*Transdisciplinary Engineering for Complex Socio-technical Systems – Real-life Applications*
*J. Pokojski et al. (Eds.)*

*doi:10.3233/ATDE200058*

# Text Analysis of Disciplinary Research Papers

Emily CAREY[1], James GOPSILL and Linda NEWNES
*University of Bath*

**Abstract.** Research literature terminology illustrates that publications claim to pertain to "disciplinary" approaches and researcher's align themselves to specific, multi-, inter- or trans-disciplinarities. Ambiguity exists in definition and application of disciplinarity, hence there is need to establish a coherent application of disciplinarity. We present results of content analysis of research literature claiming to be inter-, multi-, or transdisciplinary to assist in ascertaining commonalities or differences for those disciplinarities. We analyse the abstracts and keywords of 8834 papers, using n-grams and bi-grams, dating from 1970 until 2018, extracting a list of 76,552 terms for comparison. The top 15 most frequent terms characterise each disciplinarity and Venn diagrams of the top 15 features illustrate differences and overlap. A total of six terms appear common to all approaches in the abstracts, with four shared by multi- and inter-, two between inter- and trans-, and none common to multi- and trans-. The term "social science(s)" appears to be a unique feature in the trans- abstracts and our findings identify common text terms such as the "research" feature, common to all disciplinarities. This supports characterising the nature of transdisciplinarity and its unique differences from other approaches such as inclusion of social science(s).

**Keywords.** Transdisciplinary, Disciplinarity, Engineering Research, Content Analysis

## Introduction

The term "discipline" is defined in most dictionaries as "a branch of knowledge, typically one studied in higher education" [1]; the purpose of science is to advance knowledge within disciplines. Traditionally, a set of core disciplines exists, such as maths, physical sciences or humanities, however many newer disciplines, each with their own bodies of research literature, have emerged since the 1970's. As these disciplines have advanced, new merged disciplines such as systems engineering and education studies [1] have emerged and literature relating to types of "disciplinarities", has grown. These "disciplinarities" are differentiated from single disciplinary work by the use of prefixes such as multi- (MD), inter- (ID) or transdisciplinary (TD) [2] defining the governing principles for how disciplinary knowledge is used within and across disciplines. Hence a "disciplinarity" describes the disciplinary process or system within which academic knowledge overlaps and interacts, relating those specific rules for combining expertise or working amongst established core disciplines to create new knowledge.

As research problems and disciplines increase in complexity and branch into other fields, the projects and teams investigating them are forced to become multifaceted (or

[1] Corresponding Author, Mail: e.m.c.carey@bath.ac.uk.

multidimensional) in terms of their knowledge, skills and experience [3]. Global grand challenges and real-world projects typically involve increasing numbers of core academic disciplines and the focus is on the approaches taken to involve or incorporate disciplines and knowledge, hence the disciplinarity. This is especially true for complex societal projects that involve researchers, industry and society, and are typical in engineering [4], making disciplinarity an important issue in engineering. This is evidenced in reviews of scientific literature where, MD, ID and TD approaches have been populous [5,6] but the interchangeable use of the terms [7] mean their differences remain unclear.

The purpose of this paper is to describe the most frequent text terms used in literature to represent different disciplinarities and to empirically establish the differences and overlaps in terminology that exist. This characterisation and semantic clarity are needed to guide researchers to understand what these approaches, such as TD, should feature to solve societal problems [6,8]. To ascertain whether the research approach requires a specific disciplinarity such as MD, ID or TD, a means to understand their core features is much needed. The use of text analysis is one approach to determine what features are common or unique to a type of disciplinarity. Within this paper we describe the computational linguistics approach we have adopted for analysing the literature claiming to be MD, ID and TD. From this analysis we extract and rank the most frequent text concepts associated with each disciplinary category, identifying the terms unique or common. The results are presented in ranked tables and Venn diagrams, highlighting the unique features of MD, ID and TD approaches in literature.

