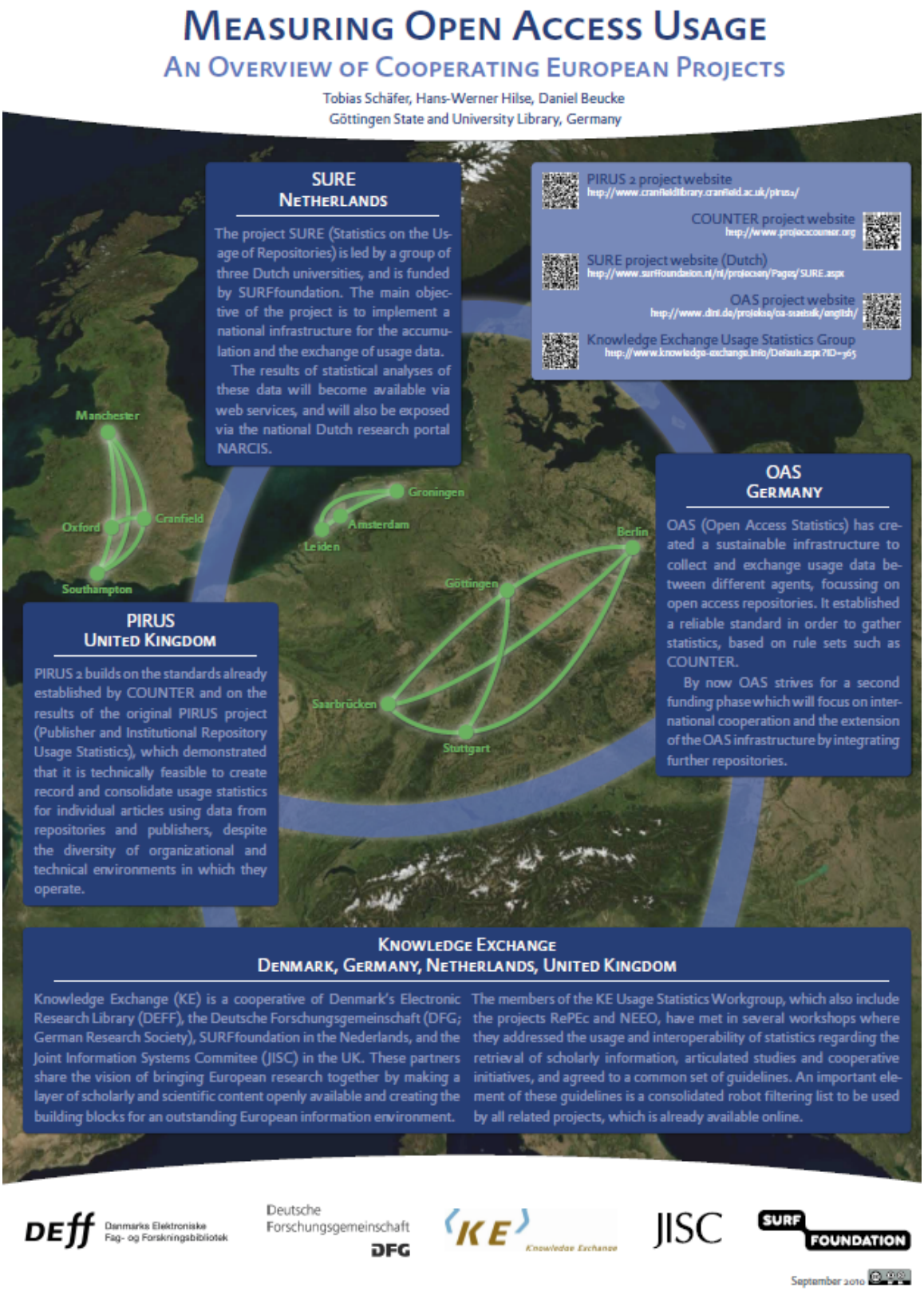
# KE Usage Statistics Guidelines

# Guidelines for the aggregation and exchange of Usage Data

Knowledge Exchange and Usage Statistics

KE is a co-operative effort between JISC, SURF, DEFF and DFG. International interoperability guidelines for the comparable exchange of usage data is one of these co-operative efforts.




# Table of contents


- Guidelines for the aggregation and exchange of Usage Data

- Table of contents
- Document information
- Abstract
- Endorsement
- 1. Introduction
- 2. Terminology and strategy
- 3. Data format
  - 3.1. Core set
    - 3.1.1. <context-object>
    - 3.1.2. <referent>
    - 3.1.3. <referringEntity>
    - 3.1.4. <requester>
    - 3.1.5. <service-type>
    - 3.1.6. <resolver> and <referrer>
  - 3.2. Extensions
    - 3.2.1. <requester>
- 4. Transfer Protocols
  - 4.1. OAI-PMH
    - 4.1.1. Requirement for metadata record identifiers
    - 4.1.2. Datestamps for records
    - 4.1.3. MetadataPrefix
    - 4.1.4. Mandated metadata in Dublin Core (DC) format
    - 4.1.5. Usage of Sets
    - 4.1.6. Deletion tracking
    - 4.1.7. Metadata formats
    - 4.1.8. Inclusion of Context Objects in OAI-PMH records
  - 4.2. SUSHI
- 5. Normalisation
  - 5.1. Double Clicks
  - 5.2. Robot filtering
    - 5.2.1. Definition of a robot
    - 5.2.2. Strategy
    - 5.2.3. Robot list schema
- 6. Legal boundaries
  - 6.1. Usage of IP addresses and the protection of a 'natural person'
- Appendices


# Document information


**Title:** KE Usage Statistics Guidelines
**Subject:** Usage Statistics, Guidelines, Repositories, Publications, Research Intelligence
**Moderator:** Peter Verhaar (KE Usage Statistics Work Group)
**Version:** 1.0
**Date published:** 2010-05-18
**Excerpt**:Guidelines for the exchange of usage statistics from a repository to a central server using OpenURL Context Objects via OAI-PMH or SUSHI.

(Optional information)
**Type:** Guidelines, info:eu-repo/semantics/technicalDocumentation
**Format:** html/text
**Identifier:** http://purl.org/REP/standards/KE Usage Statistics Guidelines
**Language:** EN
**Rights:** CC-BY
**Tags:**


| Date | version | Author | Description | PDF |
|---|---|---|---|---|
| 2010-05-26 | 1.0 | Peter Verhaar | Definitive version 1.0.; Comments made during the phone conference which took place on 25-05-2010 and which was attended by Thobias Schäfer, Hans-Werber Hilse, Jochen Schirrwagen, Marek Imialek, Paul Needham, Peter Verhaar, Maurice Vanderfeesten, Natalia Manola and Lefteris Stamatogiannakis | Download |

| | | | | |
|---|---|---|---|---|
| 2010-05-18 | 0.9.5 | Peter Verhaar | DRAFT version 1.0 ; comments and layout improvements made by Jochen Schirrwagen, Max Kemman, Peter Verhaar and Maurice Vanderfeesten | |
| 2010-04-24 | 0.9 | Peter Verhaar | Revised version, in which comments made by Benoit Pauwels, Hans-Werner Hilse, Thobias Schäfer, Daniel Metje and Paul Needham have been incorporated. | |
| 2010-04-13 | 0.2 | Maurice Vanderfeesten | Added the sections based on the Knowledge Exchange meeting in Berlin. And filled in some additional information to these sections. | |
| 2010-03-25 | 0.1 | Peter Verhaar | First draft, based on technical specifications from the OA-Statistics project (written by Daniel Metje and Hans-Werner Hilse), the NEEO project (witten by Benoit Pauwels) and the SURE project (written by Peter Verhaar and Lucas van Schaik) | |

# Abstract

Guidelines for the exchange of usage statistics from a repository to a central server using OAI-PMH and OpenURL Context Objects.

This page is maintained by: KE Usage Statistics Work Group

# Endorsement

The following projects and parties will endorse these guidelines, by applying these in their implementation.

| Project | Status |
|---|---|
| SURF Sure | |
| PIRUS2 | |
| OA-Statistics | |
| NEEO | |
| COUNTER | |

# 1. Introduction

The impact or the quality of academic publications is traditionally measured by considering the number of times the text is cited. Nevertheless, the existing system for citation-based metrics has frequently been the target of serious criticism. Citation data provided by ISI focus on published journal articles only, and other forms of academic output, such as dissertations or monographs are mostly neglected. In addition, it normally takes a long time before citation data can become available, because of publication lags. As a result of this growing dissatisfaction with citation-based metrics, a number of research projects have begun to explore alternative methods for the measurement of academic impact. Many of these initiatives have based their findings on usage data. An important advantage of download statistics is that they can readily be applied to all electronic resources, regardless of their contents. Whereas citation analyses only reveal usage by authors of journal articles, usage data can in theory be produced by any user. An additional benefit of measuring impact via the number of downloads is the fact that usage data can become available directly after the item has been placed on-line.

