

NEEO TECHNICAL GUIDELINES FOR THE EXCHANGE OF USAGE METADATA

DRAFT – Version 1.4

NEEO – WP5

Author: Benoit Pauwels

Date	Version	
9/9/2008	0.1	Initial skeleton document
20/11/2008	1.0	Introduce SWUP: the Scholarly Works Usage Community Profile
26/11/2008	1.1	Correct some errors in examples Added annex 1
17/2/2009	1.2	Description of SWUP ContextObject moved to Annex 1 Introduce NEEO-SWUP ContextObject Clearer guidelines for anonymization of NEEO-SWUP ContextObjects Clearer guidelines for filtering out robot generated NEEO-SWUP ContextObjects
11/6/2009	1.3	2.2: NEEO decides to do MD5 encryption of IP addresses of the requester of a download 2.5: NEEO decides to do robot filtering through matching of the user agent against a list of regular expressions; the list of these expressions (at date of 11/6/2009) is added as annex 2
8/9/2009	1.4	Corrected typos: ctx:referringentity should be ctx:referring-entity in figures 4 and 7. User Agent information should be given as <metadata-by-val> under the referrer entity Example of figure 7 made available on network address Introduction of idea of "session identifier" for the requester descriptor. A standardized way for constructing such an identifier is needed, but clearly out of scope of this SWUP standard.

Table of Contents

1	INTRODUCTION	3
2	NEEO-SWUP: USE OF SWUP WITHIN THE NEEO COMMUNITY	5
2.1	NEEO-SWUP ENTITIES	5
2.2	ANONYMIZATION OF NEEO- SWUP CONTEXTOBJECTS	6
2.3	GEOGRAPHICAL INFORMATION (REQUEST’S COUNTRY OF ORIGIN)	6
2.4	FILTERING OUT “DOUBLE” CLICKS	7
2.5	FILTERING OUT MACHINE GENERATED NEEO-SWUP CONTEXTOBJECTS	7
2.6	MINIMAL NEEO-SWUP CONTEXTOBJECT	8
2.7	RECOMMENDED NEEO-SWUP CONTEXTOBJECT	9
3	EXCHANGE OF NEEO-SWUP CONTEXTOBJECTS	12
3.1	OAI-PMH COMPLIANT USAGE EVENT REPOSITORY	12
3.2	OAI-PMH SPECIFICATIONS AND RECOMMENDATIONS	12
3.3	EXAMPLE OF AN OAI-PMH GETRECORD RESPONSE ENCAPSULATING A RECOMMENDED NEEO-SWUP CONTEXTOBJECT	15
4	ANNEX 1 – SWUP: THE SCHOLARLY WORKS USAGE COMMUNITY PROFILE	17
4.1	THE OPENURL CONTEXTOBJECT	17
4.2	SWUP: THE SCHOLARLY WORKS USAGE COMMUNITY PROFILE	18
4.3	SPECIFICATIONS OF THE SWUP ENTITIES	20
4.4	EXAMPLE OF A SWUP CONTEXTOBJECT ENCAPSULATED IN AN OAI-PMH GETRECORD RESPONSE	28
5	ANNEX 2 – FILTERING OF ROBOT DOWNLOAD REQUESTS – LIST OF REGULAR EXPRESSIONS	31
6	REFERENCES	32

1 Introduction

Most institutional repositories maintain a log of usage events: whenever an end user views an abstract, or wants to download an object file from the IR, a log entry is written. Such an entry typically records who requested which service for which IR item and at what moment. Based on this information an institution is capable of assessing the usage of the IR items and the quality and impact of the IR in its whole. This usage information is typically presented through some Web interface, showing for example:

- how many times every item in the IR has been read,
- which item (and by extrapolation which author, department, ...) is the most popular within the institution or within a given research domain,
- an evolution on the usage of the IR in its whole
- search results get ranked on frequency of download of the object files
- etc

In more advanced environments, mining of the usage data could yield other very interesting value-added services, like:

- the creation of a network of (clusters of) related publications: publications that are read by the same person within a certain amount of time can be considered to be similar in some way,
- recommender systems, in which the end user gets a recommendation on which other publications are of possible interest in relation to a document he wishes to retrieve

It is evident that this kind of services is of far more value to the researcher if the usage data that these are based on are obtained from a multitude of information services, including publisher repositories, institutional repositories, linking servers, etc.

Institutional repositories are information services that hold publications that are produced within the institution. If a publication is produced by multiple co-authors, it is very likely that, depending on the policy within the institution, multiple copies and/or versions and/or formats of the same publication get stored in the IR. Moreover, if the co-authors are affiliated with different institutions, then the publication will be available from multiple institutional repositories. And in most cases, a copy of the publication will also be available on a repository system of the 'official' publisher of the work. Therefore, mining of usage logs of one institutional repository cannot produce metrics that represent the complete usage of a specific item. It is clear that the usage log entries pertaining to the one publication must be aggregated from all information services where this item is held into one system, which in turn can then yield correct usage statistics.

Detection of usage trends, recommender systems, etc can only be interesting if these services are based on usage data for a representative portion of the literature within a specific domain. EO with its aim of aggregating 50.000 references is not in such a position. Other aggregators in economics, such as RePEc with its 600K database and a big worldwide user base are much better placed to take up this kind of service provision. This is what RePEc does to some extent through their LogEC service.

EO will aggregate the RePEc database and will forward its own IR based items to RePEc. The EO portal, EconPapers, and IDEAS will therefore all be services based on (practically) the same database and aiming at the same user base. It would therefore be very interesting that usage data that is captured in these different services can be compared and used to generate metrics that will inevitably be more representative of the usage of economics publications than if each of the 3 services would be doing the same job independently of the others.

Also, our research world is becoming more multi-disciplinary, and our researchers will be increasingly interested in this kind of usage services if they cross several research domains. It is important that Nereus, as a consortium of 20 and more renowned institutions in economics, can subscribe to this global move towards analysis of scholarly usage data, through its NEEO project.

Even though NEEO does not have in its plan the setup of very sophisticated services based on usage data (we will limit ourselves to presentation of access statistics, and ranking of search results on frequency of downloads of items), the EO Gateway and the individual NEEO data providers should still be prepared in order to be able to collaborate in a global setting of usage metrics as described above. NEEO can do this by adopting open standards for the normalization and exchange of their usage data.

In the remainder of this document we explain how we use the XML serialization of the OpenURL ContextObject to capture information on a usage event within an IR, and how these get exchanged with the EO gateway using the OAI-PMH protocol.

Why do we want item-level/objectfile-level usage data?

- Same raw data will permit massaging in all sorts of way afterwards in the central gateway (make click streams based on ID of user (= track usage per user), usage per item over periods of time, determine most popular origins of usage events (based on referrer ID), recommender system: publications that are accessed by the same user within a certain period are probably related?, ...)

Relationship with COUNTER project?

- typically publisher usage data is available in COUNTER format
- COUNTER uses journals as lowest level of granularity
- NEEO = item-level usage (more granular than COUNTER)
- COUNTER works on extension of its standards to item-level for journal articles
- NEEO = not only articles
- First conclusion? COUNTER records and OCO records are not easily comparable, resulting in irrelevant metrics?

Relationship with SUSHI project?

- OAI-based exchange of COUNTER records

2 NEEO-SWUP: use of SWUP within the NEEO community

2.1 NEEO-SWUP entities

The following table gives an overview of the specifications for a NEEO-SWUP *ContextObject*: it explains which SWUP *entities* and *descriptors* are needed and how to fill them in, in order to construct a NEEO-SWUP *ContextObject*. Please refer to [annex 1](#) of this document for more information on the SWUP *entities* and *descriptors*.

Please refer to paragraph [2.6](#) for the specifications of what is at minimum required in a NEEO-SWUP *ContextObject*. Paragraph [2.7](#) lists *entities* and *descriptors* which we recommend all NEEO data providers to implement.

