Is Interdisciplinarity a relevant criterion for the selection of project proposals by funding agencies?

A case study implementing a text-mining-based indicator.

Dominique Besagni

Ivana Roche
Claire François

Marianne Hörlesberger
Edgar Schiebel

Background

- A European project:

  Development and verification of a Bibliometric model for the Identification of Frontier Research
- Coordination and Support Action (CSA) for the European Research Council (ERC)
- Its goal was to infer attributes of frontier research in peer-reviewed research project proposals.
- Identification of 4 key attributes:
  - Novelty
  - Risk
  - Applicability
  - Interdisciplinarity

Interdisciplinarity indicator

- "*… it pursues questions irrespective of established disciplinary boundaries, involves multi-, inter- or trans-disciplinary research that brings together researchers from different disciplinary backgrounds, with different theoretical and conceptual approaches, techniques, methodologies and instrumentation, perhaps even different goals and motivations*", EC's High Level Expert Group report (2005)

- Hypothesis:
  - the higher the occurrence in a proposal indexing of keywords belonging to different domains, the more interdisciplinary that proposal is considered
- Calculation:
  - keywords labeling according to their statistical frequency of occurrence across all domains
  - assessment of the concentration of keywords labeled as belonging to different domains

Interdisciplinarity indicator

- We applied our methodology to a case study coming from project proposals submitted to the ERC 2009 Starting Grant Call.

	ERC StG 2009	Dataset (6 panels)
Proposals	2,503	198
Successful	244	41
Non-successful	2,259	157

- Among the 19 ERC panels representing Life Sciences (LS) and Mathematics, Physics, Chemistry, Engineering & Earth Sciences (PE), we chose 6 panels with a balance between LS and PE as well as between basic and applied fields.

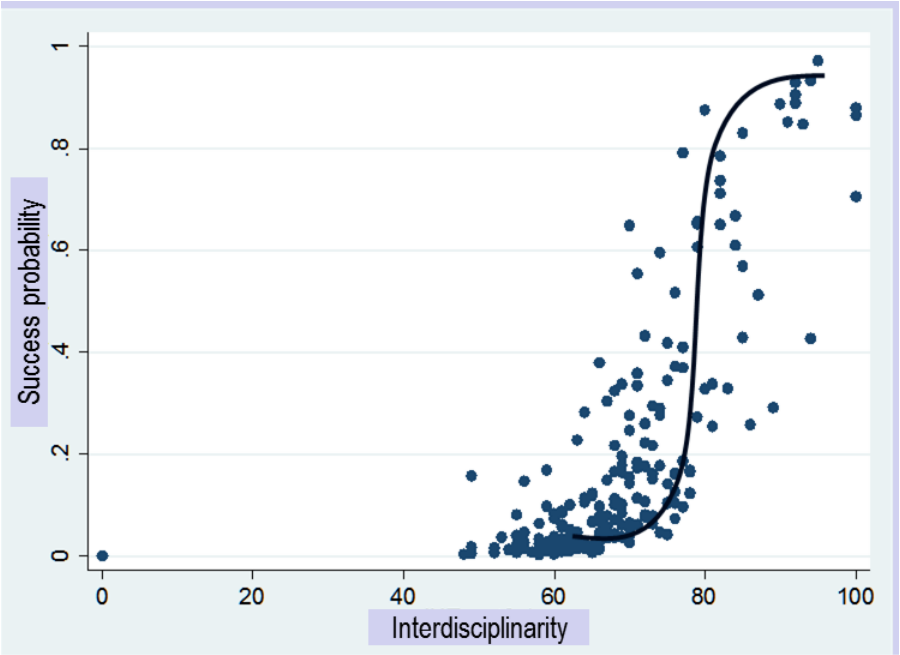
- The table on the right shows the values for ERC panel PE1 ("Mathematics and mathematical foundations"). The successful proposals are highlighted in blue.

Proposal ID	Interdisciplinarity
239885	95
239694	94
239807	94
240518	93
239748	92
239781	92
240123	92
239983	87
239784	85
240127	85
239959	82
239870	81
239902	76
240428	76
240269	73
240074	72
240223	71
240053	70
240471	70
240157	69
239800	68
239814	68
239952	67
240121	67
240693	67
240008	66
240192	66
240683	66
239737	65
240014	65
239607	64
239769	62
240265	61
239853	60
239929	60
240416	58
240459	55
240633	53
240201	49

Interdisciplinarity indicator

We used a statistical discrete choice model (DCM) to estimate the decision probability for a proposal to be accepted on the basis of measured attributes of "frontier research" and conducted an initial analysis of the ex-post comparison between the indicator-based scientometric evaluation and the empirical peer-review process.