Virtually all web servers that provide access to electronic resources record usage events as part of their log files. Such files usually provide detailed information on the items that have been requested, on the users that have initiated these requests, and on the moments at which these requests took place. One important difficulty is the fact that these log files are usually structured according to a proprietary format. Before usage data from different institutions can be compared in a meaningful and consistent way, the log entries need to be standardised and normalised. Various projects have investigated how such data harmonisation can take place. In the MESUR project, usage data have been standardised by serialising the information from log files into XML files structured according to the OpenURL Context Objects schema (Bollen and Van de Sompel, 2006). This same standard is recommended in the *JISC Usage Statistics Final Report*. Using this metadata standard, it becomes possible to set up an infrastructure in which usage data are aggregated within a network of distributed repositories.

In Europe, at least four projects have experimented with these recommendations and have actually implemented an infrastructure for the central accumulation of usage data:

1. *Publishers and Institutional Repository Usage Statistics* (PIRUS), which was funded by JISC, aims to develop COUNTER-compliant usage reports at the individual article level that can be implemented by any entity (publisher, aggregator, IR, etc.,) that hosts online journal articles. The project will enable the usage of research outputs to be recorded, reported and consolidated at a global level in a standard way. An important outcome of this project was a range of scenarios for the "creation, recording and consolidation of individual article usage statistics that will cover the majority of current repository installations" "Developing a global standard to enable the recording, reporting and consolidation of online usage statistics for individual journal articles hosted by institutional repositories, publishers and other entities (Final Report)", p.3. <http://www.jisc.ac.uk/media/documents/programmes/pals3/pirus_finalreport.pdf>.
2. Te German *OA-Statistics* <http://www.dini.de/projekte/oa-statistik/> project, which is funded DINI (Deutsche Initiative für Netzwerk Information), has set up an infrastructure in which various certified repositories across Germany can exchange their usage data.
3. In the Netherlands, the project *Statistics on the Usage of Repositories* <http://www.surffoundation.nl/nl/projecten/Pages/SURE.aspx> (SURE) has a very similar objective. The project, which is funded by SURFfoundation, aimed to find a method for the creation of reliable and mutually comparable usage statistics and has implemented a national infrastructure for the accumulation of usage data.
4. The *Network of European Economists Online* <http://www.neeoproject.eu/> (NEEO) is an international consortium of 18 universities which maintains a subject repository that provides access to the results of economic research. As part of this project, extensive guidelines have been developed for the creation of usage statistics. NEEO has also developed an aggregator for usage statistics. The central database is exposed via a web service which can provide information on the number of downloads for each publication.

Whereas these four projects all make use of the OpenURL Context Object standard, some subtle differences have emerged in the way in which this standard is actually used. Nevertheless, it is important to ensure that statistics are produced in exactly the same manner, since, otherwise, it would be impossible to compare metrics produced by different projects. With the support of Knowledge Exchange, a collaborative initiative for leading national science organisations in Europe, an initiative was started to align the technical specifications of these various projects. This document is a first proposal for international guidelines for the accumulation and the exchange of usage data. The proposal is based on a careful comparison of the technical specifications that have been developed by these three projects.

# 2. Terminology and strategy

A usage event takes place when a **user** downloads a **document** which is managed in a **repository**, or when a user views the **metadata** that is associated with this document. The user may have arrived at this document through the mediation of a **referrer**. This is typically a search engine. Alternatively, the request may have been mediated by a **link resolver**. The usage event in turn generates **usage data**.

The institution that is responsible for the repository that contains the requested document is referred to as a **usage data provider**. Data can be stored locally in a variety of formats, but to allow for a meaningful central collection of data, usage data providers must be able to expose the data in a standardised **data format**, so that they can be harvested and transferred to a central database. The institution that manages the central database is referred to as the usage data **aggregator**. The data must be transferred using a well-defined transfer protocol. The data aggregator harvests individual usage data providers minimally on a daily basis, and bears the primary responsibility for synchronising the local and the central data. Ultimately, certain **services** can be built on the basis of the data that have been accumulated (see *figure 1*)

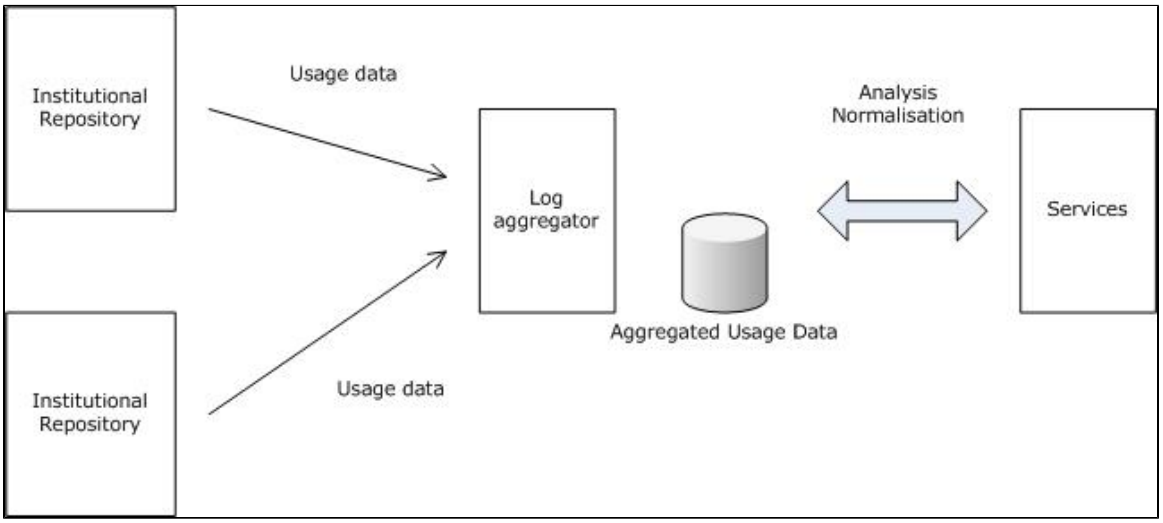



*Figure 1.*

The approach that is proposed here coincides largely with scenario B that is described in the final report of PIRUS1 (see *figure 2*). In this scenario, "the generated OpenURL entries are sent to a server hosted locally at the institution, which then exposes those entries via the OAI-PMH for harvesting by an external third party".

