DEUTSCHE INITIATIVE
FÜR NETZWERKINFORMATION E.V.

Nutzungsstatistiken elektronischer Publikationen

DFG-Projekt Open Access-Statistik (OA-S)
und
Arbeitsgruppe „Elektronisches Publizieren“

DINI Schriften 11-de
[Version 1.0, August 2009]

DEUTSCHE INITIATIVE
FÜR NETZWERKINFORMATION E.V.

Nutzungsstatistiken elektronischer Publikationen

DFG-Projekt Open Access-Statistik (OA-S)
und
Arbeitsgruppe „Elektronisches Publizieren“

DINI Schriften 11-de

[Version 1.0, August 2009]

Impressum
DINI – Deutsche Initiative für Netzwerkinformation e. V.
Geschäftsstelle
c/o Niedersächsische Staats- und Universitätsbibliothek Göttingen
Papendiek 14
37073 Göttingen
Tel.: 0551-39-3857
Fax: 0551-39-3856
E-Mail: gs@dini.de
www.dini.de

Inhaltsverzeichnis

Über DINI **4**

1 Einleitung **5**

2 Notwendige Eigenschaften der Beschreibung eines Nutzungsvorgangs **7**

2.1 Notwendiger Datenschutz 7

2.2 Zwischenspeicherung (Caching) 8

2.3 Roboterzugriff 9

2.4 Zählhorizont 9

3 Optionale Eigenschaften der Beschreibung eines Nutzungsvorgangs **9**

3.1 User-IP 9

3.2 Verweisende Ressource (Referer) 10

4 Veranschaulichung anhand einer Apache-Skizze **10**

5 Dienstart **12**

6 Beispielimplementation des Zählhorizonts **13**

7 Berechnungsstandards **14**

8 Anforderungsvergleich der verschiedenen Berechnungsstandards **16**

9 Ursachen von Messabweichungen **18**

10 Referenzen **22**

11 Anmerkung **23**

Aufnahmeantrag für die Mitgliedschaft in DINI e. V. **24**

Über DINI

Die Entwicklung der modernen Informations- und Kommunikationstechnologie verursacht einen Wandel innerhalb der Informationsinfrastrukturen der Hochschulen und anderer Forschungseinrichtungen. Dieser Wandel ist ein zentrales Thema in der deutschen Hochschullandschaft und setzt mehr als bisher Absprachen, Kooperationen, Empfehlungen und Standards voraus. Die Deutsche Initiative für Netzwerkinformation (DINI) unterstützt diese Entwicklung.

DINI wurde gegründet, um die Verbesserung der Informations- und Kommunikationsdienstleistungen und die dafür notwendige Entwicklung der Informationsinfrastrukturen an den Hochschulen sowie regional und überregional zu fördern. Durch Absprachen und Arbeitsteilung zwischen den Infrastruktureinrichtungen soll das Informationstechnik- und Dienstleistungsangebot weiter verbessert werden. Hierfür ist auch die gemeinsame Entwicklung von Standards und Empfehlungen erforderlich.

DINI ist eine Initiative der drei Partnerorganisationen:

- AMH (Arbeitsgemeinschaft der Medienzentren an Hochschulen e. V.),
- dbv (Deutscher Bibliotheksverband Sektion 4: Wissenschaftliche Universalbibliotheken) und
- ZKI (Zentren für Kommunikation und Informationsverarbeitung in Lehre und Forschung e. V.).

DINI verfolgt das Ziel,

- beispielhafte Lösungen bekannt zu machen und für die Nachnutzung zu empfehlen,
- die Erarbeitung, Anwendung und Weiterentwicklung von Standards anzuregen, zu unterstützen sowie Empfehlungen für deren Einsatz zu verbreiten,
- Kompetenzzentren zu registrieren und mit Hilfe moderner netzbasierter Instrumente bekannt zu machen,
- den übergreifenden Erfahrungsaustausch durch Tagungen, Workshops, Expertengespräche u. ä. zu verbessern,
- Förderprogramme bekannt zu machen und neue Programme anzuregen.

1 Einleitung

Elektronisches Publizieren und Open Access eröffnen neue Zugänge zu den Problemen des Wissenstransfers und des Forschungsmanagements. Forschende und Förderinstitutionen sind auf positive Bewertungen im Zuge von Qualitätssicherung und Evaluationen angewiesen, um ihre Mittel effizient und effektiv einsetzen zu können. Es ist im beiderseitigen Interesse, dass diese Bewertungsprozesse fair, valide, zeitnah und sparsam sind.

Im Bereich des Wissensmanagements stellt sich die Frage, wie Forschenden und Studierenden dabei geholfen werden kann, in der anwachsenden Flut von Informationen jene Objekte zu identifizieren, die sie benötigen.

Scientometrie und heute auch die Webometrie versuchen, die Auswirkungen neuer wissenschaftlicher Erkenntnisse innerhalb der Wissenschaft zu quantifizieren. Eine weit verbreitete Annahme ist, dass der gesellschaftliche Nutzen mit dieser Impactmessung korreliert. Das bekannteste Maß ist der Impact Factor (IF), entwickelt von Eugene Garfield. Unabhängig von seinen Vorzügen weist der Impact Factor Schwächen auf, die mit dem stetigen Wandel des Publikationsprozesses in den letzten Jahren zunehmen:

- willkürliche Stichprobengestaltung
- willkürlicher Erhebungszeitraum
- Granularität auf Zeitschriftentitelebene
- Bevorzugung englischsprachiger Publikationen
- Ausblenden von Inhalten, die nicht in bestimmten Zeitschriften erscheinen

Trotz dieser Schwächen ist der IF in einigen Wissenschaftsdisziplinen, wie z. B. den Naturwissenschaften, ein geeignetes Instrument zur Beschreibung der Leistungsfähigkeit wissenschaftlicher Zeitschriften. Sobald ein Fach jedoch vom angenommenen Publikationsverhalten abweicht, sind mit diesem Ansatz kaum noch qualitative Aussagen belegbar.

