Leitfaden
für die Teilnahme an OA-Statistik

Version 1.0, September 2012

Computer- und Medienservice der Humboldt-Universität zu Berlin
Niedersächsische Staats- und Universitätsbibliothek Göttingen
Saarländische Universitäts- und Landesbibliothek Saarbrücken
Universitätsbibliothek Stuttgart
und
Verbundzentrale des GBV (VZG)

Auf Initiative und in Abstimmung mit der
Deutschen Initiative für Netzwerkinformation e. V.

Schritte zur Anbindung des eigenen Repositoriums an die Infrastruktur von OA–Statistik

Der Dienst OA-Statistik befindet sich zurzeit noch im Testbetrieb!

Im Laufe der Projektlaufzeit werden alle erforderlichen Datenschutzmaßnahmen umgesetzt, daher können Sie unsere Daten noch nicht für den produktiven Betrieb nutzen. Sollte es zu technischen oder gesetzlichen Problemen kommen, können die beteiligten Projektpartner nicht dafür zur Verantwortung gezogen werden.

Übersicht:

1 Softwarepakete von der OA-Statistik-Projektwebsite herunterladen

Unter http://www.dini.de/projekte/oa-statistik/die-projektergebnisse/software-download/ finden Sie eine Auswahl von Softwarepaketen, die Sie für die Anbindung Ihres Repositoriums an OA-Statistik benötigen. Das Paket „Data-Provider für DSpace und WebDoc“ bietet den Quellcode mit einer Dokumentation des OAS-Data-Providers an, der für unterschiedliche Repositorien wie z.B. DSpace angepasst werden kann. Das Paket „Data-Provider für OPUS“ bietet den Quellcode des Data-Providers für OPUS-Repositorien an.

2 Anpassung des OAS-Data-Providers an das eigene Repositorium

Der OAS-Data-Provider besteht im Wesentlichen aus zwei Elementen:

- Der Logfile-Parser analysiert die Server-Logfiles, wertet ergänzende Metadaten des Repositoriums aus und bündelt diese als OpenURL ContextObjects. Diese werden lokal gespeichert.
- Der OAS-Data-Provider stellt diese OpenURL ContextObjects über ein standardisiertes OAI-Interface zur Verfügung.

Die Installation des Paketes „Data-Provider für DSpace und WebDoc“ wird im Folgenden erläutert:

Systemvoraussetzungen:

Der OAS-Data-Provider arbeitet unter einem „LAMP“-System, d.h. einem Linux-Server mit dem Apache Web Server und PHP unter den folgenden Voraussetzungen:

- PHP-Version 5.2.x oder eine neuere Version
    - Der Logfile-Parser benötigt PHP mit pcntl (bitte beachten Sie, dass für Windows keine pcntl-Erweiterung zur Verfügung steht) und CLI.
    - Der Logfile-Parser und das OAI-Interface benötigen die DOM-Erweiterung.
- Eine Datenbank zum Speichern der OpenURL ContextObjects und Metadaten: Diese muss über „PHP Data Objects“ (PDO) angesprochen werden können (z.B. MySQL, PostgreSQL, SQLite). Zudem ist ein Datenbankzugang erforderlich, der über die Rechte für das Anlegen von Tabellen verfügt.

Installation:

- Entpacken Sie die beiden Elemente Logfile-Parser und OAS-Data-Provider auf Ihrem Repositorium.
- Der Logfile-Parser kann in einem frei wählbaren Verzeichnis gespeichert werden, allerdings sollte er – aus Sicherheits- und Datenschutzgründen – nicht von außen zugänglich sein. Der OAS-Data-Provider sollte im Wurzelverzeichnis des Webservers liegen und muss von außen zugänglich sein; nutzen Sie dazu bitte .htaccess. (Siehe Absatz 5).

Konfiguration:

Für DSpace-Repositorien stellen wir eine angepasste Klasse bereit. Editieren Sie die Konfigurationsdateien im Hauptverzeichnis „Logfile-Parser“, bestehend aus einer allgemeinen Konfigurationsdatei namens „*config.php*“ und einer Repositorium - spezifischen Datei „*config-dspace.php*“.