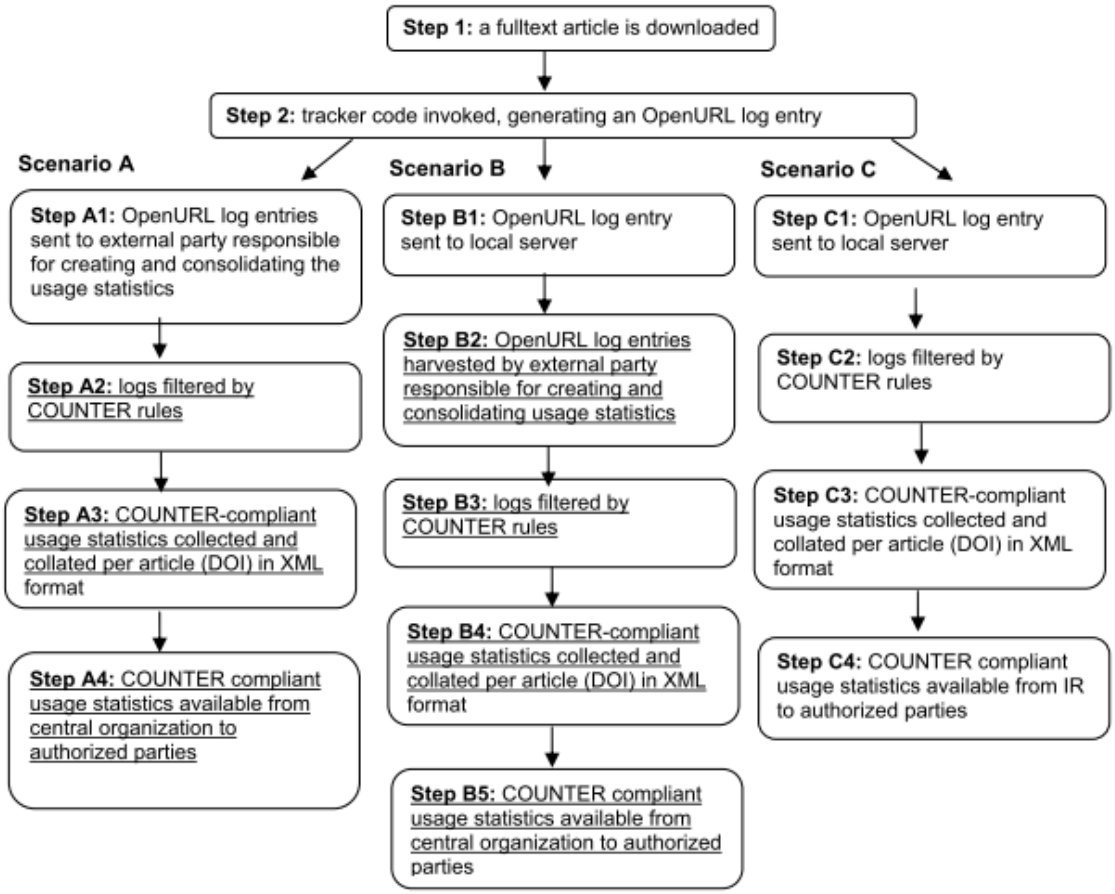



*Figure 2. PIRUS1 scenario's (as from the PIRUS final report)*

The main advantages of secanrio B is that that normalisation does not have to be carried out by individual repositories. Once the data have been received by the log aggregator, the normalisation rules can be applied consistently to all data. Since local repositories only need to make sure that their data can be exposed for harvesting, the implementation should be much easier.

# 3. Data format

To be able to compare usage data from different repositories, the data need to be available in a uniform format. This section will provide specifications for the aspects of the usage event that need to be recorded. In addition, guidelines need to be developed for the format in which this information can be expressed. Following recommendations from MESUR and the JISC Usage Statistics Project, it will be stipulated that usage events need to be serialized in XML using the data format that is specified in the OpenURL Context Objects schema. The XML Schema for XML Context Objects can be accessed at http://www.openurl.info/registry/docs/info:ofi/fmt:xml:xsd:ctx. The specifications for the use of OpenURL ContextObject in this section are more restrictive than the original schema with respect to the occurrence and the permitted values of elements and attributes.

A distinction will be made between a **core set** and **extensions**. Data in the core set can be recorded using standard elements or attributes that are defined in the OpenURL Context Object schema. The extensions are created to record aspects of usage events which cannot be captured using the official schema. They have usually been defined in the context of individual projects to meet very specific demands. Nevertheless, some of the extensions may be relevant for other projects as well. They are included here to inform the usage statistics community about the additional information that could be made available. Naturally, the implementation of all the extension elements is optional.

## 3.1. Core set

### 3.1.1. <context-object>

The OpenURL Framework Initiative recommends that each community that adopts the OpenURL Context Objects Schema should define its application profile. Such a profile must consist of specifications for the use of namespaces, character encodings, serialisation, constraint languages, formats, metadata formats, and transport protocols. This section will attempt to provide such a profile, based on the experiences from the projects PIRUS, NEEO, SURE and OA-Statistics.
The root element of the XML-document must be <context-objects>. It must contain a reference to the official schema and declare the following namespace:

| | |
|---|---|
| OpenURL Context Objects | info:ofi/fmt:xml:xsd:ctx |

Each usage event must be described in a separate <context-object> element. All individual <context-object> elements must be wrapped into the <context-objects> root element.
Each <context-object> must have a timestamp attribute, and it may optionally be given an identifier attribute. These two attributes can be used to record the request time and an identification of the usage event. Details are provided below.

#### <context-object/@timestamp> | Request Time

| | |
|---|---|
| Description | The exact time on which the usage event took place. |
| XPath | ctx:context-object/@timestamp |
| Usage | Mandatory |
| Format | The format of the request time must conform to ISO8601. The YYYY-MM-DDTHH:MM:SS representation must be used. |
| Example | 2009-07-29T08:15:46+01:00 |

#### <context-object/@identifier> | Usage Event ID

| | |
|---|---|
| Description | An identification of a specific usage event. |
| XPath | ctx:context-object/@identifier |
| Usage | Optional |
| Format | No requirements are given for the format of the identifier. If this optional identifier is used, it must be (1) opaque and (2) unique for a specific usage event.<br>In the Netherlands a MD5 Hash is generated from a concatenation of a code for the institution, the identifier of the publication and the timestamp. Other projects may of course use different ways to create identifiers, as long as these are globally unique. |
| Example | b06c0444f37249a0a8f748d3b823ef2a |

## Occurences of child elements in <context-object>

Within a <context-object> element, the following subelements can be used:

| Element name | minOccurs | maxOccurs |
|---|---|---|
| Referent | 1 | 1 |
| ReferringEntity | 0 | 1 |
| Requester | 1 | 1 |
| ServiceType | 1 | 1 |
| Resolver | 1 | 1 |
| Referrer | 0 | 1 |

# 3.1.2. <referent>

The <referent> element must provide information on the item that is requested. More specifically, it must record the following data elements.

## <referent/identifier> | Location of object file or metadata file

| | |
|---|---|
| Description | The URL of the object file or the metadata record that is requested. Since this document focuses on usage by means of the World Wide Web, there will always be one URL for each usage event. |
| XPath | ctx:context-object/ctx:referent/ctx:identifier |
| Usage | Mandatory |
| Format | URL |
| Example | https://openaccess.leidenuniv.nl/bitstream/1887/12100/1/Thesis.pdf |

## <referent/identifier> | Other identifier of requested item

| | |
|---|---|
| Description | A globally unique identification of the resource that is requested must be provided if there is one that is applicable to the item. Identifiers should be 'communication protocol'-independent as much as possible. In the case of a request for an object file, the identifier should enable the aggregator to obtain the object's associated metadata file. When records are transferred using OAI-PMH, providing the OAI-PMH identifier is mandatory. |
| XPath | ctx:context-object/ctx:referent/ctx:identifier |
| Usage | Mandatory if applicable |
| Format | URI |

| | |
|---|---|
| Example | http://hdl.handle.net/1887/12100 |

## 3.1.3. <referringEntity>

The <ReferringEntity> provides information about the environment that has directed the user to the item that was requested. This referrer can be expressed in two ways.

### <referringEntity/identifier> | Referrer URL

| | |
|---|---|
| Description | The entity which has directed the user to the referent. When the request was mediated by a search engine, record the URL provided by the HTTP referrer string. |
| XPath | ctx:referring-entity/ctx:identifier |
| Usage | Mandatory if applicable |
| Format | URL |
| Example | http://www.google.nl/search?hl=nl&q=beleidsregels+artikel+4%3A84&meta= |

## 3.1.4. <requester>

The user who has sent the request for the file is identified in the <requester> element.

### <requester/identifier> | IP-address of requester

| | |
|---|---|
| Description | The user can be identified by providing the IP-address. Including the full IP-address in the description of a usage event is not permitted by international privacy laws. For this reason, the IP-address needs to be obfuscated. The IP-address must be hashed using salted MD5 encryption. The salt must minimally consist of 12 characters. The IP address of the requester is pseudonymised using encryptions, before it is exchanged and taken outside the web-server to another location. Therefore individual users can be recognised when aggregated from distributed repositories, but they cannot be identified as 'natural persons'. This method appears to be consistent with the European Act for Protection of Personal data. The summary can be found here: ? http://europa.eu/legislation_summaries/information_society/l14012_en.htm . Further legal research is needed to determine of if this method is sufficient to protect the personal data of a 'natural person', in order to operate within the boundaries of the law. |
| XPath | ctx:context-object/ctx:requester/ctx:identifier |
| Usage | Mandatory |
| Format | A data-URI, consisting of the prefix "data:,", followed by a 32-digit hexadecimal number. |
| Example | data:,c06f0464f37249a0a9f848d4b823ef2a |

