Identifying interdisciplinary research in research projects

Peer-reviewed author version

PHAM, Hoàng Son; VANCRAEYNEST, Bram; POELMANS, Hanne; VANCAUWENBERGH, Sadia & ALI ELDIN, Amr (2023) Identifying interdisciplinary research in research projects. In: SCIENTOMETRICS, 128 (10), p. 5521-5544.

DOI: 10.1007/s11192-023-04810-6
Handle: http://hdl.handle.net/1942/40849

Identifying Interdisciplinary Research in Research Projects

Hoang-Son Pham[1,2*], Bram Vancraeynest[1,2†], Hanne Poelmans[1,2,3†], Sadia Vancauwenbergh[1,2,3†] and Amr Ali-Eldin[1,2,4†]

[1*]Centre for Research & Development Monitoring (ECOOM-UHasselt), Hasselt, 3500, Belgium.
[2]Data Science Institute, Hasselt University, Hasselt, 3500, Belgium.
[3]Directorate Research, Library, International Office, Hasselt University, Hasselt, 3500, Belgium.
[4]Computer engineering and Control systems Department, Faculty of Engineering, Mansoura University, Mansoura, 35516, Egypt .

*Corresponding author(s). E-mail(s): hoangson.pham@uhasselt.be;
Contributing authors: bram.vancraeynest@uhasselt.be; hanne.poelmans@uhasselt.be; sadia.vancauwenbergh@uhasselt.be; amr.alieldin@uhasselt.be;
†These authors contributed equally to this work.

Abstract

Identifying interdisciplinary research has become an important area of study in scientometrics. However, defining what exactly constitutes interdisciplinarity and how it manifests in research activities, such as publications or research projects, remains challenging. In this paper, we propose a mathematical modeling approach to interdisciplinarity measurement based on assessing project diversity. Particularly, we propose a novel approach that combines three indicators: the diversity of researchers, the diversity of research organizations, and the diversity of research disciplines involved in the project, to identify potentially interdisciplinary research projects. To measure diversity, we employ various methods, including distance matrix calculation, evaluation of the distance between researchers, and assessment of the relevancy of researchers' expertise to the projects. We implemented the proposed approach on two datasets; FRIS and Dimensions. We could classify the interdisciplinarity of projects into three groups - Low, Medium, and High. Empirical results analysis supports the proposed approach assumption that the diversity of research projects gets higher when the distances between disciplines in the projects increase. Further, it was shown that the diversity of researchers and organizations was strongly affected by the distance. The number of researchers and organizations had a relatively small impact on the overall diversity score. Furthermore, the relevancy weight can be incorporated as an additional factor in the measurement of interdisciplinary.

Keywords: Interdisciplinary research, Interdisciplinarity indicator, Diversity, Research collaboration, Distance metrics.

1 Introduction

Interdisciplinary research (IDR) is a mode of research that involves the combination of two or more academic disciplines into one activity (e.g., a publication, a research project) (NSF, 2005). IDR is essential to deal with boundary-spanning problems and to encourage the development of emerging research fields. Over the past few years, universities and research funding organizations have strongly encouraged interdisciplinary research at the (sub)national, European, and international levels. (Allmendinger, 2015; Wernli & Darbellay, 2016).

Although IDR has received a lot of attention in research policy, there is a lack of objective consensus in the literature as to the definition of interdisciplinarity. Various terms and definitions such as 'multidisciplinarity', 'transdisciplinarity', and 'crossdisciplinarity' as well as 'interdisciplinarity' are circulating (Choi & Pak, 2006) and are often used to refer to the same concept. In this study, we used the term 'interdisciplinarity' for consistency with other studies in the field, as it is a widely recognized term. According to NSF (2005), interdisciplinarity is defined as the integration of knowledge, tools, and methods from two or more research disciplines to address a scientific or societal issue. Interdisciplinary research involves collaboration among researchers from different disciplines to develop a comprehensive understanding of complex problems and to develop innovative solutions that cannot be achieved through the efforts of individual disciplines alone.

As stated above, "knowledge integration" is the focal point of IDR. Assessing "knowledge integration" is important and challenging work (Glänzel, 2021). According to studies of Adams, Loach, and Szomszor (2016); Zhang, Sun, Chinchilla-Rodríguez, Chen, and Huang (2018), the choice of data, the methodology, and the indicators could produce seriously inconsistent and even contradictory outcomes. The most frequently used approaches for measuring IDR are based on publications and their citations (Cassi, Mescheba, & De Turckheim, 2014; Porter, Cohen, Roessner, & Perreaul, 2007; Zhang, Rousseau, & Glänzel, 2016). These methods relied on the subject classification of the publication's references. Particularly, a publication is considered a potential IDR if the articles in its references section are relatively far from each other in terms of disciplines. The main drawback of the citation-based approach is its reliance on subject classification schemes. Using different subject classification schemes can lead to different or even inconsistent results (Rousseau, Zhang, & Hu, 2019).

Compared to the rich literature of studies that measure IDR based on citation analysis, only a few studies have explored IDR using text-based methods. Typical approaches, in this research direction, are keyword analysis and topic modeling (Ba, Cao, Mao, & Li, 2019; Bonaccorsi, Melluso, & Massucci, 2021; Nichols, 2014; Xu, Guo, Yue, Ru, & Fang, 2016). These are promising approaches, however, to efficiently identify IDR, the text-based approaches need a certain amount of high-quality text which is not always available in many databases.

Another important aspect of IDR is the collaboration between researchers with diverse disciplinary expertise (Abramo, D'angelo, & Costa, 2017; Abramo, D'Angelo, & Di Costa, 2012; Zhang et al., 2018), which is known as organizational approaches. The reasoning behind this focus is that IDR occurs when researchers from different backgrounds or even different research institutions collaborate. When these researchers or organizations contribute from various disciplinary domains, the probability of "knowledge integration," and therefore interdisciplinarity, increases (Rousseau et al., 2019). Methods investigating IDR based on this collaboration approach mainly focus on the expertise of involved researchers or research organizations. To measure this, researchers and/or research organizations need to be assigned one or more disciplines (whether or not from a predefined set of disciplines) that represent their domain of expertise. Given this information, a research activity can be considered IDR if the disciplines of the researchers or research organizations involved in that research activity are highly diverse.

In line with this organizational research direction, in this paper, we propose a novel organizational approach for identifying IDR. This approach is useful when citation data is not available, e.g., research project data. The proposed approach assumes that the level of interdisciplinary research in a project depends on the distances between its disciplines, whether they are disciplines of researchers, organizations, or the project itself. More specifically, in this work, we propose to use three indicators: diversity of researchers, diversity of organizations, and diversity of disciplines assigned to the project to identify IDR. To calculate diversity, we first propose an approach for creating a distance matrix. Then, we formulate the researchers, organizations, and assigned disciplines involved in a research project as vectors that enable the calculation of the distance between them. Additionally, we introduce a relevancy weight that evaluates the relevancy of the researchers' disciplines to their projects. These calculations allow us to determine the diversity of researchers, organizations, and disciplines involved in the projects.

Because it is still unknown how to fully capture interdisciplinarity based on a single indicator, in this study, we propose a combination of three indicators to identify potentially interdisciplinary research projects. The diversity of researchers and organizations involved in a project reflects the range of knowledge backgrounds integrated into it, while the diversity of assigned disciplines shows the range of disciplines within the project itself. We propose utilizing a combination of these elements to integrate details about both the proficiency of the researchers carrying out the research project and the substance of the project. In addition, the relevancy weight can be utilized as an additional indicator to assess the degree of IDR. To the best of our knowledge, there are no other studies that propose methods to calculate these indicators and combine them to identify IDR.

The rest of the paper is organized as follows: Section 2 introduces the data resources and the discipline classification used in this study. Section 3 presents the proposed approach to identify IDR based on research project data. It includes a description of the used approaches to calculate diversity and classify projects into several IDR levels. Section 4 presents the empirical results of this study. Section 5 concludes the paper and highlights possibilities for future studies.

