Webis-STEREO-21: Exploring Scientific Text Reuse at Scale

Scientific Authorship and Peer Review Workshop
Berlin, September 1st, 2022

Lukas Gienapp

Victor Zimmermann

Wolfgang Kircheis

Martin Potthast

UNIVERSITÄT LEIPZIG

What is Text Reuse?

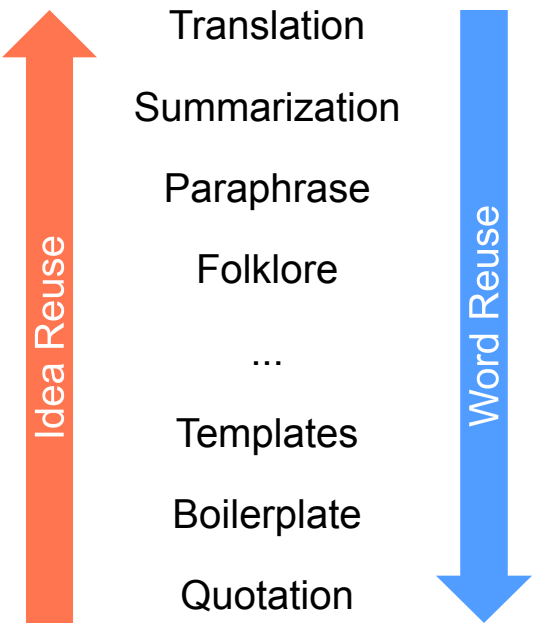
Reused text is text primarily derived from a secondary source.

It occurs in many forms:

- quotations, paraphrases, summarization, ...
  [Potthast, 2011] [Sun et al. 2015]
- template writing, boilerplates, “folklore”, ...
  [Potthast, 2011] [Anson, 2020]

It can be operationalized on different levels:

- reuse on semantic level (“idea reuse”)
- reuse on syntactic level (“word reuse”)
- both levels operationalize reuse as similarity between text



Text reuse is an essential writing technique.

Why is analyzing Text Reuse relevant?

Evolving scientific practice makes reuse a prevalent issue

- digital methods make it easier than ever to reuse texts
- scientific competition makes it an attractive option to reuse
- standardization makes it a necessity to reuse

But: we have little quantitative insight about reuse thus far

- no cross-disciplinary studies
- no large-scale studies
- no general-purpose studies

We assembled the Webis-STEREO-21 dataset to address these issues.

Important: text reuse $\neq$ plagiarism. We do not judge legitimacy!

In this talk...

(1) Collecting Data

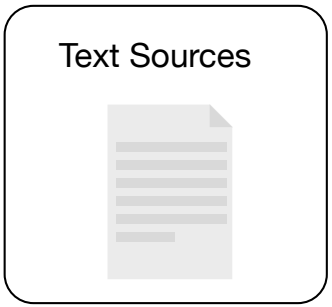
- obtain text data
- obtain metadata

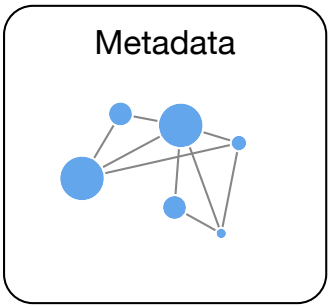
(2) Finding Text Reuse Cases

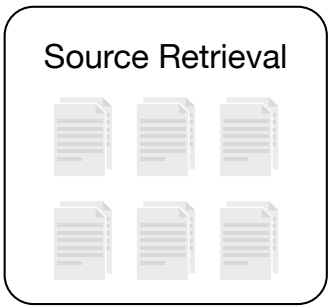
- find document pairs
- compute exact matches

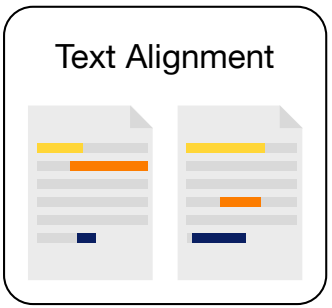
(3) Generate Insight

- analyze individual cases
- analyze general trends









Data Collection

Text Data

- 6 million unique open-access DOIs from CORE-2018
- the PDF file of each DOI is crawled from the web
- plain text is extracted and cleaned from the PDFs

Metadata

- sourced from the Microsoft Open Academic Graph [Sinha et al. 2015]
- DOI, author, title, year of each publication
- field-of-study corresponding to DFG discipline classification

Result: 4.2 million publications with plain text and metadata

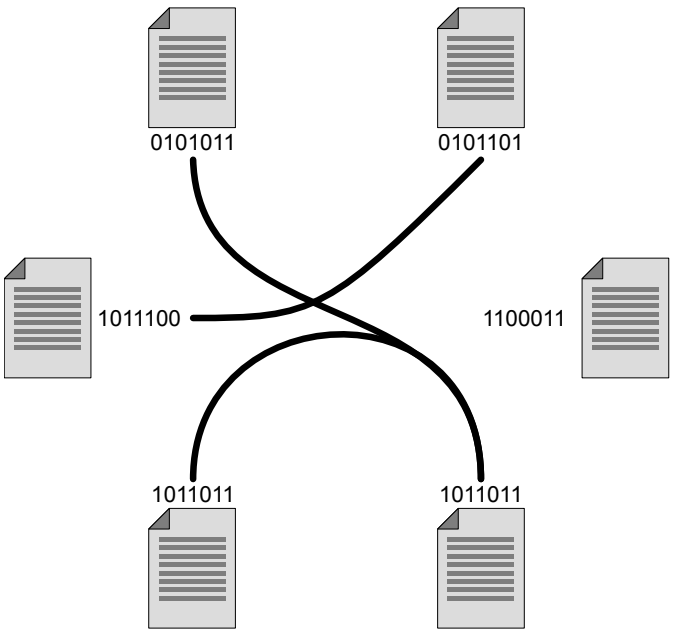
Text Reuse Detection

Source Retrieval

Given the corpus of documents

a fingerprint is calculated for each,

based on which pairs are identified.



Similar fingerprints indicate high likelihood of reuse.

But: exact reused text is unknown.