### <requester/.../dcterms:spatial> | Geographic location

| | |
|---|---|
| Description | The country from which the request originated may also be provided explicitly. |

| | |
|---|---|
| XPath | ctx:context-object/ctx:requester/ctx:metadata-by-val/ctx:metadata/dcterms:spatial<br><br>If this element is used, the <metadata> element must be preceded by<br>ctx:requester/ctx:metadata-by-val/ctx:format<br>with value<br>"http://dublincore.org/documents/2008/01/14/dcmi-terms/" |
| Usage | Optional |
| Format | A two-letter code in lower case, following the ISO 3166-1-alpha-2 standard. http://www.iso.org/iso/english_country_names_and_code_elements |
| Example | ne |

## 3.1.5. <service-type>

### <service-type/.../dcterms:type> | Request Type

| | |
|---|---|
| Description | The request type provides information on the type of user action. Currently, this element is only used to make a distinction between a download of an object file and a metadata record view. In the future, extensions can be defined for other kinds of user actions, such as downloads of datasets, or ratings. |
| XPath | ctx:context-object/ctx:service-type/ctx:metadata-by-val/ctx:metadata/dcterms:type<br><br>If this element is used, the <metadata> element must be preceded by<br><br>ctx:requester/ctx:metadata-by-val/ctx:format<br><br>with value<br><br>"http://dublincore.org/documents/2008/01/14/dcmi-terms/" |
| Inclusion | Mandatory |
| Format | One of these values must be used:<br>• info:eu-repo/semantics/objectFile or<br>• info:eu-repo/semantics/descriptiveMetadata See for explanation of these concepts info:eu-repo Object types |
| Example | info:eu-repo/semantics/objectFile |

## 3.1.6. <resolver> and <referrer>

### <resolver/identifier> | Host name

| **Host name** | |
|---|---|
| Description | An identification of the institution that is responsible for the repository in which the requested item is stored. |
| XPath | ctx:context-object/ctx:resolver/ctx:identifier |
| Usage | Mandatory |
| Format | The baseURL of the repository must be used. This must be a URI, and not only the domain name. |
| Example | http://openaccess.leidenuniv.nl/dspace-oai/request |

### <resolver/identifier> | Location of OpenURL Resolver

| | |
|---|---|
| Description | In the case of link resolver usage data, the baseURL of the OpenURL resolver must be provided. |
| XPath | ctx:context-object/ctx:resolver/ctx:identifier |
| Usage | Optional |
| Format | URL |
| Example | http://sfx.gbv.de:9004/sfx_sub/ |

### <referrer/identifier> | Referrer Identifier

| | |
|---|---|
| Description | Identification of the external service that had provided a reference to the referent, e.g. a search engine |
| XPath | ctx:context-object/ctx:referrer/ctx:identifier |
| Usage | Optional |
| Format | A URI that is registered in http://info-uri.info/registry/OAIHandler?verb=GetRecord&metadataPrefix=reg&identifier=info:sid/ |
| Example | info:sid/dlib.org:dlib<br>info:sid/google.com |

## 3.2. Extensions

## 3.2.1. <requester>

### <requested/identifier> | C-class Subnet

| | |
|---|---|
| Description | When the IP-address is obfuscated, this will have the disadvantage that information on the geographic location, for instance, can no longer be derived. For this reason, the C-Class subnet must be provided. The C-Class subnet, which consists of the three most significant bytes from the IP-address, is used to designate the network ID. The final (most significant) byte, which designates the HOST ID, is replaced with a '0'. The C-class Subnet may optionally be hashed using MD5 encryption. |
| XPath | ctx:context-object/ctx:requester/ctx:identifier If the C-Class subnet is hashed, the MD5 hash must be provided in the following element: ctx:context-object/ctx:metadata/dini:requesterinfo/dini:hashed-c If this element is used, the <metadata> element must be preceded by ctx:requester/ctx:metadata-by-al/ctx:format with value "http://dini.de/namespace/oas-requesterinfo" |
| Usage | Optional |
| Format | A data-URI, consisting of the prefix "data:,", followed either by a 32-digit hexadecimal number, or by three hexadecimal numbers separated by a dot, followed by a dot and a '0'. |
| Examples | data:,208.77.188.0<br>data:,ec17f0564f32240c0a9d848d4b823ef2a |

### <requester/.../dini:classification> | Classification of the requester

| | |
|---|---|
| Description | The user may be categorised, using a list of descriptive terms. If no classification is possible, it must be omitted. |
| XPath | ctx:context-object/ctx:requester/ctx:metadata/dini:requesterinfo/dini:classificationIf this element is used, the <metadata> element must be preceded by<br>ctx:requester/ctx:metadata-by-al/ctx:format<br>with value<br>"http://dini.de/namespace/oas-requesterinfo" |
| Usage | Optional |
| Format | Three values are allowed:<br>• "internal": classification for technical, system-internal accesses. Examples would be automated availability and consistency checks, cron jobs, keep-alive queries etc.<br>• "administrative": classification for accesses that are being made due to human decision but are for administrative reasons only. Examples would be manual quality assurance, manual check for failures, test runs etc.<br>• "institutional": classifies accesses that are made from within the institution running the service in question, regardless whether they are for administrative reasons. |
| Example | institutional |

### <requester/.../dini:hashed-session> | Hashed session of the requester

| | |
|---|---|
| Description | The identifier of the complete usage session of a given user. |
| XPath | ctx:context-object/ctx:requester/ctx:metadata/dini:requesterinfo/dini:hashed-session<br>If this element is used, the <metadata> element must be preceded by<br>ctx:requester/ctx:metadata-by-al/ctx:format<br>with value<br>"http://dini.de/namespace/oas-requesterinfo" |
| Usage | Optional |
| Format | If the session ID is a hash itself, it must be hashed. Otherwise, provide a MD5 hash of the session ID. |
| Example | 660b14056f5346d0 |

### <requester/.../dini:user-agent> | Full user agent string of the requester

| | |
|---|---|
| Description | The full HTTP user agent string |
| XPath | ctx:context-object/ctx:requester/ctx:metadata/dini:requesterinfo/dini:classification/dini:user-agent<br>If this element is used, the <metadata> element must be preceded by<br>ctx:requester/ctx:metadata-by-al/ctx:format<br>with value<br>"http://dini.de/namespace/oas-requesterinfo" |
| Usage | Optional |
| Format | String |
| Example | Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.0.6) Gecko/2009011913 Firefox/3.0.6 (.NET CLR 3.5.30729) |

# 4. Transfer Protocols

## 4.1. OAI-PMH

The data exchange between a data provider and a log aggregator may be based on the widely established OAI Protocol for Metadata Harvesting ( OAI-PMH). If this protocol is used, it must be its version 2.0.  In principle, OAI-PMH specifies a data synchronisation mechanism which supports a reliable implementation of one-way data synchronisation. This functionality also fits well for the purpose of usage data transfer. Since OAI-PMH was originally designed for the exchange of bibliographic metadata, this section will specify how OAI-PMH can be used to transfer usage data.

The general procedure is that local repositories expose the entries from their log files as OpenURL Context Objects, and that they make these available for harvesting by the log aggregator (see *figure 3*).

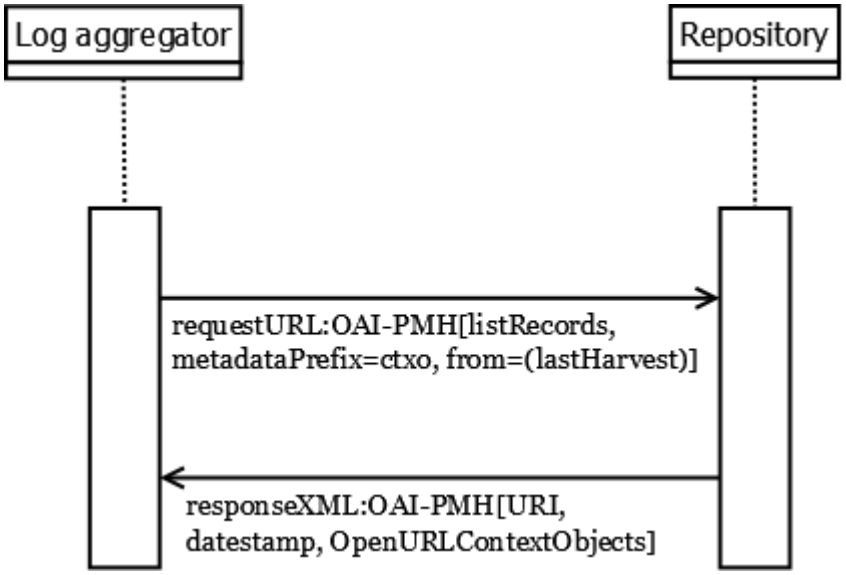



*Figure 3.*

The document-centric approach of OAI-PMH results in the following central problems when applied to usage data:

## 4.1.1. Requirement for metadata record identifiers

(see OAI-PMH, 2.4 )

Data providers must issue identifiers for data records to formally comply with OAI-PMH. THE OAI identifiers should adhere to the OAI Identifier Format, as descibed in the OAI-PMH guidelines. These identifiers are not used by the log aggregator.

