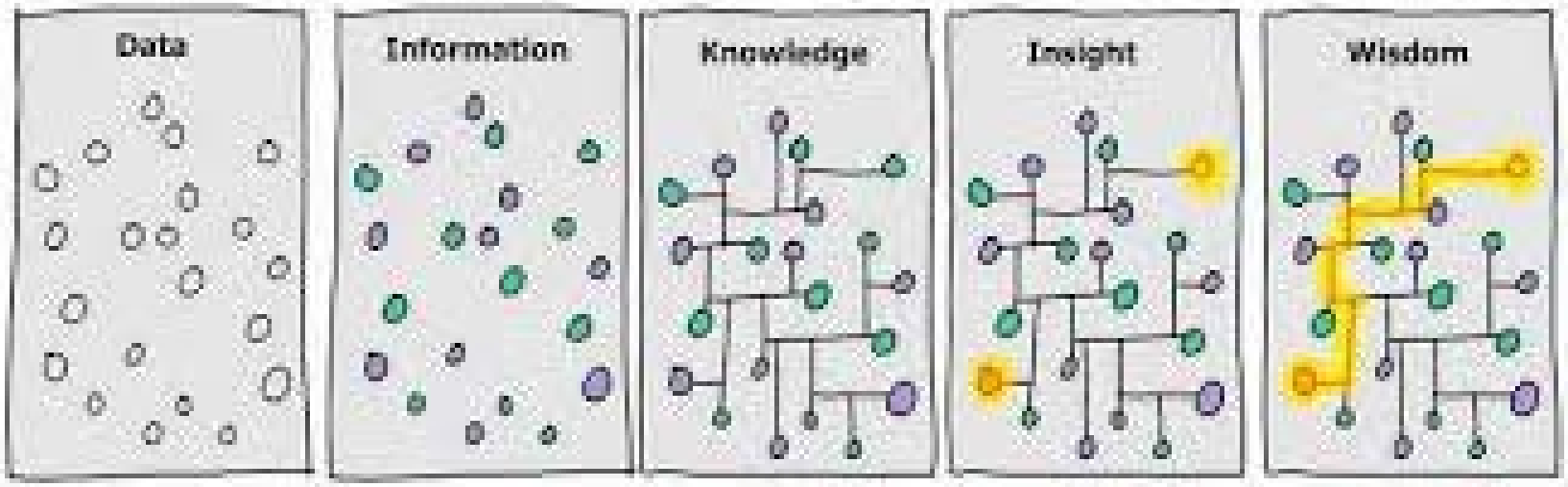
INTERDISCIPLINARITY BEYOND BIBLIOMETRICS - (IN)VALIDATION OF WEBSITE INFORMATION AS AN INDICATION OF INTERDISCIPLINARITY

Rainer Frietsch, Peter Neuhäusler, Alexander Feidenheimer, Andrea Zielinski and Antje Schimke



[Cartoon by David Somerville, based on a two pane version by Hugh McLeod.]

Fraunhofer ISI

CONTEXT AND RESEARCH QUESTION

Overall aim

- The main aim is to find (additional) data and information to measure **interdisciplinarity beyond bibliometrics**
- Demand by management board to have **KPIs**
  - „What you cannot measure, you cannot manage“
  - Provide such a KPI at the level of **research groups/institutes**
  - Measure not only the occurrence/volume of interdisciplinarity (yes/no; shares of total), but also assess the **intensity**
- Assess interdisciplinarity not only at the **output/outcome** side (e.g. by publications), but also at the **input side** -> generation versus dissemination