## 1. Content Analysis Approach

The approach taken in this paper is to create representative lists of most frequent text terms for each disciplinarity by analysing text contents of academic literature pertaining to specific disciplinarities (MD, ID and TD). Samples of literature have been created using the Scopus database as it is a broad discipline database that is an "Index to journals and conference papers across all subject areas." [9], and hence gathers text from many different disciplinary approaches. Scopus is a comprehensive, general academic publication database and is considered preferential to Web of Science (the two most extensive academic databases) as it provides 20% more coverage and incorporates a wider range of journals [9]. There are many terms used to describe disciplinary approaches but for simplicity and to minimise cross definitions in our analysis, only the core disciplinary definitions MD, ID and TD have been selected for this analysis. It is noted that multiple definitions of disciplinarity can exist within a single text and hence it was necessary to ensure that each sample of literature created for the sampling pertained only to one disciplinary approach. To ensure the samples were accurately labelled as per the authors own classification of their disciplinarity for MD, ID or TD it was necessary to create a search strategy that labelled literature pertaining to only one approach. This was created by including only literature that labelled its disciplinarity in the abstract, keywords and title. The search criteria text is illustrated below in Figure 1.

Each search was created and literature samples selected as per Figure 1. The abstract text, keywords and title were downloaded as comma-separated value (.csv) files. The underlying assumption is that the abstract, keywords and title would succinctly summarise the content of each paper [10] and hence provide enough differentiating text terms to substantiate the approach [11]. The content was further processed using

established Natural Language Processing Techniques. First, the words were all made lowercase and "stopwords"** removed. Second, stemming was applied to remove variants in syntax, such as "disciplinary" and "disciplinarity".

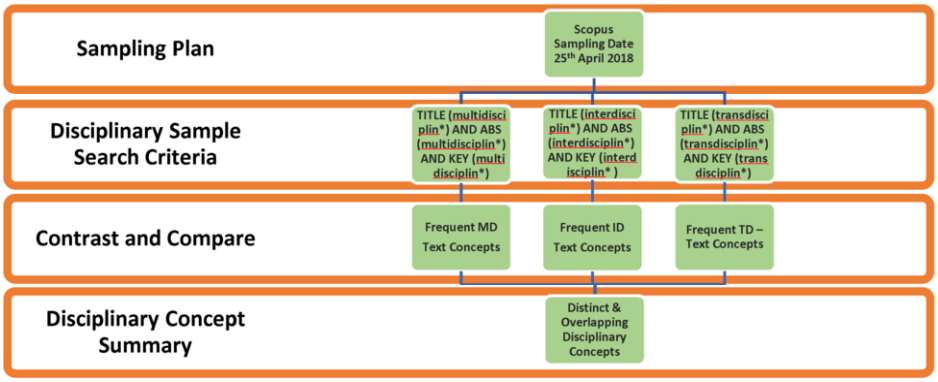


**Figure 1.** Literature sampling and research approach.

## 2. Text and Concept Analysis Results

The sampling results of the searches described in Figure 1 are shown in Table 1. If we compare the size of the Scopus literature samples obtained, encouragingly they are proportionally similar to amounts of disciplinary literature found in the work by Bruun [5], where they find that "actively" being TD forms 6% of their funding proposal sample. Their ID research formed 38-46%, however their analysis did not include any growth for the period from 2005 to 2018 and may explain this difference. Their examination of MD was not differentiated from disciplinary and therefore not possible to directly compare.

**Table 1.** Comparison of sample sizes.

| Search String & Prefix | *Trans*disciplin* | *Inter*disciplin* | *Multi*disciplin* |
|---|---|---|---|
| Literature Sample Size | 612 | 4422 | 3800 |
| Percentage of Sample | 6.92% | 50.06% | 43.02% |
| Total Keywords | 3193 | 18129 | 13935 |
| Unique Keyword Tokens | 2109 | 10706 | 8089 |
| Total Abstract Bi-grams | 7238 | 52519 | 50702 |
| Unique Abstract Bi-grams | 6018 | 37063 | 33471 |

**The Python stop-words library was used (https://pypi.org/project/stop-words/) to remove 127 commonly used words in the English language.

