Wer schreibt exzellente Artikel? Eine statistische Auswertung.

Martin Rulsch (DerHexer)
derhexer87@yahoo.de
WikiConvention 2013
Karlsruhe, 23. November

WIKIMEDIA
DEUTSCHLAND

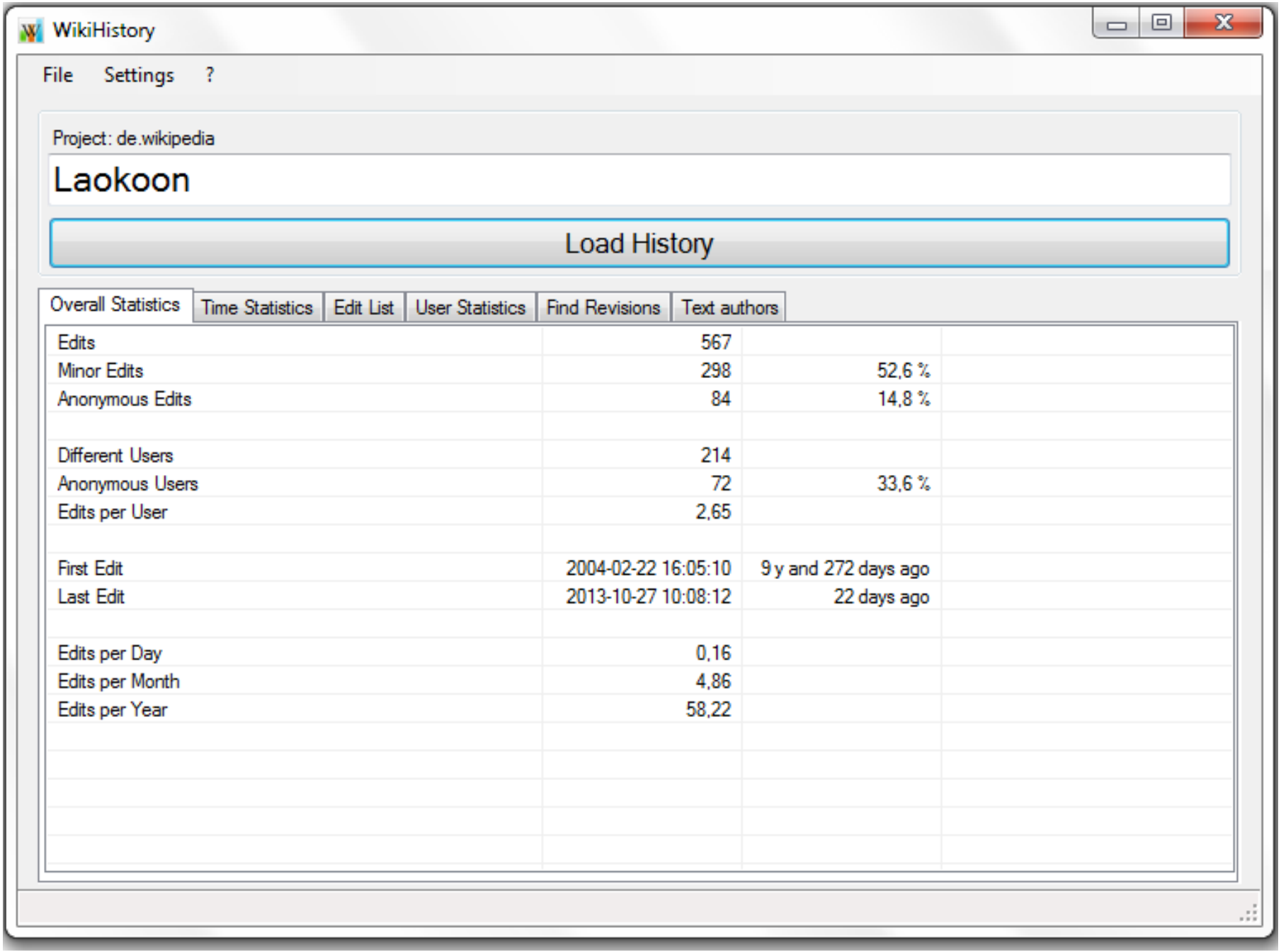
Hintergrund

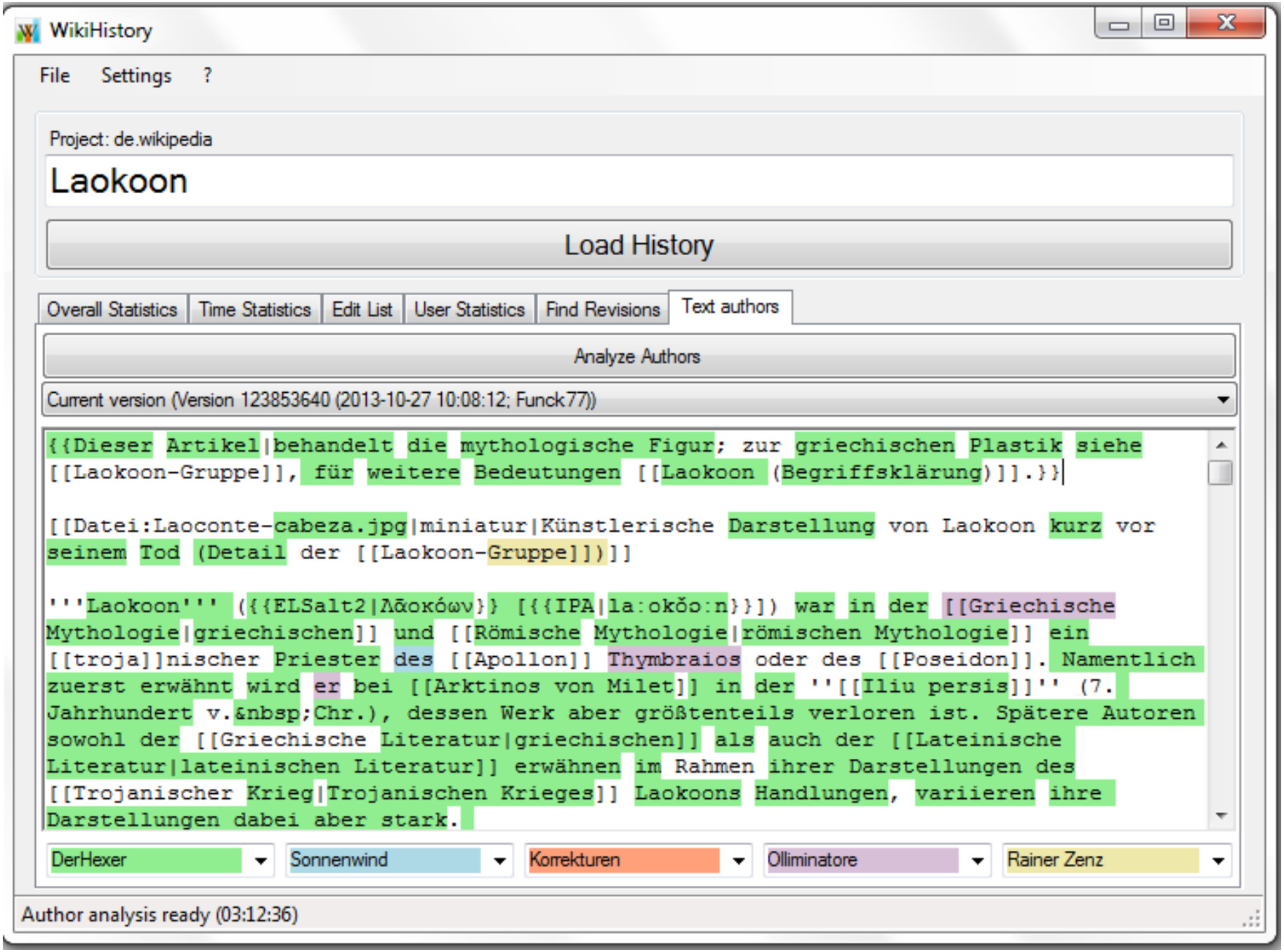
Hintergrund

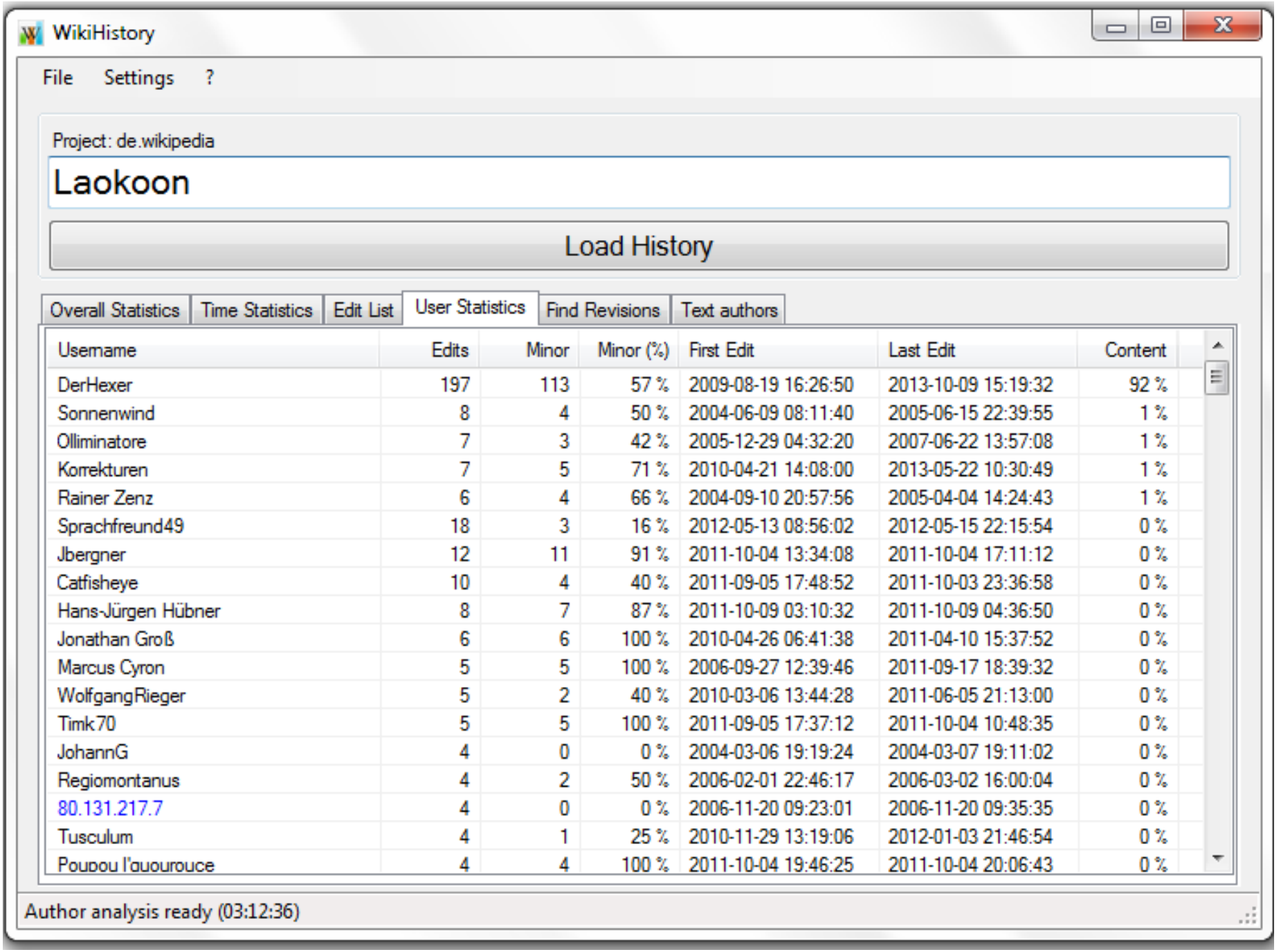
- oft Fragen, wie Qualität in der Wikipedia entsteht
- bisher kaum Untersuchungen zu lesenswerten / exzellenten Artikeln in der Wikipedia
- auch kritisches Thema der Hauptautoren wenig behandelt trotz einiger Tools – nur wenige geben Wortanteil recht präzise an → WikiTrust, APPERs WikiHistory u. a.
- [[Benutzer:Minderbinder/Hauptautoren]] hat WikiHistory für 10 Lesenswerte ausgetestet

WikiHistory

- seit 2008 in Entwicklung, neue Version demnächst veröffentlicht
- recht präzise Bestimmung von Wortanteilen
- weitere Funktionen wie Entwicklung des Artikels über Zeit, einzelne Wörter finden


WikiHistory
File
Settings
?
Project: de.wikipedia
Laokoon
Load History
Overall Statistics
Time Statistics
Edit List
User Statistics
Find Revisions
Text authors
Edits 567
Minor Edits 298 52,6 %
Anonymous Edits 84 14,8 %
Different Users 214
Anonymous Users 72 33,6 %
Edits per User 2,65
First Edit 2004-02-22 16:05:10 9 y and 272 days ago
Last Edit 2013-10-27 10:08:12 22 days ago
Edits per Day 0,16
Edits per Month 4,86
Edits per Year 58,22


WikiHistory
File
Settings
?
Project: de.wikipedia
Laokoon
Load History
Overall Statistics
Time Statistics
Edit List
User Statistics
Find Revisions
Text authors
Analyze Authors
Current version (Version 123853640 (2013-10-27 10:08:12; Funck77))
{{Dieser Artikel|behandelt die mythologische Figur; zur griechischen Plastik siehe [[Laokoon-Gruppe]], für weitere Bedeutungen [[Laokoon (Begriffsklärung)]].}}
[[Datei:Laoconte-cabeza.jpg|miniatur|Künstlerische Darstellung von Laokoon kurz vor seinem Tod (Detail der [[Laokoon-Gruppe]])]]
'''Laokoon''' ({{ELSalt2|Λαοκόων}} [{{IPA|la:okŏo:n}}]) war in der [[Griechische Mythologie|griechischen]] und [[Römische Mythologie|römischen Mythologie]] ein [[troja]]nischer Priester des [[Apollon]] Thymbraios oder des [[Poseidon]]. Namentlich zuerst erwähnt wird er bei [[Arktinos von Milet]] in der ''[[Iliu persis]]'' (7. Jahrhundert v. Chr.), dessen Werk aber größtenteils verloren ist. Spätere Autoren sowohl der [[Griechische Literatur|griechischen]] als auch der [[Lateinische Literatur|lateinischen Literatur]] erwähnen im Rahmen ihrer Darstellungen des [[Trojanischer Krieg|Trojanischen Krieges]] Laokoons Handlungen, variieren ihre Darstellungen dabei aber stark.
DerHexer
Sonnenwind
Korrekturen
Olliminatore
Rainer Zenz
Author analysis ready (03:12:36)