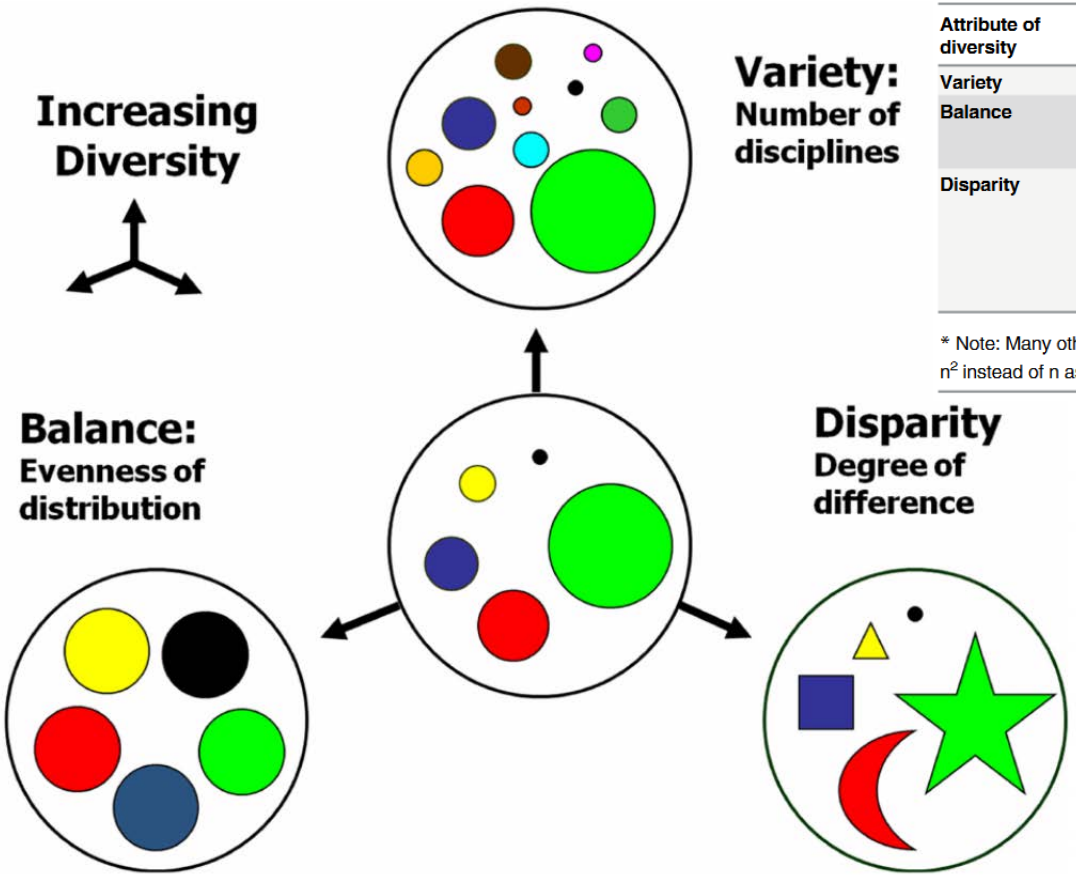
Literature Review - Definitions

- Interdisciplinarity is "bringing together and applying different aspects from several disciplines" (van den Besselar/Heimericks 2001)
  - application oriented
  - differing from the „norm"(=specialties)
- Different to multidisciplinarity (different perspectives on the same topic taken by different disciplines) or transdisciplinarity (basic research questions that cannot uniquely assigned to a single discipline) (Alvargonzález, 2011)

- National Academies (2005) define 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..." (cited after Wang, Thijs, Glänzel, 2015)

- Wang et al (2015): „…**easy to intuit but difficult to define or measure**"

Conceptual basis: Rao-Stirling-Diversity



Attribute of diversity	Operationalisation*
Variety	We use the number of distinctive WoS categories (n) cited in an article.
Balance	We use Shannon diversity (H) normalised by variety (n), where p_i is the proportion of references in WoS category i: Balance $= -\frac{1}{\ln(n)} \sum_i p_i \ln p_i$
Disparity	We use a measure of disparity is based on the average cognitive distance between WoS categories within the reference list. The cognitive distance between two disciplines is calculated as $d_{ij} = 1 - s_{ij}$ with s_{ij} being the cosine similarity between each pair of disciplines i and j. The sum is over disciplines with at least one cited reference: Disparity $= \frac{1}{n(n-1)} \sum_{ij} d_{ij}$

* Note: Many other operationalisations of these properties are possible. For example, we could have taken n^2 instead of n as variety, or the median disparity rather than the mean disparity of a reference set.

- Variety: how many types do we have
- Balance: how much of each type do we have
- Disparity: how different are the types from each other

Source: Yegros-Yegros, A., Rafols, I., D'Este, P., 2015. Does Interdisciplinary Research Lead to Higher Citation Impact? The Different Effect of Proximal and Distal Interdisciplinarity. Plos One 10 (8).