Beispiele hierfür sind:

- Veröffentlichung auf der eigenen Instituts- oder Mitarbeiter-Website (Informatik)
- lange Rezeptionszyklen (Mathematik)
- alternative Publikationswege (Geisteswissenschaften) oder
- Sprachbarrieren (Sprachwissenschaften).

Ein einzelner Kennwert ist nur in den seltensten Fällen ausreichend. Daher werden häufig mehrere Maßzahlen kombiniert, um reale Ereignisse möglichst zutreffend zu beschreiben. Brody et al. (2007) stellen dementsprechend die Frage, welche Kombination von Kennwerten für die Bewertung wissenschaftlicher Tätigkeiten

am besten geeignet ist. Jenseits des Gütekriteriums sollte hierbei nie vergessen werden, dass die Erhebung bestimmter Kennwerte sehr zeitaufwändig und ressourcenintensiv sein kann.

Statistiken über die Nutzung elektronischer Dokumente – als eine Möglichkeit – weisen einige interessante Eigenschaften auf:

- Nutzungsmessung in Echtzeit
- Objekte als Betrachtungsebene
- geringe Produktionskosten und -zeit
- verteilte Kennwertberechnung
- unabhängig von den Eigenschaften der Publikation (z. B. Herkunftsland, Sprache)
- Möglichkeit der Weiterverarbeitung (z. B. Nutzung im zeitlichen Verlauf)

Gerade für Informationsmanager ist die Weiterverarbeitung der einfachen Kennwerte, sprich Data Mining, von besonderem Interesse. Indem Verlaufsprofile, Dokumentprofile und Metadaten kombiniert werden, können Mehrwertdienste geschaffen werden.

Grundvoraussetzung im Bereich der Zugriffsmessung ist eine möglichst unveränderte Erhebung der originären Zugriffsprotokolle. Solange dies transparent geschieht (Merk et al. 2009), können in einem zweiten Schritt verschiedene Heuristiken und bereinigende Algorithmen angewandt werden, um eine faire Operationalisierung und Zählung durchzuführen. Hierbei müssen auch Heuristiken transparent beschrieben werden, und die Dokumentation eines Algorithmus sollte sich nicht auf die Kommentare im Quellcode beschränken.

Eine Reihe von Projekten (z. B. Publishing and the Ecology of European Research – PEER, Interoperable Repository Statistics – IRS) setzt sich mit dieser Fragestellungen auseinander. Zwar sind viele der grundlegenden Problemfelder dieselben, aufgrund von Unterschieden in den Zielvorstellungen und technischen Infrastrukturen werden Lösungen allerdings. wenn überhaupt, nur konzeptuell übertragbar sein.

Auf die Beschreibung eines normierten Log-Formats, wie dies in den DRIVER-Guidelines (http://www.driver-support.eu/documents/DRIVER_Guidelines_v2_Final_2008-11-13.pdf) erfolgt, wird in diesem Text verzichtet. Stattdessen werden die Art und der Umfang der Informationen beschrieben, die ein Repositorium an einen aggregierenden oder weiterverarbeitenden Dienst weitergeben sollte.

Von den bisherigen Erfahrungen und Erkenntnissen im Projekt Open Access-Statistik ausgehend, erscheint es sinnvoll, folgende Empfehlungen auszusprechen: Die Nutzung von JavaScript, POST und Cookies sollte für dokumentbezogene Handlungen, wie zum Beispiel den Abruf eines Volltexts vermieden werden, da

ein erheblicher Mehraufwand nötig ist, diese Ebenen mit den Informationen der HTTP-Get-Ebene zusammenzuführen. JavaScript-Aktionen müssten zum Beispiel, da sie auf der anfragenden Maschine ausgeführt werden, das Repositorium aktiv benachrichtigen, was durch Eingriffsmöglichkeiten von außen wiederum zu Verzerrungen führen kann.

2 Notwendige Eigenschaften der Beschreibung eines Nutzungsvorgangs

Um alle derzeit anerkannten Zugriffsstandards zu berechnen, sind folgende Daten zwingend notwendig:

- eineindeutige Bezeichnung des aufgerufenen Objekts
- Dateiformat des Objekts
- Dienstart
- Zugriffszeitpunkt, sekundengenau, inklusive Zeitverschiebungsinformation
- anfragende IP
- Hostname des anfragenden Systems
- anfragender User-Agent
- HTTP Status Code
- Zahl der übertragenen Bytes

2.1 Notwendiger Datenschutz

Die anfragende IP-Adresse darf aus Gründen des Datenschutzes nicht vorgehalten werden, stattdessen wird eine Streufunktion angewandt, sodass verschiedene Anfragen einem nutzenden System zugeordnet werden können, ohne dass dessen Identität ermittelbar ist.

Aufgrund des LogEc-Standards ist es jedoch nicht genug, einen Streuwert der IP zu speichern, zusätzlich muss ein Hash des C-Klasse-Anteils der IP gespeichert werden. Als Streuwert-Funktion empfiehlt sich SHA-256. Des Weiteren ist der Name der Top-Level-Domäne, zum Beispiel **dfg.de**, zu ermitteln, bevor durch die Umwandlung der IP-Adresse in einen Streuwert der DNS-LookUp unmöglich wird. Somit ergeben sich

- IP-Hash und
- C-Class-Hash

als dem deutschen Kontext geschuldete Konkretisierungen der anfragenden IP-Adresse.

