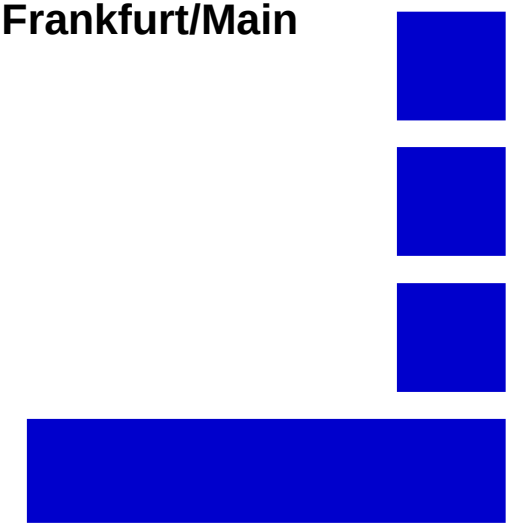
Tutorial Zugriffsstatistik

DINI-Workshop: Open Access und das DINI-Zertifikat 2007

7. und 8. Februar an der Deutschen Nationalbibliothek Frankfurt/Main



Ulrich Herb
Saarländische Universitäts- und Landesbibliothek (SULB)

Struktur

1. Status Quo: Umgang mit Webserver-Logs und Nutzungsstatistiken
2. Anforderungen des DINI Zertifikats 2007
3. Datenschutz
4. Aufbau von Webserver-Logs
5. Referenzen
6. Fragestellungen rund um Webserver-Logs
7. Anwendungen
8. Perspektiven/ offene Fragen

Ulrich Herb
Saarländische Universitäts- und Landesbibliothek (SULB)

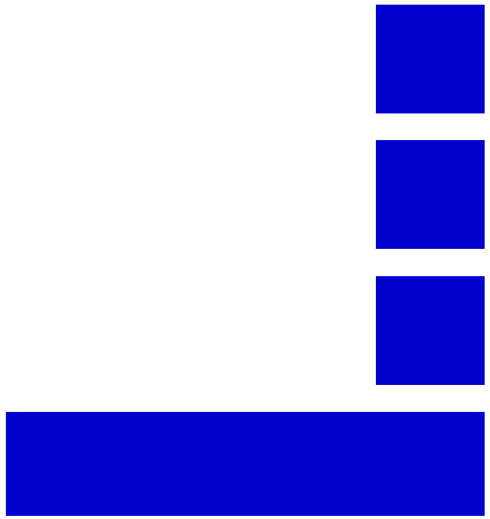
1. Status Quo: Umgang mit Webserver-Logs und Nutzungsstatistiken

A Fragen zum Repository

B Behandlung der Webserver-Logs

C Aufbereitung der Webserver-Logs für die Benutzersicht

D Infrastruktur und Darstellung



Ulrich Herb
Saarländische Universitäts- und Landesbibliothek (SULB)

2. Anforderungen des DINI Zertifikats 2007

2.7 Zugriffsstatistik

Mindeststandard

- Jeder Dokumentenserver muss im Rahmen der rechtlichen Bestimmungen eine eigene konsistente Zugriffsstatistik führen (Webserver-Log).
- Webserver-Logs sind zur längerfristigen Speicherung zu anonymisieren[21].
- Der Statistik muss eine Erklärung bzw. Dokumentation beigefügt sein, die beschreibt, nach welchen Kriterien die Statistik erstellt bzw. aufbereitet wurde. Es muss darauf hingewiesen werden, dass die Zugriffszahlen i.d.R. nicht zu einem Vergleich zwischen unterschiedlichen Dokumenten-Servern bzw. Diensten eingesetzt werden können, sondern dass die Zugriffszahlen allein eine Vergleichbarkeit der Dokumente auf einem Server erlauben.