## 4.1.2. Datestamps for records

(see OAI-PMH, 2.7.1)

OAI-PMH requires datestamps for all records of provided data. the datestamp within the OAI-PMH record header must be the time at which the Context Object or the Context Objects container has been stored in the database which feeds the OAI-PMH interface.

This information has to be kept separately from the datestamp of the usage event itself. This latter datestamp is the time at which the actual usage event took place.

The OAI-PMH specification allows for either exact-to-the-second or exact-to-the-day granularity for record header datestamps. The data providers may chose one of these possibilities. The service provider will most certainly rely on overlapping harvesting, i. e. the most recent datestamp of the harvested data is used as the "from" parameter for the next OAI-PMH query. Thus, the data provider will provide some records that have been harvested before. Duplicate records are matched by their identifiers (those in the OAI-PMH record header) and are silently tossed if their datestamp is not renewed (see notes below on deletion tracking).It is strongly recommended to implement exact-to-the-second datestamps to keep redundancy of the transferred data as low as possible.

## 4.1.3. MetadataPrefix

A KE-compliant OAI-PMH interface must support the "ctxo" metadataPrefix. In responmse to each OAI-PMH request that specifies the "ctxo" prefix, it must return KE-compliant context objects.

## 4.1.4. Mandated metadata in Dublin Core (DC) format

OAI-PMH repositories must be able to provide records with metadata expressed in Dublin Core. As a minimum, a rudimentary DC data set (identifier and description) should be provided which should describe the data offered and linked to by a certain identifier (see above regarding the identifier discussion). For creating a DC data set, follow the DRIVER guidelines. Example Warning: the XML excerpts given in these guidelines as illustrations do not necessarily contain all details regarding XML namespaces and XML schema. Nevertheless this omitted information is to be included in actual implementations and must not be considered optional.

**OAI-PMH listRecords metadataPrefix=oai_dc**

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH>
...
<record>
    <header> ... (compare notes about the record header)</header>
    <metadata>
        <dc xmlns="http://www.openarchives.org/OAI/2.0/oai_dc/"
            xmlns:dc="http://purl.org/dc/elements/1.1/">
            <identifier>ID2</identifier>
            <description> Usage Event Data for Server ... from ... until ... </description>
        </dc>
    </metadata>
</record>
...
</OAI-PMH>
```

## 4.1.5. Usage of Sets

(see OAI-PMH, 2.7.2 )

OAI-PMH optionally allows for structuring the offered data in "sets" to support selective harvesting of the data. Currently, this possibility is not further specified in these guidelines. Future refinements may use this feature, e. g. for selecting usage data for certain services. Provenance information is already included in the Context Objects.

## 4.1.6. Deletion tracking

(see OAI-PMH, 2.5.1)

The OAI-PMH provides functionalities for the tracking of deletion of records. Compared to the classic use case of OAI-PMH (descriptive metadata of documents) the use case presented here falls into a category of data which is not subject to long-term storage. Thus, the tracking of deletion events does not seem critical since the data tracking deletions would summarize to a significant amount of data.However, the service provider will accept information about deleted records and will eventually delete the referenced information in its own data store. This way it is possible for data providers to do corrections (e. g. in case of technical problems) on wrongly issued data.It is important to note that old data which rotates out of the data offered by the data provider due to its age will not to be marked as deleted for storage reasons. This kind of data is still valid usage data, but not visible anymore.The information about whether a data provider uses deletion tracking has to be provided in the response to the "identify" OAI-PMH query within the <deletedRecords> field. Currently, the only options are "transient" (when a data provider applies or reserves the possibility for marking deleted records) or "no".The possible cases are:

- Incorrect data which has already been offered by the data provider shall be corrected. There are two possibilities:
    - Re-issuing of a corrected set of data carrying the same identifier in the OAI-PMH record header as the set of data to be corrected, with an updated OAI-PMH record header datestamp
    - When the correction is a full deletion of the incorrect issued data, the OAI-PMH record has to be re-issued without a Context Object payload, with specified "<deleted>" flag and updated datestamp in the OAI-PMH record header.
- Records that fall out of the time frame for which the data provider offers data: These records are silently neglected, i. e. not offered via the OAI-PMH interface anymore, without using the deletion tracking features of OAI-PMH.

## 4.1.7. Metadata formats

(see OAI-PMH, 3.4)

All data providers have to provide support for <context-objects> aggregations. While a specific "metadataPrefix" is not required, the information about "metadataNamespace" and "schema" is fixed for implementations:

**OAI-PMH listMetadataFormats**

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH>
...
<metadataFormat>
  <metadataPrefix>ctxo</metadataPrefix>
  <schema>http://www.openurl.info/registry/docs/xsd/info:ofi/fmt:xml:xsd:ctx</schema>
  <metadataNamespace>info:ofi/fmt:xml:xsd:ctx</metadataNamespace>
</metadataFormat>
...
</OAI-PMH>
```

### 4.1.8. Inclusion of Context Objects in OAI-PMH records

Corresponding to the definition of XML encoded Context Objects as data format of the data exchanged via the OAI-PMH, the embedding is to be done conforming to the OAI-PMH:

**method 1 : all Context Objects in one OAI-PMH record : OAI-PMH listRecords metadataPrefix=ctxo**

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH>
...
<record>
    <header>
        <identifier>urn:uuid:e5d037a0-633c-11df-a08a-0800200c9a66</identifier>
        <datestamp>2009-06-02T14:10:02Z</datestamp>
    </header>
    <metadata>
        <context-objects xmlns="info:ofi/fmt:xml:xsd:ctx">
            <context-object datestamp="2009-06-01T19:20:57Z">
              ...
            </context-object>
            <context-object datestamp="2009-06-01T19:21:07Z">
              ...
            </context-object>
        </context-objects>
    </metadata>
</record>
...
</OAI-PMH>
```

## 4.2. SUSHI

OAI-PMH is a relatively light-weight protocol which does not allow for a bidirectional traffic. If a more reliable error-handling is required, the *Standardised Usage Statistics Harvesting Initiative* (SUSHI) must be used. SUSHI http://www.niso.org/schemas/sushi/ was developed by NISO (National Information Standards Organization) in cooperation with COUNTER. This document assumes that the communication between the aggregator and the usage data provider takes place as is explained in figure 4.

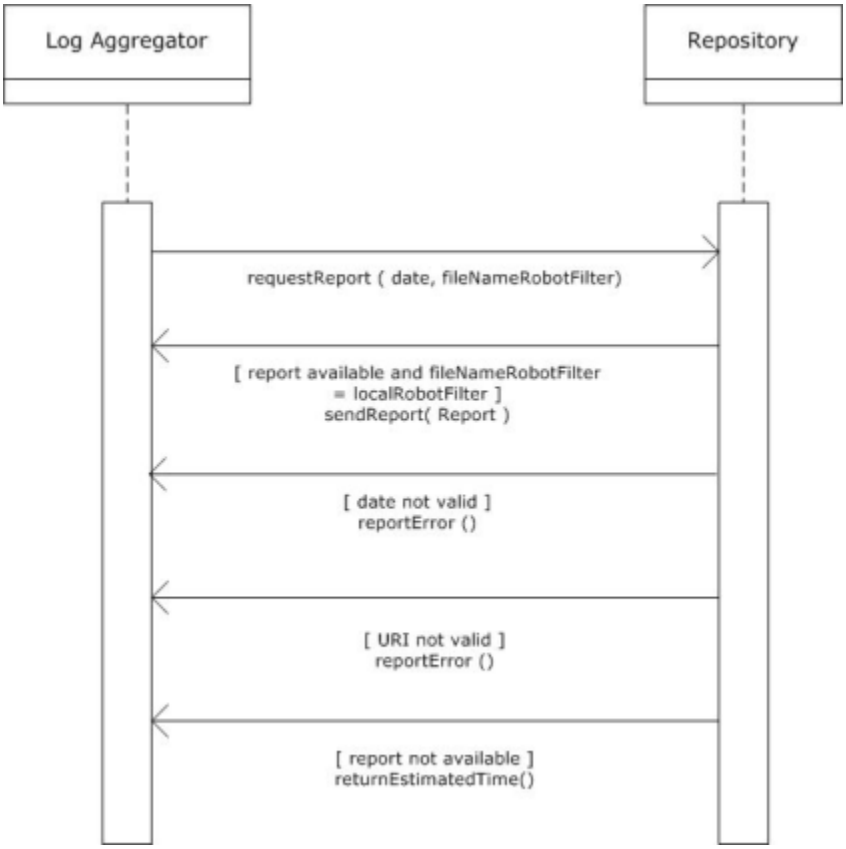



*Figure 4.*

The interaction commences when the log aggregator sends a request for a report about the daily usage of a certain repository. Two parameters must be sent as part of this request: (1) the date of the report and (2) the file name of the most recent robot filter. The filename that is mentioned in this request will be compared to the local filename. Four possible responses can be returned by the repository.

- If the filename that is mentioned in the request exactly matches the filename that is maintained locally, and if a report for the requested data is indeed available, this report will be returned immediately.
- In this protocol, only daily reports will be allowed. This was decided mainly to restrict the size of the data traffic between the servers. If a request is sent for a period that exceeds one day, an error message will be sent indicating that the date parameter is incorrect.
- If the URI of the robot filter file, for some reason, cannot be resolved to an actual file, an error message will be sent about this.
- If the parameters are correct, but if the report is not yet available, a message will be sent which provides an estimation of the time of arrival.

In SUSHI version 1.0., the following information must be sent along with each request:

- Requestor ID
- Name of requestor
- E-mail of requestor
- CustomerReference ID (may be identical to the Requestor ID)
- Name of the report that is requested
- Version number of the report
- Start and end date of the report

This request will active a special tool that can inspect the server logging and that can return the requested data. These data are transferred as OpenURL Context Object log entries, as part of a SUSHI response.
The reponse must repeat all the information from the request, and provide the requested report as XML payload
The usage data are subsequently stored in a central database. External parties can obtain information about the contents of this central database through specially developed web services. The log harvester must ultimately expose these data in the form of COUNTER-compliant reports.
Listing 1 is an example of a SUSHI request, sent from the log aggregator to a repository.

**Listing 2**

