Jeffrey Demaine
Independent researcher
Ottawa, Ontario, Canada

MAPPING SDGS BY FACULTY TO FIND NEW INTERDISCIPLINARY COLLABORATIONS, A TYPE OF LINKED LITERATURE ANALYSIS

Abstract
Partnerships are essential to achieving the United Nations' Sustainable Development Goals. In academia, interdisciplinary research can help to address complex challenges related to the Goals. This paper offers a structured approach to identifying current and potential research collaborations across faculties at a Canadian university. Publications from the Dimensions database with an SDG categorization were matched against publications indexed by the university's Research Information Management System. Potential interdisciplinary research collaborations are then identified by matching authors from different faculties who both have publications within the same research category. Intriguingly, this technique for linking potential collaborators via a shared research category is similar to the hypothesis-discovery model first proposed by Swanson in the 1980s for use in the biomedical field. The utility of this technique for inferring new relationships suggests that it is an archetypal pattern in information science which has applicability in other contexts. Indeed, interest in these techniques is growing as Large Language Models allow causal relationships to be extracted from a broader range of fields.

Overview

Advancing the United Nations Sustainable Development Goals (UN SDGs) is a focus of many research institutions of higher education. An important theme in the discussion around how universities can work towards these goals is the need for an interdisciplinary approach, as the complex solutions to the challenges of sustainable development require a combination of expertise found in different departments and faculties.

In order to better manage this shift towards more inter-faculty collaborations, institutions need to be able to track these initiatives. However, few resources exist to enable a university's administration to do this, with university rankings offering only comparisons at a global level. For example, the Times Higher Education (THE) consultancy has highlighted the role of SDGs in academia by providing 'Impact Rankings', which, starting in 2019, rank institutions by their research output as defined by the SDGs (Times Higher Education, 2022). This provides some context in which universities can benchmark their efforts against other institutions. Across all Impact Rankings criteria, McMaster rated 93.1, and is ranked ninth in Canada and 37th in the world in 2022. Part of THE's methodology includes evaluating universities' research output that is linked to SDGs. While useful to assess and promote an institution's SDG-related research publications overall, the Impact Rankings cannot provide details at the faculty or departmental level. A more granular view of an institution's research at the level of faculties and departments would provide the university's leadership with insights into the strengths and weaknesses within the institution in

terms of each SDG. This would allow the institution to be more strategic when planning future research, including fostering collaborations between faculties.

A better approach to obtaining an overview of SDG-related research is to use bibliometric metadata as indexed in several databases of academic publications (such as Web of Science, Scopus, and Dimensions). However, information about the organizational structure of the university is beyond the scope of a database of publications. This is instead the purpose of a university's *Research Information Management System* (RIMS), which provides a public-facing profile of each faculty member's research, organized by departmental affiliation. By matching the organizational affiliations of faculty members with the bibliographic metadata of their publications, a detailed picture of the strengths and weaknesses of the university in terms of SDGs can be obtained at the level of faculties and departments. This approach makes it possible to identify publications with authors from different faculties in order to highlight the interdisciplinarity of SDG research.

Methods & Results

A search of the Dimensions database for publications with at least one co-author from McMaster University from January 1st, 2018 to December 31st, 2023 (the 'study period') returned a total of 32,605 publications of the type *Research Article*, *Review Article*, or *Conference Paper*. Of these, 25,406 records could be matched by DOI to an individual in the McMaster Experts Research Information Management System (RIMS). Of these, 8,594 had also been assigned to an SDG category by Dimensions. Detailed steps necessary to replicate these results, including all data and queries, are provided in the Supplementary Material found in a Figshare repository (Demaine *et al*., 2024; DOI: 10.6084/m9.figshare.25075727).

These records were used in three analyses that provide further insight into how McMaster's research aligns with SDGs:

- *SDG publications by faculty*
  To identify which faculty or faculties are associated with each publication retrieved from Dimensions, records were augmented with their authors' faculty affiliation by matching the DOIs of the publication metadata against the publication records held in the university's *Research Information Management System* (RIMS).