Die Verwendung des Konzepts der C-Klasse ist aufgrund von Entwicklungen wie dem Classless Inter-Domain Routing (CIDR) und Internet Protocol Version 6 (IPv6) langfristig voraussichtlich problematisch.
Da dieselbe IP bei verschiedenen Repositorien den gleichen Streuwert aufweisen soll, ist es nicht möglich, so genannte Salted-Hashes zu verwenden. Bei dieser Methode würde vor der Hash-Berechnung eine jedes Mal neu, zufällig erzeugte Zeichenkette an den IP-String angehängt werden. Um jedoch wenigstens die Menge der möglichen Zustände des zu hashenden Strings zu vergrößern, empfiehlt es sich, einen über alle Repositorien konstanten String an alle IP-Adressen anzuhängen.

2.2 Zwischenspeicherung (Caching)

Um durch Caching und Prefetching für ein Repositorium nicht zählbare Rezeptionen zu verhindern, ist einer der folgenden Eigenschaftssätze im HTTP Response Header auszuweisen:

Variante 1:
- Cache-Control: must-revalidate
- ETag: ETag-Streuwert der Ressource

oder

Variante 2:
- Expires: 0
- Expires: (Zeitpunkt des HTTP Request des Klienten)
- Pragma: no-Cache
- Cache-Control: no-cache

Optionale Zusätze, die mit beiden Varianten kombinierbar sind:
- Verweise auf PDF-Dateien sind mit einem zufälligen Query-Argument zu versehen, um das Caching der Volltext-Datei zu vermeiden (z. B. beispiel.repositorium.de/volltexte/beispiel.pdf?**zufall=123**).
- Zusätzlich kann ein Verweis auf eine Grafik, die nicht existiert, in HTML-Seiten eingebunden werden. Da diese Grafik niemals ausgeliefert werden kann, wird selbst bei einer Auslieferung durch einen Cache diese eine Datei neu angefordert. Dies gilt jedoch nur, solange der Browser nicht durch Konfiguration oder Add-ons das Laden von Grafiken unterbindet.

2.3 Roboterzugriff

Es muss eine Datei robots.txt existieren, und auf diese muss entsprechend der Regeln für Roboterzugriffssteuerung verwiesen werden. Allerdings kann diese Datei, damit die Texte indexiert werden, schlicht leer sein. Wichtig ist nur, dass festgestellt werden kann, ob ein Akteur diese Datei abgerufen hat oder nicht.

2.4 Zählhorizont

Da Objekte häufig verteilt über mehrere Server zugänglich gemacht werden, ist es nötig zu erfassen, welche Ereignisse vom Repositorium erhoben werden können, um für das Objekt aussagelose Kennwerte erst gar nicht anzuzeigen. Wenn von einem Dokument zum Beispiel aufgrund eines Link Resolving keine Volltextzugriffe dem Repositorium bekannt sein können, wäre die Zahl der Volltextzugriffe ohne diese Zusatzinformation immer gleich Null, obwohl sie in Wahrheit einfach nicht bekannt, d. h. ein Missing Value, ist.
Selbstverständlich kann der Horizont nicht aus einem Access-Log ermittelt werden. Stattdessen ist diese Information als Metadatum eines Dokuments zu verstehen, das einer Sammlung durch den Service Provider zugänglich sein muss.

3 Optionale Eigenschaften der Beschreibung eines Nutzungsvorgangs

3.1 User-IP

Proxy-Server erzeugen den Anschein einer hohen und vielfältigen Aktivität ausgehend von einer einzelnen IP-Adresse. Um dies zu korrigieren, sollte versucht werden, Personen hinter dem Proxy voneinander zu unterscheiden.
Zum einen kann dies durch Übergabe einer Session-ID erfolgen. Eine andere Möglichkeit ist die Auswertung des HTTP-Headers um die IP-Adresse zu erfahren, in deren Auftrag der Proxy die Anfrage stellt. Diese IP-Adresse, ab hier User-IP genannt, muss genauso umgewandelt werden, wie im vorherigen Abschnitt für die Request-IP beschrieben. Komplikationen können allerdings auftreten, da der HTTP-Header mit sehr geringem Aufwand manipuliert und verfälscht werden kann.
Es muss betont werden, dass Session Identifier sowohl als Adresszeilenargument als auch als Cookie im Gegensatz zu IP-Adressen den datenschutzrechtlichen Anforderungen ohne Mehraufwand gerecht werden und außerdem für die Proxy-Problematik nicht anfällig sind. Aus Gründen der Effizienz und Messgenauigkeit erscheint die Verwendung von Session Identifier der alleinigen Nutzung von IP-Adressen überlegen.

3.2 Verweisende Ressource (Referer)

Das starke Interesse an zeitnahen und automatisch produzierbaren Ergänzungen zum Impact Factor kann unter anderem durch sogenannte Clickstream-Analysen bedient werden. Ähnlich wie bei Zitationen werden hier Beziehungen zwischen nacheinander besuchten Informationsobjekten konstruiert (Retrieval Coherence Assumption). Es gelten hierbei die bereits erwähnten Einschränkungen für Informationen aus dem HTTP-Header.