[21] Gemäß den datenschutzrechtlichen Bestimmungen § 4 Abs. 4 Nr. 2 TDDSG, § 18 Abs. 4 Nr. 2 MDStV. Beispiel unter http://www.zendas.de/technik/sicherheit/apache/index.html

2. Anforderungen des DINI Zertifikats 2007

Empfehlungen

- Zugriffe auf Dokumente durch automatisierte Agenten, Roboter u. ä. werden herausgefiltert[22].
- Webserver-Logs werden einheitlich nach dem Counter Code of Practice aufbereitet[23].
- Einem Dokument ist seine Zugriffsstatistik als dynamisches Metadatum zugeordnet und für den Endnutzer sichtbar.[24]

22 Z. B. http://www.robotstxt.org/wc/active/all.txt

23 Orientierung v. a. an http://www.projectcounter.org/cop_books_ref.html#rbr_5 und Anhang D http://www.projectcounter.org/cop_books_appendix_d.html

24 Z. B. in PsyDok (http://psydok.sulb.uni-saarland.de/) und MONARCH (http://archiv.tu-chemnitz.de/) jeweils von den Indexseiten (Document Splash Pages) aus realisiert.

3. Datenschutz

§ 4 Abs. 4 Nr. 2 Teledienstedatenschutzgesetz (TDDSG):

„Der Diensteanbieter hat durch technische und organisatorische Vorkehrungen sicherzustellen, dass

1. der Nutzer seine Verbindung mit dem Diensteanbieter jederzeit abbrechen kann,
2. **die anfallenden personenbezogenen Daten über den Ablauf des Zugriffs oder der sonstigen Nutzung unmittelbar nach deren Beendigung gelöscht oder gesperrt werden können**,
3. der Nutzer Teledienste gegen Kenntnisnahme Dritter geschützt in Anspruch nehmen kann,
4. die personenbezogenen Daten über die Inanspruchnahme verschiedener Teledienste durch einen Nutzer getrennt verarbeitet werden können,
5. Daten nach § 6 Abs. 2 nur für Abrechnungszwecke und
6. Nutzerprofile nach § 6 Abs. 3 nicht mit Daten über den Träger des Pseudonyms zusammengeführt werden können."

http://bundesrecht.juris.de/tddsg/__4.html

analog: § 18 Abs. 4 Nr. 2 Mediendienste-Staatsvertrag (MDStV):
http://www.datenschutz-berlin.de/recht/de/stv/mdstv.htm#nr18

3. Datenschutz

Begriffe: § 3 Abs. 4 Nr. 4 und Nr. 5 Bundesdatenschutzgesetzes (BDSG):

(4) Verarbeiten ist das Speichern, Verändern, Übermitteln, Sperren und Löschen personenbezogener Daten. Im Einzelnen ist, ungeachtet der dabei angewendeten Verfahren:

1. Speichern das Erfassen, Aufnehmen oder Aufbewahren personenbezogener Daten auf einem Datenträger zum Zwecke ihrer weiteren Verarbeitung oder Nutzung,

2. Verändern das inhaltliche Umgestalten gespeicherter personenbezogener Daten,

3. Übermitteln das Bekanntgeben gespeicherter oder durch Datenverarbeitung gewonnener personenbezogener Daten an einen Dritten in der Weise, dass

   a) die Daten an den Dritten weitergegeben werden oder

   b) der Dritte zur Einsicht oder zum Abruf bereitgehaltene Daten einsieht oder abruft,

4. Sperren das Kennzeichnen gespeicherter personenbezogener Daten, um ihre weitere Verarbeitung oder Nutzung einzuschränken,

5. Löschen das Unkenntlichmachen gespeicherter personenbezogener Daten.

http://www.gesetze-im-internet.de/bdsg_1990/__3.html

3. Datenschutz

§ 3a Bundesdatenschutzgesetzes (BDSG): Datenvermeidung und Datensparsamkeit

„Gestaltung und Auswahl von Datenverarbeitungssystemen haben sich an dem Ziel auszurichten, keine oder so wenig personenbezogene Daten wie möglich zu erheben, zu verarbeiten oder zu nutzen.

