Guadaltel

el valor de la diferencia

Nombre del Proyecto

RECOLECTA

Guía presentación de la jornada de estadísticas

Cliente

Fundación española para la ciencia y la tecnología (FECYT).

FECYT

Fecha

27/04/16

Versión

1.01

Tipo de documento

Material de formación

1. Objeto

El objetivo de este documento es servir de guía para la presentación de una jornada para dar unas directrices claras a los administradores de cada repositorio para que concluyan con éxito la implantación y validación de un sistema de generación de estadísticas así como una introducción a la plataforma de explotación de datos.

Para ello se establecen los siguientes puntos a tratar:

- Instalación del módulo proveedor de estadísticas en un repositorio
- Formato y estructura de la información
- Metadatos que conforman el sistema de recolección de estadísticas
- Ejemplos reales
- Plataforma de explotación de estadísticas

2. Instalación del módulo proveedor de estadísticas en un repositorio

El módulo proveedor de estadísticas de uso, es como su propio nombre indica, el encargado de suministrar las estadísticas de uso recogidas en un repositorio. El objetivo de de éste software es recuperar y publicar las estadísticas de uso en un formato estándar para su posterior recolección (previa validación) y explotación por terceros.

El protocolo determinado para la publicación de estadísticas de uso es OAI-PMH con metadataPrefix **OpenURL Context Object Schema** (CTXO), estándar establecido por Knowledge Exchange (KE) que surgió de la cooperación de la danesa Denmark´s Electronic Research Library (DEFF), la alemana Deutsche Forschungsgemeinschaft (DFG), la SURF Foundation (SURF) holandesa y el Joint Information System Comitee (JISC) del Reino Unido; regido por las directrices **KE Usage Statistics**, directrices para la recolección y tratamiento de estadísticas de uso de repositorios en línea. Las KE Usage Statistics son las directrices europeas más importantes y más extendidas en materia de recolección de estadísticas de uso.

Existen varios proveedores alternativos que procederemos a nombrar en el siguiente apartado, pero tras un estudio pormenorizado se concluyó que OAS Data Provider era el que mejor satisfacía las necesidades que FECYT determinó en su momento dada la heterogeneidad de gestores de repositorios existentes en su contexto.

OAS Data Provider es un proyecto alemán, iniciado en Mayo de 2008 para la recolección de estadísticas de uso de documentos alojados en repositorios institucionales. Es un proyecto apoyado por la DFG (Deutsche Forschungsgemeinschaft) Fundación Alemana de Investigación y la DINI (Deutsche Initiative Für NetzwerkInformation) Iniciativa Alemana de Información en Red.

Se instala en la misma máquina donde esté desplegado el repositorio y la interacción con éste se realiza, indirectamente, a través de los logs de Apache, aislándolo así del software de gestión de repositorios elegido.

Se necesitan los siguientes puntos clave:

- El log de Apache debe ser rotado diariamente.
- El log de Apache debe ser generado con un formato determinado.
- Se recomienda que el script de ejecución del OAS Data Provider se incluya en un cron de ejecución diaria para que diariamente se traten los ficheros de log existentes y se extraigan las estadísticas.

El OAS Data Provider lo conforman dos módulos diferentes, un módulo parseador de logs denominado LogFile Parser y un módulo publicador de estadísticas por OAI-PMH y CTXO denominado OAI Data Provider.

El mecanismo o interfaz de interoperabilidad del OAS Data Provider es OAI-PMH como ya se podía intuir. El software publica una URL mediante el protocolo OAI-PMH en la que aparecen los metadatos que describen cada uno de los eventos de uso registrados en CTXO.

2.1. Alternativas

Existen diversos software alternativos para la recopilación de estadísticas de uso y su posterior publicación por OAI-PMH en CTXO. Seguidamente se listan los más notables:

- PIRUS/IRUS-UK. Institutional Repository Usage Statistics (IRUS-UK), es una herramienta cuyo desarrollo ha sido financiado por el JISC para prestar un servicio de recolección de estadísticas de uso a los repositorios del Reino Unido dentro del proyecto RepositoryNet+ (RepNet)1. En su desarrollo ha participado el Mimas (de la Universidad de Manchester) con la colaboración de la Universidad de Cranfield y la Evidence Base (de la Universidad de Birmingham), a partir de los resultados obtenidos en el proyecto PIRUS2.

