Response to Anonymous Referee #1:

The breakdown of replies to the individual comments are below:

General comments

The paper is too long and dense in terminology and nuanced meaning. The authors have endeavoured to symbolise some of the different and complicated database terminology although possibly more is needed. The flow is logical and the writing is generally understandable although laboured in places. The figures and tables are appropriate and informative in most places; some minor improvements have been suggested. I have not checked references or URL links exhaustively.
⟶ In the updated version, we reduced the length of the paper by putting the details of Section 2.4.2 Survey Extraction in the Appendix. We addressed the dense terminology by defining all the variables in a table in Section 2.1. Conventions. We symbolized the workflows differently as you have suggested.

Specific comments

Introduction, section 1: replace first 2 sentences with 'Drilling is a process of penetrating through the ground that is capable of extracting information about rocks from various depths below the surface. This is useful for establishing the geology beneath. Drill core or cuttings can be collected thus providing samples for description, interpretation and analysis.'
⟶ This was revised in the updated version.

Introduction, section 3: The legacy data described seem to be hardcopy forms subsequently digitised. Legacy digital data also suffer from lack of standardisation, inconsistency.
⟶ Legacy digital data is included and this correction is revised in the updated version.

Introduction, line 78: These data are not 'unstructured' but they may not conform to standards or be consistently applied/described.
⟶ Unstructured data has an imprecise definition across different references. We agree there is some ambiguity as the data does reside in a relational database (as structured data is). However, the data we are dealing with requires text analysis to sort and extract data. In this study, we refer to "unstructured" to mean that the written content cannot be readily mapped onto standard database fields and not easily searchable. In the case of the free text comments, descriptors such as color, age, texture are not written in any order or using any standard. In the manuscript, we added description as to what we mean by "unstructured" for clarity, while acknowledging the potential ambiguity.

Material and Methods, Conventions: workflows need their own distinct convention (font). Later they are confusingly rendered as combinations of database table fields.
⟶ This was revised in the updated version.

Thesauri: Some of the so-called 'synonyms' actually have distinct meanings from each other, even the listed example elevation vs relative (reduced) level. Maybe qualify 'synonym' as meaning 'nearly the same' or a 'close match' for their general intent is similar e.g the elevation terms all are recording a vertical height.
⟶ Definition of "synonyms" in this work is expounded in the updated version.

Thesauri, line 245: Rather than 'The opposite is true as well' suggest you explain specifically that more than one code may refer to the same lithology. Basically there is a many-to-many relationship between code and lithology.
⟶ We agree and applied this in the updated version.

Thesauri, line 252: The CGI vocabularies support the GeoSciML and EarthResourceML (note singular) geology data models but potentially other applications. Suggest you rephrase as '...the CGI-IUGS geoscience vocabularies accessible at http://geosciml.org/resource/def/voc/'.
⟶ This was revised in the updated version.

Thesauri, Lithology Hierarchical Thesaurus: The 3-level hierarchy is highly simplified compared to CGI's Simple Lithology. Many of the 'Lithology_Subgroups' listed have parent-child relationships e.g. 'mafic_fine_grained_crystalline' is a child of 'mafic'. This should be mentioned, presumably some simplification and pragmatism is needed for your analysis.
⟶ Parents in parent-child relationships are included with 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 term alone.

Data Extraction, Collar Extraction: I wondered why this section needs to be here at all. The collar extraction isn't central to your paper focus on lithology. You don't utilise collar location in a spatial analysis or context - its only function in this paper seems to be a pre-filter for data quality. The method itself is good and useful, and ultimately important for data mining where spatial understanding is needed, just not essential for the lithology-driven analysis presented here.
⟶ We understand the concern with having this section in the paper. However, while the drill collar may appear to be a trivial example, it is to prime the reader for the more complicated following sections. We were using the collar extraction as an introduction to the database structure before transitioning to the database structure for lithology which also has two workflows. This is also to emphasize that dh2loop provides the three basic interval tables needed to import geological data into 3D modelling softwares.

