KU LEUVEN

FACULTEIT ECONOMIE EN BEDRIJFSWETENSCHAPPEN

Management, Strategy and Innovation (MSI)

Improving the precision of subject assignment for disparity measurement in studies of interdisciplinary research

Wolfgang Glänzel, Bart Thijs and Ying Huang

MSI_2104

Improving the precision of subject assignment for disparity measurement in studies of interdisciplinary research

Wolfgang Glänzel, Bart Thijs, Ying Huang

Wolfgang.Glanzel@kuleuven.be
KU Leuven, ECOOM & Dept MSI, Leuven Belgium

Abstract

Studies of interdisciplinarity research (IDR) poses severe challenges to bibliometricians. These challenges range from conceptualisation of IDR, over the definition of disciplines and the way of how research can be assigned to those, to finding the particular methods for quantifying and measuring the peculiarities of IDR. One of the key issues is the determination of granularity. This issue is twofold: The conceptional consideration should clarify the question the level at which IDR would be studied, namely as topic, subject or field interdisciplinarity, and this needs to be supported by quantitative results. The second key issue concerns the way of the assignment of the subjects that are integrated in the research, once the granularity level has been chosen.

These two questions are tackled in the present paper, which is closely linked to further studies by the authors on the effect of similarity measurement approaches on indicators (Huang et al., 2021) and on different implementations of similarity for the measurement of disparity and variety (Thijs et al., 2021). The present study proposes a multiple-generation reference model and gives solutions for individual-document based subject assignment and the calculation of cognitive distances between disciplines needed for the determination of disparity measures.

Introduction

As science increasingly deals with boundary-spanning problems, important research ideas often transcend the scope of a single discipline or program. Thus, building sustainable bridges between two or more disciplines is valuable for pushing academic capability forward and for accelerating scientific discovery. Despite the growing attention interdisciplinary research (IDR) has received, there is a lack of consensus in the literature as to the definition of "interdisciplinary" (Huutoniemi et al., 2010). The definition of a "discipline" and discussions of the varieties of interdisciplinary, multidisciplinary, and transdisciplinary research have occupied much scholarly debate (NSF, 2004).

As a multi-faceted concept, IDR can mean different things to different people. One of the most broadly-accepted definitions of IDR is set forth in a National Academies' report (COSEPUP, 2004):

> *"a mode of research by teams or individuals that integrates information, data, techniques, tools, perspectives, concepts and/or theories from two or more disciplines or bodies of specialized knowledge to advance fundamental understanding or to solve problems whose solutions are beyond the scope of a single discipline or area of research practice."*

Derived from the above definitions, knowledge integration is the essence of IDR. According to Porter et al., (2008), the knowledge to be integrated can be of various forms: ideas (such as concepts and theories), methods (techniques and tools), or data from various fields of knowledge. The research on interdisciplinarity can be approached from several different perspectives. From a conceptual viewpoint, there are two main perspectives of studying interdisciplinarity, namely the cognitive perspective on the basis of information flows across disciplines and the organisational approach on the basis of co-authors' professional skills and/or affiliation. The cognitive approach can be implemented by using two different methods

that can actually be combined into a hybrid methodology: Since abstract and citation databases provide the necessary link and textual information on citation flows and document topics, respectively, on the large scale even providing the basis of benchmarking, the cognitive approach can be considered the first option. By contrast, useful large-scale information on skills or affiliation needed for the organisational approach is only covered partially (Zhang et al., 2018). On the other hand, subject classification schemes provided with the databases or based on journal assignment and fail in the context of interdisciplinarity whenever journals have a general or even multidisciplinary scope. In order to gain reliable information on the subject of individual documents indexed in the database, a document-based improvement of the subject assignment is needed. The question arises of whether such an assignment with appropriate granularity is feasible on the large scale. We propose a parameterized rule-based model, implement different versions and study the effect of altering parameter settings. The assignment of a publication is based on the most dominant subfield(s) in the aggregated set of cited references where not only the references in the publication are taken into consideration but also those included in the cited publications.

Conceptual considerations and research design

The two central concepts related to IDR are "diversity" and "coherence" (cf. Rafols and Meyer, 2010). While Stirling (1994) distinguished has three components of diversity *variety*, *balance*, and *disparity*, Rafols (2014) proposed to subdivide the notion of coherence into the three aspects *density*, *intensity* and *disparity*. In either component, but most notably in the context of variety and disparity, the correctness and the granularity of topic identification is crucial.

The first aim of the present study is therefore twofold, on the one hand, we attempt to choose a granularity level at which the identification of IDR has still a nuts-and-bolts use and interpretation (biomedical engineering or urbanism certainly represents a different level of IDR than stochastic geometry), secondly, we try to extend existing journal-based subject classification schemes towards the individual document-level assignment. In order to achieve this objective, we will first summarise the effect of granularity on subject disparity on the basis of literature cited by articles indexed in the 2018 volume of the Web of Science Core Collection. We have applied three scientometric link-based methods to measure similarity, bibliographic coupling, co-citation and cross-citation links.