Insbesondere ist von den Möglichkeiten der Anonymisierung und Pseudonymisierung Gebrauch zu machen, soweit dies möglich ist und der Aufwand in einem angemessenen Verhältnis zu dem angestrebten Schutzzweck steht.“

-> Betreiber von Repositories sind verpflichtet, die anfallenden Nutzungsdaten zu anonymisieren.

Anonymität ist gegeben, wenn ein Personenbezug auch mit Hilfe von zusätzlichem Wissen nicht mehr hergestellt werden kann.

Es existieren Tools zur datenschutzkonformen Erstellung von Webserver-Logs:

http://www.zendas.de/technik/sicherheit/apache/index.html

Ulrich Herb
Saarländische Universitäts- und Landesbibliothek (SULB)

4. Aufbau von Webserver-Logfiles

Datenschutz: Personenbezogene und personenbeziehbare Daten

Aufbau von Webserverlogs am Beispiel von APACHEs Combined Log Format:
134.96.121.172 - - [05/Feb/2007:11:03:33 +0100] "GET /volltexte/2004/408/html/gesch.htm HTTP/1.0" 200 238810
"http://scholar.google.com/scholar?q=nachteile+und+vorteile+der+transaktionsanalyse&hl=de&lr=&start=30&sa=N" "Mozilla/4.0 (compatible; MSIE 5.5; Windows 98)"

1. Adresse des aufrufenden Hosts
2. Authentifizierungsfelder
3. Datum und Uhrzeit des Zugriffs, sowie Zeitzonendifferenz zum UTC Normal
4. Zugriffsmethode und Inhalt des HTML-Aufrufs
5. Statuscode des Webservers zum Aufruf
6. Anzahl der vom Webserver übertragenen Bytes
7. Referenzadresse der aufrufenden Seite
8. Angaben über den eingesetzten Browser und das Betriebssystem des Clients

Personenbezogene bzw. -beziehbare Daten: Felder 1, 2, 4 und 7

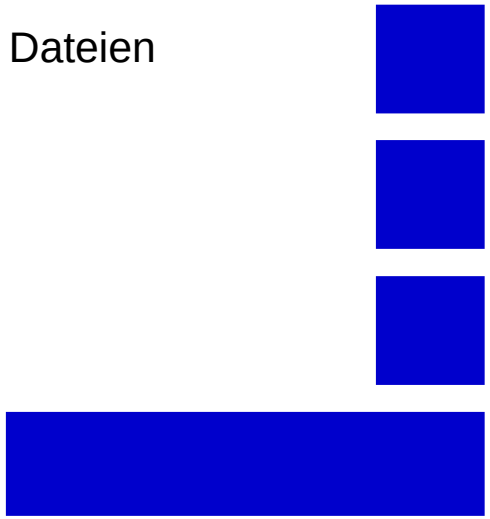
s. auch http://www.rheinland-pfalz.datenschutz.de/materialien/hinweise/info_webserverlogfiles.pdf

4. Aufbau von Webserver-Logfiles

Verzerrungen

z.B.

- Doppelclicks
- Spider/Robots
- Bei HTML-Dokumenten: eingebundene Elemente wie Grafiken erzeugen Einträge in den Logfiles
- Ggf. mehrere Einträge in Webserver-Logs beim Download größerer Dateien
- (Systemeigene) Überwachungstools



5. Referenz: COUNTER

COUNTER code of Practice for Journals and Databases/ COUNTER Code of Practice for Books and Reference Works

COUNTER berücksichtigt nur die Webserver Status Codes 200 (OK) und 304 (not modified)
http://bignosebird.com/apache/a5.shtml

Timespan bei PDFs: innerhalb 30 Sek. wird ein neues Laden als Doppelklick gewertet
Timespan bei HTML: innerhalb 10 Sek. wird ein neues Laden als Doppelklick gewertet

The Code of Practice Journals and Databases: Release 2
(http://www.projectcounter.org/r2/COUNTER_COP_Release_2.pdf)
Release 1 of the COUNTER Code of Practice for Books and Reference Works
(http://www.projectcounter.org/cop/books/cop_books_ref.pdf)

Ist COUNTER ein brauchbarer Standard?