  PRO: Bajo el amparo de las KE Usage Statistic, cumple con dichas directrices.

  CONTRA: Su instalación es compleja y requiere de conocimientos técnicos avanzados. Para la recuperación de estadísticas de uso requiere de la instalación de un software intermediario denominado “Tracker” cuya configuración depende del software de gestión de repositorios utilizado.

- SURE. Statistics on the Usage of Repositories (SURE). Ha sido promovido por la SURF Foundation dentro de su proyecto SURF Share. Une a las Universidades de Investigación, Universidades de Ciencias Aplicadas e Instituciones de Investigación de los Países Bajos (la Universidad de Leiden, la Universidad de Amsterdam, la Universidad pública de Amsterdam, y la Real Academia Holandesa de Artes y Ciencias entre otras). Todos ellas colaboran en proyectos innovadores para mejorar la calidad de la educación superior y la investigación.

  PRO: SURE utiliza los estándares OAI-PMH, SUSHI, informes COUNTER y OpenURL Context Object para sus desarrollos.

  CONTRA: El sistema se satura cuando se interactúa con un numero elevado de repositorios. Su instalación requiere de conocimientos técnicos.

- AWStats. Es una herramienta de código abierto para la generación de informes de análisis web, apta para analizar datos de servicios de Internet como: servidores web, streaming, de correo electrónico o FTP. Está enmarcada dentro del grupo de las herramientas Web Analyzer. AWStats analiza los archivos de log de los servidores, y en base a ellos produce informes HTML. Los datos son presentados visualmente en tablas de resultados y en gráficos de barra. Pueden crearse informes estáticos mediante una interfaz de línea de comando, y se pueden obtener informes bajo demanda a través de un navegador web.

  PRO: Fácil instalación. No requiere de altos conocimientos técnicos.

  CONTRA: El principal contra que presenta AWStats es que no publica los datos por OAI-PMH.

- OAS Data Provider. Ya ha sido mencionado y explicado en el apartado anterior.

2.2. Requisitos previos

No existen unos requisitos previos determinados por así decirlo. El software OAS Data Provider puede funcionar en cualquier sistema o distribución de Linux que esté capacitada para dar soporte a la tecnología LAMP (**L**inux, **A**pache, **M**ySQL y **P**HP, Perl o Python).

Se necesita un servidor Apache y PHP bajo las siguientes condiciones:

- PHP versión 5.2.x o superior
- Se requiere la extensión DOM de PHP
- Una base de datos MySQL, con permisos para la creación de tablas

Hay que tener en consideración es que la instalación del OAS Data Provider debe llevarse a cabo en la misma máquina donde se encuentre instalado el repositorio del que se quieren recuperar los datos de estadísticas de uso. Por lo tanto si el sistema estaba capacitado para incluir una instalación de un repositorio, lo estará también para incluir la instalación del OAS Data Provider.

Por otro lado, y como hemos dicho antes, OAS Data Provider obtiene la información de los ficheros logs de Apache, por lo tanto habrá que garantizar mediante la configuración oportuna, que la comunicación con el repositorio quede registrada en los logs de Apache.

En términos de memoria física, simplemente hay que garantizar que el sistema cuente con la suficiente memoria para almacenar los ficheros de logs que se van a ir registrando diariamente. Periódicamente es interesante limpiar los logs para evitar un uso de almacenamiento excesivo e innecesario.

OAS Data Provider recoge la información de los logs y la vuelca a la base de datos; desde aquí será publicada para su recolección.

2.3. Instalación

Seguidamente se detallan los pasos más importantes de la instalación del proveedor de estadísticas. La instalación se ha hecho del OAS Data Provider para un repositorio DSpace y el SO elegido es CentOS. Téngase en cuenta que ninguno de los dos son requisitos obligatorios para poder ser incluido en el sistema de recolección de estadísticas.