In this context we note that in the present paper we just summarise the major outcomes of these two objectives, while we have devoted two separate studies to the analysis of the granularity level in the context of IDR (Huang et al., 2021) and the effect of the choice of the particular scientometric method (coupling, co- and cross-citation) on the definition of the (dis-)similarity measures (Thijs et al., 2021). We decided to publish these two important aspects in separate papers as the exhausting investigation would require more space than available in this paper. The main focus of the present study is actually laid on the assignment of individual documents to disciplines, as the journal-based assignment used by the database providers proved too unspecific and thus, at least in the case of more general and multidisciplinary journals, impractical for our purpose. Nonetheless, the choices of both granularity level and link-analysis are indispensable steps in quantification and measurement of IDR and need to be decided upon before (interlinked) document are properly assigned to subjects. We decided therefore, not to remove these steps but to report them giving additional examples to those already given in the two other studies by the authors.

The following sections describe the experiment and the major outcomes.

Data sources and data processing

All documents indexed as articles in the three journal editions (SCIE, SSCI, A&HCI) of the 2018 volume of the Web of Science (WoS) Core Collection have been extracted from the

database. For evolutionary and robustness analyses the previous volumes covering the period 1999–2018 have been inclued. For the analysis of cited references, all items that are indexed in *any* WoS database are taken into consideration. As the cognitive subject scheme underlying this study, the adjusted Leuven-Budapest classification with 16 major and 74 sub-fields (disciplines) building upon the WoS Subject Categories (Glänzel et al., 2016) has been used. Figure 1 shows the subject structure of the Leuven-Budapest scheme at the sub-field level.

Methods and results

Determining the granularity level

The basic methodological idea was to conduct the research in several consecutive steps. The first one concerned the granularity proceeding from the assumption that the discriminative power of the level would not dramatically change when we adjust the journal-based assignment by an article-based version using the same hierarchic structure. We applied three scientometric methods based on bibliographic coupling (BC), co-citation (CO) and cross-citation (CC). Again, a systematic investigation of the appropriate methodology for link-based similarity measures has been conducted by Thijs et al. (2021). We here give an example to quantitatively underpin the choice of the granularity level, which, in turn, is necessary for the implementation of individual subject assignment. This, again, shows the interweaving of the tree aspects.

Since co-citations need an appropriate citation window, we have used the five-year window 2010–2014 in this case. Since we are proceeding from a vector-space model, a cosine measure was applied to determine the similarity of subjects. The cosine similarity is defined as the cosine of the angle between two vectors representing the respective documents, i.e., ratio of their scalar product and the product of their lengths. In the case of a Boolean vector space as, for instance, in the case of BC and CO, this reduces to Salton's measure, i.e., the ratio of the number of jointly cited/citing items and the geometric mean of the number of all cited/citing items. To cross-citations, we have used formula used by Zhang et al. (2010) in Eq. (1) and later applied in the context of IDR as well (cf. Zhang et al., 2016). Dissimilarities can readily be derived from the cosine measure by elementary arithmetic manipulation.

This step is required to select both the granularity level and the methodological basis for measuring the disparity aspect of subjects in possible IDR indicators. The correlation across granularity levels is strong (≥ 0.92 for major fields and disciplines and ≥ 0.82 for the subject categories and disciplines) but the slopes reveal systematic trends. These are also reflected by the means. Table 1 gives the *mean* and *minimum mean* similarity with other subjects according to the three approaches. BC provides in general stronger similarity than CC and CO at all three levels. Although the minima for the individual scientometric link-analyses are taken by different subjects, the deviation is not large because the values of G (Geo & space sciences) and L (Social sciences II), K6 (literature) and H2 (pure mathematics), and UT (poetry) and EO (classics), respectively are of similar order (EO = 0.018 in BC, H2 = 0.051 in CC and G = 0.216 in CO, which do, however, not appear in Table 1).

Table 1. Empirical similarity statistics at three granularity levels

	BC		**CC**		**CO**	
	Mean	*Minimum*	*Mean*	*Minimum*	*Mean*	*Minimum*
Major Field	0.364	0.248 (G)	0.326	0.180 (G)	0.314	0.208 (L)
Discipline	0.228	0.064 (H2)	0.180	0.038 (K6)	0.163	0.051 (H2)
Subject Category	0.139	0.007 (UT)	0.090	0.015 (EO)	0.071	0.006 (EO)

Data sourced from Clarivate Analytics Web of Science Core Collection

This confirms previous findings by Glänzel et al. (2009) concerning subject-normalised citation indicators according to which major fields proved too coarse and the lowest level, (subject categories) provide a fine-grained but very fuzzy subject assignment. Subfields could therefore serve as the favoured reference level for disciplines. This gives also some quantitative evidence to support the decision of not to go for kind of "topic interdisciplinarity" as sketched in the preliminary and more conceptual considerations above.

