithology_ Code Workflow:** Detailed_Lithology and **Comments Workflow:** Detailed_Lithology from the dataset for the fuzzy string matching assessment, we can assess if matches using the **Comments workflow** alone can sufficiently decode lithology. ~~Excluding the unmatched entries and taking only the unique combinations of Company_LithoCode, Lithology_Code Detailed_Lithology and Comments Detailed_Lithology, the dataset results into 53 unique records.~~

To be able to assess the matching we take a look at the type of matches between **Lithology_Code Workflow:** Detailed_Lithology and **Comments Workflow:** Detailed_Lithology. First, we define a **match** as retrieving an answer from the fuzzy string matching with a score greater than 80. It is important to note here that it only suggests that it succeeded to find an answer above the score threshold but not necessarily mean that it is the correct answer. To further describe the quality of a match, we modified for this purpose the following terminologies from the Simple Knowledge Organization System ~~(Miles and Bechhofer, 2009)~~(Miles and Bechhofer, 2009):

a) **Exact Match** suggests that both ~~workflows~~**Lithology Code workflow** and **Comments workflow** resulted in the same classification at all 3 levels. The match at the Detailed_Lithology level has an exact match, thus resulting to an exact match on the other two levels.
b) **Close Match** suggests that the results at the Detailed_Lithology level are related rocks and belong to the same Lithology_Subgroup. This is usually caused by differing use of lithological nomenclature.
c) **Related Match** suggests that the results at the Detailed_Lithology level are related rocks and belong to the same Lithology_Group.
d) **Broad Match** refers to the Detailed_Lithology from **Lithology Code workflow** matches to a Lithology_Subgroup in the **Comments workflow**.
e) **Narrow Match** is the logical equivalent of a Broad Match. In this case, the **Comments workflow** resulted in a Detailed_Lithology level while the **Lithology Code workflow** resulted in a Lithology_Subgroup level.
f) **Broader Match** is similar to a broad match except that the Detailed_Lithology from **Lithology Code workflow** matches to a Lithology_Group instead of a Lithology_Subgroup in the **Comments workflow**.
g) **Narrower Match** is the logical equivalent of Broader Match. The **Comments workflow** results to a Detailed_Lithology while the **Lithology_Code workflow** results to a Lithology_Group level.
h) **Failed Match** suggests all levels of both workflows do not match. This is usually attributed to contrasting information from both fields or the algorithm fails. This category is an addition to the SKOS reference.

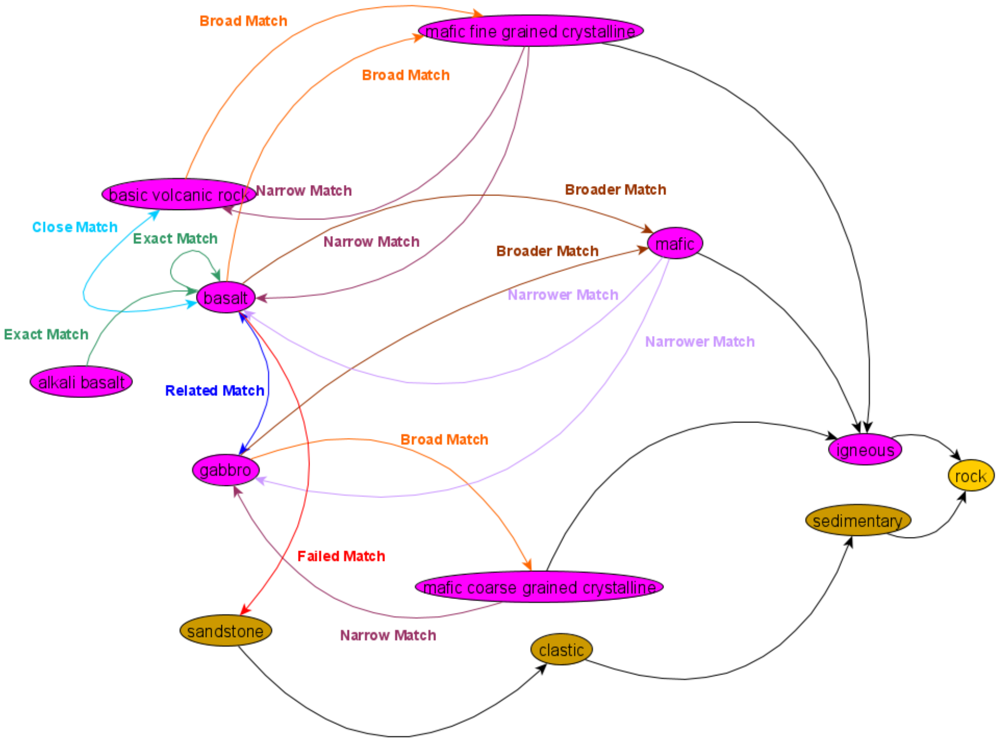
For better understanding of these relationships, examples are shown in Table 2 and Fig ~~9~~3.

Formatted: Font: Century Gothic, Bold
Formatted: Font: Century Gothic, Bold
Formatted: Font: Century Gothic, Bold
Formatted: Font: Century Gothic, Bold
Formatted: Font: Times New Roman
Formatted: Font: Century Gothic, Bold
Formatted: Font: Century Gothic, Bold
Formatted: Font: Century Gothic, Bold
Formatted: Font: Century Gothic, Bold
Formatted: Font: Century Gothic, Bold
Formatted: Font: Century Gothic, Bold
Formatted: Font: Century Gothic, Bold
Formatted: Font: Century Gothic, Bold
Formatted: Font: Century Gothic, Bold
Formatted: Font: Century Gothic, Bold
Formatted: Font: Century Gothic, Bold
Formatted: Font: Century Gothic, Bold

Table 2. Fuzzy string matching terminology used to describe the quality of matches based on the Simple Knowledge Organization System (SKOS) ~~(Miles and Beechhofer, 2009).~~(Miles and Bechhofer, 2009). The values being compared are the `Detailed_Lithology` **level for both** Lithology Code workflow **and** Comments workflow **~~(brown text)~~. The level at which the records are considered to match are in bold.** ~~**A Match retrieves an answer from the fuzzy string matching with a score greater than 80. An Exact Match suggests that both workflows resulted in the same classification at all 3 levels. A Close Match suggests that the results at the `Detailed_Lithology` level are related rocks and belong to the same `Lithology_Subgroup`. A Related Match suggests that the results at the `Detailed_Lithology` level are related rocks and belong to the same `Lithology_Group`. A Broad Match refers to the `Detailed_Lithology` from Lithology Code workflow matches to a `Lithology_Subgroup` in the Comments workflow. Narrow Match is the logical equivalent of a Broad Match. Broader Match is similar to a broad match except that the `Detailed_Lithology` from Lithology Code workflow matches to a `Lithology_Group` instead of a `Lithology_Subgroup` in the Comments workflow. Narrower Match is the logical equivalent of Broader Match. A Failed Match suggests all levels of both workflows do not match.**~~

Lithology Code Workflow: Detailed_ Lithology ~~Level~~	Comments Workflow: Detailed_ Lithology ~~Level~~	Lithology Code Workflow: Lithology_ Subgroup ~~Level~~	Comments Workflow: Lithology_ Subgroup ~~Level~~	Lithology Code Workflow: Lithology_ Group ~~Level~~	Comments Workflow: Lithology_ Group ~~Level~~	**Type of Match**
basalt	**basalt**					Exact Match
basalt	~~basic volcanic rock~~basaltoid	**mafic fine grained crystalline**	**mafic fine grained crystalline**			Close Match
basalt	gabbro	mafic fine grained crystalline	mafic coarse grained crystalline	**igneous**	**igneous**	Related Match
basalt	mafic fine grained crystalline	**mafic fine grained crystalline**	**mafic fine grained crystalline**			Broad Match
mafic fine grained crystalline	basalt	**mafic fine grained crystalline**	**mafic fine grained crystalline**			Narrow Match
basalt	mafic	mafic fine grained crystalline	mafic	**igneous**	**igneous**	Broader Match
mafic	basalt	mafic	mafic fine grained crystalline	**igneous**	**igneous**	Narrower Match
basalt	sandstone	mafic fine grained crystalline	clastic	igneous	sedimentary	Failed Match

Field Code Changed
Formatted: Font: Century Gothic
Formatted: Font: Century Gothic
Formatted: Not Highlight
Formatted: Font: Century Gothic, Highlight
Formatted: Font: Century Gothic, Highlight
Formatted: Font: Century Gothic, Highlight
Formatted: Font: Century Gothic, Highlight
Formatted: Font: Century Gothic, 10 pt, English (United Kingdom)
Formatted: Font: Century Gothic, 10 pt, English (United Kingdom)
Formatted: Font: Century Gothic, 10 pt, English (United Kingdom)
Formatted: Font: Century Gothic, 10 pt, English (United Kingdom)
Formatted: Font: Century Gothic, 10 pt, English (United Kingdom)
Formatted: Font: Century Gothic, 10 pt, English (United Kingdom)
Formatted: Font: Times New Roman
Formatted: Font: Lucida Console, 10 pt, English (United Kingdom)
Formatted: Font: Lucida Console, 10 pt, English (United Kingdom)
Formatted: Font: Lucida Console, 10 pt, English (United Kingdom)
Formatted: Font: Lucida Console, 10 pt, English (United Kingdom)
Formatted ... [1]
Formatted ... [2]
Formatted ... [4]
Formatted ... [6]
Formatted ... [3]
Formatted ... [5]
Formatted: Font: +Body (Calibri)
Formatted Table
Formatted: Font: +Body (Calibri)
Formatted: Font: +Body (Calibri)
Formatted: Font: +Body (Calibri)
Formatted Table
Formatted: Font: +Body (Calibri)
Formatted: Font: +Body (Calibri)
Formatted: Font: +Body (Calibri)
Formatted: Font: +Body (Calibri)
Formatted: Font: +Body (Calibri)



Broad Match
mafic fine grained crystalline
Broad Match
basic volcanic rock
Narrow Match
Broader Match
Close Match
Exact Match
Narrow Match
mafic
Broader Match
basalt
Narrower Match
Exact Match
Narrower Match
alkali basalt
Related Match
Broad Match
igneous
gabbro
rock
sedimentary
Failed Match
mafic coarse grained crystalline
sandstone
Narrow Match
clastic

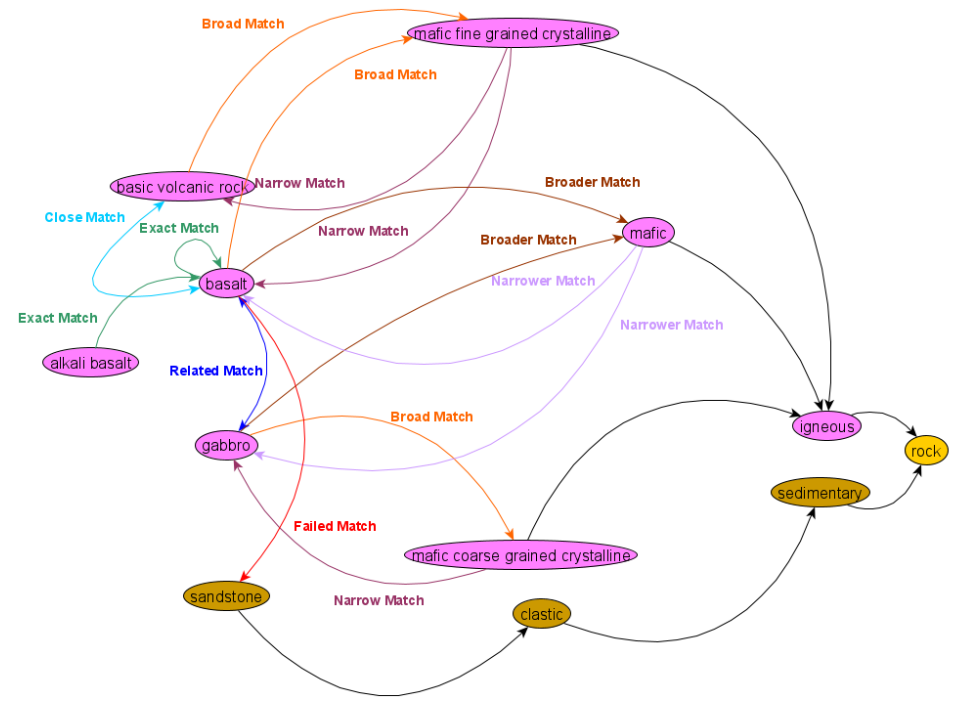


Figure ~~10.~~3. SKOS graph showing the semantic, associative and hierarchical relationship in the Lithology Hierarchical Thesaurus. In this example, terms "basalt" and "alkali basalt" are judged to be sufficiently the same to assert an Exact Match relationship (in green). "basic volcanic rock" however is considered a Close Match (in cyan) and "gabbro" a Related Match (in blue). "mafic fine grained crystalline" and "mafic coarse grained crystalline" are broader concepts, thus considered a Broad Match (in orange) to "basalt" and "gabbro" respectively. Broader Match (in brown) are similar to Broad Matches but are used to refer a wider semantic difference between the two concepts. Narrow Matches (in light purple) and Narrower Matches (in dark purple) are the logical equivalent of Broad Match and Broader Match. Failed Matches is used to describe unrelated matches.

~~34 out of the 53 unique entries (64%) result to matches between the `Detailed Lithology` and `Comments Detailed_Lithology`.~~

~~**26 of which are Exact Matches, 19 unique entries are Close Matches and 26% percent are Failed Matches. The Failed Matches are due to unrelated descriptions in the Comments field which was used to obtain the results in `Comments_Detailed_Lithology`. An example of this is the interval is logged as "ironstone" but the comments contain "mafic schist". Another less common reason is the `Company_LithoCode` is repeated in the `Comments`. An example of this is would be an interval logged as "colluvium" and the Comments as "COL". The Comments workflow will result to "coal" instead.**~~

~~## 5 Fuzzy String Matching Results~~

~~We present results from the data extraction using both workflows: **Lithology Code** and **Comments**. The overlaps between these two workflows suggest that the user may need to make choices to identify which is better suited for matching in their area of interest. To better understand the difference between these results, we take a looks at the matching overlaps between the two workflows (3,074 entries). These matching overlaps are used to compare and describe the fuzzy string matching using the decoding the lithology codes and using the comments free text descriptions.~~

~~**Exact Matches**: Of the total matched entries, 944 were Exact Matches (31%) (Table). The Exact Matches are ideal outcomes as both workflows resulted in exactly the same answers.~~

Formatted: Caption, Line spacing: 1.5 lines

Formatted: Font: Times New Roman

Formatted: Font: Times New Roman

Formatted: Font: Century Gothic

Formatted: Heading 2, No bullets or numbering

Formatted: Font: Century Gothic, Bold

Formatted: Font: Century Gothic, Bold

Close Matches: The Close Matches are common for coarse-grained igneous rocks, clastic sedimentary rocks, surficial residual rocks and filling structures. The coarse-grained igneous rocks such as gabbro, gabbroid and dolerites are used interchangeably in both fields. `Comments` can contain terminologies such as "gabbroic", "granophyric gabbro to dolerite", "intrusive granitoid to gabbro" resulting to close matches. Similar cases are observed between granodiorite and granite and between peridotite and coarse-grained ultramafic rocks. For clastic sedimentary rocks, the Close Matches are a result of gradation of grain size in the comments field. For example, an interval logged as mudstone is then described in the comments as "mudstone to sandstone" or "intercalated with siltstone". These comments will result in "sandstone" and "siltstone", respectively. Both clastic sedimentary rocks but not an Exact Match to mudstone. Metasediments and quartz veins occur together and what is described last dictates the `Detailed_Lithology` classification. Surficial rocks such as soil, duricrust, colluvium, laterite, calcrete, ferricrete and cover are used loosely or occur together resulting to multiple combination of these Close Matches.

Related Matches: 60 entries (3%) resulted in related matches. For igneous rocks, this result is observed when the comments field use rock type descriptors such as "komatitic", "basaltic" and "doleritic". An example would be an interval logged as dolerite and is then described in the comments as "dolertic basalt". This would result in dolerite in the **Lithology Code workflow** and "basalt" in the **Comments workflow**. Both lithologies are igneous, however have different composition and textural implications. For sedimentary rocks, **Lithology Code workflow** results to sedimentary rocks classified based on grain size as they have been logged ("gravel", "mud"). The comments field contains compositional descriptions such as "with silcrete" or "minor chert". In this case, the comments workflow will result in "silcrete" and "chert". Both workflows will result in sedimentary rocks, but the Lithology Code workflow will result in "clastic" rocks while the comments workflow will classify these to "siliceous" at the `Lithology_Subgroup` level. The related matches for structures occur across coincident lithologies such as "mylonite", "vein", "fault" and "breccia" which could either be "fillings" or "fault_rock" at the `Lithology_Sugbroup`.

Broad and Narrow Matches: No broad matches were noted and only one narrow match was obtained (Table 3). The interval was logged as "ironstone" with "BIF" in the comments, "ironstone" being a more general description for "banded iron formation".

Broader and Narrower Matches: More common cases are Broader and Narrower Matches indicate that there is a bigger relationship gap between the data in the lithology and comments field. Broad matches are a result of low detail comments. For example, an interval logged as "gabbro" is described as "medium-grained mafic", "massive mafic", "rich mafic". The inverse is noted for narrower matches, the interval is logged as "sediment" but in the comments the interval is described as "siliceous sediments".

Failed Matches: 1,694 entries resulted in Failed Matches (55%). Failed Matches occur when the lithology and comments field contain different information. This could be because of the lithology contains the main lithology while the comments contains all other lithologies intercalated in the interval. Another reason is the lithology field is relogged based on adjacent intervals without amending the comments. "Mudstone" had failed matches with a wide range of lithologies, such as: "amphibolite", "dolerite", "saprolite", "duricrust", "laterite", "banded iron formation", "chert", "phyllite", "schist", "vein". The same is observed for igneous rocks such as: "coarse-grained-ultramafic-rock". For "chert", the failed matches are within a range of sedimentary rocks: "alluvium" and "mud", "amphibolite" and "massive sulphide", "carbonate", "vein", "pegmatite".