En concreto el OAS Data Provider interacciona con el repositorio sobre el que se quieren recoger las estadísticas; parseando los logs que se registran en un servidor Apache que debe interceptar todo el tráfico que se genere en el repositorio. Esto hace que OAS Data Provider pueda trabajar con la mayor parte de tipos de gestores de repositorios existentes sólo realizando pequeños ajustes en los scripts de parseo de logs.

2.3.1. Instalación de Apache

Instalamos el servidor de Apache. En CentOS.

```
yum install httpd
```

Editamos el fichero httpd.conf. Este fichero contiene la configuración del log que registra Apache. Se debe establecer el formato del log de Apache de tal forma que registre la información necesaria para representar correctamente cada evento de uso producido.

```
vim /etc/httpd/conf/httpd.conf
```

Sustituimos el atributo LogFormat:

```
[...]
<IfModule log_config_module>
    #
    # The following directives define some format nicknames for use with
    # a CustomLog directive (see below).
    #
    LogFormat "%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{User-Agent}i\"" combined
    LogFormat "%h %l %u %t \"%r\" %>s %b" common
[...]
```

Por el siguiente:

```
[...]
<IfModule log_config_module>
    #
    # The following directives define some format nicknames for use with
    # a CustomLog directive (see below).
    #
    LogFormat "%h %l %u %t \"%r\" %>s %O \"%{Referer}i\" \"%{User-Agent}i\" \"%{JSESSIONID}C\"" combined
    LogFormat "%h %l %u %t \"%r\" %>s %b" common
[...]
```

Lo que se persigue con este cambio es incluir información sobre la sesión del usuario que generó el evento. Esta información podrá ser utilizada para filtrar multiclics y para la detección de usuarios únicos.

Procedemos a configurar el servidor Apache para que intercepte las peticiones al repositorio. Creamos el fichero.

```
touch /etc/httpd/conf.d/dspace.conf
```

Lo editamos incluyendo el siguiente contenido:

```
<VirtualHost *:80>

    ServerAdmin admin@mail.com

    ServerName dspace


    Include conf.d/dspace.redirecciones

</VirtualHost>
```

Creamos el fichero.

```
touch /etc/httpd/conf.d/dspace.redirecciones
```

Lo editamos con el siguiente contenido (aunque dependerá de la configuración de cada repositorio):

```
RewriteRule ^/$ /xmlui [R]


ProxyRequests Off

ProxyPreserveHost On


#########################################################

#########################################################


ProxyPass /xmlui ajp://localhost:8009/xmlui

ProxyPassReverse /xmlui ajp://localhost:8009/xmlui


#########################################################

#########################################################


ProxyPass /oai ajp://localhost:8009/oai

ProxyPassReverse /oai ajp://localhost:8009/oai


ProxyPass /jspui ajp://localhost:8009/jspui

ProxyPassReverse /jspui ajp://localhost:8009/jspui


ProxyPass /solr ajp://localhost:8009/solr
```

```
ProxyPassReverse /solr ajp://localhost:8009/solr
```

Iniciamos el servicio Apache.

```
systemctl start httpd.service
```

Ponemos en la configuración que el servicio se inicie por defecto:

```
systemctl enable httpd.service
```

Pasemos a ver ahora un punto clave dentro de la instalación. Se debe establecer el rotado del log de Apache. El rotado de logs es importante porque se pretende no saturar de ficheros logs el sistema. Con el rotado facilitaremos, por un lado el tratamiento de la información (limitando el tamaño de estos), su organización (pudiendo determinar en cada momento cuales han sido procesados), y por otro lado, el limpiado periódico de información que ya haya sido procesada.