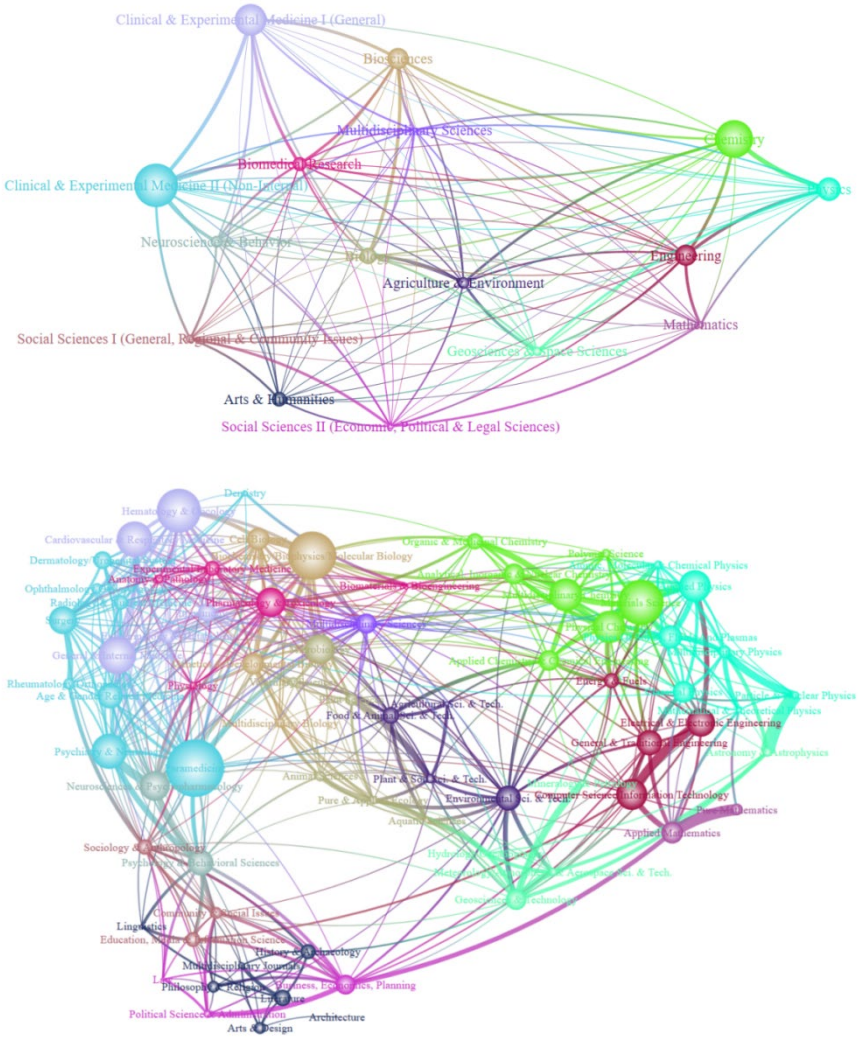

Figure 1. The cross-citation based structure of the Leuven-Budapest scheme (1999–2018) based on major fields (top) and subfields (bottom)
Data sourced from Clarivate Analytics Web of Science Core Collection

Figure 1 shows the disciplinary structure of the WoS at the broad field level (15 major fields) and subfield level (74 disciplines). The detailed scheme for these hierarchic levels is given in the Appendix.

Individual document based subject assignment

The next step towards creating the groundwork of variety and disparity measurement concerns the individual subject assignment of articles and their cited references. Variety (and balance) is based on the number and distribution of disciplines the knowledge of which is integrated in published research results, whereas disparity takes also their dissimilarity into account. While

the first aspect can be studied by analysing, e.g., the disciplines to which the cited references belong, the later aspect requires the knowledge of the disciplinary structure of the complete database (cf. Figure 1). Because of the strong correlation and robustness of all three methods, any of those are suitable for the creation of a 'global' (dis-)similarity matrix. We decided to choose BC, since this does not require any particular citation window and most documents have sufficiently long reference lists. At this point, we have to make a distinction between the reference items used for BC, the particular topics of which are not relevant for the link analysis, and those used to improve subject assignment and to detect topics of knowledge integration in IDR. This forms a straight continuation and update of the idea proposed by Glänzel & Schubert (2003) in connections with the creation of cognitive but bibliometrics-aided subject classification scheme. Since many cited documents are published in general journals such as *Physical Review Letters* in physics, *JACS* or *Angewandte Chemie* in chemistry or even multidisciplinary journals like *Science*, *Nature*, *PNAS US* or *PLoS One* with no specific subject profile, we have to detect the topic of those items by analysing their own references. Therefore, we introduce the multiple-generation reference model to classify individual scientific publications. In this step, we adopt a full-counting method for the assignment classification system and we track the relationship between the original publication and the cited references' sub-fields (1st generation), but also the cited reference's sub-fields of those cited references (2nd generation) and so on (3rd generation).

The classification model we propose is a parameterized rule-based model. This approach allows us to implement different versions and study the effect of altering parameter settings. In short, a publication is assigned to the most dominant subfield(s) in the aggregated set of cited references. Cited subfields are ranked based on the normalized share they take in the total set across multiple generations. The selection of the most dominant subfields can be based on a particular threshold or a combination of rules or judgements.

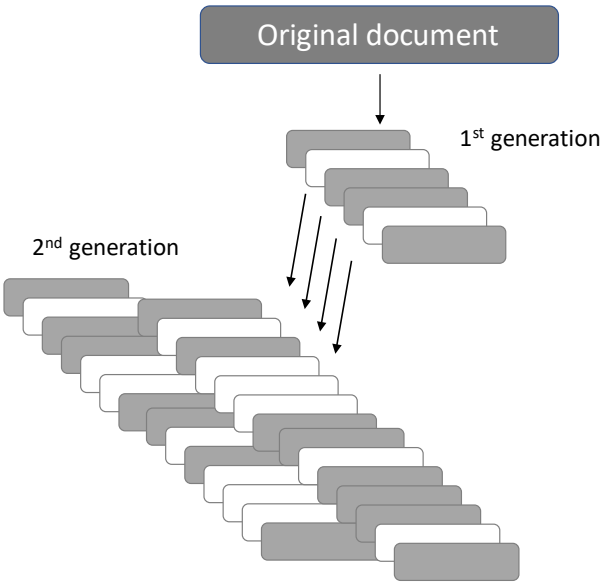

Figure 2. The multi-generation reference model supporting individual document subject-assignment based on two generations with 'active' (dark) and 'non-source' items (light).