Formatted: Font: Century Gothic, Bold

Formatted: Font: Century Gothic, Bold

Formatted: Font: Century Gothic, Bold

Table 3. Distribution of matches across the Fuzzy String Matching Dataset. A total of 45% of the unique records were matched reasonably, 31% of which are Exact Matches, 6% Close Matches, 3% Related Matches, 3% Broader Matches and 3% Narrower Matches.

The matching results can be visualized as confusion matrices, which are typically used in machine learning to compare the performance of an algorithm versus a known result. In this case, we are comparing the performance of the string matching using the **Comments workflow** against the results from the **Lithology Code workflow**. From the 3,074 unique records, we use a total of 1,200 samples for the confusion matrices. The reason for this difference is the limitation of building a confusion matrix wherein both workflows look at the same classes. Each row of the matrix represents the matched lithology from the **Comments workflow** while each column represents the matched lithology from the **Lithology Code workflow**. The diagonal elements represent the count for which the **Comments workflow** class is equal to the **Lithology Code workflow**. The off-diagonal elements are those that are misclassified by the **Comments workflow**. The higher the diagonal values of the confusion matrix the better, indicating many correct matches. The confusion matrices show normalisation by class support size. This kind of normalisation addresses the class imbalance and allowallows better visual interpretation of which class is being misclassified. The color of the cell represents the normalised count of the records to address the uneven distribution of records across different classes. Relying on one metric to assess the matching can be misleading, therefore, we would like to use four metrics: accuracy, precision, recall and F1 score. It is worth mentioning that a small support influences the precision and/or recall. However, this is the nature of using real-world geological logs as more detail is given to particular lithologies or areas depending on the interest of the study.

3 Case Study: Yalgoo-Singleton Greenstone Belt

3.1 Study Area

In this paper, we demonstrate the application of *dh2loop* to data from the Yalgoo-Singleton greenstone belt (YSGB) (Fig. 4), a geologically complex, largely heterogeneous and highly mineralized arcuate granite-greenstone terrane, in the western Youanmi Terrane, Yilgarn Craton in Western Australia (Anand and Butt, 2010). The YSGB has good range of different lithologies in the area. Igneous rocks occur as extensive granitoid intrusions emplaced between 2700 and 2630 Ma (Myers, 1993), as well as ultramafic to mafic volcanic rocks formed as extensive submarine lavas and local eruptive centres of felsic and mafic volcanic rocks. Some layered gabbroic sills intruding the greenstone are also observed. Sedimentary rocks formed in broad basins during tectonic and volcanic quiescence consist of mostly banded iron formation (BIF) and felsic volcaniclastic rocks. The greenstone belt is metamorphosed to greenschist facies (Barley et al., 2008). The area is also covered by deeply weathered regolith which conceals mineral deposits hosted by the underlying bedrock. Regolith contains signatures of mineralisation that are distal signatures of possible economically significant deposits (Cockbain, 2002). Furthermore, the YSGB is a major target for exploration as it has considerable resources of gold, nickel, bauxite, as well as lesser amounts of a wide range of other commodities (Cockbain, 2002). It hosts multiple mineral deposits ranging from volcanogenic massive sulphide (Golden Grove, Gossan Hill), orogenic gold (Mt. Magnet), banded iron formations (Mount Gibson, Karara, Extension Hill). The geological and structural complexity, including its relevance to mineral exploration makes the YSGB a reasonable and sensible area to test the *dh2loop* thesauri, matching and upscaling.

Formatted: Font: Century Gothic, Bold

Formatted: Font: Century Gothic, Bold

Formatted: Font: Century Gothic, Bold

Formatted: Font: Century Gothic, Bold

Formatted: Font: Century Gothic, Bold

Formatted: Font: Century Gothic, Bold

Formatted: Font: Century Gothic, Bold

Formatted: Space Before: 12 pt, After: 12 pt, Line spacing: single, Don't keep lines together

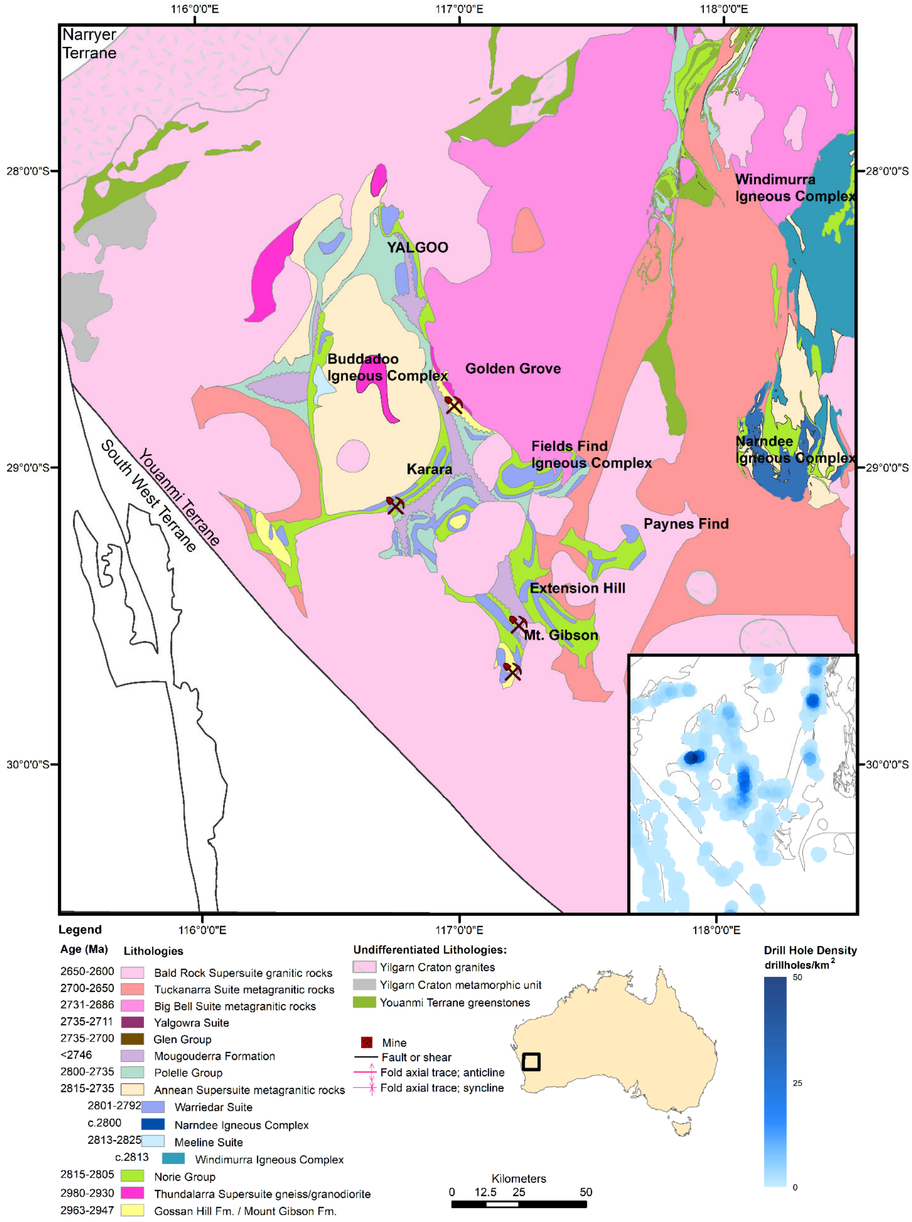


Figure 4. The map shows the Yalgoo-Singleton greenstone belt highlighting the different mines and prospects in the area. The inset map shows the heterogeneous distribution and drill hole density from the legacy data available from the WAMEX database.

extracted from the *collar* table with 73,881 drill holes with 769,981 rows of information from *collarattr*. It includes the location of the collar both in geographic and projected coordinated systems, relative level (RL) and maximum depth (MaxDepth). The discarded values come from 8 companies. disregarded and absent RL or MaxDepth values. The "clean" collar export file contains at least either a value for RL or MaxDepth. The reasoning behind keeping records with at least one of the two field is there are other ways to extract for RL or MaxDepth from the database. RL values can be extracted from digital terrain models and MaxDepth values can be taken for the largest ToDepth values from the other tables.

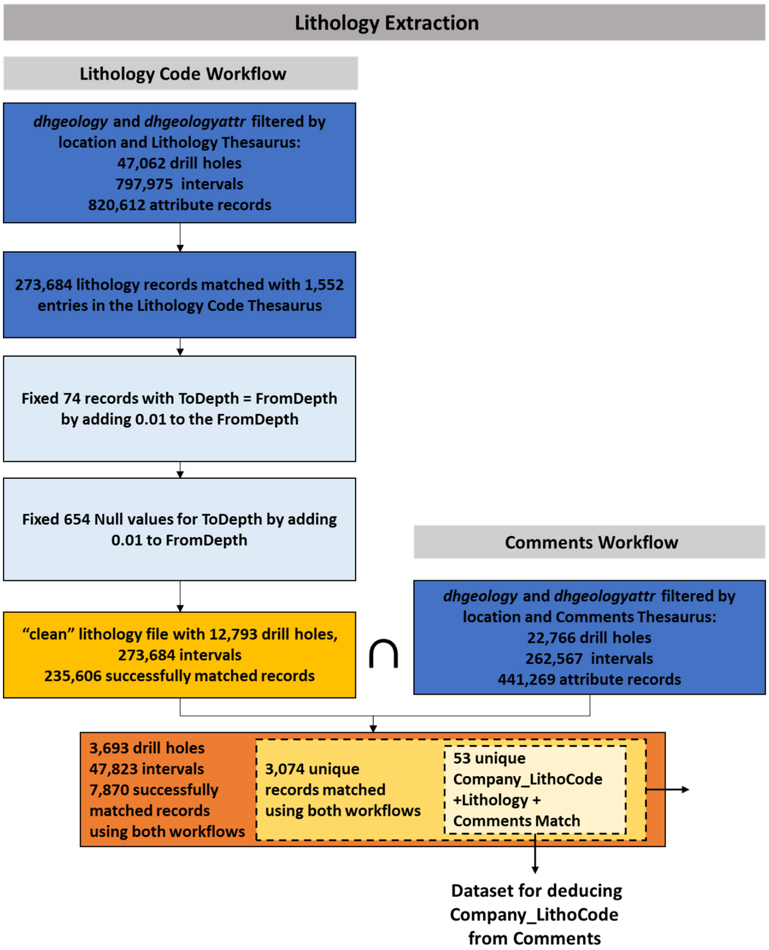


Figure 5. Extraction of the collar, survey and lithology data for the YSGB. . The collar extraction resulted in a collar file with 68,729 drill holes from the *collar* table with 73,881 drill holes with 769,981 rows of information from *collarattr*. The "clean" collar export file contains at least either a value for RL or MaxDepth. Survey extraction in YSGB resulted in 126,669 survey depth information across 45,708 drill holes. The *dhsurvey* table contained 146,713 survey depth intervals (from 45,708 drill holes) with corresponding 850,507 entries of supplementary survey information in *dhsurveyattr*. 77 Azimuth values greater than 360 and 152 values. Lithology extraction is divided into two workflows. For the Lithology Code workflow, the extraction starts with filtering the *dhgeology* and *dhgeologyattr* table by the location extents and the Drill Hole Lithology Thesaurus. The *dhgeology* table contained 47,062 drill holes across 115 companies with 797,975 lithology depth intervals with corresponding 820,612 entries of lithology information in *dhgeologyattr*. These records are matched with the entries from the Lithology Code

Formatted: Font: Times New Roman, Highlight
Formatted: Font: Times New Roman, Highlight
Formatted: Font: Times New Roman, Highlight
Formatted: Font: Times New Roman, Highlight
Formatted: Font: Times New Roman, Highlight
Formatted: Not Highlight
Formatted: Not Highlight
Formatted: Not Highlight
Formatted: Font: Lucida Console
Formatted: Not Highlight
Formatted: Font: Lucida Console
Formatted: Not Highlight
Formatted: Font: Lucida Console
Formatted: Not Highlight
Formatted: Font: Century Gothic
Formatted: Not Highlight
Formatted: Font: Times New Roman, Not Italic
Formatted: Not Highlight

thesaurus. 74 records had equal FromDepth and ToDepth values. 654 had values for FromDepth but null values for ToDepth. The Lithology Code workflow resulted to 235,606 records successfully matched in the fuzzy string matching. The Comments workflow extracts the records from the *dhgeology* and *dhgeologyattr* table as well, but this time using the Drill Hole Comments Thesaurus. 47,823 records are present in both workflows, 7,870 records of which are successfully matched. The 3,074 unique entries from this is used as the dataset for the fuzzy string matching assessment.

These discarded values involved 228 holes across 10 companies. . The discarded values correspond to 94 drill holes across 5 companies.

3.2 Lithology Extraction: Lithology Code Workflow and Comments Workflow

Lithology extraction is divided into two workflows. For the **Lithology Code workflow**, the extraction starts with filtering the *dhgeology* and *dhgeologyattr* table by the location extents and the Drill Hole Lithology Thesaurus. The *dhgeology* table contained 47,062 drill holes across 115 companies with 797,975 lithology depth intervals with corresponding 820,612 entries of lithology information in *dhgeologyattr*. These records are matched with the entries from the Drill Hole Lithology Codes Thesaurus resulting to 273,684 matched records. The FromDepth and ToDepth for these records are then validated. 74 records had equal FromDepth and ToDepth values. 654 had values for FromDepth but null values for ToDepth. For both cases, ToDepth is calculated as FromDepth+0.01. The cut-off value of 80 is used for the string matching based on the performance of the matching on a subset of 1,548 unique lithology codes from the Golden Grove area (Fig. 6). The **Lithology Code workflow** resulted to 273,684 intervals across 12,793 drill holes wherein 235,606 records are successfully matched in the fuzzy string matching. The remaining 546, 819 entries did not obtain a match with a score greater than 80. An example of unmatched entries is provided in Table 2.

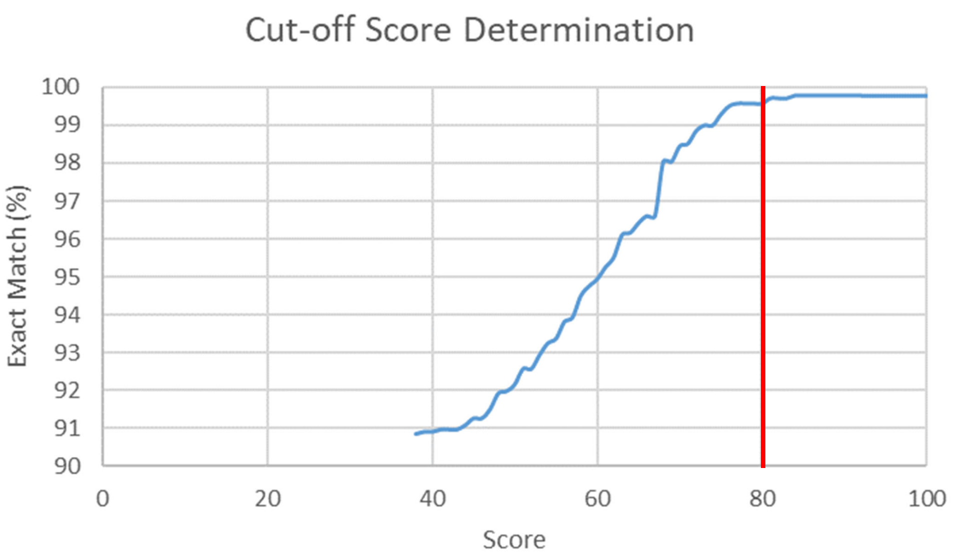


Figure 6. The user-defined cut-off score of 80 is chosen based on the results of the testing different cut-offs on a smaller dataset within the YSGB area. As seen in this figure, the number of exact matches plateau at a score of 80. This relationship may vary depending on the datasets available in the area. Thus, this cut-off value is user-defined and is best to test the matching performance on a subset in the user's area.

Formatted: Not Highlight

Formatted: Not Highlight

Formatted: Heading 2

Formatted: Font: Century Gothic, Bold

Formatted: Font: Lucida Console, Italic

Formatted: Font: Lucida Console

Formatted: Font: Lucida Console

Formatted: Font: Lucida Console

Formatted: Font: Lucida Console

Formatted: Caption

The **Comments workflow** extracts the records from the *dhgeology* and *dhgeologyattr* table as well, but this time using the Drill Hole Comments Thesaurus. For YSGB, the database has 262,567 records across 22,766 drill holes with free text descriptions. 47,823 records are present in both workflow. Since the free text descriptions are extracted here to compare their results from fuzzy string matching, only 7,870 records that also matched (both have a score greater than 80) in the **Lithology Code workflow** are retained.

3.3 Fuzzy String Matching Results

We present results from the data extraction using both workflows: **Lithology Code** and **Comments**. The dataset for the fuzzy string matching assessment consists only of the unique records matched on both **Lithology Code workflow** and **Comments workflow** (3,074 records). It is visually checked from the records that the **Lithology Code workflow:** `Detailed_Lithology` results are sound classifications of the `Company_Litho`. This is done to make sure that these results could be considered as the "true value" in the fuzzy string matching assessment. The overlaps between these two workflows suggest that the user may need to make choices to identify which is better suited for matching in their area of interest. To better understand the difference between these results, we looked at the matching overlaps between the two workflows (3,074 entries). These matching overlaps are used to compare and describe the fuzzy string matching using the decoding the `Company_LithoCode` and using `Comments`.