Ulrich Herb
Saarländische Universitäts- und Landesbibliothek (SULB)

6. Fragestellungen rund um Webserver-Logs

Wozu? Warum sollen Nutzungsstatistiken gesammelt und standardisiert werden?

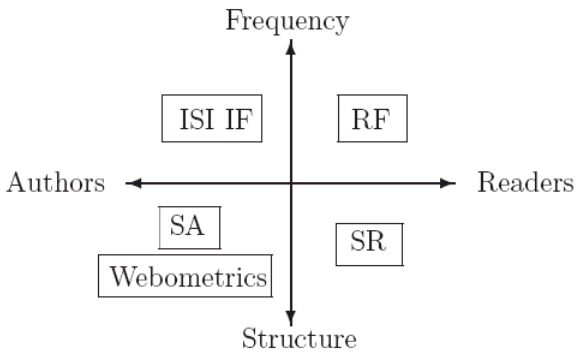
- Evaluation: Research Assessment, Research Management, Repository Marketing

- Zugriffsstatistiken als ergänzendes Impact-Maß

Brody, T., Harnad, S. and Carr, L. (2006): Earlier Web Usage Statistics as Predictors of Later Citation Impact. Journal of the American Association for Information Science and Technology (JASIST) 57(8) pp. 1060-1072.

- Zugriffsstatistiken als Basis alternativer Impact-Faktoren

Bollen, J., Van De Sompel, H., Smith, J.A., Luce, R. (2005): Toward alternative metrics of journal impact: A Comparison of download and citation data. Information Processing and Management 41(6) pp. 1419-1440.



Ulrich Herb
Saarländische Universitäts- und Landesbibliothek (SULB)

6. Fragestellungen rund um Webserver-Logs

Was? Welche Informationen sollen gesammelt werden?

- Was soll unter Nutzung verstanden werden?

- Was soll gemessen werden?

- Welche Items sind interessant und sollen bei den Zugriffstatistiken berücksichtigt werden?

- Für wen werden die Informationen gesammelt?

- Welche Statistiken können erstellt werden?
  Gesamtstatistiken vs. dokumentbezogene Statistiken

- Welche Zusatzinformationen – neben der reinen Downloadhäufigkeit – können von Interesse sein?

- Wie können Verzerrungen beseitigt werden?

Ulrich Herb
Saarländische Universitäts- und Landesbibliothek (SULB)

7. Anwendungen

Welche Anwendungen/ Services sind - aufbauend auf Zugriffstatistiken – möglich oder wünschenswert?

- Recommender
- Collection Management
- Evaluierung
- Research Assessment
- Austausch mit Verlagsstatistiken
- VG Wort Vergütung
- Ergänzung anerkannter bibliometrischer Verfahren
- Etablierung neuer Impact-Maße

Ulrich Herb
Saarländische Universitäts- und Landesbibliothek (SULB)

8. Perspektiven/ offene Fragen

Wie groß ist das Vertrauen in die Daten und Statistiken?
- Wie kann das Vertrauen gestärkt werden?
- Wird es möglich sein, neue Impact-Maße zu etablieren?
- Wie können Missbrauch und Manipulation der Daten verhindert werden? (Transparenz)
- Standards und Techniken müssen offen sein
- Wird es Audits geben müssen?
- Standards und Durchführung der Audits müssen transparent sein
-Technische Umstellungen (Updates, Server-Arbeiten) dürfen nicht zu Ausfällen/ Verzerrungen führen

Generieren eines Raw-Log-Formats, das in beliebige andere Formate (z.B. COUNTER. AWStats,) umgewandelt werden kann?
- das Format sollte langzeitarchivierungsfähig sein
- Anonymisierung erlauben / shibboleth compliance
- Identifikation von Nutzern/Sessions erlauben

Welche Items/Einheiten werden gezählt?
- Granularität (journals, articles, Bücher etc.)
- Wie können die Daten auf dem übergeordneten Level aggregiert werden (Journal, Buch)?