- *Identification of potential collaborations*
  In addition to categorizing publications by SDGs, Dimensions also lists publications according to other ontologies. To further refine the results, Australian and New Zealand Standard Research Classification (ANZSRC 2020) "Fields of Research" categories were included as part of the metadata for each record (Australian Bureau of Statistics, 2020). The ANZSRC is a hierarchy of 1,754 specific research Fields organized into 190 broader Groups that are in turn collected into 23 general research Divisions. Dimensions assigns two levels of ANZSRC codes to publications: two-digit Divisions (e.g. "34 Chemical Sciences"), four-digit Groups (e.g. "3402 Inorganic Chemistry"). Potential co-authors are identified by selecting publications with matching SDG and ANZSRC Field of Research classifications, but where the authors' faculty affiliations are different.

- *Expert input in selecting collaborators*

The manipulation of structured information should be used to inform, but not replace, expert input. The final step of this process is to provide the managers and administrators who understand the needs of the institution with a range of choices in an accessible format. By combining the ability to process large quantities of data with the contextual knowledge of humans can insights be drawn.

SDG publications by faculty

Because a publication may be about multiple SDGs as well as being authored by researchers from different faculties, the 8,594 publications map to 9,345 categorizations-by-affiliations, allowing comparisons of faculties by their output in each SDG. For brevity this data is not shown, but the main points are highlighted:

- The distribution of publications by faculty is highly skewed, with the Faculty of Health Sciences producing 6,588-publications (70.5% of McMaster's total output).
- Of these, 88.7% address SDG *#3 - Good Health and Well Being*.
- A distant second place is the Faculty of Engineering with 1,179 publications (12.6% of total output), most of which (725; 61.5%) fall into Goal *#7- Affordable and Clean Energy*.

To tease out the patterns of collaboration within the 9,345 categorizations and affiliations, the publications were split into those involving only one faculty, and those resulting from inter-faculty collaborations. Again, because some publications correspond to more than one SDG, the 8,594 publications linked with a faculty member in McMaster Experts were assigned to 8,919 SDG categories by Dimensions (that is; 325 publications were assigned to two or more SDG categories). Of these, 8,432 (94.5%) were produced by a single faculty. Only 487 (5.5%) were the result of a collaboration between two or three faculties (see Table 1). The three areas of research with the most inter-faculty collaborations are:

- *SDG 3: Good Health and Well Being*: Of the 6,420 publications related to SDG 3, 6,064 (94.5%) were produced by a single faculty. Only 356 (5.5%) involved researchers from two or three faculties. The greatest number of these collaborations brought together researchers from the faculties of Science and of Health Sciences.

- *SDG 4: Quality Education*: Among all 540 publications related to SDG 4, a total of 503 (93.1%) were written by researchers from within the same faculty, and 37 publications (6.8%) involved researchers from two or three faculties. These were sprinkled across a range of collaborators, mostly involving the Faculty of Health Sciences.

- *SDG 7: Affordable and Clean Energy*: Among all 815 publications related to SDG 7, a total of 787 (96.5%) were the result of single-faculty research, with only 28 (3.4%) bringing together researchers from two or three faculties. Interestingly, all but two involved the faculties of Engineering and Science.

SDG	*Matched in RIMS*	*Single Faculty*			*Two+ Faculty*		
		Pubs.	*Cites*	*FCR*	*Pubs.*	*Cites*	*FCR*
1 No Poverty	10	6	5	4.5	4	10	4.8
2 Zero Hunger	103	97	25.2	11.9	6	2.2	0.4

SDG	Matched in RIMS	Single Faculty			Two+ Faculty		
		Pubs.	Cites	FCR	Pubs.	Cites	FCR
3 Good Health and Well Being	6,420	6,064	23.2	12.3	356	14.3	6.9
4 Quality Education	540	503	11.7	6.8	37	5.2	4.3
5 Gender Equality	106	99	18	15.2	7	7.9	3.3
6 Clean Water and Sanitation	66	61	13.3	5.1	5	2.8	0.7
7 Affordable and Clean Energy	815	787	19.3	6.8	28	21.6	5.5
8 Decent Work and Economic Growth	63	59	9.7	5.4	4	4.5	5.7
9 Industry, Innovation and Infrastructure	52	50	31.8	20.8	2	12	9.6
10 Reduced Inequalities	43	41	14.6	10.2	2	5	4.8
11 Sustainable Cities and Communities	66	56	10.9	6.8	10	11	6.8
12 Responsible Consumption and Production	29	27	29.4	3.7	2	2	0.3
13 Climate Action	238	223	24.4	6.4	15	30.3	5.8
14 Life Below Water	55	55	15	6			
15 Life on Land	123	119	26.7	8	4	9.3	8.7
16 Peace, Justice and Strong Institutions	178	175	7.7	5.5	3	2.3	
17 Partnerships for the Goals	12	10	19.7	10.9	2	15.5	8.6
Total	**8,919**	**8,432**			**487**		
Average			*18*	*8.6*		*9.7*	*5.1*

Table 1. Comparing the number of publications, average times cited ("Cites"), and Field Citation Ratio ("FCR") by SDG, for publications from one and two (or more) faculties. Note that publications can be assigned to more than one faculty.