```
<?xml version="1.0" encoding="UTF-8"?>
<soap:Envelope xmlns:soap="http://schemas.xmlsoap.org/soap/envelope/"
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:schemaLocation="http://schemas.xmlsoap.org/soap/envelope/ http://schemas.xmlsoap.org/soap/envelope/">
 <soap:Body>
  <ReportRequest
   xmlns:ctr="http://www.niso.org/schemas/sushi/counter"
   xsi:schemaLocation="http://www.niso.org/schemas/sushi/counter
   http://www.niso.org/schemas/sushi/counter_sushi3_0.xsd"
   xmlns="http://www.niso.org/schemas/sushi"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" >
   <Requestor>
    <ID>www.logaggregator.nl</ID>
    <Name>Log Aggregator</Name>
    <Email>logaggregator@surf.nl</Email>
   </Requestor>
   <CustomerReference>
    <ID>www.leiden.edu</ID>
    <Name>Leiden University</Name>
   </CustomerReference>
   <ReportDefinition Release="urn:robots-v1.xml" Name="Daily Report v1">
    <Filters>
     <UsageDateRange>
      <Begin>2009-12-21</Begin>
      <End>2009-12-22</End>
     </UsageDateRange>
    </Filters>
   </ReportDefinition>
  </ReportRequest>
 </soap:Body>
</soap:Envelope>
```

Note that the intent of the SUSHI request above is to see all the usage events that have occurred on 21 December 2009. The SUSHI schema was originally developed for the exhchange of COUNTER-compliant reports. In the documentation of the SUSHI XML schema, it is explained that COUNTER usage is only reported at the month level. In SURE, only daily reports can be provided. Therefore, it will be assumed that the implied time on the date that is mentioned is 0:00. The request in the example that is given thus involves all the usage events that have occurred in between 2009-12-21T00:00:00 and 2002-12-22T00:00:00.
As explained previously, the repository can respond in four different ways. If the parameters of the request are valid, and if the requested report is available, the OpenURL ContextObjects will be sent immediately. The Open URL Context Objects will be wrapped into element <Report>, as can be seen in listing 2.

**Listing 3**

```
<?xml version="1.0" encoding="UTF-8"?>
<soap:Envelope xmlns:soap="http://schemas.xmlsoap.org/soap/envelope/"
                                xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                                xsi:schemaLocation="http://schemas.xmlsoap.org/soap/envelope/ http://schemas.
xmlsoap.org/soap/envelope/">
        <soap:Body>
                <ReportResponse xmlns:ctr="http://www.niso.org/schemas/sushi/counter"
                                                xsi:schemaLocation="http://www.niso.org/schemas/sushi
/counter http://www.niso.org/schemas/sushi/counter_sushi3_0.xsd"
                                                xmlns="http://www.niso.org/schemas/sushi"
                                                xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" >
                        <Requestor>
                                <ID>www.logaggregator.nl</ID>
                                <Name>Log Aggregator</Name>
                                <Email>logaggregator@surf.nl</Email>
                        </Requestor>
                        <CustomerReference>
                                <ID>www.leiden.edu</ID>
                                <Name>Leiden University</Name>
                        </CustomerReference>
                        <ReportDefinition Release="urn:DRv1" Name="Daily Report v1">
                                <Filters>
                                        <UsageDateRange>
                                                <Begin>2009-12-22</Begin>
                                                <End>2009-12-23</End>
                                        </UsageDateRange>
                                </Filters>
                        </ReportDefinition>
                        <Exception>
                                <Number>1</Number>
                                <Message>The range of dates that was provided is not valid. Only daily
reports are
                                available.</Message>
                        </Exception>
                </ReportResponse>
        </soap:Body>
</soap:Envelope>
```

If the begin date and the end date in the request of the log aggregator form a period that exceeds one day, an error message must be sent. In the SUSHI schema, such messages may be sent in an <Exception> element. Three types of errors can be distinguished. Each error type is given its own number. An human-readable error message is provided under <Message>.

**Listing 4**

```
<?xml version="1.0" encoding="UTF-8"?>
<soap:Envelope xmlns:soap="http://schemas.xmlsoap.org/soap/envelope/"
                                xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                                xsi:schemaLocation="http://schemas.xmlsoap.org/soap/envelope/ http://schemas.
xmlsoap.org/soap/envelope/">
        <soap:Body>
                <ReportResponse xmlns:ctr="http://www.niso.org/schemas/sushi/counter"
                                                xsi:schemaLocation="http://www.niso.org/schemas/sushi
/counter http://www.niso.org/schemas/sushi/counter_sushi3_0.xsd"
                                                xmlns="http://www.niso.org/schemas/sushi"
                                                xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" >
                        <Requestor>
                                <ID>www.logaggregator.nl</ID>
                                <Name>Log Aggregator</Name>
                                <Email>logaggregator@surf.nl</Email>
                        </Requestor>
                        <CustomerReference>
                                <ID>www.leiden.edu</ID>
                                <Name>Leiden University</Name>
                        </CustomerReference>
                        <ReportDefinition Release="urn:DRv1" Name="Daily Report v1">
                                <Filters>
                                        <UsageDateRange>
                                                <Begin>2009-12-22</Begin>
                                                <End>2009-12-23</End>
                                        </UsageDateRange>
                                </Filters>
                        </ReportDefinition>
                        <Exception>
                                <Number>1</Number>
                                <Message>The range of dates that was provided is not valid. Only daily
reports are
                                available.</Message>
                        </Exception>
                </ReportResponse>
        </soap:Body>
</soap:Envelope>
```

A second type of error may be caused by the fact that the file that is mentioned in the request can not be accessed. In this situation, the response will look as follows:

**Listing 5**