The analysis of keywords and abstracts followed the "bag of words" model. Keywords were split based on the Scopus keyword delimiter ";" and combined to form a list of keywords for each of the three datasets. These were subsequently reduced based on the re-occurrence of terms producing a word occurrence vector for each. Abstracts were split using a regular expression to identify bi-grams and allowing for overlaps. This formed the list of words that were reduced based on the re-occurrence of terms producing a word occurrence vector. An example of such a bi-gram result would be "sustainability science" rather than "science" (n-grams), which alone could be misinterpreted without

the context of "sustainability". The resulting frequency ratios for text terms are normalised by literature publication to represent accurately the relevance of each term proportionally over the entire sample.

The distributions in Figure 2 show the keyword and abstract bi-gram occurrence in MD, ID and TD approaches, similar to the "power law" expression of terms used the work by Liu [11]. This simply illustrates that the majority of text concepts in the resulting tables lie within the first 10% of the samples (shown to the left of the dividing line) and hence there are few very significant concepts in the literature. The relevance weighting or proportion of terms occurring to the right of this line means these terms may be of little significance to describe the samples and are unlikely to be represented in our frequent text terms.

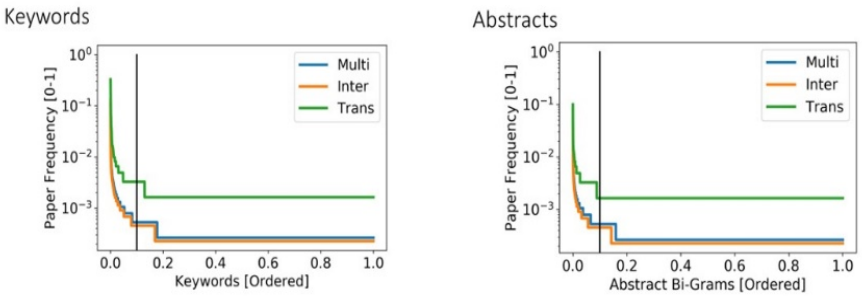


**Figure 2**. Distribution of text occurrences in MD, ID and TD samples.

The temporal distribution of the sample of literature, shown in Figure 3 below, ranges from 1970 through to 2018 and illustrates that there has been much growth in the utilisation of disciplinary references in the academic literature. Whilst references were first made in the 1970's the growth in MD and ID literature since approximately 1992 far exceeds that of the TD literature.

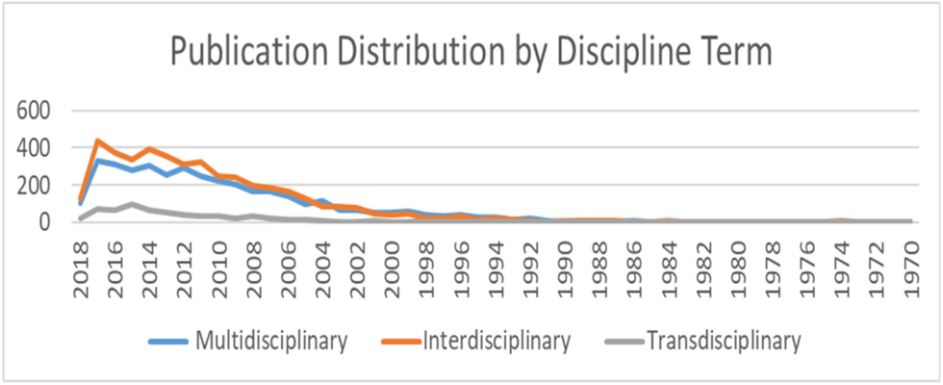


**Figure 3.** Temporal distribution of literature samples.

### *2.1 Keywords and Abstracts*

The top 10 sample results of the Keyword and Abstracts processing are shown in Table 2 and 3 respectively, with the most frequently occurring terms being listed first. The

number of unique tokens processed for the Keywords and Abstracts are listed in Table 1. The representative frequencies are shown in ratio values of the overall samples to enable direct comparison. From these tables it is evident in the TD sample, that there is cross reference to other disciplinarities (use of ID) in the text terms used, substantiating the interchangeable use of terms issue. Due to the extensive size of complete samples, full results are available upon request.