Figure 2 illustrates the 2-generation approach. The grey-shaded references symbolise the "active" references, i.e. those that are indexed in the database, the white one stand for non-source references the assignment of which is often unclear and which are therefore ignored.

Of course, each additional generation can increase the number of disciplines contributing to the integration of knowledge but could weaken their direct influences and thus increases

fuzziness. Table 2 shows the distribution within the major fields of the similarity between the discipline profiles of the first two generations of cited references, where the share of the records within a certain similarity range in all records belonging to the corresponding field is calculated. According to the results, in some fields there is less integration of knowledge from other disciplines over reference generations than in other fields. The corresponding field codes can be found in the Appendix. The more skewed the distribution, the lower is the fuzziness and vice versa. Therefore, lower weights can be given to 'indirect' references, which are represented by higher generations and the precision of which in terms of the assignment of the original document decreases with the order of the generation.

Table 2. The distribution of discipline similarity between 1st and 2nd generation references by major fields, where red stands for strong and blue for week, both with grading ranging from pronounced (dark) to less distinctive (light)

Field	[.95,1]	[.9, .95)	[.85, .9)	[.8, .85)	[.75, .8)	[.7, .75)	[.65, .7)	[.6, .65)	.55, 1.6)	[.5, .55)	[.0, .5)
A	0.456	0.268	0.120	0.062	0.034	0.020	0.012	0.008	0.006	0.004	0.009
B	0.484	0.282	0.119	0.054	0.026	0.014	0.008	0.004	0.003	0.002	0.003
C	0.510	0.249	0.109	0.054	0.029	0.017	0.010	0.007	0.004	0.003	0.007
E	0.578	0.199	0.090	0.047	0.027	0.019	0.011	0.008	0.006	0.004	0.011
G	0.729	0.139	0.055	0.028	0.015	0.010	0.007	0.005	0.004	0.003	0.007
H	0.640	0.164	0.076	0.041	0.023	0.020	0.010	0.007	0.006	0.004	0.010
I	0.467	0.280	0.124	0.058	0.029	0.016	0.009	0.006	0.004	0.002	0.005
K	0.368	0.160	0.108	0.077	0.050	0.061	0.032	0.026	0.029	0.023	0.065
L	0.692	0.144	0.064	0.033	0.019	0.015	0.009	0.006	0.005	0.004	0.010
M	0.501	0.241	0.113	0.058	0.032	0.019	0.012	0.008	0.005	0.004	0.008
N	0.675	0.195	0.067	0.029	0.014	0.008	0.005	0.003	0.002	0.001	0.003
P	0.535	0.228	0.101	0.052	0.029	0.018	0.011	0.008	0.005	0.004	0.009
R	0.339	0.295	0.162	0.086	0.047	0.027	0.016	0.010	0.006	0.004	0.008
Y	0.444	0.230	0.117	0.067	0.040	0.030	0.019	0.014	0.011	0.008	0.022
Z	0.484	0.262	0.113	0.056	0.031	0.019	0.011	0.007	0.005	0.004	0.008

Data sourced from Clarivate Analytics Web of Science Core Collection

The general formula of normalizing the share of a research field or discipline in our multiple-generation model is as follows:

$$NFS^{(n)}{}_i = w_1 FS^{(1)}{}_i + w_2 FS^{(2)}{}_i + \ldots + w_n FS^{(n)}{}_i,$$

with

- n is the number of cited reference generations considered.
- $NFS^{(n)}{}_i$ denotes the normalized share of category i aggregated over n generation.
- $FS^{(k)}{}_i$ denotes the share of cited category i in the total number of cited categories at generation k.
- w_k refers to the normalizing factor or weight attributed to generation k. it takes a value between 0 and 1.
- $w_1 + w_2 + \ldots + w_n = 1$.

Based on the values that are assigned to the normalizing factors or weights special cases can be identified:

- If w_k=1 for any value of k where $1 \leqslant k \leqslant n$, only the k^{th} generation is taken into consideration.
- If w_k = 1/n for any value of k where $1 \leqslant k \leqslant n$, every generation is taken into consideration and treated equally.
- If $w_k = w_1/k$, for any value of k where $1 \leqslant k \leqslant n$, every generation is taken into consideration and the k^{th} generation has the $1/k$ times of share compared to the 1^{st} generation.
- If $w_k = w_1^k$, for any value of k where $1 \leqslant k \leqslant n$, each generation is taken into consideration and the factor for each additional generation is multiplied by w_1.

Next, the cited categories or fields are ranked in descending order based on their normalized shares. Finally, the highest ranked category or categories are attributed to the publication after the application of judgement rules. The normalized share for the highest ranked subfield is given by $NFS^{(n)}_{r=1}$

In this study, we consider only the first two generations and report on three weight-allocation schemes (or models) as shown in Table 3. References to *multi-disciplinary journals* in the last considered generation were ignored as being unspecific. The weights and their ratios have been determined on an empirical basis and simple arithmetic, where, for instance, M3 takes the potentiating number of references into account.