WikiHistory
File
Settings
?
Project: de.wikipedia
Laokoon
Load History
Overall Statistics
Time Statistics
Edit List
User Statistics
Find Revisions
Text authors
Username
Edits
Minor
Minor (%)
First Edit
Last Edit
Content
DerHexer 197 113 57 % 2009-08-19 16:26:50 2013-10-09 15:19:32 92 %
Sonnenwind 8 4 50 % 2004-06-09 08:11:40 2005-06-15 22:39:55 1 %
Olliminatore 7 3 42 % 2005-12-29 04:32:20 2007-06-22 13:57:08 1 %
Korrekturen 7 5 71 % 2010-04-21 14:08:00 2013-05-22 10:30:49 1 %
Rainer Zenz 6 4 66 % 2004-09-10 20:57:56 2005-04-04 14:24:43 1 %
Sprachfreund49 18 3 16 % 2012-05-13 08:56:02 2012-05-15 22:15:54 0 %
Jbergner 12 11 91 % 2011-10-04 13:34:08 2011-10-04 17:11:12 0 %
Catfisheye 10 4 40 % 2011-09-05 17:48:52 2011-10-03 23:36:58 0 %
Hans-Jürgen Hübner 8 7 87 % 2011-10-09 03:10:32 2011-10-09 04:36:50 0 %
Jonathan Groß 6 6 100 % 2010-04-26 06:41:38 2011-04-10 15:37:52 0 %
Marcus Cyron 5 5 100 % 2006-09-27 12:39:46 2011-09-17 18:39:32 0 %
WolfgangRieger 5 2 40 % 2010-03-06 13:44:28 2011-06-05 21:13:00 0 %
Timk70 5 5 100 % 2011-09-05 17:37:12 2011-10-04 10:48:35 0 %
JohannG 4 0 0 % 2004-03-06 19:19:24 2004-03-07 19:11:02 0 %
Regiomontanus 4 2 50 % 2006-02-01 22:46:17 2006-03-02 16:00:04 0 %
80.131.217.7 4 0 0 % 2006-11-20 09:23:01 2006-11-20 09:35:35 0 %
Tusculum 4 1 25 % 2010-11-29 13:19:06 2012-01-03 21:46:54 0 %
Poupou l'quourouce 4 4 100 % 2011-10-04 19:46:25 2011-10-04 20:06:43 0 %
Author analysis ready (03:12:36)

WikiHistory

- Beispiel:
  - aktuelle (3.) Version: Die Wikipedia ist eine freie Enzyklopädie und darüber hinaus ein Communityprojekt.
  - 1. Version: Wikipedia ist eine Enzyklopädie.
  - 2. Version: Die Wikipedia ist ein Communityprojekt und darüberhinaus eine Enzyklopädie.
- Analyse:
  - Übereinstimmungsprüfung vom größten zum kleinsten Segment (vollständiger Artikel zu einzelnem Wort)

WikiHistory

- Analyse der 1. Version „Wikipedia ist eine Enzyklopädie.“
  - ist Version 3 komplett in Version 1 enthalten? → nein
  - sind „Die Wikipedia ist eine freie Enzyklopädie und darüber hinaus ein“ darin, oder „Wikipedia ist eine freie Enzyklopädie und darüber hinaus ein Communityprojekt.“? → nein
  - → 1. Fund: [Wikipedia ist eine]
  - → 2. Fund: [Enzyklopädie]

WikiHistory

- Analyse der 2. Version „Die Wikipedia ist ein Communityprojekt und darüberhinaus eine Enzyklopädie.“
  - sind die noch nicht an die 1. Version vergebenen Teile hier zu finden?
  - ➔ 1. Fund: [ein Communityprojekt]
  - ➔ 2. Fund: [Die]
  - ➔ 3. Fund: [und]

WikiHistory

- Analyse der 3. Version „ Die Wikipedia ist eine freie Enzyklopädie und darüber hinaus ein Communityprojekt.“
  - sind die noch nicht an die 1. und 2. Version vergebenen Teile hier zu finden?
  - ➔ 1. Fund: [darüber hinaus]
  - ➔ 2. Fund: [freie]

WikiHistory

- Ergebnis
  - [Die] [Wikipedia ist eine] [freie] [Enzyklopädie] [und] [darüber hinaus] [ein Communityprojekt].
- daraus ergibt sich ein prozentualer Zeichenanteil für die jeweiligen Autoren → wenn Version 1 und 3 vom gleichen Autor wären, wäre natürlich beides farblich gleich markiert:
  - [Die] [Wikipedia ist eine] [freie] [Enzyklopädie] [und] [darüber hinaus] [ein Communityprojekt].

Grenzen

- Rundung + Sortierung 1. nach Prozent, 2. nach Editanzahl
- kleine Wörter werden eher früheren Autoren zugewiesen → dadurch gegebenenfalls überrepräsentiert
- Rechtschreibkorrekturen werden dem Korrekteur zugewiesen
- Wortumstellungen → Sinnveränderungen?
- es werden jedoch Zurücksetzungen korrekt ausgeklammert
- → trotzdem ziemlich präzise (kleinere Fehler wurden inzwischen korrigiert) + das Beste, was wir haben

Untersuchung

Untersuchungsgegenstand

- Exzellente Artikel gemäß [[Kategorie:Wikipedia:Exzellent]]: rund 2.250 Artikel
  - von 10.000 bis 350.000 Bytes, von 30 bis 7000 Versionen, von 10 bis 2300 Bearbeitern, von 100 bis 100.000 Zugriffen pro Monat
- Untersuchung
  - mit WPPageHistStat, stats.grok.se und v. a. WikiHistory
  - seit Juli 2013 – vermehrt seit September 2013, mittlerweile alle ausgewertet, Dauer von 30 Sekunden bis zu mehreren Tagen je Artikel

Untersuchungsgegenstand

- gesammelte Daten anhand eines Beispiels

LNr	Lemma	Revision	Erstell-datum	Exzellenz-datum	Bytes	Versionen
1078	Laokoon	123853640	2004-02	2011-10	147.244	567

Benutzer	Zugriffe/ Monat	Zeichenanteil 1. Autor	Edits 1. Autor	Name 1. Autor	Meiste Edits
214	3756	91 %	197	DerHexer	1

Zeichenanteil 2. Autor	Edits 2. Autor	…	Zeichenanteil 5. Autor	Edits 5. Autor
1 %	8	…	1 %	6

Untersuchungsgegenstand

- weitere zentrale Werte:
  - durchschnittliche Bearbeitungen je Benutzer (2,65)
  - Prozentanteil der Bearbeitungen des 1. und 2. Autors an der Gesamtbearbeitungszahl (35 %; 1 %)
  - Zeichenanteil, Bearbeitungen und Prozentanteil der Bearbeitungen an der Gesamtbearbeitungszahl des 1. und 2. Autors sowie des 1. bis 5. Autors (92 %; 205; 36 %;; 95 %; 225; 40 %)
  - Differenz zwischen Zeichenanteil des 1. und 2. Autors (90 %)

Zentrale Ergebnisse

Zentrale Ergebnisse

Methode	Bytes	Bearbei-tungen	Benutzer	Bearbei-tungen / Benutzer	Zugriffs-zahlen
Arithm. Mittel	61.693	622	226	3,27	4.831
Geom. Mittel	52.032	415	144	2,89	1.744
Median	52.086	393	134	2,58	1.483
Maximum	367.137	7.006	2.293	69,40	105.766
Minimum	10.136	31	10	1,26	121

Zentrale Ergebnisse

Methode	Zeichen-anteil 1. Autor	Edit-anteil 1. Autor	Zeichen-anteil 2. Autor	Edit-anteil 2. Autor
Arithm. Mittel	70 %	31 %	8 %	6 %
Geom. Mittel	65 %	22 %	6 %	3 %
Median	76 %	27 %	6 %	3 %
Maximum	100 %	87 %	45 % (3.: 28 %, 4.: 16 %, 5.: 15 %)	61 %
Minimum	8 %	0,04 %	0 %	0,03 %

Zentrale Ergebnisse

Methode	Zeichen-anteil 1.–2. Autor	Edit-anteil 1.–2. Autor	Zeichen-anteil 1.–5. Autor	Edit-anteil 1.–5. Autor
Arithm. Mittel	78 %	37 %	86 %	44 %
Geom. Mittel	76 %	30 %	85 %	38 %
Median	84 %	34 %	91 %	42 %
Maximum	100 %	97 %	100 %	98 %
Minimum	14 %	0,3 %	24 %	2 %

Zentrale Ergebnisse

- weitere zentrale Ergebnisse:
  - in 87,9 % der untersuchten Artikel war der Hauptautor auch der Bearbeiter mit den meisten Bearbeitungen
  - der Abstand des Zeichenanteils vom 1. zum 2. Autor beträgt im Durchschnitt 62 %, der Median 69 %
  - es gibt über 725 Autoren, die mindestens einen exzellenten Artikel geschrieben haben – im Schnitt rund 3,1
  - es gibt über 110 Autoren, die mehr als 5, über 45, die mehr als 10, und über 15, die mehr als 20, und 6, die mehr als 30 exzellente Artikel geschrieben haben

Zentrale Ergebnisse

- die Person mit den meisten exzellenten Artikeln (72 – mehr als 20 mehr als Nummer 2!) hat nicht die meisten Bytes als Hauptautor beigetragen (mit >40 % Abstand Nummer 3) und recht geringen Zeichenanteil
- die Top 20 (minimal 16 exzellente Artikel) erstellen innerhalb von 7,5 Jahren ihre exzellenten Artikel – Maximum: 10, Minimum: 3 Jahre
- ihre Artikel sind im Schnitt 15 % kleiner als der Durchschnitts-Exzellente-Artikel, haben 25 % weniger Versionen, 20 % weniger beteiligte Benutzer und geringere Zugriffszahl – sonst alles maximal 10 % Abweichung nach oben oder unten

Zentrale Ergebnisse

- Zeichenanteil des 1. Autors

≥95 %	≥90 %	≥75 %	≥50 %	≥25 %
7,6 %	20,2 %	52,0 %	81,5 %	95,7 %

- Editanteil des 1. Autors

≥75 %	≥50 %	≥25 %	≥10 %	≥5 %
2,4 %	19,4 %	53,4 %	83,5 %	93,3 %

Zentrale Ergebnisse

- Zeichenanteil mehrerer Autoren
  - oberes Ende

≥2 Autoren ≥30 %	≥2 Autoren ≥10 %	≥3 Autoren ≥20 %	≥3 Autoren ≥10 %	≥3 Autoren ≥5 %	≥5 Autoren ≥5 %
3,2 %	29,3 %	0,4 % (8)	5,8 %	27,7 %	3,8 %

  - unteres Ende

≥1 Aut. ≤1 %	≥2 Aut. ≤1 %	≥3 Aut. ≤1 %	≥4 Aut. ≤1 %	≥1 Aut. 0 %	≥2 Aut. 0 %	≥3 Aut. 0 %
58,1 %	42,9 %	26,6 %	10,6 %	15,8 % (max. 4)	9,4 %	4,4 %

Zentrale Ergebnisse

- viele weitere Fragemöglichkeiten:
  - 11 Artikel, in denen 1. und 2. Autor gleichen Zeichenanteil haben
  - bei 58,1 % der exzellenten Artikel hat der 5. Autor weniger als 2 % beigetragen, nur in 3,8 % mehr als 5 %
  - bei 19,2 % der exzellenten Artikel hat der 1. Autor ≥50 % Bearbeitungsprozent
  - bei 11,7 % der exzellenten Artikel hat der 2., 3., 4. oder 5. Autor mehr Bearbeitungen getätigt als der 1. Autor
  - bei 37,8 % der exzellenten Artikel haben der 1.–5. Autor über 50 % der Bearbeitungen getätigt

Datums-angaben

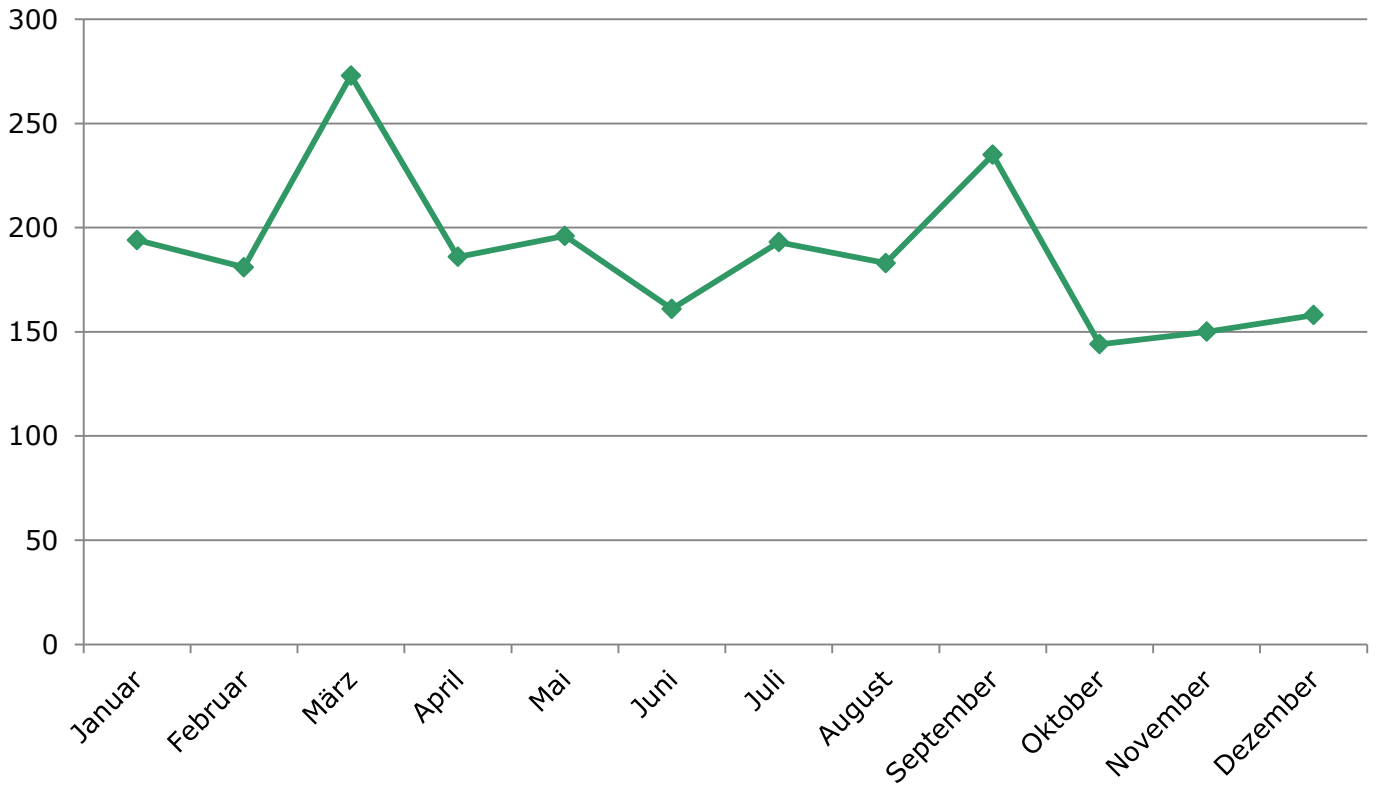
Artikel nach Jahr

x-Achse: Erstellungsjahr des exz. Artikels
y-Achse: Anzahl der in diesem Jahr erstellten exz. Artikel