2 Data resource

In this study, we mainly used project data available on the FRIS. The FRIS platform offers details about the research that is (partially) financed by the public sector in Flemish, Belgium, including information about researchers, research institutions, projects, and publications. Over 40 data providers, including organizations involved in both research performance and funding, contribute to the FRIS. The platform is a valuable resource for the Flemish government, which uses it to gather insights for policy-making, generate reports, conduct analyses, and monitor trends. The main goal of FRIS is to accurately reflect the state of research in the Flanders at any given time. To obtain this, all FRIS data providers push information from their institutional research information systems to the FRIS platform and incrementally, in real-time, propagate data changes to FRIS. The FRIS portal hence always demonstrates the most recent, up-to-date information, that is harmonized between data providers. The FRIS portal currently (on the date when data was collected[1]) contains information on 42298 researchers, 2201 organizations, 56200 projects, and 582159 publications. These objects have logical relationships. For example, given a project, we can determine related information such as researchers, organizations, etc. Each project relates to one or some researchers that are in turn affiliated with one or more organizations.

[1] data was collected on March 23^{th} 2023

Each object in the FRIS is assigned one or more research disciplines. This is obligatory since January 2019. To label research objects with research disciplines, FRIS makes use of the Flemish Research Discipline Standard ("Vlaamse Onderzoeksdiscipline Standaard", abbreviated VODS, in Dutch) (Vancauwenbergh & Poelmans, 2019). The VODS is the result of the integration of the three Flemish research discipline classifications: FWO, FRIS, and VLIR discipline classifications, and is based on the structure of the Fields of Research and Development (FORD) list of OECD (OECD, 2015). The VODS contains four hierarchical levels, reflecting research disciplines to a different level of granularity and containing 7, 42, 382, and 2493 disciplines, respectively. The first level corresponds to the six fields of science of the OECD FORD classification (i.e., Natural sciences (01), Engineering and technology (02), Medical & health sciences (03), Agricultural and veterinary sciences (04), Social sciences (05), Humanities and the arts (06)), expanded with one extra discipline to label administrative and technical research personnel (i.e., General and logistic services (07)). The second level contains the main disciplinary fields (i.e., Mathematical sciences (0101), Information and computing sciences (0102), Physical sciences (0103), ...) while the third and fourth levels correspond to more granular subfields. Most objects in FRIS have a level four discipline attached to them.

3 Method

This section presents a novel method to identify IDR projects. In particular, IDR is evaluated based on three indicators: 1) the diversity of the researchers, 2) the diversity of the research organizations, and 3) the diversity of disciplines assigned to the project.

A range of techniques to quantify diversity has been identified in the literature (Glänzel, 2021; Q. Wang & Schneider, 2018). In this work, however, we adopted the Rao-Stirling diversity index (Stirling, 2007) to calculate the diversity of researchers, the diversity of organizations, and the diversity of disciplines assigned to the project. The Rao-Stirling diversity index provides a more robust and nuanced measure of interdisciplinarity than others such as the Simpson index (Simpson, 1949) or Shannon entropy (Ortiz-Burgos, 2016), making it a valuable tool for researchers studying interdisciplinary collaboration and innovation (Porter & Rafols, 2009). It is well-suited for measuring interdisciplinarity as it considers not only the number of categories (variety) and their probability distribution (balance) but also incorporates the pairwise distances (disparity) between them (Rafols & Meyer, 2010). Suppose we have N categories, denoted by $1, 2, \ldots, N$. Let p_i be the probability distribution of the number of elements in category i and the total number of elements ($p_i = x_i/X$; $X = \sum x_i$), and let m_{ij} be the distance between categories i and j. The Rao-Stirling diversity index is calculated as follows:

Table 1: Notations used in the paper.

Notation	Description
v	vector of disciplines
n	number of researchers in a project
m	number of organizations in a project
k	number of disciplines assigned to a project
N	number of disciplines in data
Δ_R	diversity of researchers
Δ_O	diversity of organizations
Δ_D	diversity of disciplines
Δ_{RW}	diversity of researchers with relevancy weight
Δ_{OW}	diversity of organizations with relevancy weight
λ	distance between two discipline vectors
θ	relevancy weight of disciplines of a researcher and disciplines of a project
M	distance matrix
m_{ij}	distance between two disciplines i and j

$$\Delta = \sum_{ij} p_i * p_j * m_{ij}. \tag{1}$$

In this study, the categories can be considered as either the researchers, organizations, or disciplines assigned to the project. To apply the Rao-Stirling diversity index to calculate the diversity of researchers, organizations, and disciplines, we need to calculate the related factors: variety, balance, and disparity. Particularly, we need to calculate the number, frequency of researchers/organizations/disciplines, and especially the distance between researchers/organizations/disciplines. In the following sections, we introduce some terminology and the approach used to calculate these factors. To easily read the following sections, we present a list of notations used in the paper in Table 1.

3.1 Distance matrix

A distance matrix reflects the mutual similarity or distance between disciplines in the applied subject scheme. It is an important part of diversity calculation. As mentioned in study of Thijs, Huang, and Glänzel (2021), there are various methods used to create a distance matrix such as bibliographic coupling, co-citation, and cross-citation. These approaches can be applied on different research classification schemes such as Web of Science Categories, Leuven-Budapest classification scheme (Glänzel & Schubert, 2003). It is important to properly choose a method and classification scheme as it may lead to large differences in the obtained diversity scores (J. Wang, Thijs, & Glänzel, 2015).

In theory, to measure the similarity between disciplines, most approaches rely on citation-based analysis (Leydesdorff & Rafols, 2009; Zhang et al., 2016). For example, co-citation analysis is a method used in scientometrics to examine the relationship between scientific articles based on their shared citations. By applying co-citation analysis, a distance matrix can be created to represent the similarity or dissimilarity between pairs of articles. This is achieved by constructing a co-citation matrix that counts the number of times each pair of articles is co-cited. If two articles from different research disciplines have a high co-citation count, it suggests a frequent citation pattern, indicating a potential relationship or connection between the disciplines.

Similarly, this study assumes that two disciplines are considered similar if they frequently co-occur. In this context, we developed a discipline distance matrix as a tool to measure interdisciplinarity in research projects based on projects metadata only and not on publications data. Consequently, we employ the co-occurrence of disciplines among researchers who collaborate in the projects to create the distance matrix. This approach proves particularly valuable when citation information is unavailable, such as in research information systems that store project data but lack citation data. Specifically, the construction of the distance matrix proceeds as follows. We first define the probability distribution of disciplines of a researcher as follows:

$$v = (v_1, v_2 \ldots, v_N) \in \mathbb{R}^N, \tag{2}$$

where each coordinate v_i represents the probability of a possible discipline, such that $\sum v_i = 1$. The probability of a discipline, v_i, is the proportion of that discipline within the total probability of all disciplines. For example, suppose a researcher has three discipline codes: $(0101, 0102, 0103)$, the vector is $v = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0 \ldots, 0)$.

We consider each pair of researchers in a project as a collaboration. Each researcher is represented by a discipline vector (e.g., p, q). The weight of the collaboration between discipline i and j within researchers p and q is calculated as follows:

$$w_{p,q}(i,j) = \begin{cases} p_{v_i} \cdot q_{v_j} + p_{v_j} \cdot q_{v_i} & \text{if } v_i \neq v_j, \\ p_{v_i} \cdot q_{v_i} & \text{if } v_i = v_j \end{cases} \tag{3}$$

For each collaboration within each project between two individuals, we count the collaborations between their respective disciplines. This process allows us to create a co-occurrence matrix. In the co-occurrence matrix, each cell c_{ij} represents the total number of weighted collaborations between disciplines i and j.

After obtaining the co-occurrence matrix, we utilize the cosine similarity measure (Salton & Buckley, 1988) to compute the similarity between two disciplines, which yields a similarity matrix. Subsequently, the distance matrix is constructed by subtracting the similarity matrix from 1.

3.2 Disparity calculation

This section presents a novel approach to calculate the disparity between two researchers/organizations which plays an important role in the calculation of the diversity of the researchers/organizations. To do that, we first formalize the expertise of the researcher/organization as a vector. The disparity between two researchers/organizations is then considered as the distance between two representative vectors.