Table 3. The weight-allocation model for normalising the share of research fields

Wight model	Formula	w_1	w_2
M1	$w_1 = 1$	1.000	0.000
M2	$w_2 = 1$	0.000	1.000
M3	$w_2 = w_1^2$	0.618	0.382

For the assignment of the sub-fields to individual papers, we have implemented these selection rules:

- The individual assignment to a paper is limited to three sub-fields (i.e., disciplines) according to their frequency ranks i = 1, 2, 3.
- A publication can be uniquely assigned to discipline a only if none of the higher ranked subfields has a normalized share which is at most 0.67 times the subsequent one.
- Assignment of additional disciplines is done, if the above share is larger than 0.67 following the same algorithm.
- The procedure is to be stopped after at most three fields have been assigned. Otherwise, assignment is stopped by the procedure whenever $NFS^{(2)}_{r=i+1}/NFS^{(2)}_{r=i} \leqslant 2/3$ for any $i \leq 3$.
- Unassigned documents can still be assigned manually, but are very likely to be truly interdisciplinary themselves. These cased proved to be rather rare.

Table 4 provides a concise formalised view of the complete procedure of subfield assignment to individual papers.

We add a sample of the identified papers published in Nature to illustrate the procedure of assigning the field, shown in Table 5. Note, that the weight type here is M1, i.e., only considering the first-generation reference. As our purpose is to allocate up to three fields to the individual scientific publication, the fields labeled “Multidisciplinary Sciences (X0)” should be removed.

Table 4. Overview of the complete procedure of subfield assignment to individual papers

1st round judgement	2nd round judgement	3rd round judgement	Assignment
$NFS^{(2)}_{r=2}/NFS^{(2)}_{r=1} \leqslant 2/3$			Field 1
$2/3 < NFS^{(2)}_{r=2}/NFS^{(2)}_{r=1} \leqslant 1$	$NFS^{(2)}_{r=3}/NFS^{(2)}_{r=2} \leqslant 2/3$		Field 1 Field 2
$2/3 < NFS^{(2)}_{r=2}/NFS^{(2)}_{r=1} \leqslant 1$	$2/3 < NFS^{(2)}_{r=3}/NFS^{(2)}_{r=2} \leqslant 1$	$NFS^{(2)}_{r=4}/NFS^{(2)}_{r=3} \leqslant 2/3$	Field 1 Field 2 Field 3
$2/3 < NFS^{(2)}_{r=2}/NFS^{(2)}_{r=1} \leqslant 1$	$2/3 < NFS^{(2)}_{r=3}/NFS^{(2)}_{r=2} \leqslant 1$	$2/3 < NFS^{(2)}_{r=4}/NFS^{(2)}_{r=3} \leqslant 1$	Unassigned

Table 5. The sample to illustrate the procedure of assigning the field

ISI	Total Refs.	WoS Refs.	Field 1	Field 2	Field 3	Field 4	Assigned Field(s)
000419769300025	71	59	G2 (0.389)	X0 (0.254)	G4 (0.152)	G3 (0.085)	G2
000419769300035	37	35	Z3 (0.314)	B1 (0.229)	R3 (0.086)	B2 (0.057)	Z3; B1
000419769300037	59	58	X0 (0.268)	B2 (0.232)	B1 (0.214)	B3 (0.107)	X0; B2; B1
000419769300030	45	43	C6 (0.246)	C4 (0.174)	P1 (0.159)	X0 (0.145)	Unassigned
000419769300031	43	35	X0 (0.192)	P6 (0.154)	C4 (0.154)	C6 (0.154)	Unassigned

Data sourced from Clarivate Analytics Web of Science Core Collection

In this paper, we propose three ways to deal with the "Multidisciplinary Sciences (X0)" during the procedure. Samples are shown in Table 6.

Table 6. Comparative results of the procedure for individual subject assignment using different methods to resolve X0 assignment

Method	Weight model	Rank1		Rank2		Rank3		Rank4		Assigned Field(s)
		Share	Field	Share	Field	Share	Field	Share	Field	
Original	M1	0.268	X0	0.232	B2	0.214	B1	0.107	B3	X0; B2; B1
	M2	0.238	B2	0.219	X0	0.211	B1	0.107	I4	B2; X0; B1
	M3	0.247	B2	0.206	B1	0.139	X0	0.135	I4	B2; B1; X0
Remove X0 after original result	M1	-	-	0.232	B2	0.214	B1	0.107	B3	B2; B1
	M2	0.238	B2	-	-	0.211	B1	0.107	I4	B2; B1
	M3	0.247	B2	0.206	B1	-	-	0.135	I4	B2; B1
Remove X0 before calculating shares	M1	0.317	B2	0.293	B1	0.146	B3	0.122	I4	B2; B1
	M2	0.306	B2	0.272	B1	0.135	I4	0.125	B3	B2; B1
	M3	0.287	B2	0.239	B1	0.156	I4	0.09	B3	B2; B1
Remove X0 after calculating shares	M1	0.232	B2	0.214	B1	0.107	B3	0.089	I4	B2; B1
	M2	0.238	B2	0.211	B1	0.107	I4	0.096	B3	B2; B1
	M3	0.247	B2	0.206	B1	0.135	I4	0.077	B3	B2; B1

Data sourced from Clarivate Analytics Web of Science Core Collection

Table 7 gives an impression on the effect of the chosen formula and weight scheme on the share of possible individual assignment for four selected journals. Using the M2 and M3 alone usually resulted in lower shares, for instance, of about 0.87 for the journals *Nature*, *Science* and

PNAS, only the combination of M1, M2 and M3 resulted in significant improvement of the share. The same applied to the general journals like JACS in chemistry but even in the more specific journals with already high w_1 scores, the (M1 + M2 + M3) combination resulted in further improvement (cf. JASIST in information and library science).

