OpenURL Context Objects

1. Aspects to be recorded

A usage event takes place when a **user** downloads a **document** which is managed in an **institutional repository**, or when a user view the **meta data** that is associated with this document, see also figure 1.

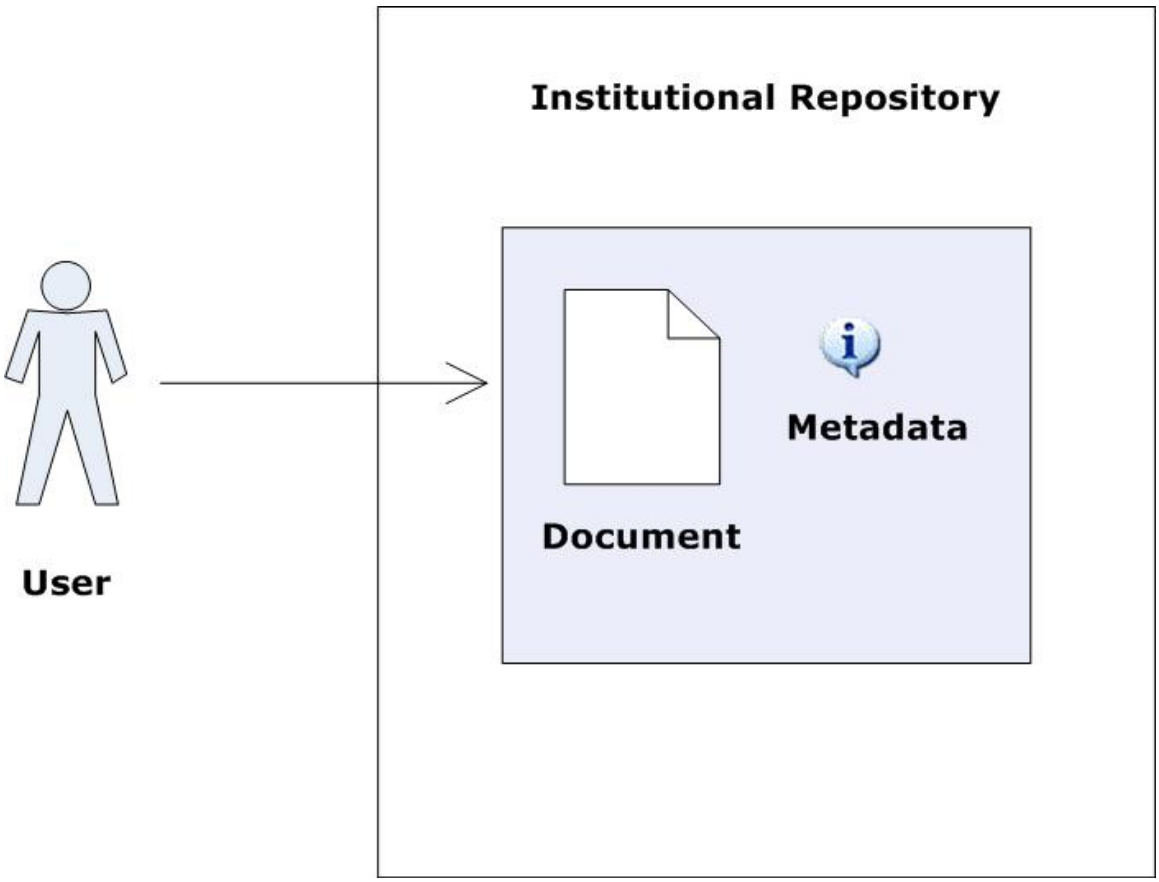

Figure 1.

The user may have arrived at this document through the mediation of a **referrer**. This is typically a search engine, such as Google or Yahoo.

The aim of the SURFShare SURE project is to describe the highlighted entities that are involved in the usage event as fully as possible. The event took place in a certain country, at a certain time on a certain date, and from a machine which a certain IP-address. In the *JISC Usage Statistics Review*,[1|http://www.surffoundation.nl/#_ftn1] which was released in September 2008, it is stipulated that the following items of information must minimally be captured:

- Who (Identification of user/session)
- What (Item identification)
- Type of request performed (e.g. full-text, front-page, including failed/partially fulfilled requests)
- When (Date and time)
- Usage event ID

In line with these JISC recommendations, it has been decided in the SURF SURE project to provide the following data elements:

Data element	*Description*
IP-address of requestor	Providing the full IP-address is not permitted by international copyright laws. For this reason, the IP-address needs to be encrypted.
C-class Subnet	When the IP-address is encrypted, this will have the disadvantage that information on the geographic location, for instance, can no longer be derived. For this reason, the Class subnet, being the first three most significant bits from the IP-address must also be provided.
Geographic location	The country from which the request originated is also provided explicitly.
Persistent identifier of requested document	See also section 4.
URL of document	See also section 4.
Date and time of the request	
Request Type	It must be clear if a document was downloaded or if its metadata was viewed.
Host name	The institution that is responsible for the repository in which the requested document is stored.
Usage event ID	Unique number for a specific usage event.
Referrer URL	The URL which was received from the referring entity, if it was used. This URL often contains the search terms that were typed in by the user.
Referrer name	A classification of the referrer, based on a short list of search engines.