	Mandatory Optional	Constraints on descriptors
Attributes of <context-object>		
<i>Identifier</i>	M	
<i>Timestamp</i>	M	ISO8601-conformant datetime in the YYYY-MM-DD or YYYY-MM-DDTHH:MM:SSZ representation (i.e. time expressed in UTC (Coordinated Universal Time))
Entity		
<i>Referent</i>	M (exactly 1)	The following 2 <ctx:identifier> <i>descriptors</i> must be present: <ul style="list-style-type: none"> Identifier of the object file downloaded (must correspond to the value of the "ref" attribute in the NEEO-DIDL descriptor for the objectFile¹) OAI identifier of the publication (of which the specific object file is downloaded²) More <ctx:identifier> <i>descriptors</i> can be present. A good candidate is the identifier used in the root Item of the NEEO-DIDL representation of the publication ³ .
<i>ReferringEntity</i>	O	If this <i>entity</i> is used in a NEEO-SWUP <i>ContextObject</i> , then at least one <ctx:identifier> must be present, holding the URL of the web page from which the download request was initiated.
<i>Requester</i>	O	If this <i>entity</i> is used in a NEEO-SWUP <i>ContextObject</i> , then at least one <ctx:identifier> must be present, holding the MD5-encryption of the IP address of the browser from which the usage event occurs. Geographical information can be included as a <ctx:metadata-by-val> <i>descriptor</i> . Please refer to paragraph 2.3 .
<i>ServiceType</i>	O	The EO Gateway will consider all incoming NEEO-SWUP <i>ContextObjects</i> as representations for download events. This <i>entity</i> , if present, is ignored by the EO Gateway.
<i>Resolver</i>	M (exactly 1)	Exactly 1 <ctx:identifier> <i>descriptor</i> must be present, holding the OAI baseURL for the repository that

¹ Xpath: "DIDL/Item/Item[Descriptor/Statement/type='info:eu-repo/semantics/objectFile']/Component/Resource@ref"

² Xpath: "OAI-PMH/{GetRecord|ListRecords}/record/header/identifier"

³ Xpath: "DIDL/Item/Descriptor/Statement/dii:Identifier"

		generates this <i>ContextObject</i> . This OAI baseUrl must correspond exactly to the one given in the Admin file of the corresponding NEEO partner institution.
<i>Referrer</i>	0	If this <i>entity</i> is used in a NEEO-SWUP <i>ContextObject</i> , then at least one <code><ctx:metadata-by-val></code> must be present, holding a string that corresponds to the User-Agent header, which is transmitted in the HTTP transaction that generates the download. Please refer to figure 4 for an example.

2.2 Anonymization of NEEO- SWUP ContextObjects

In order to not infringe on privacy laws, it is important that the local IRs anonymize their usage data before it is exposed for aggregation, so that any information that could link back to a person is removed from the usage data.

Within the NEEO community we decide to use MD5 encryption⁴ on the IP address of the user, expressed through the *Requester entity* in the NEEO-SWUP *ContextObject*. An MD5 hash is a 128-bit number, usually expressed as a string of 32 hexadecimal digits.

```
<ctx:requester>
  <ctx:identifier>b06c0444f37249a0a8f748d3b823ef2a</ctx:identifier>
</ctx:requester>
```

Figure 1: Example of a Requester entity identifying a user through an MD5-encrypted IP address

2.3 Geographical information (request's country of origin)

Since the IP address of a requester is transmitted in an encrypted way, it is impossible for the EO Gateway to determine the country of origin of the request. It is the responsibility of each IR to determine the country. This can be easily done using the free GeoIP API's and associated database from MaxMind Inc., available for a lot of OS and programming languages⁵.

The country should be expressed as a `<ctx:metadata-by-val>` *descriptor* on a *Requester entity*, according to the ISO3166⁶ standard and in upper-case.

```
<ctx:requester>
  <ctx:identifier>b06c0444f37249a0a8f748d3b823ef2a</ctx:identifier>
  <ctx:metadata-by-val>
    <ctx:format>http://purl.org/dc/terms/</ctx:format>
```

⁴ It is well known that MD5 encryption is not 100% secure, meaning that the 32-digits strings can be decrypted into a valid IP address. It is technically possible to raise this level of security through the use of a salt on the MD5 encryption, or indeed by using other (more complex) encryption algorithms, like SHA1. However, we need to make sure that all NEEO partners are able to implement this encryption on each of their respective IR platforms, which are written in different programming languages, and hence for which it needs to be checked whether the appropriate libraries are available.

Clearly a balance needs to be found between technical complexity of implementation and the level of security that is legally requested or acceptable.

⁵ MaxMind GeoIP: <http://www.maxmind.com/>

⁶ ISO3166 country codes:

http://www.iso.org/iso/country_codes/iso_3166_code_lists/english_country_names_and_code_elements.htm

```
<ctx:metadata xmlns:dcterms="http://purl.org/dc/terms/">
  <dcterms:spatial xsi:type="dcterms:ISO3166">FR</dcterms:spatial>
</ctx:metadata>
</ctx:metadata-by-val>
</ctx:requester>
```

Figure 2: Example of a Requester entity identifying a user through an MD5-encrypted IP address, with geographical information expressed as an ISO3166 country code

2.4 Filtering out "double" clicks

Double clicks represent consecutive downloads of the same file by the same user within a certain timeframe. This kind of downloads should only be counted once: it is judged not to count all these downloads, since they can impossibly represent 'real' downloads, being too near in time: the same person cannot read the same text within such a short timeframe.

Some discussion in the literature exists on the 'timeframe', ranging from 10 seconds (COUNTER) to 1 month (LogEC). Another issue to address is the hiding of the real origin of the request behind proxies of all sorts, making it not trivial to decide whether a click is coming from the same requester or not.

It is decided that double clicks are filtered out by the EO Gateway. No action is required from the NEEO IRs.

2.5 Filtering out machine generated NEEO-SWUP ContextObjects

Bots, spiders and web crawlers access publications for indexing purposes: typical examples are Google (Scholar), the EO Gateway,... The number of events that are triggered by machines would typically exceed the number of human usage events by many factors. However, the EO Gateway is not interested in these robot usage events, since the objective is to demonstrate the frequency that a given publication is being 'used' by a human being.

Therefore, in order not to inflict a big unnecessary burden on both the data providers and the EO Gateway, which respectively need to generate the NEEO-SWUP *ContextObjects* (in the case of the IR), and receive, store (and throw away most of these) NEEO-SWUPs (in the case of the EO Gateway), it is very important that this filtering out of machine usage events be solved in a correct way on each IR installation.

Furthermore, each data provider needs to do this filtering according to the same specifications: if one data provider filters out robot accesses from the EO Gateway but not from Google Scholar, and the others do, download rates for a publication might be seriously distorted and metrics might be completely irrelevant.

Within the NEEO project we decide to do this filtering based on a list of regular expressions, to be used for identification of robots/spiders/linkcheckers/... through matching on the *User-Agent* HTTP header⁷. Please refer to Annex 2 for a complete list of these regular expressions at the date of 11/6/2009.

⁷ <http://www.robotstxt.org/> -- <http://www.robotstxt.org/db/all.txt>; <http://www.botsvsbrowsers.com/>; <http://www.jafsoft.com/searchengines/webbots.html>; cleaned up version of the aggregation of these lists is available in the robots.pm Perl script of the AWStats software (enhanced by AnteZeta, available at <http://www.antezeta.com/awstats/robots.html>)

Using this list, we have been able to positively match 579 out of 656 user agents of which we know that they represent robots.⁸ Conclusion is therefore that this list is at least a good starting point; although it will be periodically updated through mining of the central EO database of log entries and/or through input from other projects, repositories, etc...

Each data provider must filter out NEEO-SWUP *ContextObjects* using the following algorithm:

Do not expose the NEEO-SWUP *ContextObject* if the following condition is true:

- the User-Agent value contained in the <ctx:identifier> of the *Referrer entity* matches an entry in the EO regular expressions list

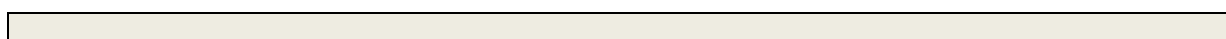
The data provider must make sure to obtain a fresh copy of the EO "regular expressions" list into his local installation at regular intervals.

