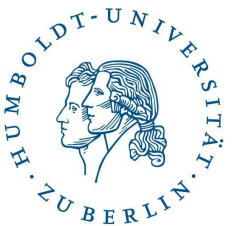
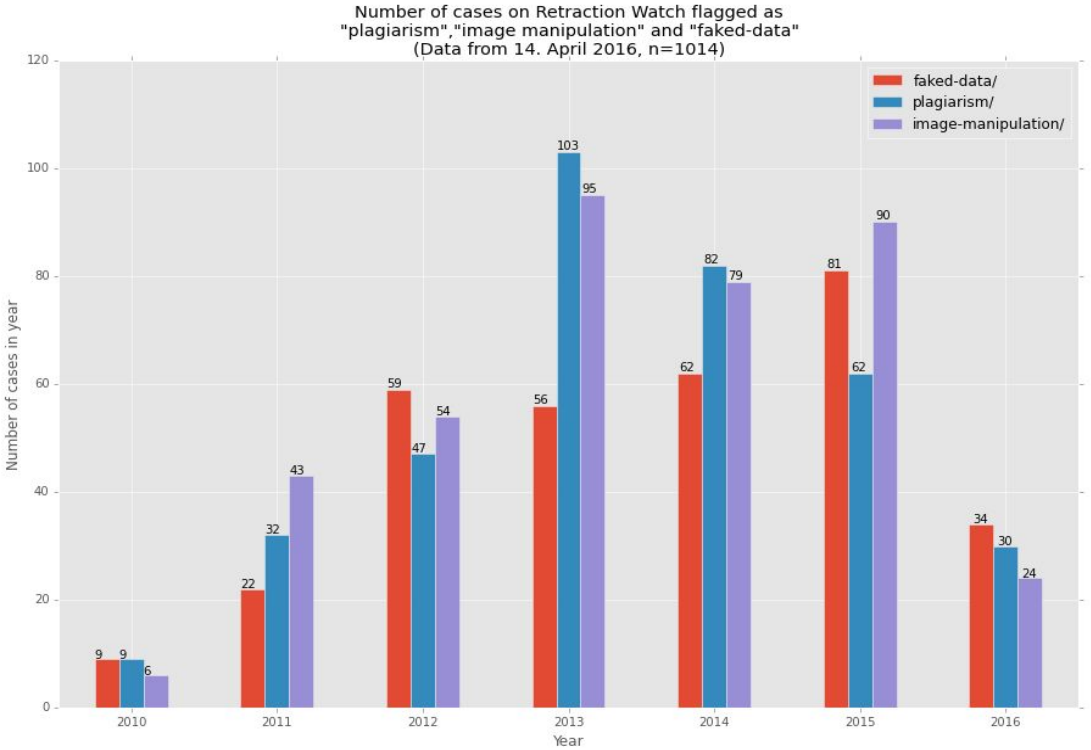


Measuring Plagiarism

By Michael Seadle
Humboldt-Universität zu Berlin
Berlin School of Library and Information Science
& HEADT Centre

Retraction Watch statistics



Graphic credit: Heinz-Alexander Fütterer.

Research Integrity issues 1

Publishers, editors and scholars tend to think of research integrity as a black and white binary choice: either researchers have violated integrity or have not. This is too simple and does not reflect the real world.

Just as the scanning process involves categorizing -- measuring -- the differences between shades, a systematic approach to research integrity should involve measuring the amount of deviation between suspected problems and an expected ideal. This is surprisingly hard.

Research Integrity issues 2

Concepts like “true” and “original” are limits that researchers can approach but, like an asymptote, can never completely reach. No scholarly research is totally original because scholarship in the sciences as in the social sciences and humanities rightly builds on previous work.

The truth of a claim or of a set of data or of an image depends on the social acceptance of the circumstances of their origin. Originality represents an increment toward a new level of “truth”.