Text Reuse Detection

Text Alignment

Exact reuse cases are found in 3 steps:

(1) Chunking

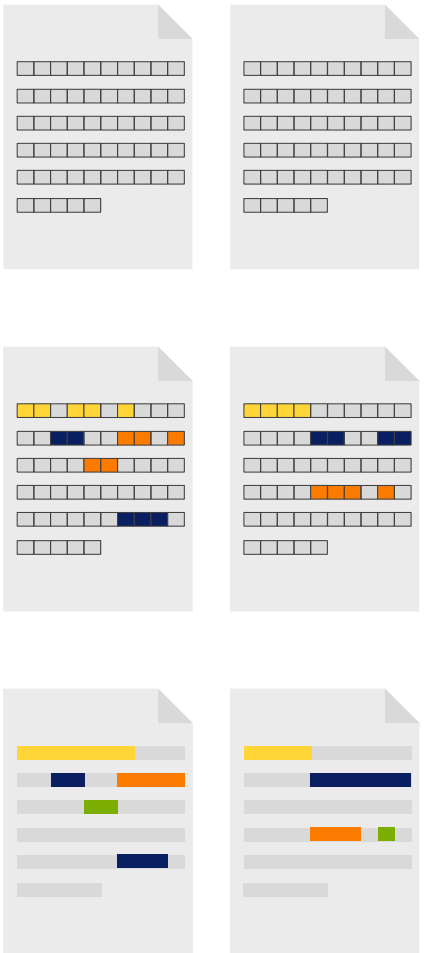
- text is separated into small chunks
- in our case, overlapping 8-word spans

(2) Seeding

- all chunks in text A are compared to all chunks in text B with a similarity measure
- in our case, chunks had to have at least 7 common words

(3) Extending

- chunks in close proximity in one of the texts are joined into larger passages



- Scientific Text Reuse in Open-Access
- 4.2 million unique OA publications
- 91 million cases of reused passages
- cases in 46 scientific fields of study
- cases spanning over the last 150 years
- zenodo.org/record/5575285

Case Examples

Text Recycling

...mplementation of the TBETI (AFBETI) and TFETI (AFFETI) methods. The direct solvers of SPS systems are useful also to the solution of eigenvalue problems with a singular matrix . The results are of importance for the solution of semicoercive contact problems of elasticity with ''floating'' bodies , when it is not possible to avoid manipulations with positive semidefinite matrices by application of the FETI-DP method . More stable, but more laboriou...

...I) method and tested on the solution of a number of engineering problems . The results are of special importance for the solution of semicoercive contact problems of elasticity with 'floating' bodies , when it is not possible to avoid manipulations with positive-semidefinite matrices by application of the FETI-DP method . Moreover, our experiments show that our variant of the Farhat and Géradin method can be used to get effectively the subdomain TFETI and TBETI flexibility matrices that are better conditioned than the corresponding FETI-DP matrices. The approach presented here can be useful also to the solution of eigenvalue problems with a singular matrix and to the solution of highly ill-conditioned problems. This research has been supported by the gran...

- same authors, same topic, different publications, one cites the other
- verbatim copying of (partial) sentences

Case Examples

Boilerplate & Templates

... that the associative processes that give rise to false memories are also enhanced by future planning . Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made ....

...erranean University in Cyprus for the support provided to develop the research reported in this paper . Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made ....

... new localization on the surgical scar. G, gestation; P, parity; TLH, total laparoscopic hysterectomy . Written informed consent was obtained from the patient for publication of this case and accompanying images. A copy of the written consent is available for review by the Editor-in-Chief of this journal . The author declares that they have no competing interests. AS and ER are two major, laparoscopic gy...

...followups of patients diagnosed with LCA for subsequent development of additional tumors is mandatory . Written informed consent was obtained from the patient for publication of this case report and any accompanying images. A copy of the written consent is available for review by the Editor-in-Chief of this journal . Authors' contributions SC reviewed relevant literature and wrote the initial draft. MP reviewed the...

- standardized texts with little modification; no original source is apparent

Case Examples

Plagiarism (?)

...bility of the experimental results can be definitely inferred. reports the analysis of residuals. The normal probability plots display residuals that roughly follow straight lines and, consequently, are normally distributed, with moderate departures from normality. The graphs of the residuals versus the fitted values display residuals that are almost randomly scattered about zero. The detected patterns indicate the presence of a certain uneven spreading of residuals across fitted values. In particular, going towards lower values of the coating thickness, the residuals increase. This can be however expected, being the coating process more and more troublesome to control at any time thinner coating thickness are desired. The histograms of residuals show the bell-shaped distribution of the residuals, with no

... . Figs. 6 and 7 report the residual plots for both CHDFB and EFB coating process. In both cases, the normal probability plots display residuals that roughly follow straight lines and, consequently, are normally distributed, with moderate departures from normality. The graphs of the residuals versus the fitted values display residuals that are randomly scattered about zero. The detected patterns indicate the lack of fanning or uneven spreading of residuals across fitted values, with no evidence of missing terms or outliers. The histograms of residuals show the bell-shaped distribution of the residuals, with no skewness. A couple of modest outliers can be detected in for ARTICLE IN PRESS residuals in EFB coating process. The plots of residuals in the order of the corresponding observations fluctuate in a

- longer passages copied and modified (∼5 more sentences omitted)
- different authors, no attribution given
- **Important**: STEREO-21 does **not** judge the legitimacy of text reuse

Analysis

How prevalent is text reuse in scientific writing?

- frequency over time
- frequency by discipline

When does text reuse occur?

- relative: year delta between the two publications in a case

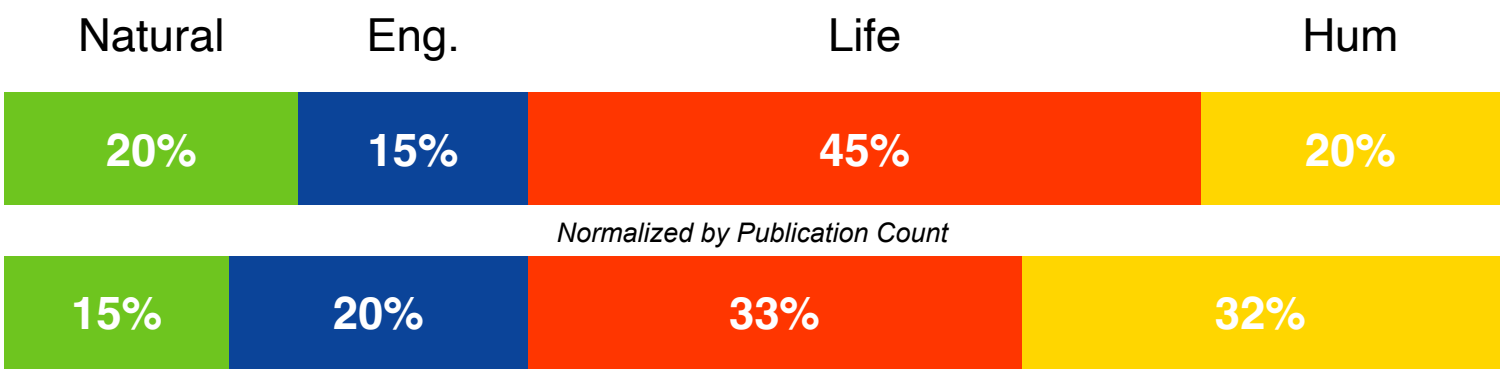
How long are reused passages?

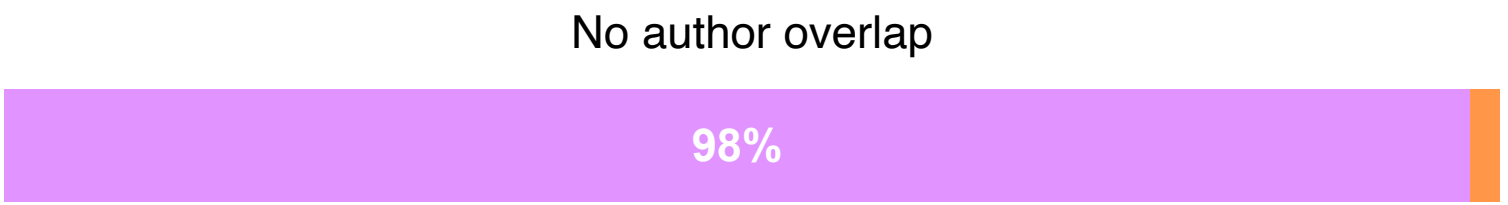
- spread: length of reused passages as character count
- spread differing by author relation, discipline, ...?