4 Veranschaulichung anhand einer Apache-Skizze

Die Tabelle beschränkt sich auf die Nennung der Direktive, welche die nötige Information und eventuell weitere Daten enthält. Technische Operationen, um den relevanten Teilstring zu ermitteln, werden nicht angegeben. Verbreiteten Feldern wird der Vorzug vor exotischen Direktiven gegeben. Bei der Implementierung ist zu entscheiden in welcher Granularität Informationen erhoben werden sollen.

OA-S-Feldname	Direktive im Apache Log	Beispiel	Anmerkung
Dokumentidentifikator	%r	GET **beispiel.reposito-rium.de**/abstract_page.php?**id=0815** HTTP/1.1	1
Dateiformat	%r	GET beispiel.repositorium.de/volltexte/beispiel.**pdf** HTTP/1.1	
Dienstart	%r	GET beispiel.reposito-rium.de/**abstract_page**.php?**id=0815** HTTP/1.1	2
Zugriffszeitpunkt	%t	**[30/Feb/2008:00:02:13 +0100]**	
Request-IP	%h	sha-256(**254.254.254.254**+jgdlhgdfjbdyjfbnydkbn)	3

C-Klasse-Subnetz	%h	sha-256(**254.254.254**+jgdlhgdfjbdyjfbnydkbn)	
Hostname	LookUp von %a bzw. %h	host(**254.254.254.254**) bzw. www.**klientendomäne.de**	4
Session Identifier	%r	GET beispiel.repositorium.de/abstract_page.php?id=0815&**SESSION_ID=12345** HTTP/1.1	
User Agent	%{User-Agent}i	**ELinks/0.9.3 (textmode; Linux 2.6.9-kanotix-8 i686;)**	
HTTP Status Code	%>s	**200**	
Bytes sent	%b	**815**	
Referrer	%{Referer}i	**beispiel.repositorium.de**/abstract_page.php?id=0815	
Referring Entitiy	%{Referer}i	**anderes.repositorium.de/verweis.php?id=777**	
User-IP	%{X-Forwarded-For}i	sha-256(**127.0.0.1**+jgdlhgdfjbdyjfbnydkbn)	

1 Die Dokument-Nummer in diesem Beispiel ist nur im Kontext des Repositoriums eineindeutig. Zu übertragen ist eine kontextfreie Signatur, die sich in der Regel aus der Dokumentnummer generieren lässt. URNs sind ungeeignet, da ein Informationsobjekt unter Umständen mehrere URNs besitzen kann.

2 Welche HTTP-Anfragen, welche Art von Dienst (z. B. Abstractpage, Volltext, Kapitel eines Volltexts) implizieren, kann nur lokal erkannt werden und muss dann in ein kontrolliertes Vokabular übersetzt werden.

3 Falls HostnameLookup aktiviert ist, muss die IP mit der Direktive %a statt %h oder wahlweise einem Reverse-Lookup ermittelt werden.

4 Den Hostnamen nicht im Moment der Logerzeugung, sondern erst im Zuge der Datenaufbereitung zu bestimmen, entlastet den Webserver und wird deshalb empfohlen.

Zusammengefasst kann das Log-Format wie folgt im Apache Webserver konfiguriert werden:

Variante 1 (nachträglicher Hostname-LookUp):

```
%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{User-Agent}
i\" \"%{X-FORWARDED-FOR}i\|%{X_FORWARDED_FOR}i|%{CLIENT_
IP}i\"
```

oder

Variante 2 (LookUp per Apache Direktive zum Zeitpunkt des Requests inklusive Bestimmung der IP um den ReverseLookUp einzusparen):

```
%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{User-Agent}
i\" \"%{X-FORWARDED-FOR}i\|%{X_FORWARDED_FOR}i|%{CLIENT_
IP}i\" %a
```

Informationen, die aus HTTP-Header-Feldern ausgelesen werden, werden zwischen Anführungszeichen eingeschlossen. Dies hilft Angriffe auf die Datenkonsistenz zu verhindern, da Header-Felder durch Eingriffe von außen beliebig befüllt werden können.
Abzüglich der letzten Direktivengruppe in Anführungszeichen handelt es sich bei Variante 1 um das Combined Log-Format, welches weit verbreitet ist und die als notwendigen eingestuften Informationen vollständig erfasst.

5 Dienstart

kontrollierter Begriff	Kürzel	Definition	Beispiel
Repository	R	Das HTTP-Objekt dient dem Zugang zu den Dokumenten	Suchmaske
Abstract	A	Das HTTP-Objekt enthält den Abstract, welcher die Ressource in ihrer Gänze beschreibt, und optional weitere Metadaten.	Abstractpage

Fulltext	F	Das HTTP-Objekt umfasst den vollen Inhalt des Dokuments.	PDF-Datei
Embedded	E	Das Objekt ist Teil der Präsentation eines HTTP-Objekts aber auf der Dateiebene unabhängig	Grafikdatei, die von einem HTML-Text aus referenziert wird
Chapter	C	Das HTTP-Objekt ist nur ein Teil des Volltext-Inhalts, ist jedoch nicht durch Metadaten von den anderen Teilobjekten unterscheidbar.	einzelnes Kapitel eines Buches in einer eigenen Datei
Appendix	X	Das HTTP-Objekt ergänzt die eigentliche Ressource, ist jedoch inhaltlich/formal unabhängig	Video-Datei, die den Ablauf eines Experiments veranschaulicht

6 Beispielimplementation des Zählhorizonts

Der Zählhorizont könnte dokumentweise erfasst werden und bestünde dann aus einem Tupel der Dienstarten, die vom Logging-Mechanismus des Repositoriums erfasst werden. Unter gewissen Umständen kann das Repositorium nur das Absetzen der Anfrage, jedoch nicht den Erfolg, das heißt den HTTP-Statuscode, erfassen. In diesem Fall ist die entsprechende Dienstart durch ein Apostroph zu kennzeichnen.