Wenn Sie einen anderen Server benutzen, müssen Sie diese Klassen ggf. selbst erstellen und konfigurieren. Die vom „*oasparser-webserver-standard.php*“ erstellten OpenURL ContextObjects stellen keinen ServiceType und keinen Identifier außer der URI bereit. Deswegen muss „*logfile-parser/lib/oasparser-webserver-standard.php*“ durch eine Repositorien-spezifische Klasse in „*logfile-parser/lib/oasparser-webserver-<name>.php*“ erweitert werden.

Diese benötigt wiederum ihre eigene Programmbibliothek („*logfile-parser/lib/identifiers/lib-<name>.php*“) um Metadaten zu erfassen, die nicht durch das Logfile ermittelt werden können sowie eine Konfigurationsdatei namens „*logfile-parser/config-<name>.php*“.
Binden Sie die Datei *lib-<name>.php* mit „*require_once*“ in „*logfile-parser/lib/oasparser-webserver-<name>.php*“ ein und legen Sie in „*config-<name>.php*“ auf die gleiche Weise Ihren „*oasparser*“ fest.

Bitte legen Sie eine Datenbank und einen Nutzer für Ihre OpenURL ContextObjects an. Weisen Sie diesem neuen Nutzer die Datenbank zu (mit „grant“).

Sobald dieser Schritt durchgeführt ist, initialisieren Sie die Datenbank, um die Tabelle für die OpenURL ContextObjects automatisch zu erstellen. Rufen Sie dazu „*log2ctx -O*“ auf. Die Konfigurationsdatei *config.php* wird standardmäßig verwendet, benutzen Sie „*log2ctx -c*“ für eine benutzerdefinierte Konfigurationsdatei. Der Parameter - i wird benutzt, um den OpenURL

ContextObjects eine Zeitangabe zuzuweisen, über die veraltete OpenURL ContextObjects identifiziert und gelöscht werden können. Weitere Informationen finden Sie in der readme-Datei („*readme.txt*“).

Falls Sie Schwierigkeiten mit der automatischen Erstellung der Tabelle haben sollten, können Sie diese auch manuell anlegen:

```
CREATE TABLE `contextobjects_test` (`timestamp` int(11) DEFAULT NULL, `identifier` varchar(255) DEFAULT NULL,`line` int(11) DEFAULT NULL,`data` blob, KEY timestamp (timestamp)) ENGINE=MyISAM DEFAULT CHARSET=latin1;
```

Es gibt zwei verschiedene Unix-Shell-Scripts, die mit dem Logfile-Parser zusammenarbeiten:

- In „*update-script.sh*“ wird beschrieben, wie Metadaten gesammelt und Webserver-Logfiles in OpenURL ContextObjects konvertiert werden. Passen Sie diese Datei für Ihren OAS-Data-Provider an.
- Passen Sie bitte auch „*cleanup-script.sh*“ an, und verwenden Sie diese, um veraltete OpenURL ContextObjects zu entfernen.

Beide Skripte sollten regelmäßig mit Hilfe eines Cron-Jobs aufgerufen werden.

3 Aktualisierung des SALTs

Zur Anonymisierung der übertragenen Logdateninformationen wird neben einem SHA-Hashing auch ein SALT verwendet, dessen Wert monatlich wechselt. Der Salt besteht aus einer zufällig gewählten Zeichenfolge. Er wird an die IP angehängt, bevor diese gehasht wird. Wenn alle beteiligten Repositorien den gleichen Salt und das gleiche Hashverfahren benutzen, können anonymisierte IP-Adressen Repositorien-übergreifend nachverfolgt werden (anonymisierte Daten). Der jeweils aktuelle Wert des SALTs ist unter https://oas.sulb.uni-saarland.de/salt/salt_value.txt zu finden.

Da die VZG als Anbieter des OAS-Service-Providers aus datenschutzrechtlichen Gründen den SALT nicht kennen darf, stellt die SULB Saarbrücken diesen zur Verfügung. Der Wert wird automatisch um 0:00 Uhr an jedem Monatsersten auf einem Server in Saarbrücken aktualisiert. Im Installationspaket „Data-Provider für DSpace und WebDoc“ ist dieses Skript bereits enthalten, es kann aber auch unter https://oas.sulb.uni-saarland.de/salt/get_salt.sh herunter geladen werden. In diesem Fall muss die heruntergeladene Datei mit der Extension .txt in eine Datei mit der Extension .php umbenannt werden. In beiden Fällen muss die Datei für den eigenen Server konfiguriert werden. Das Skript testet, ob sich der Wert tatsächlich im Vergleich zum bisherigen lokal gespeicherten Wert (sofern vorhanden) geändert hat und startet ansonsten weitere Aktualisierungsversuche, so dass auch im Falle von Netzwerkproblemen o.ä. eine - wenn auch verzögerte - Aktualisierung des SALTs ohne manuelles Eingreifen gewährleistet ist. Daher sollte das Skript ebenfalls um Mitternacht an jedem Monatsersten mit einem Cron-Befehl aufgerufen werden.