3.2.1 Discipline vector

In this study, we assume that disciplines related to a researcher can be inferred from other resources such as disciplines of their affiliations, disciplines of projects that they worked on, disciplines of co-authors in projects, disciplines of publications, and disciplines of co-authors in publications. Suppose a researcher associates with the following disciplines: disciplines of organizations (denoted by $v_{\_organization}$), disciplines of projects (denoted by $v_{\_project}$), disciplines of co-authors in projects (denoted by $v_{\_co-author-project}$), disciplines of publications (denoted by $v_{\_publication}$), and disciplines of co-authors in publications (denoted by $v_{\_co-author-publication}$), the disciplines of the researcher can be calculated as follows:

$$\begin{aligned} v = & v_{\_organization} * w_{\_organization} + \\ & v_{\_project} * w_{\_project} + \\ & v_{\_co-author-project} * w_{\_co-author-project} + \\ & v_{\_publication} * w_{\_publication} + \\ & v_{\_co-author-publication} * w_{\_co-author-publication}, \end{aligned} \tag{4}$$

where w_i is a weight of the corresponding vector and $\sum w_i = 1$.

From the dataset and by using linear regression, we can find the values of the weighting factors in Equation (4). In the experimental work section, we further show how the predictability of researcher disciplines can be improved using neural networks.

3.2.2 Distance calculation

Each researcher or organization is defined as a vector with each element representing a probability distribution of a discipline from a set of N disciplines. The distance between two researchers or two organizations is defined as the distance between two representative vectors. In this work, to calculate the distance between two discipline vectors, we propose to use the **Wasserstein distance** (Olkin & Pukelsheim, 1982). The Wasserstein distance, also known as the Earth Mover's distance, is a distance metric that is particularly well-suited for comparing probability distributions over a given space M which is the distance matrix in this work.

One of the main advantages of the Wasserstein distance is that it can capture the underlying geometry of the probability distributions being compared. This is because it takes into account the actual locations of the probability mass, rather than just comparing histograms or probability density functions. This makes it a very powerful tool for comparing distributions that may be fundamentally different in shape or structure. Another advantage of the Wasserstein distance is that it is a stable distance metric. This means that small changes in the input distributions will result in small changes in the distance, which makes it a useful tool for applications where small differences in the distributions are important. Additionally, Wasserstein distance also has nice mathematical properties. It's a metric that is stronger than triangle inequality, which means that it's relatively smoother.

To calculate the Wasserstein distance between v_p and v_q, we need to compare how the probabilities are distributed across the different positions or values that the random variable can take. We do this by constructing a matrix of distances M, where m_{ij} is the distance between position i of v_p and position j of v_q. Next, it solves a linear programming problem to find the optimal "transport plan" for moving the probabilities from v_p to v_q in a way that minimizes the total distance traveled. This transport plan tells us how much probability needs to be moved from each position i of v_p to each position j of v_q to transform v_p into v_q while conserving the total mass of each distribution. The optimal solution of the linear programming problem gives us the minimum cost of transporting the mass of v_p to v_q, which is the Wasserstein distance between v_p and v_q.

More specifically, let $M \in \mathbb{R}^{N \times N}$ be the distance matrix such that m_{ij} is the distance between discipline i and j; let v_p and v_q be respectively probability distribution discipline vectors; the Wasserstein distance between v_p and v_q is the minimum cost of moving plan between v_p and v_q. It is calculated as follows:

$$\lambda_{pq} = min \sum P_{ij} m_{ij}, \tag{5}$$

where P_{ij} is the amount of mass to be transported from position i in v_p to position j in v_q.

3.2.3 Diversity calculation

Researcher diversity calculation

To calculate the diversity of researchers, we first create a vector representing the disciplines for each researcher participating in the project, as defined by Equation (2). We then apply the Rao-Stirling diversity index to these vectors to calculate the diversity of researchers. Specifically, let $R = (f_1, f_2, \ldots, f_n)$ be the frequency of discipline vectors of researchers, the diversity of R, denoted by Δ_R, is calculated as follows:

$$\Delta_R = \sum_{ij} f_i * f_j * \lambda_{ij}, \tag{6}$$

where f_i and f_j are the frequency of the discipline vectors i, j. λ_{ij} is the distance between two discipline vectors i and j. Note that to avoid repeated submissions, in this equation, the index i takes values from 1 to $n-1$, while the index j takes values from $i+1$ to n.

As an example, consider four researchers with their disciplines as follows: $r_1 = (0101)$, $r_2 = (0102)$, $r_3 = (0101, 0102)$, $r_4 = (0101, 0102)$. The discipline vectors of these researchers are $v_{r_1} = (1, 0, ...)$, $v_{r_2} = (0, 1, ...)$,

$v_{r_3} = (0.5, 0.5, ...)$, $v_{r_4} = (0.5, 0.5, ...)$. Since v_{r_3} and v_{r_4} are identical, we consider them as 1 vector. As a result, the discipline vectors $R = (f_1, f_2, f_3)$ where $f_1 = 0.25$, $f_2 = 0.25$, and $f_3 = 0.5$. The diversity of researchers is calculated as $\Delta_R = 0.25 * 0.25 * \lambda_{r_1 r_2} + 0.25 * 0.5 * \lambda_{r_1 r_3} + 0.25 * 0.5 * \lambda_{r_2 r_3}$.

Organization diversity calculation

To calculate the diversity of the organizations involved in a project, we first collect the research disciplines of each organization from its profile. For each organization, we then create a discipline vector, as defined by Equation (2). Suppose $O = (f_1, f_2, ..., f_m)$ be the frequency of discipline vectors of m organizations. The diversity of O, denoted by Δ_O, can be calculated by using the Equation (6) where f_i and f_j represent the frequency of the discipline vectors of organizations i and j, respectively. λ_{ij} represents the distance between the discipline vectors of organizations i and j.

Discipline diversity calculation

To calculate the diversity of disciplines assigned to the project, we take into account the number, frequency, and distance between disciplines. To do that, for each project, we collect its related disciplines and create a discipline vector, as defined by Equation (2). We then apply the Rao-Stirling diversity index to this vector to calculate the diversity of disciplines. Specifically, let $D = (p_1, p_2, \ldots, p_N)$ be discipline vector of project p, the diversity of D, denoted by Δ_D, is calculated as follows:

$$\Delta_D = \sum_{i,j} p_i * p_j * m_{ij}, \tag{7}$$

where p_i and p_j are the frequency of disciplines i and j. m_{ij} is the distance between disciplines i and j, which can be obtained directly from the distance matrix M.

3.3 Relevancy weight

In this work, we employed the Rao-Stirling method to assess the diversity of researchers and organizations. One notable characteristic of this measure is that any alteration in factors such as variety or disparity will result in a change in diversity. For instance, a project with a greater number of researchers involved would exhibit higher researcher diversity compared to a project with a smaller research team. In other words, this approach might inadvertently assign less importance to smaller universities or organizations with fewer researchers, while favoring those with larger research teams and organizations. Consequently, the evaluation of IDR could be biased if universities attempt to manipulate the results by assigning researchers with diverse disciplines or from diverse organizations in their projects in order to obtain a higher score and secure additional funding

To address this limitation, one potential solution is to incorporate a relevancy weight for the researchers or organizations involved in the project. This weight can serve as an additional metric to assess the degree of IDR within the projects. In practice, the disciplines of researchers in a research project are more or less related to the project's disciplines. To assess the relevance of researchers in the project, we propose to assign a relevancy weight to each researcher/organization participating in the project. The relevancy weight can be defined by the minimum distance between disciplines of the project and the disciplines of the researcher. If at least one of the disciplines of the researcher is close to any of the disciplines of the projects, the value of the relevancy weight will be high. Particularly, consider two sets of disciplines: $p = (dp_1, dp_2, dp_N)$ and $r = (dr_1, dr_2, dr_N)$. The relevancy weight is determined by the minimum distance between each pair of disciplines from sets p and r. It is calculated as follows:

$$\theta_{pr} = 1 - min(distance(dp_i, dr_i)), \tag{8}$$

where i ranges from 1 to N. dp_i and dr_i represent the i^{th} discipline in sets p and r, respectively. The $distance(dp_i, dr_i)$ represents the distance between the i^{th} disciplines, which can be calculated using the distance metric M.