**Table 2.** Keywords: Top 10 occuring text concepts.

| MD | Ratio | ID | Ratio | TD | Ratio |
|---|---|---|---|---|---|
| multidisciplinary | 0.131 | interdisciplinary | 0.161 | transdisciplinarity | 0.388 |
| multidisciplinary design optimization | 0.074 | interdisciplinarity | 0.147 | transdisciplinary | 0.212 |
| multidisciplinary team | 0.059 | interdisciplinary research | 0.052 | transdisciplinary research | 0.120 |
| multidisciplinary approach | 0.039 | collaboration | 0.027 | interdisciplinarity | 0.078 |
| multidisciplinary care | 0.032 | education | 0.025 | sustainability | 0.055 |
| multidisciplinary treatment | 0.032 | interdisciplinary collaboration | 0.022 | interdisciplinary | 0.047 |
| multidisciplinary teams | 0.023 | interdisciplinary education | 0.020 | complexity | 0.039 |
| multidisciplinary design optimization (mdo) | 0.017 | interdisciplinary approach | 0.016 | collaboration | 0.039 |
| multidisciplinarity | 0.016 | interdisciplinary team | 0.015 | sustainable development | 0.026 |
| multidisciplinary optimization | 0.015 | communication | 0.012 | education | 0.022 |

**Table 3.** Abstract: Top 10 occuring text concepts.

| MD | Ratio | ID | Ratio | TD | Ratio |
|---|---|---|---|---|---|
| multidisciplinary design | 0.083 | interdisciplinary research | 0.054 | transdisciplinary research | 0.099 |
| design optimization | 0.057 | interdisciplinary approach | 0.035 | transdisciplinary approach | 0.050 |
| multidisciplinary team | 0.055 | paper describes | 0.029 | paper presents | 0.039 |
| multidisciplinary approach | 0.043 | paper presents | 0.028 | case study | 0.035 |
| paper presents | 0.040 | interdisciplinary team | 0.028 | paper describes | 0.021 |
| paper describes | 0.029 | health care | 0.023 | transdisciplinary approaches | 0.021 |
| results show | 0.022 | interdisciplinary collaboration | 0.020 | article describes | 0.019 |
| multidisciplinary optimization | 0.021 | case study | 0.017 | social sciences | 0.016 |
| health care | 0.020 | article describes | 0.016 | research project | 0.016 |
| multidisciplinary care | 0.019 | article presents | 0.012 | paper explores | 0.016 |

### *2.2 Combined Disciplinary Text Terms*

In Table 4 and 5 , we further process the results of our text frequency analysis to reduce the terms to those most representative of MD, ID and TD. These have been calculated using the ratios shown in Tables 2 and 3, with some manual post processing of the samples. The final text frequency lists are the result of manual processing of Tables 2 and 3 results to standardise singular and plural forms, merge synonyms, nouns, gerunds,

abbreviations and acronyms [11]. Additional "stopwords" and misleading terms such as "multidisciplinary" or alternative references to the sample designation have also manually been removed from the results to reduce to those concepts most relevant.