Data Extraction, Survey Extraction: Ditto, I wondered why this section needs to be here at all. If it is retained then suggest the 4th field should be ‘Inclination’ not ‘Dip’. Dip is a measurement of the slope of a planar surface feature whereas inclination refers to the plunge of a linear feature. Additionally, how consistent are WAMEX records around positive inclinations meaning upwards-directed drill holes?
⟶ We addressed this by putting the bulk of the Survey extraction as part of the Appendix. The term “dip” is also changed into “inclination” in both the paper and the code. In this dataset, we obtained 1014 dip values that are between 0 to 90 (from 684 holes, 23 companies, 28 reports). These examples of underground holes drilled upwards are accommodated in the workflow, so long as the metadata and data appropriately describe them as such. Detecting where an upwards directed hole is when not reported as such is beyond the scope of the paper, and second-guesses the database.

Data Extraction, Survey Extraction: The ‘Calculated X, Y, Z values’ are not particularly helpful or necessary in this Survey table i.e. only recording collar and the end of hole locations. Survey tables more typically describe changing azimuths and inclinations with ‘depth’ (i.e. account for curved drill holes) – does WAMEX not do this?
⟶ WAMEX does describe changing azimuths and inclinations with depth. The Calculated X, Y, Z is a functionality made available as it is also used in the Lithology Table. This is shown here as some software accept calculated XYZ as input parameters for survey. Since we moved this section to the Appendix, we think it be alright to include this detail.

Data Extraction, Lithology Extraction: The fuzzywuzzy algorithm appears to be repeating pre-processing already mentioned in the previous paragraph (line 419-424).
⟶ The pre-processing cleans up the text, while the fuzzywuzzy algorithm matches the text to a dictionary. We added this text to the introduction of Section 2.4.

Data Extraction, Line 432-433: What does ‘Since the sorted intersection component of token_set(), will result in an exact match…’ mean? Elaborate or explain more clearly.
⟶ Sorted intersection tokens are the similar tokens (characters) between the two strings, thus it always equates to an exact match (=100). The remainder component is what lowers the score. We included this in the text.

Data Extraction, Line 439: What is an ‘intersection token’?
⟶ Intersection token are the similar tokens between the two strings. We added this into the text.

Data Extraction, Table 1: The column of ticks and crosses is unexplained. In two cases the lower score is ticked implying it is the preferred result?
⟶ Yes, we expanded the caption to clear this. This was revised in the updated version.

Data Extraction, Line 453: ‘Andesitic basalt’ is an unfortunate example since ‘basaltic andesite’ is an established volcanic rock name. Would basaltic andesite wrongly revert to andesite in this classification process?
⟶ This raises a good point. We suggest to sticking with the “andesitic basalt” example, but we further explained that in the case of “basaltic_andesite” it will not be simplified into andesite, as the thesaurus includes an entry for basaltic_andesite, since it is an established volcanic rock name as you have mentioned.

Data Extraction, Figure 8: I struggled to understand this graph. How can data with 100% Exact Match score only 80%?
⟶ The graph shows that using a smaller dataset, we could notice that using a score cut-off of 80, ~99% are returned as exact matches. This exercise tests at what cut-off score does the number of exact matches plateau. This is to avoid using a very stringent 100 cut-off to capture Exact Matches. This cut-off score is a parameter that can be changed and is dependent on the data being processed. We added this into the figure caption for more clarity.

Data Extraction Results, Unique Lithology Code Results: Database table field names seem to be inconsistent e.g. Company_LithoCode vs Company_Lithology vs Lithology_Code. Suggest careful check to ensure consistency of use otherwise confusing for the reader.
⟶ This was revised in the updated version.

Data Extraction Results, Unique Lithology Code Results: Workflows such as ‘Lithology_Code Detailed_Lithology’ need to be distinctly symbolised. At the moment they look like co-joined database table names without an obvious algorithm progression between them. Suggest also where these are mentioned you mention ‘workflow’ or ‘workflows’
⟶ This was revised in the updated version.

Data Extraction Results, Table 2: Struggling to understand why row 3 is a Close Match when the almost identical row 5 is a Broad Match? If anything the ‘basic volcanic rock’, not being a recognised Lithology_SubGroup member, is broader rather than closer than ‘mafic fine grained crystalline’.
⟶ We agree with the comment. We changed “basic volcanic rock” to “basaltoid” to better illustrate the difference.

Data Extraction Results, Fuzzy String Matching Results: Mentions of ‘comments’ should be ‘Comments’ in most cases, possibly with special font.
⟶ This was be revised in the updated version.