Proposed measures of interdisciplinarity

- Bibliometric indicators
- Patent indicators
- Indicators based on funding data
- Indicators based on employment (Fraunhofer internal)

Correlation & factor analyses

„Interdisciplinarity Index“ (composite indicator)

Proposed measures of interdisciplinarity

- Bibliometric indicators
- Patent indicators
- Indicators based on funding data
- Indicators based on employment (Fraunhofer internal)
- **Topic model of publications**
- **Topic model of websites**

Correlation & factor analyses

„Interdisciplinarity Index“ (composite indicator)

Factor Analysis & Correlation of factors

Variable	Factor 1	Factor 2	Factor 3	Uniqueness
Avg. nr. of fields (B1)	-0.2729		0.2536	0.837
Share foreign field cit (B3)		0.706		0.480
Share ref. to foreign fields (B4)		0.647		0.575
Heterogeneity pub (B5)		0.604		0.627
Spread IPC (4-digit) (P1)			0.836	0.272
Spread NACE (Top5 exkl.) (P2)		-0.360	0.621	0.458
Avg. nr. project partners (Fkat1)	1.151			-0.338
Avg. nr. project partners pub. (Fkat2)	0.369	-0.489	-0.215	0.579
Avg. nr. project partners priv. (Fkat3)	0.724	0.284		0.392
Avg. nr. LPS classes (Fkat4)			-0.288	0.883
Avg. nr. of staff across study areas (F2)			0.252	0.925
Spread of staff across study areas (F4)	-0.265		0.424	0.727

- **Three factors**
  1. Cooperation behavior
  2. Knowledge generation and –dissemination
  3. Discipline and field structures (technologies, sectors, study areas)

Note: Factor analysis based on »iterated principal factors«, values after varimax-rotation, values below +/- 0,2 are not shown. Source: Scopus, PATSTAT, BvD - ORBIS, BMBF funding catalogue, Fraunhofer internal data, calculations by Fraunhofer ISI.

Fraunhofer ISI

Insights so far...

- ... interdiciplinarity is a **multidimensional** subject
- ... we need a measure (**KPI**) that reflects this multidimensionality
- ... interdisciplinarity of what?
  - ... **institutes** or **research groups/departments** versus papers/disciplines
    - demarcation and aggregation problem
  - ... input/generation versus output/dissemination
- ... we aim not only for **volume** (variety) and **concentration** (balance), but also for assessing the **intensity** (disparity)

⇒ Topic modelling of publications a) for benchmarking purposes and b) to develop a measure of intensity (disparity)

⇒ Topic modelling of websites a) for generating an additional and "publicly available indicator and b) to develop a measure of intensity (disparity)

TOPIC MODELLING OF PUBLICATIONS

Topic modeling of Fraunhofer Publications (2016-2018)

- Cleaning & Preprocessing:
  - Select title and abstract of Fraunhofer's English publications (2016-2018)
  - Set all characters to lower
  - remove punctuations & digits
  - remove English stopwords
  - remove words of length shorter than 3 characters
  - apply lemmatization using Mechura (2016) list (https://michmech.github.io/) in R's 'lexicon' package
  - additionally create bi- & trigrams

- Document-Term-Matrix: 13,310 docs x 61,912 terms

Fraunhofer ISI

Topic modeling of Fraunhofer Publications (2016-2018)