**Table 4.** Most frequent keyword concepts.

| MD | %age | ID | %age | TD | %age |
|---|---|---|---|---|---|
| design optimisation | 9.1 | research | 6.0 | research | 23.00 |
| team/teams | 8.2 | collaboration | 4.9 | sustainability/sustainable development/sustainability science | 9.7 |
| approach | 3.9 | education | 4.5 | complexity | 3.9 |
| care | 3.2 | team/teams | 2.7 | collaboration | 2.6 |
| treatment | 3.2 | communication | 2.4 | education | 2.2 |
| optimisation | 2.5 | approach | 1.6 | knowledge integration | 2.1 |
| rehabilitation | 2.5 | higher education | 1.0 | evaluation | 2.1 |
| breast cancer/cancer | 2.3 | treatment | 1.0 | approach | 1.9 |
| design | 1.5 | rehabilitation | 1.0 | participation | 1.8 |
| collaborative optimisation | 1.4 | team work | 0.9 | climate change | 1.8 |
| chronic pain | 1.4 | curriculum | 0.9 | methodology | 1.6 |
| quality of life | 1.0 | chronic pain | 0.9 | epistemology | 1.4 |
| education | 1.0 | care | 0.9 | health | 1.4 |
| clinic | 0.9 | sustainability | 0.9 | integration | 1.3 |
| communication | 0.6 | learning | 0.8 | creativity | 1.3 |

**Table 5.** Most frequent abstract concepts.

| MD | %age | ID | %age | TD | %age |
|---|---|---|---|---|---|
| design | 8.3 | research | 5.4 | research/research project | 11.5 |
| team/teams | 7.1 | team/teams | 4 | approach/approaches | 7.1 |
| design optimisation | 5.7 | approach | 3.6 | case study/ies | 4.7 |
| approach | 4.4 | case study/ies | 2.5 | social science/s | 2.6 |
| optimisation | 2.1 | health care | 2.4 | research process | 1.5 |
| health care | 2.0 | collaboration | 2 | climate change | 1.5 |
| care | 1.9 | work | 1 | knowledge | 1.3 |
| treatment | 1.7 | mental health | 0.9 | health care | 1.3 |
| design variables | 1.4 | different disciplines | 0.9 | action research | 1.3 |
| optimisation problem | 1.3 | team members | 0.8 | process | 1.1 |
| design process | 1.3 | treatment | 0.8 | knowledge integration | 1.1 |
| case study | 1.2 | design | 0.8 | model | 1.1 |
| collaborative optimisation | 1.2 | course | 0.8 | conceptual framework | 1.0 |
| confidence interval | 1.2 | higher education | 0.8 | across disciplines | 1.0 |
| proposed method | 1.2 | learning | 0.7 | different disciplines | 1.0 |
| | | nature | 0.7 | sustainable development | 1.0 |

The number of terms that have been merged and removed in the manual post processing mean few concepts remained to describe each sample adequately. Hence a practical approach was taken to include the resulting top most frequent 15 text concepts for each of the disciplinarities for equal comparison. Where there are more terms included in the list, this is as a result of an equal ratio weighting in the 15$^{th}$ term, and all equally frequent concepts are included for completeness. A final list of text terms to describe each of the three disciplinarities is shown in Tables 4 and 5.

There are independent text terms emerging for each of the disciplinarities, however, there is considerable overlap with terms appearing in multiple lists. This is evident in Figures 4, 5, in the Venn diagrams illustrating both independent and overlapping terms. The frequency of terms in Tables 4 and 5 indicates those terms most significant across and within the samples as they are normalized by the number of papers in each sample.