The figure on the right shows the relation between the value of Interdisciplinarity and the success probability of proposals predicted by the DCM for the whole dataset. The indicator fits the theoretical logistic curve, as confirmed by statistical tests.



Aim of the study

Having defined and used that indicator successfully, we wanted to see if we could apply the same principle to a different set of project proposals:

- from the e-Corda (External COmmon Research DAtabase) database produced by the EC, *collecting information related to all project proposals submitted for grant at a project Call published in the 7th FP (2007-2013)*,
- where the content of each proposal is represented by "keywords" identified by text-mining tools.

Methodology (1)

We used the approach of the diffusion model where the diffusion degree of each keyword is obtained by applying a statistical filtering to identify terms describing a domain specificity.

We selected a set of project calls having a common and easily identifiable theme (i.e. “Health” or “ICT”). Each theme is a domain, or “home field”, of the diffusion model.

Each proposal is assigned to the “home field” corresponding to the theme of the project call where it came from.

Using the extracting module of the BibTechMon tool on the corpus of project proposals of the selected calls, we obtained noun groups that we use as keywords.

Methodology (2)

The raw data are cleansed to eliminate non-pertinent strings (as punctuation marks, numbers, XML tags, etc.) and to homogenize under the same form the different variants of a keyword (e.g. plural to singular form).

The "cleansed" keywords then were assigned to a "home field" in function of the relative frequency of their occurrence in the different "home fields" (by the way of the proposals).

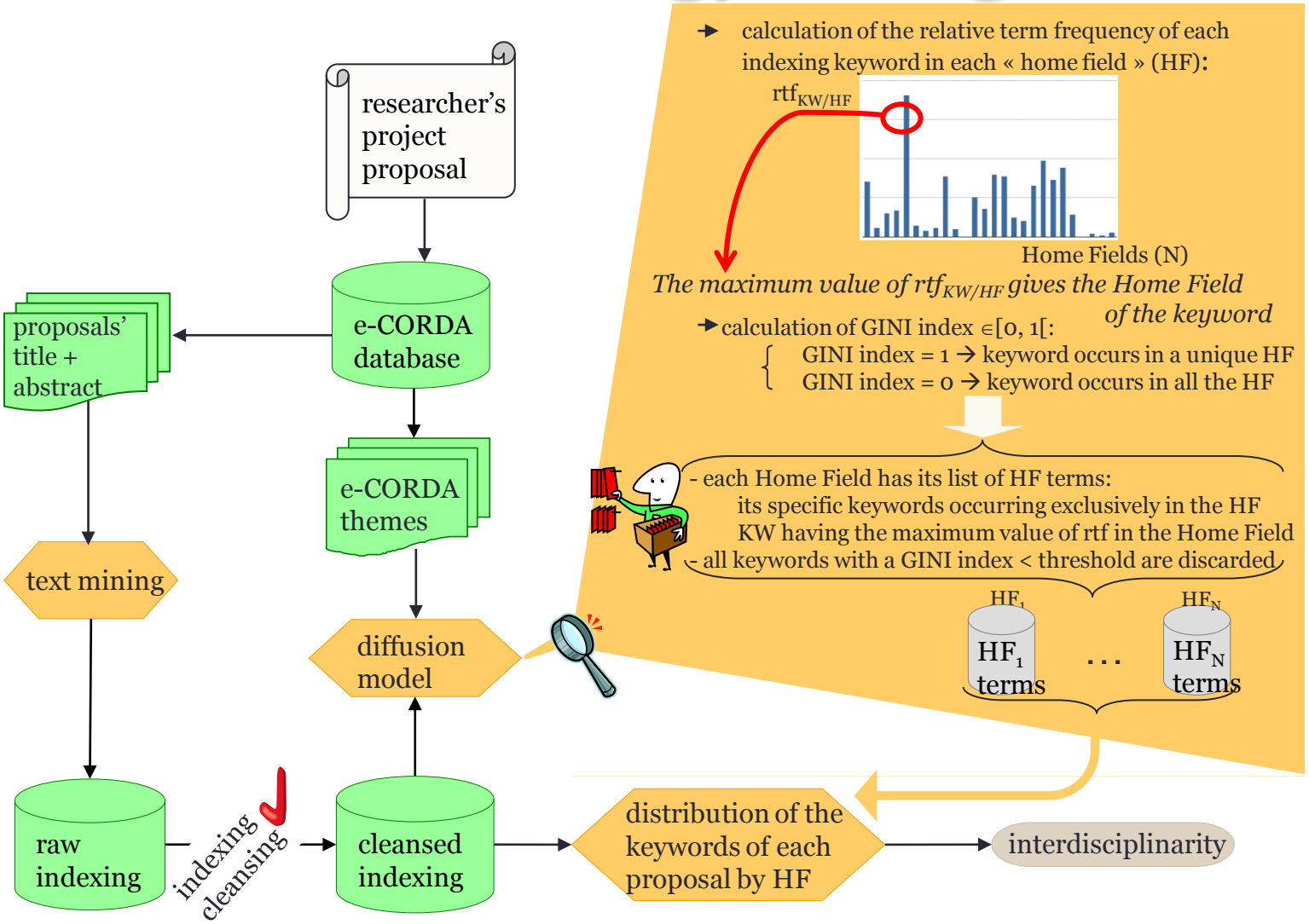
We calculated for each keyword its Gini index to weed out all the keywords that were far too widespread.