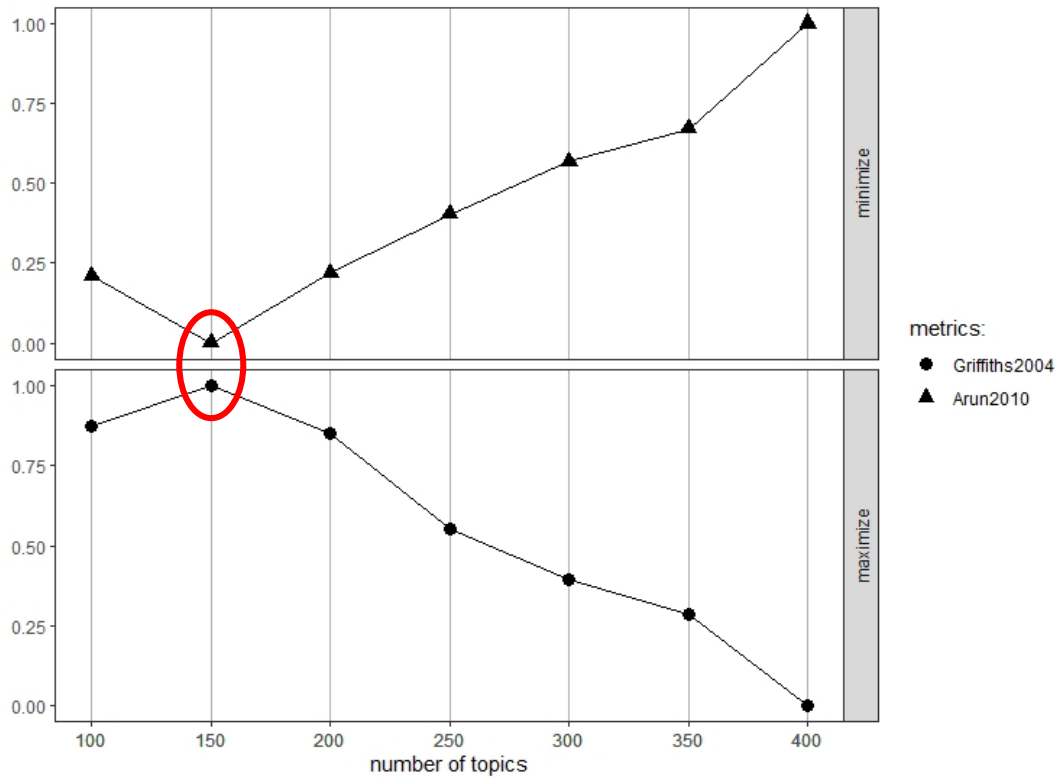
- Model specification:
  - Latent Dirichlet allocation (LDA) using Gibbs sampling approach in R's 'topicmodels' package (Hornik and Grün 2011)
  - Iterations = 500
  - burnin = 480 (first 480 are discarded, as the predictions gets better with each iteration)
  - keep = 10 (the log-likelihood values are stored every keep iteration)

- find optimal number of topics using R's 'ldatuning' package

How to select the optimal number of topics

- Griffiths (2004): Based on <u>maximising</u> the likelihood
- Arun (2010): Based on <u>minimising</u> Kullback-Leibler divergence

→ Both metrics suggest an optimal number of topics around 150



Topic modeling of Fraunhofer Publications (2016-2018)

Aggregation of results on institute level: an example

- An aggregation of document-topic probabilities by Institute

Fraunhofer Institute for Systems and Innovation Research (Fraunhofer ISI)

Topic	ISI
115: innovation technology policy	48.65
68: generation flexibility energy_storage	26.77
69: road traffic automotive	12.71
132: paper current issue	8.192
60: sustainability recycle raw	5.509
127: model simulation result	2.381
149: propose value perform	2.297
148: production process manufacture	2.054
138: scientific management seps	1.281
56: secure sensitive privacy_preserve	1.213

Fraunhofer Institute of Optronics, System Technologies, and Image Exploitation (Fraunhofer IOSB)

Topic	IOSB
47: result detect accuracy	47.1489
37: image camera resolution	30.16104
91: service architecture smart	18.17851
142: turbulence adaptive optical	14.89642
127: model simulation result	9.061806
98: provide visual track	8.681775
45: analytics analyze technique	8.628316
49: radar target system	7.213561
148: production process manufacture	7.04657
84: scheme solution constraint	6.961542

Note: Only document-topic probabilities higher then .05 are considered. This means that at least 5% of the document must be described by the topic

Topic modeling of Fraunhofer Publications (2016-2018)

- Aggregation of results on institute level: an example

Main topic of Fraunhofer ISI

Institute	115: innovation technology policy
ISI	48.65018703
FOKUS	11.19490711
IAO	6.918144918
IMW	6.653945398
FIT	6.565905483
IPK	3.843405521
Center	2.509943231
ISE	2.260740947
IPA	2.083519473
IML	1.923107279

Main topic of Fraunhofer IOSB

Institute	47: result detect accuracy
IOSB	47.14889832
IGD	15.01740801
HHI	12.25904758
IIS	7.851103626
MEVIS	6.554158609
FKIE	6.297668294
IAIS	5.764062824
IDMT	5.003685466
IPK	3.060924242
Center	2.66254539

→ Potential collaboration partners within Fraunhofer Society?