Ulrich Herb
Saarländische Universitäts- und Landesbibliothek (SULB)

8. Perspektiven/ offene Fragen

Policy
- Muss/ soll es Policies über das Sammeln und die Verwendung von Daten geben?
- Ist es zulässig, Zugriffsdaten anzuzeigen?
- Ist Datenaustausch zulässig?
- Ist die Aggregation von Daten auf einem übergeordneten Level möglich/erlaubt?
- Unter welchen Bedingungen?
- Welche Statistiken sollen zugänglich sein?

Wenn es eine Policy geben sollte, müssen auch folgende Punkte geklärt sein:
- Kommerzielle Weiterverwertung der Daten (OA!)
- Die Übereinstimmung mit Rechtsvorschriften
- Welche Standards werden beim Erstellen der Statistiken angewandt?
- Wie wird das Einhalten der Standards überprüft? (Audit?)
- Wie wird die (Weiter-)Verwendung der Daten kontrolliert?
- Wie erfolgt die Sammlung und Aggregation der Daten?

8. Perspektiven/ offene Fragen

Zugriffe durch Robots müssen aus den LogFiles entfernt werden (auf lokalem oder auf einem übergeordneten Level?)
- Vorteile bei Bereinigung auf übergeordnetem Level sind bessere Robots-Erkennung (bei Robots, die sich nicht zu erkennen geben über robots.txt oder Namen) sowie verringerter Administrationsaufwand auf lokaler Ebene.
z.B. Name des user agent, ip in awstats, >XXXX items pro Tag/Woche
Definition eines Guides, der Vorgaben/Richtlinien zur Eliminierung von Robots macht?

Ansätze::
1. user agent = wget et al.?
2. Greift die ip/ der user agent auf robots.txt zu?
3. Führt AWStats die ip in der list of robots?
4. Lädt die IP > 1000 items pro Tag herunter? (# Zeitrahmen muss begründet werden)

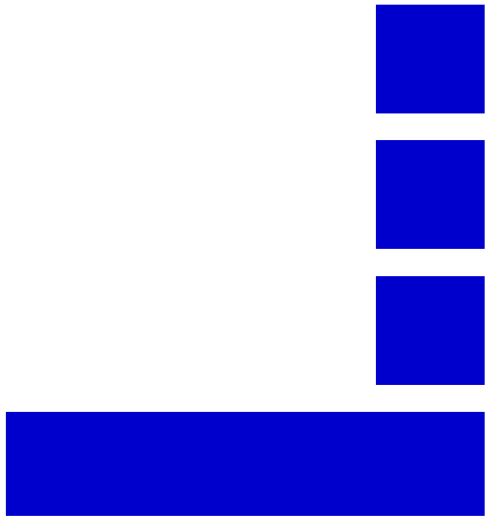
De-Duplizierung bei Sammlung der Nutzungsdaten verschiedener IRs:
- wie erfolgt De-Duplizierung?
- wo erfolgt die De-Duplizierung?
- Orientierung an LANL bX, Google Scholar, CiteSeer

Datenaustausch
-Wie werden die Daten mit Verlagsdaten/-statistiken zusammengeführt/ verglichen?
- Wie findet der Datenaustausch statt? In beide Richtungen?
- Verlage müssten ebenfalls Policy für die Verwendung der Nutzungsdaten haben
- Welche Quellen – außer Verlagen – kommen in Frage?

Ulrich Herb
Saarländische Universitäts- und Landesbibliothek (SULB)

Vielen Dank für Ihre Aufmerksamkeit!

Email
u.herb@sulb.uni-saarland.de



Ulrich Herb
Saarländische Universitäts- und Landesbibliothek (SULB)