3.4 IDR identifying

To identify IDR in research projects, we propose to use a combination of three indicators: the diversity of researchers, the diversity of organizations, and the diversity of attached disciplines. Projects with larger diversity scores are evaluated as having a higher probability of being IDR compared to others with smaller diversity scores. We can sort the projects based on the mean of these three diversity scores and indicate which projects are more likely to be IDR than others. However, for analysis purposes, it is necessary to classify projects into different categories based on their diversity scores. Each category can be considered an IDR level with a certain probability of containing interdisciplinary research projects. Since various IDR measurement strategies may use different scales, we propose to categorize projects based on the diversity scores into three levels - Low, Medium, and High - in relation to one another.

4 Experimental work

4.1 Experimental setup

4.1.1 Data

As a case study, we selected projects available in the FRIS portal with the following criteria: 1) start date on 01/01/2019 or later, 2) the number of researchers in the project is larger than one, and 3) excluding projects that contain discipline `0700 General and logistic services`. As a result, 3379 projects were selected for the analysis. Statistics on the input data are shown in Table 2.

Table 2: Statistics on input data.

	#Researchers	**#Organizations**	**#Disciplines**
mean	2.75	1.67	1.64
std	1.89	1.28	0.97
min	2.00	1.00	1.00
max	36.00	24.00	13.00

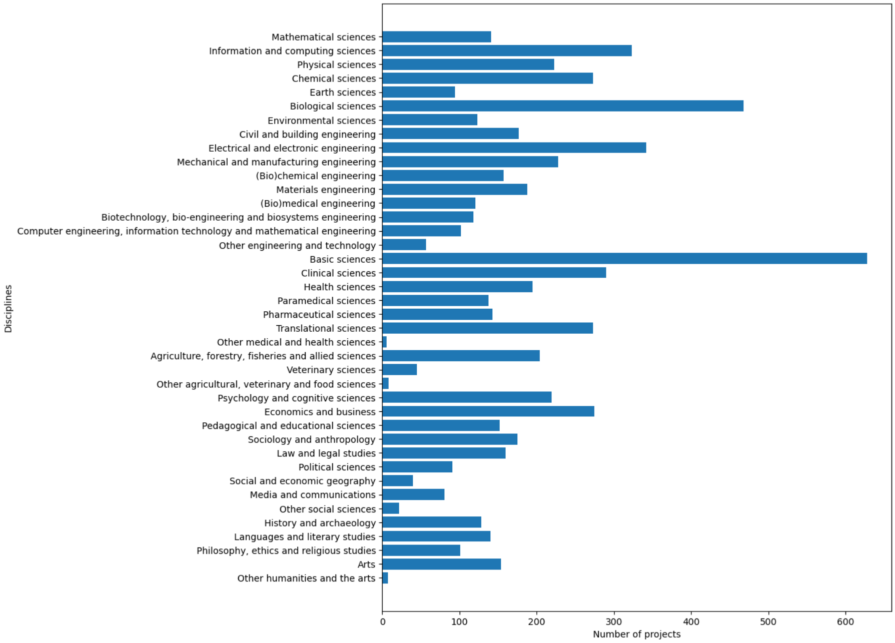


Fig. 1: Projects distributed over the disciplines.

4.1.2 Granularity of disciplines

In this study, we focused on the second level of VODS which includes 41 disciplines. Note that we already excluded the `General and logistic services` because it is not an actual research discipline. More granular levels (the third and/or fourth levels) were mapped onto their corresponding, hierarchical higher second-level disciplines. For example, suppose a researcher r has a set of disciplines: $(01010101, 01010103, 01020201)$, then the disciplines of r will be reduced to the second level codes as $(0101, 0101, 0102)$.

We used the second level of VODS because it reflects the main research fields of research objects. The first level is too general, whereas the third and fourth levels are too specific. This choice of granularity of disciplines was also comparable with other studies. For instance, the study of Zhang et al. (2018) made use of the ECOOM's 68 sub-disciplines (Glänzel & Schubert,

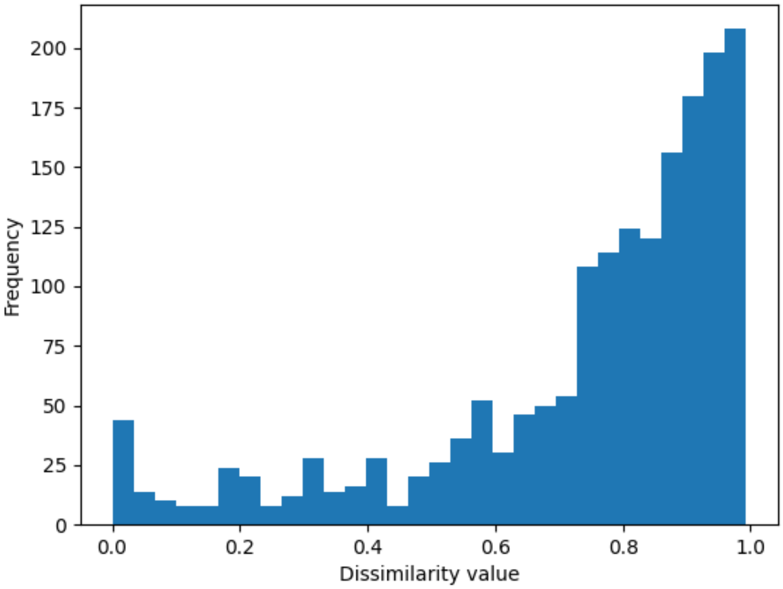


Fig. 2: Distribution of dissimilarity values.

2003) to classify affiliations, whereas a study of Rafols and Meyer (2010) used 175 subject categories in ISI to analyze reference list. These categories can be compared and matched. For example, any disciplines from 175 categories in ISI can each be assigned to one of 68 ECOOM's sub-disciplines. Similarly, for each discipline of 68 ECOOM's sub-disciplines or 175 ISI categories, we can convert it into an equivalent discipline in the second level of VODS. These concordances tables are available and can be downloaded by researchers in the Flemish region. At the moment, there is no automatic mechanism applied but we plan to do that in the future. Projects distributed over the disciplines are illustrated by Fig. 1.

4.1.3 Distance matrix

To calculate the diversity of researchers, organizations, and assigned disciplines, we created the distance matrix based on collaborations of researchers in 20,776 projects with the 41 research fields in the second level of the VODS. Distribution of dissimilarity values is shown in Fig. 2. As can be seen, the values are variably distributed; however, the majority of values are close to one. In other words, the distribution of the values in the distance matrix is negatively skewed, and the majority of the values are concentrated towards the higher end of the distribution. A high degree of dissimilarity reflected in the distance matrix indicates that the dataset consists of diverse research areas. This could be due to various reasons, such as a wide range of research topics, different methodologies, or diverse research disciplines within the dataset. These findings accurately reflect the disciplines within the second level of VODS, as this level encompasses the main research disciplines. Optimization and evaluation of the distance matrix are beyond the scope of this work; however, further evaluation and enhancements of the distance matrix will be conducted in future research.

4.1.4 Discipline vector

Recall that in the FRIS data, all objects, e.g., projects, organizations, publications, and projects, were assigned one or more discipline codes. In addition, these objects were logically linked to each other. For example, given a project, we can get related researchers, organizations, and discipline codes. Further, given a researcher, we can get his/her research groups, publications, projects, and disciplines from the profile. Furthermore, given a publication, we can get a list of researchers; etc. This logical relationship allows us to collect disciplines and aggregate the discipline vector of the researcher.

In this study, we collected the disciplines of the researchers from various objects such as projects, and publications. We assumed that data integrity is taken care of by any RIS that stores publication data such as FRIS. Collecting projects or publications data from different external data sources is not in the scope of this paper but we plan in the future to investigate ways to harmonize publications metadata collected from different data sources.