Data Extraction Results, lines 611, 622: These results are suboptimal. You discuss this later but it seems your method is sometimes picking a subordinate lithology rather than the dominant lithology.
⟶ We kept the discussion of the results as part of the discussion.

Data Extraction Results, line 653: I wasn't clear what 'the limitation' is – processing?
⟶ The limitation we are describing is that it is not possible to compare the matches for the cases where only one of the workflows arrive with a match. We expounded on this in the updated version manuscript.

Discussion, Assessment of String Matching Results (line 844): Need to qualify that the 'classification of structures and textures and metamorphic rocks' has higher confidence in the study area dataset, not necessarily in others. I'm sure there will metamorphic-dominated terranes where the subordinate igneous rocks will be classified with higher confidence.
⟶ We expounded on this in the updated version.

Technical corrections

line 44: delete ', particularly as it is likely to have been conducted by tens to hundreds of geologists...' with something like 'as all logging geologists have their own personal biases.
⟶ This was revised in the updated version.

Line 49-50: delete 'even detection of'
⟶ This was revised in the updated version.

line 57: The semi-automatic methods also are poor at describing textural characteristics (foliation, banding, grainsize variation)
⟶ This was revised in the updated version.

line 70" Delete 'Elizabeth'?
⟶ This was revised in the updated version.

line 105: limitations -> limitation
⟶ This was revised in the updated version.

line 198: replace 'that occurred' with 'were emplaced'
⟶ This was revised in the updated version.

line 199: replace 'ultramafic mafic' with 'ultramafic to mafic' and 'local centres' with 'local eruptive centres'.
⟶ This was revised in the updated version.

line 201: replace 'volcanoclastic' with volcaniclastic'
⟶ This was revised in the updated version.

line 203: delete 'profiles' and delete 'both'. Break sentence after 'bedrock' and start next with 'Regolith...'
⟶ This was revised in the updated version.

line 209: suggest replacing 'complexity' with 'diversity'
⟶ We changed it to "structural complexity". "Diversity" does not capture the complexity in relationship between these lithologies.

Figure 3 needs a unit to describe dill hole density e.g. per square kilometre
⟶ This was revised in the updated version.

line 254: Insert after ‘Added records’ ‘with examples’
⟶ This was revised in the updated version.

line 270: Replace ‘GeoSciML’ with ‘the CGI-IUGS Simple Lithology vocabulary http://resource.geosciml.org/classifier/cgi/lithology’
⟶ This was revised in the updated version.

line 276: Suggest deleting second half of sentence i.e. after ‘dictionary’
⟶ This was be revised in the updated version.

line 294: Delete orphan ‘And’
⟶ This was revised in the updated version.

Figure 4: Lighten purple shade (or whiten text within)
⟶ This was revised in the updated version.

line 358: Replace ‘Dip: it is the inclination angle perpendicular to the azimuth…’ with ‘Inclination: the plunge angle of the drill hole relative to horizontal…’.
⟶ This was revised in the updated version.

line 358-360: Replace sentence ‘A positive value indicates an upward-directed drill hole and a negative value indicates a drill hole directed downwards.’
⟶ This was revised in the updated version.

line 411: Replace ‘The string followed by key phrases such as…’ with ‘The string preceded by key phrases such as…’
⟶ This was revised in the updated version.

line 415: Does ‘tokens with less than three characters’ mean or include short words?
⟶ Yes, it does include 2 letter words. But most of two-letter words are prepositions (to, in, at, etc.). Only obvious issue should be “aa flows”, which has not been observed as terminology used in the logs.

Line 434: Insert ‘method’ after ‘ratio()’.
⟶ This was revised in the updated version.

Line 511: Font change in ‘Company_Lithology’.
⟶ This was revised in the updated version.

Line 570: Where is the ‘brown text’ in Table 2?
⟶ This was revised in the updated version. We will symbolize it as bold to avoid confusion.

Figure 10: Lighten purple shade (or whiten text within).
⟶ This was revised in the updated version.

Line 598: replace ‘take a look’ with ‘looked’
⟶ This was revised in the updated version.

Line 663: replace ‘couple of’ with ‘four’
⟶ This was revised in the updated version.

Line 671: replace ‘trumps’ with ‘trump’
⟶ This was revised in the updated version.

Line 833: What ‘information being fed itself’ mean?
⟶ This was revised in the updated version.

Line 887: delete ‘a couple of’
⟶ This was revised in the updated version.