Preliminary results of the project OA-Statistik

Goals of OA-Statistics

We aim to produce valid and reliable document usage statistics based solely on information gathered from the HTTP layer.
There are two main issues addressed by all existing standards which generate the bulk of the necessary corrections:

- Identification of non-human access
- Multi-Click correction

Besides this, we investigate the amount of data and effort necessary to produce complex statistics, for example, click-streams, without violating privacy laws. At the bottom of this page there is a comparison table including links to all standards mentioned. A detailed description of OA-S can be found at http://www.dini.de/projekte/oa-statistik/#c1203
Usage statistics - and even more important raw usage data - have to be described on an abstract level. It is not sufficient to define a derivative of the Apache Access Log as there is a multitude of different software solutions in use to operate a full text repository. Many do not even produce a log file let alone utilise an Apache Server.

Information needed to generate COUNTER, LogEc and IFABC

Note: The field names might still be subject to change as the project goes on.

OA-S-Fieldname	Description	COUNTER	LogEc	IFABC*
Document-Identifier	non-ambigious label identifying the full text	needed	needed	needed
File Format	File format of server reply (e.g. HTML orPDF)	needed	needed	needed
Service Type	nature of server reply (e.g. full text,ab-stract)	needed	needed	-
Time of Request	Time of request processing to the second	needed	needed	needed
IP	IP-Adress of user (Client)	needed	needed	IF Session-Identifier is not available: needed
Session-Identifier	server generated non-ambiguous session/visit label	optional	-	IF IP is not available: needed
User Agent	User-Agent-String of the requesting client	needed	needed	IF Session-ID is not available: needed
HTTP Status Code	Server-Status-Code of the HTTP-Requests	needed	needed	needed
Bytes sent	server reply size	-	-	IF File Format is not HTML: needed

Additional pieces of information which comply with OpenURL Context Objects

The following fields are important to our advanced research interests and thus implemented from the beginning.

Referrer	non-ambigious identifier of the server which created the ContextObject
Referring Entitiy	non-ambigious label of the object of origin (e.g. the Abstract Page which links to the full text file)

Additional suggestions

States and properties of the repository software have to be delivered from the available data.
Examples:

- Focus Page in Search Result Paging View
- ID of the current document
- Search arguments and result presentation
- Abstract Page vs. Fulltext Page
- Administrative actions
- Document upload
- Metadata allocation

There should be reliable information about the origin of the client (i.e. the referrer). For example, it should be possible to tell whether a client accessed the file via the frontpage or via a link in the repository's RSS-Feed.
In case of multiple server logs it is mandatory to synchronize the system time on all associated repository servers.

Table of Web Usage Standards

Provider URL	Counting Clause	Multi-Click Time Span	User Identification	Crawler Clause	Crawler Identification	Crawler Count Report
Counter Code of Practice Draft 3	HTTP Status Code is 200 or 304.	for HTML 10s; for PDF 30s	at least IP, preferably Session	robots, prefetches, caching, federated searches(n.a.)	Black-List, client HTTP header	separate report
About LogEc	HTTP Status Code is 200, 206, 301, 302 or 304.	one calendar month	IP	robots, automated downloads (wget)	Access of robots.txt; # of requests 10,000 items /month; C-Class access 10% of stock; known robot-Domain/IP	separate column in report
Interoperable Repository Statistics	HTTP Status code is 200 on abstract or full-text page	24 hours	IP	search engine crawlers + automated	AWStats' black list	discarded
AWStats	Default: HTTP Status codes (200;304)	Default: 1 hour	IP	search engine crawlers	Black-List	separate column in report
IFABC	HTML: Tracking Pixel; Other: bytes transferred 95% of file size	Each Pageview is counted only once per visit. Visit means series of clicks coming from one IP-Number/Session-ID less than 30 minutes apart.	IP+User-Agent; Cookie-Session, Login-Session	search engine crawlers; automated downloads (optional)	proprietary Blacklist	discarded