We also take a look at the unique combinations of `Company_LithoCode`, `Company_Litho`, **Lithology Code workflow:** `Detailed_Lithology` and **Comments workflow:** `Detailed_Lithology` (53 unique records from the 3,074 records). 34 out of the 53 unique entries (64%) result to matches between the **Lithology Code Workflow:** `Detailed_Lithology` and **Comments Workflow**: `Detailed_Lithology`, 26 of which are Exact Matches, 19 unique entries are Close Matches and 26% percent are Failed Matches. The Failed Matches are due to unrelated descriptions in the Comments field which is used to obtain the results in **Comments Workflow**: `Detailed_Lithology`. An example of this is the interval is logged as "ironstone" (`Company_Litho`) but Comments contains "mafic schist". Another less common reason is the `Company_LithoCode` is repeated in the `Comments`. An example of this is would be an interval logged as "colluvium" and the Comments as "COL". The **Comments workflow** will result to "coal" instead.

Exact Matches: Of the total matched entries, 944 are Exact Matches (31%) (Table 2). The Exact Matches are ideal outcomes as both workflows resulted in exactly the same answers.

Close Matches: The Close Matches are common for coarse-grained igneous rocks, clastic sedimentary rocks, surficial residual rocks and filling structures. The coarse-grained igneous rocks such as gabbro, gabbroid and dolerites are used interchangeably in both fields. `Comments` can contain terminologies such as "gabbroic", "granophyric gabbro to dolerite", "intrusive granitoid to gabbro" resulting to close matches. Similar cases are observed between granodiorite and granite and between peridotite and coarse-grained ultramafic rocks. For clastic sedimentary rocks, the Close Matches are a result of gradation of grain size in the `Comments`. For example, an interval logged as mudstone (`Company_Litho`) is then described in `Comments` as "mudstone to sandstone" or "intercalated with siltstone". `Comments` entries like this will result in "sandstone" and "siltstone", respectively. Both clastic sedimentary rocks but not an Exact Match to mudstone. Metasediments and quartz veins occur together and what is described last dictates the `Detailed_Lithology` classification. Surficial rocks such as soil, duricrust, colluvium, laterite, calcrete, ferricrete and cover are used loosely or occur together resulting to multiple combination of these Close Matches.

Formatted: Font: Century Gothic, Bold

Formatted: Heading 2, No bullets or numbering

Formatted: Font: Century Gothic, Bold

Formatted: Font: Century Gothic, Bold

Formatted: Caption, Line spacing: 1.5 lines

Formatted: Font: 10 pt, Not Bold

Formatted: Font: 10 pt, Not Bold

Formatted: Font: Times New Roman, 10 pt, Not Bold

Formatted: Font: 10 pt, Not Bold

Formatted: Font: Century Gothic, 10 pt

Formatted: Font: Century Gothic, 10 pt, Not Bold

Formatted: Font: 10 pt, Not Bold

Related Matches: 60 entries (3%) resulted in related matches. For igneous rocks, this result is observed when Comments use rock type descriptors such as "komatitic", "basaltic" and "doleritic". An example would be an interval logged as dolerite and is then described in Comments as "dolertic basalt". This would result in dolerite in the **Lithology Code workflow** and "basalt" in the **Comments workflow**. Both lithologies are igneous, however have different composition and textural implications. For sedimentary rocks, **Lithology Code workflow** results to sedimentary rocks classified based on grain size as they have been logged ("gravel", "mud"). The comments contains compositional descriptions such as "with silcrete" or "minor chert". In this case, the **Comments workflow** will result in "silcrete" and "chert". Both workflows will result in sedimentary rocks, but the **Lithology Code workflow** will result in "clastic" rocks while the **Comments workflow** will classify these to "siliceous" at the Lithology_Subgroup level. The related matches for structures occur across coincident lithologies such as "mylonite", "vein", "fault" and "breccia" which could either be "fillings" or "fault_rock" at the Lithology_Sugbroup.

Formatted: Font: Century Gothic, Bold

Formatted: Font: Century Gothic, Bold

Formatted: Font: Century Gothic, Bold

Broad and Narrow Matches: No broad matches are noted and only one narrow match is obtained (Table 3). The interval is logged as "ironstone" with "BIF" in comments, "ironstone" being a more general description for "banded iron formation".

Broader and Narrower Matches: More common cases are Broader and Narrower Matches indicate that there is a bigger relationship gap between the data in Company_Litho and Comments. Broad matches are a result of low detail free text descriptions in Comments. For example, an interval logged as "gabbro" is described as "medium-grained mafic", "massive mafic", "rich mafic". The inverse is noted for narrower matches, the interval is logged as "sediment" but in Comments the interval is described as "siliceous sediments".

Failed Matches: 1,694 entries resulted in Failed Matches (55%). Failed Matches occur when Company_Litho and Comments contain different information. This could be because the Company_Litho contains the main lithology while Comments contains all other lithologies intercalated in the interval. Another reason is the Company_Litho is relogged based on adjacent intervals without amending Comments. "Mudstone" had failed matches with a wide range of lithologies, such as: "amphibolite", "dolerite", "saprolite", "duricrust", "laterite", "banded iron formation", "chert", "phyllite", "schist", "vein". The same is observed for igneous rocks such as: "coarse-grained-ultramafic-rock". For "chert", the failed matches are within a range of sedimentary rocks: "alluvium" and "mud", "amphibolite" and "massive sulphide", "carbonate", "vein", "pegmatite".

Table 3. Distribution of matches across the Fuzzy String Matching Dataset. A total of 45% of the unique records are matched reasonably, 31% of which are Exact Matches, 6% Close Matches, 3% Related Matches, 3% Broader Matches and 3% Narrower Matches.

Type of Match	Number of Entries	Percent
Exact Match	944	31%
Close Match	197	6%
Related Match	60	3%
Broad Match	0	0%
Narrow Match	1	0%
Broader Match	84	3%
Narrower Match	95	3%
Failed Match	1694	55%

TOTAL	3,074	100%

The matching results are visualized as confusion matrices, comparing the performance of the string matching using the **Comments workflow** against the results from the **Lithology Code workflow**. From the 3,074 unique records, we use a total of 1,200 samples for the confusion matrices. The reason for this difference is the limitation of building a confusion matrix wherein both workflows look at the same classes~~Relying on one metric to assess the matching can be misleading, therefore, we would like to use a couple of metrics: accuracy, precision, recall and F1 score. Accuracy sums the true positives and true negatives and puts this number in the contrast of all matches:~~

$$\sout{Accuracy = \frac{True\ Positive + True\ Negative}{Positive + Negative}}$$

~~Precision is a useful metric in cases where false positives are a higher concern than false negatives. The precision of the matching is the true positives divided by the sum of the true positives and false positives. The precision measures the fraction of correctly classified are positive:~~

$$\sout{Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}}$$

~~Recall is a useful metric in cases where false negatives trumps false positives. It is the percentage of total relevant results correctly classified while precision is the percentage relevant results. It is computed as:~~

$$\sout{Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}}$$

~~F1-score is a combined metric of precision and recall. It takes their harmonic mean, thus it is maximum when precision is equal to recall. However, the interpretability of the F1-score is poor. Its formula is written as:~~

$$\sout{Matching\ F1\ Score = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}}$$

~~5~~, and ensuring that both workflows produce a match.

3.3.1 Structure and Texture

While geological structures are not lithologies, they are sometimes described in lithological logs (Fig ~~11~~7). Structures common in the YSGB area are faults and veins. Figure ~~11~~ 7 shows the confusion matrix for the structures and textures. The vertical axis represents the matches from the **Lithology Code workflow** while the horizontal axis for the results from the **Comments workflow**. We consider a dataset of 52 unique records where we are trying to assess if the **Comments workflow** results to the same classification as the **Lithology Code workflow**. Figure ~~11~~ 7 shows that there are 6 records classified as "fault" and 46 records as "vein". When looking at the classification of "faults" we can say that there are 2 records that are true positives. 46 records are true negative pairs, as in this 2x2 matrix, if it is not a "fault", it is a "vein". True negatives together with true positives are the Exact Matches and suggests that the **Comments workflow** identified it correctly. To have a better look at the parts that ~~were~~are not classified correctly we look at the false positives and false negatives. False positives represent the number of records classified as "fault" but based on the **Lithology Code workflow** are not. In this case, there are no false positive values. False negatives represent number of records classified as "vein" but are actually "faults" based on the **Lithology Code workflow**.

A total of 48 Exact Matches ~~were~~are noted, 46 records of which are "veins" and 2 records are "faults". This can be surmised by looking into the diagonal cells. The rest of the "veins" (4 records) are Related Matches as "faults". They are considered Related Matches as faults and veins tend to coexist in nature. In addition, faults often occur as fault zones, with infill clay or

Formatted: Heading 3

Formatted: Font: Century Gothic, Bold

Formatted: Font: Century Gothic, Bold

Formatted: Font: Century Gothic, Bold

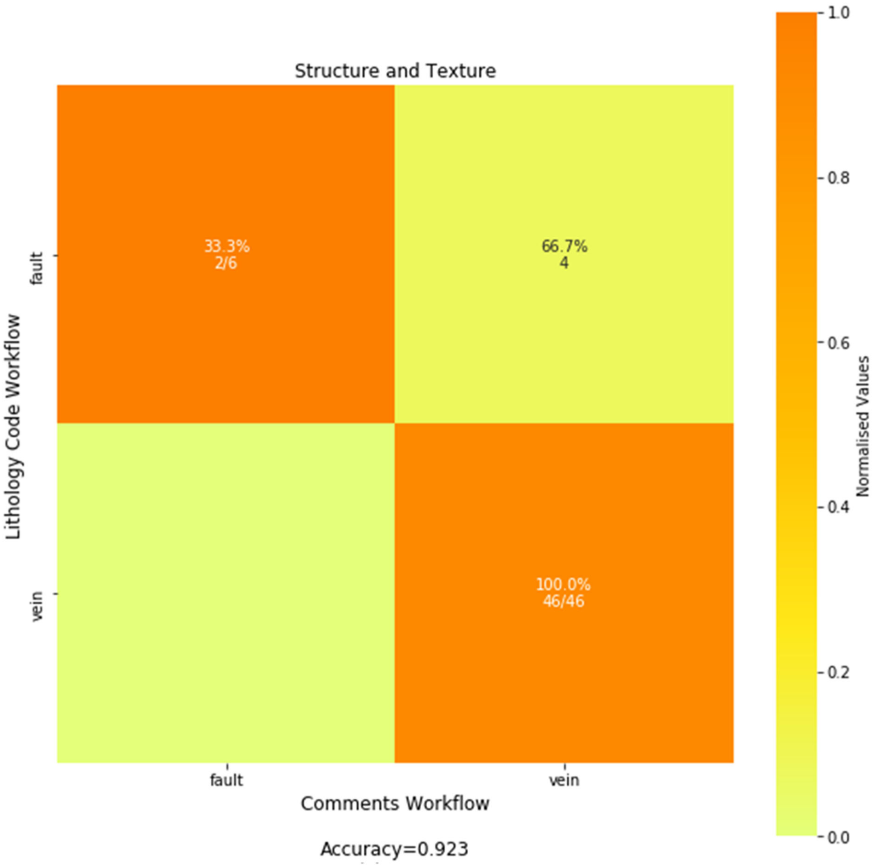
Formatted: Font: Century Gothic, Bold

Formatted: Font: Century Gothic, Bold

Formatted: Font: Century Gothic, Bold

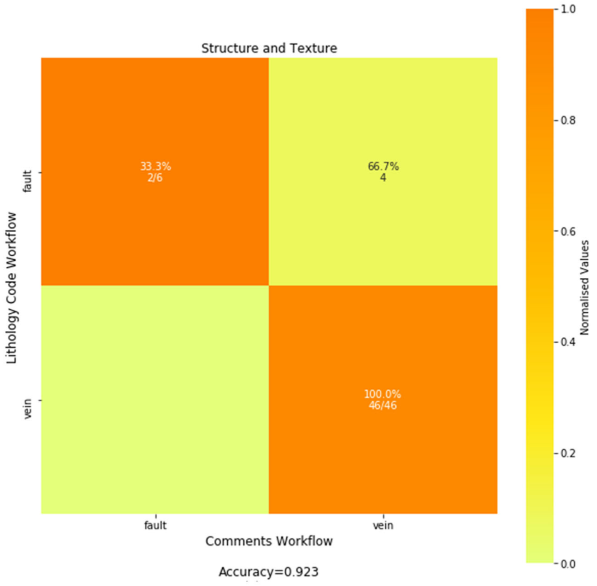
Formatted: Font: Century Gothic, Bold

silica vein sulphides which are described in ~~the comments~~Comments that then obscures the classification. These structure-related lithological descriptions can be used as proxies in further geological studies.



Accuracy=0.923

	precision	recall	f1-score	support
fault	1.00	0.33	0.50	6
vein	0.92	1.00	0.96	46
macro avg	**0.96**	**0.67**	**0.73**	**52**
weighted avg	**0.93**	**0.92**	**0.91**	**52**



	precision	recall	f1-score	support
fault	1.00	0.33	0.50	6
vein	0.92	1.00	0.96	46
macro avg	**0.96**	**0.67**	**0.73**	**52**
weighted avg	**0.93**	**0.92**	**0.91**	**52**

Figure ~~11.~~7. Confusion matrix for structure and texture comparing the fuzzy string matching results from the Lithology_ Code workflow (vertical axis) and Comments workflow (horizontal axis). The heatmap shows the values normalised to the support size to address the imbalance between classes. The values shown in the cells indicate the number of samples classified for the class. Empty cells indicate zero samples. The Structures and Texture Lithology_Group had an accuracy of 92.3% across 52 samples, 46 for veins and 6 for faults.

~~5~~3.3.2 Igneous Rocks

The confusion matrix for igneous rocks considers a dataset of 218 unique records (Fig ~~12~~8). Dealing with a larger matrix is not as straight-forward as the previous matrix. When looking at the classification of a single lithology, the true positives are where both axes refer to the same class. For example, for "basalt" there are 15 records of true positives which correspond to the Exact Matches. The false positives are the sum of all the other entries along the corresponding vertical axis and the false negatives are the sum of all the entries along the corresponding horizontal axis. The sum of all the other cells represent the true negatives. For "basalt", there are 15 true positives, 13, false positives, 15 false negatives and 175 true negatives. This results to 54% classification precision for "basalt".

This statistic is helpful in quantifying the performance of the classification. However, what it does not capture is the semantic and hierarchical relationship of the false negative pairs. As shown in Figure ~~12~~8, 3 records ~~were~~are classified as "komatiite" and 12 records ~~were~~are classified as "mafic". The "komatiite" matches are a result of when ~~the comments~~Comments describe the basalts as "komatiitic basalts". This can be considered as a Related Match. The 12 records which ~~were~~are classified as "mafic: are considered "Broader Match". For the false positive values, the "mafic" records are Narrower Matches while the "dolerite" is a Related Match. These quantitative assessment of the matches show us that although the matching is not perfect, the context of the misclassification is not severe.

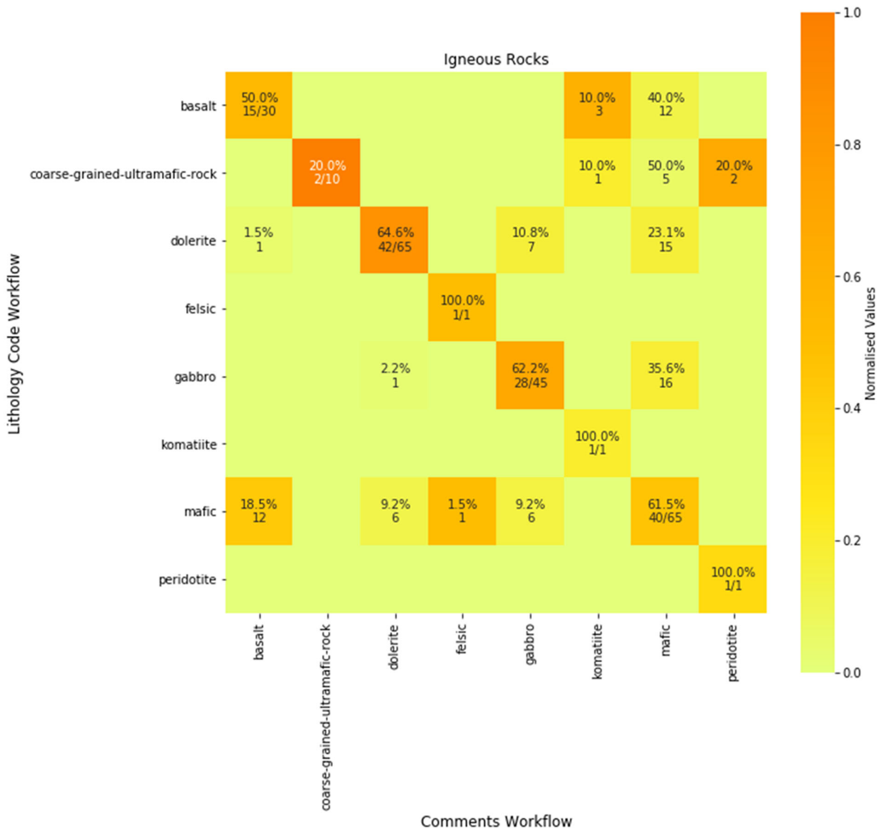
Formatted: Font: Century Gothic

Formatted: Font: Century Gothic

Formatted: Font: Century Gothic

Formatted: Heading 3

“Dolerite” is the most common igneous rock matched. This could be attributed to the sampling bias towards dolerite as it is often targeted by drilling as they are used as targeting criteria for gold mineralisation (Groves et al., 2000). Given that dolerites can be described by their mafic component or be confused as gabbro when weathered, the descriptions contain strings “mafic” and “gabbro” which explain Close and Broader Matches. Gabbros are also common in the YSGB. Some of the “gabbros” ~~were~~are classified as “mafic” in the `Comments Detailed_Lithology`. This is another example of a Broader Match. However, it is important to note that although it is not an Exact Match, a Broader Match can be useful in geological studies relating to rock composition as gabbros are members of mafic rocks. 40% of the igneous rock that ~~were~~are mismatched at the `Detailed_Lithology` level ~~were~~are Broader Matches (matches correctly at `Lithology_Group`).