Fraunhofer ISI

Topic modeling: next steps

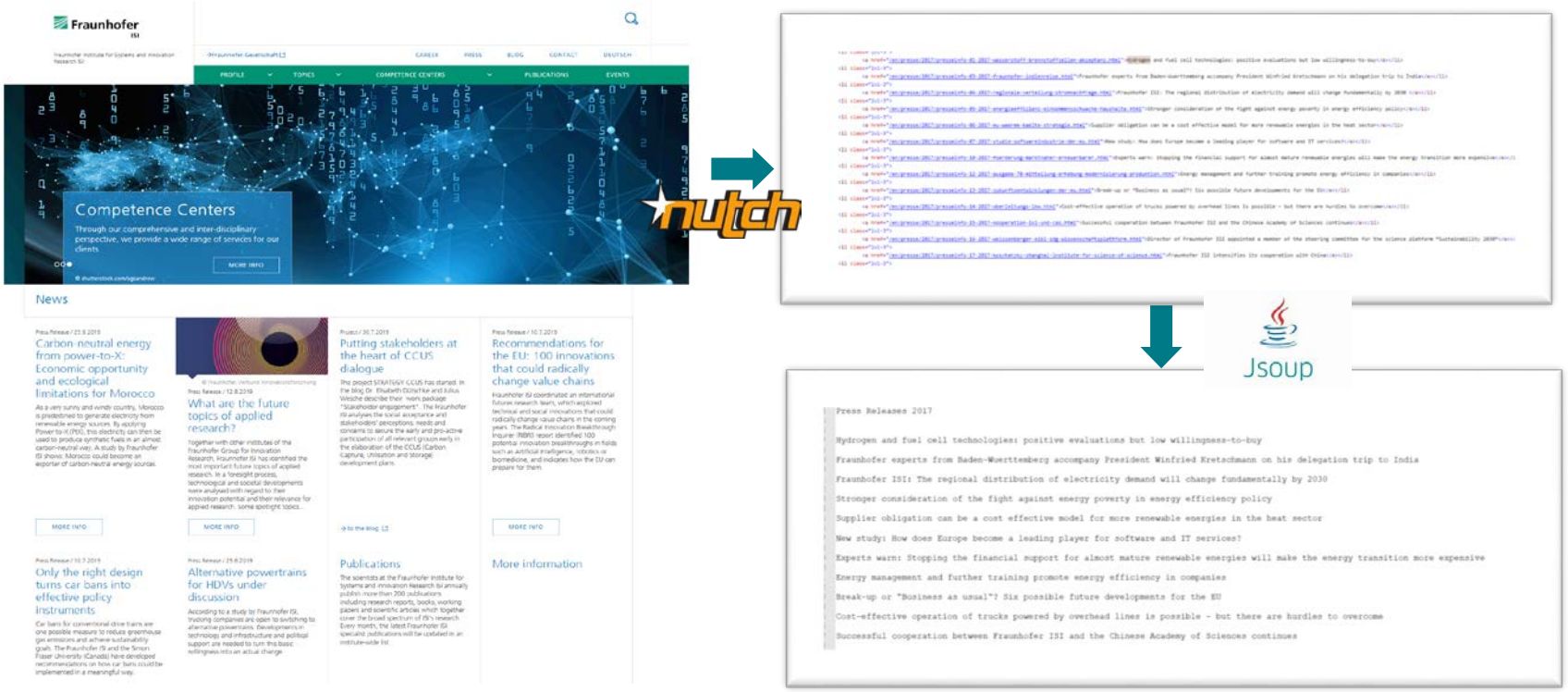
- Train a model on a set of publications worldwide
- Build a topic x topic co-occurrences matrix as benchmark
- Apply the model to Fraunhofer publications and build a second topic x topic co-occurrences matrix
- compare both and calculate similarity(?)

TOPIC MODELLING OF WEBSITES

Reseach Questions

- Is it possible to compare the interdisciplinarity of institutions based on their websites?
- Is it possible to extract a set of meaningful topics from Institutes websites?
  - Problem: Sparse textual data
  - Needs manual inspection of clusters
- How to categorize the discrete disciplines?
  - Classification system of Web of Science, etc.?
  - Needs a research category classifier
- What is the best measure for interdiciplinarity based on topic models?
  - Needs adaptation of diversity measure for topic models, e.g. Stirling's measure to model different aspects of interdisciplinarity