Artikel nach Jahr

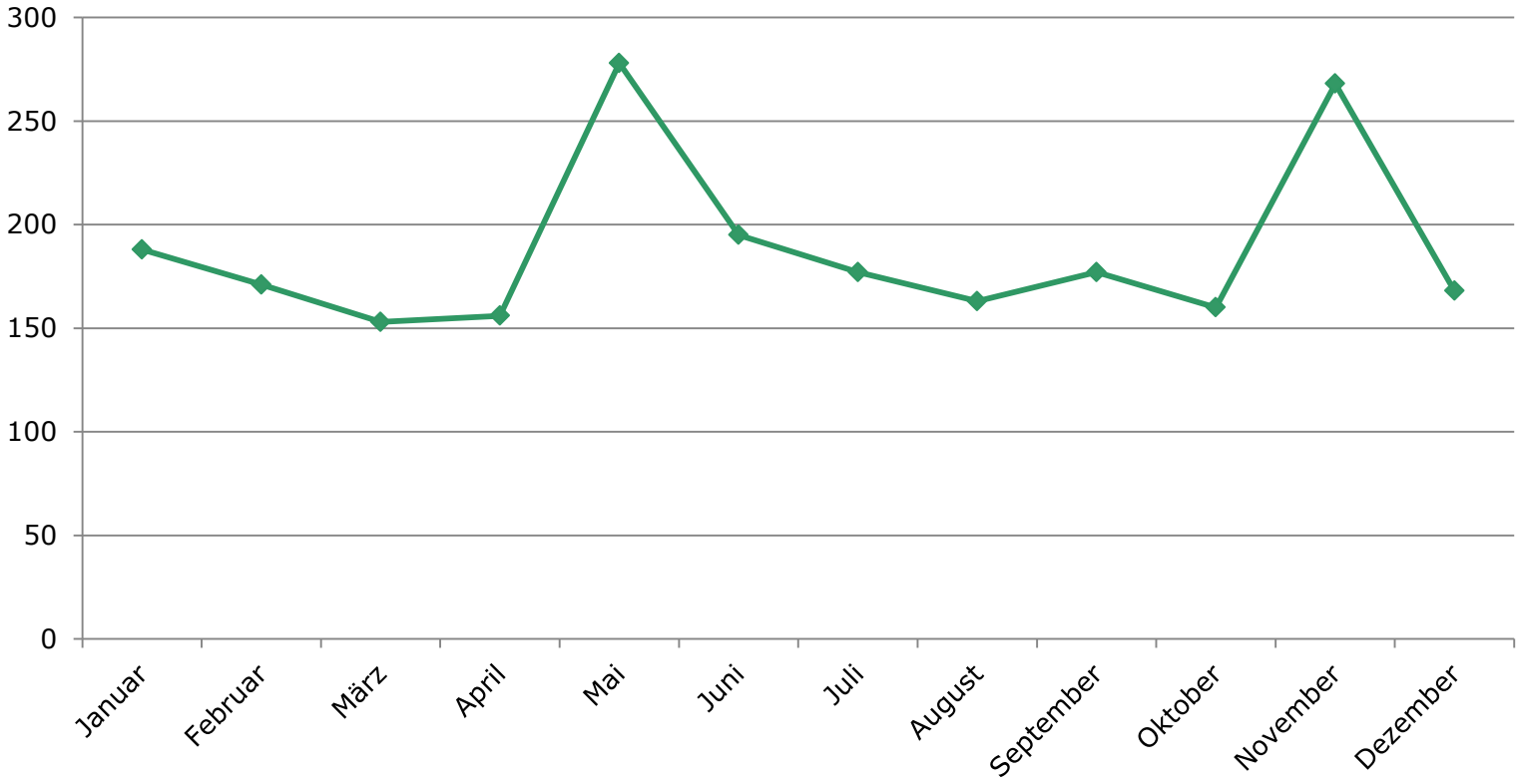
x-Achse: Jahr, in dem der Artikel exzellent wurde
y-Achse: Anzahl der Artikel, die in diesem Jahr exzellent wurden

Artikel nach Monat



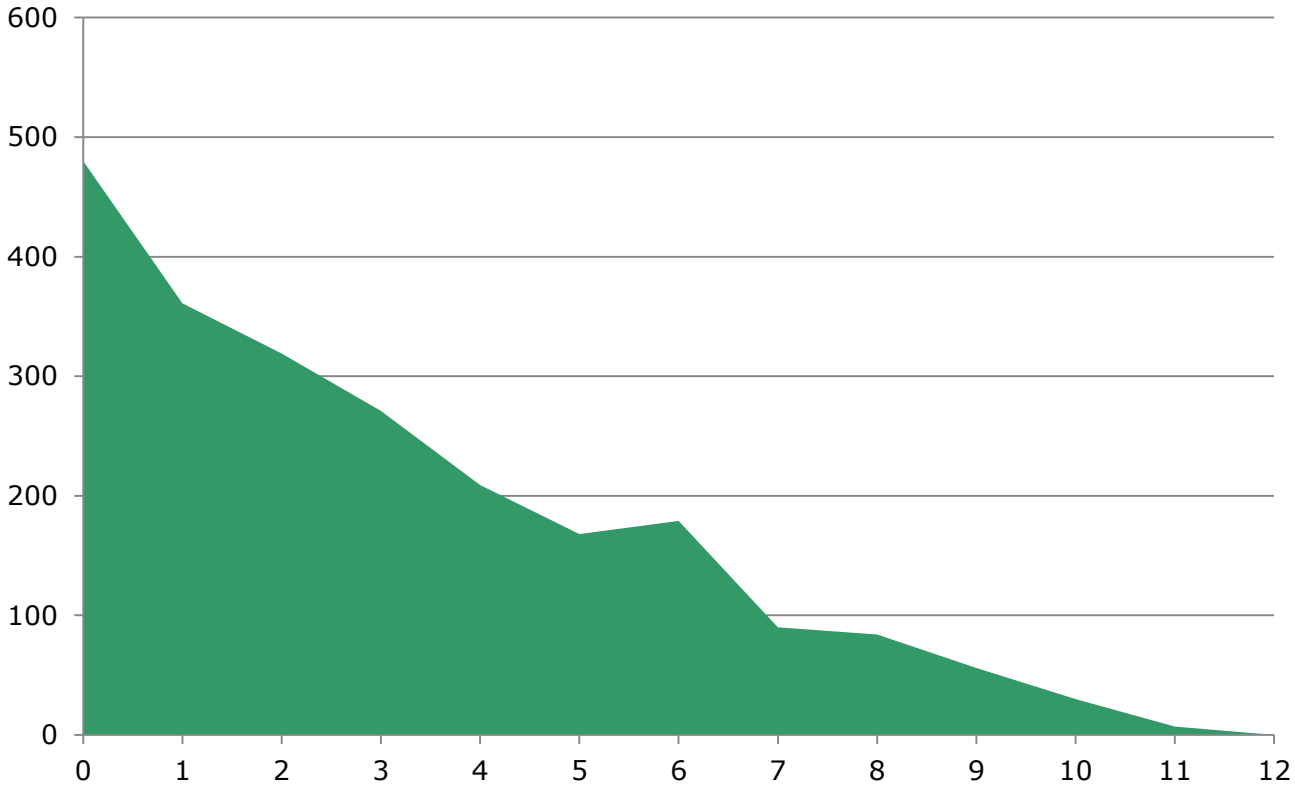
x-Achse: Erstellungsmonat des exz. Artikels
y-Achse: Anzahl der in diesem Monat erstellten exz. Artikel

Artikel nach Monat



x-Achse: Monat, in dem der Artikel exzellent wurde
y-Achse: Anzahl der Artikel, die in diesem Monat exzellent wurden

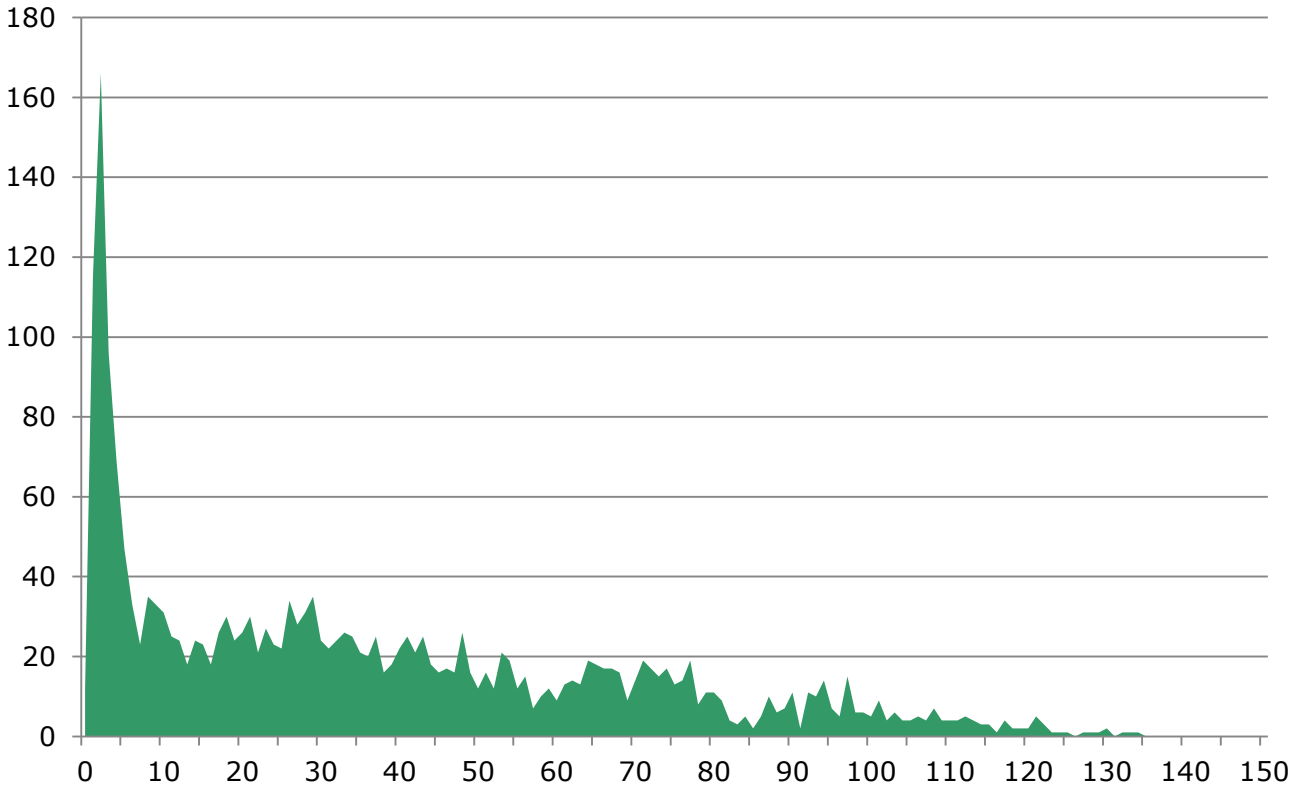
Dauer bis zur Exzellenz



Arithm. Mittel	3,0 Jahre
Maximum	11 Jahre
Minimum	0 Jahre

x-Achse: Jahre zwischen Erstellung und Exzellenzwerdung
y-Achse: Anzahl der Artikel in diesen Jahren