Then for each proposal, we calculated the interdisciplinarity indicator as the share of keywords belonging to a "home field" different from the proposal's own "home field".

It is a value:

- from 0: *all keywords representing the proposal content come from the "home field" of the proposal*,
- to 1: *all keywords representing the proposal content come from "home fields" different of the proposal's own "home field"*.

The methodology at a glance



Data source (1)

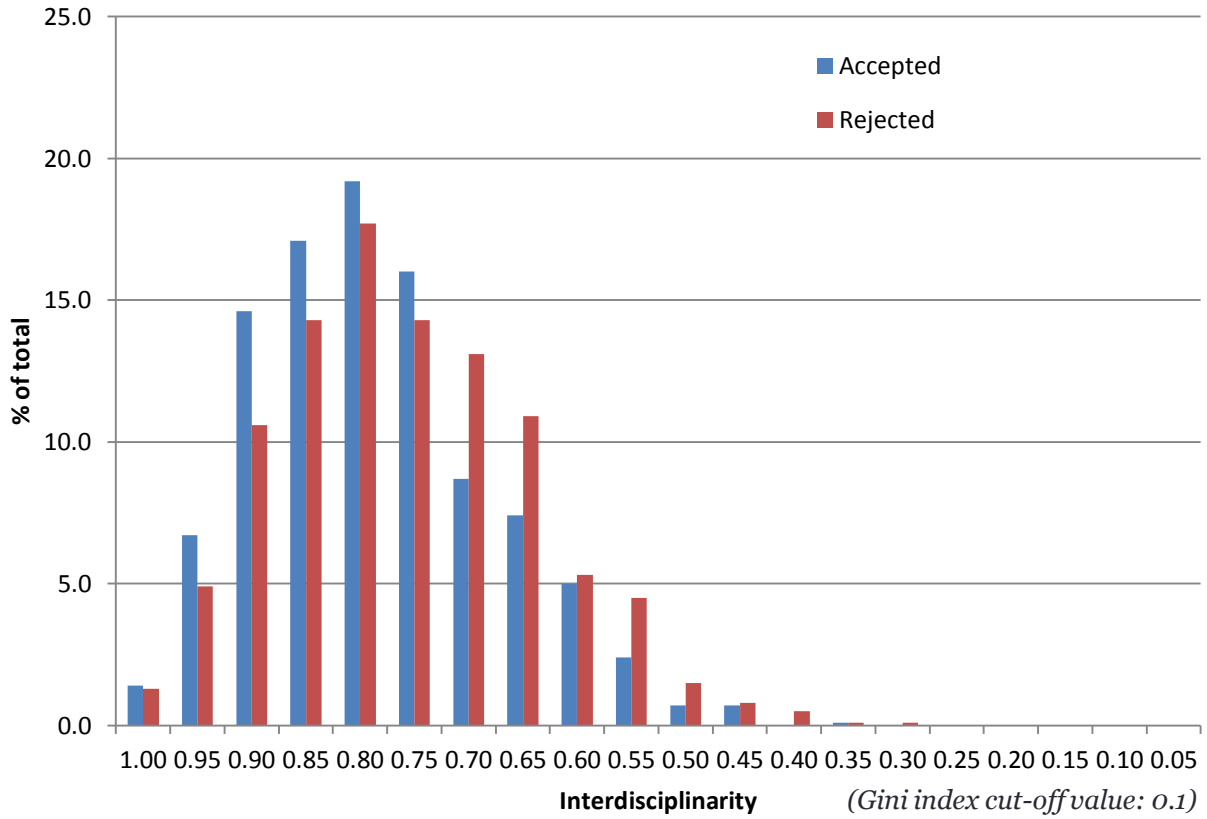
- Database:
  - e-CORDA 2007-2011
  - 327 project calls
  - 102,688 records
- Filtering:
  - Reject of general, regional or heterogeneous calls, as well as ERC calls
- Corpus:
  - 170 project calls
  - 34,739 records (33,549 eligible)
  - 11 themes