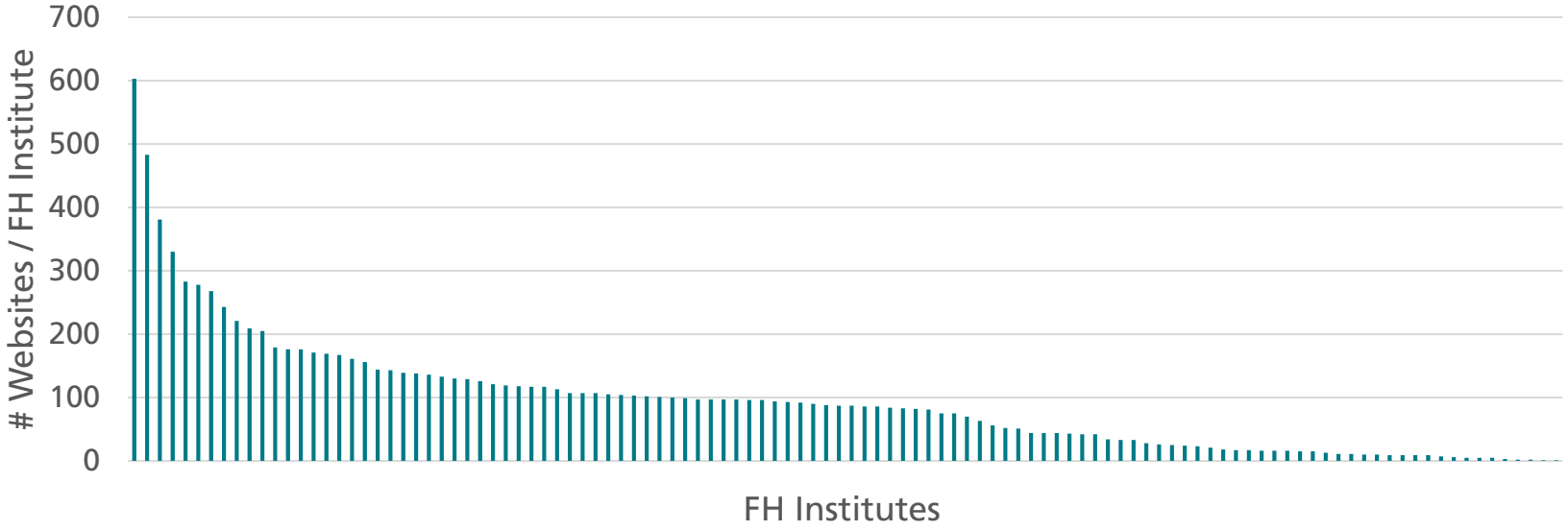
Identification of relevant Websites and Extraction of Textual Content