Accuracy=0.596

	precision	recall	f1-score	support
basalt	0.54	0.50	0.52	30
coarse-grained-ultramafic-rock	1.00	0.20	0.33	10
dolerite	0.86	0.65	0.74	65
felsic	0.50	1.00	0.67	1
gabbro	0.68	0.62	0.65	45
komatiite	0.20	1.00	0.33	1
mafic	0.45	0.62	0.52	65
peridotite	0.33	1.00	0.50	1
macro avg	0.57	0.70	0.53	218
weighted avg	0.66	0.60	0.60	218

Figure 12.8. Confusion matrix for igneous rocks comparing the fuzzy string matching results from the Lithology_Code workflow (vertical axis) and Comments workflow (horizontal axis). The heatmap shows the values normalised to the support size to address the imbalance between classes. The values shown in the cells indicate the number of samples classified for the class. Empty cells indicate zero samples. ~~The accuracy is 59.6%, with a weighted average precision of 66% and recall of 60%. These results were taken~~

Formatted: Font: Century Gothic

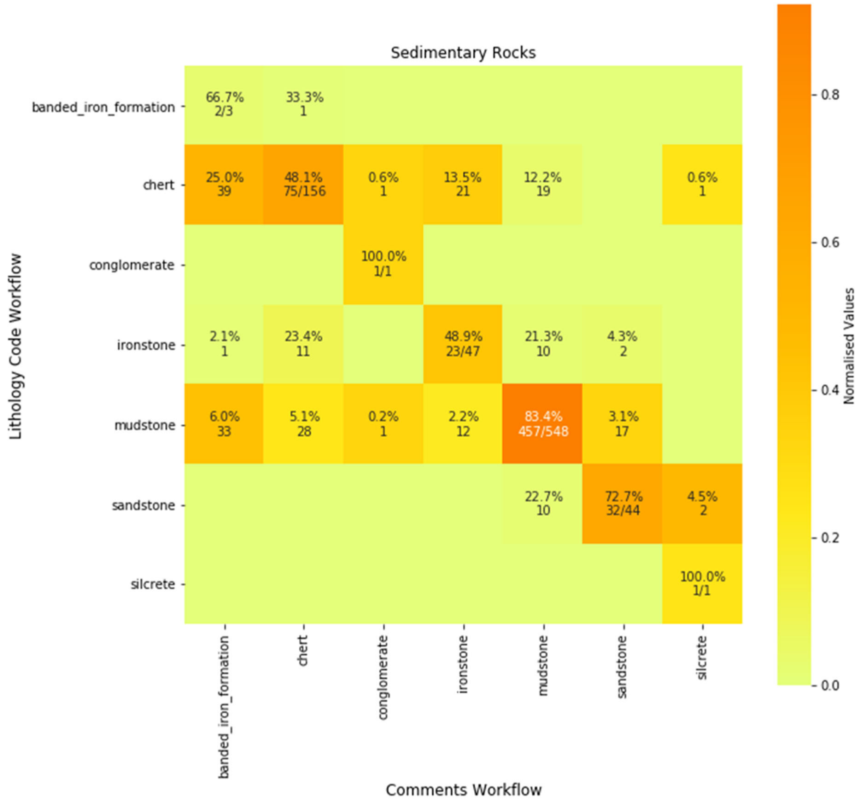
Formatted: Font: Century Gothic

from a subset of 218 samples across 8 classes. "Coarse-grained-ultramafic-rock" has a precision of 1 that implies there are no False Positives.

53.3.3 Sedimentary Rocks

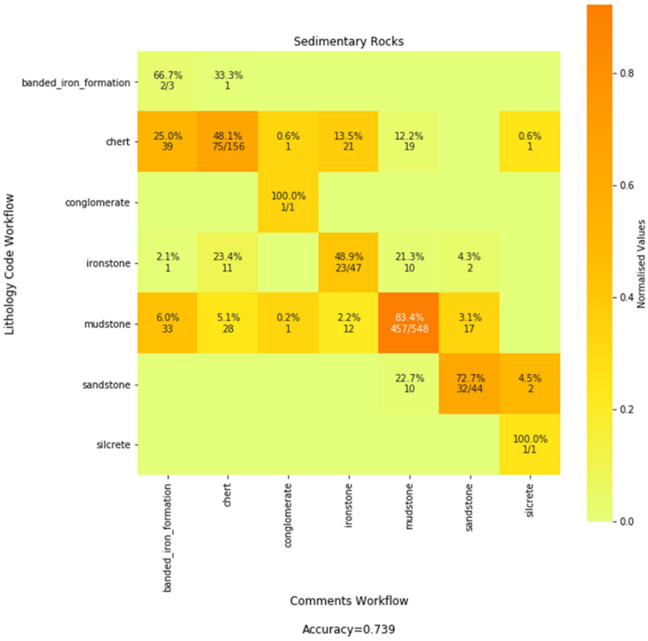
The largest Lithology_Group of the lithological entries relates to sedimentary rocks (800 entries) (Fig 139). 457 of the 800 entries are true positive classification of mudstones. Mudstones are common as shale beds. Mudstones resulted in Related Matches with "chert" and "ironstone". The misclassification occurs when the logs describe intervals wherein the mudstone occurs together and is intercalated with these lithologies. A few mudstones (17) are matched as sandstone due to textural and grain-size descriptors (Close Match). 48% of the cherts are resulted in Exact Matches. 39 records of cherts resulted in Failed Matches as their Detailed_Lithogy level matched with "banded iron formation", it occurs when intercalated together such as "cherts with BIF" or as include string descriptors such as "BIF-fy".

Formatted: Level 3



Accuracy=0.739

	precision	recall	f1-score	support
banded_iron_formation	0.03	0.67	0.05	3
chert	0.65	0.48	0.55	156
conglomerate	0.33	1.00	0.50	1
ironstone	0.41	0.49	0.45	47
mudstone	0.92	0.83	0.88	548
sandstone	0.63	0.73	0.67	44
silcrete	0.25	1.00	0.40	1
macro avg	0.46	0.74	0.50	800
weighted avg	0.82	0.74	0.77	800



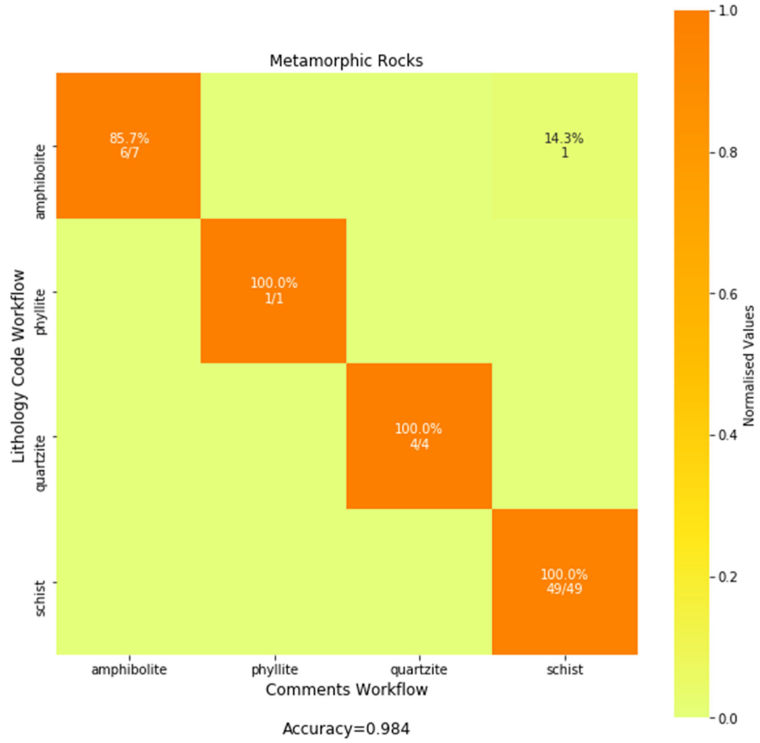
	precision	recall	f1-score	support
banded_iron_formation	0.03	0.67	0.05	3
chert	0.65	0.48	0.55	156
conglomerate	0.33	1.00	0.50	1
ironstone	0.41	0.49	0.45	47
mudstone	0.92	0.83	0.88	548
sandstone	0.63	0.73	0.67	44
silcrete	0.25	1.00	0.40	1
macro avg	0.46	0.74	0.50	800
weighted avg	0.82	0.74	0.77	800

Figure ~~13.~~9. Confusion matrix for sedimentary rocks comparing the fuzzy string matching results from the Lithology~~-~~ Code workflow (vertical axis) and Comments workflow (horizontal axis). The heatmap shows the values normalised to the support size to address the imbalance between classes. The values shown in the cells indicate the number of samples classified for the class. Empty cells indicate zero samples. ~~The accuracy is 73.9%, with a weighted average precision of 82% and recall of 74%. These results were taken from a subset of 800 samples across 7 classes.~~

~~5~~3.3.4 Metamorphic Rocks

Out of a total of 61 metamorphic rock entries, 60 ~~were~~are matched correctly (Fig ~~14~~10). Most of these ~~were~~are "schists" as the YSGB area is rich in talc-carbonate schists. The `Company_~~LithologyLitho~~Litho` entry "amphibolite mica schist" which ~~was~~is matched as "amphibolite" matches as "schist" in the **Comments workflow**. This is considered a Related Match.

Formatted: Font: Century Gothic
Formatted: Font: Century Gothic
Formatted: Font: Century Gothic
Formatted: Level 3
Formatted: Font: Century Gothic, Bold



	precision	recall	f1-score	support
amphibolite	1.00	0.86	0.92	7
phyllite	1.00	1.00	1.00	1
quartzite	1.00	1.00	1.00	4
schist	0.98	1.00	0.99	49
macro avg	0.99	0.96	0.98	61
weighted avg	0.98	0.98	0.98	61

Figure ~~14.~~10. Confusion matrix for metamorphic rocks comparing the fuzzy string matching results from the Lithology_Code workflow (vertical axis) and Comments workflow (horizontal axis). The heatmap shows the values normalised to the support size to address the imbalance between classes. The values shown in the cells indicate the number of samples classified for the class. Empty cells indicate zero samples. ~~The accuracy is 98.4%, with a weighted average precision of 98% and recall of 98%. These results were taken from a subset of 61 samples across 4 classes.~~

~~5~~3.3.5 Surficial Rocks

Fuzzy string matching accuracy of surficial rocks scored a 45% on a total of 69 entries (Fig ~~15~~11). Saprolites ~~were~~are matched as saprolite (Exact Match), rock (Failed Match) and saprock (Close Match). In instances where saprock ~~was~~is inputted as "sap rock", it results to a failed match as "rock". "Soil" is commonly used in logs to refer to the first intercept of highly weathered, clay-rich and unidentifiable intercept. "Soil" ~~was~~is classified with the highest variability of terms: "soil" (Exact Match), "rock" (Failed Match), "duricrust" (Close Match), "colluvium" (Related Match) and "calcrete" (Close Match). "Laterite" ~~was~~is

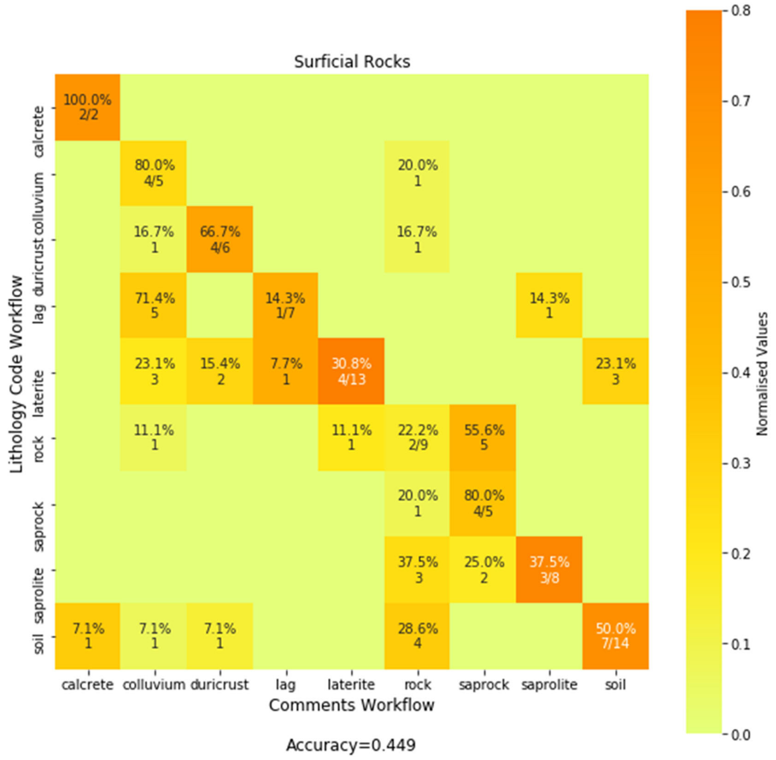
Formatted: Font: Century Gothic

Formatted: Font: Century Gothic

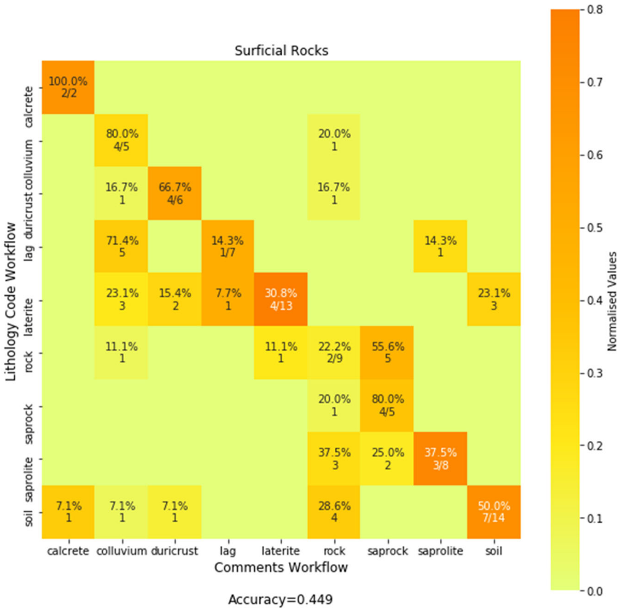
Formatted: Font: Century Gothic

Formatted: Level 3

matched to “colluvium” (Related Match), “duricrust” (Close Match) and “lag” (Close Match). “Lag” generally matches with “colluvium: (Related Match). However, when described in ~~the comments~~Comments, it can be associated with its protolith which results into a Failed Match as “rock”.



	precision	recall	f1-score	support
calcrete	0.67	1.00	0.80	2
colluvium	0.27	0.80	0.40	5
duricrust	0.57	0.67	0.62	6
lag	0.50	0.14	0.22	7
laterite	0.80	0.31	0.44	13
rock	0.17	0.22	0.19	9
saprock	0.36	0.80	0.50	5
saprolite	0.75	0.38	0.50	8
soil	0.70	0.50	0.58	14
macro avg	0.53	0.53	0.47	69
weighted avg	0.57	0.45	0.45	69



	precision	recall	f1-score	support
calcrete	0.67	1.00	0.80	2
colluvium	0.27	0.80	0.40	5
duricrust	0.57	0.67	0.62	6
lag	0.50	0.14	0.22	7
laterite	0.80	0.31	0.44	13
rock	0.17	0.22	0.19	9
saprock	0.36	0.80	0.50	5
saprolite	0.75	0.38	0.50	8
soil	0.70	0.50	0.58	14
macro avg	0.53	0.53	0.47	69
weighted avg	0.57	0.45	0.45	69

Figure ~~15.~~11. Confusion matrix for surficial comparing the fuzzy string matching results from the Lithology– Code workflow (vertical axis) and Comments workflow (horizontal axis). The heatmap shows the values normalised to the support size to address the imbalance between classes. The values shown in the cells indicate the number of samples classified for the class. Empty cells indicate zero samples~~. The accuracy is 44.9%, with a weighted average precision of 57% and recall of 45%. These results were taken from a subset of 69 samples across 9 classes.~~

Formatted: Font: Century Gothic

Formatted: Font: Century Gothic

Formatted: Font: Century Gothic

4 Discussion

4.1 *dh2loop* Functions and Notebooks

6

6.1 Data Extraction

dh2loop supports data extraction of collar, survey and lithology interval tables. The main consideration in the data extraction was that the data retrieved was complete, relevant and useful. We would rather throw erroneous or questionable data out and have the rest with a high level of confidence, than the other way around. 93% of the available collar data in the area was extracted successfully. This can be improved by implementing alternative ways for retrieving `RL` and `MaxDepth` values. For example, if no `RL` values are fetched from the database, it could be fetched from open-source digital terrain models (DTM) and/or SRTM (Shuttle Radar Topography Mission). As for missing `MaxDepth` values, the maximum `ToDepth` values in the survey and/or interval tables could be used.

The survey extraction rate of 86% was fairly good. *dh2loop* ensures that the `Azimuth` and `Dip` values are sensible measurements before including them into the extracted output file. An improvement that could be implemented is to run an assessment on the deflection angles for each drill hole and flag intervals with unrealistic deflection angles.

The lithology extraction using the Lithology Code workflow shows that the bottle neck to its extraction rate is the extensiveness of the Lithology Code thesaurus. Since the thesaurus did not have the information for all companies in the area, only 34% of the available information was retrieved. The extraction results for the Comments workflow cannot be compared with the Lithology Code workflow as only the intersection of both workflows was considered in this study.

6.2 Thesauri

dh2loop provides the user with 9 thesauri that deal with the extraction of collar, survey and lithology interval tables. For extraction of other properties, such as downhole alteration, geochemistry, mineralogy and structures, at least one thesaurus is needed for each attribute we would like to export. These thesauri are built manually by inspecting all the terminologies available in the database. Although, creating them can be tedious, updating an existing thesaurus is as simple as adding and/or removing a word to the list. There are many other properties available in the database that could be exploited using the existing methodology, thus there is an incentive in finding a way to improve the methodology of building these thesauri. Analysis on the syntax of the existing thesauri may help in automating creation of other thesauri.