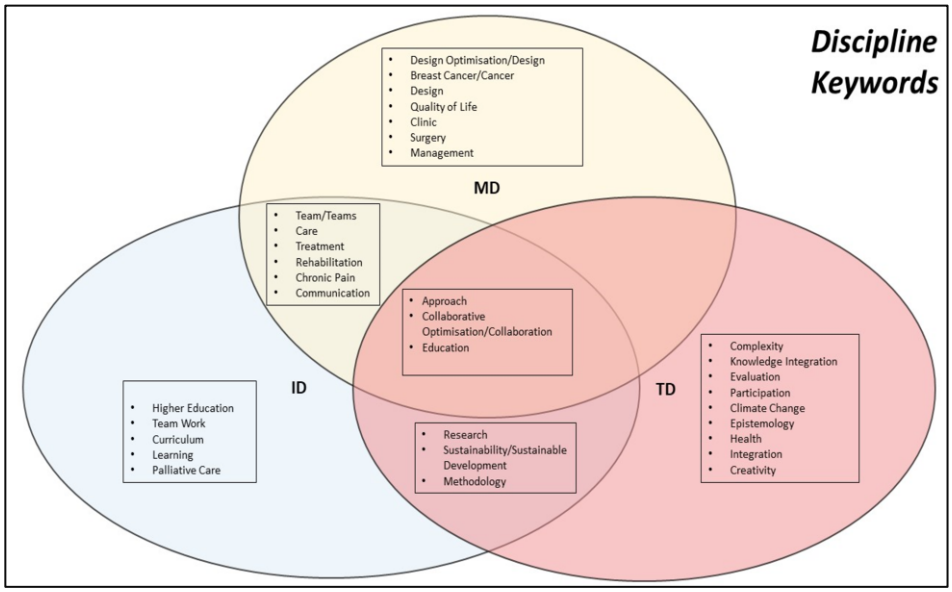


**Figure 4**. Keywords Overlap.

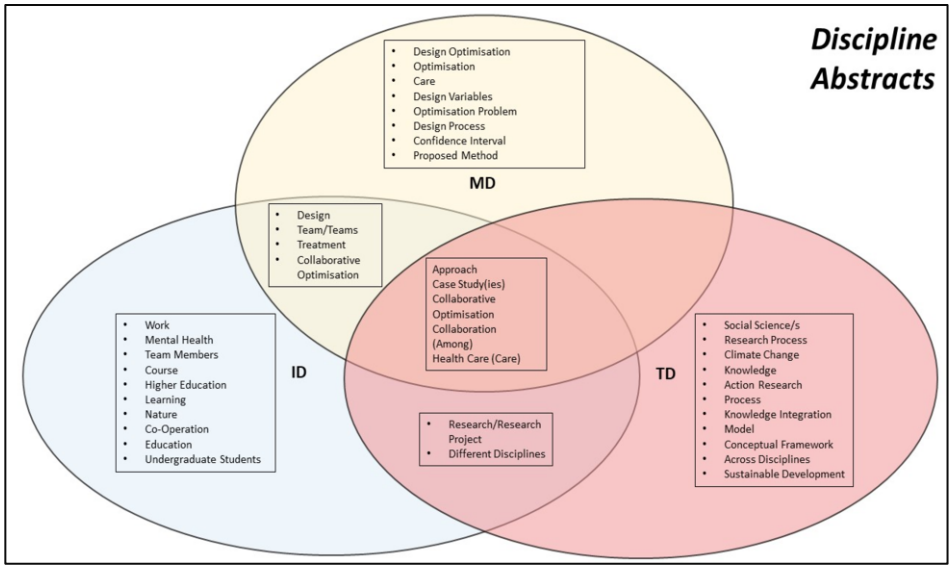


**Figure 5**. Abstracts overlap.

In Tables 4 and 5 the rate of occurrence of some terms in the "TD" category appear higher than the rate of the same term in the other two categories for example "research" and as the results have been normalized this could indicate significant category features. However, this correlation is observed for academic literature samples and is arguably expected. The authors suggest that this is a limit of the data sample and that further sampling of wider disciplinary text is needed to investigate any hypothesized relationships. Similarly, results suggest the most frequent terms are statistically significant features in respective samples and further work should measure relative significance by comparing frequency distance measures across more or less frequent text terms and disciplinary samples. For example, the relative high frequency of "climate change" in the TD sample, could be compared to its frequency in the ID sample and to other text terms and the natural next stage of this research is to investigate features that appear strongly correlated to samples. Whilst beyond the extent of this work, they could indicate important features and is briefly discussed with a highlight of interesting terms that have potential to form descriptive features for each disciplinarity.