Context – various disciplines

Context matters, and the standards for research integrity vary from discipline to discipline. In broad terms the sciences care more about whether the data in a scholarly work lead to a valid and reproducible conclusion, and they care less whether the text contains standard phrases that can be found in other works.

In the humanities and some social sciences (anthropology, for example) originality in the language matters more, since language is for them a key tool.

Overall quantity

The overall number of pages with suspected copying and the percentage of copied text are common metrics. In cases where whole paragraphs were copied word-for-word these numbers can indicate plagiarism and are often taken as conclusive. Such situations are rare, though. More often the copying is less contiguous – sentences or phrases rather than paragraphs – and then the situation can be more ambiguous, especially in sciences with a great deal of standardized language. As a general rule, over 10% could be a problem, over 30% is definitely a problem, and under 10% is often ignored.

Time, place, circumstance

Standards have changed over time. The use of quotation marks for any text taken from another author is taken for granted today, but has not always been. Not all editors, publishers, and thesis/dissertation advisors have exactly the same standards for how complete references should be. Information that is well known in a field may be more acceptable for a well-known expert to relate without footnotes than for a student or a novice. A useful metric might look at the overall pattern of references in a work, compared to others from the same time, place, and circumstance.

Restatements

Plagiarism generally implies missing references. Sometimes references are present, but without a clear link to a copied passage. Many authors believe that a restatement with a general reference somewhere in the text suffices – this was more common in the past. Here a measure of the degree of overlap is needed. How many overlapping words in what sequence were used. Clearly any restatement must reuse some of the same words and phrases. A measure of the acceptance of these sort of restatements over time would provide context.

Standard units and contiguousness

The number of copied sentences and phrases in a standard unit is a potential metric. The standard unit could be a paragraph (though they are quite variable) or a page (again there is some variability) or a fixed number of words.

Contiguousness matters. Separate single words or short phrases may be unproblematic, multiple words or phrases may not be, especially from a single source. A count of units with copying over a certain percentage could facilitate decision-making.

Content and context

All copying may not be the same. Copying in a literature review may reflect a good memory for text or an attempt to state technical matters correctly. It is hard to describe research results without using some of the same words. Copying in a statistical results section may reflect a standardized vocabulary for the kind of test or the discipline.

Here the metrics require comparison with other texts form the same discipline and the same type of content to determine reliably whether the overlap in vocabulary is greater than usual.

Specific case

Normally I cannot talk about specific cases because they are confidential, but the accusations against a member of the government of the state of Mecklenburg-Vorpommern have been in the press.

There is a real possibility that the accusations are political and related to the upcoming elections, but they could also be true. Staff are investigating this case now. Much of what follows comes from work on that case.

General knowledge

When are facts, and the formulations involving those facts, so standard that an author can reproduce them exactly without any actual copying? Here is an example

“Mecklenburg-Vorpommern ist seit dem 03.10.1990 das nordöstlichste Bundesland Deutschlands.” [Mecklenburg-Vorpommern is the most northeastern federal state of Germany since 03 October 1990.]

A metric could be the frequency these words appear in the literature of the discipline – or more broadly. A search in Google for the phrase (without quotes) gets 14 hits.

Grayscale for a plagiarism case

Copying in suspect text (not reference text) with no reference	
Significant copying: multiple **identical** sentences in a paragraph	3
some copying: ≥5 **contiguous** words within a sentence of ≥ 5 words	2
near copying: multiple (≥ 3) exact phrases (≥ 3 words) **overlap** in contiguous sentences in a paragraph	1
similarity 1: several (<3) exact phrases (≥ 3 words) overlap in contiguous sentences in a paragraph or more than 9 words in sequence	0
similarity 2: many (≥5) of the exact same words (excluding function words) in contiguous sentences in the same paragraph	0
similarity 2: topic overlap with facts and standard phrases	0
Standard words or phrases	0
General knowledge / facts	0

Questions?

Contact information: Michael Seadle

Seadle at hu-berlin.de

HEADT Centre
Institut für Bibliotheks- und Informationswissenschaft