Data source (2)

	All	Accepted	Rejected
Energy	1690	436	1254
Environment	2377	476	1901
Food & biotechnology	2336	1349	987
Health	5308	948	4360
ICT	10269	1816	8453
Nanosciences	4845	807	4038
Nuclear technologies	216	143	73
Security	1341	292	1049
Social sciences & humanities	1879	254	1625
Space	687	319	368
Transport	2601	891	1710
Total	33549	7731	25818

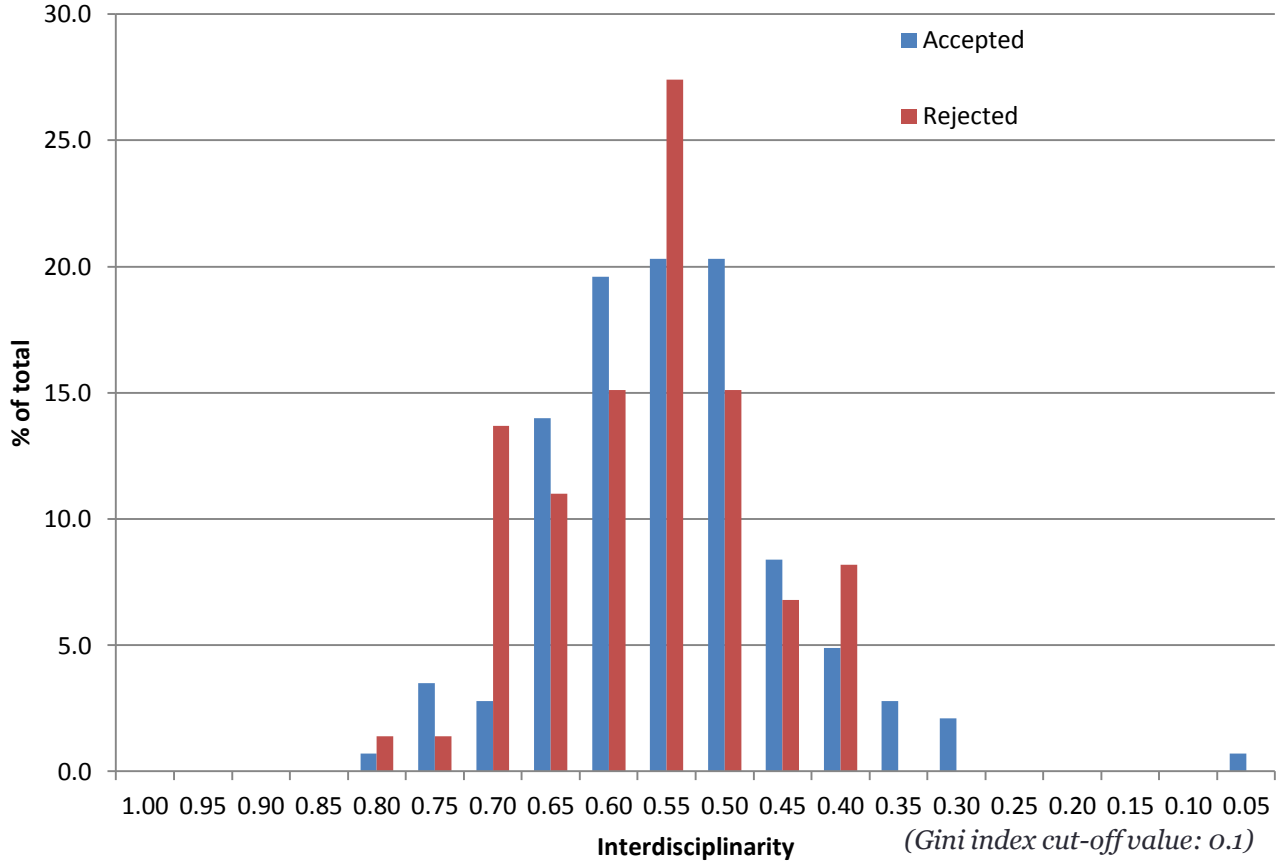
Results (1)

Distribution of accepted/rejected proposals in function of their interdisciplinarity for the domain "**Nanosciences**".