Fraunhofer ISI

Training Data - Corpus Statistics

- Scraped Fraunhofer Websites: approx. 10K URLs



Fraunhofer ISI

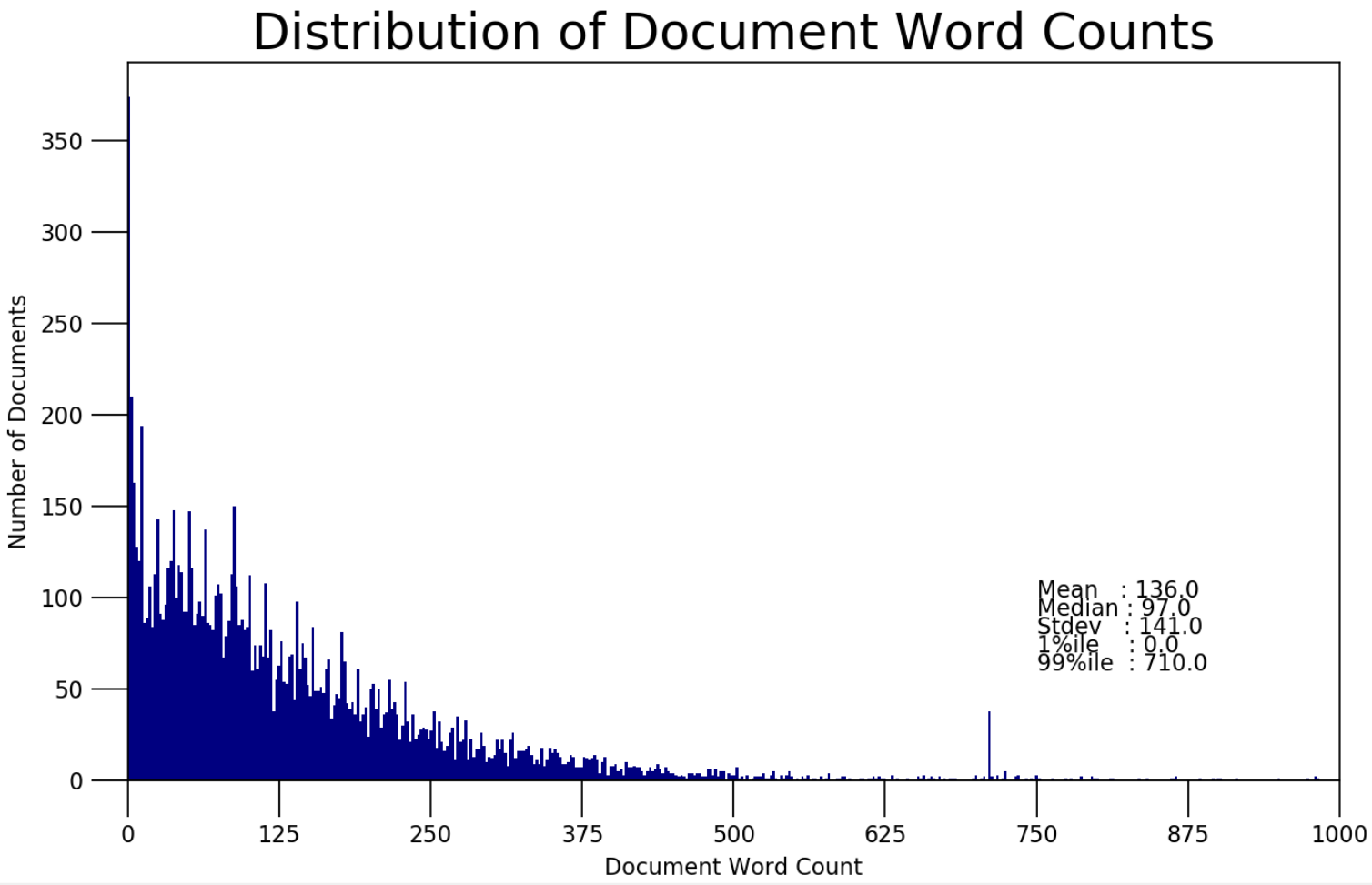
Training Data - Corpus Statistics

- Snapshot of the Data (approx. 23 MB)

10276	https://www.wki.fraunhofer.de/en/departments/zeluba/profile	Profile - Fraunhofer WKI . Where am I? . Profile . In the "Center for Light and Environmentally-Friendly Structures ZELUBA", we develop solutions for the construc
10277	https://www.wki.fraunhofer.de/en/departments/zeluba/profile/bmel-fnr-junior-re	Junior research group - Fraunhofer WKI . Where am I? . The Fraunhofer WKI and the Institute of Building Materials, Concrete Construction and Fire Safety (iBMB
10278	https://www.wki.fraunhofer.de/en/departments/zeluba/profile/contact	Contact - Fraunhofer WKI . Where am I? . Contact . Please contact us, we are looking forward to your inquiry! . Our department: . Contact Press / Media . Head o
10279	https://www.wki.fraunhofer.de/en/departments/zeluba/profile/equipment	Equipment - Fraunhofer WKI . Where am I? . Technical Equipment . Our department: . Contact . Contact Press / Media . Head of department . Fraunhofer Institu
10280	https://www.wki.fraunhofer.de/en/departments/zeluba/profile/publications	Publications - Fraunhofer WKI . Where am I? . Publications . Here you will find current publications resulting from the research work carried out by our departme
10281	https://www.wki.fraunhofer.de/en/departments/zeluba/profile/research-projects	Research projects - Fraunhofer WKI . Where am I? . Research projects . Here you will find details of selected research projects. Please contact us if you require m
10282	https://www.wki.fraunhofer.de/en/departments/zeluba/profile/research-projects/	Beech hybrid element for modern wooden construction - Fraunhofer WKI . Where am I? . Research project . Hybrid structural element for modern wood constru
10283	https://www.wki.fraunhofer.de/en/departments/zeluba/profile/research-projects/	Fire behaviour of synthetic leather - Fraunhofer WKI . Where am I? . Research project . Improving the fire behaviour of synthetic leather . Due to its aesthetic ap
10284	https://www.wki.fraunhofer.de/en/departments/zeluba/profile/research-projects/	Fire protection coating for ultra-high performance concrete (UHPC) - Fraunhofer WKI . Where am I? . Research project . Development of an intumescent fire prot
10285	https://www.wki.fraunhofer.de/en/departments/zeluba/profile/research-projects/	Fungal infestation in fire protection coatings in outdoor areas - Fraunhofer WKI . Where am I? . Research project . Fungal infestation in fire protection coatings in
10286	https://www.wki.fraunhofer.de/en/departments/zeluba/profile/research-projects/	Glimmer and smolder behavior of thermal insulation materials - Fraunhofer WKI . Where am I? . Research project . Glimmer and smolder behavior of thermal in
10287	https://www.wki.fraunhofer.de/en/departments/zeluba/profile/research-projects/	Moment connectors - Fraunhofer WKI . Where am I? . Research project . Development of ultra-high-performance moment connectors . The aim of the project is
10288	https://www.wki.fraunhofer.de/en/departments/zeluba/profile/research-projects/	Moment connectors - Fraunhofer WKI . Where am I? . Research project . Development of ultra-high-performance moment connectors . The aim of the project is
10289	https://www.wki.fraunhofer.de/en/departments/zeluba/profile/services	Services - Fraunhofer WKI . Where am I? . Services . Further tasks within the department include examinations concerning the issuing of general building authori
10290	https://www.wki.fraunhofer.de/en/departments/zeluba/simulation	Simulation - Fraunhofer WKI . Where am I? . Research topic . Numerical Simulation . In research and development, numerical simulation is a very popular tool in
10291	https://www.wki.fraunhofer.de/en/events	Events - Fraunhofer WKI . Where am I? . Results of 6 . Results per page . The K 2019 will present the latest innovations and trends from the plastics and rubber in
10292	https://www.wki.fraunhofer.de/en/events/2018/11ewbps	11th European Wood-based Panel Symposium - Fraunhofer WKI . Where am I? . You would like to obtain information on the most important trends and current