2. Source of information

All Dutch repositories make use of Apache server software for the maintenance of their repository websites. Each usage event that takes place generates an entry in the Apache logging files. These logging files will be used in the SURF SURE project as the primary source of information for usage statistics.

Figure 2 below contains a typical entry from an Apache log file.

```
13/Jul/2009:09:14:16 +0200] 193.173.52.133 TLSv1 RC4-MD5 openaccess.leidenuniv.nl "GET /bitstream/1887/3674/1/360_138.pdf HTTP/1.1" 722168 - "http://www.google.nl/search?hl=nl&q=beleidsregels+artikel+4%3A84&meta=" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727)" 200 20352 15
```

Figure 2.

From figure 2, it can be seen that the aspects which were mentioned at the end of section 2 can normally be derived relatively easy from the log file.

3. Formatting guidelines

Different institutions may also have configured the logging facilities of their servers in different ways. Because of this, there may occasionally be variations in way in which, for instance, the time and the date are formatted. To avoid problems arising from variations in formatting, this section provides guidelines on the format in which the mandatory and optional data element need to be provided. A general principle, however, is that information should be passed along to the central database as 'pure' as possible, so that analysis can take place centrally and consistently.

IP-address of requestor	
Description	The IP-address must be hashed using MD5 encryption.
Usage	Mandatory
Format	32 hexadecimal numbers.
Example	c06f0464f37249a0a9f848d4b823ef2a

C-class Subnet	
Description	The first three bytes of an IP-address, which are used to designate the network ID. It is similar to the IP-address, with the crucial difference that the final (most significant) byte, which designates the HOST ID is replaced with a '0'.
Usage	Mandatory
Format	Three decimal numbers separated by a dot, followed by a dot and a '0'.
Example	118.94.150.0

Geographic location		
Description	The country from which the request was	
Usage	Mandatory	
Format	A two-letter code in lower case, following the ISO 3166-1-alpha-2 code.[2	http://www.surffoundation.nl /#_ftn2\ ]
Example	ne	

URL of document	
Description	The document identifier provides a globally unique identification of the resource that is requested.
Usage	Mandatory
Format	The identifier must be given in the form of a URL. At the level of the service aggregator, this URL must be connected to the object's URN, so that the associated metadata can be obtained. The URN can be found by retrieving the URL of the object file in the DIDL file that is maintained by KNAW.
Example	/bitstream/1887/12100/1/Thesis.pdf

OAI-PMH identifier

Description	The URN of the publication. This identifier must be the same as the identifier in the DIDL representation of the publication.
Usage	Optional
Format	URN
Example	http://hdl.handle.net/1887/12100

Request Type	
Description	The request type specifies whether an object file was downloaded or whether a metadata record was viewed.
Inclusion	Mandatory
Format	Two values are allowed: "objectFile" or "metadataView"
Example	ObjectFile

Host name	
Description	An identification of the repository that has recorded the usage event
Usage	Mandatory
Format	URL of the receiving institution's repository
Example	www.openaccess.leidenuniv.nl

Request Time	
Description	The exact time on which the usage event took place.
Usage	Mandatory
Format	The request time must be given in a format that that conforms to ISO8601. The YYYY-MM-DDTHH:MM:SSZ representation must be used. Note that this format may differ from the format that is provided in the Apache log file.
Example	2009-07-29T08:15:46+01:00

Referrer URL	
Description	The environment which has directed the user to the requested object. This usually refers to the search engine which the client has used to find the object.
Usage	Optional
Format	URL
Example	http://www.google.nl/search?hl=nl&q=beleidsregels+artikel+4%3A84&meta="

Referrer Name	
Description	The referrer must be categorised on the basis of a limited list of known referrers.
Usage	Optional

Format	The following values are allowed: "google", "google scholar", "bing", "yahoo", "altavista"
Example	google

Usage Event ID	
Description	Unique identification of the usage event. This identification will be generated, and it can not be derived from the Apache log file.
Usage	Mandatory
Format	The identifier will be formed by combining the item, the date and a three-letter code for the institution. Next, this identifier will be encrypted using MD5, so that the identifier becomes a 32-digit number (hexadecimal notation).
Example	b06c0444f37249a0a8f748d3b823ef2a

4. Normalisation

- The SURE Statistics project will attempt to restrict its focus to requests which have consciously been initiated by human users. For this reason, automated visits by internet robots must be filtered from the data as much as possible. The Log Aggregator must maintain a file which list the names of internet robots that individual repositories must use during the filtering of their results. The name of this file must indicate its version. The name of the first file that will be published will be robots-v1.xml. Repositories can use the version indication in the filename to check if they are working with the most recent list of internet robots.
- If a single user clicks repeatedly on the same document within the same 24 hours, this should be counted as a single request.
- One single publication may be split into a set different files. The impact of such variations in the organisation of complex objects must be nullified. The consultation of a part should count towards the statistic of the whole. It should make no difference if a publication consists of one pdf-files or of multiple pdf-files.