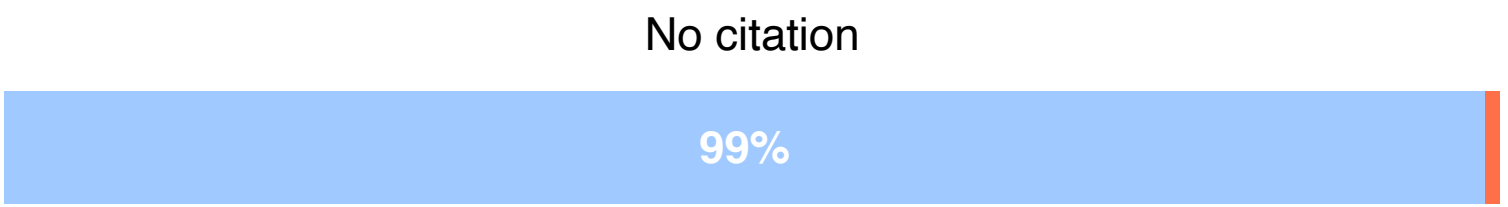
Where does reuse occur?

- location of reuse in publications
- relative: 0 $\rightarrow$ begin, 1 $\rightarrow$ end

Case Distribution

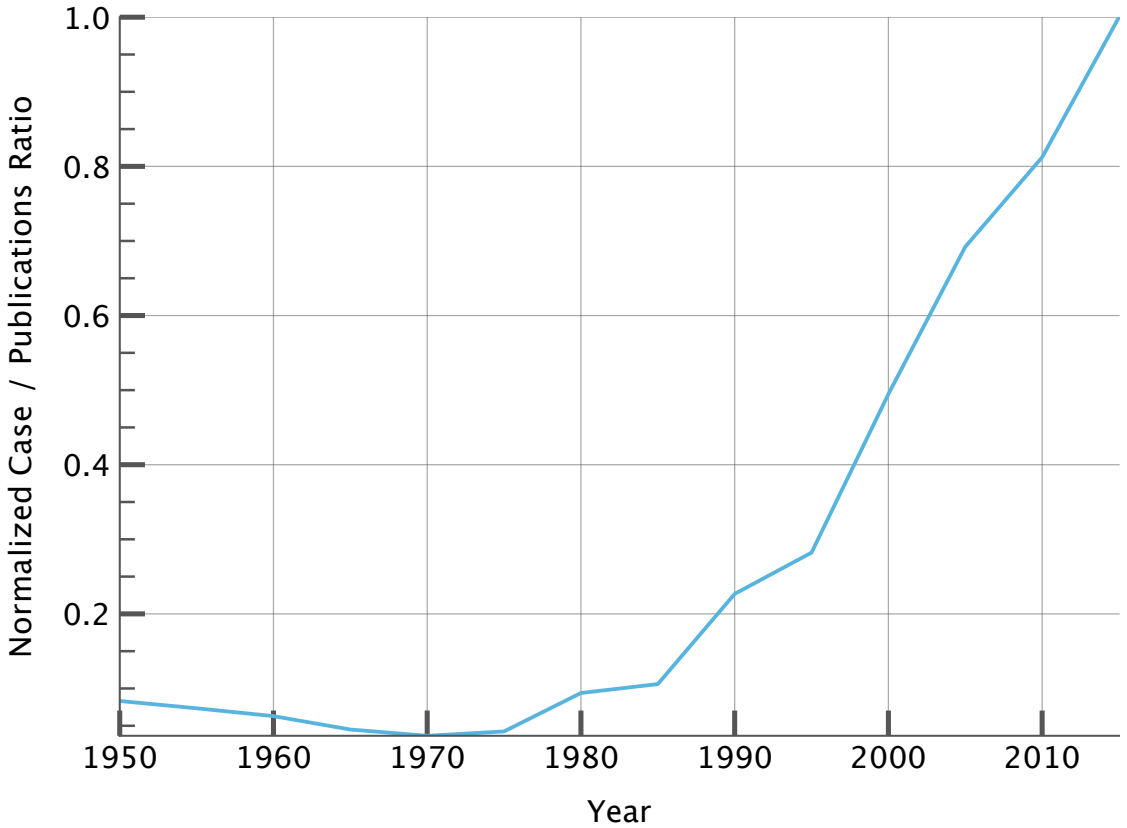






How frequent is text reuse?

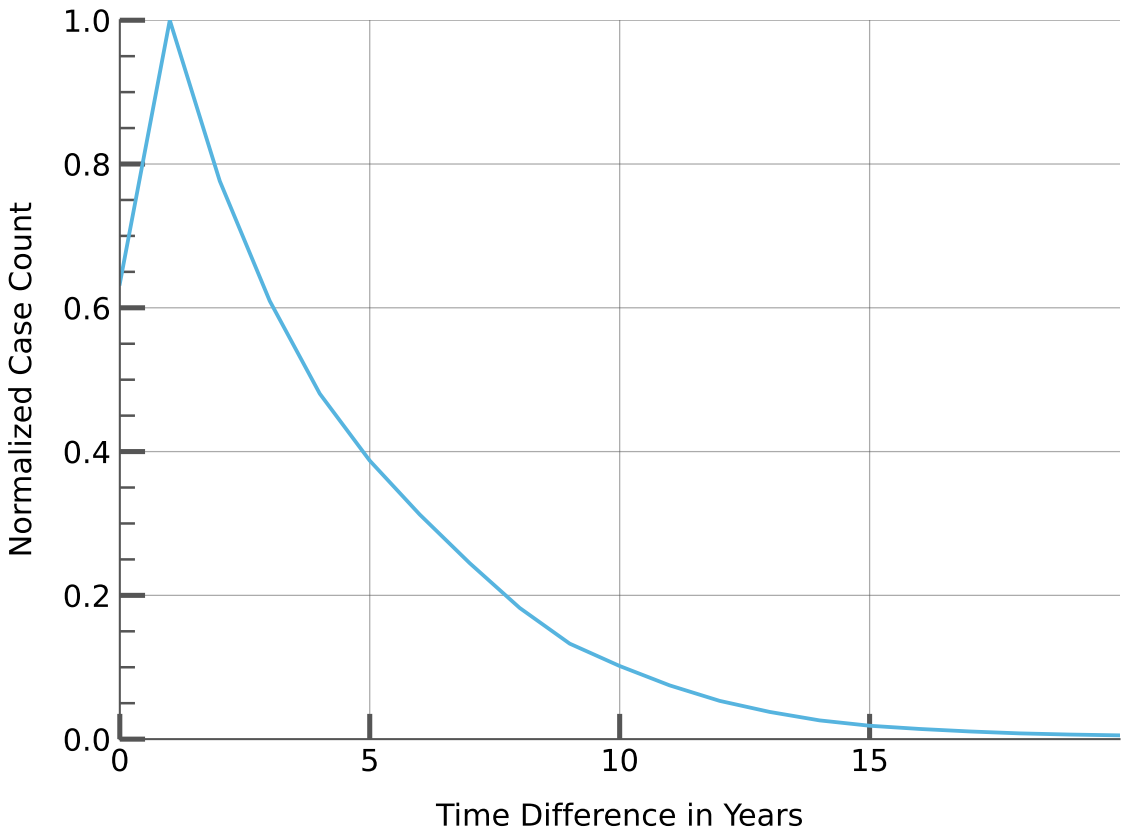
Cases by Year



- yearly number of cases normalized by number of publications is growing
- year-over-year increase is faster for cases than for publications

When does text reuse occur?