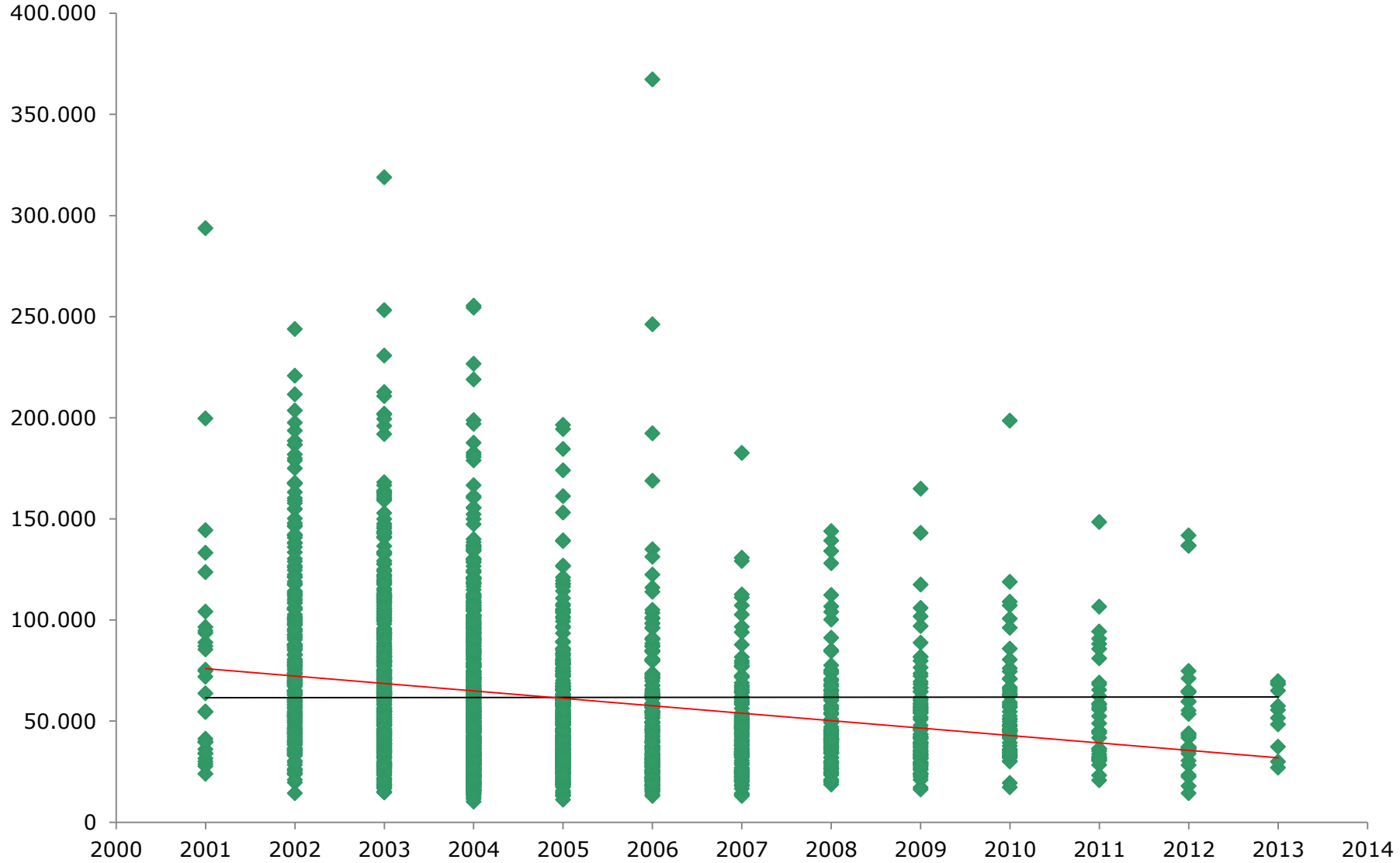
Dauer bis zur Exzellenz



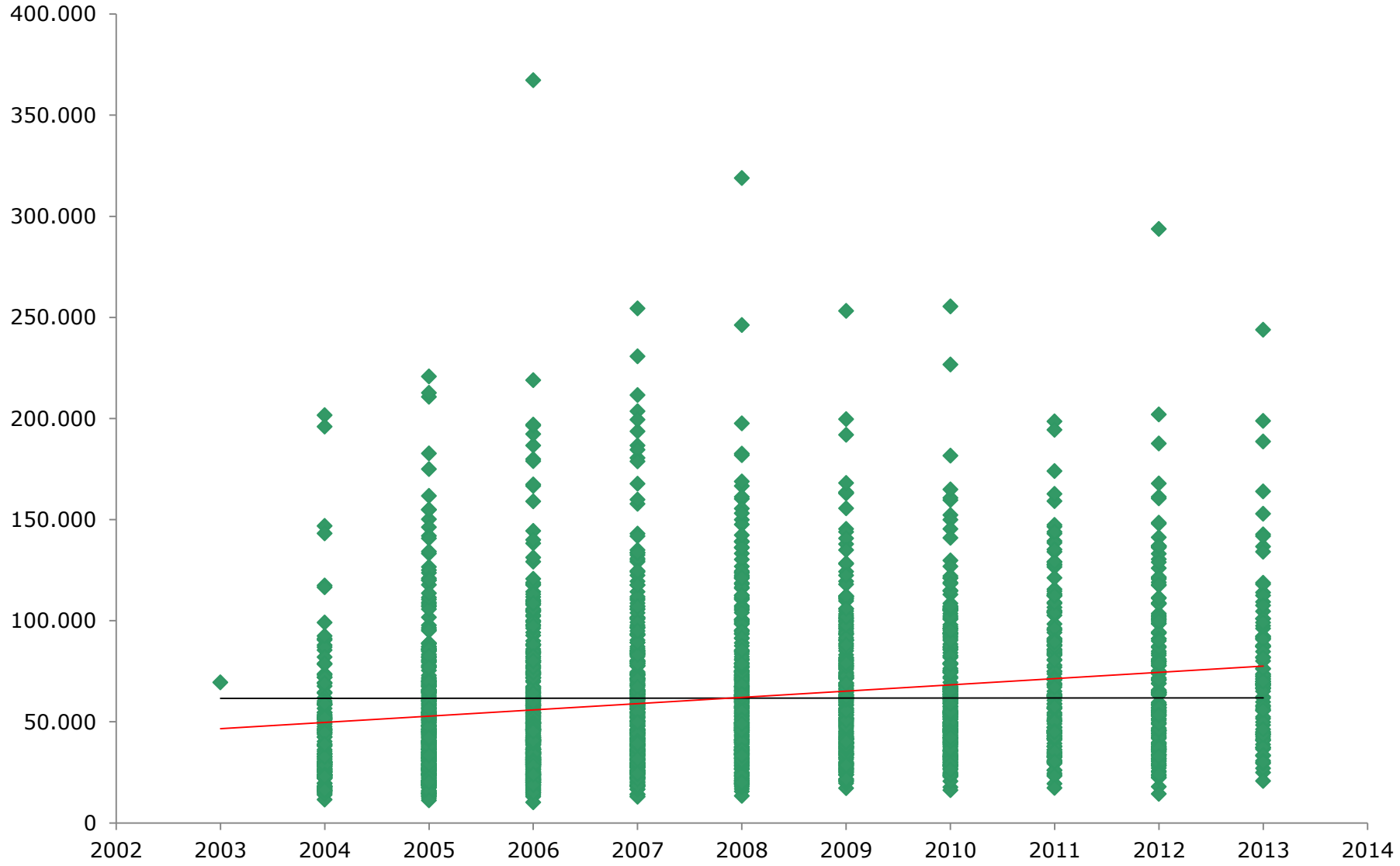
Arithm. Mittel	36,3 Monate
Maximum	134 Monate
Minimum	0 Monate

x-Achse: Monate zwischen Erstellung und Exzellenzwerdung
y-Achse: Anzahl der Artikel in diesen Monaten