Table 7. Overview of the complete procedure of subfield assignment to individual papers

	Weight model	Original	Remove X0 from original result	Remove X0 before calculating shares	Remove X0 after calculating shares
Nature (X0) (8259)	M1	0.834	0.781	0.906	0.906
	M2	0.778	0.755	0.874	0.872
	M3	0.784	0.781	0.873	0.873
	(M1+M2)	0.861	0.825	0.927	0.927
	(M1+M2+M3)	0.913	0.890	0.956	0.956
JASIST (E1,Y1) (1788)	M1	0.967	0.961	0.971	0.971
	M2	0.938	0.933	0.942	0.942
	M3	0.915	0.911	0.924	0.924
	(M1+M2)	0.975	0.970	0.977	0.977
	(M1+M2+M3)	0.987	0.984	0.988	0.988
BOI (B0,B1,E1, H1,Z3) (7296)	M1	0.779	0.768	0.805	0.805
	M2	0.740	0.735	0.782	0.780
	M3	0.787	0.784	0.867	0.867
	(M1+M2)	0.845	0.837	0.870	0.868
	(M1+M2+M3)	0.927	0.923	0.952	0.952
JACS (C0) (28141)	M1	0.889	0.885	0.902	0.902
	M2	0.861	0.860	0.875	0.875
	M3	0.839	0.839	0.848	0.848
	(M1+M2)	0.912	0.909	0.920	0.920
	(M1+M2+M3)	0.940	0.939	0.945	0.945

Data sourced from Clarivate Analytics Web of Science Core Collection

About 5% of articles published in general and multi-disciplinary journals could not be assigned individually. These papers proved highly interdisciplinary, mostly with no dominant discipline in the cited information sources so that the procedure, which was originally designed to be used in the context of IDR studies, helps directly identify interdisciplinary research quasi as by-product. Below we list four documents published in PLoS ONE and Nature just as typical examples of such research. Neither the cited references nor titles/abstracts and the authors' affiliations point to one specific discipline but immediately reveal the interdisciplinary nature of the underlying research:

- WOS:000383255900083 – Abundant topological outliers in social media data and their effect on spatial analysis (PLOS ONE)
- WOS:000391217400045 – Narrative style influences citation frequency in climate change science (PLOS ONE)
- WOS:000268938300034 – Dense packings of the Platonic and Archimedean solids (Nature)
- WOS:000376883000004 – The Kinome of pacific oyster Crassostrea Gigas, its expression during development and in response to environmental factors (PLOS ONE)

Figure 6 finally shows the subject profiles after individual subject assignment on the basis of the described iterative process of reference analysis. Above all, the three big multi-disciplinary journals reflect a large spectrum of (unequally distributed) disciplines. With the application of the models and methods described above, we have all prerequisites for the creation of the tools to measure the main aspects of IDR, most notably variety and disparity of integrated knowledge. However, this will be part of a separate study.