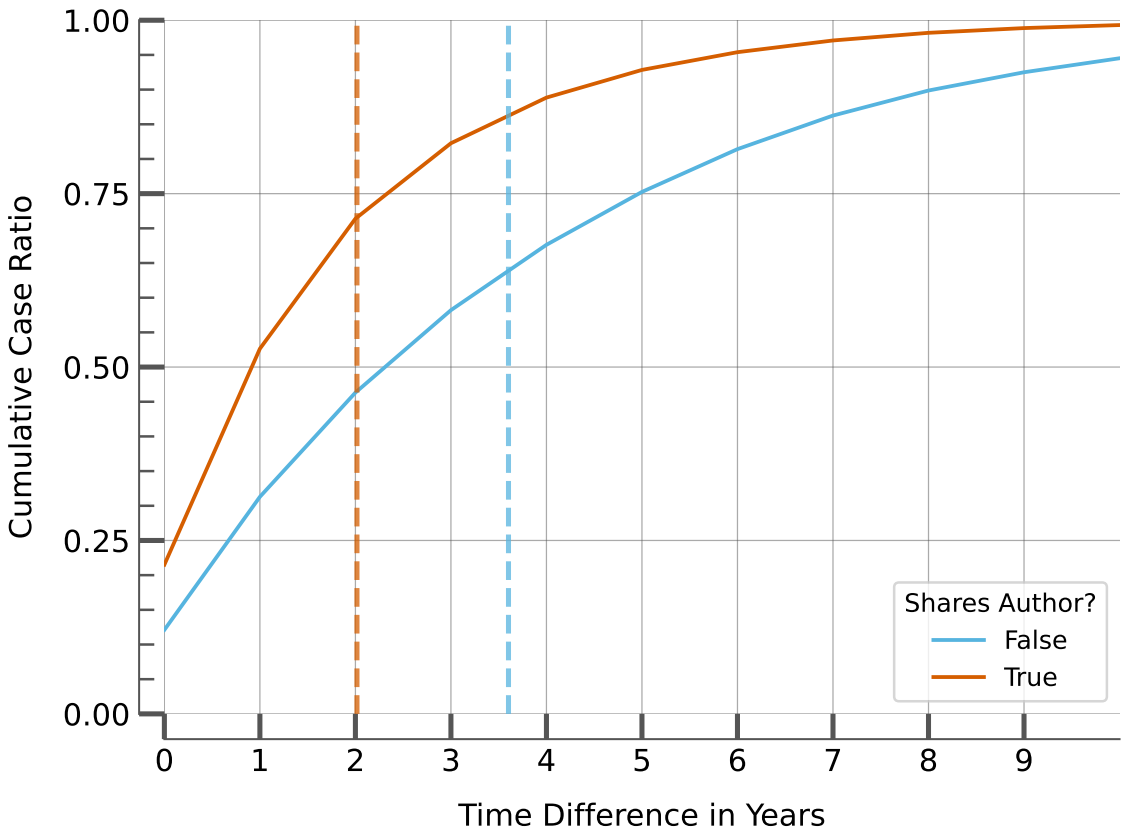
Cases by Year Delta and Author



The majority of reuse occurs in close proximity time-wise.

When does text reuse occur?

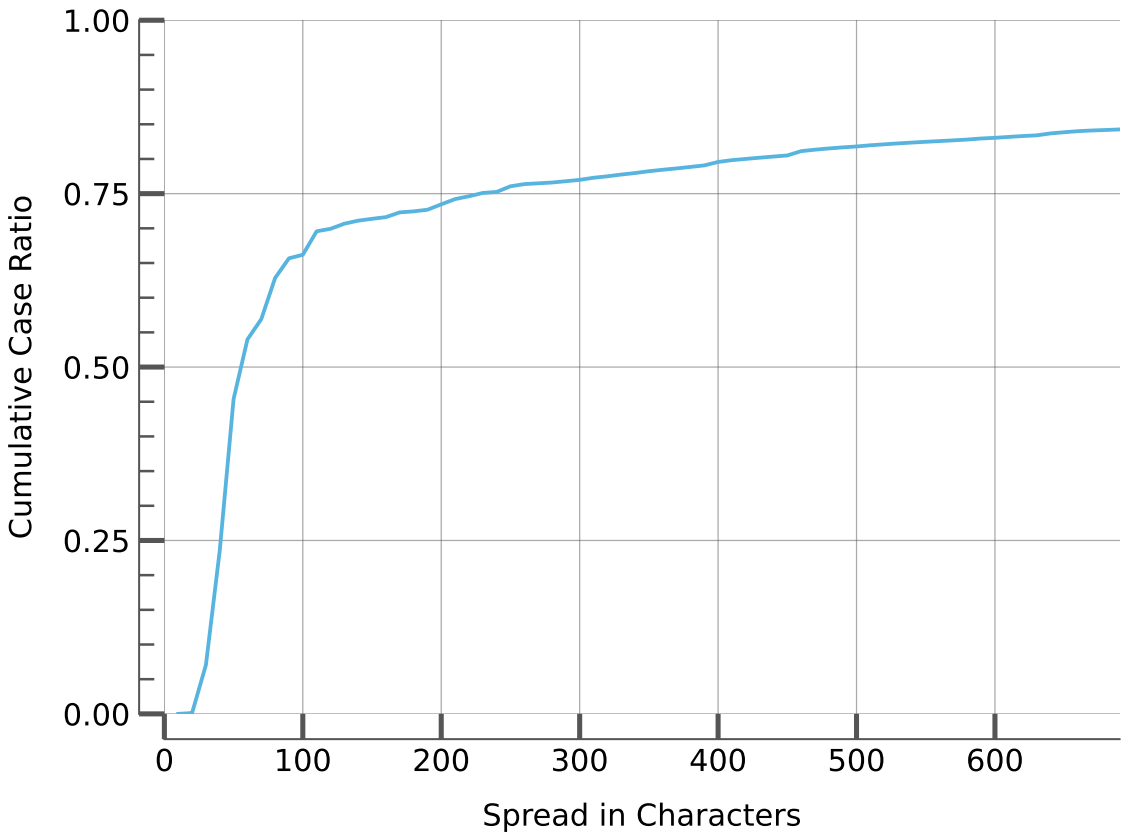
Cases by Year Delta and Author



The reuse time gap is lower when the sources share an author (“text recycling”).