Editamos el fichero /etc/logrotate.d/httpd:

```
/var/log/httpd/*log {
    missingok
    dateext
    size 1k
    copytruncate
    create 0664 apache apache
    rotate 10
    postrotate
    /bin/systemctl reload httpd.service > /dev/null 2>/dev/null || true
    endscript
}
```

La configuración anterior hace que se incluya, en el nombre del fichero de log, información de la fecha, se rote el log cuando su tamaño alcance 1 kilobytes y se determine el número máximo de ficheros de log permitidos antes de proceder a sobrescribirlos.

Esta configuración dependerá de las capacidades del sistema en el que se tiene instalado el repositorio (y por ende el data provider) y del software de parseo de logs utilizado. En nuestro caso, se sigue la **recomendación** de OAS y por ello se establece el parámetro size 1k.

2.3.2. Instalación de PHP

Instalamos PHP

```
yum install php
```

Instalamos los módulos de PHP para MySQL

```
yum install php-mysql
```

Instalamos un módulo más, necesario en CentOS, para la visualización de las páginas con PHP.

```
yum install php-theseer-fDOMDocument.noarch
```

2.3.3. Instalación de OAS Data Provider

Copiamos y descomprimimos los paquetes que conforman la instalación.

```
oai-data-provider -> /var/www/html/

logfile-parser -> /opt/download_dataprovider/
```

Encontramos tres ficheros config.php que tenemos que editar para incluir la configuración del repositorio sobre el que se quiere extraer la información y el path de los ficheros de logs desde los que se extraerá ésta.

```
/opt/download_dataprovider/logfile-parser/config.php

/opt/download_dataprovider/logfile-parser/config-dspace.php

/var/www/html/oai-data-provider/config.php
```

El fichero `/opt/download_dataprovider/logfile-parser/config-dspace.php` es el único dependiente del gestor de repositorio utilizado. Los desarrolladores de OAS lo suministran compatible con repositorios DSpace y WebDoc y por lo tanto requerirá de algunos ajustes para poder ser utilizado en e-Prints, DigiTools o Fedora.

2.3.4. Instalación de MySQL

Instalamos el servidor de base de datos.

```
yum install mysql-server
```

Creamos la base de datos en MySQL y el propietario:

Entramos en MySQL:

```
mysql
```

Creamos el usuario:

```
mysql> create user 'oasuser'@'localhost' identified by 'admin';

Query OK, 0 rows affected (0,00 sec)
```

Creamos las base de datos;

```
mysql> create database oas;

Query OK, 1 row affected (0,00 sec)
```

Le damos permisos de uso y privilegios:

```
mysql> grant usage on *.* to 'oasuser'@'localhost' identified by 'admin';

Query OK, 0 rows affected (0,00 sec)


mysql> grant all privileges on oas.* to 'oasuser'@'localhost';
```

```
Query OK, 0 rows affected (0,00 sec)


mysql> flush privileges;

Query OK, 0 rows affected (0,00 sec)
```

Iniciamos la base de datos. Para ello OAS dispone de un script que realiza su inicialización:

```
cd /opt/download_dataprovider/logfile-parser


php log2ctx.php -O
```

Comprobamos en base de datos que se ha creado la tabla **contextobjects** donde se consolidarán los datos:

```
mysql> show databases;

+------------------------------+

| Database                     |

+------------------------------+

| information_schema           |

| mysql                        |

| oas                          |

| performance_schema           |

+------------------------------+

4 rows in set (0,00 sec)


mysql> use oas;

Reading table information for completion of table and column names

You can turn off this feature to get a quicker startup with -A

Database changed

mysql> show tables;

+------------------------------+

| Tables_in_oas                |

+------------------------------+

| contextobjects               |

+------------------------------+

1 row in set (0,00 sec)
```

NOTA: Para más detalles sobre la instalación se puede acceder aquí.

2.3.5. Consolidación de la información

Llegados a este momento nuestro sistema ya dispondrá de un proveedor de datos estadísticos y por lo tanto podremos consolidar en la base de datos la información de eventos de uso que estarán registrados en los

logs.