The Hierarchical Lithology thesaurus puts equal weight on each of the entries in the thesaurus. Knowing the geology in a user's area, the matching can be improved by adding more weight to prevalent lithologies through adding a bonus score.

6.3 Assessment of String Matching Results

Formatted: Heading 2, No bullets or numbering

Formatted: Heading 2

Formatted: Font: Century Gothic

Formatted: English (Australia)

The string matching results highlights that geological drill core logging is prone to human error and bias, and result to incorrect logs. Sometimes even if the data is available and correct, it is not in format that can be directly extracted. For example, the Comment field/s are filled with a string description such as "same as above" and "-do-". Currently, for this case, *dh2loop* returns without a match. In the future, we could be able to search through the previous entries to retrieve the correct lithology. Furthermore, the code does not handle and check for inconsistencies in the logs. It only addresses the inconsistencies in nomenclature and not the logging itself. The string matching misclassification results illustrate that importance in the consistency and level of detail being put into logging and identifies differences in convention or uncoordinated logging among geologists. *dh2loop* provides a notebook that demonstrates using *striplog* to improve the consistency of the logs through data pruning and annealing. In the future, the geochemical compositions can be used to counter check and lithology assigned to the interval.

Comparing the string matching between the Lithology Code and Comments workflow, the Lithology code workflow results to a higher matching rate, 86% of the extracted data is successfully matched. Comparing this subset to the Comments workflow, the matching rate is much lower at 16%. This shows that the Lithology Code workflow, while potentially tedious, results into a higher percentage of successful matches. However, if we are considering a regional study involving multiple companies and drilling campaigns, building thesauri can be time-consuming depending on the size of the region being studied, number of attributes of interest, number of companies and drilling campaigns. This could range from a couple of hours to months. It can also be tedious as it involves inputting errors and inconsistencies as well as exhausting all permutations for decision-tree based logging systems

Comment matching provides a quicker way to standardize and classify rocks. The comprehensive clean-up dictionary allows assists in improving the matching accuracy. Given the context that we are dealing with legacy data, an extraction rate of 16% from the Comments is not bad at all. With minimal effort, we obtain additional geological data wherein, although of a smaller percentage (31% of Exact Matches) but with reasonably high confidence in its quality. It is important to note that most of the time Failed Matches are not a result of the limitations of the algorithm but of the information being fed itself. Inconsistent logs (Company_Litho data is different from Comment) usually occur when:

1. The logs were post-processed and correlated with the rest of the hole or neighbouring drill holes and changes were made to the Company_Litho but none on the Comments field.
2. The Comments would have more level of detail than the Company_Litho. In this case, we may get a lithology at Lithology_Subgroup from the Lithology Code workflow and a Detailed_Lithology from the Comments workflow.
3. The Company_Litho would have more level of detail than the Comments
4. Comments contains the description of the whole intercept, which could include a contact of two lithologies or intercalating lithologies.

Formatted: English (United Kingdom)

Formatted: Font: Century Gothic

Formatted: Font: Century Gothic

Formatted: Font: Century Gothic

Formatted: Font: Century Gothic

Formatted: Heading 2, No bullets or numbering

Formatted: Font: Century Gothic

Formatted: Font: Century Gothic

Formatted: Heading 2

From the results of the confusion matrix (Sect. 5), some rock groups are more sensitive to these inconsistencies than others. There is higher confidence in the classification of structures and textures and metamorphic rocks. The user should be more careful when dealing with igneous, sedimentary and surficial rocks. They are more difficult to classify as the way they are described are highly variable between different geologists. For structure-related lithological descriptions the small number of misclassifications occur where faults, veins and fillings coexist. For metamorphic rocks, entries like "mica amphibolite schist" can cause Broader Matches with the confusion of whether to classify it as "amphibolite" or "schist". "Schist" is a textural term of medium grade metamorphic rock with a medium to coarse-grained foliation defined by micas while "amphibolite" is a compositional term representing a granular metamorphic rock which mainly consists of hornblende and plagioclase. One should be wary about these possibilities as they may impact the interpretation of the geology in the area. For sedimentary rocks, the lack of a standard syntax as to how comments are recorded impacts the classification. Descriptions of intercalated lithologies or presence of major and minor lithology can result to Failed Matches. Igneous rocks perform fairly well, most of what is not captured as Exact Matches are captured at least as Broader Matches. These are usually related to either an inconsistent level of detail between the fields or rock types used as descriptors ("komatiitic", "andesitic", basaltic").

Low matching accuracy in surficial rocks can be attributed to the lack of universally agreed terminology for: deeply weathered regolith; poorly-defined and misapplied surficial rock nomenclature; wide range and variation of materials within the regolith and; difficulty in bulk mineral identification from macroscopic samples. Furthermore, since the degree of weathering of minerals generally increases from the bottom to the top of in-situ weathering profiles, the intermixing of strongly weathered and less weathered grains may cause confusion (Cockbain, 2002). Ubiquitous, highly variable and less interesting lithologies also cause mismatches. An example of this is "soil". Soils are technically are not rocks but is commonly used in logs to refer to the first intercept of the regolith or to describe highly weathered, clay-rich and unidentifiable intercept. Soils vary in character from thin, coarse-grained, poorly differentiated lithosols to thick, well-differentiated silt and clay-rich soils. Soils were classified with the highest variability of terms: "soil", "rock", "duricrust", "colluvium" and "calcrete". There are also certain lithologies with ambiguous nomenclature conventions, like "laterite", "duricrust", "lag". Some geologists use laterite to refer to the whole lateritic profile (ferruginous zone, mottled zone, and saprolite) while others to refer to the ferruginous zone (Eggleton, 2001). Ironcrust, duricrust, lateritic gravels and lag are commonly used interchangeably. Duricrust and ironcrust are terms to describe ferruginous indurated accumulations at or just below the surface. The difference in usage of the term laterite and the interchangeability of duricrust and lag explains the misclassification of "laterite" to "colluvium", "duricrust" and "lag". Another example is "saprolite" and "saprock". They are ambiguous terminologies as they both represent the lower horizons of lateritic weathering profiles, with saprolites having more than 20% of weatherable minerals altered and saprock having less than 20% of the weatherable minerals being altered (Eggleton, 2001). This arbitrary limit makes the terminology used in the logs easily interchangeable, thus affecting the `Detailed_Lithology` matching.

Ideally, a combination of the Lithology Code and Comments workflow should result in a more robust classification. This will also allow the user to have a better look at the result of both workflows and decide what is appropriate for one's purpose.

6.4 Value of the Lithological Information Extracted for Multiscale Analyses

Formatted: Font: Century Gothic

Formatted: English (Australia)

The *dh2loop* lithology export provides a standardized lithological log across different drilling campaigns. This information can readily imported into 3D visualization and modelling software. This allows for drill hole data to be incorporated into 3D modelling, providing better subsurface constraints, especially at a regional scale. It also allows the user to decide on the lithological resolution necessary for their purpose. It provides a three-level hierarchical scheme: `Detailed_Lithology`, `Lithology_Subgroup` and `Lithology_Group` that can be used as an input to multiscale geological modelling. *dh2loop* can be improved by correlating the these lithologies to their corresponding stratigraphic formations. Having the spatial extents of the different geological formations and their lithological assemblages (GSWA Explanatory Notes System) as well as a couple of stratigraphic drill holes, it may be possible to infer the corresponding stratigraphic formation.

Formatted: Heading 2

6.7 *dh2loop* Functions and Notebooks

The *dh2loop* library supports a workflow that extracts, processes and classifies lithological logs (Appendix A3A4). This library wasis built to extract drill hole logs from the WAMEX database. The assumptions made in the entire workflow attempts to replicate the thought process of a geologist performing the data extraction, data quality checks and lithological log classification manually. However, it can be adapted for other geological relational databases or from other table formats. An example using comma separated values tables (CSVs) is shown in the notebook: Exporting and Text Parsing of drill hole Data Demo.

In addition to the data extraction, downhole desurveying and lithological matching functions discussed, *dh2loop* also provides functionalities and a notebook demonstrating harmonization of drill hole data. This is useful for combining and correlating drill hole exports of different properties such as lithology, assays and alteration. It is also possible to export this information in Visualization Toolkit format (.VTK). It also provides a notebook that demonstrates the application of *lasio* and *striplog* on *dh2loop* interval table exports. WAMEX reports can also be interactively downloaded through a notebook provided in the package.

74.2 Thesauri

Formatted: Heading 2

dh2loop provides the user with 9 thesauri that deal with the extraction of collar, survey and lithology interval tables. For extraction of other properties, such as downhole alteration, geochemistry, mineralogy and structures, at least one thesaurus is needed for each attribute we would like to export. These thesauri are built manually by inspecting all the terminologies available in the database. Although, creating them can be tedious, updating an existing thesaurus is as simple as adding and/or removing a word to the list. There are many other properties available in the database that could be exploited using the existing methodology, thus there is an incentive in finding a way to improve the methodology of building these thesauri. Analysis on the syntax of the existing thesauri may help in automating creation of other thesauri.

The Hierarchical Lithology thesaurus puts equal weight on each of the entries in the thesaurus. Knowing the geology in a user's area, the matching can be improved by adding more weight to prevalent lithologies through adding a bonus score.

4.3 Data Extraction

Formatted: Heading 2

dh2loop supports data extraction of collar, survey and lithology interval tables. The main consideration in the data extraction is that the data retrieved is complete, relevant and useful. We would rather throw erroneous or questionable data out and have the rest with a high level of confidence, than the other way around. 93% of the available collar data in the area. This can be improved by implementing alternative ways for retrieving `RL` and `MaxDepth` values. For example, if no `RL` values are fetched from the database, it could be fetched from open-source digital terrain models (DTM) and/or SRTM (Shuttle Radar

Topography Mission). As for missing `MaxDepth` values, the maximum `ToDepth` values in the survey and/or interval tables could be used.

The survey extraction rate of 86% values are sensible measurements before including them into the extracted output file. An improvement that could be implemented is to run an assessment on the deflection angles for each drill hole and flag intervals with unrealistic deflection angles.

The lithology extraction using the **Lithology Code workflow** shows that the bottle neck to its extraction rate is the extensiveness of the Drill Hole Lithology Codes Thesaurus. Since the thesaurus did not have the information for all companies in the area, only 34% of the available information is retrieved. The extraction results for the **Comments workflow** cannot be compared with the **Lithology Code workflow** as only the intersection of both workflows is considered in this study.

4.4 Assessment of String Matching Results

The number of successful matches are dependent on the selected cut-off score. The selection of a cut-off score is a balance between the number of matched records and the exact match percentage. In this case study, we selected a cut-off score of 80 since this is where the number of exact matches plateaus (Fig. 6). A lower cut-off score could be used, depending on the familiarity to the data and/or purpose of drillhole processing. For our case, we wanted to be as conservative as possible without being too stringent (cut-off score 100).

The string matching results highlights that geological drill core logging is prone to human error and bias, and result to incorrect logs. Sometimes even if the data is available and correct, it is not in format that can be directly extracted. For example, `Comments` are filled with a string description such as "same as above" and "-do-". Currently, for this case, *dh2loop* returns without a match, as replacing "same as above" requires building a dictionary for all possible permutations to refer to this. This is not included in the scope of this work. In the future, we could be able to search through the previous entries to retrieve the correct lithology. Furthermore, the code does not handle and check for inconsistencies in the logs. It only addresses the inconsistencies in nomenclature and not the logging itself. The string matching misclassification results illustrate that importance in the consistency and level of detail being put into logging and identifies differences in convention or uncoordinated logging among geologists. *dh2loop* provides a notebook that demonstrates using *striplog* to improve the consistency of the logs through data pruning and annealing. In the future, the geochemical compositions can be used to counter check and lithology assigned to the interval.

Comparing the string matching between the **Lithology Code workflow** and **Comments workflow**, the **Lithology code workflow** results to a higher matching rate, 86% of the extracted data is successfully matched. Comparing this subset to the **Comments workflow**, the matching rate is much lower at 16%. This shows that the **Lithology Code workflow**, while potentially tedious, results into a higher percentage of successful matches. However, if we are considering a regional study involving multiple companies and drilling campaigns, building thesauri can be time-consuming depending on the size of the region being studied, number of attributes of interest, number of companies and drilling campaigns. This could range from a couple of hours to months. It can also be tedious as it involves inputting errors and inconsistencies as well as exhausting all permutations for decision-tree based logging systems. The thesauri provided by *dh2loop* could serve as a starting point to automate this process using recent advances in NLP and machine learning.

String matching using `Comments` provides a quicker way to standardize and classify rocks. The comprehensive Clean-up Dictionary allows assists in improving the matching accuracy. Given the context that we are dealing with legacy data, an

Formatted: Font: Century Gothic, Bold

Formatted: English (Australia)

Formatted: Normal

Formatted: English (United Kingdom)

Formatted: Font: Century Gothic, Bold

Formatted: Font: Century Gothic, Bold

Formatted: Font: Century Gothic, Bold

Formatted: Font: Century Gothic, Bold

extraction rate of 16% Although it is a low extraction rate, there is value in being able to obtain 7,870 records more than what is previously deemed "unusable". With minimal effort, we obtain additional geological data wherein, although of a smaller percentage (31% of Exact Matches) but with reasonably high confidence in its quality. It is important to note that most of the time Failed Matches are not a result of the limitations of the algorithm but of the legacy geological logs itself. Inconsistent logs (`Company_Litho` data is different from `Comments`) usually occur when:

1. The logs are post-processed and correlated with the rest of the hole or neighbouring drill holes and changes are made to the `Company_Litho` but none on the `Comments` field.
2. The `Comments` would have more level of detail than the `Company_Litho`. In this case, we may get a lithology at `Lithology_Subgroup` from the **Lithology Code workflow** and a `Detailed_Lithology` from the **Comments workflow**.
3. The `Company_Litho` would have more level of detail than the `Comments`
4. `Comments` contains the description of the whole intercept, which could include a contact of two lithologies or intercalating lithologies.

From the results of the confusion matrix (Sect. 3.3), some rock groups are more sensitive to these inconsistencies than others. There is higher confidence in the classification of structures and textures and metamorphic rocks in the study area dataset, not necessarily in others. There could be metamorphic-dominated terranes where the subordinate igneous rocks will be classified with higher confidence. The user should be more careful when dealing with sedimentary and surficial rocks. They are more difficult to classify as the way they are described are highly variable between different geologists. For structure-related lithological descriptions the small number of misclassifications occur where faults, veins and fillings coexist. For metamorphic rocks, entries like "mica amphibolite schist" can cause Broader Matches with the confusion of whether to classify it as "amphibolite" or "schist". "Schist" is a textural term of medium grade metamorphic rock with a medium to coarse-grained foliation defined by micas while "amphibolite" is a compositional term representing a granular metamorphic rock which mainly consists of hornblende and plagioclase. One should be wary about these possibilities as they may impact the interpretation of the geology in the area. For sedimentary rocks, descriptions of intercalated lithologies or presence of major and minor lithology can result to Failed Matches. The lack of a standard syntax as to how free text descriptions are recorded impacts the classification. This procedure provides a basis for creating a pre-standard. Not so much providing a guide of practice but highlighting what should not be done and what practices create ambiguity. Standardization will definitely reduce subjectivity and is for the geological surveys to decide and implement. It is also important to note that a "standard" would be tricky to achieve as the information and level of detail contained in logs is highly dependent on the purpose of the study. Igneous rocks perform fairly well, most of what is not captured as Exact Matches are captured at least as Broader Matches. These are usually related to either an inconsistent level of detail between the fields or rock types used as descriptors ("komatiitic", "andesitic", basaltic").

Low matching accuracy in surficial rocks can be attributed to the lack of universally agreed terminology for: deeply weathered regolith; poorly-defined and misapplied surficial rock nomenclature; wide range and variation of materials within the regolith and; difficulty in bulk mineral identification from macroscopic samples. Furthermore, since the degree of weathering of minerals generally increases from the bottom to the top of in-situ weathering profiles, the intermixing of strongly weathered and less weathered grains may cause confusion (Cockbain, 2002). Ubiquitous, highly variable and less interesting lithologies also cause mismatches. An example of this is "soil". Soils are technically are not rocks but is commonly used in logs to refer to the first intercept of the regolith or to describe highly weathered, clay-rich and unidentifiable intercept. Soils vary in character from thin, coarse-grained, poorly differentiated lithosols to thick, well-differentiated silt and clay-rich soils. Soils are classified with the highest variability of terms: "soil", "rock", "duricrust", "colluvium" and "calcrete". There are also certain

Formatted: Font: Century Gothic, Bold

Formatted: Font: Century Gothic, Bold

lithologies with ambiguous nomenclature conventions, like “laterite”, “duricrust”, “lag”. Some geologists use laterite to refer to the whole lateritic profile (ferruginous zone, mottled zone, and saprolite) while others to refer to the ferruginous zone (Eggleton, 2001). Ironcrust, duricrust, lateritic gravels and lag are commonly used interchangeably. Duricrust and ironcrust are terms to describe ferruginous indurated accumulations at or just below the surface. The difference in usage of the term laterite and the interchangeability of duricrust and lag explains the misclassification of “laterite” to “colluvium”, “duricrust” and “lag”. Another example is “saprolite” and “saprock”. They are ambiguous terminologies as they both represent the lower horizons of lateritic weathering profiles, with saprolites having more than 20% of weatherable minerals altered and saprock having less than 20% of the weatherable minerals being altered (Eggleton, 2001). This arbitrary limit makes the terminology used in the logs easily interchangeable, thus affecting the `Detailed_Lithology` matching.

Ideally, a combination of the **Lithology Code workflow** and **Comments workflow** should result in a more robust classification. This will also allow the user to have a better look at the result of both workflows and decide what is appropriate for one’s purpose.