Beispiele:

RAFECX	vollständiges Logging eines Anbieters mit Gesamt- und Kapitelfassung
RA---	virtuelle Bibliothek mit ausgelagerten Volltexten
RAF--	Repositorium, das keine Objekte bestehend aus mehreren Dateien zulässt
RAF'C'X'	virtuelle Bibliothek, welche die Anfragen nach Datei-Ressourcen über eine interne Redirect-Page leitet
--F--	Reines Textlager

7 Berechnungsstandards

Liste der relevanten Projekte und Initiativen, die bereits einen Standard zur Generierung von Volltext-Nutzungsstatistiken aus Zugriffsdaten formuliert haben oder im Moment formulieren:

URL des Statistikanbieters	**Zählkriterium**	**Multiklick-zeitspanne**
COUNTER http://www.projectcounter.org/code_practice.html	HTTP Status Code ist 200 oder 304	für HTML 10s; für PDF 30s
LogEc http://logec.repec.org/about.htm	HTTP Status Code ist 200, 206, 301, 302 oder 304.	Ein kalendarischer Monat
IFABC http://www.ifabc.org/standards.htm	HTML: Zählpixel; Andere: Traffic-Volumen 95% der Dateigröße	Besuchszeit: Aufenthalt auf dem Repositorium plus 30min
Interoperable Repository Statistics http://trac.eprints.org/projects/irstats/wiki	HTTP Status Code ist 200 für Abstract- oder Volltext-Seite	24 Stunden
AWStats http://awstats.sourceforge.net/#DOC	Default: HTTP Status Code 200 oder 304	Default: 1 Stunde

- Ein IFABC-Aufenthalt (Visit) ist durchgehend und aktiv, solange nicht mehr als 30 Minuten zwischen zwei Seitenaufrufen liegen.

Besucher-identifizierung	Crawlerdefinition	Crawleridentifikation	Crawlerzählung
mindestens IP, besser Session	Robots, Prefetches, Caching, Federated Searches (n.a.)	Black-List, Client HTTP Header	in separaten Berichtsvariablen
IP	Robots, automatisierte Downloads (wget)	Zugriff auf Robots.txt; Zahl der Zugriffe > 10,000 items/month; C-Class Access > 10% des Bestands; Hostname; IP	eigene Spalte in Statistiken
IP+User-Agent; Cookie-Session, Login-Session	Robots; automatische Downloads (optional)	proprietäre Blacklist	werden nicht berichtet
IP	Robots; automatische Downloads	AWStats Blacklist	Crawlerzugriffe werden nicht berichtet.
IP	Robots	Black-List	eigene Spalte in Statistiken

8 Anforderungsvergleich der verschiedenen Berechnungsstandards

Die Feldnamen dienen nur der Anschaulichkeit und sollen die Kommunikation erleichtern.

OA-S-Feldname	Beschreibung
Dokumentidentifikator	eindeutiger Bezeichner des Volltextes
Dateiformat	Dateiformat der Serverantwort (z. B. HTML oder PDF)
Dienstart	Art der Antwort (z. B. Volltext oder Abstractpage)
Zugriffszeitpunkt	sekundengenaue Zeit der Serverantwort
Request-IP	IP-Adresse der Anfrage (bzw. ihres Proxy)
C-Klasse-Subnetz	C-Klasse-Anteil der Request-IP
Hostname	Zur IP gehöriger Name der Hosting Domain
Session Identifier	vom Server vergebene Session-ID
User Agent	User-Agent-String des Users (Clients)
HTTP Status Code	Server-Status-Code des HTTP-Requests
Bytes sent	Größe der Serverantwort
Referrer	Bezeichnung des Servers, der das OpenURL ContextObject erzeugt
Referring Entitiy	eineindeutiger Bezeichner (z. B. Dateipfad und -name) des Objekts, das die anfragende Entität auf das aktuelle Dokument verwies
User-IP	IP-Adresse des Systems, in deren Auftrag ein Proxy einen Request ausführt

Anmerkungen:

- Falls auf lokaler Ebene die IPs nur als Hash gespeichert werden, ist für LogEc nicht nur ein Hash der IP, sondern auch ein Hash des Netzwerkteils (C-Klasse-Subnetz) der IP nötig.
- Der Hostname kann sowohl mit einem nachträglichen LookUp – vor Umwandlung der IP in den Hash-Wert – als auch mithilfe von Apache Direktiven ermittelt werden.

COUNTER	LogEc	IFABC	Clickstream
nötig	nötig	nötig	nötig
nötig	-	nötig	-
-	nötig	optional	-
nötig	nötig	nötig	optional
nötig	nötig	falls Session-ID nicht vorliegt: nötig	falls Session-ID nicht vorliegt: nötig
-	nötig	-	-
-	nötig	-	-
optional	-	falls IP und UA nicht vorliegen: nötig	falls IP nicht vorliegt: nötig
nötig	nötig	falls Session-ID nicht vorliegt: nötig	optional
nötig	nötig	nötig	nötig
-	-	falls Dateiformat ungleich HTML: nötig	-
-	-	-	nötig
-	-	-	nötig
-	-	-	falls Session-ID nicht vorliegt: empfohlen