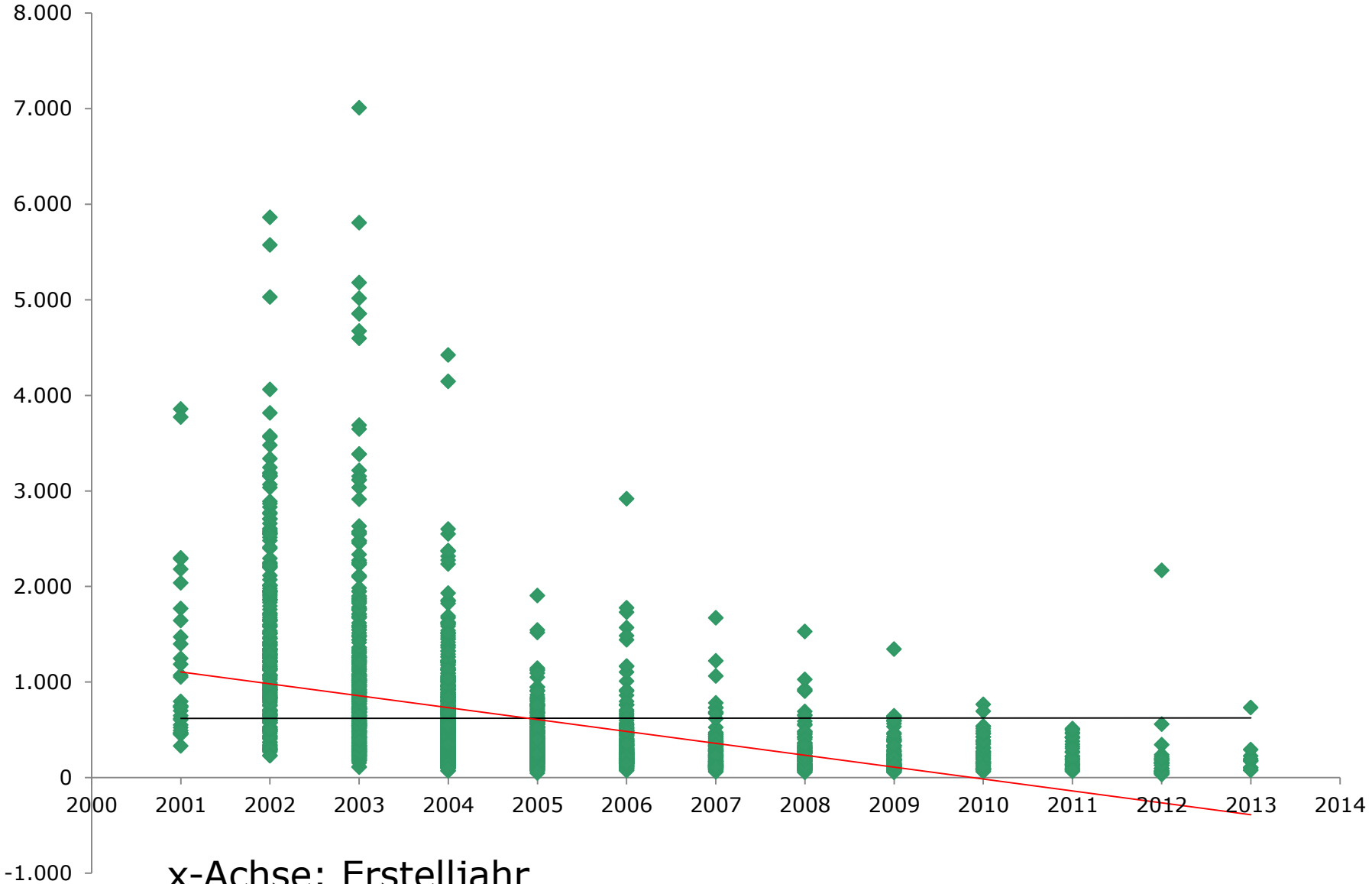
Entwicklungen über die Jahre hinweg



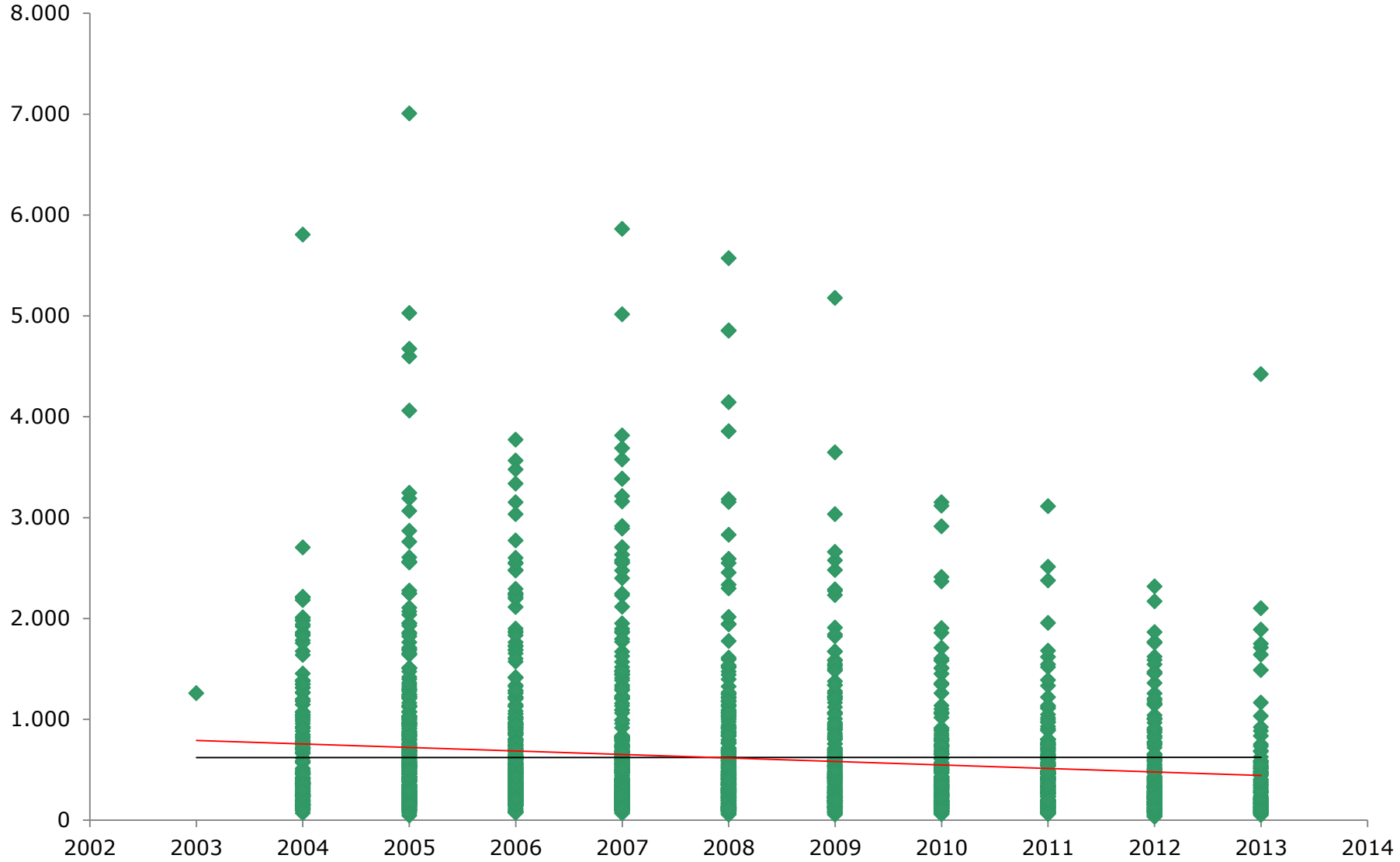
x-Achse: Erstelljahr
y-Achse: Bytes



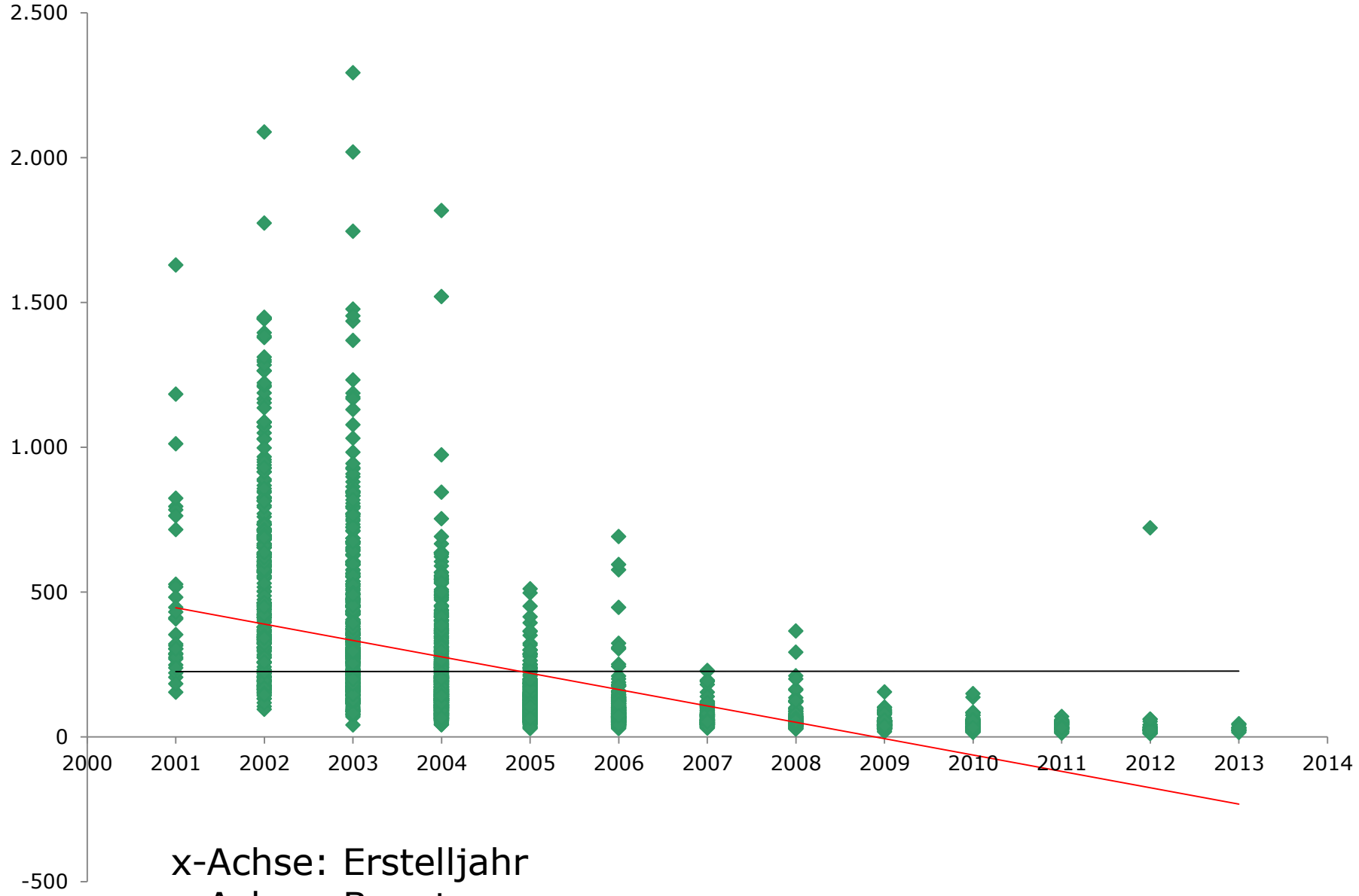
x-Achse: Exzellenzjahr
y-Achse: Bytes



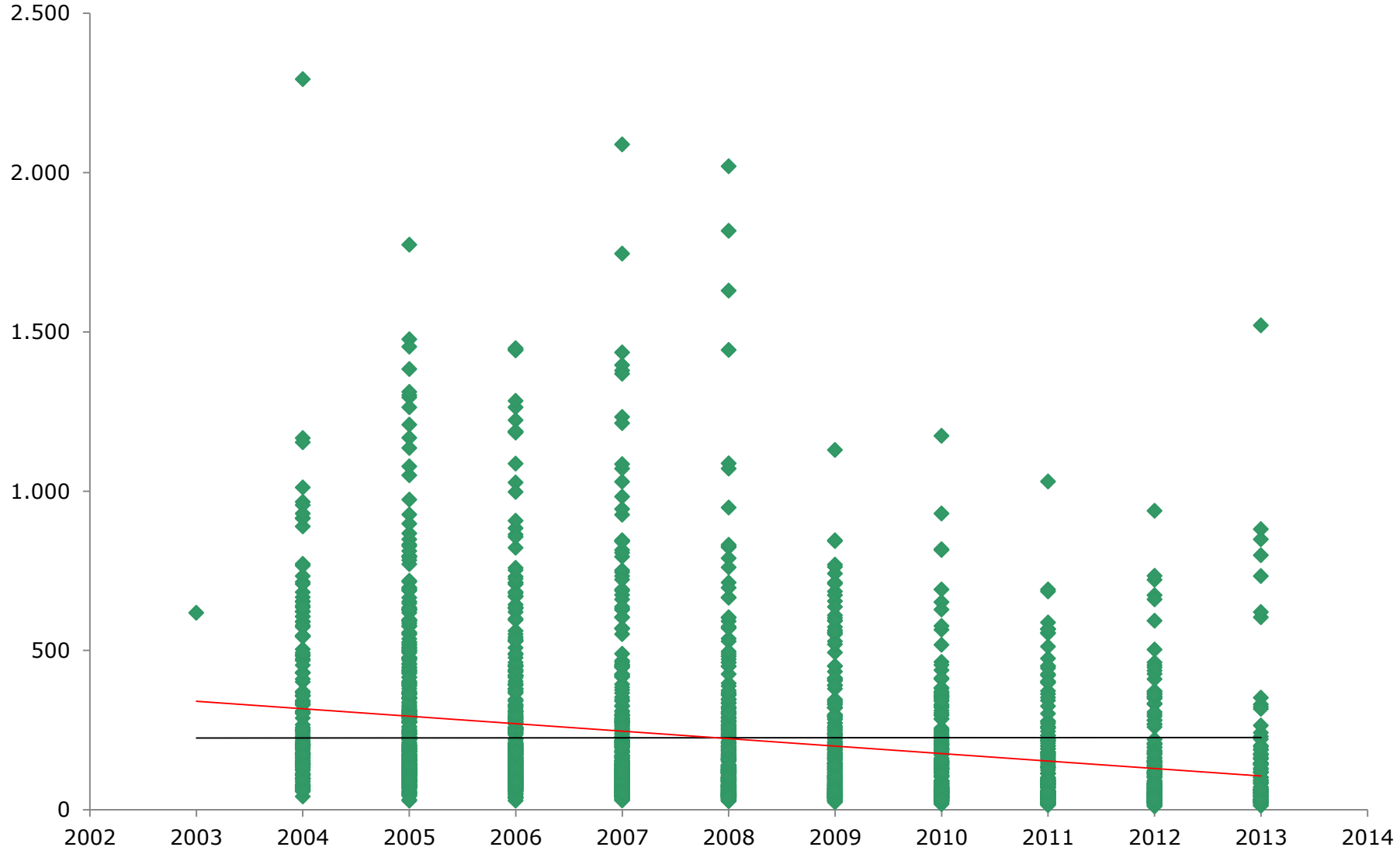
x-Achse: Erstelljahr
y-Achse: Versionen



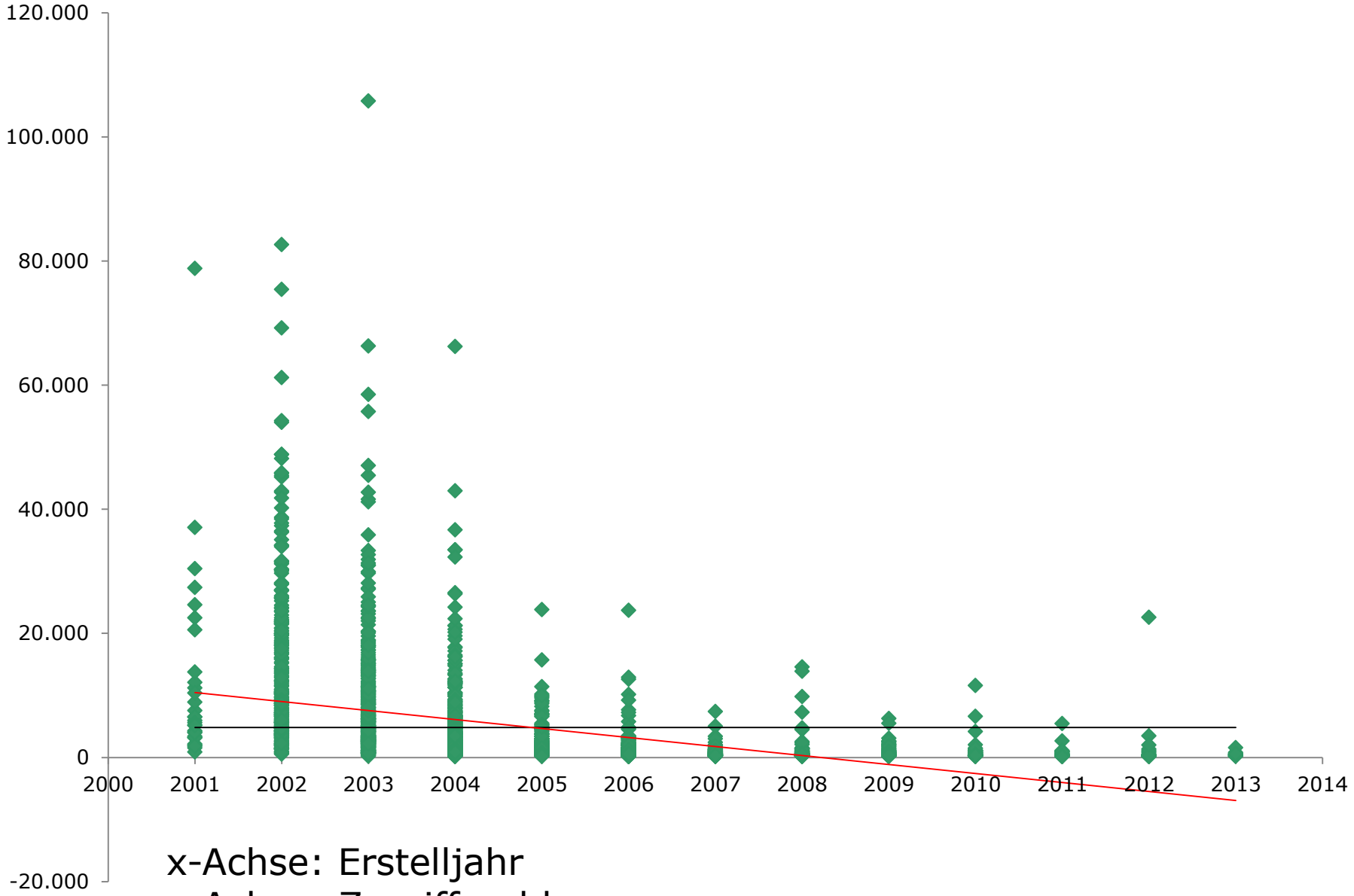
x-Achse: Exzellenzjahr
y-Achse: Versionen



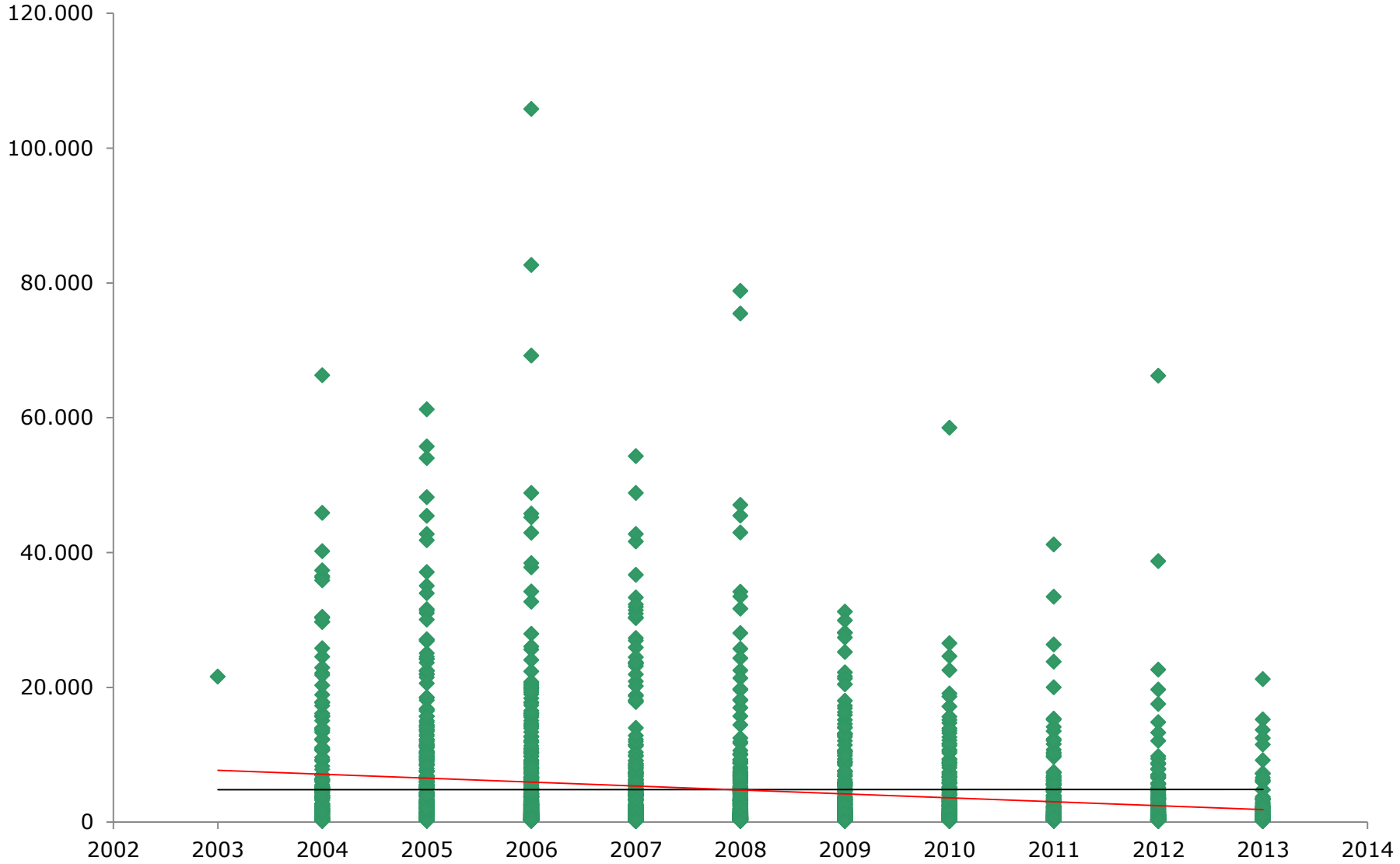
x-Achse: Erstelljahr
y-Achse: Benutzer



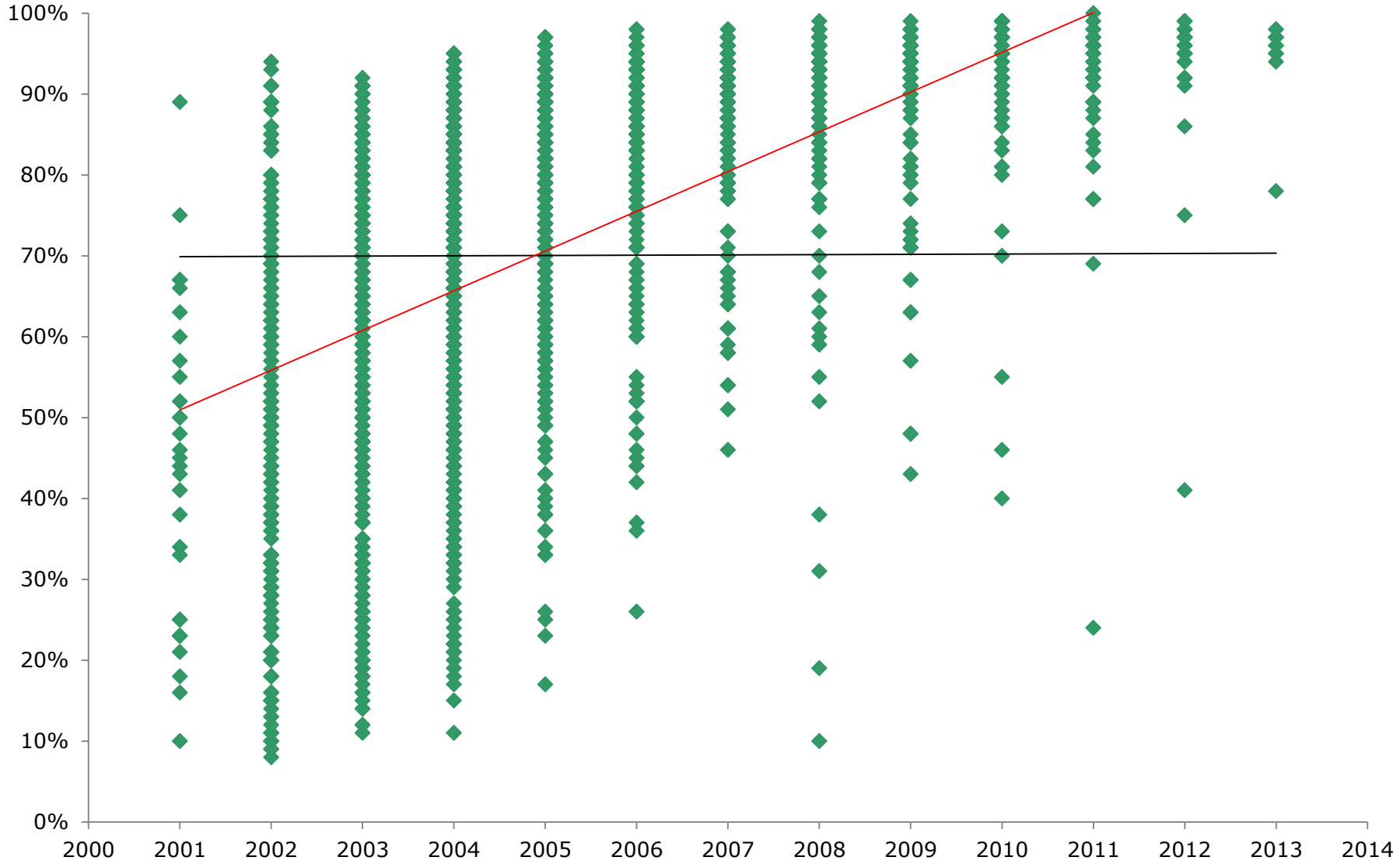
x-Achse: Exzellenzjahr
y-Achse: Benutzer