Hay que enfatizar en que **la información sobre los accesos registrados que se encuentra en la base de datos del proveedor de datos, es propiedad de los propios repositorios y está en su haber decidir qué quieren hacer con ella.** Es decir, si se desea, se puede explotar dicha información como se desee, pero hay que tener presente que estas estadísticas aún no tienen la consideración de "válidas" puesto que todavía no se han eliminado los eventos producidos por robots webs ni filtrado los multiclics.

3. Formato y estructura de la información

3.1. Descripción de los formatos

La recolección de estadísticas de acceso recogidas en cada uno de los repositorios que participan en Recolecta, se realiza desde el servidor central de recolección perteneciente a FECYT, basándose en la recolección de metadatos OpenURL Context Object (CTXO) a través del protocolo OAI-PMH (en su versión 2.0).

La recolección se hará mediante una URL de recolección que cada repositorio debe facilitar. Esta puede ser distinta a la utilizada para recolección de metadatos en Dublin Core. Por aclarar este punto. El requisito es que si se realiza la petición `?verb=ListMetadataFormats` ya sea a la URL del repositorio o a la URL de recolección de estadísticas, la respuesta debe contener la siguiente información:

```
<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>
```

Donde el valor del campo:

metadataPrefix, corresponde al prefijo que identifica los metadatos OpenURL Context Object.

schema, corresponde al esquema oficial utilizado para los metadatos OpenURL Context Object.

metadataNamespace, corresponde al namespace utilizado para los metadatos OpenURL Context Object.

Es decisión de los gestores de los repositorios el incluir el metadataPrefix ctxo entre los metadatos soportados por su herramienta de gestión de repositorios o determinar una interfaz diferente que sólo suministre el nuevo metadataPrefix. La experiencia nos dice que no es trivial incluir el nuevo metadataPrefix entre los disponibles en la propia herramienta de gestión del repositorio; y por lo tanto NO siendo un requisito obligatorio, no siempre merece la pena intentarlo.

Las respuestas a las peticiones que se hagan a través de la URL de recolección deberán ser ficheros con formato estructurado (formato XML) con las siguientes consideraciones:

Es **obligatorio** el uso de la codificación UTF-8 con los caracteres en Unicode. Esta recomendación debe estar muy presente en los administradores puesto que un carácter que no se encuentre en UTF-8 provocaría, por el propio formato de los datos, un error en la sintaxis de la respuesta siendo imposible procesar el contenido. El incumplimiento de este requisito puede derivar en errores como el siguiente:

```
Error de lectura XML: mal formado Ubicación: https://repositorio.pruebas.es Número de línea 6, columna 18:

        <body>NO COMETER MÓS ERRORES</body>
----------------------^
```

Es **recomendable** utilizar una URL para publicación de estadísticas que no sea susceptible de ser modificada, es decir, que sea independiente del lugar donde se encuentre desplegado y que pueda ser utilizada aunque cambie el servicio de ubicación.

La utilización del software OAS Data Provider garantiza el cumplimiento de todos los aspectos antes descritos.

4. Metadatos que conforman el sistema de recolección de estadísticas

4.1. Directrices sobre los metadatos

Para asegurar unos umbrales de **calidad mínima de los datos de estadísticas** de acceso que permitan una comparación fiable entre los datos recogidos, se han creado unas **directrices internacionales** que intentan reglar los metadatos con los que representar cada uno de los eventos de uso.

Estas directrices se conocen como directrices Knowledge Exchange (en adelante KE). A continuación se detallan todos los **metadatos indicados por KE:**