9 Ursachen von Messabweichungen

Einerseits können Aktivitäten auf einem Repositorium durch Automatisierung einfach beobachtet werden. In den meisten Fällen ist schon alleine aus Sicherheitsgründen eine solche Überwachung vorgeschrieben. Hiervon ausgehend ist es nur ein kleiner Schritt hin zu einer Informationssammlung, die den eigenen Bedürfnissen gerecht wird.
Andererseits erzeugen sowohl gewohnheitsbedingte Handlungen als auch fremde automatisierte Prozesse Verzerrungen, sprich Messfehler, die berücksichtigt werden müssen, wenn die Zugriffsstatistiken valide sein sollen.
Man kann diese Problemstellung im Licht der Signalentdeckungstheorie betrachten:

	Zugriff gezählt	**Zugriff ignoriert**
Echte Nutzung	Korrekte Erhöhung	unangebrachte Beibehaltung
Scheinnutzung	unangebrachte Erhöhung	Korrekte Beibehaltung

Es gibt eine Reihe von technischen, organisationsbedingten oder menschlichen Fehlerquellen, die zu Verzerrungen bei der Datenerhebung führen können:

1. Repositorien-unabhängige Verbreitung von Informationsobjekten
   - Caching (sowohl transparente Cache-Server als auch Proxy-Server)
   - Weitergabe von Ausdrucken
   - Weitergabe von Dateien (z. B. als E-Mail-Anhang oder über ein Netzlaufwerk)
2. Mehrere Client-PCs bzw. mehrere Benutzende teilen sich eine IP
   - CIP-Räume
   - Thin Clients
   - intransparente Proxy-Server
   - DHCP in institutionellen Funknetzen
3. Benutzende teilen sich einen Zugang
   - Recherche-Terminals in Bibliotheken
   - Arbeitsgruppen-Accounts
   - Privat-PCs

Der Einsatz eines Session Identifier würde die Probleme, welche durch gemeinsame IP-Nutzung entstehen, lösen. Allerdings muss betont werden, dass geteilte IPs kein generelles Problem sind, sondern schlicht der Operationalisierung der Besucher-Identität durch die IP-Adresse widersprechen. Das ursprüngliche Argument, dass die IP-Adresse weniger Varianz aufweist als ein Session Identifier, ist schwach,

wenn man bedenkt, dass viele Nutzer aus dem akademischen Umfeld zwischen mehreren PCs und tragbaren Rechnern hin und her wechseln.
Das gemeinsame Nutzen von Accounts bzw. die Unterbindung dieses Verhaltens ist ein Aspekt des Identitätsmanagments einer Institution, vergleichbar mit zwei Personen, die sich einen Bibliotheksausweis teilen. Diese Praxis wirkt sich grundlegend auf die praktische Sicherheit, den Datenschutz und Gebührenbuchungen aus und ist auf der Seite der Repositorien nicht lösbar.
Verzerrungen, die durch Caching entstehen, können in gewissem Umfang durch die Einhaltung der HTML-RFCs reduziert werden. Die Vorratshaltung durch einen Cache kann verboten beziehungsweise der Client zur Revalidierung aufgefordert werden, indem das Repositorium entsprechende Anweisungen im Header überträgt (siehe oben, 2.2 Caching). Unglücklicherweise respektieren nicht alle Agierenden diese Anweisungen, unter anderem aus Profitinteressen. Der Einsatz von Techniken wie Zählpixeln (z. B. der IVW / IFABC), würde diese Zahl reduzieren. Allerdings weisen solche Methoden auch Mängel auf und sind in der Regel nur in HTML-Objekten einsetzbar.
Das informelle Weiterverbreiten von Dateien und Ausdrucken sollte mit der Zeit von alleine zurückgehen, je besser es den Anbietern gelingt, Zugangshürden und Medienbrüche zu reduzieren. Wenn Objektidentifikatoren verlässlich in eine Ressource aufgelöst werden können und der Dateizugriff unkompliziert ist, wird die Wissenschaftsgemeinschaft von alleine die informelle – und unzuverlässige – Weitergabe wegen Verschiebungen im Kosten-Nutzen-Verhältnis unterlassen.
Unangebrachte Erhöhungen der Nutzungszahl sind quantitativ betrachtet ein größeres Problem als nicht gezählte echte Zugriffe.
Der größte Teil der Scheinnutzungen kann wie folgt klassifiziert werden:

1. Harvester (Synonyme: Roboter, Crawler, Spider, Wanderer)
2. Prefetching
   - Prefetching durch den Browser
   - Prefetching durch einen dedizierten Server
3. Automatische Downloads
4. Mehrfach-Klicks in der Browserumgebung
5. wiederholte Anfragen wegen Unachtsamkeit, Arbeitsunterbrechungen etc.
   - Wechsel der Applikation (z. B. von einem Browser-Add-on zu einem eigenständigen Programm)
   - voreiliges Schließen des Dokuments und anschließende Wiederholung der Anfrage

Den auf menschlichem Verhalten basierenden Problemen der Klassen 3.–5. versucht man mit so genannten Multiklickintervallen zu begegnen: Mehrere Anfragen der gleichen Person innerhalb dieses Zeitintervalls werden als ein Nutzungsereignis gewertet. Verschiedene Annahmen über das zu messende menschliche Verhalten bedingen verschiedene Zeitintervalle:

- Als Reaktion auf die Angewohnheit, immer doppelt zu klicken, ist ein Intervall von zehn Sekunden völlig ausreichend.
- Um die tatsächliche Rezeption oder intellektuelle Durchdringung eines Artikels zu operationalisieren, kann es sinnvoll sein, die HTTP-Requests mehrerer Tage oder sogar Wochen zusammenzufassen.