How long are reused passages?

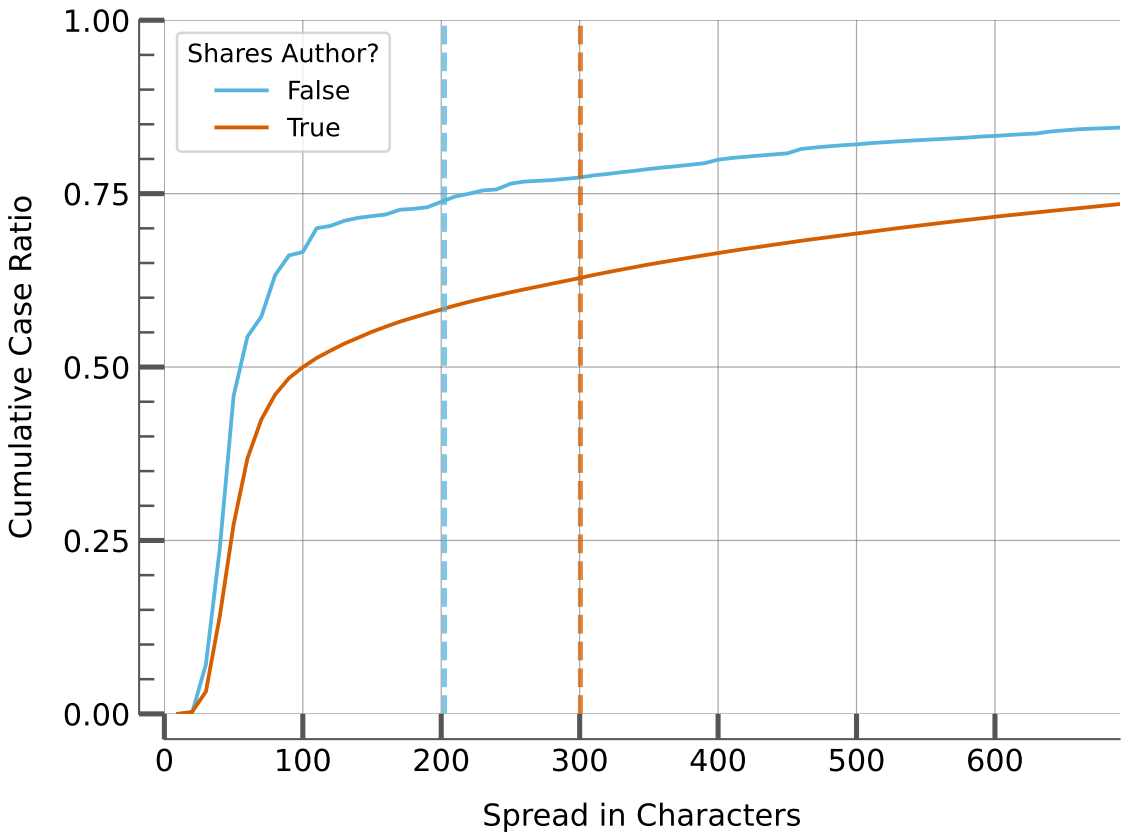
Overall Spread



Reuse is short – 75% of cases are below 200 characters (∼35 words).

How long are reused passages?

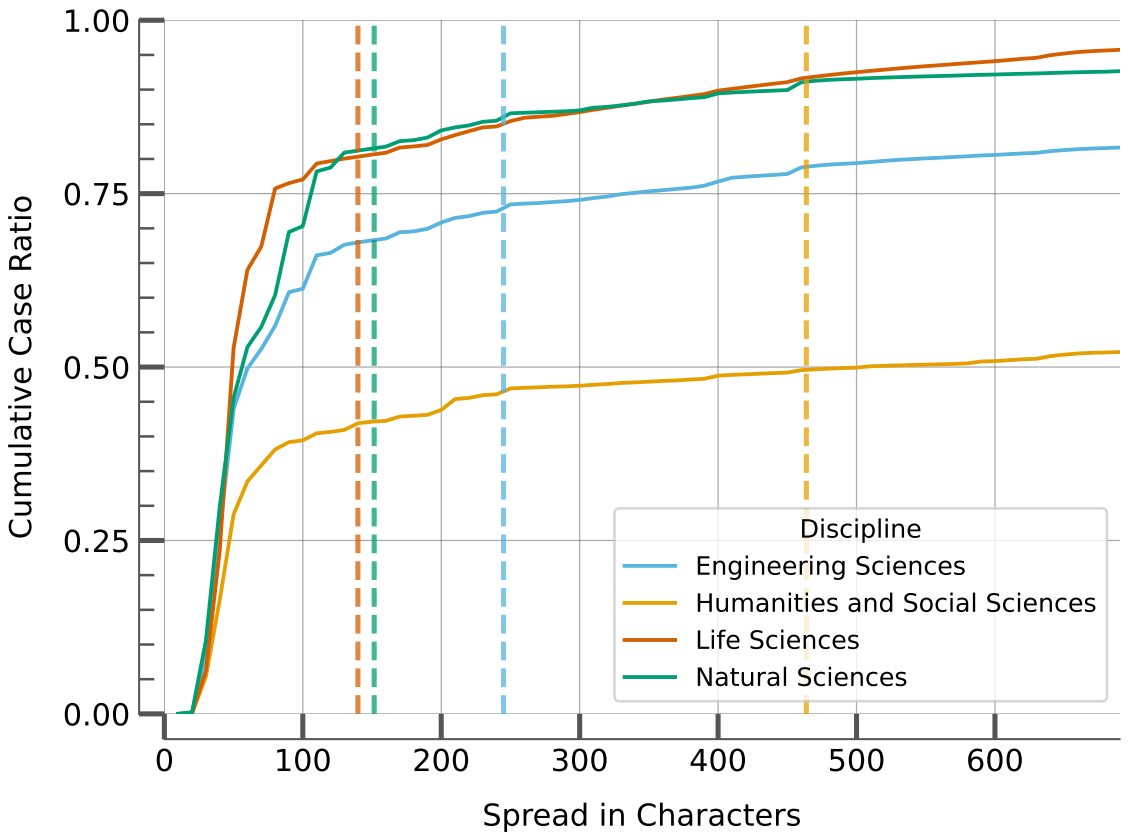
Spread by Author Relation



In cases with a shared author, the reused passages tend to be longer.

How long are reused passages?