2.6 Minimal NEEO-SWUP ContextObject

The following table gives an overview of the specifications for a minimal NEEO-SWUP *ContextObject*: at minimum these *entities* and *descriptors* must be delivered to the EO Gateway. It remains however to be encouraged for NEEO institutional repositories to deliver [recommended NEEO-SWUP ContextObjects](#), in order for the EO Gateway to be able to build more advanced services.

Constraints on descriptors	
Attributes of <context-object>	
<i>Identifier</i>	
<i>Timestamp</i>	ISO8601-conformant datetime in the YYYY-MM-DD or YYYY-MM-DDTHH:MM:SSZ representation (i.e. time expressed in UTC (Coordinated Universal Time))
Entity	
<i>Referent</i>	The following 2 <ctx:identifier> <i>descriptors</i> must be present: <ul style="list-style-type: none"> • Identifier of the object file downloaded (must correspond to the value of the "ref" attribute in the NEEO-DIDL descriptor for the objectFile⁹) • OAI identifier of the publication (of which the specific object file is downloaded¹⁰)
<i>Resolver</i>	Exactly 1 <ctx:identifier> <i>descriptor</i> must be present, holding the OAI baseURL for the repository that generates this <i>ContextObject</i> . This OAI baseURL must correspond exactly to the one given in the Admin file of the corresponding NEEO partner institution.

The following figure gives an example of a minimal NEEO-SWUP *ContextObject*, representing a download event for an object file available at <http://di-pot.ulb.ac.be:8080/dspace/handle/2013/789/1/article.pdf> of a publication with identifier "info:hdl:2013/789", which is exposed through an OAI-PMH interface as a record with OAI identifier "oai:di-pot.ulb.ac.be:2013/789".



⁸ Many thanks to the University of Minho for making their valuable data available to us

⁹ Xpath: "DIDL/Item/Item[Descriptor/Statement/type='info:eu-repo/semantics/objectFile']/Component/Resource@ref"

¹⁰ Xpath: "OAI-PMH/{GetRecord|ListRecords}/record/header/identifier"

```

<ctx:context-object
  identifier="dspace-usage.downloads.ulb.ac.be:1"
  timestamp="2008-11-05T13:15:30Z">

  <ctx:referent>
    <ctx:identifier>oai:di-pot.ulb.ac.be:2013/789</ctx:identifier>
    <ctx:identifier>
      http://di-pot.ulb.ac.be:8080/dspace/handle/2013/789/1/article.pdf
    </ctx:identifier>
  </ctx:referent>

  <ctx:resolver>
    <ctx:identifier>
      http://di-pot.ulb.ac.be:8080/dspace-oai/request
    </ctx:identifier>
  </ctx:resolver>

</ctx:context-object>

```

Figure 3: Example of a minimal NEEO-SWUP ContextObject (declarations of namespaces omitted, newlines and whitespaces added, for readability)

2.7 Recommended NEEO-SWUP ContextObject

The following table gives an overview of the specifications for a recommended NEEO-SWUP *ContextObject*.

This NEEO-SWUP *ContextObject* contains the information that is recommended to transmit to the EO Gateway, in order for this latter to be able to build more advanced services (such as, for example, presenting information in the EO portal on the origins of the download requests, thanks to the presence of the *ReferringEntity*).

Constraints on descriptors	
Attributes of <context-object>	
<i>Identifier</i>	
<i>Timestamp</i>	ISO8601-conformant datetime in the YYYY-MM-DD or YYYY-MM-DDTHH:MM:SSZ representation (i.e. time expressed in UTC (Coordinated Universal Time))
Entity	
<i>Referent</i>	Exactly 1 referent. The following 2 <ctx:identifier> <i>descriptors</i> must be present: <ul style="list-style-type: none"> Identifier of the object file downloaded (must correspond to the value of the "ref" attribute in the NEEO-DIDL descriptor for the objectFile¹¹) OAI identifier of the publication (of which the specific object file is downloaded¹²) More <ctx:identifier> <i>descriptors</i> can be present. A good candidate is the identifier used in the root Item of the NEEO-DIDL representation of the publication ¹³ .

¹¹ Xpath: "DIDL/Item/Item[Descriptor/Statement/type='info:eu-repo/semantics/objectFile']/Component/Resource@ref"

¹² Xpath: "OAI-PMH/{GetRecord|ListRecords}/record/header/identifier"

¹³ Xpath: "DIDL/Item/Descriptor/Statement/dii:Identifier"

<i>ReferringEntity</i>	At least one <ctx:identifier> must be present, holding the (XML-encoded) URL of the web page from which the download request was initiated.
<i>Requester</i>	At least one <ctx:identifier> must be present, holding the MD5-encryption of the IP address of the browser from which the usage event occurs. Geographical information can be included as a <ctx:metadata-by-val> <i>descriptor</i> . Please refer to paragraph 2.3 .
<i>Resolver</i>	Exactly 1 <ctx:identifier> <i>descriptor</i> must be present, holding the OAI baseURL for the repository that generates this <i>ContextObject</i> . This OAI baseURL must correspond exactly to the one given in the Admin file of the corresponding NEEO partner institution.
<i>Referrer</i>	At least one <ctx:metadata-by-val> must be present, holding a string that corresponds to the User-Agent header, which is transmitted in the HTTP transaction that generates the download.

The following figure gives an example of a recommended NEEO-SWUP *ContextObject*, representing a download event for an object file available at <http://di-pot.ulb.ac.be:8080/dspace/handle/2013/789/1/article.pdf> of a publication with identifier "info:hdl:2013/789", which is exposed through an OAI-PMH interface as a record with OAI identifier "oai:di-pot.ulb.ac.be:2013/789".

```
<ctx:context-object
  identifier="dspace-usage.downloads.ulb.ac.be:1"
  timestamp="2008-11-05T13:15:30Z">

  <ctx:referent>
    <ctx:identifier>oai:di-pot.ulb.ac.be:2013/789</ctx:identifier>
    <ctx:identifier>
      http://di-pot.ulb.ac.be:8080/dspace/handle/2013/789/1/article.pdf
    </ctx:identifier>
  </ctx:referent>

  <ctx:referring-entity>
    <ctx:identifier>
      http://www.google.be/search?sourceid=navclient&hl=fr&ie=UTF-
      8&rlz=1T4GGIH_frBE212BE212&q=Economic+Risk+in+Hydrocarbon+
      Exploration%e2%80%9d%2c+
    </ctx:identifier>
  </ctx:referring-entity>

  <ctx:requester>
    <ctx:identifier>
      b06c0444f37249a0a8f748d3b823ef2a
    </ctx:identifier>
    <ctx:metadata-by-val>
      <ctx:format>http://purl.org/dc/terms/</ctx:format>
      <ctx:metadata>
        <dcterms:spatial xsi:type="dcterms:ISO3166">BE</dcterms:spatial>
      </ctx:metadata>
    </ctx:metadata-by-val>
  </ctx:requester>

  <ctx:resolver>
    <ctx:identifier>
      http://di-pot.ulb.ac.be:8080/dspace-oai/request
    </ctx:identifier>
```

```
</ctx:resolver>

<ctx:referrer>
  <ctx:metadata-by-val>
    <ctx:format>http://purl.org/dc/terms/</ctx:format>
    <ctx:metadata>
      <dc:identifier>
        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)
      </dc:identifier>
    </ctx:metadata>
  </ctx:metadata-by-val>
</ctx:referrer>

</ctx:context-object>
```

*Figure 4: Example of a recommended NEEO-SWUP ContextObject
(declarations of namespaces omitted, newlines and whitespaces added, for readability)*

3 Exchange of NEEO-SWUP ContextObjects

The NEEO-SWUP *ContextObjects* are exchanged over OAI-PMH, version 2.0.¹⁴

3.1 OAI-PMH compliant usage event repository

OAI-PMH selective harvesting on date is crucial, since we can expect a lot of traffic, and therefore do not wish to re-harvest complete histories of usage events of an information system.