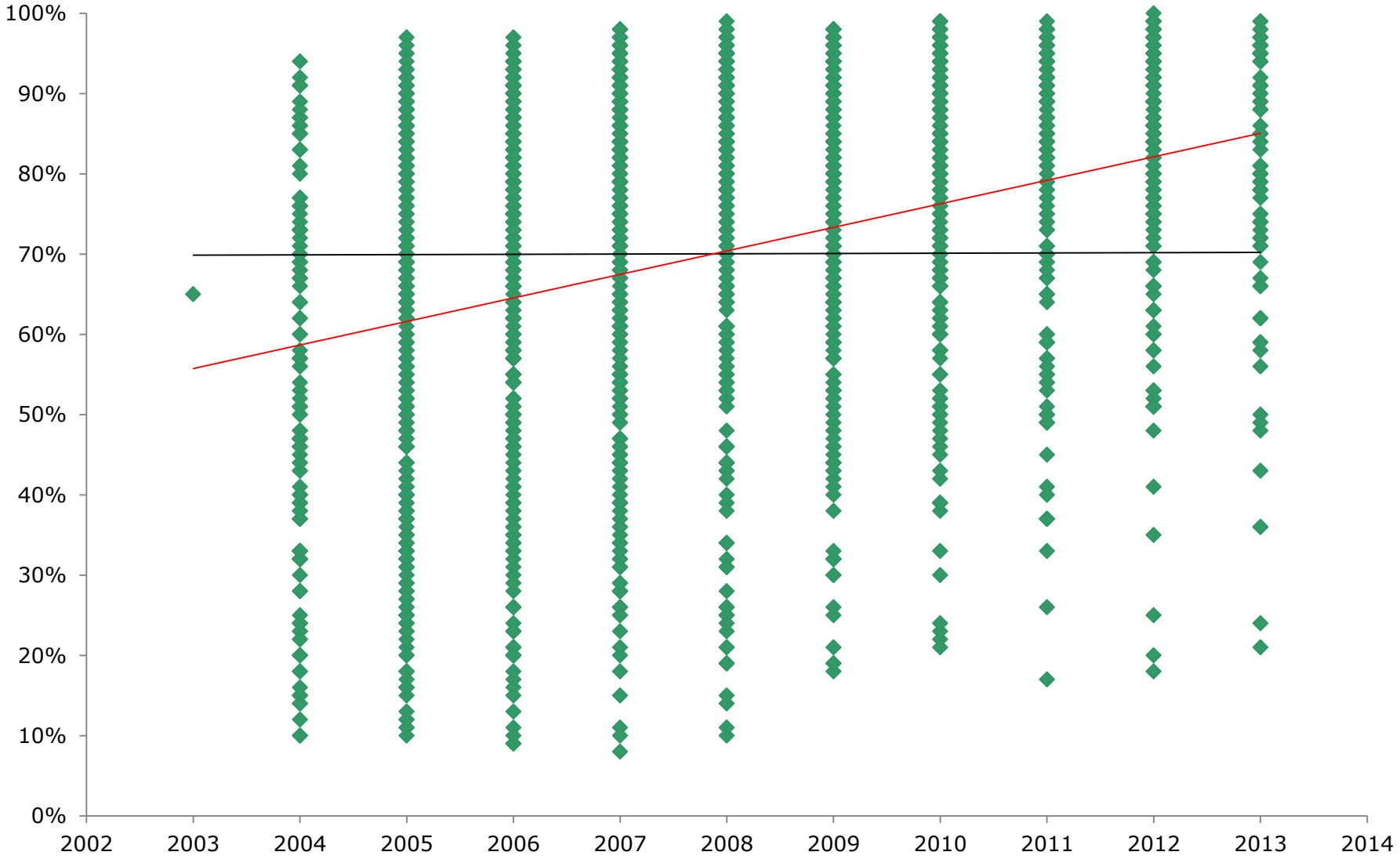
x-Achse: Erstelljahr
y-Achse: Zugriffszahl