Identification of potential collaborations

Given this set of publications that have been matched to SDGs and ANZSRC Field of Research, how might the university identify potential research collaborations that cross disciplinary boundaries? Using this dataset in an SQL query to pair faculty members, the records of past publications can be re-purposed as a tool for planning new research. Potential co-authors are identified where both the SDG and ANZSRC Field of Research match, but where the authors' faculty affiliations are different. To ensure that only new collaborations are identified, the pairs of authors who have appeared in previous publications are excluded from the matching query. Because such a large proportion of McMaster's publications are categorized as SDG 3, for practical purposes the matching was limited to the 16 other SDGs. The result is 8,571 pairs of authors, along with their respective departmental and faculty affiliations, who have not already collaborated, but who have the potential to do so (see Figure 1).

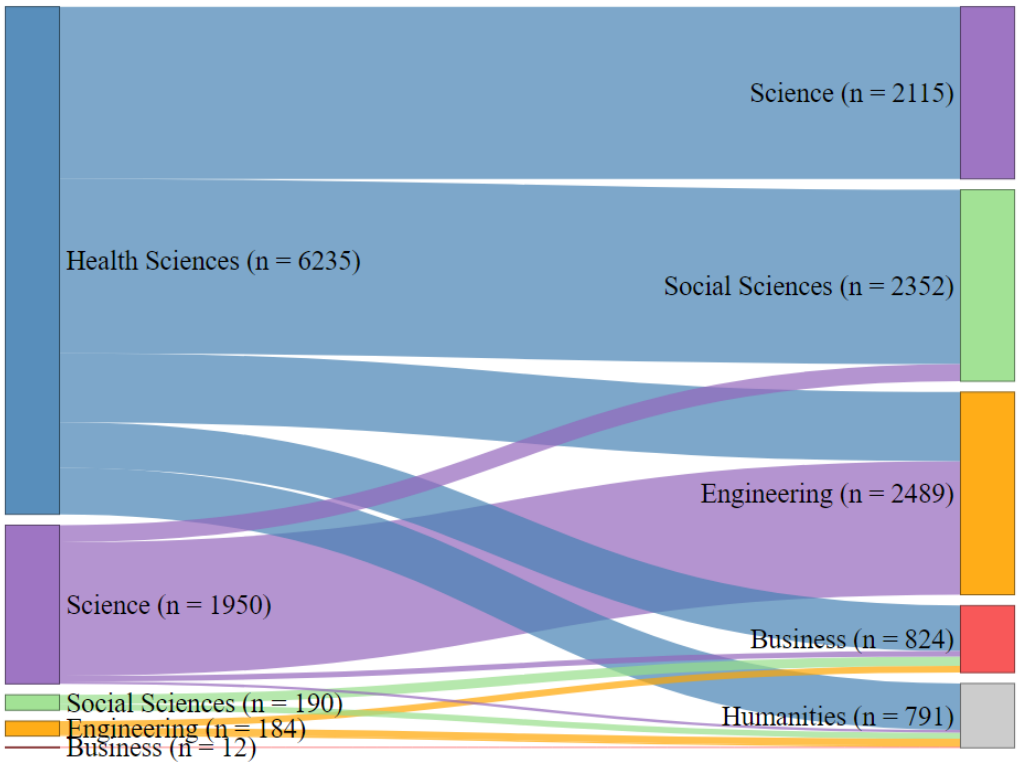

Figure 1. Potential collaborations between faculties by matching authors by SDG (less SDG 3) and ANZSRC field.

Expert input in selecting collaborators

The following example illustrates how the resulting dataset of paired authors can be used. Consider a case in which the Research Office has identified a new grant opportunity for interdisciplinary research on the topic of "sustainable transportation". How might the university identify faculty members around which a grant application could be based?