It is therefore advisable that a database of usage events is built at the data provider side, and that this database be fed from entries in log files, which are typically used to record usage events: this could be web server logs, IR software logs, etc.

This database must be made available as an OAI-PMH compliant repository, exposing usage events as NEEO-SWUP *ContextObjects*, and should permit a selection on a datetime, as such able to respond to OAI-PMH `ListRecords` or `ListIdentifiers` requests (from the EO Gateway) that specify the `'from'` and `'until'` parameters.

3.2 OAI-PMH specifications and recommendations

The transportation of the NEEO-SWUP *ContextObjects* to the EO Gateway is done according to the OAI-PMH protocol, in its version 2.0. We summarize here specific guidelines and recommendations that need to be considered for the implementation of this OAI-PMH interface.

3.2.1 MetadataPrefix naming

A NEEO-SWUP compliant OAI-PMH interface **must** support the `"swup"` `metadataPrefix`, and in response to any OAI-PMH request specifying the `"metadataPrefix=swup"` parameter **must** deliver NEEO-SWUP *ContextObjects* in the OAI-PMH response.

3.2.2 Identify response

- *deletedRecord*
Usage events that have happened at some point in time can never be 'undone'. Therefore, the OAI-PMH interface can safely set this feature to `"no"`.
- *adminEmail*
Every NEEO data provider must provide name and email of a person responsible for the correct functioning of the OAI-PMH interface of the usage event repository. In case of malfunctioning of the harvesting process, this person will be asked to check

¹⁴ If experience would show that too many NEEO-SWUP *ContextObjects* need to be exchanged and that this is generating too many OAI-PMH request/response cycles, we might consider an alternative solution for the exchange. A proposal is that the data provider prepares files, which are conformant to the OAI-PMH `ListRecords` response, containing records that are structured as NEEO-SWUP *ContextObjects*. These files need to be made available on a web server at the data provider side. We then need to imagine a file nomenclature, making it possible for the EO Gateway to determine which files need harvesting each time it visits the web server... and/or use the 'HEAD' HTTP request in order to determine the datetime of last modification of a file? Some research to do here...

XML validation errors in the logs generated on the EO Gateway, and will be asked to correct bugs in the implementation of the crosswalk on its usage event repository.

A typical OAI-PMH Identify response would look like this:

```
<Identify>
  <repositoryName>DI-fusion download events</repositoryName>
  <baseURL>
    http://difusion.ulb.ac.be:8080/dspace-oai-downloads/request
  </baseURL>
  <protocolVersion>2.0</protocolVersion>
  <adminEmail>difusion-help@ulb.ac.be</adminEmail>
  <earliestDatestamp>2008-01-01T00:00:00Z</earliestDatestamp>
  <deletedRecord>no</deletedRecord>
  <granularity>YYYY-MM-DDThh:mm:ssZ</granularity>
  <description> ... </description>
</Identify>
```

*Figure 5: Example of an OAI-PMH Identify response
(declaration of namespaces omitted, newlines and whitespaces added, for readability)*

3.2.3 OAI-PMH identifier

OAI-PMH identifiers should be unique and persistent, at least within the NEEO community, and worldwide. The OAI identifiers used in NEEO should adhere to the "OAI identifier format" as described on <http://www.openarchives.org/OAI/2.0/guidelines-oai-identifier.htm>.

As a recommendation: this identifier may be constructed based on the identifier of the NEEO-SWUP *ContextObject* (set in the identifier attribute of the root <context-object> element), by pre-pending it with the string "oai:" (look at the next figure for an example).

```
<GetRecord>
  <record>
    <header>
      <identifier>oai:difusion-usage.downloads.ulb.ac.be:1232</identifier>
      <datestamp>2008-10-27T18:29:07Z</datestamp>
    </header>
    <metadata>
      <ctx:context-object
        version="Z39.88-2004"
        identifier="difusion-usage.downloads.ulb.ac.be:1232"
        timestamp="2008-11-05T13:15:30Z">
        <ctx:referent>...</ctx:referent>
        <ctx:service-type>...</ctx:service-type>
      </ctx:context-object>
    </metadata>
    <about>
      ...
    </about>
  </record>
</GetRecord>
```

*Figure 6: Example of an OAI-PMH GetRecord response for a SWUP ContextObject
(declaration of namespaces omitted, newlines and whitespaces added, for readability)*

3.2.4 OAI-PMH set(s)

The concept of OAI-PMH sets gives the possibility to an IR administrator to divide its IR content into logical collections. A service provider that aggregates records from such an IR can specify the OAI-PMH set(s) it wants to harvest through the `set` parameter on the `listRecords` and `listIdentifiers` OAI-PMH verbs.

NEEO IRs can use this facility to limit harvesting to a part (or several parts) of their IR content. In the context of a NEEO repository of usage events, one could imagine the definition of sets based on the following criteria:

- events for all / NEEO publications
- download / abstract-view events
- machine-triggered / human-triggered events
- combination of the above criteria

3.2.5 Resumption token lifespan

All NEEO data providers should respect a reasonable time for the resumption token to be kept alive: recommended is 24 hours.

3.2.6 Harvest batch size

Every NEEO data provider should deliver OAI-PMH records in batches of a reasonable size. Two considerations need to be made here:

- we can expect quite some traffic: download events can occur frequently in any NEEO institutional repository, resulting in a lot of NEEO-SWUP *ContextObjects* to be exchanged over this OAI-PMH interface
- it is expected that the size of the OAI-PMH responses to be limited: only minimal NEEO-SWUP *ContextObjects* will in most cases be transmitted.

Taking into account these two considerations, we recommend a harvest batch size of up to 1000 records.

3.2.7 Frequency of harvesting

In order to limit the amount of data to be exchanged in one harvest operation, it is recommended that incremental harvesting of NEEO-SWUP *ContextObjects* be done on (at least) a weekly basis.

Also, given the rather big amount of data that needs to be exchanged, a re-harvest of a complete history of usage events from an institutional repository should really be the exception.

Frequency of (bulk and/or incremental) harvesting needs to be agreed upon with the EO support desk at Tilburg University (economistsonline@uvt.nl).

3.2.8 XML validation of ingested OAI-PMH records

All ingested OAI-PMH records are validated against the XML Schemas that are used within the record (OAI-PMH, SWUP *ContextObject*, DIDL, MODS, etc).

Ingested records that fail to validate are refused in the EO gateway. The administrator of the IR is advised by the EO support desk (economistsonline@uvt.nl) of this shortcoming through an email message.

3.3 Example of an OAI-PMH GetRecord response encapsulating a recommended NEEO-SWUP ContextObject

This example (with proper XML encoding) is also available on <http://homepages.ulb.ac.be/~bpauwels/NEEO/WP5/neeswup-example.xml>.

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
    http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2002-02-08T08:55:46Z</responseDate>
  <request verb="GetRecord"
    identifier=" oai:difusion-usage.downloads.ulb.ac.be:1232"
    metadataPrefix="swup">
    http://bib15.ulb.ac.be:8080/dspace-oai-downloads/request
  </request>
  <GetRecord>
  <record>
  <header>
    <identifier>oai:arXiv.org:cs/0112017</identifier>
    <datestamp>2001-12-14</datestamp>
    <setSpec>cs</setSpec>
    <setSpec>math</setSpec>
  </header>
  <metadata>
  <ctx:context-object
    xmlns:ctx="info:ofi/fmt:xml:xsd:ctx"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="info:ofi/fmt:xml:xsd:ctx
    http://www.openurl.info/registry/docs/xsd/info:ofi/fmt:xml:xsd:ctx"
    identifier="dspace-usage.downloads.ulb.ac.be:1"
    timestamp="2008-11-05T13:15:30Z">
  <ctx:referent>
    <ctx:identifier>oai:di-pot.ulb.ac.be:2013/789</ctx:identifier>
    <ctx:identifier>
      http://di-pot.ulb.ac.be:8080/dspace/handle/2013/789/1/article.pdf
    </ctx:identifier>
  </ctx:referent>
  <ctx:referring-entity>
    <ctx:identifier>
      http://www.google.be/search?sourceid=navclient&hl=fr&ie=UTF-
      8&rlz=1T4GGIH_frBE212BE212&q=Economic+Risk+in+Hydrocarbon+
      Exploration%e2%80%9d%2c+
    </ctx:identifier>
  </ctx:referring-entity>
  <ctx:requester>
    <ctx:identifier>
      b06c0444f37249a0a8f748d3b823ef2a
    </ctx:identifier>
    <ctx:metadata-by-val>
    <ctx:format>http://purl.org/dc/terms/</ctx:format>
    <ctx:metadata
      xmlns:dcterms="http://purl.org/dc/terms/"
      xsi:schemaLocation="http://purl.org/dc/terms/
      http://dublincore.org/schemas/xmls/qdc/dcterms.xsd"
      <dcterms:spatial
        xsi:type="dcterms:ISO3166">BE</dcterms:spatial>
      </ctx:metadata>
    </ctx:metadata-by-val>
  </ctx:context-object>
  </record>
  </GetRecord>
</OAI-PMH>
```