For each researcher, we collected disciplines from six objects: the profile, organizations, projects, co-authors on projects, publications, and co-authors on publications. We first checked whether there was any dependency between these disciplines. To do this, we created a matrix for each researcher; a 2-dimensional array with N rows and six columns. Each row represents a discipline, and the values of the columns are the frequency of the discipline appearing in the six objects. We repeated this process for each researcher. At the end of this process, a 2-dimensional array with $n * N$ rows and six columns was created, where n is the number of researchers and N is the number of disciplines. In this matrix, we excluded rows containing only zeros. In this experiment, we created the matrix based on 1309 researchers who worked on 500 projects. The achieved matrix contained 12,844 rows and six columns. The Pearson correlation scores of the six objects are shown in Table 3. The correlation scores between the profiles' disciplines and the other factors were above 0.5, except for publications, which was 0.13. All p-values obtained from the correlation test were less than 0.01, indicating that the results were significant.

Further analysis was conducted to see whether disciplines of the profile can be predicted from the other factors. To do that, we applied 1) multilayer perceptron neural networks (MLP) and 2) Long short-term memory neural networks (LSTM) to train the prediction model on the matrix. We used disciplines of the profile as the output variable and other factors: organization, project, co-authors on projects, publications, and co-authors on publications

Table 3: Person correlation scores of six variables.

	profile	**organization**	**project**	**co-project**	**pub**	**co-pub**
profile	1	0.62	0.66	0.73	0.13	0.50
organization		1	0.55	0.64	0.12	0.52
project			1	0.72	0.15	0.49
co-project				1	0.14	0.60
pub					1	0.19
co-pub						1

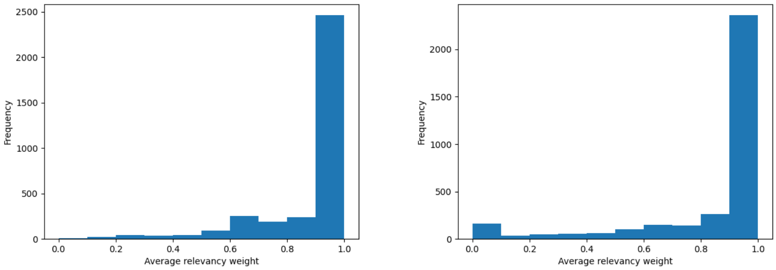
All p-values related to the coefficient scores are statistically significant at a level of less than 0.01.

as predictor variables. The Mean Square Error (MSE) scores for MLP and LSTM were 0.02 and 0.05, respectively, indicating that the profile disciplines could be accurately predicted for the other research objects associated with the researcher.

In this work, we used the true labels (i.e., disciplines assigned to the profiles) to create the discipline vectors of researchers. The disciplines vectors of organizations were calculated based on disciplines available in organizations' profiles. Similarly, the disciplines vectors of projects were calculated based on the disciplines assigned to the projects.

4.1.5 IDR identifying

To assess the IDR in the input project data, we calculated diversity for researchers, organizations, and assigned disciplines. In this work, the diversity of researchers and organizations was calculated by Equation (6), while the diversity of disciplines was determined using Equation (7). Furthermore, in order to assess the effectiveness of the relevancy weight, we computed the relevancy weight for each researcher/organization involved using Equation (8). We used the average scores of the three diversity measures to classify projects into three IDR levels - Low, Medium, and High. Projects with the lowest diversity score were assigned to the low group, while those with the highest diversity score were assigned to the high group. To classify projects into three groups, we applied the characteristic score and scale (CSS) algorithm (Glänzel & Schubert, 1988). CSS is a method for scaling and ranking items or variables based on their relative importance or strength in a given dataset. The algorithm works by first calculating the mean and standard deviation of each item, and then computing a characteristic score for each item based on its deviation from the mean. The characteristic score is a measure of the relative importance of an item compared to the others in the dataset. After the characteristic scores are computed, the CSS algorithm then scales the items so that they have a common range of values. This allows the items to be directly compared to each other and ranked based on their characteristic scores.



(a) Distribution in researchers. (b) Distribution in organizations.

Fig. 3: Distribution of average relevancy weights.

4.2 Experimental results

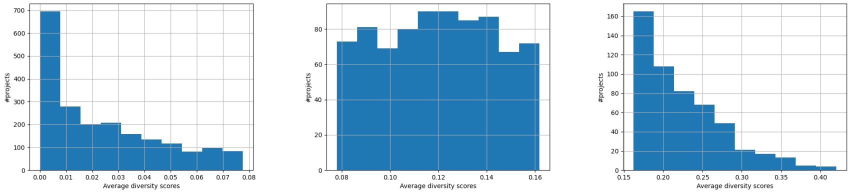
4.2.1 Relevancy weight

We first present the results of the relevancy weight calculation. Recall that we computed the relevancy weight for each researcher/organization in each project, and also determined the average values for these weights. The distributions of average relevancy weights for researchers and organizations involved in projects are illustrated in Fig. 3. As can be seen, a significant majority of projects exhibit high researcher relevancy weights (Fig. 3a): 95.64% have weights higher than 0.5, with 60.43% having exact matches, and only a small fraction (4.35%) having weights lower than 0.5. Similarly, for organizations (Fig. 3b), 62.27% have exact matches, 89.35% have relevancy weights larger than 0.5, and 10.65% have weights smaller than 0.5. These findings indicate that the disciplines of most researchers/organizations are relevant to the disciplines of the projects. However, there is a small percentage of researchers/organizations whose disciplines are less relevant to the projects' disciplines. Upon further analysis of projects with low researcher relevancy weights, we observed instances where a researcher was added to the project due to their profile expertise being closely aligned with the project scope. However, there were cases where the researcher was associated with a discipline not explicitly specified or may have a missing discipline in their profile that would be a better match for the project. This subset represents 4.35% of the total number of projects, and our model could not differentiate them from irrelevant researchers. Although this group of researchers falls outside the scope of our current work, we plan to conduct further studies to include them in the future.

These findings demonstrate that the relevancy weight does not significantly influence the IDR calculation results, as a considerable percentage of researchers and organizations were found to be highly related to the projects. However, in cases where the IDR is high while the relevancy weight is low or a large number of researchers or organizations are not closely associated with the

Table 4: Statistics on the output.

Group	**Number of projects**	**Percentage(%)**
High	532	15.74
Medium	794	23.50
Low	2053	60.76



(a) Low (b) Medium (c) High

Fig. 4: Distribution of average diversity scores in the three groups.

projects, additional measures should be implemented to verify the accuracy of the IDR calculation within those specific projects.

4.2.2 IDR classification

We used the average score of three diversity scores (Δ_R, Δ_O, Δ_D) to classify projects into three groups: low, medium, high. The descriptive statistics for each of the three groups are shown in Table 4. The number of projects across the three groups varies greatly, with only a small proportion being evaluated as having high diversity score. Specifically, the high group includes 532 projects, comprising only 15.74% of the total input projects. In contrast, the low group comprises a large number of projects (2053 projects, accounting for 60.76% of the total). The medium group includes 794 projects which account for 23.50% of the total projects. These findings indicate that relatively few projects were evaluated as having high diversity scores. We further showed the distribution of average IDR scores within these groups in Fig. 4. As shown in Fig. 4a, most of the projects in the low group have small average diversity scores, with many scores close to zero. For the medium group, depicted in Fig. 4b, the average diversity scores are equally distributed across the projects. In the high group, represented in Fig. 4c, a few projects have high average diversity scores. However, the majority of the projects in this group have average diversity scores of 0.15 and 0.3.

4.2.3 Results evaluation

In this section, we further analyzed the results to evaluate the performance of the proposed method in identifying IDR. Recall that each project available on

FRIS has been labeled with one or more disciplines. Based on the assigned disciplines, one could have a good indication of how diverse a research project is. For example, a project that belongs to two disciplines such as `(Bio)medical engineering` and `Computing sciences` could be considered as high IDR since this project was related to research fields that are very different from each other. In contrast, a pair of disciplines `Translational Sciences` and `Basic sciences` are very close. Thus, projects assigned by these disciplines could be considered as low IDR. Based on this observation, we evaluated the performance of the proposed method by assessing how it could identify projects that were assigned multiple dissimilar disciplines.