## 3. Discussion

The growth in the TD literature is minor and Bruun [5] find similarly that only 6% is actively TD. ID in our sample is lower than Bruun [5], which could demonstrate a move towards TD research [6,8]. Alternatively, it could show a pragmatic move by researchers to focus on methods called for by funders in the Grand challenges [4]. It is also possible that the definitions of TD in literature are tightening up and becoming more specific. This variation in definition could skew the temporal distributions of labels that are assigned by researchers and publishers over logitudinal periods of time and hence make interpretation difficult.

It is notable that many of the samples contained multiple references to alternative disciplinarities, such as references in the ID sample to TD or MD. This is demonstrated in Tables 2 and 3, where TD texts make reference to ID. Although not shown in the reduced results, this is also evident in the ID sample, where it also references MD and TD. This evidences the cross over in the use of terms and the potential confusion for researchers about the type of work they may be conducting [7]. It does, however, mean that clarity in definition to use such approaches are much needed for scientists and publishers. The absence of overlap between MD and TD exists in both Figure 4 and 5. This suggests the boundary between them is distinct and hence very well defined in literature. The same absence of overlap in keywords indicates that researchers are clear in their own minds of the difference in these approaches as they are self-selected fields. Hence research should focus upon the definitions of ID and TD approaches where the overlap is prevalent.

The prevalence of '*team/teams'* in MD and ID in Tables 4 and 5, could represent an insular project with rigid bounds focussing on only clearly defined teams. The absence of the term in TD text could characterise the nature of the teams and boundaries being far less distinct and adding to complexity. This in practice may make TD research much less objective and the expected outcomes hard to define.

The prevalence of the term "*research*", whilst perhaps expected in research literature resources, demonstrates a need for higher levels of disciplinarity such as TD, especially in the case of the Grand challenges and societal problems [4]. There is a need to widen the research scope to study those industries, customers and societies that are involved in

projects, and in the same way better the research into the role of wider partners and consumers potentially contributing to better outcomes across the entire design and manufacturing chain.

The term *"social science/s"* appears as fourth most important and the term "climate change" as the sixth most frequent in the abstracts of TD literature suggestive of significance and does not appear in the frequent text for ID and MD. If the Venn illustrations were redrawn using the top 40+ and 30+ most frequent terms (respectively) the illustration would still represent these findings. This indicates a high measure of frequency distance referred to in the results, indicating these are necessary features of TD classifications. This fits with descriptions from the OECD [2] of a TD system and is certainly a statistically significant finding for the literature samples we have chosen. These frequency distance measures could be different for alternative samples for example industrial literature and should be further explored.

Additional characteristic text terms appear in the tables for each of the disciplinary categories and appear intrinsic to each. These could represent groups of features that can be used to distinguish between approaches identifying shared attributes. Awareness of context dependent terms such as *"health"* that are also prevalent may then direct the approach suitable for certain types of project.

## 4. Conclusions

Difficulties in using the correct terminology to describe research approaches is a problem not yet solved. This paper has sought to establish experimentally the text terminology used to describe differing disciplinarities. The disciplinarities analysed included MD, ID and TD, using literature samples from the Scopus database and using an automatic text content and frequency analysis approach. The combined automatic and manual text frequency analysis created lists of text concepts to describe each sample (see Figure 4 and 5). Using this approach, it has been possible to identify sets of text terms that characterise the core features for each of the disciplinarities.

Findings also showed that there are terms used commonly across the disciplinarities, such as *"research"* and context dependent terms such as *"health"*. These terms may be expected in academic literature such as *"research"* or "*health"* and do indicate the focus or context for the projects described in the literature. However, it does indicate that for wider reaching disciplinarities such as TD, we need to explore text from industry or society to fully represent the extent of these approaches.

The features we have identified through our analysis can be utilised by researchers to categorise more accurately their own disciplinary approaches or to identify the most suitable disciplinary approach needed for their research projects. Our findings provide an empirical evidence-based approach to characterising disciplinarities and with further analysis we expect these features to emerge more robustly. This includes extension of the work to create lists of statistically correlated sets of core features and to find text samples that can be analysed to represent the wider communities and their participation in disciplinary approaches. For example, the current push within engineering research for TD approaches is to solve complex problems, using the results from our text frequency analysis it is possible to suggest that to apply a TD approach it would be necessary to involve the social sciences.