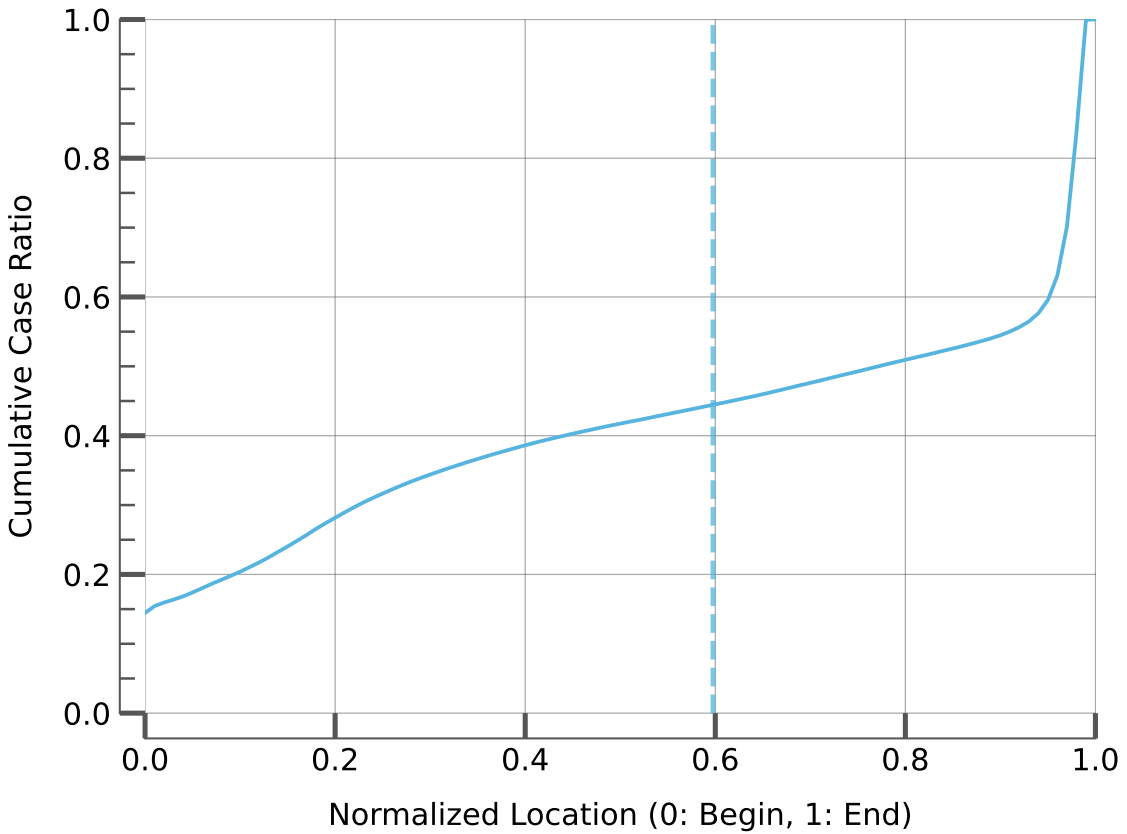
Spread by Discipline



Reuse in Humanities is twice as long as in Engineering Sciences on average; three times as long as Life Sciences and Natural Sciences.

Where does reuse occur?

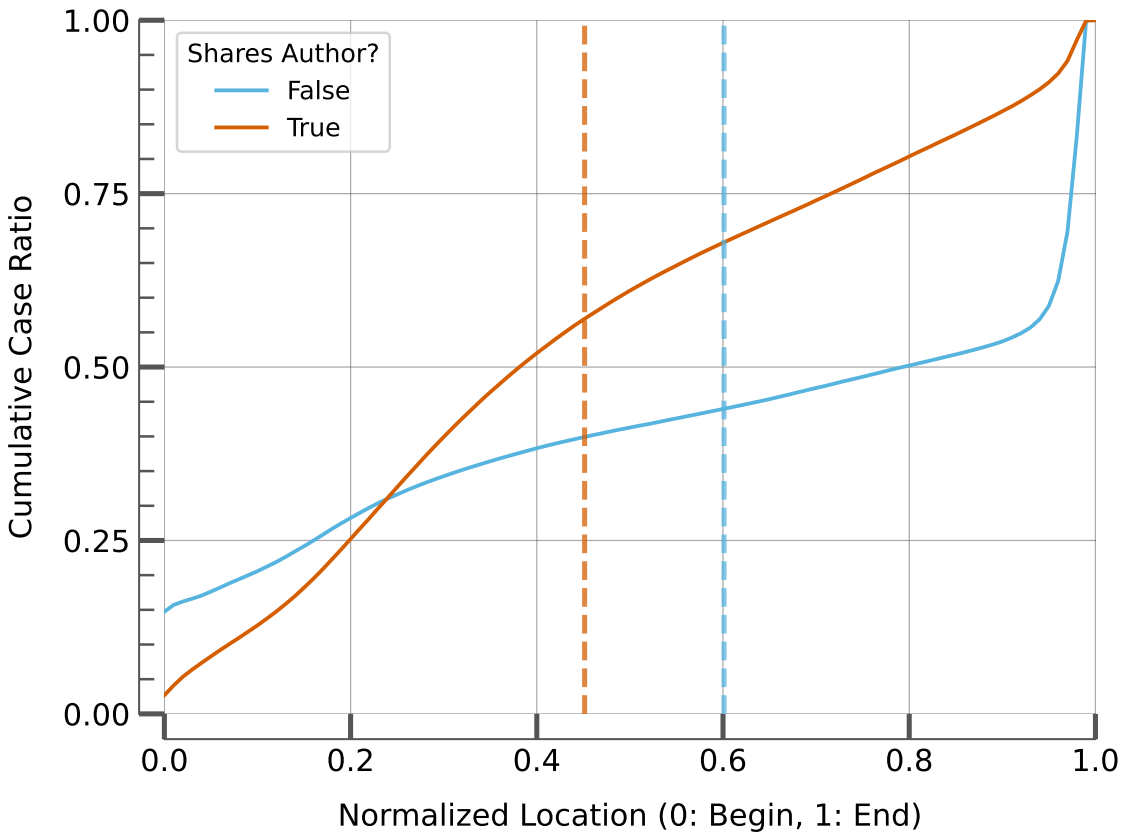
Overall Location



Location of reuse is distributed evenly throughout publications, with a sharp increase at the end.

Where does reuse occur?