6. Data format

In compliance with the JISC Usage Statistics Review, individual usage events need to be serialized in XML using the syntax that is specified in the OpenURL Context Objects schema.[3|http://www.surffoundation.nl/#_ftn3] This section will describe a recommended practice for the use of this schema.

The root element of the XML-document must be <context-objects>. It must contain a reference to the official schema and declare two namespaces: xmlns:ctx="info:ofi/fmt:xml:xsd:ctx" and xmlns:dcterms=http://dublincore.org/documents/2008/01/14/dcmi-terms/.

Each usage event must be described in a separate <context-object> element, which must appear as a direct child of <context-objects>. Two attributes must be used:

- The time and date on which the usage event took place must be recorded in the *timestamp* attribute.
- The identification of the usage event must be captured in an attribute with the name *identifier*.

Within <contextobject>, a number of elements can be used which describe the context of the usage event. The names of these elements are as follows:

- <refererent>: the object that was downloaded or viewed
- <requester> refers to the agent that has initiated the usage event
- <serviceType>: the type of service that was requested
- <referrer>: the system that has forwarded the reader to the downloaded object
- <resolver>: the institution that provides access to the requested item and which has received the usage event.

Information about these contextual entities can be given in four different ways. Firstly, they can be characterised using an <identifier>. Secondly, metadata can be included literally by wrapping these into the file using a <metadata-by-val> element. Thirdly, a reference to metadata stored elsewhere can be included by using <metadata-by-ref>. A fourth method is the use of the element <private-data>. In the SURE Statistics project, only the first two methods shall be used. Listing 1 is an example of a full OpenURL Context Object document.

```
<?xml version="1.0" encoding="UTF-8"?>
<ctx:context-objects xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xmlns:dcterms="http://dublincore.org/documents/2008/01/14/dcmi-terms/"
   xsi:schemaLocation="info:ofi/fmt:xml:xsd:ctx
   http://www.openurl.info/registry/docs/info:ofi/fmt:xml:xsd:ctx"
   xmlns:ctx="info:ofi/fmt:xml:xsd:ctx">
   <ctx:context-object identifier="c06f0464f37249a0a9f848d4b823ef2a"
      timestamp="2009-07-13T09:14:16+02:00">

      <ctx:referent>
         <ctx:identifier>https://openaccess.leidenuniv.nl/dspace/
            handle/1887/584</ctx:identifier>
         <ctx:identifier>http://hdl.handle.net/
            handle/1887/584</ctx:identifier>
      </ctx:referent>

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

      <ctx:requester>
         <ctx:identifier>c06f0464f37249a0a9f84
                     8d4b823ef2a</ctx:identifier>
         <ctx:identifier>118.94.150.0</ctx:identifier>
         <ctx:metadata-by-val>
            <ctx:format>
              http://dublincore.org/documents/dcmi-terms/
            </ctx:format>
            <ctx:metadata>
               <dcterms:spatial>nl</dcterms:spatial>
            </ctx:metadata>
         </ctx:metadata-by-val>
      </ctx:requester>

      <ctx:service-type>
         <ctx:metadata-by-val>
            <ctx:format>
              http://dublincore.org/documents/dcmi-terms/
            </ctx:format>
            <ctx:metadata>
               <dcterms:type>objectFile</dcterms:type>
            </ctx:metadata>
         </ctx:metadata-by-val>
      </ctx:service-type>

      <ctx:resolver>
         <ctx:identifier>openaccess.leidenuniv.nl</ctx:identifier>
      </ctx:resolver>

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

Listing 1.

- Under <referent>, the two identifiers for the requested document must both be given in separate <identifier> elements.
- Element <referring-entitity> contains information on the referrer. The URL that was received from the referrer and the classification of the search engine, if it was used, must both be given in an <identifier> element.

- The <requester>, the agent who has requested the <refererent> must be identified by providing the C-class Subnet, and the encrypted IP-address must both be given in separate <identifier>s. In addition, the name of the country where the request was initiated must be provided. The <metadata-by-val> element must be used for this purpose. The country must be given in <dcterms:spatial>. The dcterms namespace must be declared in the <format> element as well.
- The DC metadata term "type" is used to clarify whether the usage event involved a download of a object file or a metadata view.
- Finally, an <identifier> for the institution that provided access to the downloaded document must be given within <resolver>.
  ----[1|http://www.surffoundation.nl/#_ftnref1] http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2007/usagestatisticsreview.aspx
  [2|http://www.surffoundation.nl/#_ftnref2] http://www.iso.org/iso/english_country_names_and_code_elements
  [3|http://www.surffoundation.nl/#_ftnref3] The XML Schema for XML Context Objects can be accessed at http://www.openurl.info/registry/docs/info:ofi/fmt:xml:xsd:ctx

Application profiles

SURFshare use of Usage Statistics Exchange