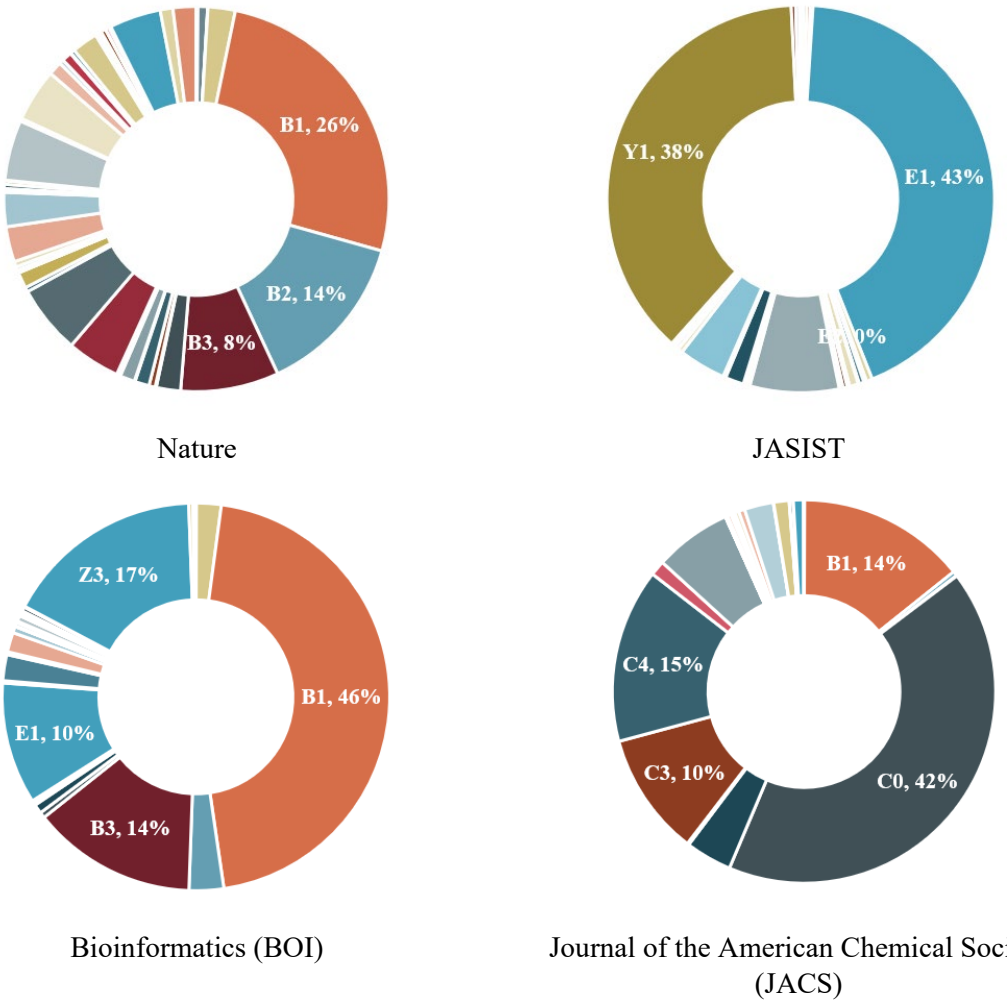

Figure 6. Visualisation of the subject profiles of several analysed multi- and interdisciplinary journals using 74 disciplines in the sciences, social sciences and humanities
Data sourced from Clarivate Analytics Web of Science Core Collection

Summary of main findings and conclusions

The experiment has provided three major results, firstly, bibliographic coupling, cross-citation and co-citations provided similar results in the analysis of subject (dis-)similarity of the WoS documents space, where BC proved a suitable and robust methodological basis at all granularity levels. Secondly, the granularity sub-field level with 74 disciplines in the sciences, socials sciences and humanities proved most suited for the large-scale analysis of IDR of individual documents and provides quantitative support to the conceptual consideration on what level the integration of knowledge may be studied. Thirdly and finally, the iterative process of weighted multi-generation reference analysis resulted in the applicability of the journals-based ECOOM classification scheme to individual-document assignment for ≥ 95%

of documents. Recalling the objective of this study, namely improving the precision of disparity measurement in studies of interdisciplinary research, the achieved level of about 95% is sufficient: On the one hand, the method helps build reliable (dis-)similarity matrices for providing the basis for developing disparity measures and essentially improves the accuracy of the subject assignment of cited references (on the basis of the 2nd reference generation) for the measurement of variety but can, on the other hand, also be used to assign the documents themselves to a number of particular subjects. The remaining documents proved to be truly interdisciplinary and require further (qualitative) investigation of knowledge integration. This could be done on the basis of cited literature standing for knowledge that has become integrated into the research in question in conjunction with text analysis using the title, abstracts and keywords or, whenever available, the full texts of the documents. The results of this study form the groundwork for important future tasks: The main future objective is, of course, the creation of measures of variety and disparity with full implementation of the methodology described in the present study. This will be achieved in a separate paper on the different implementations of similarity for disparity and variety measures (Thijs et al., 2021). A secondary task is the individual subject assignment of all papers indexed in the WoS to disciplines according to the ECOOM classification scheme with the possibility of supplementary attribution of IDR labels to documents for policy-relevant applications.

Acknowledgement

The research underlying this study is done within the framework of the project "Interdisciplinarity & Impact" (2019-2023) funded by the Flemish Government. We would like to thank Wouter Jeuris and Lin Zhang for their inspiring discussions.

References

Huutoniemi, K., Klein, J. T., Bruun, H., et al. (2010). Analyzing interdisciplinarity: Typology and indicators. *Research Policy*, 39(1), 79-88.

COSEPUP (2004). *Facilitating interdisciplinary research. National Academies*. "Committee on Facilitating Interdisciplinary Research, Committee on Science, Engineering, and Public Policy", Washington: National Academy Press, p. 2.

Glänzel, W., & Schubert, A. (2003), A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, 56(3), 357-367.

Glänzel, W., Schubert, A., Thijs, B., & Debackere, K. (2009). Subfield-specific normalized relative indicators and a new generation of relational charts: Methodological foundations illustrated on the assessment of institutional research performance. *Scientometrics*, 78(1), 165–188.

Glänzel, W., Thijs, B. & Chi, P.S. (2016). The challenges to expand bibliometric studies from periodical literature to monographic literature with a new data source: The Book Citation Index. *Scientometrics*, 109(3), 2165–2179.

Huang, Y., Glänzel, W., Thijs, B., Porter, A.L., & Zhang, L. (2021), *The comparison of various similarity measurement approaches on interdisciplinary indicators*. FEB Research Report MSI_2102, Report No. MSI_2102.

Porter, A. L., Roessner, D. J., & Heberger, A. E. (2008). How interdisciplinary is a given body of research? *Research Evaluation*, 17(4), 273-282.

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, 82(2), 263-287.