The findings presented within this paper present the exploratory results from analysing the academic research literature and the natural progression would to be to

increase the literature sampling in a longitudinal study. Hence, in the absence of long study periods, the next practical steps in the process is to extend analysis to full text analysis or to widen the sampling beyond academic literature to industry literature. Over the next 18 months we will be applying a clustering method to identify and define themes within the text concepts and create associations to other literature related to disciplinarity. Through this evolving analysis, the communities understanding and definitions for each type of disciplinarity will be enhanced with the aim of enabling a scientific consensus to be formed on where to apply MD, ID or TD approaches to solve engineering design problems. Once these clear definitions exist we can move on to be able to create appropriate assessments and methods for measuring the benefits of such disciplinary approaches.

## Acknowledgement

The work reported in this paper was undertaken as part of the Designing the Future: Resilient Trans-Disciplinary Design Engineers Project, at the Universities of Bath, Bristol and Surrey. The project is funded by the Engineering and Physical Sciences Research Council (EPSRC) Grant EP/R013179/1.

## References

[1] S, Areekkuzhiyil, Emergence of New Disciplines, *Edutracks*, Vol: 17, 2017. No:4. pp 20-22.
[2] CERI, *Interdisciplinarity: Problems of teaching and research in Universities*, 1972.
[3] K. Huutoniemi, J. Thompson Klein, H, Bruun and J. Janne, Analyzing interdisciplinarity: Typology and indicators, *Research Policy,* Vol. 39, 2010. pp. 79-88.
[4] Royal Academy Of Engineering (RAEng), 2019, *Global Grand Challenges Summit*, Accessed: 19022020. [Online]. Available: https://www.raeng.org.uk/policy/partnerships/international-policy-and-development/ggcs/2019/.
[5] H. Bruun, J.I. Hukkinen, K.I. Huutoniemi, J.Thompson Klein, *Promoting Interdisciplinary research. The Case of the Academy of Finland 2005,* The Academy of Finland, Helsinki, 2005.
[6] S. Lattanzio, E. Carey, A. Hultin, R. Imani Asrai, M. McManus, N. Mogles, G. Parry and L.B. Newnes, Transdisciplinarity Within the Academic Engineering Literature*, International Journal of Agile Systems and Management*, 2020 (awaiting publication).
[7] C. Pohl, C and G.H. Hadorn, Principles for Desigining Transdisciplinary Research, Swiss Academies of Arts and Sciences, Bern, 2007.
[8] N. Wognum, C. Bil, F. Elgh, M. Peruzzini, J. Stjepandic and W.J.C. Verhagen, Transdisciplinary systems engineering: implications, challenges and research agenda, *International Journal of Agile Systems and Management*, Vol. 12, 2019, pp. 58-89.
[9] A, Aghaei Chadegani, H. Salehi, M. Yunus, H. Farhadi, M. Fooladi, M. Farhadi and N. Ale Ebrahim , A Comparison between Two Main Academic Literature Collections: Web of Science and Scopus Databases, *IDEAS Working Paper Series from RePEc,* 2013.Royal Academy Of Engineering (RAEng), 2019, *Global Grand Challenges Summit*, Accessed: 19022020. [Online]. Available: https://www.raeng.org.uk/policy/partnerships/international-policy-and-development/ggcs/2019/.
[10] Y, Ding, G, Gobinda, G, Chowdhury, F, Schubert, Bibliometric cartography of information retrieval research by using co-word analysis, *Journal for Information Processing and Management*, Vol. 37, 2001. pp. 817-842.
[11] Y, Liu, J, Goncalves, D, Ferreira, B, Xiao, S, Hosio, and V, Kostakos, Chi 1994-2013: Mapping two decades of intellectual progress through co-word analysis. *In Proceedings of the 32nd annual ACM conference on human factors in computing systems*, 2014. pp. 3553–3562.