Parámetro		Uso	Descripción
context-object	**timestamp**	Obligatorio	Fecha y hora en que tuvo lugar el evento de uso. Debe estar en formato ISO8601. Por ejemplo: 2009-07-29T08:15:46 +01:00
	identifier	Opcional	No existe un formato dado para el identificador (Usage Event ID), pero de usarse debe estar codificado y ser único. En los Países Bajos por ejemplo se utiliza un hash MD5 que se genera a partir de una concatenación del código de la institución, el identificador de la publicación y la fecha y la hora.
referent	**identifier**	Obligatorio	Identificador URI de la publicación. Tras el URI de una publicación puede estar tanto la URL del archivo físico (Full Text) como la URL de los metadatos que se solicitaron. Por ejemplo: https://openaccess.leidenuniv.nl/bitstream/1887/12100/1/Thesis.pdf Opcionalmente si aplica, se puede incluir un identificador DOI adicional. Por ejemplo: http://hdl.handle.net/10272/6
referringEntity	**identifier**	Obligatorio si aplica	La dirección URL que redirigió al usuario hasta la petición actual.
requester	**identifier**	Opcional	Dirección IP cifrada desde donde se realizó la petición.
	spatial	Opcional	Código en formato ISO 3166-1-alpha-2 del país desde el que se originó la solicitud. Por ejemplo: ES
	hashed-c	Opcional	Máscara de subred desde donde se realizó la petición. Al encontrarse la IP cifrada, la máscara de subred puede ser usada para obtener el país origen.
	hashed-session	Opcional	Cifrado de la sesión de usuario que realizó la petición. Es optativo pero recomendable para el filtrado de multiclics.
	classification	Opcional	Clasificación del usuario (internal, administrative, institutional), en caso de ser posible.
	user-agent	Opcional	Cadena completa del cliente HTTP utilizado.
service-type	**dcterms:type**	Obligatorio	Provee información sobre el tipo de petición realizada por el usuario, descarga de fichero o lectura de metadatos.
resolver	**identifier**	Obligatorio	URL base del repositorio que contiene el ítem que ha sido accedido.
referrer	**Identifier**	Opcional	La identificación del servicio exterior que haya proporcionado la reseña al elemento, por ejemplo, un motor de búsqueda. Debe estar conformado según el formato http://info-uri.info/registry/OAIHandler?verb=GetRecord&metadataPrefix=reg&identifier=info:sid/

A partir de las anteriores, se establecen las directrices de Recolecta para la recolección de estadísticas. **Estas directrices deben ser seguidas en su totalidad, para conseguir cumplir con la validación que realiza FECYT antes de proceder a la recolección de los datos de estadísticas de uso.**

Sobre el protocolo OAI-PMH

Regla	Tipo	Descripción	Descripción del error	Comentarios
OAI-PMH válida	Obligatorio	El formato del XML debe ser válido conforme a su XSD.	El formato del protocolo OAI-PMH no es válido conforme a su XSD.	Sin el formato correcto, los recolectores no pueden extraer el contenido de los registros incluidos en la petición.
Formato de metadatos	Obligatorio	Debe soportar el metadataPrefix CTXO.	El formato de metadatos con prefijo CTXO no es soportado.	El formato con prefijo CTXO es el utilizado para la recolección de estadísticas de acceso.
Uso de sets	Opcional	El protocolo puede soportar opcionalmente la inclusión de sets para filtrar los resultados.	No está soportado el uso de sets en su repositorio.	Pueden utilizarse los conjuntos para realizar la recolección selectiva de los registros.
Política de borrado	Obligatorio	La política de eliminación del repositorio sólo puede ser “transient” o “no”.	El formato de eliminar estadísticas con el modo “persistent” no está permitido.	Otra política de eliminación no está permitida.
Inclusión de context-objects	Obligatorio	Los objetos estadísticos deben estar incluidos dentro de los metadatos en un objeto global <context-objects>	Los metadatos de estadísticas no están englobados en un objeto <context-objects>	En correspondencia con la definición de CTXO como formato de datos para el intercambio de las estadísticas a través del protocolo OAI-PMH; la incorporación de registros se debe hacer conforme a su XSD, englobando en un objeto <context-objects> que contenga todas las estadísticas de acceso para un registro. Opcionalmente se puede crear un objeto <context-objects> por cada evento.
ListMetadataFormats debe contener CTXO	Obligatorio	El protocolo OAI-PMH debe publicar en su listado de metadatos aceptados el de CTXO.	No se encuentra el formato de metadatos CTXO.	El formato de metadatos de CTXO debe aparecer entre los disponibles en la petición verb=ListMetadataFormats incluyendo el XSD y el namespace correspondientes.

