Interdisciplinarity beyond bibliometrics - (in)validation of website information as an indication of interdisciplinarity

Background

The National Academies (2005) defined interdisciplinary research as a "...mode of research by teams or individuals that integrates information, data, techniques, tools, perspectives, concepts, and/or theories from two or more disciplines..." (p.2). In sum, interdisciplinarity is the integration and use of different aspects - competences, perspectives, methods, knowledge - from different disciplines. The main task of the project underlying this presentation at GTM2019 is to measure and assess the level of interdisciplinarity of research units - universities, research institutes, or even groups within them - beyond bibliometrics.

Based on Stirling's (2007) conceptualization of diversity, we expect interdisciplinarity as being a multi-dimensional concept. We thereby follow similar applications of Stirling's approach e.g. by Wang et al. (2015) or by Roessner et al. (2009), but aim to go beyond these applications in the context of interdisciplinarity as they mainly used bibliometric data (Wang et al.) or bibliometric and qualitative data (Roessner et al.). The use of bibliometric data in the analysis of interdisciplinarity is widespread (Mugabushaka et al. 2016; Nagaoka/Kwon 2006; Shafique 2010; Small 2010; Besselar und Heimericks 2001; Steele und Stier 2000; Rinia et al. 2001). However, we intend to go beyond bibliometric data only and aim at using additional data sources (patents, annual reports, project databases, or websites).

The concrete research agenda of this presentation is the extraction and analysis of indications of interdisciplinarity from websites by all Fraunhofer institutes. Tests of usability and reliability as well as a (in)validation of the newly extracted indicators will be conducted based on existing indicators of interdisciplinarity.

Methods

In a previous study we were able to use internal data by Fraunhofer - e.g. on the disciplinary structure of the research staff, on the research partners in public research projects, on patent applications, and on the scientific publications - for the assessment of the level of interdiciplinarity of about 70 Fraunhofer institutes. The latter indicator group of bibliometric data covered 1) the average number of fields per publication, 2) the share of citations from publications outside the papers' disciplines/research fields, 3) shares of references to papers from other disciplines/research fields, and 4) the heterogeneity (Herfindahl-Hirschman index) of the reference lists. We correlated all indicators to check for distinction and overlap. In addition, we conducted a factor analysis to extract distinct latent factors of interdisciplinarity based on the common variance of the sometimes highly correlated individual variables.

What is new to the presentation suggested here is to use text data for the identification of interdisciplinarity and assess the intensity of the interdisciplinarity of each institute. For this purpose, we scraped the websites of all Fraunhofer institutes. This unstructured data will be

analysed using text mining approaches. We start with simple keyword searches to identify relevant parts of the text and to assess the meaningfulness of the dataset as such. In a second step, we use TF-IDF for the identification of relevant and discriminating terms referring to interdisciplinarity. By this analytical step, we intend to identify the level of self-description or self-ascription of interdisciplinarity by the institutes. In a third analytical step we will conduct a topic modelling approach to count the mentioning of distinct topics on the institutes' websites. At a later stage, we intend to broaden the data sources and include websites from other research institutes in Germany (e.g. Max-Planck and some university institutes). However, as we cannot assume that the Fraunhofer websites have a sufficiently broad coverage of scientific and technological fields, we need to be careful with the interpretation of the set of 'within'-topics. To overcome this potential shortcoming, we might resort to a list of about 400 topics that were extracted from publications and patents in all FP7 projects. This dataset is available to us from another project. We then simply count the occurrences of distinct research fields from this list of 400 topics.

In sum, we expect to extract four different kinds of indicators based on the websites of Fraunhofer institutes. The indicators will mainly be counts, but we might also check concentration and distribution measures (Herfindahl-Hirschman, TF-IDF directly). In a final step, we will correlate these new indicators with the above-mentioned indicators stemming from the previous project, thereby (in)validating the text-based indicators of the interdisciplinarity of research institutes and assessing its additionality or overlap mainly with broadly used bibliometric indicators of interdisciplinarity.

First Results

The results of the previous study suggest a rather disperse distribution of interdisciplinarity in the group of about 70 Fraunhofer institutes. The factor analysis revealed three different latent factors that we interpret as three dimensions of interdisciplinarity. First, the collaboration and exchange pattern (based on the data of collaboration partners in public research projects) was extracted. The second factor highly loaded on the bibliometric indicators, which we interpreted as capacities of interdisciplinary knowledge creation and absorption. The third factor reflects the field/technological/disciplinary structure of the research unit, referring to the average number of fields per publication, per patent, or per research project.

Outlook

We expect our analyses mainly to be of conceptual/methodological value. The fact that we have in-depth data access and insights into Fraunhofer offers enormous potential for the identification of several dimensions of interdisciplinarity. As such data is not easily available and maybe even not available at all for most research units worldwide, the assessment of their interdisciplinarity might be a huge challenge in most cases. However, if we are able to show that text-based - websites or annual reports are more easily accessible - approaches substitute (or supplement) some of the indicators, we would be able to generalize our analytical approach and assess the level of interdisciplinarity of any research unit/entity worldwide.