Rafols, I. (2014). *Knowledge integration and diffusion: Measures and mapping of diversity and coherence*. In: Ding Y., Rousseau R., Wolfram D. (eds) Measuring scholarly impact (pp. 169-190): Springer, Cham.

Stirling, A. (1994). Diversity and ignorance in electricity supply investment: Addressing the solution rather than the problem. *Energy Policy*, 22(3), 195-216.

Thijs, B., Huang, Y. & Glänzel, W. (2021), *Comparing different implementations of similarity for disparity and variety measures in studies on interdisciplinarity*. FEB Research Report MSI_2103, Report No. MSI_2103.

Zhang, L., Janssens, F., Liang, L.M., & Glänzel, W. (2010). Journal cross-citation analysis for validation and improvement of journal-based subject classification in bibliometric research. *Scientometrics*, 82(3), 687–706.

Zhang, L., Rousseau, R., & Glänzel, W. (2016). Diversity of references as an indicator of the interdisciplinarity of journals: taking similarity between subject fields into account. *Journal of the association for information science and technology*, 67(5), 1257-1265.

Zhang, L., Sun, B., Chinchilla-Rodríguez, Z., Chen, L., & Huang, Y. (2018). Interdisciplinarity and collaboration: on the relationship between disciplinary diversity in departmental affiliations and reference lists. *Scientometrics*, 117(1), 271-291.

Appendix

The revised Leuven-Budapest classification scheme according to Glänzel et al. (2016)

THE LEUVEN – BUDAPEST CLASSIFICATION SCHEME FOR THE SCIENCES, SOCIAL SCIENCES AND HUMANITIES

0. MULTIDISCIPLINARY SCIENCES
- X0 multidisciplinary sciences

1. AGRICULTURE & ENVIRONMENT
- A1 agricultural science & technology
- A2 plant & soil science & technology
- A3 environmental science & technology
- A4 food & animal science & technology

2. BIOLOGY (ORGANISMIC & SUPRAORGANISMIC LEVEL)
- Z1 animal sciences
- Z2 aquatic sciences
- Z3 microbiology
- Z4 plant sciences
- Z5 pure & applied ecology
- Z6 veterinary sciences

3. BIOSCIENCES (GENERAL, CELLULAR & SUBCELLULAR BIOLOGY; GENETICS)
- B0 multidisciplinary biology
- B1 biochemistry/biophysics/molecular biology
- B2 cell biology
- B3 genetics & developmental biology

4. BIOMEDICAL RESEARCH
- R1 anatomy & pathology
- R2 biomaterials & bioengineering
- R3 experimental/laboratory medicine
- R4 pharmacology & toxicology
- R5 physiology

5. CLINICAL AND EXPERIMENTAL MEDICINE I (GENERAL & INTERNAL MEDICINE)
- I1 cardiovascular & respiratory medicine
- I2 endocrinology & metabolism
- I3 general & internal medicine
- I4 hematology & oncology
- I5 immunology

6. CLINICAL AND EXPERIMENTAL MEDICINE II (NON-INTERNAL MEDICINE SPECIALTIES)
- M1 age & gender related medicine
- M2 dentistry
- M3 dermatology/urogenital system
- M4 ophthalmology/otolaryngology
- M5 paramedicine
- M6 psychiatry & neurology
- M7 radiology & nuclear medicine
- M8 rheumatology/orthopedics
- M9 surgery

7. NEUROSCIENCE & BEHAVIOR
- N1 neurosciences & psychopharmacology
- N2 psychology & behavioral sciences

8. CHEMISTRY
- C0 multidisciplinary chemistry
- C1 analytical, inorganic & nuclear chemistry
- C2 applied chemistry & chemical engineering
- C3 organic & medicinal chemistry
- C4 physical chemistry
- C5 polymer science
- C6 materials science

9. PHYSICS
- P0 multidisciplinary physics
- P1 applied physics
- P2 atomic, molecular & chemical physics
- P3 classical physics
- P4 mathematical & theoretical physics
- P5 particle & nuclear physics
- P6 physics of solids, fluids and plasmas

10. GEOSCIENCES & SPACE SCIENCES
- G1 astronomy & astrophysics
- G2 geosciences & technology
- G3 hydrology/oceanography
- G4 meteorology/atmospheric & aerospace science & technology
- G5 mineralogy & petrology

11. ENGINEERING
- E1 computer science/information technology
- E2 electrical & electronic engineering
- E3 energy & fuels
- E4 general & traditional engineering

12. MATHEMATICS
- H1 applied mathematics
- H2 pure mathematics

13. SOCIAL SCIENCES I (GENERAL, REGIONAL & COMMUNITY ISSUES)
- Y1 education, media & information science
- Y2 sociology & anthropology
- Y3 community & social issues

14. SOCIAL SCIENCES II (ECONOMIC, POLITICAL & LEGAL SCIENCES)
- L1 business, economics, planning
- L2 political science & administration
- L3 law

15. ARTS & HUMANITIES
- K0 multidisciplinary
- K1 arts & design
- K2 architecture
- K3 history & archaeology
- K4 philosophy & religion
- K5 linguistics
- K6 literature

MANAGEMENT, STRATEGY AND INNOVATION (MSI)
Naamsestraat 69 bus 3535
3000 LEUVEN, Belgium
tel. + 32 16 32 67 00
msi@econ.kuleuven.be
https://feb.kuleuven.be/research/MSI/