```
<?xml version="1.0" encoding="UTF-8"?>
<soap:Envelope xmlns:soap="http://schemas.xmlsoap.org/soap/envelope/"
                                xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                                xsi:schemaLocation="http://schemas.xmlsoap.org/soap/envelope/ http://schemas.
xmlsoap.org/soap/envelope/">
        <soap:Body>
                <ReportResponse xmlns:ctr="http://www.niso.org/schemas/sushi/counter"
                                                xsi:schemaLocation="http://www.niso.org/schemas/sushi
/counter http://www.niso.org/schemas/sushi/counter_sushi3_0.xsd"
                                                xmlns="http://www.niso.org/schemas/sushi"
                                                xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" >
                        <Requestor>
                                <ID>www.logaggregator.nl</ID>
                                <Name>Log Aggregator</Name>
                                <Email>logaggregator@surf.nl</Email>
                        </Requestor>
                        <CustomerReference>
                                <ID>www.leiden.edu</ID>
                                <Name>Leiden University</Name>
                        </CustomerReference>
                        <ReportDefinition Release="urn:DRv1" Name="Daily Report v1">
                                <Filters>
                                        <UsageDateRange>
                                                <Begin>2009-12-22</Begin>
                                                <End>2009-12-23</End>
                                        </UsageDateRange>
                                </Filters>
                        </ReportDefinition>
                        <Exception>
                                <Number>2</Number>
                                <Message>The file describing the internet robots is not accessible.</Message>
                        </Exception>
                </ReportResponse>
        </soap:Body>
</soap:Envelope>
```

When the repository is in the course of producing the requested report, a response will be sent that is very similar to listing 5. The estimated time of completion will be provided in the <Data> element. According to the documentation of the SUSHI XML schema, this element may be used for any other optional data.

**Listing 6**

```
<?xml version="1.0" encoding="UTF-8"?>
<soap:Envelope xmlns:soap="http://schemas.xmlsoap.org/soap/envelope/"
                                xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                                xsi:schemaLocation="http://schemas.xmlsoap.org/soap/envelope/ http://schemas.
xmlsoap.org/soap/envelope/">
        <soap:Body>
                <ReportResponse xmlns:ctr="http://www.niso.org/schemas/sushi/counter"
                                                xsi:schemaLocation="http://www.niso.org/schemas/sushi
/counter http://www.niso.org/schemas/sushi/counter_sushi3_0.xsd"
                                                xmlns="http://www.niso.org/schemas/sushi"
                                                xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" >
                        <Requestor>
                                <ID>www.logaggregator.nl</ID>
                                <Name>Log Aggregator</Name>
                                <Email>logaggregator@surf.nl</Email>
                        </Requestor>
                        <CustomerReference>
                                <ID>www.leiden.edu</ID>
                                <Name>Leiden University</Name>
                        </CustomerReference>
                        <ReportDefinition Release="urn:DRv1" Name="Daily Report v1">
                                <Filters>
                                        <UsageDateRange>
                                                <Begin>2009-12-22</Begin>
                                                <End>2009-12-23</End>
                                        </UsageDateRange>
                                </Filters>
                        </ReportDefinition>
                        <Exception>
                                <Number>3</Number>
                                <Message>The report is not yet available. The estimated time of completion is
                                provided under "Data".</Message>
                                <Data>2010-01-08T12:13:00+01:00</Data>
                        </Exception>
                </ReportResponse>
        </soap:Body>
</soap:Envelope>
```

Error numbers and the corresponding Error messages are also provided in the table below.

| *Error number* | *Error message* |
|---|---|
| 1 | The range of dates that was provided is not valid. Only daily reports are available. |
| 2 | The file describing the internet robots is not accessible |
| 3 | The report is not yet available. The estimated time of completion is provided under "Data" |

# 5. Normalisation

## 5.1. Double Clicks

If a single user clicks repeatedly on the same item within a given amount of time, this should be counted as a single request. This measure is needed to minimise the impact of conscious falsification by authors. There appears to be some difference as regards the time-frame of double clicks. The table below provides an overview of the various timeframes that have been suggested.

| | |
|---|---|
| COUNTER | 10 sec for a HTML-resource; 30 sec for a PDF |
| LogEC | 1 month |
| AWStats | 1 hour |

| | |
|---|---|
| IFABC | 30 minutes |

Individual usage data providers should not filter doubles clicks. This form of normalisation should be carried out on a central level by the aggregator.

By default the KE guidelines follow the COUNTER rules, in order to deliver statistics that can be compared to those of publishers.

# 5.2. Robot filtering

## 5.2.1. Definition of a robot

The "user" as defined in section 2 of this report is assumed to be a human user. Consequently, the focus of this document is on requests which have consciously been initiated by human beings. Automated visits by internet robots must be filtered from the data as much as possible.

**Definition of a "robot" according to robotstxt.org**

A robot is a program that automatically traverses the Web's hypertext structure by retrieving a document, and recursively retrieving all documents that are referenced.
- http://www.robotstxt.org/faq/what.html, also used as definition by (Geens, 2006), (Heinonen, 1996)

## 5.2.2. Strategy

It is decided to make a distinction between two 'layers' of robot filtering (see also figure 5):

1. Local repositories should make use of a "core" list of robots. It was agreed that a list can probably be created quite easily by combining entries from the lists that are used by COUNTER, AWStats, Universidade do Minho and PLoS. This basic list will filter about 80% of all automated visits.
2. Dedicated service providers can carry out some more advanced filtering, on the basis of sophisticated algorithms. The specification of these more advanced heuristics will be a separate research activity. Centralised heuristics should improve confidence in the reliability of the statistics.

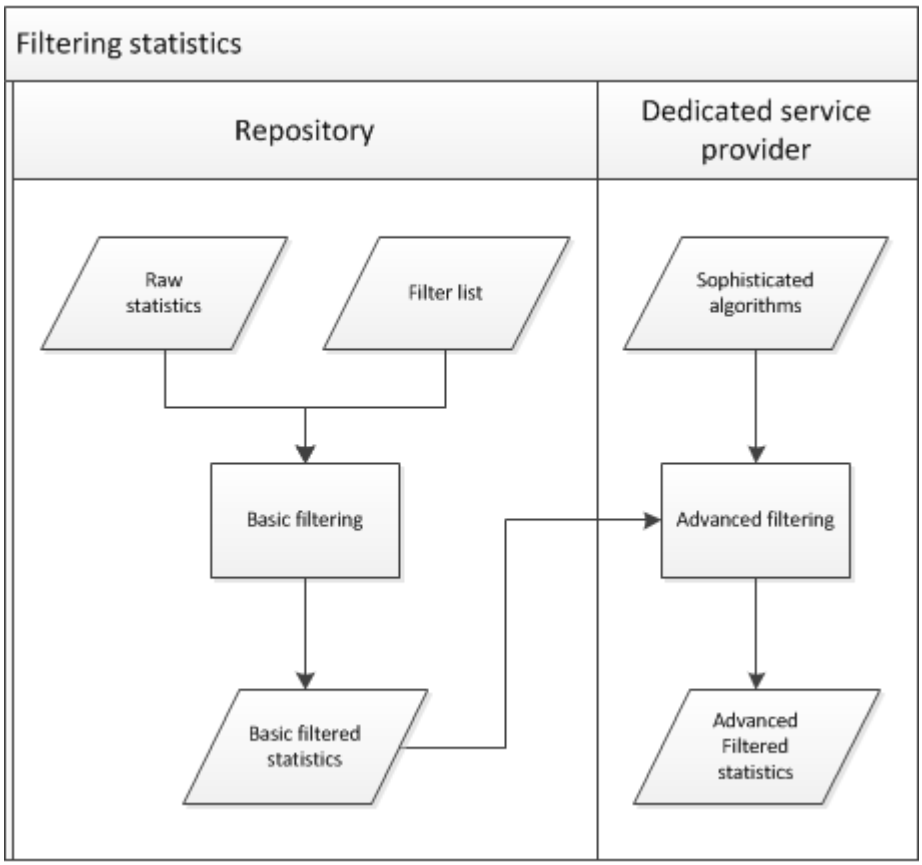



*Figure 5.*

Internet robots will be identified by comparing the value of the User Agent HTTP header to regular expressions which are present in a list of known robots which is managed by a central authority. All entries are expected to conform to the definition of a robot as provided in section 5.2.1. All institutions that send usage data must first check the entry against this list of internet robots. If the robot is in the list, the event should not be sent. It has been decided not to filter robots on their IP-addresses. The reason for this is that IP-addresses change very regularly, and this would make the list very difficult to maintain.

In the study of (Geens, 2006), using the user agent field in the log file resulted in a recall of merely 26.56%, with a precision of 100%.

As an alternative, identifying robots by analyzing the following 4 components resulted in a recall of 73%, with a precision of 100%:

- identifying who accessed robots.txt, which are usually not normal users but bots
- identifying with an IP-address list of known bots
- identifying with the user agent field
- identifying who accesses pages with a HEAD method, where normal users perform the GET method

More alternatives are given the report, perhaps an interesting read

- Max Kemman

## 5.2.3. Robot list schema

The robot list must meet the following requirements:

1. It must be possible to 'timestamp' the list so that agents can refer to specific versions.
2. The list must be in a machine-readable format, and preferably in XML. The list that is currently maintained by COUNTER is a word file.
3. The extended list which is created by KE partners must be approved by COUNTER. Institutions that make use of the extended list should also be able to pass the COUNTER audit.
4. It must be possible to indicate the 'source' of each entry in the list (e.g. "COUNTER", "AwStats", "Plos", etc.)
5. It must be possible to access the robot list on the basis of a persistent URI.
6. It must be possible to manage different versions of the robot list. The most recent version must always be available through a uniform URL.

To implement requirement 5, the following mechanism will be implemented:

- The current version of the list can be reached by placing /current/ in the local path of the URI, e.g.: http://purl.org/robotslist/current/robotlist.xml

# 6. Legal boundaries

## 6.1. Usage of IP addresses and the protection of a 'natural person'

The IP address of the requester is pseudonymised using encryptions, before it is exchanged and taken outside the web-server to another location. Therefore individual users can be recognised when aggregated from distributed repositories, but cannot be referred back to a 'natural person'. This method may seem consisted with the European Act for Protection of Personal data. The summary can be found here: ?http://europa.eu/legislation_summaries/information_society/l14012_en.htm. Further legal research needs to be done if this method is sufficient to protect the personal data of a 'natural person', in order to operate within the boundaries of the law.

> In these guidelines the IP addresses are pseudonymized using a Salted MD5 hash encryption.

# Appendices

**Appendix A: Sample OpenURL Context Object File**