Formatted: Font: Century Gothic, Bold

Formatted: English (Australia)

4.5 Value of the Lithological Information Extracted for Multiscale Analyses

The *dh2loop* lithology export provides a standardized lithological log across different drilling campaigns. This information can readily imported into 3D visualization and modelling software. This allows for drill hole data to be incorporated into 3D modelling, providing better subsurface constraints, especially at a regional scale. It also allows the user to decide on the lithological resolution necessary for their purpose. It provides a three-level hierarchical scheme: `Detailed_Lithology`, `Lithology_Subgroup` and `Lithology_Group` that can be used as an input to multiscale geological modelling. *dh2loop* can be improved by correlating the these lithologies to their corresponding stratigraphic formations. Having the spatial extents of the different geological formations and their lithological assemblages (GSWA Explanatory Notes System) as well stratigraphic drill holes, it may be possible to infer the corresponding stratigraphic formation.

5 Conclusions

The *dh2loop* library is an open-source library that extracts geological information from a legacy drill hole database. This workflow has the following advantages:

1. Maximizes the ~~decades~~amount of legacy geoscientific data available for analysis and modelling.
2. ~~Gains~~Provides better subsurface characterization~~, where data is available~~ and critical inputs to 3D geological modelling
3. Standardizes geological logs across different drilling campaigns, a necessary but typically time-consuming and error-prone activity
4. Provides ~~s~~a set of complementary thesauri that ~~can~~are easily ~~be~~ updated and are individually useful references.

   Formatted: English (Australia)

~~5. Provides additional subsurface constraints which are critical for 3D geological modelling~~

~~6.~~5. Implements a hierarchical classification scheme that can be used as an input to multiscale geological modelling

~~7.~~6. Classification results can also be used as a tool to improve future geological logging works by revealing common errors and sources of inconsistencies

Code and Data Availability

dh2loop is a free, open-source python library licensed under the MIT License. It is hosted on the GitHub repository https://github.com/Loop3D/dh2loop and can be cited as http://doi.org/10.5281/zenodo.4043568.

Author Contribution

M. Jessell contributed the original idea, which lwased further developed by R. Joshi. K. Madaiah developed the code. M. Jessell, M. Lindsay, G. Pirot provided guidance and direction in the research. R. Joshi prepared the manuscript with contributions from all co-authors. Lastly, M. Jessell supervised the entire process.

Acknowledgements

The research was carried out while the first author was supported in receipt of Scholarship for International Research Fees (Australian Government Research Training Program Scholarship) and Automated 3D Geology Modelling PhD Scholarship (University Postgraduate Award) at the University of Western Australia. The work has been supported by the Mineral Exploration Cooperative Research Centre whose activities are funded by the Australian Government's Cooperative Research Centre Program. This is MinEx CRC Document 2020/***. This work was also done with the Loop Consortium (http://loop3d.org) as part of an international effort to found a new open-source platform to build the next generation of 3D geological modelling tools. Mark Lindsay is funded by ARC Discovery DE190100431. We would also like to acknowledge Tim Ivanic for his inputs on the geology of the Yalgoo-Singleton greenstone belt.

Appendix A: *dh2loop* package information

A1 Conventions and Terminologies

Convention	Usage in the paper	Description/Repository
Python libraries are written in italics	*dh2loop*	*dh2loop* stands for drill hole data extracted into a 3D modelling input format, compatible with/for the Loop platform (Ailleres et al., 2019). It is a drill hole processing tool that integrates published dictionaries, glossaries and/or thesauri to and improve standardize highly subjective use of terminology and idiosyncratic logging methods and classify lithological logs.
	fuzzywuzzy	Python package for fuzzy logic for string matching (Cohen, 2011)
	pandas	Python package for data analysis and manipulation (McKinney, 2011)
	psycopg2	Python package for PostgreSQL database adapter for python
	numpy	*numpy*
	nltk	Python package for Natural Language Toolkit
	pyproj	Python package for cartographic projections and coordinate transformations library
Python functions are written in italics followed by an open and close parenthesis	*ratio ()*	*fuzzywuzzy* functions
	partial_ratio ()	
	token_set_ratio ()	
	token_sort_ratio ()	
	partial_token_set_ratio ()	
Database tables are written in Lucida Console Italics	*`collar`*	It contains main collar information
	`collarattr`	It contains collar additional information
	`dhsurvey`	It contains main survey information
	`dhsurveyattr`	It contains survey additional information
	`dhgeology`	It contains geology information
	`dhgeologyattr`	It contains additional geology information
Database table fields are written in Lucida Console	`CollarID`	It is the primary key from the *`collar`* table. It is the Unique ID field that identifies drill hole It is used to associate data in different tables with a single drill hole.
	`HoleID`	This is the drill hole name as the company would internally identify the drill hole.
	`Longitude`	The geographical longitude coordinate locating the collar of the drill hole.
	`Latitude`	The geographical latitude coordinate locating the collar of the drill hole.
	`CompanyID`	Unique ID field that identifies the company used
	`DHSurveyID`	Unique ID field that identified unique drill hole and depth location
	`Depth`	It refers to the downhole depth where the survey measurement is taken (meters)
	`DHGeologyID`	Unique ID field that identified unique drill hole and depth interval
	`FromDepth`	The start/from and end/to downhole depth values (meters)
	`ToDepth`	The end/to downhole depth values (meters)
Output fields are written in Lucida Console	`RL`	Relative Level refers to the Z coordinate of the collar location (meters).
	`MaxDepth`	This refers to the maximum downhole length (meters) drilled for a drill hole, commonly referred as the end-of-hole.
	`X`	It is the calculated Northing (meters)
	`Y`	It is the calculated Easting (meters)
	`Z`	It is the calculated Z position (meters)
	`Azimuth`	It is the trend direction indicated by an angle between 0-360 degrees from the north going clockwise.

Formatted: Not Highlight

Formatted: Not Highlight

Formatted: Not Highlight

	`Inclination`	It is the plunge angle of the drill hole relative to horizontal indicated by an angle between -90 to 90. It is measured from the horizontal plane, thus a positive value indicates an upward-directed drill hole and a negative value indicates a drill hole directed downwards.
	`Company_LithoCode`	This fetches the lithology codes that are typically three-letter codes using the Drill Hole Lithology Thesaurus.
	`Company_Litho`	This value is fetched by matching the `CompanyID` and `Company_LithoCode` to the Drill Hole Lithology Codes Thesaurus.
	`Comments`	It is the free text descriptions from *`dhgeologyattr`*
	`Detailed_Lithology`	This value is the lowest level lithology matched through fuzzy string matching.
	`Lithology_Subgroup`	This value is the subgroup level lithology matched through fuzzy string matching.
	`Lithology_Group`	This value is the highest/group level lithology matched through fuzzy string matching.
Workflows are written in Century Gothic Bold	**Lithology Code workflow**	Workflow to decode **`Company_LithoCode`**
	Comments workflow	Workflow to decode **`Comments`**
Thesurus (https://github.com/Loop3D/dh2loop/blob/master/thesauri/)	Drill Hole Collar Elevation Thesaurus	https://github.com/Loop3D/dh2loop/blob/master/thesauri/thesaurus_collar_elevation.csv
	Drill Hole Maximum Depth Thesaurus	https://github.com/Loop3D/dh2loop/blob/master/thesauri/thesaurus_collar_maxdepth.csv
	Drill Hole Survey Azimuth Thesaurus	https://github.com/Loop3D/dh2loop/blob/master/thesauri/thesaurus_survey_azimuth.csv
	Drill Hole Survey Inclination Thesaurus	https://github.com/Loop3D/dh2loop/blob/master/thesauri/thesaurus_survey_inclination.csv
	Drill Hole Lithology Thesaurus	https://github.com/Loop3D/dh2loop/blob/master/thesauri/thesaurus_geology_lithology.csv
	Drill Hole Comments Thesaurus	https://github.com/Loop3D/dh2loop/blob/master/thesauri/thesaurus_geology_comment.csv
	Drill Hole lithology Codes Thesaurus	https://github.com/Loop3D/dh2loop/blob/master/Thesauri/thesaurus_geology_lithology_code.csv
	Clean-up Dictionary	https://github.com/Loop3D/dh2loop/blob/master/thesauri/thesaurus_cleanup.csv
	Lithology Hierarchical Thesaurus	https://github.com/Loop3D/dh2loop/blob/master/thesauri/thesaurus_geology_hierarchical.csv

Formatted: Line spacing: 1.5 lines

Formatted: Line spacing: 1.5 lines

Field Code Changed

A2 Installation and Dependencies

Installing *dh2loop* can be done by cloning the GitHub repository with $ git clone https://github.com/Loop3D/dh2loop.git and then manually installing it by running the python setup script in the repository: $ python setup.py install

~~A2~~

It primarily depends on a number of external open-source libraries:

1. *fuzzywuzzy* (https://github.com/seatgeek/fuzzywuzzy) which uses fuzzy logic for string matching (Cohen, 2011)
2. *pandas* (https://pandas.pydata.org/) for data analysis and manipulation (McKinney, 2011)
3. *psycopg2* (https://pypi.org/project/psycopg2/), a PostgreSQL database adapter for python (Gregorio and Varrazzo, 2018)
4. *numpy* (https://github.com/numpy/numpy)
5. *nltk* (https://github.com/nltk/nltk), the Natural Language Toolkit is a suite of open source Python modules, data sets, and tutorials supporting research and development in Natural Language Processing (Loper and Bird, 2002).
6. *pyproj* (https://github.com/pyproj4/pyproj), python interface to PROJ (cartographic projections and coordinate transformations library)

Code describing basic drill hole operations, such as desurveying (process of translating collar (location) and survey data (azimuth, inclination, length) of drill holes into XYZ coordinates in order to define its 3D geometry of the non-vertical borehole), is heavily inspired from *pyGSLIB* drill hole module (Martínez-Vargas, 2016). The *pyGSLIB* drillhole module is re-written into python to make it more compact with less dependencies and tailor it to the data extraction output.

Formatted: Not Highlight

Formatted: Normal, Indent: Left: 0.63 cm, No bullets or numbering

Formatted: Normal, No bullets or numbering

A3 Documentation

dh2loop's documentation provides a general overview over the library and multiple in-depth tutorials. The tutorials are provided as Jupyter Notebooks, which will provide the convenient combination of documentation and executable script blocks in one document. The notebooks are part of the repository and located in the notebooks folder. See http://jupyter.org/ for more information on installing and running Jupyter Notebooks.

~~A3~~

A4 Jupyter notebooks

Jupyter notebooks are provided as part of the online documentation. These notebooks can be executed in a local python environment (if the required dependencies are correctly installed). In addition, static versions of the notebooks can currently be inspected directly on the *github* repository web page or through the use of *nbviewer*.

1. WAMEX Interactive report downloads (https://github.com/Loop3D/dh2loop/blob/master/notebooks/0_WAMEX_Downloads_Interactive.ipynb)
2. Exporting and text parsing of drill hole data from PostgreSQL database (https://github.com/Loop3D/dh2loop/blob/master/notebooks/1_Exporting_and_Text_Parsing_of_Drillhole_Data_From_PostgreSQL.ipynb)
3. Exporting and Text Parsing of drill hole Data Demo (https://github.com/Loop3D/dh2loop/blob/master/notebooks/2_Exporting_and_Text_Parsing_of_Drillhole_Data_Demo.ipynb)
4. Harmonizing drill hole data (https://github.com/Loop3D/dh2loop/blob/master/notebooks/3_Harmonizing_Drillhole_Data.ipynb)

Appendix B: Thesauri

A few examples of each thesauri are shown below. Each thesauri includes alternate nomenclature and spelling. The complete thesauri are available at: https://github.com/Loop3D/dh2loop/blob/master/thesauri/

B1 Drill hole Collar Elevation

(https://github.com/Loop3D/dh2loop/blob/master/thesauri/thesaurus_collar_elevation.csv)

1. """RL"""
2. ADJ_RL
3. Adjusted_RL
4. AMG_mRL
5. Approx RL
6. Arbitary_RL
7. Best_RL
8. COLL_RL
9. Collar Elevation
10. Collar RL
11. Collar RL (m)
12. Corrected_RL
13. DB_RL
14. DEM_RL
15. DGPS_Elevation_MGA94 Zone 51
16. DGPS RL
17. DTM RL
18. Elevaion
19. Elevat
20. Elevati
21. Elevatio
22. Elevation
23. elvation
24. Lidar_RL
25. Local RL
26. MGA RL
27. MGA_Elev
28. MGA_Elevation
29. MGA_RL_Z
30. MGA_Z50_RL
31. MGA94_RL
32. MINE_RL
33. mRL
34. NAT RL
35. Orig RL
36.
37. Orig_Reg_RL
38. R.L
39. R.L.
40. Raw_RL
41. Real RL
42. ref_mRL
43. RL
44. Surveyed RL
45. UTMElev
46. UTMmRL
47. WGS84_WORLD_LL_Cale_Z
48. z RL
49. Z_(RL)
50. ZCOLLAR_RL
51. ZMINE_RL

B2 Drill hole Maximum Depth Thesaurus

(https://github.com/Loop3D/dh2loop/blob/master/thesauri/thesaurus_collar_maxdepth.csv)

1. """DEPTH"""
2. "Depth
3. """Final
4. AC Depth
5. Actual Depth
6. DD_Depth
7. DDH_Depth
8. Depth
9. Depth - m
10. DEPTH (m)
11. Depth (EOH)
12. Depth (metres)
13. Depth_D
14. Depth_DD
15. Drill Depth
16. Drilled_Depth
17. End of hole depth
18. END_DEPTH
19. EOH Depth
20. EOH Depth (Metres)
21. F/Depth
22. F/Depth(m)
23. F_Depth
24. FIN_DEPTH
25. Final Depth
26. Final Depth (m)
27. Final_dpth
28. Finl_Depth
29. HDEPTH
30. Hole depth
31. MAX DEPTH
32. max_depth (m)
33. Max_Depth m
34. Maximum Depth
35. T_depth
36. TD
37. toatl_depth
38. TOT DEPTH
39. TOT_DEPTH_M
40. Total Depth
41. Total Depth (m)
42. Total Depth Drilled m
43. Total Depth M
44. Total Hole Depth

B3 Drill hole Survey Azimuth Thesaurus (https://github.com/Loop3D/dh2loop/blob/master/thesauri/thesaurus_survey_azimuth.csv)

1. AMG AZIMUTH
2. AMG_azim
3. AMGAZM
4. Aximuth
5. Aximuth_gyro
6. AZ
7. Az_AMG
8. Az_grid
9. AZ_LOCAL
10. AZ_Mag
11. AZ_MINE
12. AZI GRID
13. AZI(T)
14. Azi_Mag
15. AZI_MGA
16. AZIM
17. Azim mag
18. Azim_AMG
19. Azim_Local
20. Azim_M
21. AZIMUTH
22. Azimuth Local
23. Azimuth Mag
24. Azimuth(T)
25. AzLoc
26. GDA_Az
27. GILBEYS AZI
28. GRID AZI
29. GRID_AZ
30. GridAzim
31. LOCAL AZIMUTH
32. Local_Az
33. Local_Azi
34. Local_Azim
35. LOCALAZID
36. LocAzim
37. LOCAZM
38. Mag_Az
39. Mag_Azim
40. Mag_Azimu
41. MagAzi
42. MAGAZM
43. Magnetic AZI
44. MGA AZI
45. MGA Azimuth
46. MGA94_Az
47. NAT_Azimuth
48. Nominal Az
49. Nominal AZI
50. Orig AZI
51. Orig Azimuth
52. Orig_Azim
53. ORIG_AZIMU
54. Project_Azim
55. Ref. AZI
56. ref_azim
57. REG_AZIM
58. UTM_Az
59. UTM_Azi
60. UTM_Azimuth
61. WMC AZI

B4 Drill hole Survey Dip Thesaurus (https://github.com/Loop3D/dh2loop/blob/master/thesauri/thesaurus_survey_dip.csv)

1. Dip
2. Dip (deg.)
3. Dip_2
4. INC
5. Inclination
6. DIP_camera
7. Dip_gyro
8. DIP_LOCAL
9. Nominal_Dip
10. DIP_Surtron

B5 Drill hole Lithology Thesaurus (https://github.com/Loop3D/dh2loop/blob/master/thesauri/thesaurus_geology_lithology.csv)

1. $Lith
2. %Maj
3. %Maj_Lith
4. %Major Lith
5. 1_LithCode
6. 1_RootCode
7. F_lithology
8. GeoCode
9. Geol
10. Geol Code

Geol General

11. GEOL_Rock1
12. geol_type
13. GEOL1
14. Geological Unit

15. Geological_code
16. GEOLOGY
17. Geology Code
18. HOST_LITH1
19. Intermediate Rock Type
20. lihological
21. Lilh1
22. lit
23. Lit_1
24. lith
25. Lith_Code
26. Lith 1
27. Lith 1 Rock Type
28. Lith Maj
29. Lith Major
30. Lith_Cat
31. Lith_Codes
32. Lith_Maj_1
33. Lith_PrimaryCode
34. LITH_Protolith
35. LITH_TYPE
36. Lith01
37. lith-1
38. Lith1_Code1
39. LithCode1_A
40. lithcode1_main
41. LithCode1_T
42. LithCode1_V
43. LithCodeSy
44. Lithgen
45. LithGenrl
46. Lithic_code
47. Lithlogy 1
48. LithMajor1
49. LithMin1
50. Litho
51. LITHO_1
52. Litho Code
53. Litho Type
54. LITHO_PLOT
55. Lithological Unit
56. LITHOLOGY
57. Lithology_rock
58. Lithoology
59. Main Geol Unit
60. Main Lithology
61. MAJ LITH
62. Maj Lithcode
63. Maj. Lithology
64. Maj. Rock
65. Major Lithology
66. Major Rock
67. Major Rock Type
68. Primary Lith
69. Primary RockType
70. ROCK
71. ROCK CODE
72. Rock Group
73. ROCK NAME
74. Rock type code
75. Rock Type Major
76. Rock Unit
77. Rock_id
78. ROCK-CODE
79. RockLithCode
80. RockMain
81. RockMajor
82. Root_Code
83. Root_Lith
84. Root_rock
85. Wiluna Lithology Code
86. WMC ROCK Code
87. wmc_lith1