```

</ctx:requester>
<ctx:resolver>
  <ctx:identifier>
    http://di-pot.ulb.ac.be:8080/dspace-oai/request
  </ctx:identifier>
</ctx:resolver>
<ctx:referrer>
  <ctx:metadata-by-val>
    <ctx:format>http://purl.org/dc/elements/1.1/</ctx:format>
    <ctx:metadata
      xmlns:dc="http://purl.org/dc/elements/1.1/"
      xsi:schemaLocation="http://purl.org/dc/elements/1.1/
        http://dublincore.org/schemas/xmls/qdc/dc.xsd">
      <dc:identifier>
        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)
      </dc:identifier>
    </ctx:metadata>
  </ctx:metadata-by-val>
</ctx:referrer>
</ctx:context-object>
</metadata>
</record>
</GetRecord>
</OAI-PMH>

```

*Figure 7: recommended NEEO-SWUP ContextObject encapsulated in an OAI-PMH GetRecord response
XML encoding left out, newlines and whitespaces added, for readability.*

4 Annex 1 – SWUP: the Scholarly Works Usage Community Profile

4.1 The OpenURL ContextObject

Each institutional repository platform probably stores its usage metadata in a proprietary manner, be it in a web service log or an IR specific log. When it comes to exchanging usage metadata with an aggregator such as the EO gateway (or any other system that wishes to build services on top of IR usage data), we need to agree on a common specification for the representation of this data, to which each of the data providers in the network need to adhere.

Any log with usage data typically records the following information:

- an identification of the item that was used
- a indication of the date and time at which this item was used
- an identifier of an end user who used this item
- an indication of what type of usage has been done (typically a download request)

In some cases the logs can contain additional information, such as an identification of the service from where the usage request was made by an end user.

Underneath you find two typical log entries: one from an Apache web server log, another from a DSpace log. In both cases the above information can be recognized: item-id, request datetime, user-id, and request type.

```
193.172.168.174 - - [03/Mar/2008:15:10:30 +0100] "GET /show.cgi?fid=63617
HTTP/1.1" 200 930558
```

Figure 8: Typical Apache entry log

```
2008-03-03 02:55:07,816 INFO org.dspace.app.webui.servlet.RetrieveServlet
@anonymes:session_id=926496725D2E8BE7C554E7415270F984:view_bitstream:bitst
ream_id=5576
```

Figure 9: Typical Dspace entry log

Making this kind of usage information available to the world outside of an IR, means that we need an XML-serializable data structure which can incorporate information on a request made by a user within a specific context for a specific item. The OpenURL ContextObject as defined within the OpenURL Framework¹⁵ fulfils exactly this requirement and has also been recommended by the JISC Usage Statistics Review project¹⁶ as the standard to be used for the exchange of item-level usage data.

¹⁵ The OpenURL Framework for Context-Sensitive Services :

http://www.niso.org/kst/reports/standards?step=2&gid=None&project_key=d5320409c5160be4697dc046613f71b9a773cd9e

¹⁶ JISC Usage Statistics Review project :

<http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2007/usagestatisticsreview.aspx>

Moreover, the MESUR project¹⁷, conducted at the Los Alamos National Laboratory, has already built various services based on the OAI exchange of usage data formatted as OpenURL ContextObjects.

An *OpenURL ContextObject* is defined as a data structure that holds information on the following 6 *entities*:

- *Referent*: this entity corresponds to the resource which this *ContextObject* is about
- *ReferringEntity*: an entity that references the *Referent*
- *Requester*: an entity that describes the resource that requests services pertaining to the *Referent*
- *ServiceType*: type of service requested
- *Resolver*: a resource that can deliver the requested services
- *Referrer*: a resource that generates the ContextObject

“The descriptions of the *ReferringEntity*, the *Requester*, the *ServiceType*, the *Resolver*, and the *Referrer* express the *Context* in which the *Referent* is referenced and in which the request for services pertaining to the *Referent* takes place.”¹⁸

Each of these *entities* is described through *descriptors*, which can be of 4 different types:

- *identifier*: identifier for the *entity*
- *metadata-by-val*: metadata about the *entity*; the metadata is included ‘by-value’ in the ContextObject
- *metadata-by-ref*: metadata about the *entity*; the metadata is available at a network location
- *private-data*: metadata about the *entity*; the format is not defined within the OpenURL Framework (but rather defined within a specific community)

The *OpenURL Framework Initiative* stipulates that a community can define an *OpenURL Framework Application* by selecting specific values for each of its core components: *Namespaces*, *Character Encodings*, *Serialization*, *Constraint Languages*, *ContextObject Formats*, *Metadata Formats*, and *Transports*. The aggregate of these selections is known as a *Community Profile*, and can be registered within the *Framework*.

In the next chapter we define such a *Community Profile* that we will refer to as “SWUP”, the Scholarly Works Usage community Profile. It defines the characteristics of an *OpenURL Framework Application*, which aims at a normalized description of usage events of scholarly works.

4.2 SWUP: the Scholarly Works Usage Community Profile

Any *OpenURL Community Profile* defines the core characteristics of an *Application* as a list of *Registry* entries. The following table shows the *Registry Identifiers* used in SWUP.

Core Component	Registry Entry	Registry Identifier
<i>Namespaces</i>
<i>Character Encodings</i>	UTF-8 Unicode	info:ofi/enc:UTF-8
<i>Serialization</i>	W3C XML 1.0	info:ofi/fmt:xml

¹⁷ MESUR project: <http://www.mesur.org>

¹⁸ Extract from : “The OpenURL Framework for Context-Sensitive Services”, at : http://www.niso.org/kst/reports/standards?step=2&gid=None&project_key=d5320409c5160be4697dc046613f71b9a773cd9e

<i>Constraint Language</i>	W3C XML Schema	info:ofi/fmt:xml:xsd
<i>ContextObject Format</i>	XML ContextObject Format	info:ofi/fmt:xml:xsd:ctx
<i>Metadata Formats</i>	XML Metadata Format for Scholarly ServiceTypes	info:ofi/fmt:xml:xsd:sch_svc
	Open Archives Initiative Unqualified Dublin Core	info:ofi/fmt:xml:xsd:oai_dc
	MODS Metadata Object Description Schema. Version 3.2	info:ofi/fmt:xml:xsd:mods
	XML Metadata Format for Scholarly Works designed as Compound Objects, using DIDL according to the NEEO application profile ¹⁹	info:ofi/fmt:xml:xsd:sch_didl Not registered... yet
<i>Transports</i>

Exactly 1 *ContextObject* may be represented per XML instance, and the following constraints are defined for the *entities* contained in this *ContextObject*: it must have exactly 1 *Referent*. All other *entities* are optional. SWUP also imposes the presence of attributes *identifier* and *timestamp* in the root <context-object> XML element.

The following table gives an overview of the *attributes* and *entities* as they are used in a SWUP *ContextObject*. Each of these is then addressed in some detail in the following chapter, explaining constraints and recommendations for its *descriptors*.

	Description	Mandatory Optional	Number Allowed
Attributes of <context-object>			
<i>version</i>	Version of OpenURL Framework Standard. Fixed value: "z39.88-2004"	O	<= 1
<i>identifier</i>	An identifier for the usage event	M	= 1
<i>timestamp</i>	Date and time at which the usage event occurred	M	= 1
Entity			
<i>Referent</i>	The publication for which the usage event occurred	M	= 1
<i>ReferringEntity</i>	The entity managed by the <i>Referrer</i> through which the usage event was initiated	O	<= 1
<i>Requester</i>	Person or resource that caused the usage event	O	<= 1
<i>ServiceType</i>	Type of usage event. Typical services are: "abstract view" and "full"	O	<= 1

¹⁹ NEEO Technical Guidelines, DIDL and MODS Application Profiles:
http://homepages.ulb.ac.be/~bpauwels/NEEO/WP5/WP5_Technical_guidelines.pdf

	text view"		
<i>Resolver</i>	The resource that delivers the requested service. This corresponds to the service that permits the download or abstract view.	0	<= 1
<i>Referrer</i>	The resource from which the requester initiated the usage event. This corresponds to the browser of the user (typically identified through the User-Agent HTTP header)	0	<= 1

The *OpenURL Framework Standard* also specifies that the above selections of *Registry Identifiers* and constraints must be available in a machine readable way as an *XML Document*. This file is available on <http://homepages.ulb.ac.be/~bpauwels/NEEO/WP5/swup.xml>.²⁰

4.3 Specifications of the SWUP entities

4.3.1 identifier

This must be a globally unique identifier which unambiguously identifies the usage event. Examples are:

- a UUID²¹ generated at the time of creation of the *ContextObject*
- an identifier constructed through concatenation of the DNS entry of the server that creates the *ContextObject*, followed by the value of a primary key in a locally maintained SQL database.

4.3.2 timestamp

This is the datetime at which the usage event took place.

Format: ISO8601-conformant datetime in the YYYY-MM-DD or YYYY-MM-DDTHH:MM:SSZ representation (i.e. time expressed in UTC (Coordinated Universal Time)).

The following figure shows a *ContextObject* representing a usage event that took place on November 5, 1994, 8:15:30 am, US Eastern Standard Time. The identifier is constructed as a concatenation of a DNS name and a locally generated (and locally guaranteed) unique number.