To identify the disciplines that frequently co-occurred in projects within each group, we utilized the widely-used association rule mining algorithm (Han, Pei, & Yin, 2000) commonly employed in the field of data mining. Association rule mining is a data mining technique used to discover associations or relationships between variables in large datasets. The goal is to identify frequent patterns or co-occurrences of items in the data. The technique works by examining the frequency of items that appear together in transactions or records and then generates rules that describe the relationships between these items. In this study, we consider each project as a record, and disciplines in the projects as items. In the context of association rule mining, a confidence measure is a metric that is used to evaluate the strength of a relationship between two items in a dataset. The confidence of a rule `"A => B"` is defined as the proportion of transactions that contain both A and B over the proportion of transactions that contain A. Confidence measures are commonly used in association rule mining to identify the most important and reliable rules and to filter out spurious or uninteresting rules that have low confidence values (more details about association rule mining algorithm and constraints can be found in the study of Han et al. (2000)).

To identify pairs of disciplines that co-occur in projects, we used association rule mining, treating disciplines as items and projects as transactions. Specifically, we applied the FPGrowth algorithm (Han et al., 2000) with a minimum confidence threshold of 10%. With the specified threshold, we identified 35 association rules, as shown in Table 5. These rules indicate that certain pairs of academic disciplines frequently appear together in research projects with low distance scores. For example, the rules `Biological sciences => Basic sciences`, `Translational sciences => Basic sciences`, and `Clinical sciences => Basic sciences` were found in many projects. These three pairs of disciplines all have distance scores below 0.2, which suggests a strong association between them.

We further evaluated to see the correlation between *Occ*, *Dist*, and diversity scores as well as with other factors such the balance, variety, and disparity. To do that, for each association rule, we first selected the projects that included both disciplines in that association rule. Then we calculated all factors related to projects such as diversity, balance, variety, and disparity. Table 6 shows the results of this calculation. In the table, we used fr_i, fo_i, and fp_i to indicate the

Table 5: Association rules within the projects.

No	Association rule	Occ	$Conf$	$Dist$
1	Biological sciences =>Basic sciences	95	0.24	0.16
2	Translational sciences =>Basic sciences	94	0.4	0.12
3	Clinical sciences =>Basic sciences	87	0.36	0.09
4	Translational sciences =>Clinical sciences	52	0.22	0.04
5	Information and computing sciences => Electrical and electronic engineering	49	0.18	0.43
6	Clinical sciences =>Biological sciences	44	0.18	0.2
7	Pharmaceutical sciences =>Basic sciences	44	0.37	0.21
8	Agriculture, forestry, fisheries and allied sciences => Biological sciences	43	0.26	0.38
9	Mechanical and manufacturing engineering => Electrical and electronic engineering	42	0.22	0.1
10	(Bio)chemical engineering =>Chemical sciences	41	0.29	0.04
11	Materials engineering =>Chemical sciences	40	0.25	0.33
12	Physical sciences =>Electrical and electronic engineering	40	0.22	0.54
13	Mathematical sciences =>Information and computing sciences	36	0.31	0.37
14	Physical sciences =>Chemical sciences	36	0.2	0.53
15	Biotechnology, bio-engineering and biosystems engineering => Biological sciences	34	0.34	0.54
16	Health sciences =>Basic sciences	32	0.2	0.28
17	Chemical sciences =>Biological sciences	32	0.14	0.63
18	Computer engineering, information technology and mathematical engineering =>Electrical and electronic engineering	30	0.33	0.17
19	Translational sciences =>Biological sciences	29	0.12	0.16
20	Paramedical sciences =>Clinical sciences	28	0.27	0.09
21	(Bio)chemical engineering =>Materials engineering	28	0.2	0.36
22	History and archaeology =>Arts	26	0.24	0.15
23	Pharmaceutical sciences =>Translational sciences	25	0.21	0.2
24	Sociology and anthropology =>Economics and business	25	0.18	0.63
25	Computer engineering, information technology and mathematical engineering =>Information and computing sciences	24	0.27	0.18
26	(Bio)medical engineering =>Basic sciences	24	0.23	0.28
27	Earth sciences =>Environmental sciences	24	0.29	0.33
28	Materials engineering => Mechanical and manufacturing engineering	24	0.15	0.5
29	Materials engineering =>Electrical and electronic engineering	24	0.15	0.58
30	Paramedical sciences =>Basic sciences	23	0.22	0.11
31	Pharmaceutical sciences =>Clinical sciences	22	0.18	0.18
32	(Bio)medical engineering =>Translational sciences	21	0.2	0.3
33	Health sciences =>Clinical sciences	21	0.13	0.31
34	Earth sciences =>Biological sciences	21	0.25	0.54
35	Mechanical and manufacturing engineering => Information and computing sciences	21	0.11	0.59

Occ: occurrences, $Conf$: confident score, $Dist$: distance score

average frequency of researchers, organizations, and disciplines, respectively. Meanwhile, n, m, and k indicated the average number of researchers, organizations, and disciplines, respectively. The symbols λ indicated the average distance between researchers/organizations. Lastly, m_{ij} indicated the average distance between disciplines i and j.

We used the Pearson correlation method to determine the correlation between these factors. The correlation scores between these factors are illustrated in Table 7. Note that in this table, the values shown in the cells below are p-values. The results showed that Occ had a negative correlation with diversity scores. Specifically, the correlation scores between Occ and Δ_R, Δ_O, and Δ_D were -0.48, -0.39, and -0.53, respectively. The correlation scores between

Table 6: Diversity score and related factors associated with association rules.

No	Δ_R	fr_i	n	λ_r	Δ_O	fo_i	m	λ_o	Δ_D	fp_i	k	m_{ij}
1	0.08	0.56	2.03	0.28	0.01	0.17	0.41	0.05	0.31	0.38	3.16	0.49
2	0.09	0.45	2.6	0.3	0.04	0.2	0.74	0.15	0.32	0.36	3.32	0.49
3	0.07	0.56	2.19	0.26	0.05	0.21	0.71	0.19	0.38	0.33	3.33	0.57
4	0.08	0.48	2.7	0.28	0.04	0.2	0.83	0.16	0.3	0.31	4.04	0.41
5	0.09	0.49	2.49	0.31	0.05	0.2	0.89	0.2	0.31	0.41	2.77	0.52
6	0.08	0.5	2.25	0.31	0.02	0.13	0.55	0.08	0.34	0.39	2.83	0.55
7	0.07	0.57	1.9	0.26	0.05	0.18	0.9	0.2	0.37	0.32	3.43	0.54
8	0.08	0.49	2.3	0.28	0.05	0.16	0.78	0.18	0.18	0.38	2.9	0.26
9	0.08	0.5	2.38	0.27	0.06	0.23	1.03	0.2	0.29	0.35	3.18	0.44
10	0.11	0.43	2.75	0.38	0.09	0.28	1.4	0.31	0.34	0.36	3.05	0.53
11	0.11	0.43	2.67	0.36	0.08	0.23	1.13	0.28	0.28	0.37	3.04	0.43
12	0.09	0.43	2.71	0.31	0.06	0.23	1.12	0.22	0.34	0.37	2.97	0.54
13	0.05	0.5	2.42	0.17	0.03	0.28	1.13	0.11	0.14	0.39	2.83	0.22
14	0.05	0.46	2.57	0.16	0.03	0.24	1.02	0.12	0.16	0.34	3.2	0.24
15	0.05	0.45	2.38	0.17	0.02	0.15	0.55	0.09	0.18	0.31	3.55	0.25
16	0.08	0.41	3.35	0.26	0.05	0.2	1.12	0.16	0.26	0.41	2.7	0.43
17	0.09	0.5	2.4	0.29	0.03	0.11	0.57	0.11	0.26	0.36	3.21	0.37
18	0.11	0.47	2.48	0.39	0.05	0.2	0.96	0.19	0.4	0.31	3.61	0.58
19	0.09	0.46	2.7	0.3	0.03	0.15	0.63	0.12	0.24	0.34	3.59	0.34
20	0.1	0.47	2.39	0.35	0.05	0.23	0.87	0.21	0.35	0.36	3.3	0.53
21	0.09	0.46	2.32	0.32	0.06	0.2	0.79	0.23	0.27	0.34	3.21	0.41
22	0.11	0.4	3	0.34	0.07	0.32	1.42	0.24	0.27	0.33	3.63	0.38
23	0.09	0.39	3	0.29	0.06	0.28	1.29	0.22	0.27	0.34	3.33	0.4
24	0.04	0.47	2.61	0.14	0.03	0.28	1.22	0.09	0.11	0.35	3.23	0.16
25	0.07	0.47	3.13	0.24	0.06	0.17	1.31	0.19	0.27	0.31	3.69	0.38
26	0.04	0.48	2.48	0.13	0.03	0.21	1.13	0.11	0.16	0.35	3.48	0.23
27	0.08	0.43	2.95	0.27	0.05	0.29	1.68	0.19	0.19	0.34	3.16	0.29
28	0.05	0.41	2.84	0.16	0.03	0.3	1.21	0.13	0.13	0.37	3.04	0.19
29	0.07	0.53	4.24	0.2	0.04	0.13	1.81	0.13	0.28	0.31	3.76	0.4
30	0.05	0.48	2.46	0.18	0.04	0.3	1.29	0.16	0.15	0.37	3.25	0.21
31	0.07	0.41	3.14	0.21	0.06	0.35	2.14	0.2	0.18	0.33	3.36	0.25
32	0.06	0.41	2.87	0.2	0.04	0.32	1.38	0.15	0.14	0.31	3.44	0.2
33	0.05	0.44	2.88	0.2	0.02	0.29	1.52	0.08	0.17	0.34	3.2	0.25
34	0.11	0.48	2.43	0.35	0.06	0.26	1.09	0.2	0.4	0.39	2.87	0.65
35	0.09	0.42	2.58	0.3	0.07	0.33	1.54	0.21	0.25	0.38	2.85	0.38