Zum Zugriff wird ein Benutzername, der aus der Email-Adresse des jeweiligen Ansprechpartners des zugehörigen Repositoriums besteht, sowie ein Passwort benötigt. Sollten Sie diese Zugangsdaten nach einer Registrierung (siehe Punkt 8) nicht automatisch zugeschickt bekommen, wenden Sie sich bitte an: oas@sulb.uni-saarland.de

4 Nutzung der Datei Robots.txt

Für die Erkennung von automatisierten Zugriffen (Suchmaschinenrobotern etc.) auf die Dokumente des Repositoriums sollte eine Datei robots.txt auf Ihrem Webserver existieren. Diese kann auch leer sein. Wird diese Datei aufgerufen, so ist dieser Akteur vermutlich ein Roboter und kann als solcher zentral auf unserem OAS-Service-Provider klassifiziert und herausgefiltert werden.

5 Geschützter Ort für den OAS-Data-Provider mit .htaccess

Aus Datenschutzgründen dürfen weder Ihre ursprünglichen Logdaten noch die für OA-Statistik aufbereiteten Logdaten in Form von OpenURL ContextObjects von Dritten erreichbar sein. Daher muss das Verzeichnis des OAS-Data-Providers mit Hilfe von .htaccess vor fremden Zugriffen geschützt werden. Bitte wenden Sie sich an uns, um die Zugänge zu diesem Verzeichnis mit uns abzusprechen.

Kontakt:

Marco Recke
Verbundzentrale des GBV
Abteilung Digitale Bibliothek
Platz der Göttinger Sieben 1
37073 Göttingen
Tel.: 0551/3912323
E-Mail: recke@gbv.de

6 Apache Konfiguration: Log-Konfiguration

Der Apache-Webserver Ihres Repositoriums sollte so konfiguriert sein, dass Logfiles geschrieben werden (oder in eine Ausgabe umgeleitet werden können).

Für die Log-Einstellung wird das „Combined Log Format“ im Modul mod_log_config benötigt:

```
LogFormat "%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{User-agent}i\"" combined
```

Diese Einstellung beinhaltet zusätzlich zum „Common Log Format“ Informationen über den Referer (die direkt zuvor besuchte Website) und den User-Agent, der zur Roboteridentifikation benötigt wird.

7 Implementation eines Opt-Outs

Da Sie Nutzungserfassungen durchführen, müssen Sie aus Datenschutzgründen Ihren Nutzerinnen und Nutzern ein „Opt-Out“ anbieten. Dies könnte z.B. ein Button sein, über den ein Cookie auf den Computer Ihrer Nutzer abgelegt wird, der wiederum vom Ihrem Webserver ausgelesen wird und verhindert, dass die IP-Adressen Ihrer Nutzer aufgezeichnet werden.

8 Registrierung Ihres OAS-Data-Providers

Bitte nutzen Sie das Registrierungsformular auf unserer Website, wenn Sie die Schritte 1-7 zur Anbindung Ihres Repositoriums an die Infrastruktur von OA–Statistik erfolgreich durchgeführt haben.

http://www.dini.de/projekte/oa-statistik/die-technik/registrierung/

9 Anbieten der Ergebnisse auf Ihrem Repositorium

Den teilnehmenden Repositorien werden für die einzelnen Dokumente tagesgenaue Zugriffszahlen bereitgestellt. Dies können zum einem Datenbankexporte für alle bereits erfassten Identifier des OAS-Service-Providers in den Formaten JSON oder XML sein. Diese sind über eine Schnittstelle abrufbar. Zum anderen wird von OA-Statistik ein iFrame angeboten, das Sie in Ihr Repositorium einbinden und dem von Ihnen gewählten Personenkreis sichtbar machen können. Als Parameter muss der Identifier des Dokuments mitgegeben werden.