The list of 17,142 distinct faculty members (i.e. double the number of 8,571 pairs of authors) can be filtered by criteria relevant to the hypothetical grant opportunity. In this case, the SDGs 7 ("Affordable and Clean Energy") and 11 ("Sustainable Cities and Communities") were selected. By themselves, these SDGs are far too broad to identify the topic of sustainable transportation. But a manageable set of results can be arrived at by leveraging the ANZSRC fields of research, which offer a more granular classification of publications. Three relevant fields were chosen:

- 3304 Urban and Regional Planning
- 3509 Transportation, Logistics and Supply Chains
- 4011 Environmental Engineering

Finally, as the goal is to identify interdisciplinary collaborations, two different faculties are selected. For this example, the faculties of *Engineering* and *Science* are likely to be the most relevant.

The resulting list of matches reveals 19 people from the Faculty of Engineering and 9 from the Faculty of Science whose research would seem to be aligned. At this stage, the manipulation of metadata and sorting of spreadsheets reaches its limits. From this point, the expertise and judgement of managers much be relied on to read the published research and to infer relationships. Rather than presenting management with a spreadsheet of names, the relationships between potential co-authors can be visualized as a network. In Figure 2, those associated with the Faculty of Engineering are represented by orange nodes, and those in Science by purple nodes:

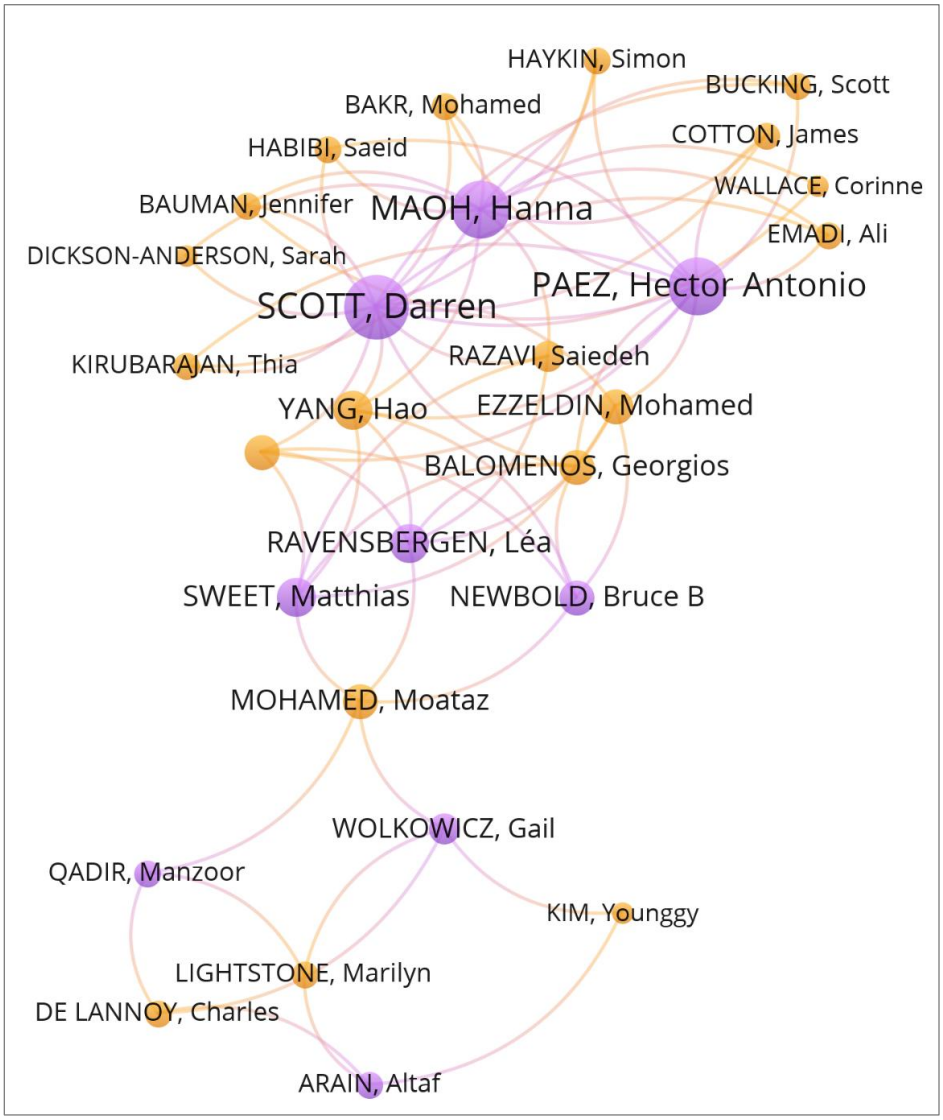

Figure 2. Potential new co-authors on the topic of "sustainable transportation" from the faculties of Engineering (orange) and Science (purple).