Fraunhofer ISI

Training Data - Corpus Statistics



Topic Model

- Pre-processing
  - Compute unigrams, bigrams and trigrams
  - Lemmatization (spacy, en_core_web_sm)
  - Remove stop-words (i.e., stopwords from nltk.corpus , apply POS filter)
  - Compute tf.idf scores (weights of terms within topics)
- Size of Dictionary computed from Websites: 35,920 unigrams, bigrams or trigrams
- Compute baseline LDA model
  - Implementation: *gensim.models.ldamodel.LdaModel*

Identified Topics

Topic Number	Top 10 Terms in Topic
1	"ionic_conductor" + "organophilic" + "contactor" + "dehydration" + "encrustation" + "fracke" + "geothermic" + “bioenergetic" + "mcfc" + "reconversion"
2	"research" + "product" + “process" + "system" + “develop" + “project" + "image" + “industry“ + "development" + "datum"
3	information" + “new" + "application" + "result" + "monitoring" + "find" + "offer" + "time" + "increase" + "enable"
4	"generate" + "display" + "transfer" + "respect" + "analyze" + “publication" + “fallturm_breman" + "clinician" + "microdisplay" + "pc"
5	...

CONCLUSION AND OUTLOOK

Measuring the intensity of interdisciplinarity

1. Similar approach to Nichols, L. (2014): A topic model approach to measuring interdisciplinarity at the National Science Foundation, *Scientometrics,* 100, pp. 741-754.
   - However, this needs an assignment of the websites to scientific disciplines (classification)
2. Alternativ approach: weighted average distance of topics
   - Use the co-occurence probability of topics in all German/global Scopus publications
   - Index, calculated as the inverse of this co-occurence
   - As topics are assigned to the text corpus probabiistically (not uniquely), we need to weight the co-occurrence by this
   - This might overcomplicate the issue ... however, we give it a try
   - Challenge: How to assign website topics to publication topics?

Conclusions and outlook

- Interdisciplinarity is a multidimensional construct

  → uni-dimensionsal view does not adequately mirror interdisciplinarity (Stirling)

  → single datasources (e.g. publication) might not be sufficient
- We already found three latent constructs within the concept of "interdisciplinarity"
  - a) cooperation behavior, b) knowledge generation and dissemination, c) discipline and field structure
- As we are working at the level of institutes / research groups, a clear demarcation of the entity is necessary (works well for Fraunhofer or Max-Planck)
- How to demarcate departments / faculties in universities?
  - almost impossible in bibliometric data
  - hope is: if we use websites, this problem can be solved on the URL-level