```
<ctx:context-object
  version="Z39.88-2004"
  identifier="dspace-usage.downloads.ulb.ac.be:1"
  timestamp="2008-11-05T13:15:30Z">
  <ctx:referent>...</ctx:referent>
</ctx:context-object>
```

²⁰ The SWUP profile is not yet registered, and hence this XML specification of the profile could need revision.

²¹ UUID: Universal Unique Identifier : <http://en.wikipedia.org/wiki/UUID>

4.3.3 Referent

The *Referent* represents the item or publication for which a usage event took place. Exactly one **must** be present in a SWUP *ContextObject*.

It is important that the referent be identified as unambiguous as possible, especially through globally unique identifiers (such as DOI, NBN, etc); so that a service provider that aggregates usage event information from a multitude of institutional repositories (or any information service such as a publisher e-journals website) that expose its usage events as SWUP *ContextObjects*, can compare the harvested usage data, and present interesting services, such as a global usage download rate for a certain publication.

Two examples:

- consider the following scenario: an institutional repository contains a postprint copy of an article which is also made available through the SpringerLink service. Both the IR and Springer service expose the information of each download action for this article from any user as a SWUP *ContextObject*, which contains the DOI of the article. An aggregator that harvests these SWUP *ContextObjects* can now discover that the corresponding download clicks were aimed at the same (although, in this case, different versions of the) article, and hence present a more global figure with respect to the visibility of this work.
- <http://www.sciencedirect.com/science/book/9780124441651>: this web page presents the table of contents of the e-book with title "Economic Risk in Hydrocarbon Exploration", with links to a PDF version for each of the chapters. A click on one of these links (probably) generates an entry in a log file of the ScienceDirect service, based on which Elsevier can create a SWUP *ContextObject*, stipulating this ebook as *Referent*.

http://www.google.be/search?sourceid=navclient&hl=fr&ie=UTF-8&rlz=1T4GGIH_frBE212BE212&q=Economic+Risk+in+Hydrocarbon+Exploration%e2%80%9d%2c+ : this webpage shows the first 10 hits in a result set generated upon a Google search for the phrase "Economic Risk in Hydrocarbon Exploration". One of the entries points to the full text of the same e-book, however available as a (scanned and hence different) version on books.google.be. Again, a click on this entry permits books.google.be to generate a SWUP *ContextObject* with the same *Referent* as above (namely the specific e-book).

If both SWUP *ContextObjects* contain the ISBN of the e-book expressed as one of the <ctx:identifier> *descriptors* for the *Referent*, an aggregator that gathers SWUP *ContextObjects* from Elsevier's Scencedirect and Google's Book Search services, is now able to present cumulative statistics for this e-book based on download events generated from these two information systems.

The *OpenUrl Framework Standard* proposes 4 types of *descriptors* for the description of the *Referent*. SWUP imposes the following constraints on these *descriptors* for a *Referent*:

- <ctx:identifier>

At least one identifier **must** be given. Recommended is to include as many as possible globally unique and widely known identifiers (like DOI, NBN, RePEc handle, ISBN, LCCN, etc). Within certain communities it can be useful to include identifiers of both the item/publication and of the object file.

It is **recommended** to specify these identifiers within the "info:ofi/nam" namespaces, if these are defined in the *OpenURL Framework Registry*²².

```
<ctx:identifier>
  info:ofi/nam:info:doi:10.1016/j.jeconom.2008.09.002
</ctx:identifier>

<ctx:identifier>info:ofi/nam:urn:ISBN:978-0-12-444165-1</ctx:identifier>

<ctx:identifier>RePEc:osi:journl:v:4:y:2008:p:13-33</ctx:identifier>
```

Figure 11: Examples of <ctx:identifier> descriptors for a Referent

- <ctx:metadata-by-ref>

These *descriptors* are used to describe the *Referent* through a reference to (typically) a fragment of XML that sits around on some network location. Usage of this *descriptor* is optional in SWUP.

Multiple of these *descriptors* may be included in a SWUP *ContextObject*. As an example:

- <ctx:format> set to "info:ofi/fmt:xml:xsd:sch_didl"²³; and
- <ctx:location> being an URL pointing at the DIDL representation of the publication, according to the NEEO Application profiles²⁴

```
<ctx:metadata-by-ref>
  <ctx:format>info:ofi/fmt:xml:xsd:sch_didl</ctx:format>
  <ctx:location>
    http://bib11.ulb.ac.be:8080/dspace-
      oai/request?verb=GetRecord&metadataPrefix=didl&identifier=oai:di-
      pot.ulb.ac.be:2013/6407
  </ctx:location>
</ctx:metadata-by-ref>
```

Figure 12: Example of a <ctx:metadata-by-ref> descriptor for a Referent; expressed according the info:ofi/fmt:xml:xsd:sch_didl metadata format (XML encoding left out, for readability)

- <ctx:metadata-by-val>

These *descriptors* are used to describe the *Referent* through the inclusion of an XML fragment in the SWUP *ContextObject*. Typically this would be expressed according to widely known *MetadataFormats*, such as "info:ofi/fmt:xml:xsd:oai_dc" (Dublin

²² Namespace entries in the OpenURL Framework Registry are available at : http://alcme.oclc.org/openurl/servlet/OAIHandler?verb=ListRecords&metadataPrefix=oai_dc&set=Core:Namespaces

²³ The "info:ofi/fmt:xml:xsd:sch_didl" Metadata Format is not yet registered in the OpenURL Framework Registry

²⁴ NEEO Technical guidelines, DIDL and MODS application profiles : http://homepages.ulb.ac.be/~bpauwels/NEEO/WP5/WP5_Technical_guidelines.pdf

Core)²⁵ or "info:ofi/fmt:xml:xsd:mods" (MODS)²⁶. Usage of this *descriptor* is optional in SWUP.

SWUP does not impose the usage of <ctx:metadata-by-val>, but if used, it is recommended that this metadata contains the following minimal information in order to be useful for an aggregator (*using MODS entries*):

- o <mods:titleInfo>: title of the publication
- o <mods:originInfo>: year of publication
- o <mods:genre>: type of publication (recommended values: as registered under the info:eu-repo/semantics sub-namespace)
- o <mods:relatedItem type="host"> (if applicable): information on the host of the publication (for example, information on a journal in the case of an article)