B6 Drill hole Comments Thesaurus
(https://github.com/Loop3D/dh2loop/blob/master/thesauri/thesaurus_geology_comment.csv)

1. Comment
2. COMMENTS
3. D_stLITHCOMMETNS
4. Description
5. INTRCPT_COMMENT
6. LITH_COMMENT

B7 Drill hole Lithology Codes Thesaurus
(https://github.com/Loop3D/dh2loop/blob/master/Thesauri/thesaurus_geology_lithology_code.csv)

CompanyID -> Company_LithoCode > Company_Litho

1. 551 > BIF > Banded iron formation
2. 1311 > BIF > Banded iron formation
3. 551 > CS > Saprolite, undifferntiates
4. 2551 > Cs > Cambrian Sediment
5. 551 > CSM > Mafic saprolite
6. 2551 > Csm > Cambrian Sediment Limestone
7. 2551 > Cv > Cambrian Vein material
8. 2790 > CV > Colluvium
9. 551 > CY > Clay, undifferentiated
10. 2551 > Cy > Cambrian Mylonite
11. 2551 > Hp > Carboniferous Massive sulphide
12. 2790 > HP > Hardpan

13. 2551 > Lc > Oligocene Chemical Sediments
14. 3053 > Lc > undiferentiated laterite clay
15. 2551 > Ls > Oligocene Sediment
16. 3053 > Ls > undiferentiated laterite sand
17. 1311 > MB > Basalt
18. 2551 > Mb > Miocene Mafic Extrusive
19. 369 > PH > Phyllite
20. 2551 > Ph > Proterozoic Hornfels
21. 2551 > Qc > Quaternary Chemical Sediments
22. 3053 > Qc > undifferentiated recent surficial deposit clay
23. 2551 > Qs > Quaternary Sediment
24. 3053 > Qs > undifferentiated recent surficial deposit sand
25. 551 > QV > Quartz Vein
26. 2551 > Qv > Quaternary Vein material
27. 369 > S > Undifferentiated sediments
28. 3053 > S > undifferentiated sediment
29. 2551 > Sh > Silurian Hornfels
30. 3053 > Sh > shale
31. 551 > SS > Sandstone
32. 2551 > Ss > Silurian Sediment
33. 369 > TCC > Channel clays
34. 1311 > TCC > Tertiary palaeochannel clay
35. 369 > TCS > Channel sands
36. 1311 > TCS > Tertiary palaeochannel clayey sand
37. 369 > TLC > Lake clays
38. 2551 > Tlc > Triassic Lamprophyre/Kimberlites Carbonatite
39. 369 > ALL > Alluvium
40. 2790 > ALLU > Alluvium
41. 2790 > AMPH > Amphibolite
42. 369 > MAA > Amphibolite
43. 369 > MBA > Basalt
44. 2790 > CALC > Calcrete
45. 369 > LCZ > Calcrete
46. 11410 > CL > Clay
47. 2790 > CLAY > Clay
48. 369 > COL > COLLUVIUM
49. 2790 > COLL > Colluvium
50. 369 > FDI > Diorite
51. 2790 > IDO > Diorite
52. 3053 > Bdo > dolerite
53. 2790 > DOLR > Dolerite
54. 1311 > MD > Dolerite
55. 3053 > Bgb > gabbro
56. 2790 > GABR > Gabbro
57. 2790 > MDG > Gabbro
58. 11410 > MG > Gabbro
59. 369 > MGA > gabbro
60. 369 > FGR > Granite
61. 11410 > GR > Granite
62. 2790 > GRAN > Granite
63. 2790 > HARD > Hardpan
64. 369 > LDH > Hardpan
65. 2790 > NMO > Mottled zone
66. 2790 > WMZ > Mottled zone
67. 369 > FPG > Pegmatite
68. 2790 > PEGM > Pegmatite
69. 2790 > QZVN > Quartz Vein
70. 2790 > VQZ > Quartz Vein
71. 2790 > FSS > Sericite schist
72. 551 > SESCH > Sericite Schist
73. 2790 > UMS > Serpentinite
74. 3053 > Us > serpentinite
75. 369 > SSH > Shale
76. 369 > LSZ > Silcrete
77. 2790 > SILC > Silcrete
78. 2790 > SLS > Siltstone
79. 369 > SSTS > Siltstone
80. 369 > M > Undifferentiated Mafic Rocks
81. 2790 > MOO > Undifferentiated mafic rock

B8 Clean up dictionary (https://github.com/Loop3D/dh2loop/blob/master/thesauri/thesaurus_cleanup.csv)

B9 Lithology Hierarchical Thesaurus

(https://github.com/Loop3D/dh2loop/blob/master/thesauri/thesaurus_geology_hierarchical.csv)

1735

Formatted: Section start: New page

Appendix C:

C1 Configuration file

```
#Extents to query
minlong=115.5
maxlong=118
minlat=-30.5
maxlat=-27.5

#src_pro,Dst_proj
src_csr = 4326
dst_csr = 28350

#ExportFiles
export_path='../data/export_db/'
DB_Collar_Rl_Log = export_path + 'DB_Collar_Rl_Log.log'
DB_Collar_Maxdepth_Log = export_path + 'DB_Collar_maxdepth_Log.log'
DB_Collar_Export=export_path+'DB_Collar_Export.csv'
DB_Survey_Export=export_path+'DB_Survey_Export.csv'
DB_Survey_Export_Calc=export_path+'DB_Survey_Export_Calc.csv'
CET_Litho=export_path+'CET_Litho.csv'
DB_Lithology_Export=export_path+'DB_Lithology_Export.csv'
DB_Lithology_Export_Backup=export_path+'DB_Lithology_Export_Backup.csv'
DB_Lithology_Upscaled_Export=export_path+'DB_Lithology_Upscaled_Export.csv'
Upscaled_Litho_NoDuplicates_Export = export_path+'Upscaled_Litho_NoDuplicates_Export.csv'
DB_Lithology_Export_Calc=export_path+'DB_Lithology_Export_Calc.csv'
DB_Lithology_Export_VTK=export_path+'DB_Lithology_Export.vtp'

print('Default parameters loaded from DH2_LConfig.py:')
with open('../notebooks/DH2_LConfig.py', 'r') as myfile:
    data = myfile.read()
    print(data)
    myfile.close()
print('\nModify these parameters in the cell below')
```

C2 Fuzzy String Matching Pseudocode

```
DEFINE FUNCTION Attr_val_With_fuzzy():
Bestmatch is -1
Bestlitho is ‘ ‘
list top
I is 0
list ←attr_val_sub_list

open csv file Attr_val_fuzzy for writing .
write csv file heading CollarID','code','Attr_val','cleaned_text','Fuzzy_wuzzy','Score'

Convert list Var.Attr_val_Dic to  list of list Attr_val_Dic_new
For each element Attr_val_Dic_ele in Attr_val_Dic_new
        cleaned_text_1= call clean_text user defined function with attribute_value
        cleaned_text_1=call tokenize_and_lemma python function with cleaned text
        cleaned_text_1
        cleaned_text=join each word of cleaned_text_1 with space as one string.
        Words is replace slashes by space using re pattern, strip leading and trailing spaces.
        Words  is strip \n\r and split on space.
        for each element Litho_dico_ele in Var.Litho_dico
                litho_words is Litho_dico_ele with lower case, rstrip \n\r,replace ( or) by space and split
                on space
                scores = call python process.extract with arguments cleaned_text, litho_words,
                scorer=fuzz.token_set_ratio
                for sc in each scores
                        if(sc[1]>bestmatch): #better than previous best match
                                bestmatch =  sc[1]
                                bestlitho=litho_words[0]
                                top.append([sc[0],sc[1]])
                                if(sc[0]==words[last]): #bonus for being last word in phrase
                                        bestmatch=bestmatch*1.01
                        elif (sc[1]==bestmatch): #equal to previous best match
                                if(sc[0]==words[last]): #bonus for being last word in phrase
                                        bestlitho=litho_words[0]
                                        bestmatch=bestmatch*1.01
                                else:
                                        top.append([sc[0],sc[1]])
        if bestmatch >80:
                write bestlitho and bestmatch along with required data to csv file
                clear top
                CET_Litho as ' '
                Bestmatch  as -1
                Bestlitho as ''
        else
                write ‘Other’ as bestlitho with bestmatch along with required data to csv file
                clear top
                CET_Litho as ' '
                Bestmatch  as -1
                Bestlitho as ''
```

Appendix B: Collar and Survey Data Extraction

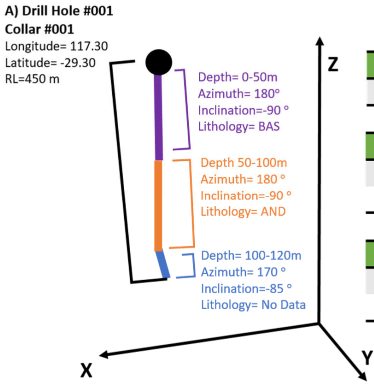
B1 Simplified example of a drill data

Simplified example of a drill hole (1.A) and its corresponding interval tables collar (1.B), survey (1.C) and lithology (1.D). The black circle denotes the collar location of the drill hole which is obtained from a collar table (1.B). The purple line represents the first downhole interval taking its deviation data from the survey table (1.C) and the lithology information from

Formatted: Heading 1, Space After: 0 pt, Line spacing: single

Formatted: Font: Not Italic

Formatted: Justified, Space After: 0 pt, Line spacing: 1.5 lines the lithology table (1.D). The same applies for the second interval (orange line) and third interval (blue line). The orange line follows the same trajectory as the first interval as it uses the same entry in the survey table (1.C). The blue line has no lithology data as this information is not present in the lithology table (1.D). The `MaxDepth` denotes the total drill length (1.B).



B) Collar Data

CollarID	HoleID	Longitude	Latitude	RL	MaxDepth
001	DDH_001	117.30	-29.30	450	120

C) Survey Data

CollarID	Depth	Azimuth	Inclination
001	0	180	-90
001	100	170	-85

D) Lithology Data

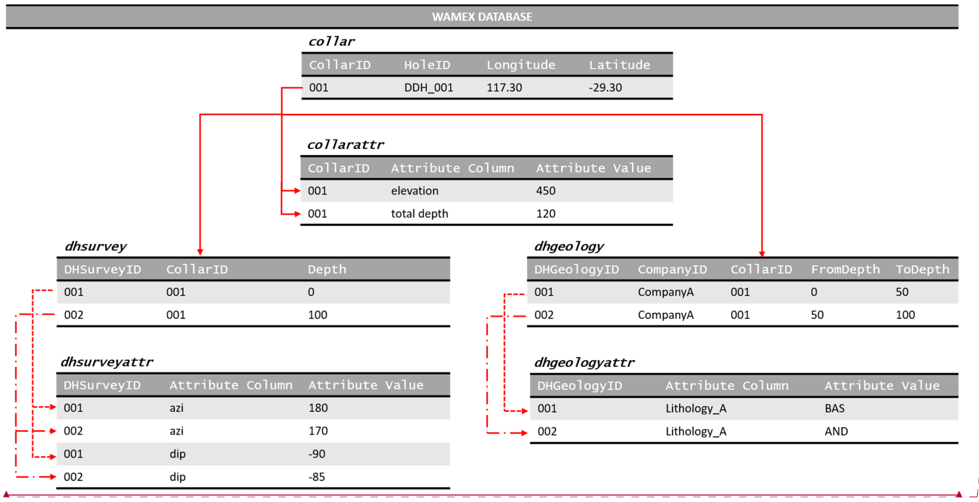
CompanyID	CollarID	FromDepth	ToDepth	Attribute Value	Company_Litho
CompanyA	001	0	50	BAS	Hi-Mg basalt
CompanyA	001	50	100	AND	Andesite

Formatted: Not Highlight

B2 Simplified WAMEX database schema

Formatted: Heading 2

Simplified WAMEX database schema showing the one-to-many relationship between the *collar* table and the *collarattr* table (red solid line). *collarattr* stores other attributes that describe each unique drill hole, such as maximum depth and elevation. The figure also shows the relationship between the *collar* table and the other interval tables such as *dhsurvey*, *dhsurveyattr*, *dhgeology*, *dhgeologyattr*. The deviation of the drill hole is stored in a table, *dhsurvey*, with a primary key (`DHSurveyID`) that refers to each unique depth of a drill hole. This primary key has a many-to-one relationship with collar, as there are multiple depth measurements for each drill hole. Furthermore, *dhsurvey* also has a one-to-many relationship with table *dhsurveyattr*, which stores additional attribute information regarding survey, such as azimuth and inclination readings. The example shows the relationship between tables for the first (red dashed line) and second interval (red dashed-dot line). Each drill hole in the WAMEX database is identified by its geographic coordinates and a unique ID (`CollarID`) in the collar table. The drill hole 3D geometry is described in the survey tables (*dhsurvey*, *dhsurveyattr*). This similar relationship is maintained with interval tables, except that the primary key (e.g.`DHGeologyID`) is used to refer a unique downhole interval rather than a depth measurement. For lithological information, we refer to tables: *dhgeology* and *dhgeologyattr*. *dhgeologyattr* which contain information such as rock names and free text descriptions while *dhgeology* provides information to which hole and interval depth that data refers to. This information can be joined and extracted through SQL (Structured Query Language) queries.



B3 Collar Extraction

Collar extraction workflow fetches the `CollarID`, `HoleID`, `Longitude` and `Latitude` information from the *`collar`* table (a, red), while the corresponding `RL` and `MaxDepth` values are fetched from the *`collarattr`* table using the Drill Hole Collar Elevation Thesaurus (b, blue) and Drill Hole Maximum Depth Thesaurus (c, orange). With the minimum input of a region of interest, the *dh2loop* library exports a Comma-Separated Values file (CSV) listing the drill holes in the area with the following information:

1. `CollarID`: The `CollarID` for a drill hole is identical in all tables in order for data to be associated with that drill hole.
2. `HoleID`: This is the drill hole name, as the company would internally identify the drill hole.
3. `Longitude` and `Latitude`: Both values are expressed in WGS 1984 lat/long (EPSG:4326).
4. Relative level (`RL`): We use RL here to refer to elevations of survey points with reference to the mean sea level. This definition of RL is equivalent to the elevation values used in DEMs. This value is extracted by using the Drill Hole Collar Elevation Thesaurus to filter the values referring to relative level. More than one value can be fetched due to duplicate company submissions or multiple elevation measurements, in which case the code retains the value with most decimal places assuming higher precision corresponds to better accuracy. If no elevation values are fetched from the database the entire record is skipped. Non-numeric values are also ignored.
5. Maximum depth (`MaxDepth`): This value is extracted by using the Drill Hole Maximum Depth Thesaurus. Due to duplicate company submissions, there can be more than one value fetched. Since there is no submission date information, the code takes the value with largest value assuming it is the latest submission.

Formatted: Not Highlight

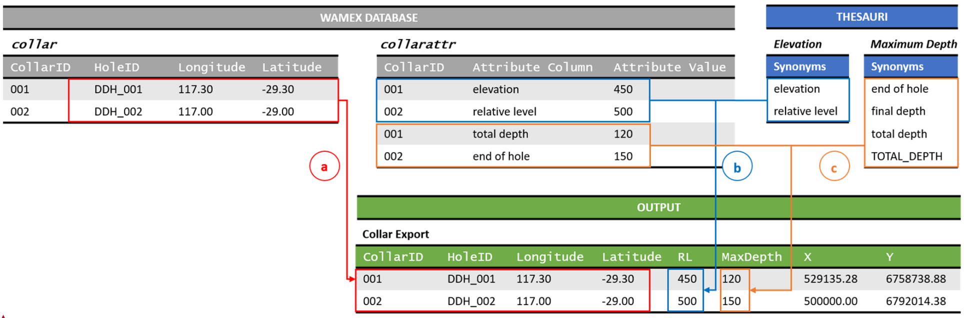
Formatted: Not Highlight

Formatted: Not Highlight

Formatted: Not Highlight

Formatted: Not Highlight

6. Calculated X, Y values of projected coordinates: These values are commonly calculated and used to be able to plot the drill hole in a metric system to be able to accurate display and measure distance within and between drill holes. The projection system used in the calculation is based on the input specified in the configuration file.



Extraction of the collar data for YSGB resulted in a collar file with 68,729 drill holes. This information is extracted from the *collar* table with 73,881 drill holes with 769,981 rows of information from *collarattr*. It includes the location of the collar both in geographic and projected coordinated systems, relative level (RL) and maximum depth (MaxDepth). A total of 136,100 records for RL are retrieved from the database, 1,526 of which are disregarded: 846 records for having an RL value greater than 10,000 meters and 680 non-numeric records. These discarded values are retrieved from the attribute column "RL_Local". In spite of it being an isolated issue for "RL_Local", the attribute column is retained as it is retrieved sensible values for other companies. The discarded values are limited to data from two companies (4085, 4670) for RL attribute columns "TD" and "DEPTH". A total of 58,706 records for MaxDepth are retrieved from the database: 58,642 of which are extracted as is, while 64 entries are disregarded for having a value of -999. The discarded values come from 8 companies. Null values are disregarded and absent RL or MaxDepth values. The "clean" collar export file contains at least either a value for RL or MaxDepth. The reasoning behind keeping records with at least one of the two field is there are other ways to extract for RL or MaxDepth from the database. RL values can be extracted from digital terrain models and MaxDepth values can be taken for the largest ToDepth values from the other tables. 93% of the available collar data in the area is extracted successfully. This can be improved by implementing alternative ways for retrieving RL and MaxDepth values. For example, if no RL values are fetched from the database, it could be fetched from open-source digital terrain models (DTM) and/or SRTM (Shuttle Radar Topography Mission). As for missing MaxDepth values, the maximum ToDepth values in the survey and/or interval tables could be used.