x-Achse: Exzellenzjahr
y-Achse: Zugriffszahl

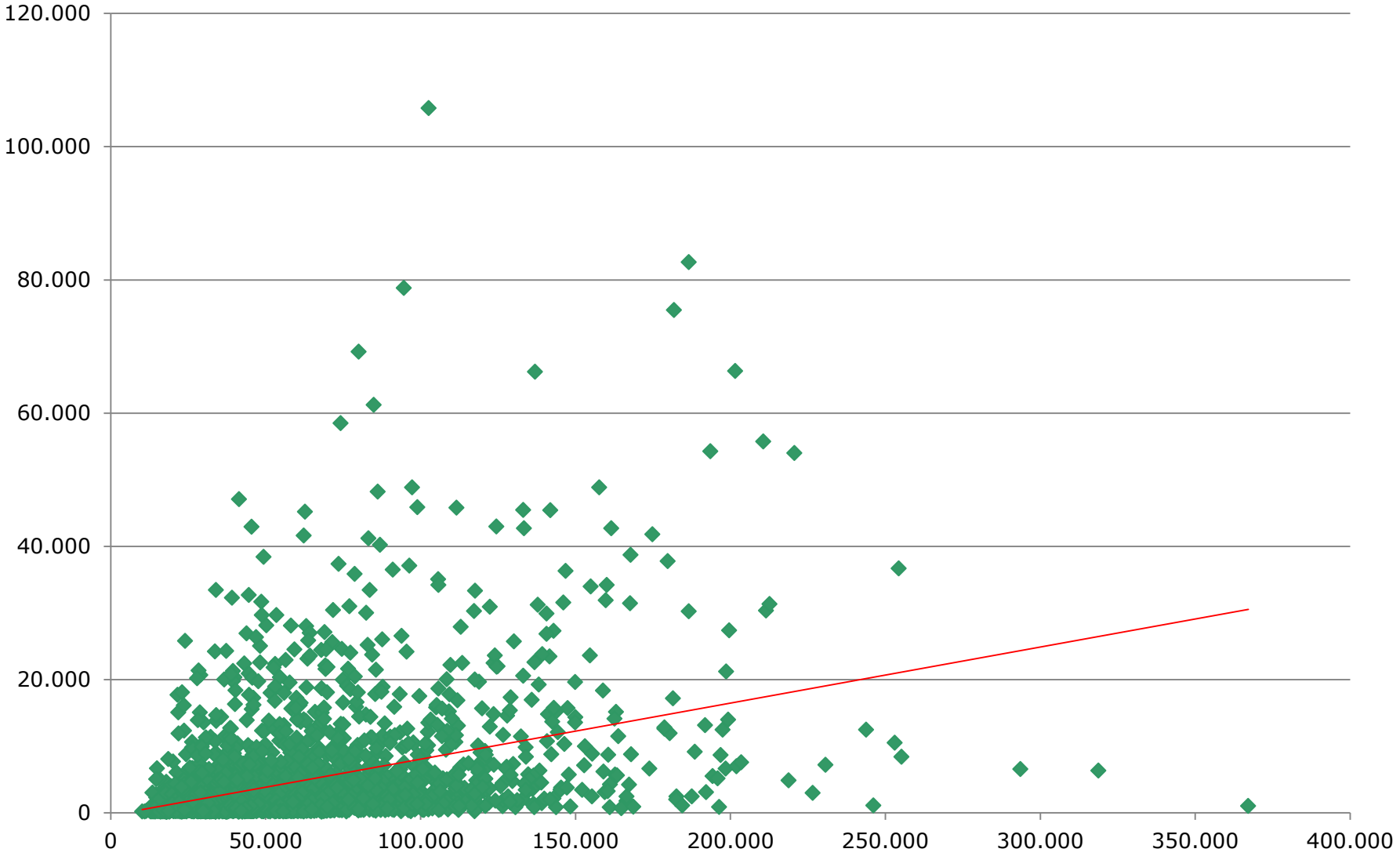


x-Achse: Erstelljahr
y-Achse: Zeichenanteil des 1. Autors

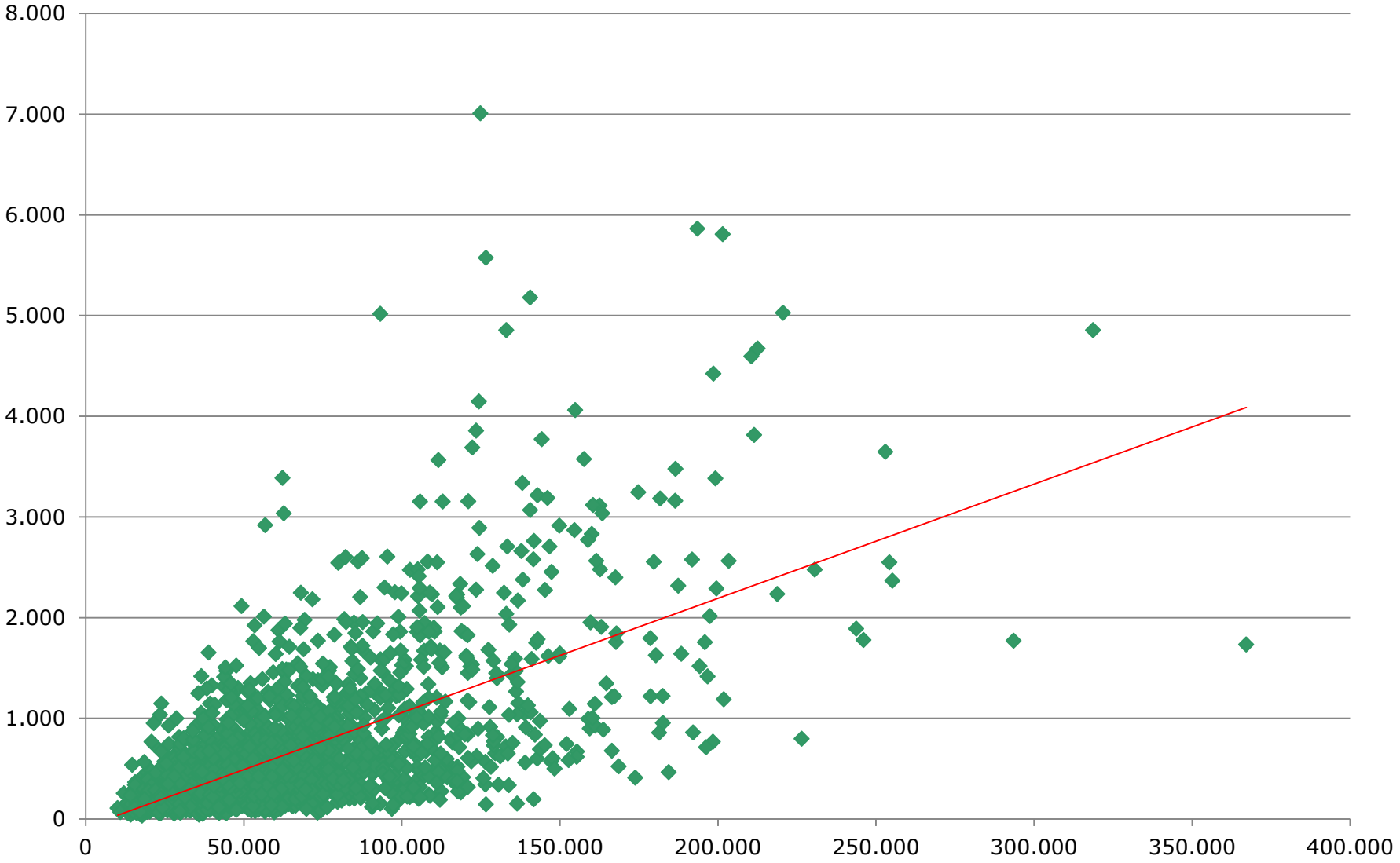


x-Achse: Exzellenzjahr
y-Achse: Zeichenanteil des 1. Autors

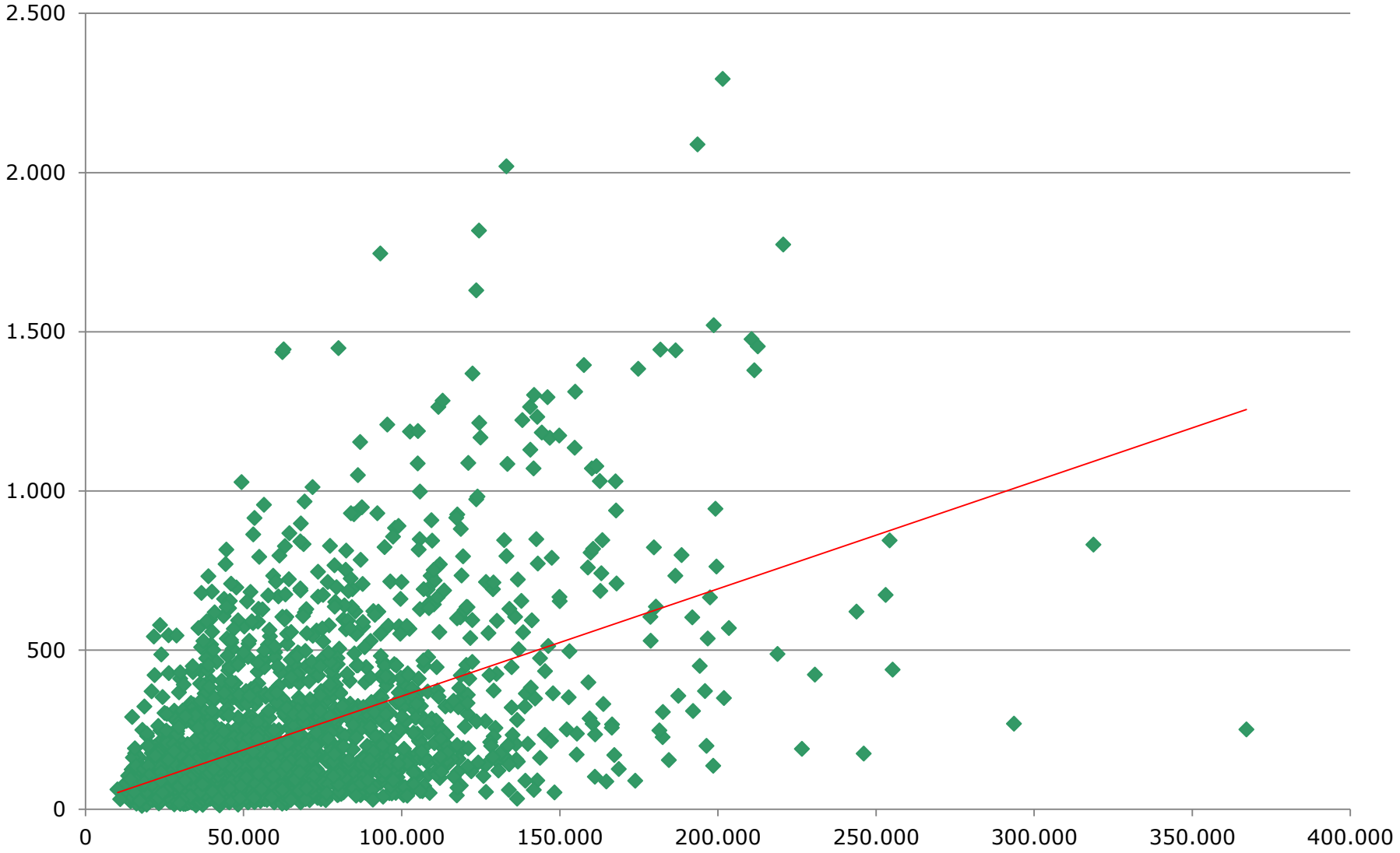
weitere Korrelationen



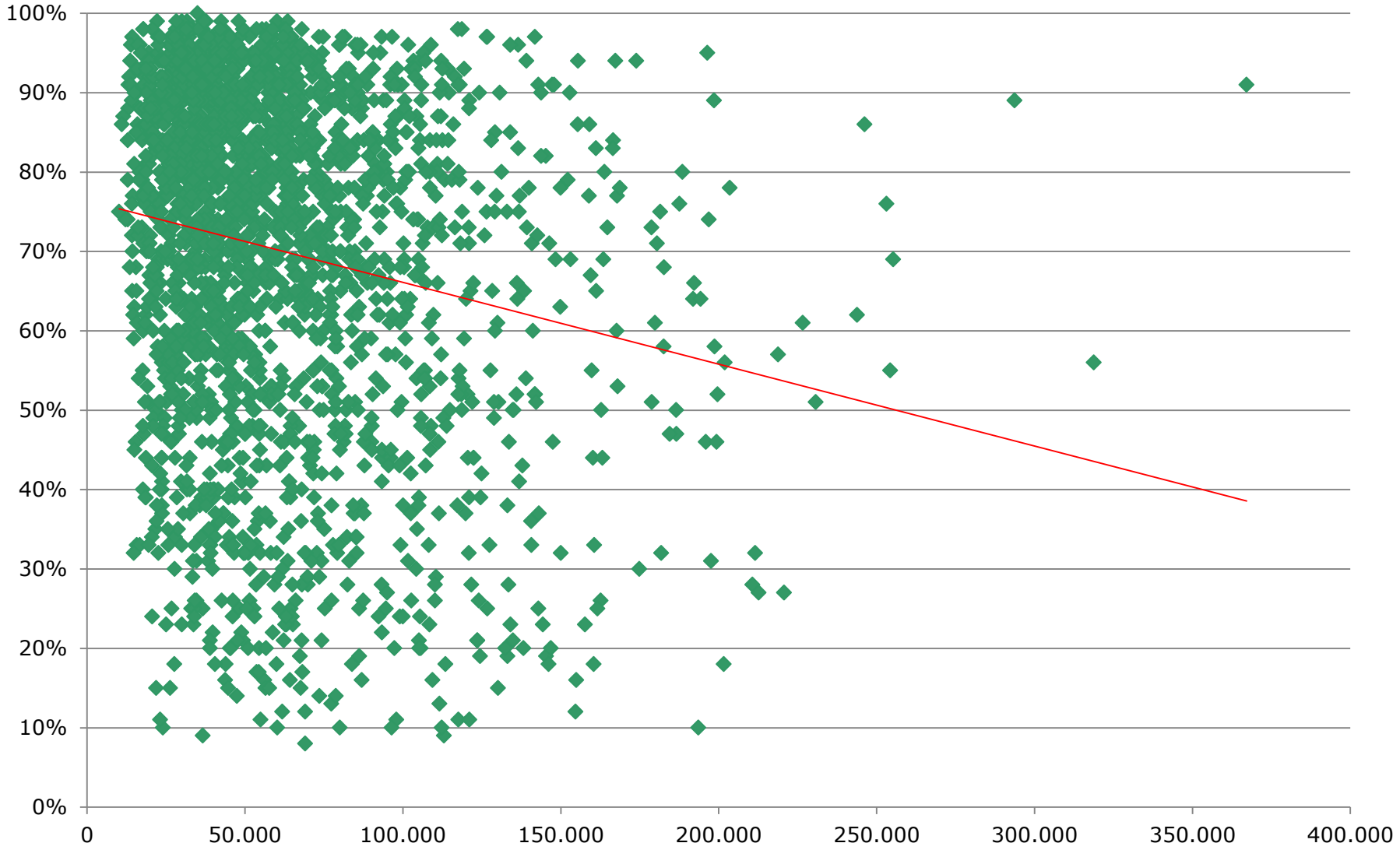
x-Achse: Bytes
y-Achse: Zugriffszahl



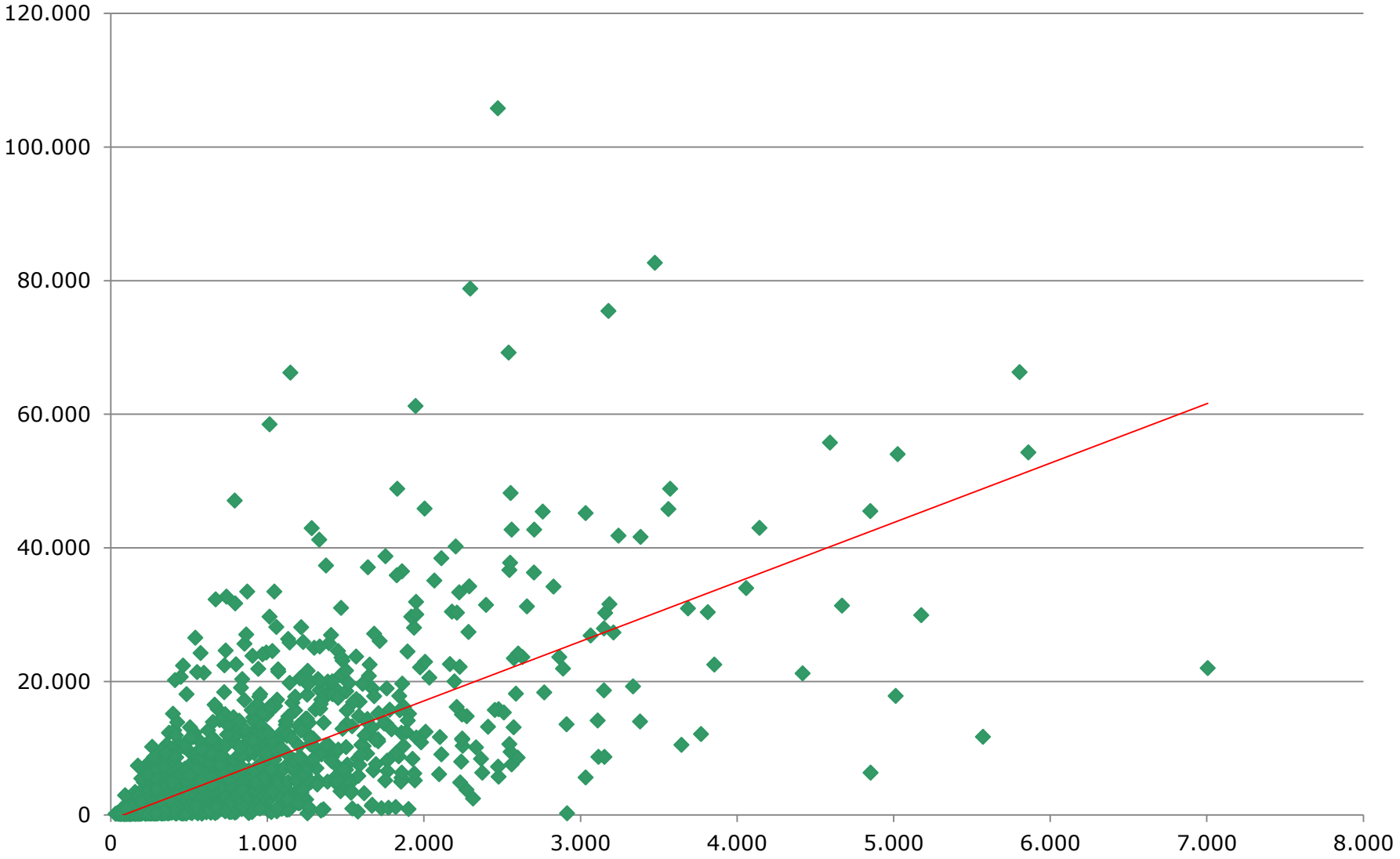
x-Achse: Bytes
y-Achse: Versionen



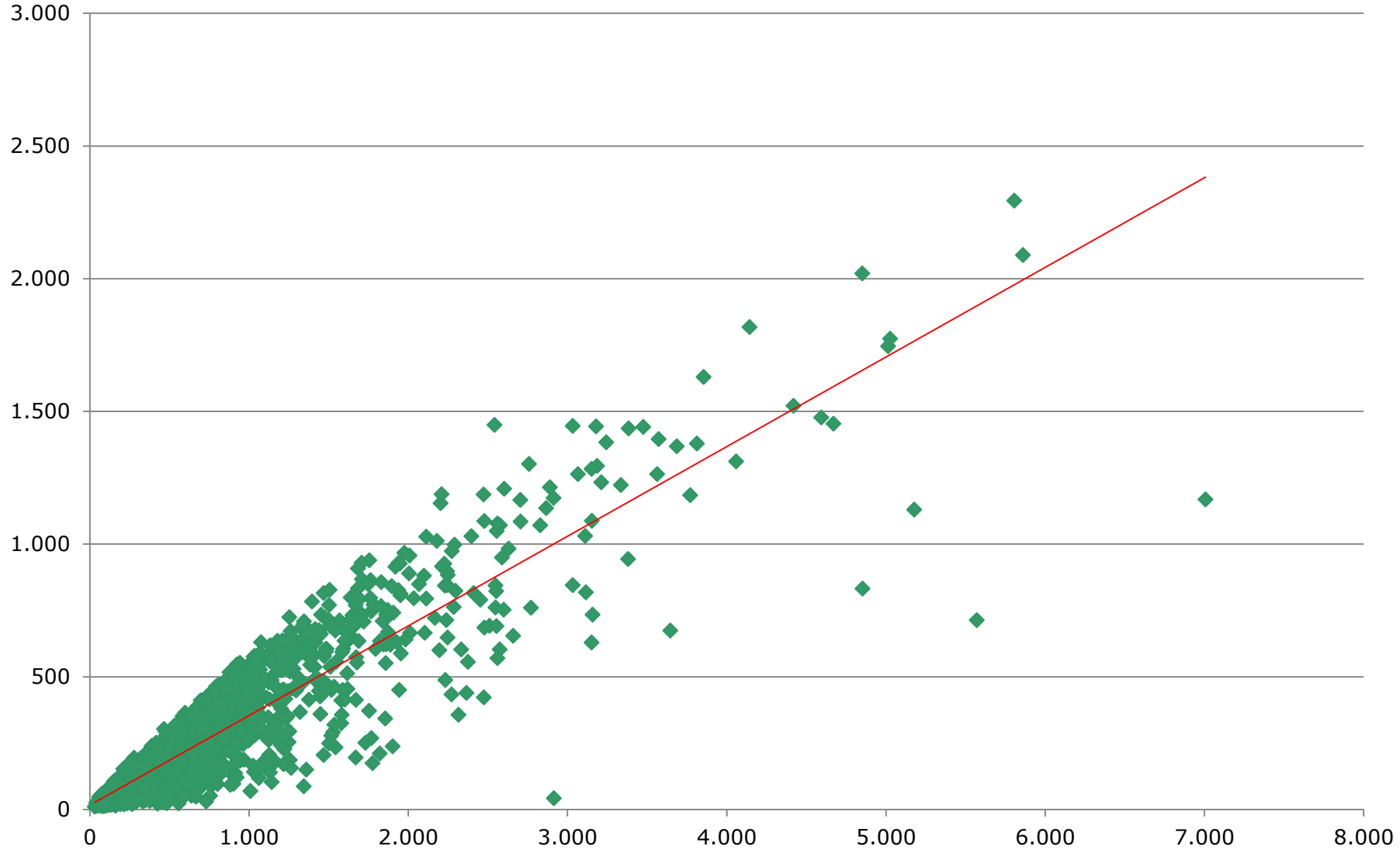
x-Achse: Bytes
y-Achse: Benutzer



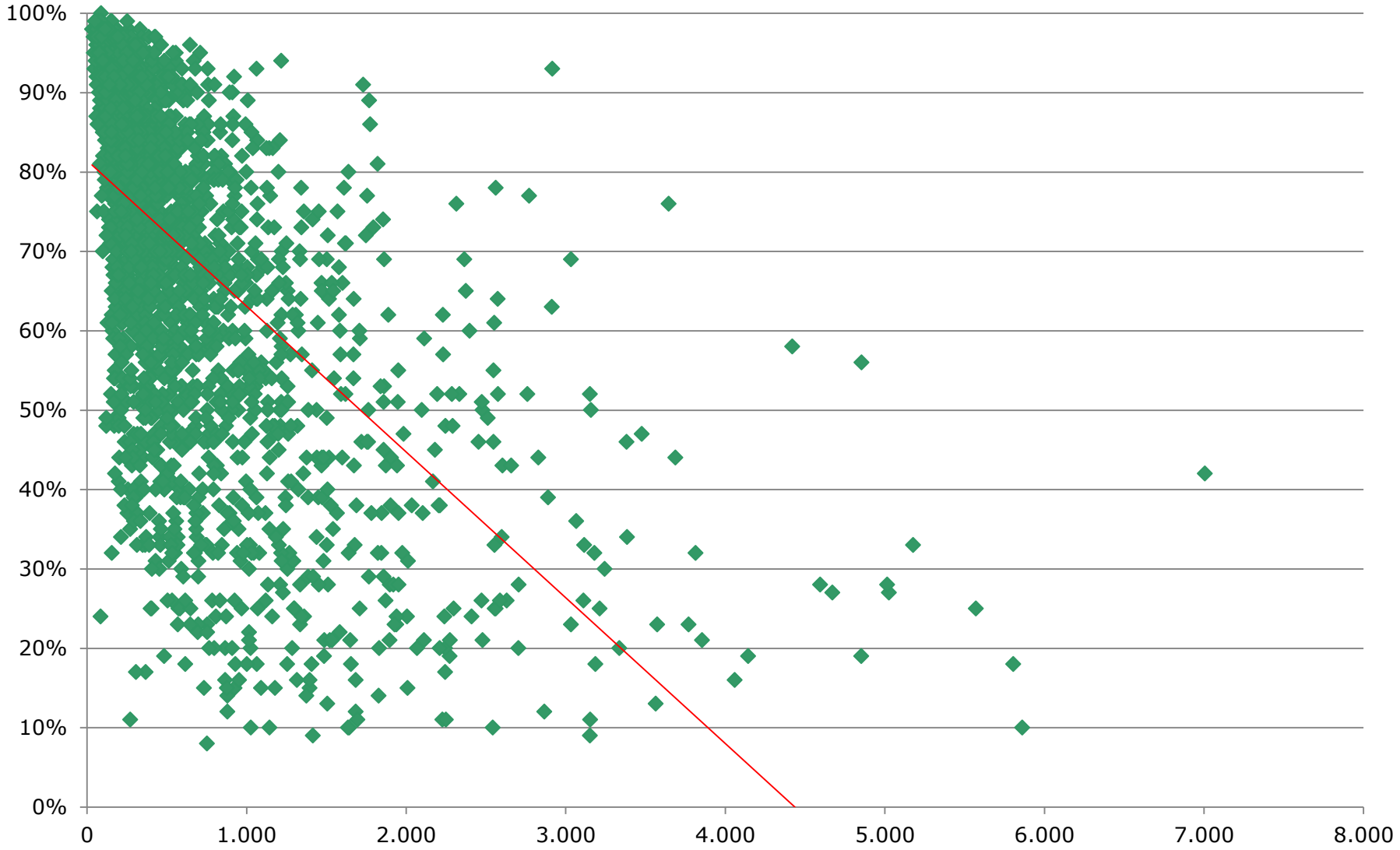
x-Achse: Bytes
y-Achse: Zeichenanteil 1. Autor