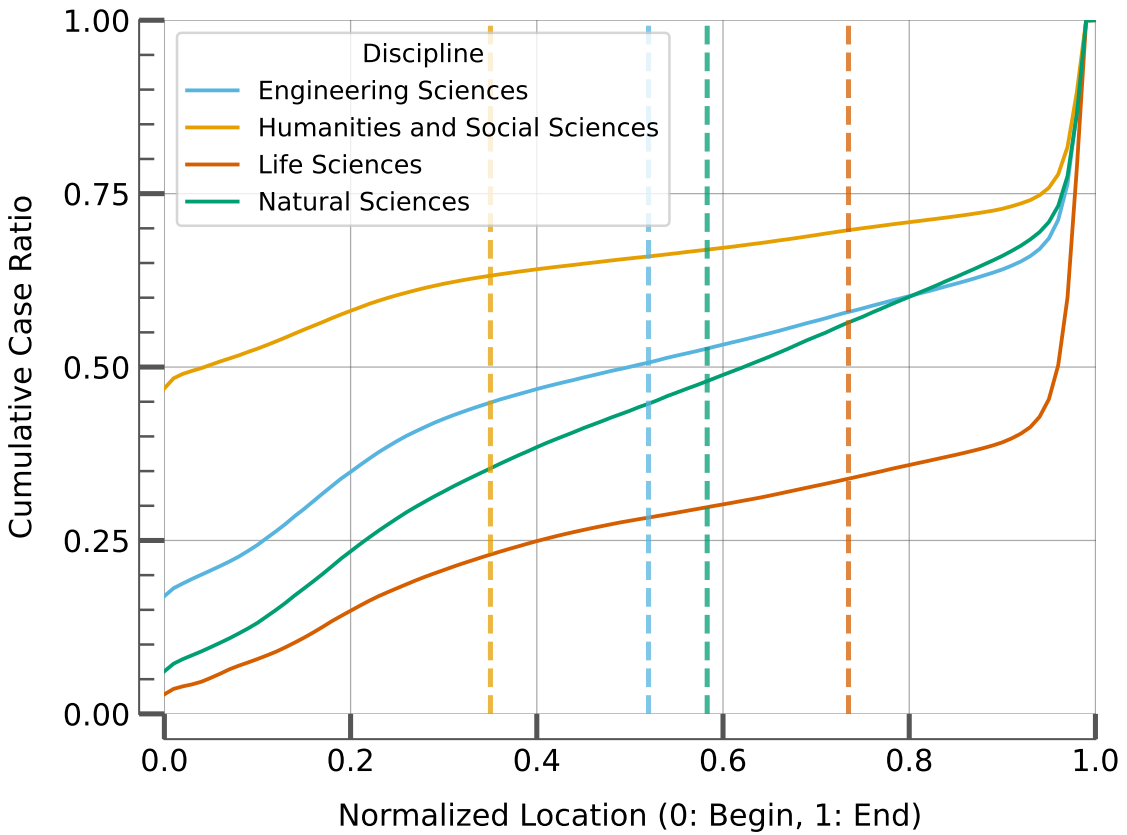
Location by Author Relation



For shared authors, reuse occurs earlier in an article.

Where does reuse occur?

Location by Discipline



More pronounced end spike for Life Sciences (contrib., COI, ethics statements); early onset for Humanities is skewed data (big publisher boilerplate text).

In-Progress and Future Work

Data

- data cleaning
- curation of focused subsets

Access

- development of specialized search engine
- access to subsets for small-scale analysis

Analysis

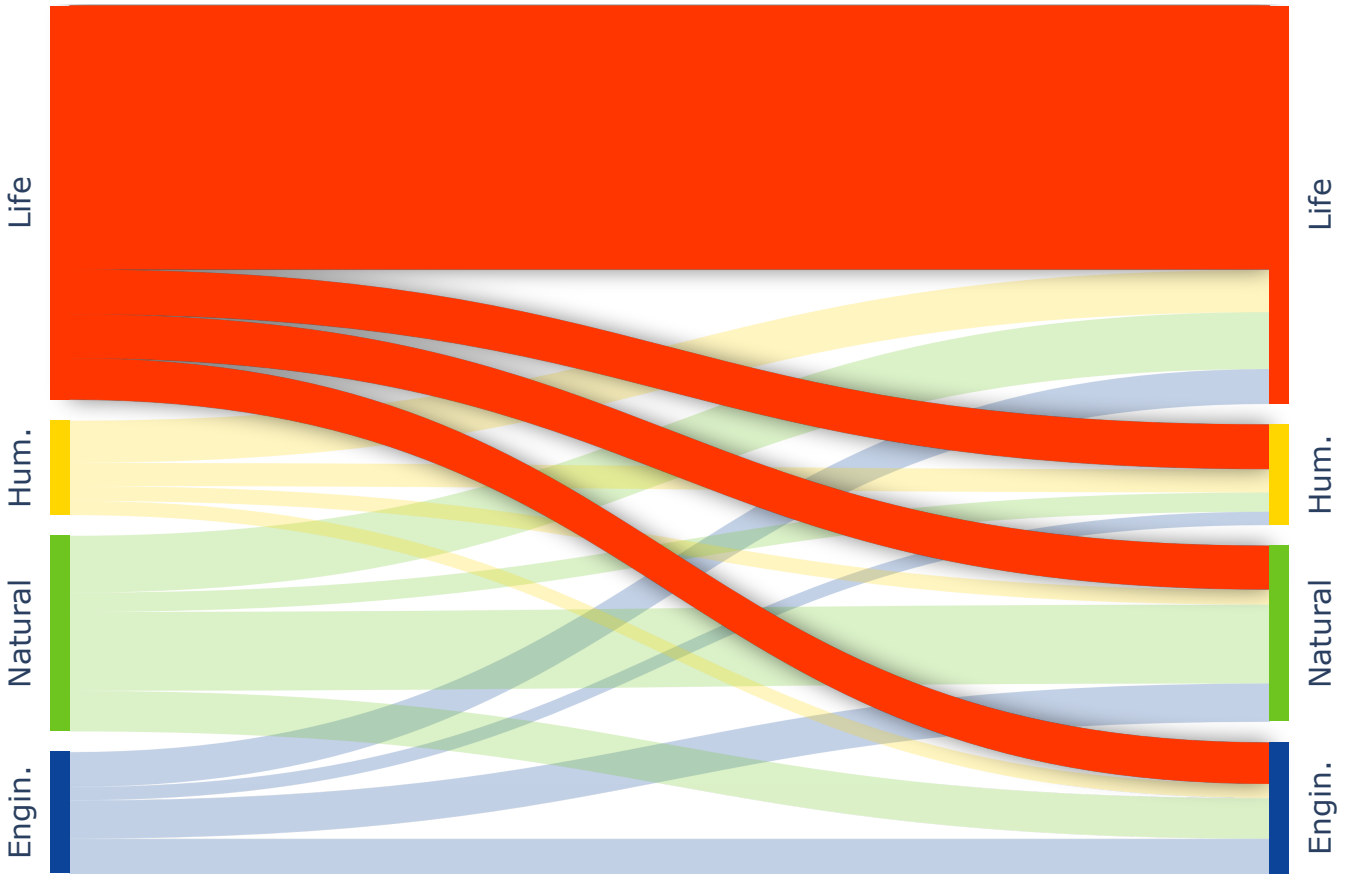
- classify cases into types of reuse
- hypothesis tests, graph-based analysis
- qualitative analysis

Thank you for your attention!