```
<ctx:metadata-by-val>
  <ctx:format>info:ofi/fmt:xml:xsd:mods</ctx:format>
  <ctx:metadata>
    <mods:titleInfo>
      <mods:title>
        On linear models with rational expectations which admit
        a unique solution
      </mods:title>
    </mods:titleInfo>
    <mods:genre>
      info:eu-repo/semantics/article
    </mods:genre>
    <mods:originInfo>
      <mods:dateIssued encoding="iso8601">1984</mods:dateIssued>
    </mods:originInfo>
    <mods:relatedItem type="host">
      <mods:titleInfo>
        <mods:title>European economic review</mods:title>
      </mods:titleInfo>
      <mods:identifier type="uri">urn:issn:0014-2921</mods:identifier>
      <mods:part>
        <mods:detail type="volume">
          <mods:number>24</mods:number>
        </mods:detail>
        <mods:extent unit="pages">
          <mods:start>103</mods:start>
          <mods:end>111</mods:end>
        </mods:extent>
      </mods:part>
    </mods:relatedItem>
  </ctx:metadata>
</ctx:metadata-by-val>
```

Figure 13: Example of a <ctx:metadata-by-val> descriptor for a Referent; expressed according the info:ofi/fmt:xml:xsd:mods metadata format (declarations of namespace omitted, for readability)

²⁵ "info:ofi/fmt:xml:xsd:oai_dc" metadata format:

http://alcme.oclc.org/openurl/servlet/OAIHandler?verb=GetRecord&metadataPrefix=oai_dc&identifier=info:ofi/fmt:xml:xsd:oai_dc

²⁶ "info:ofi/fmt:xml:xsd:mods" metadata format:

http://alcme.oclc.org/openurl/servlet/OAIHandler?verb=GetRecord&metadataPrefix=oai_dc&identifier=info:ofi/fmt:xml:xsd:mods

4.3.4 ReferringEntity

This is the resource from which the usage event pertaining to the *Referent* was initiated. This would typically be a web page, be it static or a dynamic one that, for example, holds a search result list in an information portal.

Revisiting the examples introduced under the “Referent” section above: the following two URLs are the *ReferringEntities* in the SWUP *ContextObjects* that would be generated through the ScienceDirect and Google Book Search services:

- <http://www.sciencedirect.com/science/book/9780124441651>
- http://www.google.be/search?sourceid=navclient&hl=fr&ie=UTF-8&rlz=1T4GGIH_frBE212BE212&q=Economic+Risk+in+Hydrocarbon+Exploration%e2%80%9d%2c+

A *ReferringEntity* is optional in a SWUP *ContextObject*, but if used, it **must** be specified through at least one `<ctx:identifier>` descriptor.

The first URL would show up in a SWUP *ContextObject* as follows:

```
<ctx:referring-entity>
  <ctx:identifier>
    http://www.sciencedirect.com/science/book/9780124441651
  </ctx:identifier>
</ctx:referring-entity>
```

Figure 14: Example of a ReferringEntity entity

4.3.5 Requester

Information on the user who initiated the usage event can be very interesting to collect. It will permit, for example, to filter out ‘false’ and ‘double clicks’ (i.e. when the user revisits the same full-text within a certain timeframe, it can be advised to count these successive clicks as just one download event), or it can help to determine geographical spread of usage, based on request’s country of origin.

Another possible service could be to track the behaviour (typically referred to as the ‘click stream’) of a user, as (s)he navigates through the information space visiting abstracts and full texts of scholarly works. This could be valuable input to recommender systems through which a user gets recommendations on information resources to visit, based on navigation behaviours of other users.

A typical identifier of a requester would be an IP address of a system from which the user initiated the usage event. This information is commonly available in log entries of probably all web-based information systems. However several problems need to be addressed in this respect:

- *Proxy servers* can hide the real identity of the user. Some proxy software platforms have configuration possibilities that permit the IP address of the original user still to be carried and therefore to be visible to the system that generates a SWUP *ContextObject*. But not all do so.
- *Bots, spiders, crawlers*: usage events can be initiated by machines as opposed to human beings. We are typically only interested in the human interaction, as this reveals the ‘real’ visibility of a work within the scholarly community. It is therefore desirable to filter out usage events from these machines. This supposes that we know how these machines can be identified.

- *Privacy*: IP addresses are not anonymous; this is a problem (in some countries) as it infringes on privacy laws. This problem can be solved through anonymization of IP addresses. This could be fairly easily done at creation time of the SWUP *ContextObject*, for example through MD5 encryption: the information on the requester is anonymized to the outside world. It is however not clear to the author if these 'simple' procedures are sufficient enough to solve the infringement problems on privacy laws: further legal advice seems necessary, and this is probably to be done on a per-country basis.

Another way of identifying a requester is through a (anonymous) session identifier, to be expressed as a `<ctx:identifier>` *descriptor*. In order for this identifier to reveal cross-system behaviour of users, a standardized way for the construction of these identifiers would need to be agreed upon. This is out of the scope of this standard.²⁷

A *Requester* is optional in a SWUP *ContextObject*, but if used, it **must** be specified through at least one `<ctx:identifier>` descriptor. The following figure shows a *Requester* entity: the user is identified through an IP address.

```
<ctx:requester>
  <ctx:identifier>urn:ip:164.15.4.89</ctx:identifier>
</ctx:requester>
```

Figure 15: Example of a Requester entity, identifying the requester through an IP address.

4.3.6 ServiceType

SWUP does not impose the usage of a *ServiceType* entity, but if used, it **must** contain exactly one `<ctx:metadata-by-val>` *descriptor*, specifying the type of service request, according to the "info:ofi/fmt:xml:xsd:sch_svc" *MetadataFormat*. At the date of 13/2/2009, OFI-registered service types of interest in the SWUP context are: "fulltext" and "abstract"; but SWUP is not limited to these types.

Specific communities could agree on a default value in case the *ServiceType* entity is not used in a SWUP *ContextObject*.

```
<ctx:service-type>
  <ctx:metadata-by-val>
    <ctx:format>info:ofi/fmt:xml:xsd:sch_svc</ctx:format>
    <ctx:metadata>
      <sv:fulltext xmlns:sv="info:ofi/fmt:xml:xsd:sch_svc"
        xsi:schemaLocation="...">
        yes
      </sv:fulltext>
    </ctx:metadata>
  </ctx:metadata-by-val>
</ctx:service-type>
```

Figure 16: "fulltext" ServiceType entity (declaration of schemaLocation omitted, for readability)

4.3.7 Resolver

²⁷ A W3C working draft exists on this issue: <http://www.w3.org/TR/WD-session-id>.

This entity is used to identify the resource or system that delivers the requested service, typically a download of the text of a scholarly work. Such systems are institutional repositories or any repository that is capable of exposing scholarly works through a web service, such as Elsevier's ScienceDirect²⁸, JSTOR²⁹, or YouTube³⁰.

SWUP does not impose the usage of a *Resolver entity*, but if used, it **must** be specified through at least one <ctx:identifier> *descriptor*. These identifiers could be for example a value within the "info:sid" namespace³¹ or the OAI baseURL of an institutional repository.

```
<ctx:resolver>
  <ctx:identifier>
    info:sid/www.sciencedirect.com
  </ctx:identifier>
</ctx:resolver>

<ctx:resolver>
  <ctx:identifier>
    http://bib11.ulb.ac.be:8080/dspace-oai/request
  </ctx:identifier>
</ctx:resolver>
```

Figure 17: Two examples of resolver entities: Elsevier's ScienceDirect and an institutional repository

4.3.8 Referrer

This entity represents the system from which the usage event was initiated. This corresponds to the web browser of the user who requested, for example, the download of a file.

A *Referrer* is optional within a SWUP *ContextObject*, but, if used, it **must** be specified through at least one <ctx:metadata-by-val> descriptor. A browser can typically be identified through the User-Agent HTTP header.