Die zweite Gruppe von Verzerrungen setzt sich aus den Aktivitäten von Automaten zusammen.

Eine Indexierung durch Suchmaschinen ist selbstverständlich im Interesse der Repositorien, da hierdurch die Sichtbarkeit der angebotenen Inhalte erhöht wird. Sogar vom Standpunkt des Administrators aus kann es interessant sein, die Wege der Harvester nachzuverfolgen, da hierdurch Schwachstellen und Konstruktionsfehler im eigenen Angebot aufgedeckt werden können Die Bewegungsmuster von Suchmaschinen dienen in der Regel dazu, ein Angebot vollständig zu indexieren, beziehungsweise einen solchen Gesamtindex zu aktualisieren. Es findet keine Nutzung im Sinne einer intellektuellen Auseinandersetzung statt.. Aus diesem Grund sind diese automatisierten Zugriffe grundlegend von menschlichen Zugriffen abzugrenzen.

Nicht alle Roboter geben sich deutlich als solche zu erkennen. Zudem unterscheiden sich selbst bei solchen, die sich zu erkennen geben die Methoden der Identifizierung teils erheblich, was bei der Entwicklung von entsprechenden Filtern berücksichtigt werden muss. Es gibt jedoch eine Reihe von Verfahren, mithilfe derer Crawler durch verschiedene Eigenschaftskombinationen (Tan & Kumar, 2002) oder Bayessche Netze (Stassopoulou & Dikaiakos, 2007) erkannt werden können.

Zukünftige Implementationen von Prefetching müssen beobachtet werden, um Erkennungskriterien formulieren zu können. Ähnlich wie bei automatischen Downloads muss außerdem entschieden werden, wie die Unsicherheit berücksichtigt werden soll, ob ein automatisch angefordertes Dokument tatsächlich rezipiert wurde oder nicht.

Automatische Abrufe werden von den verschiedenen Standards unterschiedlich gewertet:

- Für LogEc ist ein automatischer Download immer eine Scheinnutzung, auch wenn das heruntergeladene Dokument am nächsten Tag gelesen wird.
- COUNTER sieht die automatische Beschaffung als gleichwertig zum manuellen Dokumentaufruf.
- IFABC schließlich überlässt es dem Anbieter automatische Anfragen als Nutzung zu werten oder auch nicht.

Die Verteilungsstruktur der Dokumente über mehrere Institutionen erschwert die Aggregation von Kennwerten sowie die Erzeugung von Meta-Berichten:
Ein Objekt kann über mehrere Server zugänglich sein, zum Beispiel über ein institutionelles und ein fachliches Repositorium. In einem solchen Fall können nicht alle Zugriffsereignisse von einem Anbieter protokolliert werden, da einige Dienstarten gar nicht vor Ort angefallen sind, sondern weitergeleitet wurden.
Ohne zusätzliche Informationen ist es jedoch nicht möglich zu entscheiden, ob zwei Dokumente tatsächlich unterschiedlich sind, oder ob zwei Abstract-Seiten auf dieselbe Datei auf demselben Speichermedium verweisen. Bis heute gibt es keinen etablierten Standard, der die Relationen zwischen verschiedenen Versionen, wie zum Beispiel Preprint und Postprint einer Publikation abbilden kann. Da Texte sich auf der Dateiebene stark unterscheiden können, auch wenn sie inhaltlich identisch sind, bieten Streuwerte der eigentlichen Dateien keinen Ausweg.
Durch die unterschiedliche Schwerpunktsetzung verschiedener Organisationen kann es sogar geschehen, dass ein Dokument durch voneinander abweichende Metadatensätze und sogar unterschiedliche konstante Identifikatoren, sprich URN oder DOI, beschrieben wird. Dies erschwert die Aggregation der Nutzungsereignisse über Anbietergrenzen hinweg zusätzlich. Eine autorenzentrierte Statistik zum Beispiel ist somit nicht möglich, wenn einige Arbeiten dieser Person in einem Repositorium liegen, dessen Inhalte auf Zeitschriftenebene und nicht auf Artikelebene organisiert sind.
Diensteanbieter wie Elsevier (Scopus) oder Thomson Reuters (ISI Web of Science) erstellen Zitationsgraphen und -statistiken für individuelle Autoren. Auch sie sind allerdings auf die ihnen zugängliche Dokumentenmenge, sprich ihr eigenes Angebot, beschränkt.
Man muss allerdings betonen, dass die hier geschilderten Verzerrungen weder schwerwiegender noch geringfügiger sind als die Messungenauigkeiten bei etablierten Methoden der Scientometrie.
Solche methodischen Probleme können, wie oben bereits geschildert, ausgeglichen werden, indem verschiedene Methoden zusammen genutzt werden. So kann mithilfe von Kreuzvalidierung und Inter-Rater-Reliabilität ein genaueres Bild der

tatsächlichen Abläufe erstellt werden. Die Einführung von in Echtzeit ermittelten Nutzungsstatistiken ist ein möglicher Ansatz der Wirkungsbeschreibung wissenschaftlicher Publikationen, der gegenüber traditionellen Ansätzen wie dem IF eine qualitativ andere Perspektive erschließt (Bollen 2009).