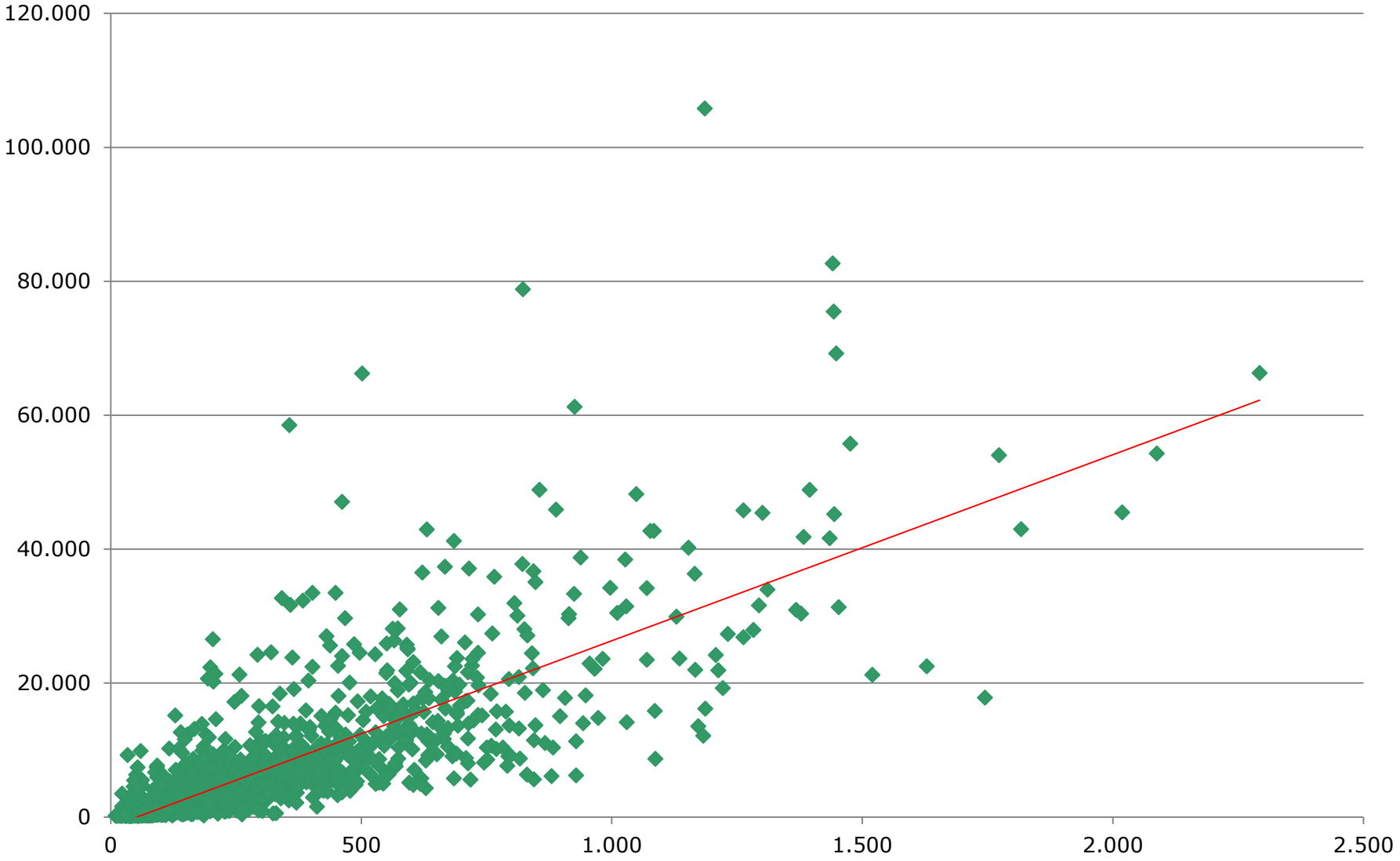
x-Achse: Versionen
y-Achse: Zugriffszahl



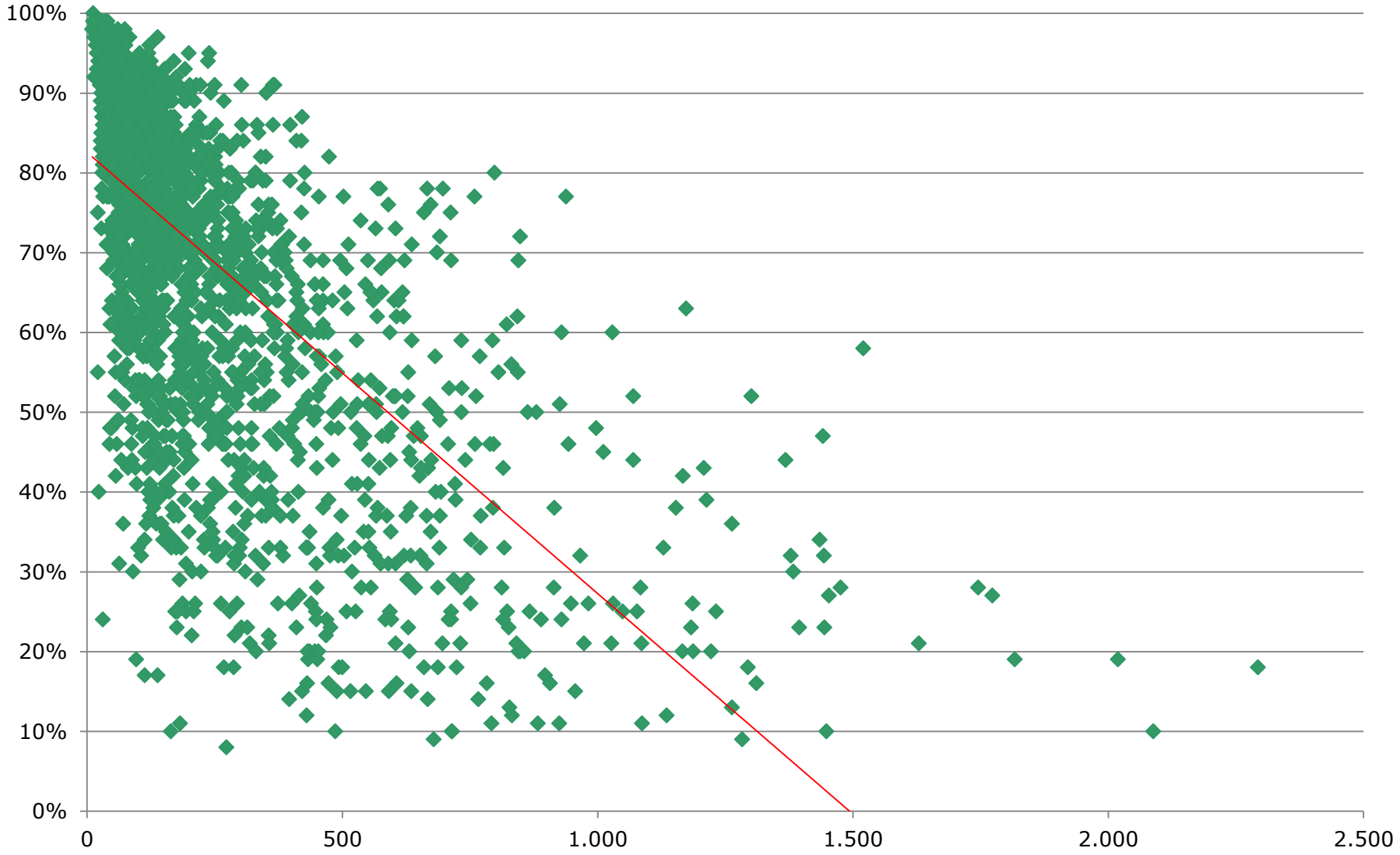
x-Achse: Versionen
y-Achse: Benutzer



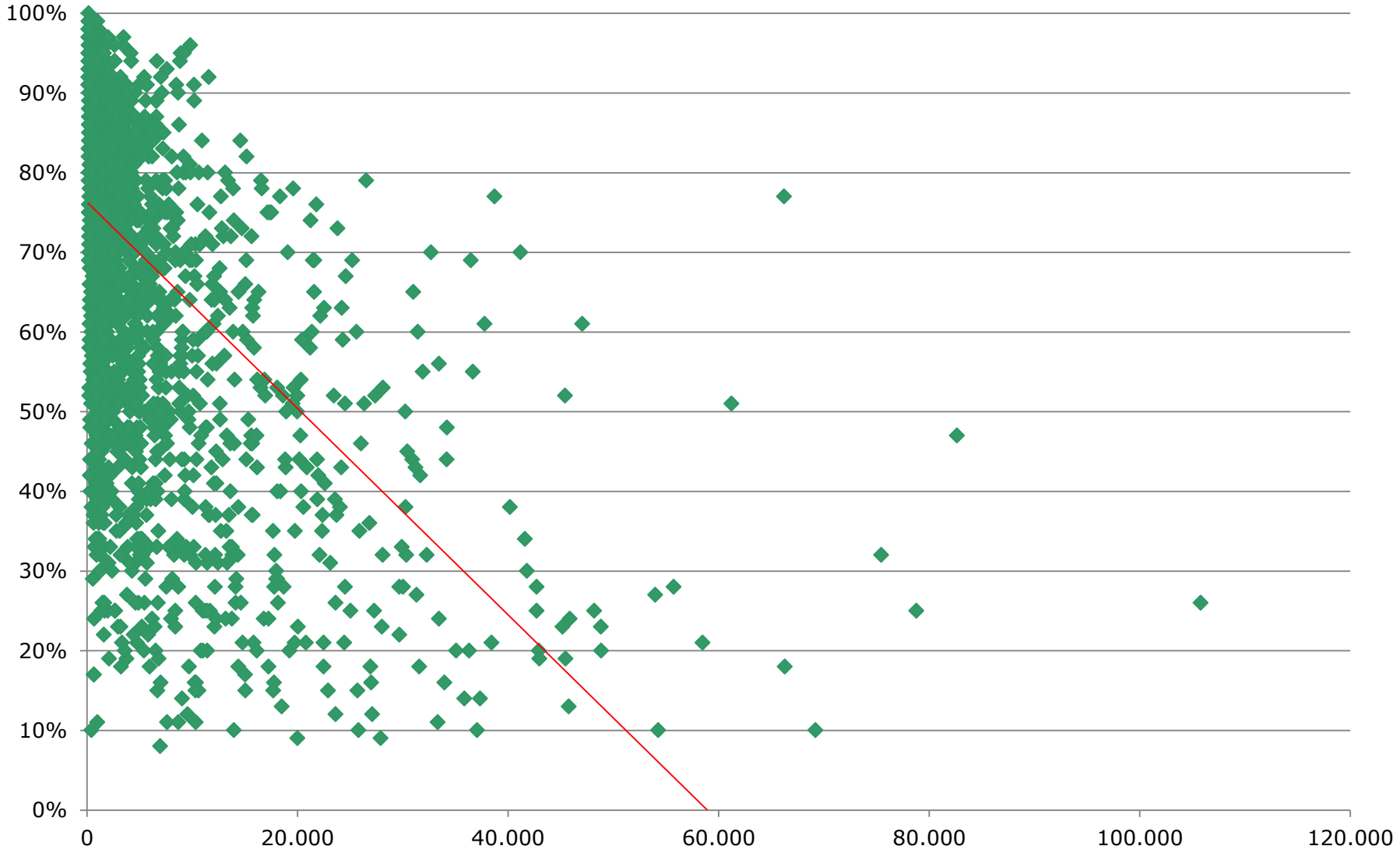
x-Achse: Versionen
y-Achse: Zeichenanteil 1. Autor



x-Achse: Benutzer
y-Achse: Zugriffszahl



x-Achse: Benutzer
y-Achse: Zeichenanteil 1. Autor



x-Achse: Zugriffszahl
y-Achse: Zeichenanteil 1. Autor

Korrelationen in Zahlen

- untersucht die jeweils 20 Artikel mit der größten und kleinsten Anzahl {resp. die Artikel über bzw. unter dem Durchschnitt} (Bytes, Versionen, Benutzer, Bearbeitungen/Benutzer, Zugriffszahlen, Zeichenanteil des 1. Autors und Bearbeitungsanteil)
  - Artikel mit vielen Bytes haben auch viele Versionen und Zugriffe (400–500 %) und einen geringeren Zeichenanteil des 1. Autors und sind früher entstanden – Artikel mit besonders wenigen Bytes wurden aber besonders früh exzellent
  - Artikel mit vielen Versionen resp. Benutzern haben einen signifikant kleineren Zeichenanteil des 1. Autors (<50 %), dafür signifikant größere Zugriffszahlen (>650%) und sind zudem früher erstellt (über zwei Jahre gegenüber dem Schnitt)

Korrelationen in Zahlen

- Artikel mit einem hohen Bearbeitungen-je-Benutzeranteil haben einen deutlich höheren Zeichenanteil des 1. Autors, besonders wenige beteiligte Benutzer und Zugriffe (<20 %) und sind im Schnitt über 4 Jahre später entstanden (und wurden rund 2 Jahre später exzellent); bei 50 % der Artikel mit einem geringen derartigen Anteil sind IPs Hauptautoren
- Artikel mit einer hohen Zugriffszahl haben einen deutlich geringen Zeichenanteil des 1. Autors (<45 %), dafür viele Versionen (>200 %) und viel mehr Versionen und Bearbeiter (je >500 %) – sie werden früher erstellt und exzellent
- Artikel mit einem hohen Zeichenanteil des 1. Autors sind deutlich kleiner (~ 75 %) und haben signifikant weniger Versionen, Benutzer und Zugriffe (~ 20–5 %) – sind über 5 Jahre später erstellt und über 3 Jahre später exzellent geworden als der Schnitt; dafür wurden Artikel mit einem geringen Zeichenanteil des 1. Autors im Schnitt mehr als 2 Jahre eher erstellt und exzellent, bei 25 % von ihnen sind IPs Hauptautoren

Wer schreibt exzellente Artikel? Eine statistische Auswertung.

Martin Rulsch (DerHexer)
derhexer87@yahoo.de
WikiConvention 2013
Karlsruhe, 23. November

WIKIMEDIA
DEUTSCHLAND

Appendix: Artikel bei Exzellenzwerdung

Methode	Bytes	Zeichen-anteil 1. Autor	Zeichen-anteil 1.–2. Autor	Zeichen-anteil 1.–5. Autor
Arithm. Mittel	54.255	78 %	86 %	92 %
Geom. Mittel	45.716	75 %	85 %	92 %
Median	45.610	84 %	91 %	96 %
Maximum	301.724	100 %	100 %	100 %
Minimum	5.121	9 %	12 %	21%

Appendix: Artikel bei Exzellenzwerdung

- bei 97,5 % der exzellenten Artikel entspricht die Person mit dem höchsten Zeichenanteil der heutigen Person mit eben diesem (nur bei 57 Artikeln hat er gewechselt)
- 12,6 % der exzellenten Artikel haben von Exzellenzwerdung bis heute Bytes verloren (290 Artikel, min. +293.772 Bytes, min. –105.353 Bytes, Schnitt +7.742 Bytes)
- bei 3,3 % der Artikel hat die Person mit dem höchsten Zeichenanteil diesen noch ausbauen können (76 Artikel)
- im Schnitt gewinnt ein Artikel 19 % an Bytes nach Exzellenzwerdung dazu (max. 972 %, min. –35 %)