The similarities illustrated in Figure 2 suggest which collaborations one might explore. At the top of the network, many people from the Faculty of Engineering cluster around *Darren Scott, Hector Antonio Paez*, and *Hanna Maoh*. Consulting the McMaster Experts RIMS system, a sample of *Hanna Maoh*'s previous publications illustrates how his research aligns well with the hypothetical grant opportunity:

> "Battery electric vehicle acquisition timeframes in Canadian fleets" (*Transportation Planning and Technology*)
>
> "Examining the Variability of Crossing Times for Canadian Trucks at the Three Major Canada–U.S. Border Crossings" (*Professional Geographer*)

In order to make this an inter-faculty collaboration, a researcher from the Faculty of Engineering is selected for comparison. *Saiedeh Razavi* is located nearby *Maoh* in the network, and it seems clear that their research is indeed similar:

> "Adoption patterns of autonomous technologies in Logistics: evidence for Niagara Region." (*Transportation Letters*)
>
> "Transportation data visualization with a focus on freight: a literature review" (*Transportation Planning and Technology*)

Thus, by combining techniques to manipulate the metadata with the interpretation of textual meaning, a process is arrived at that successfully identified two researchers who, despite not having co-authored together seem to be publishing on a similar topic. This alignment suggests that they would be ideal collaborators on which a grant proposal could be based.

This example illustrates how the leadership of the university can, by selecting a few criteria and then performing a quick scan of the paired authors' publications, arrive at new insights into the untapped potential collaborations that exist across campus. Once the metadata has been compiled, no particular technical skills are required to identify matching authors based on common research interests. It is straightforward to create a dashboard that would provide a user-friendly interface, allowing managers to filter the data with a few clicks.

Parallels with Linked-Literature Analysis

The analytical technique described above for identifying potential collaborators from different departments who are doing research that is aligned according to research classification (i.e. SDG and ANZSRC) can be characterized as a Venn diagram in which Professor A is linked to Professor C via a common (intermediary) research category B. This is significant because this same pattern is at the core of an intriguing information science technique first proposed in the 1980s. The fact that this same technique can be applied in the context of managing research suggests that it could be applied in a broader range of analyses as well.

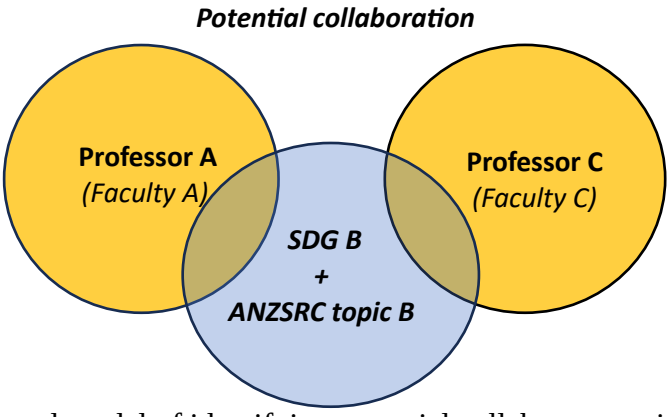

Figure 3. Conceptual model of identifying potential collaborators with distinct affiliations as joined by a shared topic. Note the A-B-C linkage.

In 1986, Don Swanson postulated that Fish Oil could be used to treat Raynaud's Disease because they both co-occurred in PubMed with the concept of Blood Circulation (Swanson, 1986). At that time, PubMed contained no records that listed both Fish Oil and Raynaud's Disease as MeSH terms simultaneously. Could there be a causal relationship between these concepts because they are linked via a common third term? Lab research subsequently confirmed Swanson's hypothesis that Raynaud's Disease could indeed be treated with Fish Oil. Several other hypotheses were generated in this way, with subsequent studies confirming a causal biomedical effect (Swanson, 1988; Swanson, 1990). The success of Swanson's "Linked Literature Analysis" technique suggested that latent discoveries exist in the scientific literature and launched the field of "Undiscovered Public Knowledge" (UPK). A subsequent experiment by Demaine, Martin, and De Bruijn (2003) found that by limiting PubMed searches in the form "(A ∩ C) ∪ (C ∩ B)" to articles published before a given year could identify 8% more "A ∪ B" in subsequent years. These are the hypothesized links between concepts that turn out to be true. This suggests that discovering likely hypotheses can be automated by using the LLA technique in brute-force searches of the scientific literature.