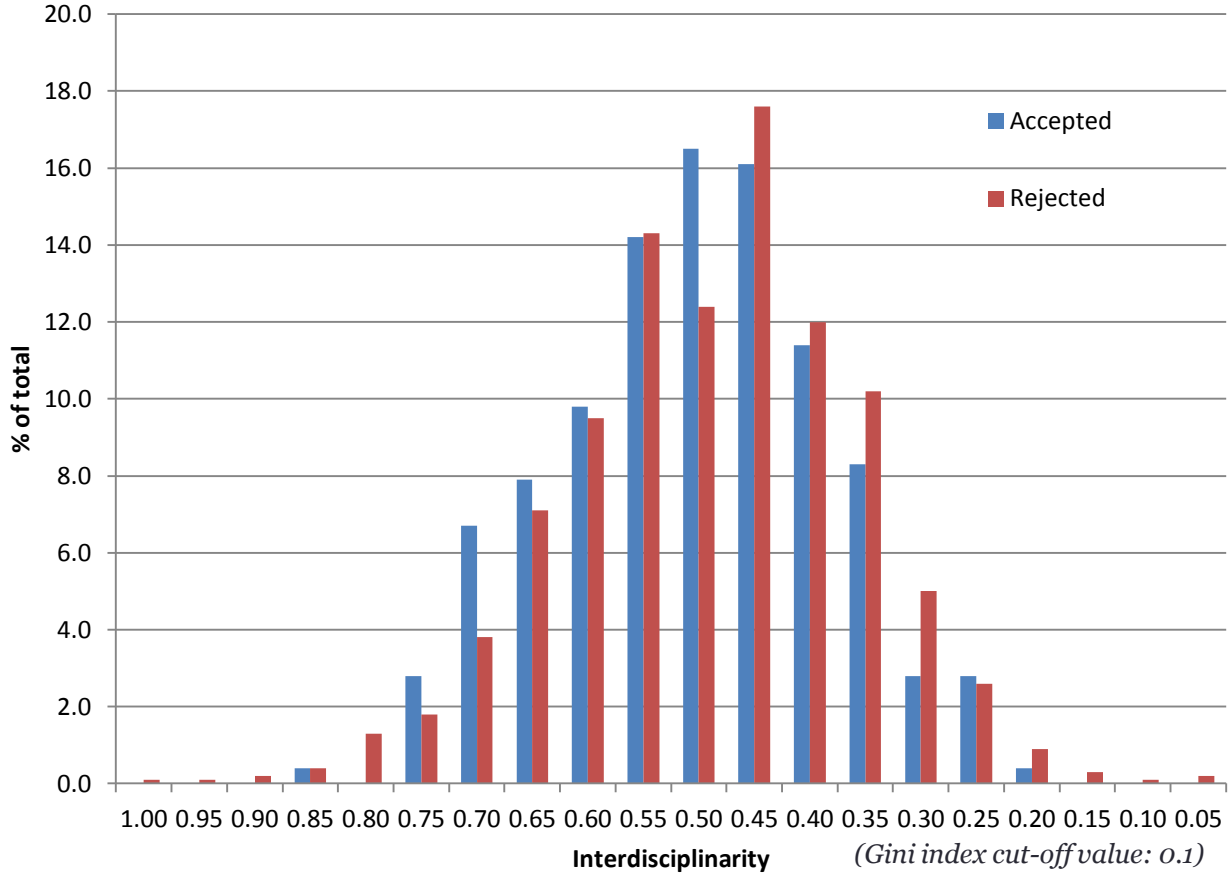
Results (2)

Distribution of accepted/rejected proposals in function of their interdisciplinarity for the domain "**Nuclear technologies**"



Results (3)

Distribution of accepted/rejected proposals in function of their interdisciplinarity for the domain "**Social sciences & humanities**"



(Gini index cut-off value: 0.1)

Conclusion

We developed an indicator:

- based on content analysis,
- to categorize project proposals,
- without scientific expertise.

We used a text-mining technique to extract noun groups that represent the content of each proposal ⇒ keywords.

For most domains, results show that the more interdisciplinary a proposal is, the more likely it is to be accepted, but some domains do not follow this pattern.

But after that first experiment, we have mostly questions:

- Was it the right set of data?
- How to improve the results, specially with NLP techniques?
- Is there an added value?

Acknowledgements

This work originated from the DBF (Development and Verification of a Bibliometric Model for the Identification of Frontier Research - http://www.ait.ac.at/dbf) project within the CSA of the Ideas specific Programme of the EU's 7th Framework Programme

This work is made possible with the provision of data coming from e-CORDA database produced by EC (https://webgate.ec.europa.eu/e-corda/)

Thank you!

[ivana.roche; dominique.besagni; claire.francois]@inist.fr
[marianne.horlesberger; edgar.schiebel]@ait.ac.at