```
<ctx:referrer>
  <ctx:metadata-by-val>
    <ctx:format>http://purl.org/dc/elements/1.1/</ctx:format>
    <ctx:metadata xmlns:dc="http://purl.org/dc/elements/1.1/"
      xsi:schemaLocation="http://purl.org/dc/elements/1.1/
        http://dublincore.org/schemas/xmls/qdc/dc.xsd">
      <dc:identifier>
        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)
      </dc:identifier>
    </ctx:metadata>
  </ctx:metadata-by-val>
```

²⁸ Elsevier's ScienceDirect: <http://www.sciencedirect.com/>

²⁹ JSTOR: <http://www.jstor.org/>

³⁰ YouTube: <http://www.youtube.com/>

³¹ "info:sid" namespace : <http://info-uri.info/registry/OAIHandler?verb=GetRecord&metadataPrefix=reg&identifier=info:sid/>

```
</ctx:referrer>
```

Figure 18: A FireFox browser as a Referrer entity

4.4 Example of a SWUP ContextObject encapsulated in an OAI-PMH GetRecord response

This example shows a download event for an item which is known under the following identifiers:

- info:hdl:2013/6407
- RePEc:eee:eecrev:v:24:y:1984:i:1:p:103-111
- info:doi:10.1016/0014-2921(84)90015-1

It includes all declarations of namespaces, is XML encoded and validates³² against all XML schemas used. Some whitespaces and newlines have been added for readability.

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH
  xmlns="http://www.openarchives.org/OAI/2.0/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2002-02-08T08:55:46Z</responseDate>
  <request verb="GetRecord"
    identifier=" oai:difusion-usage.downloads.ulb.ac.be:1232"
    metadataPrefix="swup">
    http://bib15.ulb.ac.be:8080/dspace-oai-downloads/request
  </request>
  <GetRecord>
    <record>
      <header>
        <identifier>oai:arXiv.org:cs/0112017</identifier>
        <datestamp>2001-12-14</datestamp>
        <setSpec>cs</setSpec>
        <setSpec>math</setSpec>
      </header>
      <metadata>
        <ctx:context-object
          xmlns:ctx="info:ofi/fmt:xml:xsd:ctx"
          xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
          xsi:schemaLocation="info:ofi/fmt:xml:xsd:ctx
http://www.openurl.info/registry/docs/xsd/info:ofi/fmt:xml:xsd:ctx"
          version="Z39.88-2004"
          identifier="dspace-usage.downloads.ulb.ac.be:1"
          timestamp="2008-11-05T13:15:30Z">
          <ctx:referent>
            <ctx:identifier>info:hdl:2013/6407</ctx:identifier>
            <ctx:identifier>RePEc:eee:eecrev:v:24:y:1984:i:1:p:103-111
              </ctx:identifier>
            <ctx:identifier>info:doi:10.1016/0014-2921(84)90015-1
              </ctx:identifier>
          <ctx:metadata-by-val>
            <ctx:format>info:ofi/fmt:xml:xsd:mods</ctx:format>
          <ctx:metadata>
            <mods:mods version="3.2"
              xmlns:mods="http://www.loc.gov/mods/v3"
              xsi:schemaLocation="http://www.loc.gov/mods/v3
http://www.loc.gov/standards/mods/v3/mods-3-2.xsd">
              <mods:titleInfo>
```

³² using “Altova XMLSpy Professional Edition 2008 sp1”

```

    <mods:title>
      On linear models with rational expectations which admit
      a unique solution
    </mods:title>
  </mods:titleInfo>
  <mods:genre>
    info:eu-repo/semantics/article
  </mods:genre>
  <mods:originInfo>
    <mods:dateIssued encoding="iso8601">1984</mods:dateIssued>
  </mods:originInfo>
  <mods:relatedItem type="host">
    <mods:titleInfo>
      <mods:title>European economic review</mods:title>
    </mods:titleInfo>
    <mods:identifier type="uri">
      urn:issn:0014-2921
    </mods:identifier>
    <mods:part>
      <mods:detail type="volume">
        <mods:number>24</mods:number>
      </mods:detail>
      <mods:extent unit="pages">
        <mods:start>103</mods:start>
        <mods:end>111</mods:end>
      </mods:extent>
    </mods:part>
  </mods:relatedItem>
</mods:mods>
</ctx:metadata>
</ctx:metadata-by-val>
<ctx:metadata-by-ref>
  <ctx:format>info:ofi/fmt:xml:xsd:sch_didl</ctx:format>
  <ctx:location>
    http://bib11.ulb.ac.be:8080/dspace-
    oai/request?verb=GetRecord&metadataPrefix=didl&
    identifier=oai:di-pot.ulb.ac.be:2013/6407
  </ctx:location>
</ctx:metadata-by-ref>
</ctx:referent>
<ctx:requester>
  <ctx:identifier>urn:ip:164.15.4.89</ctx:identifier>
</ctx:requester>
<ctx:service-type>
  <ctx:metadata-by-val>
    <ctx:format>info:ofi/fmt:xml:xsd:sch_svc</ctx:format>
  <ctx:metadata>
    <sv:fulltext
      xmlns:sv="info:ofi/fmt:xml:xsd:sch_svc"
      xsi:schemaLocation="info:ofi/fmt:xml:xsd:sch_svc
      http://www.openurl.info/registry/docs/xsd/
      info:ofi/fmt:xml:xsd:sch_svc">
      yes
    </sv:fulltext>
  </ctx:metadata>
</ctx:metadata-by-val>
</ctx:service-type>
<ctx:referrer>
  <ctx:identifier> 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)
</ctx:identifier>

```

```
    </ctx:referrer>  
  </ctx:context-object>  
</metadata>  
</record>  
</GetRecord>  
</OAI-PMH>
```

5 Annex 2 – Filtering of robot download requests – list of regular expressions

Total of 50 regular expressions (as of 11/6/2009):

```
celestial
daumoa
robots
crawler
appie
architext
jeeves
bjaaland
contentmatch
ferret
gulliver
virus[_+ ]detector
harvest
htdig
linkwalker
lilina
lycos[_+ ]
moget
myweb
nomad
scooter
slurp
^voyager\/
weblayers
libwww
nutch
spider
google
bot
yacy
yahoo
findlinks
lwp
heritrix
urllib
xenu
larbin
onetszukaj
linkcheck
kyluka
blaiz\-bee
webreaper
yandex
rss
java\/
intute
scirus
mimas
mail.ru
scientificcommons
```

6 References

- Johan Bollen & Rick Luce (2002). Evaluation of Digital Library Impact and User Communities by Analysis of Usage Patterns. *D-Lib Magazine, June 2002, Volume 8 Number 6 (ISSN: 1082-9873)*
(<http://dlib.org/dlib/june02/bollen/06bollen.html>).
- JISC Usage Statistics Review project. *Results from the JISC Usage Statistics Workshop, Berlin, 7-8 July 2008*
(<http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/usagestatisticsreviewreport.pdf>)
- Bollen, Johan & Herbert Van de Sompel (2006). An architecture for the aggregation and analysis of scholarly usage data. *JCDL '06: Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, Chapel Hill, NC, USA. 298-307.
(<http://arxiv.org/abs/cs/0605113v1>)
- The OpenURL Framework for Context-Sensitive Services. ANSI/NISO Standard Z39.88-2004 (2005). National Information Standards Organization, Bethesda, Maryland, U.S.A. (ISSN: 1041-5653)
(http://www.niso.org/standards/resources/Z39_88_2004.pdf)
- Registry for the OpenURL Framework - ANSI/NISO Z39.88-2004
(<http://www.openurl.info/registry>)
- NEEO Technical guidelines, DIDL and MODS application profiles. eContentPlus EC funded NEEO project (2008-2010).
(http://homepages.ulb.ac.be/~bpauwels/NEEO/WP5/WP5_Technical_guidelines.pdf)