10 Referenzen

Projekt Open-Access-Statistik. URL: http://www.dini.de/projekte/oa-statistik/ [Stand: 23.07.2009]

RFC 2616: Hypertext Transfer Protocoll -- HTTP/1.1. Memo. URL: http://www.w3.org/Protocols/rfc2616/rfc2616.html [Stand: 20.10.2008]

mod_log_config in the Apache documentation. URL: http://httpd.apache.org/docs/2.2/mod/mod_log_config.html [Stand: 28.11.2008]

A Standard for Robot Exclusion. URL: http://www.robotstxt.org/orig.html [Stand 28.11.2008]

Bollen, J. (2009) Studying scientific activity from large-scale usage data. CERN workshop on Innovations in Scholarly Communication (OAI6). URL: http://indico.cern.ch/materialDisplay.py?contribId=22&sessionId=8&materialId=slides&confId=48321 [Stand 23.07.2009]

Brody, T., Carr, L., Harnad, S. & Swan, A. (2007) Time to Convert to Metrics. Research Fortnight. pp. 17-18. URL: http://eprints.ecs.soton.ac.uk/14329/ [Stand: 24.03.2009]

Counting Online Usage of Networked Information Resources (COUNTER). URL: http://www.projectcounter.org [Stand 23.07.2009]

Garfield, E. (1955) Citation indexes to science: a new dimension in documentation through association of ideas. Science. Volume 122. pp. 108-111. URL: http://garfield.library.upenn.edu/essays/v6p468y1983.pdf [Stand 23.07.2009]

International Federation of Audit Bureaus of Circulations (IFABC) URL: http://www.ifabc.org/ [Stand 23.07.2009]

LogEc: Access Statistics for Participating RePEc Services. URL: http://logec.repec.org/ [Stand: 24.03.2009]

Marcum, J. (1960) A statistical theory of target detection by pulsed radar. IEEE Trans. Info. Thry. Volume 6. Issue 2. pp. 59-267

Merk, C., Scholze, F. & Windisch, N. (2009) Item-level usage statistics: A review of current practices and recommendations for normalization and exchange. Library Hi Tech. Volume 27. Issue 1. pp. 151-162 DOI: 10.1108/07378830910942991

(http://www.emeraldinsight.com/10.1108/07378830910942991) [Stand: 18.05.2009]
Publisher and Institutional Repository Usage Statistics (PIRUS). URL: http://www.jisc.ac.uk/whatwedo/programmes/pals3/pirus.aspx [Stand: 18.05.2009]
Stassopoulou, A. & Dikaiakos, M. D. (2007) A probabilistic reasoning approach for discovering web crawler sessions. Lecture Notes in: Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Volume 4505 LNCS, 2007, S. 265-272. URL: http://www.scopus.com/scopus/inward/record.url?eid=2-s2.0-38049025300&partnerID=40&rel=R8.2.0 [Stand: 30.10.2008]
Tan, P.-N. & Kumar, V. (2002) Discovery of web robot sessions based on their navigational patterns". In: Data Mining and Knowledge Discovery, 2002,6 (1), S. 9-35. URL: http://www.scopus.com/scopus/inward/record.url?eid=2-s2.0-0036109905&partnerID=40&rel=R8.2.0 [Stand: 30.10.2008]

11 Anmerkung

Diese Empfehlung zur Erstellung und Standardisierung von Nutzungsmessungen elektronischer Publikationen entstand im Rahmen des von der DFG geförderten Projekts Open Access-Statistik (OA-S).

Autoren

Ulrich Herb, Saarländische Universitäts- und Landesbibliothek
(u.herb@sulb.uni-saarland.de)
Björn Mittelsdorf, Saarländische Universitäts- und Landesbibliothek
(b.mittelsdorf@sulb.uni-saarland.de)
Daniel Metje, Niedersächsische Staats- und Universitätsbibliothek Göttingen
(metje@sub.uni-goettingen.de)
Matthias Schulze, Universitätsbibliothek Stuttgart
(matthias.schulze@ub.uni-stuttgart.de)
Wolfram Horstmann, Universitätsbibliothek Bielefeld
(wolfram.horstmann@uni-bielefeld.de)
Susanne Dobratz, Humboldt-Universität zu Berlin
(dobratz@cms.hu-berlin.de)
Frank Scholze, Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg
(frank.scholze@mwk.bwl.de)

Aufnahmeantrag für die Mitgliedschaft in DINI e.V.

(auch online unter http://www.dini.de/mitgliedschaft/mitgliedsantrag/)

Angaben zum Antragsteller:
Name:
Vorname:
Sind Sie Bevollmächtigte/r der antragstellenden Institution? ☐ Ja ☐ Nein
Institution:
URL der Institution:

Die antragstellende Institution ist Mitglied in:
☐ AMH ☐ dbv ☐ ZKI ☐ Fachgesellschaften und Sonstige

Anzahl der Beschäftigtenvollzeitäquivalenz (BVZÄ):

Weitere Angaben (entweder zu Ihrer Person oder der Institution):
Anschrift (Straße, Nummer):
PLZ, Ort:
Telefon:
Fax:
E-Mail-Adresse:

Wer soll Mitglied werden
☐ Hochschule ☐ Institution ☐ Fachgesellschaft ☐ Sie selbst

Welche Art der Mitgliedschaft wünschen Sie?
Zur Definition der Mitgliedschaft siehe Satzung § 3
☐ Ordentliches Mitglied ☐ Assoziiertes Mitglied

Bemerkungen

Ort, Datum

Unterschrift

[Version 1.0, August 2009]

Impressum
DINI – Deutsche Initiative für Netzwerkinformation e. V.
Geschäftsstelle
c/o Niedersächsische Staats- und Universitätsbibliothek Göttingen
Papendiek 14
37073 Göttingen
Tel.: 0551-39-3857
Fax: 0551-39-3856
E-Mail: gs@dini.de
www.dini.de