While Swanson's LLA technique was intended to facilitate biomedical discoveries, the method described in this article is intended to facilitate the management of academic research. As a variation of the original LLA technique, potential collaborators from different faculties whose research occurs is the same SDG are identified. The applicability of this technique in a bibliometric context suggests that the methods of UPK are generalizable to areas beyond bioinformatics.

Similar patterns can be applied to Scholarly Communication

LLA is just one of the ways in which metadata can be repurposed to uncover latent connections in the research literature. The broader field of UPK offers other models for inferring relationships from metadata. If the identification of potential collaborators across faculties is an example of how one technique from the field of UPK can be re-purposed to help universities manage research, it seems likely that other models for extracting meaning from metadata could be applied in a similar way. Smalheiser (2017) points out several other techniques for inferring new knowledge from publicly-accessible bibliographic metadata:

- One-node A-B-C (Swanson's original Linked-Literature Analysis)
- Two-node A-B-C (Smalheiser's ARROWSMITH tool: https://arrowsmith.psyvh.uic.edu)

- Multi-step paths (Baek et al., 2017; Hossain et al., 2012; Sebastian, Siew, & Orimaye, 2017)
- Ranking of shared/implicit relationships (Wren et al. 2004)

Borrowing these approaches from the field of UPK brings a whole new type of methodology to the analysis of scholarly communication. Instead of using publication metadata to measure institutional performance in a retrospective sense, patterns within the same metadata can be leveraged to produce more forward-looking insights. Promising collaborations and new directions for research can be extracted from the academic literature, enabling the management of a university to plan ahead based on uncovering connections that were previously hidden.

Templates for interacting with Large Language Models

The examples described here rely on structured metadata in order to permit the research to be grouped by distinct categories. Then the techniques of UPK can be applied to predict potential connections between the literature groupings. For these relationships to have any value, the metadata must be sufficiently curated to enable the semantic meaning to be uncovered. For this reason, Swanson's LLA technique relied on PubMed MeSH terms (Medical Subject Headings), a highly structured ontology of pre-defined keywords that are assigned to the PubMed record of each article. Subsequent research in the broader field of UPK has typically (but not exclusively) been contained to the biomedical field because the success of the techniques can be understood through their causality. It is easy to see how the success of a "cure X is related to disease Y via the effect Z" model of knowledge discovery can be tested.

Until recently, interest in LLA and UPK was waning as the researchers who were active in this field in the 1990s and early 2000s wound down their careers. Even in its heyday, Swanson's technique remained a curiosity within information science and never really caught on with the medical research community (Spasser, 1997). However, in the past few years the advent of interactive artificial intelligence tools based on large language models (LLMs) has opened the door to the popularization of advanced hypothesis-generation techniques. Crucially, the LLMs function by inferring semantic relationships from raw text. Where once LLA was restricted to the biomedical field because of its reliance on the cataloguing of articles by MeSH terms, LLMs calculate the meaning between concepts in any context, and from these causal linkages can be generated. This will allow UPK techniques to be used with any set of documents.

Just as revolutionary as the technical underpinnings of AI language models, their ease of use allows their use by non-specialists. Whereas earlier research in UPK was confined to those with advanced coding skills as well as an interest in information science, users can now interact with powerful new AI tools simply by issuing a series of questions and prompts. If linked-literature techniques such as those invented by Swanson and Smalheiser can be formulated as commands and questions, users might be able to generate new insights in any field and make discoveries without the need for complex programming. Indeed, recent publications in information science suggest that AI has arrived just in time to prevent the concept of UPK from fading into history. A team from China recently developed DiscipLink, an LLM-powered tool that helps researchers identify potentially relevant topics in other (interdisciplinary) fields (Zheng et al. 2024). And a large international team has just released a survey of the use of LLMs for hypothesis generation (Alkan et al. 2025).