Backup Slides

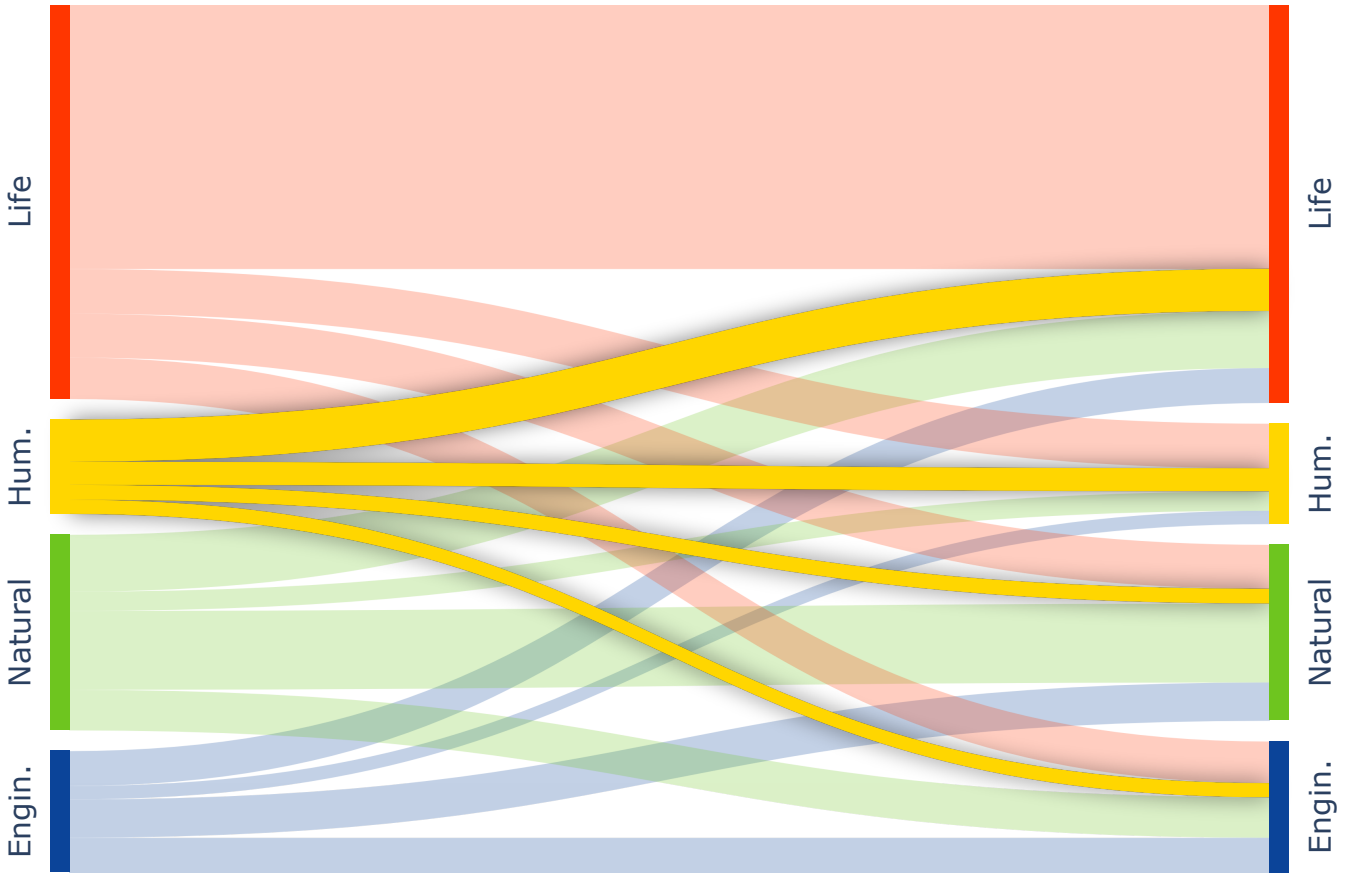
Inter-Disciplinary Reuse

Life Sciences



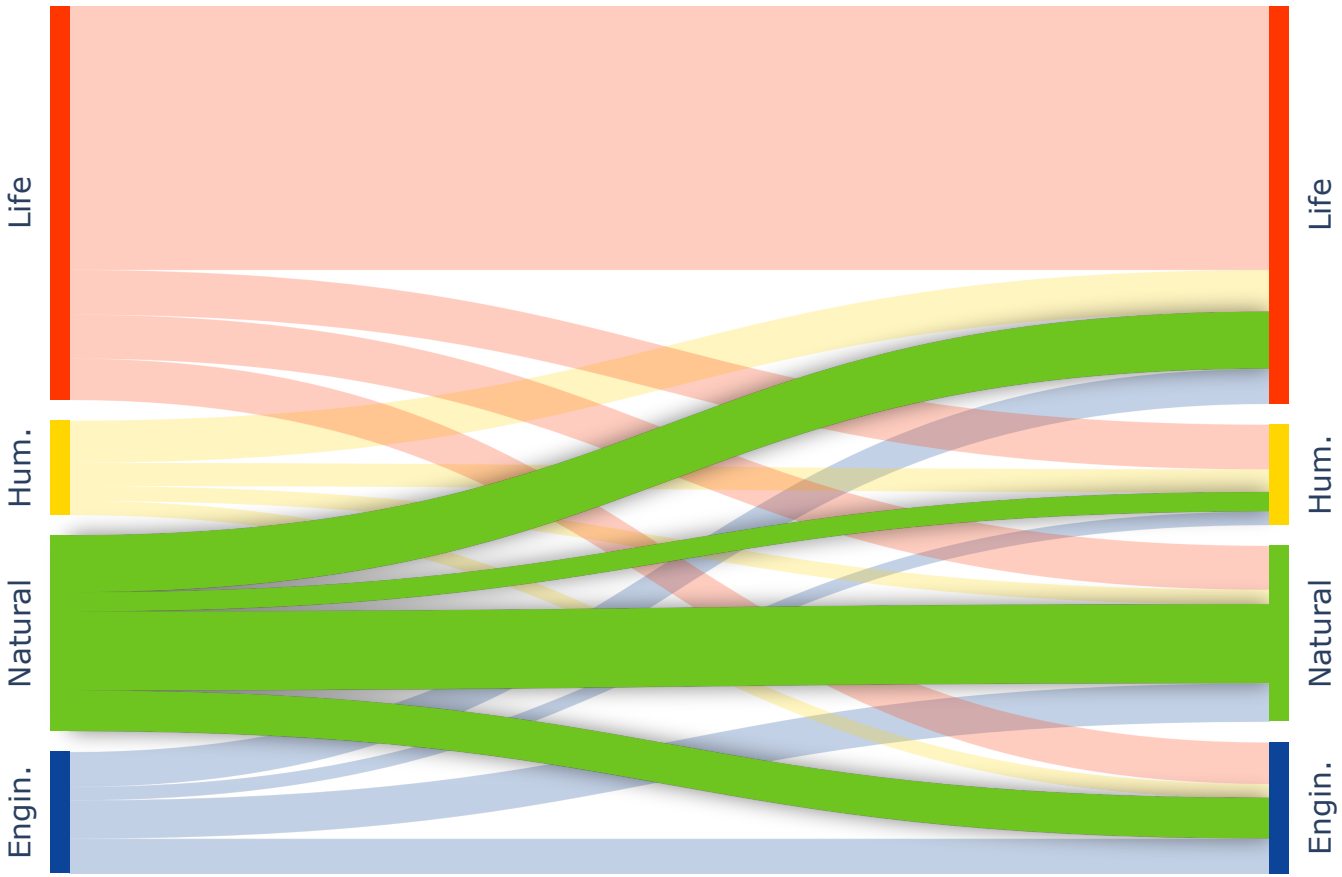
Inter-Disciplinary Reuse

Humanities & Social Sciences



Inter-Disciplinary Reuse

Natural Sciences



Inter-Disciplinary Reuse

Engineering Sciences