Dist and Δ_R, Δ_O, and Δ_D were positive with the scores of 0.57, 0.33, and 0.90, respectively. This outcome supports the proposed approach because the proposed approach assumes that the diversity of the project gets higher when the distances between disciplines in the projects increase. We also observed a correlation between Δ_R and other related factors such as frequency, number, and distance. Δ_R showed a significant correlation with the average distance, with score of 0.97. However, the correlation score between Δ_R and frequency and the number of researchers did not satisfy the significance testing score (0.05). This finding indicates that Δ_R is influenced by the distance between researchers, but not by the number of researchers.

Similarly, the Δ_O strongly correlates with the distance between organizations, with correlation score of 0.97. Δ_O also correlates with the number of organizations, although not as strongly, with a correlation score of 0.41. Δ_D was significantly correlated with distance (m_{ij}), but it was not correlated with other factors such as frequency (fp_i) or the number of disciplines (k).

These findings suggest that the diversity scores are not equally related to other factors and that some factors may have a stronger influence on

Table 7: Correlation scores between factors.

	Occ	$Dist$	Δ_R	fr_i	n	λ_r	Δ_O	fo_i	m	λ_o	Δ_D	fp_i	k	m_{ij}
Occ	1.0	-0.35 0.04	-0.48 0.0	-0.04 0.84	-0.08 0.65	-0.48 0.0	-0.39 0.02	0.17 0.32	-0.05 0.75	-0.41 0.01	-0.53 0.0	0.29 0.09	-0.33 0.05	-0.47 0.0
$Dist$		1.0	0.57 0.0	0.31 0.07	-0.22 0.21	0.64 0.0	0.33 0.05	-0.23 0.18	-0.24 0.16	0.4 0.02	0.9 0.0	0.15 0.4	-0.17 0.33	0.93 0.0
Δ_R			1.0	-0.15 0.39	-0.02 0.91	0.97 0.0	0.64 0.0	-0.1 0.56	-0.12 0.51	0.67 0.0	0.75 0.0	0.16 0.36	-0.1 0.59	0.74 0.0
fr_i				1.0	-0.44 0.01	-0.06 0.73	-0.36 0.03	-0.59 0.0	-0.49 0.0	-0.33 0.05	0.34 0.04	-0.01 0.96	0.09 0.59	0.32 0.06
n					1.0	-0.16 0.35	0.18 0.29	0.16 0.35	0.7 0.0	0.09 0.61	-0.21 0.24	-0.24 0.16	0.26 0.14	-0.22 0.2
λ_r						1.0	0.59 0.0	-0.15 0.38	-0.22 0.2	0.63 0.0	0.8 0.0	0.17 0.32	-0.13 0.46	0.79 0.0
Δ_O							1.0	0.35 0.04	0.41 0.02	0.97 0.0	0.38 0.02	0.01 0.94	-0.11 0.54	0.37 0.03
fo_i								1.0	0.69 0.0	0.32 0.06	-0.38 0.03	0.03 0.86	-0.17 0.34	-0.35 0.04
m									1.0	0.31 0.07	-0.33 0.05	-0.2 0.25	0.06 0.72	-0.34 0.05
λ_o										1.0	0.43 0.01	-0.01 0.95	-0.08 0.66	0.42 0.01
Δ_D											1.0	0.04 0.82	0.02 0.89	0.99 0.0
fp_i												1.0	-0.89 0.0	0.17 0.33
k													1.0	-0.11 0.52
m_{ij}														1.0

diversity scores than others. These results confirmed that the diversity of researchers/organizations was significantly impacted by the distance between their disciplines. However, the number of researchers/organizations also had a small impact on the diversity score. If the aim is to understand the overall diversity of a project, it might be useful to consider all three indicators: Δ_R, Δ_O, and Δ_D as they capture different aspects of diversity.

4.2.4 Evaluation on another dataset

To further evaluate the proposed approach, we implemented it on research projects available on Dimensions data[2]. Dimensions is a large research information database containing millions of records related to research projects and publications. Each publication or project in Dimensions was classified using the Australian and New Zealand Standard Classification (Australian_Bureau_of_Statistics, 2020). For each project, we can extract a list of involved researchers and a list of assigned disciplines. The organizations were not included in the project's metadata; therefore, in this experimental work, we did not calculate the diversity of organizations. For analysis purposes, we selected 1449 projects that contained at least two researchers and at least two disciplines. For each project, we calculated the diversity of researchers (Δ_R), the diversity of assigned disciplines (Δ_D), and other related factors such as frequency, number, distance. The details of the experimental setup and results

[2] https://www.dimensions.ai/

Table 8: Correlation scores between factors on Dimensions data.

	Δ_R	fr_i	n	λ_r	Δ_D	fd_i	k	m_{ij}
Δ_R	1	-0.62	**0.54**	**0.93**	0.09	-0.01	0	0.13
fr_i		1	-0.84	-0.47	0.06	-0.03	0.03	0.05
n			1	0.28	-0.04	0.01	-0.02	-0.05
λ_r				1	0.1	-0.02	0.01	0.15
Δ_D					1	-0.8	**0.8**	**0.73**
fd_i						1	-0.99	-0.18
k							1	0.18
m_{ij}								1

of the IDR identification are not presented here since this experiment aimed to briefly show the correlation between diversity scores and other related factors. The Pearson correlation scores between these factors are illustrated in Table 8. All p-values related to the coefficient scores (in bold) are statistically significant at a level of less than 0.01. As shown in the results, the Δ_R was found to be significantly correlated with λ_r, with coefficient score of 0.93. This indicates that the diversity of researchers was affected by the distance between them. The number of researchers participating in the projects also have an impact on the diversity score. The correlation score between Δ_R and number of researchers was 0.54. Moreover, a significant correlation was found between Δ_D and the number of disciplines as well as the distance between them. The coefficient scores between Δ_D and k and m_{ij} were 0.80 and 0.73, respectively. This implies that the diversity of disciplines was influenced by both the number of disciplines and the distance between them. This result reinforces the conclusion we drew from the FRIS data results, which indicated that the diversity scores were more influenced by the distance between researchers than the number of researchers participating in the project.