Conclusion

We have seen how – by enhancing the metadata of publications with author affiliations at the sub-institutional level – not only can each faculty's output be categorized by SDG, but that same metadata can be repurposed to identify potential new collaborations between researchers in different faculties whose work falls into the same SDG and ANZSRC categories. This linking of distinct groups through a common third category is a variation of the Linked-Literature Analysis technique of generating hypotheses from PubMed metadata. This suggests that LLA can be applied to other administrative and bibliometric tasks. Moreover, the ease of use of generative AI tools that are specific to academic literature enables non-specialists to implement various techniques of Undiscovered Public Knowledge for these purposes.

Acknowledgments

The author would like to acknowledge and thank Kate Whalen Ph. D, and Yash Bhatia, who collaborated on the original publication upon which this work is based.

References

Alkan, A. K., Sourav, S., Jablonska, M., Astarita, S., Chakrabarty, R., Garuda, N., Khetarpal, P., et al. (2025). A survey on hypothesis generation for scientific discovery in the era of Large Language Models. *arXiv*. https://doi.org/10.48550/arXiv.2504.05496.

Australian Bureau of Statistics. (2020). *Australian and New Zealand Standard Research Classification (ANZSRC).* Retrieved from https://www.abs.gov.au/statistics/classifications/australian-and-new-zealand-standard-research-classification-anzsrc/latest-release

Baek, S. H., Lee, D., Kim, M., Lee, J, H., & Song, M. (2017). Enriching plausible new hypothesis generation in PubMed. *PloS One, 12*(7): e0180539. https://doi.org/10.1371/journal.pone.0180539.

Demaine, J., Martin, J., & De Bruijn, B. (2003). Haystacks and hypotheses. *Proceedings of the American Society for Information Science and Technology*, *40*(1), 59–64. https://doi.org/10.1002/meet.1450400107

Demaine, J., Bhatia, Y., & Whalen, K. (2024). Mapping publications by sustainable development goal at the faculty level to highlight inter-faculty collaborations. *International Journal of Sustainability in Higher Education.* https://doi.org/10.1108/IJSHE-01-2024-0058

Hossain, M. S., Gresock, J., Edmonds, Y., Helm, R., Potts, M., & Ramakrishnan, N. (2012). Connecting the dots between PubMed abstracts. *PLOS ONE, 7*(1): e29509. https://doi.org/10.1371/journal.pone.0029509

Sebastian, Y., Siew, E., & Orimaye, S. O. (2017). Learning the heterogeneous bibliographic information network for literature-based discovery. *Knowledge-Based Systems*, *115,* 66–79. https://doi.org/10.1016/j.knosys.2016.10.015

Smalheiser, N. R. (2017). Rediscovering Don Swanson: The past, present and future of literature-based discovery. *Journal of Data and Information Science, 2*(4), 43–64. https://doi.org/10.1515/jdis-2017-0019

Spasser, M. A. (1997). The enacted fate of undiscovered public knowledge. *Journal of the American Society for Information Science, 48*(8), 707–17. https://doi.org/10.1002/(SICI)1097-4571(199708)48:8<707::AID-ASI4>3.0.CO;2-W

Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, *30*(1), 7–18. https://doi.org/10.1353/pbm.1986.0087

Swanson, D. R. (1988). Migraine and magnesium: Eleven neglected connections. *Perspectives in Biology and Medicine, 31*(4), 526–557. https://doi.org/10.1353/pbm.1988.0009

Swanson, D. R. (1990). Somatomedin C and arginine; Implicit connections between mutually-isolated literatures. *Perspectives in Biology and Medicine, 33*(2), 157-186. https://doi.org/10.1353/pbm.1990.0031

Times Higher Education. (2022, June 8). *About the Times Higher Education World University Rankings*. Retrieved from Times Higher Education World University Rankings.

United Nations. (2015). *Transforming Our World: The 2030 Agenda for Sustainable Development.*

Wren, J. D., Bekeredjian, R., Stewart, J. A., Shohet, R. V., & Garner, H. R. (2004). Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics*, *20*(3), 389–98. https://doi.org/10.1093/bioinformatics/btg421

Zheng, C., Zhang, Y., Huang, Z., Shi, C., Xu, M., & Ma, X. (2024). DiscipLink: Unfolding interdisciplinary information seeking process via human-AI co-exploration. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, 1–20. UIST '24. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3654777.3676366.