```
<?xml version="1.0" encoding="UTF-8"?>
<context-objects xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xmlns:dcterms="http://dublincore.org/documents/2008/01/14/dcmi-terms/"
        xmlns:sv="info:ofi/fmt:xml:xsd:sch_svc"
        xsi:schemaLocation="info:ofi/fmt:xml:xsd:ctx http://www.openurl.info/registry/docs/info:ofi/fmt:xml:
xsd:ctx"
        xmlns="info:ofi/fmt:xml:xsd:ctx">
    <context-object timestamp="2009-07-29T08:15:46+01:00" identifier="b06c0444f37249a0a8f748d3b823ef2a">

        <referent>
            <identifier>https://openaccess.leidenuniv.nl/bitstream/1887/12100/1/Thesis.pdf</identifier>
            <identifier>http://hdl.handle.net/1887/12100</identifier>
        </referent>

        <referring-entity>
            <identifier>http://www.google.nl/search?hl=nl&amp;q=beleidsregels+artikel+4%3A84&amp;meta="<
/identifier>
            <identifier>info:sid/google</identifier>
        </referring-entity>

        <requester>
            <identifier> data:,b505e629c508bdcfbf2a774df596123dd001cee172dae5519660b6014056f53a</identifier>

            <metadata-by-val>
                <format>http://dini.de/namespace/oas-requesterinfo</format>
                <metadata>
                    <requesterinfo xmlns="http://dini.de/namespace/oas-requesterinfo">
                        <hashed-ip>data:,b505e629c508bdcfbf2a774df596123dd001cee172dae5519660b6014056f53a<
/hashed-ip>
                        <hashed-c>data:,d001cee172dae5519660b6014056f5346d05e629c508bdcfbf2a774df596123d<
/hashed-c>
                        <hostname>uni-saarland.de</hostname>
                        <classification>institutional</classification>
                        <hashed-session>660b14056f5346d0</hashed-session>
                        <user-agent>mozilla/5.0 (windows; u; windows nt 5.1; de; rv:1.8.1.1) gecko/20061204<
/user-agent>
                    </requesterinfo>
                </metadata>
            </metadata-by-val>
        </requester>

        <service-type>
            <metadata-by-val>
                <format>http://dublincore.org/documents/2008/01/14/dcmi-terms/</format>
                <metadata>
                    <dcterms:format>info:eu-repo/semantics/objectFile</dcterms:format>
                </metadata>
            </metadata-by-val>
        </service-type>

        <resolver>
            <identifier>http://www.worldcat.org/libraries/53238</identifier>
        </resolver>

        <referrer>
            <identifier>info:sid/dlib.org:dlib</identifier>
        </referrer>

    </context-object>
</context-objects>
```

**Appendix B: Schema for Robot filter List**

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" elementFormDefault="qualified">

        <xs:element name="exclusions">
                <xs:complexType>
                        <xs:sequence>
                                <xs:element ref="sources"/>
                                <xs:element ref="robot-list"/>
                        </xs:sequence>
                        <xs:attributeGroup ref="attlist.exclusions"/>
                </xs:complexType>
        </xs:element>

        <xs:attributeGroup name="attlist.exclusions">
                <xs:attribute name="version" type="xs:string"/>
                <xs:attribute name="datestamp" type="xs:date"/>
        </xs:attributeGroup>

        <xs:element name="sources">
                <xs:complexType>
                        <xs:sequence>
                                <xs:element ref="source" minOccurs="0" maxOccurs="unbounded"/>
                        </xs:sequence>
                </xs:complexType>
        </xs:element>

        <xs:element name="source">
                <xs:complexType>
                        <xs:simpleContent>
                                <xs:extension base="xs:string">
                                        <xs:attribute name="id" type="xs:ID" use="required"/>
                                        <xs:attribute name="name" type="xs:string"/>
                                        <xs:attribute name="version" type="xs:string"/>
                                        <xs:attribute name="datestamp" type="xs:date"/>
                                </xs:extension>
                        </xs:simpleContent>
                </xs:complexType>
        </xs:element>

        <xs:element name="sourceRef">
                <xs:complexType>
                        <xs:simpleContent>
                                <xs:extension base="xs:string">
                                        <xs:attribute name="id" type="xs:IDREF" use="required"/>
                                </xs:extension>
                        </xs:simpleContent>
                </xs:complexType>
        </xs:element>

        <xs:element name="robot-list">
                <xs:complexType>
                        <xs:sequence>
                                <xs:element ref="useragent" minOccurs="0" maxOccurs="unbounded"/>
                        </xs:sequence>
                </xs:complexType>
        </xs:element>

        <xs:element name="useragent">
                <xs:complexType>
                        <xs:sequence>
                                <xs:element ref="regEx"/>
                                <xs:element ref="sourceRef" minOccurs="0" maxOccurs="unbounded"/>
                        </xs:sequence>
                </xs:complexType>
        </xs:element>

        <xs:element name="regEx" type="xs:string"/>

</xs:schema>
```

**Appendix C: Sample Robot filter list**

```
<?xml version="1.0" encoding="UTF-8"?>

<exclusions xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                        xsi:noNamespaceSchemaLocation="robotlist.xsd"
                        version="1.0"
                        datestamp="2010-04-10">

        <sources>
                <source id="l1" name="COUNTER" version="R3" datestamp="2010-04-01">COUNTER list of internet
robotos</source>
                <source id="l2" name="PLOS">PLOS list of internet robotos</source>
        </sources>

        <robot-list>
                <useragent>
                        <regEx>[^a]fish</regEx>
                        <sourceRef id="l2"/>
                </useragent>
                <useragent>
                        <regEx>[+:,\.\;\/-]bot</regEx>
                        <sourceRef id="l2"/>
                </useragent>
                <useragent>
                        <regEx>acme\.spider</regEx>
                        <sourceRef id="l2"/>
                </useragent>
                <useragent>
                        <regEx>Brutus\/AET</regEx>
                        <sourceRef id="l1"/>
                        <sourceRef id="l2"/>
                </useragent>
                <useragent>
                        <regEx>Code\sSample\sWeb\sClient</regEx>
                        <sourceRef id="l1"/>
                </useragent>

        </robot-list>

</exclusions>
```

https://www.surfgroepen.nl/sites/surfshare/ArticleWorkspace/Gedeelde%20documenten/OpenRepositories2010/posters/OR2010-PosterPresentation-DRAFT-Linking%20institutional%20repositories%20to%20the%20Open%20Data%20Web-Thomas%20Place,%20Maurice%20Vanderfeesten.pdf