5 Conclusion

In this paper, we proposed an approach for identifying IDR in projects available on the RIS based on an organizational approach. In particular, we proposed approaches to calculate the diversity of researchers, the diversity of organizations, and the diversity of disciplines attached to a project. To calculate diversity, different approaches were proposed to 1) calculate the distance matrix, 2) calculate the distance between two researchers or two organizations, and 3) calculate the relevancy between researchers' disciplines and the project's disciplines. These calculations play an important role in the proposed diversity calculation approach. The degree of IDR of a project was evaluated based on the combination of these three indicators. In addition, to identify IDR in projects, we proposed classifying them into appropriate groups based on their diversity scores, such as Low, Medium, and High. We implemented the proposed method on a large number of projects available on FRIS and Dimensions. The results showed that the proposed method could properly classify projects into three levels of IDR. The empirical analysis of findings confirms the suggested approach's assumption that the diversity of research projects increases as the distance between disciplines in the projects grows. The relationship between diversity scores and other factors varies, and certain factors may exert a greater impact on diversity scores than others. To gain a comprehensive understanding of a project's diversity, it could be valuable to take into account all three indicators: Δ_R, Δ_O, Δ_D. Furthermore, the relevancy weight can be incorporated as an additional factor in the measurement of IDR. If the IDR is high and the relevancy weight is low or a substantial number of researchers or organizations are not closely associated with the projects, it is advisable to employ additional measures to validate the accuracy of the IDR calculation within those specific projects.

Although the combination of indicators could give a good indication of the IDR level of a project, the proposed method still has some limitations. First of all, the proposed method can be directly applied to identify IDR in projects that are available on the FRIS portal since this research portal has some advantages (e.g., all objects are labeled with a research field classification, objects have logical relationships, etc.). However, we have shown that it is possible to predict researchers' profiles based on various other factors. This can be particularly useful for RIS systems that do not store researchers' profiles. The second limitation is that the proposed method did not distinguish the knowledge contribution of young (e.g., Ph.D. students) and senior researchers (e.g., promoters) to a project. In practice, senior researchers could potentially contribute more knowledge than young researchers to research activity. As a result, the knowledge contributed to the project could be significantly affected by the disciplines of promoters.

Measuring IDR in projects available on an inter-organizational research portal is challenging. Although the proposed method has been demonstrated to work well on the FRIS and Dimensions databases, future research should consider different approaches. For now, we apply linear regression and neural networks to predict researchers' disciplines without the need to assess the weight values. In the future, we intend to apply optimization techniques such as genetics or meta-heuristics to find the optimal values of the weights in Equation (4). In addition, further studies should be conducted to optimize and evaluate the distance matrix and the relevancy weight in order to improve the IDR measurements. To further evaluate the performance of the proposed method, we plan to evaluate it on various databases as well as compare it to other approaches, e.g., citation-based approach, and text-based approach. In addition, more robust indicators should be investigated to better identify potential IDR. The first focus is taking into account the relationship between authors to distinguish the knowledge contribution of the contributors to the project. With this information, the diversity of researchers can be calculated more precisely. Another research direction is analyzing the diversity and network coherence of concepts embedded in the abstract of the project. Using both diversity and network coherence of concepts would be a possible way to identify potential IDR in projects. Combining these approaches with the proposed organizational approach in this paper could be potentially used to evaluate IDR in practice.

Acknowledgments

This study was supported by The Expertise Center for Research and Development Monitoring (ECOOM), Flanders, Belgium. The authors would like to acknowledge Dimensions for granting access to their database. Further, the authors would like to thank the reviewers for the feedback and comments that they have made to improve the paper.

Conflict of interest

The authors have no competing interests to declare that are relevant to the content of this article.

Supplementary information

- Dataset can be found at: https://zenodo.org/record/8189455
- Source code can be found at: https://github.com/phamsonit/organizational_approach_for_IDR_calculation

References

Abramo, G., D'angelo, C.A., Costa, F. (2017). Do interdisciplinary research teams deliver higher gains to science? *Scientometrics*, *111*(1), 317–336.

Abramo, G., D'Angelo, C.A., Di Costa, F. (2012). Identifying interdisciplinarity through the disciplinary classification of coauthors of scientific publications. *JASIST*, *63*(11), 2206-2222.

Adams, J., Loach, T., Szomszor, M. (2016). *Digital research report: Interdisciplinary research - methodologies for identification and assessment* (Tech. Rep.). 10.6084/M9.FIGSHARE.4270289

Allmendinger, J. (2015). *Quests for interdisciplinarity: a challenge for the era and horizon 2020.* Retrieved from https://ec.europa.eu/

Australian_Bureau_of_Statistics (2020). Australian and new zealand standard research classification (anzsrc). https://arxiv.org/abs/https://www.abs.gov.au/statistics/classifications/australian-and-new-zealand-standard-research-classification-anzsrc/latest-release.

Ba, Z., Cao, Y., Mao, J., Li, G. (2019). A hierarchical approach to analyzing knowledge integration between two fields–a case study on medical informatics and computer science. *Scientometrics*, *119*(3), 1455–1486.

Bonaccorsi, A., Melluso, N., Massucci, F.A. (2021). Detecting interdisciplinarity in top-class research using topic modeling. *Issi2021* (Vol. 42, pp. 169–160). Heidelberg: Springer.

Cassi, L., Mescheba, W., De Turckheim, E. (2014). How to evaluate the degree of interdisciplinarity of an institution. *Scientometrics*, *101*(3), 1871-1895.

Choi, B.C.K., & Pak, A.W.P. (2006). Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: 1. definitions, objectives, and evidence of effectiveness. *Clinical and Investigative Medicine*, *29*(6), 351–364.

Glänzel, W. (2021, 09). Various aspects of interdisciplinarity in research and how to quantify and measure those. *Scientometrics*.

Glänzel, W., & Schubert, A. (1988). Characteristic scores and scales in assessing citation impact. *Journal of Information Science*, *14*(2), 123-127.

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

Han, J., Pei, J., Yin, Y. (2000). Mining frequent patterns without candidate generation. *Proceedings of the 2000 acm sigmod international conference on management of data* (p. 1–12). Association for Computing Machinery.

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the isi subject categories. *Journal of the American Society for Information Science and Technology*, *60*(2), 348-362.

Nichols, L.G. (2014). A topic model approach to measuring interdisciplinarity at the national science foundation. *Scientometrics*, *100*(3), 741–754.

NSF (2005). *What is interdisciplinary research?* Retrieved from https://www.nsf.gov/od/oia/additional_resources/interdisciplinary_research/definition.jsp

OECD (Ed.). (2015). *Frascati manual 2015.* OECD. 10.1787/9789264239012-en

Olkin, I., & Pukelsheim, F. (1982). The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, *48*, 257-263.

Ortiz-Burgos, S. (2016). Shannon-weaver diversity index. In M.J. Kennish (Ed.), *Encyclopedia of estuaries* (pp. 572–573). Dordrecht: Springer Netherlands.

Porter, A.L., Cohen, A.S., Roessner, J.D., Perreaul, M. (2007). Measuring researcher interdisciplinarity. *Scientometrics*, *72*(1), 117-147.

Porter, A.L., & Rafols, I. (2009). Is science becoming more interdisciplinary? measuring and mapping six research fields over time. *Scientometrics*, *8*(3), 719-745.

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, *8*, 263–287.

Rousseau, R., Zhang, L., Hu, X. (2019). Knowledge integration: Its meaning and measurement. In *Springer handbook of science and technology indicators* (pp. 69–94).

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513-523.

Simpson, E. (1949). Measurement of diversity. *Nature*, *163*.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface*, *4*(15), 707-719.

Thijs, B., Huang, Y., Glänzel, W. (2021). Comparing different implementations of similarity for disparity measures in studies on interdisciplinarity. *Issi2021* (Vol. 42). Heidelberg: Springer.

Vancauwenbergh, S., & Poelmans, H. (2019). The creation of the flemish research discipline list, an important step forward in harmonising research information. *Procedia Computer Science*, *146*, 265-278.

Wang, J., Thijs, B., Glänzel, W. (2015). Interdisciplinarity and impact: Distinct effects of variety, balance, and disparity. *PLOS ONE*, *10*(5), 1-18.

Wang, Q., & Schneider, J.W. (2018). Consistency of interdisciplinarity measures. *CoRR*, *abs/1810.00577*. Retrieved from http://arxiv.org/abs/1810.00577 https://arxiv.org/abs/1810.00577

Wernli, D., & Darbellay, F. (2016). *Interdisciplinarity and the 21st century research-intensive university.* (source at www.leru.org)

Xu, H., Guo, T., Yue, Z., Ru, L., Fang, S. (2016). Interdisciplinary topics of information science: A study based on the terms interdisciplinarity index series. *Scientometrics*, *106*(2), 583–601.

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. *JASIST*, *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*.