B4 Survey Extraction

Survey extraction workflow fetches the DHSurveyID, CollarID and Depth information from the *dhsurvey* table (a, red), while the corresponding Azimuth and Inclination values are fetched from the *dhsurveyattr* table using the Drill Hole Survey Azimuth Thesaurus (b, blue) and Drill Hole Survey Inclination Thesaurus (c, orange). With the same inputs defined in the configuration file, the *dh2loop* library outputs a survey CSV file containing the following information: CollarID, Depth, Azimuth, Inclination and Calculated X, Y, Z values. The workflow accommodates for underground holes drilled upwards as long as the metadata and data appropriately describe them as such. For all properties, all non-numeric values are ignored. For Depth, negative values are replaced by their absolute value. This assumption is made as some drill holes have negative depth information and it is technically not possible to have a negative length. This is done by some companies to denote that the depth measure is going upwards (usually for underground probing drill holes) rather than downhole. For Azimuth, the code fetches values between 0-360 degrees, thus ignoring values greater than 360. Values between -360 to 0 are assumed to be counter-clockwise from the north. If there is no survey information for a drill hole present in collar, the

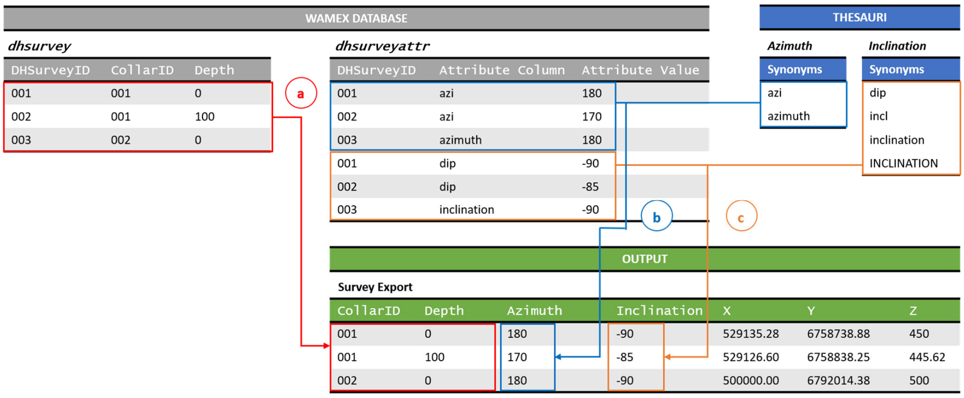
Formatted: Not Highlight

Formatted: Not Highlight

Formatted: Normal

Formatted: Not Highlight

azimuth value is set to 0. The X, Y, Z, values are calculated using the minimum curvature basing the code off the *pyGSLIB* drill hole module.



For the survey extraction, the `dhsurvey` table contained 146,713 survey depth intervals (from 45,708 drill holes) with corresponding 850,507 entries of supplementary survey information in `dhsurveyattr`. Survey extraction in YSGB resulted in 126,669 survey depth information across 45,708 drill holes with azimuth (-52.5 to 359) and inclination measurements (0-90) for each depth interval. A total of 517,592 records for `Azimuth` are retrieved from the database. 77 `Azimuth` values greater than 360 are retrieved and thus disregarded. 152 values are non-numeric values and are also disregarded. These discarded values involved 228 holes across 10 companies. A value of 0 is assigned to missing `Azimuth` values. A total of 118,223 records for `Inclination` are fetched from the database, 118,138 of which are extracted as is, while 95 entries are disregarded for having a value greater than 90. A values of -90 is assigned as the default for `Inclination`. The discarded values correspond to 94 drill holes across 5 companies. The survey extraction rate of 86% is fairly good. *dh2loop* ensures that the `Azimuth` and `Inclination` values are sensible measurements before including them into the extracted output file. An improvement that could be implemented is to run an assessment on the deflection angles for each drill hole and flag intervals with unrealistic deflection angles.

Formatted: Normal

References

~~Amorin, R.: Application of Minimum Curvature method to Well-path Calculations, 2009.~~

Ailleres, L., Jessell, M., de Kemp, E., Caumon, G., Wellmann, F., Grose, L., Armit, R., Lindsay, M., Giraud, J., Brodaric, B., Harrison, M., and Courrioux, G.: Loop - Enabling 3D stochastic geological modelling, ASEG Extended Abstracts, 2019, 1-3, 10.1080/22020586.2019.12072955, 2019.

Anand, R. R., and Butt, C. R. M.: A guide for mineral exploration through the regolith in the Yilgarn Craton, Western Australia, Australian Journal of Earth Sciences, 57, 1015-1114, Pii 929860728
10.1080/08120099.2010.522823, 2010.

Arabjamaloei, R., Edalatkha, S., Jamshidi, E., Nabaei, M., Beidokhti, M., and Azad, M.: Exact Lithologic Boundary Detection Based on Wavelet Transform Analysis and Real-Time Investigation of Facies Discontinuities Using Drilling Data, Petroleum Science and Technology, 29, 569-578, Pii 933125287
10.1080/10916460903419206, 2011.

Barley, M. E., Brown, S. J. A., Krapez, B., and Kositcin, N.: Physical volcanology and geochemistry of a Late Archaean volcanic arc: Kurnalpi and Gindalbie Terranes, Eastern Goldfields Superterrane, Western Australia, Precambrian Research, 161, 53-76, 10.1016/j.precamres.2007.06.019, 2008.

Chace, F. M.: Abbreviations in field and mine geological mapping, Economic Geology, 51, 712-723, 1956.

Cockbain, A. E.: Regolith geology of the Yilgarn Craton - Introduction, Australian Journal of Earth Sciences, 49, 1-1, DOI 10.1046/j.1440-0952.2002.00913.x, 2002.

Cohen, A.: FuzzyWuzzy: Fuzzy string matching in python, ChairNerd Blog, 22, 2011.

Culshaw, M. G.: From concept towards reality: developing the attributed 3D geological model of the shallow subsurface, Quarterly Journal of Engineering Geology and Hydrogeology, 38, 231-+, Doi 10.1144/1470-9236/04-072, 2005.

Eggleton, R. A.: The regolith glossary, Cooperative Centre for Landscape Evolution and Mineral Exploration, National Capital Printing: Canberra, 2001.

~~Elizabeth Griffin, R.: When are Old Data New Data?, GeoResJ, 6, 92-97, 10.1016/j.grj.2015.02.004, 2015.~~

Emelyanova, I., Pervukhina, M., Clennell, M., and Dyt, C.: Unsupervised identification of electrofacies employing machine learning, 2017.

Erkan, G., and Radev, D. R.: LexRank: Graph-based lexical centrality as salience in text summarization, Journal of Artificial Intelligence Research, 22, 457-479, DOI 10.1613/jair.1523, 2004.

Fuentes, I., Padarian, J., Iwanaga, T., and Vervoort, R. W.: 3D lithological mapping of borehole descriptions using word embeddings, Computers & Geosciences, 141, ARTN 104516
10.1016/j.cageo.2020.104516, 2020.

Gillespie, M., and Styles, M.: BGS rock classification scheme, Volume 1. Classification of igneous rocks, 1999.

Griffin, R. E.: When are Old Data New Data?, GeoResJ, 6, 92-97, 10.1016/j.grj.2015.02.004, 2015.

Groves, D. I., Goldfarb, R. J., Knox-Robinson, C. M., Ojala, J., Gardoll, S., Yun, G. Y., and Holyland, P.: Late-kinematic timing of orogenic gold deposits and significance for computer-based exploration techniques with emphasis on the Yilgarn Block, Western Australia, Ore Geology Reviews, 17, 1-38, Doi 10.1016/S0169-1368(00)00002-0, 2000.

Hall, M., and Keppie, F.: Striplog: new open source software for handling and analysing discontinuous and qualitative data, 2016.

Hallsworth, C. R., and Knox, R.: BGS rock classification scheme. Volume 3, classification of sediments and sedimentary rocks, 1999.

Higgins, R. F., and Mehta, S.: SeatGeek, 2018.

Hill, E. J., Robertson, J., and Uvarova, Y.: Multiscale hierarchical domaining and compression of drill hole data, Computers & Geosciences, 79, 47-57, 10.1016/j.cageo.2015.03.005, 2015.

Hill, E. J., Pearce, M. A., and Stromberg, J. M.: Improving Automated Geological Logging of Drill Holes by Incorporating Multiscale Spatial Methods, Mathematical Geosciences, 1-33, 10.1007/s11004-020-09859-0, 2020.

Jallan, Y., Brogan, E., Ashuri, B., and Clevenger, C. M.: Application of Natural Language Processing and Text Mining to Identify Patterns in Construction-Defect Litigation Cases, Journal of Legal Affairs and Dispute Resolution in Engineering and Construction, 11, Unsp 04519024
10.1061/(Asce)La.1943-4170.0000308, 2019.

Kumari, S., Mohan, A., and Saberwal, G.: Hidden duplicates: 10s or 100s of Indian trials, registered with ClinicalTrials.gov, have not been registered in India, as required by law, PLoS One, 15, e0234925, 10.1371/journal.pone.0234925, 2020.

Lark, R. M., Thorpe, S., Kessler, H., and Mathers, S. J.: Interpretative modelling of a geological cross section from boreholes: sources of uncertainty and their quantification, Solid Earth, 5, 1189-1203, 10.5194/se-5-1189-2014, 2014.

Le Vaillant, M., Hill, J., and Barnes, S. J.: Simplifying drill-hole domains for 3D geochemical modelling: An example from the Kevitsa Ni-Cu-(PGE) deposit, Ore Geology Reviews, 90, 388-398, 10.1016/j.oregeorev.2017.05.020, 2017.

Lin, C.-Y., and Hovy, E.: Automatic evaluation of summaries using n-gram co-occurrence statistics, Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, 2003, 150-157,

Lindsay, M.: Geological Interpretation of Geophysics Support from Machine Learning of Drillcore, 2019.

Liu, K., Hogan, W. R., and Crowley, R. S.: Natural Language Processing methods and systems for biomedical ontology learning, J Biomed Inform, 44, 163-179, 10.1016/j.jbi.2010.07.006, 2011.

Liu, T., and Guo, J.: Text similarity computing based on standard deviation, International Conference on Intelligent Computing, 2005, 456-464,

Formatted: Section start: Continuous

Loper, E., and Bird, S.: NLTK: the natural language toolkit, arXiv preprint cs/0205028, 2002.
McKinney, W.: pandas: a foundational Python library for data analysis and statistics, Python for High Performance and Scientific Computing, 14, 2011.
McMillan, A., and Powell, J.: British Geological Survey Rock Classification Scheme: The Classification of Artificial (man made) Ground and Natural Superficial Deposits: Applications to Geological Maps and Datasets in the UK, British Geolgoical Survey Research Report RR/99/4, 1999.
Miles, A., and Bechhofer, S.: SKOS simple knowledge organization system reference, W3C recommendation, 18, W3C, 2009.
Müller, T., Cotterell, R., Fraser, A., and Schütze, H.: Joint lemmatization and morphological tagging with lemming, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, 2268-2274,
Myers, J.: Precambrian Tectonic History of the West Australian Craton and Adjacent Orogens, Annu Rev Earth Pl Sc, 21, 453-485, 10.1146/annurev.ea.21.050193.002321, 1993.
Okuda, T., Tanaka, E., and Kasai, T.: A method for the correction of garbled words based on the Levenshtein metric, IEEE Transactions on Computers, 100, 172-178, 1976.
Otter, D. W., Medina, J. R., and Kalita, J. K.: A Survey of the Usages of Deep Learning for Natural Language Processing, IEEE Trans Neural Netw Learn Syst, 10.1109/TNNLS.2020.2979670, 2020.
Padarian, J., and Fuentes, I.: Word embeddings for application in geosciences: development, evaluation and examples of soil-related concepts, 10.5194/soil-2018-44, 2019.
Park, S. H., Ryu, K. H., and Gilbert, D.: Fast similarity search for protein 3D structures using topological pattern matching based on spatial relations, Int J Neural Syst, 15, 287-296, 10.1142/S0129065705000244, 2005.
Qiu, Q., Xie, Z., Wu, L., and Tao, L.: Dictionary-Based Automated Information Extraction From Geological Documents Using a Deep Learning Algorithm, Earth and Space Science, 7, 10.1029/2019ea000993, 2020.
Ralph, J.: Mindat. org–the mineral database, Mindat, Surrey, England, 2004.
Raymond, O., Duclaux, G., Boisvert, E., Cipolloni, C., Cox, S., Laxton, J., Letourneau, F., Richard, S., Ritchie, A., and Sen, M.: GeoSciML v3. 0-a significant upgrade of the CGI-IUGS geoscience data model, EGUGA, 2711, 2012.
Recasens, M., Danescu-Niculescu-Mizil, C., and Jurafsky, D.: Linguistic models for analyzing and detecting biased language, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2013, 1650-1659,
Richard, S. M., Boisvert, E., Brodaric, B., Cox, S., and Duffy, T.: GeoSciML–a GML application for geoscience information interchange, Philadelphia Annual Meeting, 2007, 47-59,
Riganti, A., Farrell, T. R., Ellis, M. J., Irimies, F., Strickland, C. D., Martin, S. K., and Wallace, D. J.: 125 years of legacy data at the Geological Survey of Western Australia: Capture and delivery, GeoResJ, 6, 175-194, 10.1016/j.grj.2015.02.015, 2015.
Rivera-Quiroz, F. A., and Miller, J.: Extracting Data from Legacy Taxonomic Literature: Applications for planning field work, Biodiversity Information Science and Standards, 3, 10.3897/biss.3.37082, 2019.
Robertson, S.: BGS rock classification scheme. Volume 2, Classification of metamorphic rocks, 1999.
Rosenbaum, M. S., McMillan, A. A., Powell, J. H., Cooper, A. H., Culshaw, M. G., and Northmore, K. J.: Classification of artificial (man-made) ground, Engineering Geology, 69, 399-409, 10.1016/S0013-7952(02)00282-X, 2003.
Ross, P. S., Bourke, A., and Fresia, B.: A multi-sensor logger for rock cores: Methodology and preliminary results from the Matagami mining camp, Canada, Ore Geology Reviews, 53, 93-111, 10.1016/j.oregeorev.2013.01.002, 2013.
Rothwell, R. G., and Rack, F. R.: New techniques in sediment core analysis: an introduction, New Techniques in Sediment Core Analysis, 267, 1-29, Doi 10.1144/Gsl.Sp.2006.267.01.01, 2006.
Schetselaar, E. M., and Lemieux, D.: A drill hole query algorithm for extracting lithostratigraphic contacts in support of 3D geologic modelling in crystalline basement, Computers & Geosciences, 44, 146-155, 10.1016/j.cageo.2011.10.015, 2012.
Simons, B., Boisvert, E., Brodaric, B., Cox, S., Duffy, T. R., Johnson, B. R., Laxton, J. L., and Richard, S.: GeoSciML: enabling the exchange of geological map data, ASEG Extended Abstracts, 2006, 1-4, 2006.
Smith, M. J., Keesstra, S., and Rose, J.: Use of legacy data in geomorphological research, GeoResJ, 6, 74-80, 10.1016/j.grj.2015.02.008, 2015.
Vearncombe, J., Conner, G., and Bright, S.: Value from legacy data, ~~Transactions of the Institutions of Mining and Metallurgy Section B-Applied Earth Science~~T I Min Metall B, 125, 231-246, 10.1080/03717453.2016.1190442, 2016.
Vearncombe, J., Riganti, A., Isles, D., and Bright, S.: Data upcycling, Ore Geology Reviews, 89, 887-893, 10.1016/j.oregeorev.2017.07.009, 2017.
Wang, C., and Ma, X.: Text Mining to Facilitate Domain Knowledge Discovery, in: Text Mining-Analysis, Programming and Application, IntechOpen, 2019.
Wilbur, W. J., and Sirotkin, K.: The Automatic Identification of Stop Words, Journal of Information Science, 18, 45-55, Doi 10.1177/016555159201800106, 1992.
Zhou, C. Y., Ouyang, J. W., Ming, W. H., Zhang, G. H., Du, Z. C., and Liu, Z.: A Stratigraphic Prediction Method Based on Machine Learning, Appl Sci-Basel, 9, ARTN 3553
10.3390/app9173553, 2019.
Zhou, Q., Liu, H. H., Bodvarsson, G. S., and Oldenburg, C. M.: Flow and transport in unsaturated fractured rock: effects of multiscale heterogeneity of hydrogeologic properties, J Contam Hydrol, 60, 1-30, 10.1016/s0169-7722(02)00080-3, 2003.
Zhu, G., Gao, M., Kong, F., and Li, K.: Application of Logging While Drilling Tool in Formation Boundary Detection and Geo-steering, Sensors (Basel), 19, 10.3390/s19122754, 2019.

Formatted: English (Australia)

Formatted: Heading 1

Page 27: [1] Formatted	Ranee Joshi	12/04/2021 10:03:00 AM

Font: Lucida Console, 10 pt, English (United Kingdom)

Page 27: [2] Formatted	Ranee Joshi	12/04/2021 10:03:00 AM

Font: Lucida Console, 10 pt, English (United Kingdom)

Page 27: [3] Formatted	Ranee Joshi	12/04/2021 10:03:00 AM

Font: Lucida Console, 10 pt, English (United Kingdom)

Page 27: [4] Formatted	Ranee Joshi	12/04/2021 10:03:00 AM

Font: Lucida Console, 10 pt, English (United Kingdom)

Page 27: [5] Formatted	Ranee Joshi	12/04/2021 10:03:00 AM

Font: Lucida Console, 10 pt, English (United Kingdom)

Page 27: [6] Formatted	Ranee Joshi	12/04/2021 10:03:00 AM

Font: Lucida Console, 10 pt, English (United Kingdom)