Sobre los metadatos OpenUrl Context Object

Regla	Tipo	Descripción	Descripción del error	Comentarios
Timestamp del context-object	Obligatorio	El atributo timestamp es obligatorio y debe seguir el formato ISO 6801.	El campo timestamp no se encuentra definido o se encuentra definido conforme a un formato que no es el aceptado.	La ausencia o formato defectuoso del atributo timestamp impide detectar en el momento en el que se produjo la petición, imposibilitando detectar si se trata de una petición válida o no.
EJEMPLO	`<context-object timestamp="2013-11-15T07:26:41+00:00" ...>`			
Identifier del context-object	Opcional	Es recomendable incluir un identificador único de context-object basado en una concatenación del código de la institución, el identificador de la publicación y la fecha y hora.	El atributo identifier de la etiqueta context-object no se encuentra definido.	El atributo identifier del campo context-object permite identificar repeticiones erróneas por fallo del software de captura de estadísticas. Este atributo ayuda en la detección de duplicados, aunque si no aparece se puede detectar mediante la fecha de acceso y la ip/sesión del usuario.
EJEMPLO	`<context-object ... identifier="b06c0444f37249a0a8f748d3b823ef20">`			
referent:identifier	Obligatorio	El atributo referent:identifier es obligatorio, único y debe coincidir con el URI único del documento en el repositorio.	El campo referent:identifier no se encuentra definido o aparece repetido.	La ausencia del campo referent:identifier no permite conocer el objeto digital al que se encuentra asociada la petición. Asimismo, este identificador debe coincidir con el URI único que tiene definido el ítem en el protocolo OAI-PMH para permitir el enriquecimiento de metadatos y la posterior explotación de los datos.

EJEMPLO	`<referent>`
	`<identifier>http://dlib.org/dspace/handle/123456789/6</identifier>`
	`</referent>`			

referringEntity:identifier	Obligatorio si aplica	Es recomendable incluir el campo referringEntity indicando la URL a través de la cual se accedió al objeto digital o a sus metadatos.	No se encuentra definido el campo referringEntity.	El atributo referringEntity permite identificar el lugar a través del cual el usuario ha accedido al recurso (portal de acceso, buscadores, etc.). Permite obtener estadísticas de acceso de usuario a través de buscadores a los que se encuentra conectado el repositorio.

EJEMPLO	`<referring-entity>`
	`<identifier>http://dlib.org/dspace/handle/123456789/3</identifier>`
	`</referring-entity>`			

Uso de las extensiones de DINI	Obligatorio	El atributo ctx:requester/ctx:metadata-by-val/ctx:format debe contener el valor de las extensiones DINI ("http://dini.de/namespace/oas-requesterinfo")	El formato de los metadatos CTXO no incluye las extensiones DINI para incluir información adicional a las peticiones.	La inclusión de las extensiones de DINI, contempladas en la normativa de KE (Knowledge Exchange) permite incluir información adicional necesaria para el correcto filtrado de las peticiones. Para definir estas extensiones, debe incluir el campo ctx:requester/ctx:metadata-by-val/ctx:format con el valor ("http://dini.de/namespace/oas-requesterinfo").
EJEMPLO	<requester> <identifier>data:,2e9463d760b56822bc69388fe5edea825c4f644a585b</identifier> <metadata-by-val> <format>http://dini.de/namespace/oas-requesterinfo</format> ... </metadata-by-val> </requester>			
requester:identifier debe ser un hash.	Obligatorio	El atributo requester:identifier deberá ser obligatorio, único y deberá encontrarse cifrado mediante un algoritmo de hash (MD5, SHA-1, etc.) con el formato **data:,hash(32 bits)**. Durante el cálculo del hash se debe usar un salteado de al menos 12 caracteres.	El campo requester:identifier no se encuentra definido, aparece múltiples veces o no se encuentra conforme al formato establecido.	**¿Por qué debe ser obligatorio y no opcional como indica KE?** La ausencia del campo requester:identifier no permitiría identificar al usuario que realizó la petición, imposibilitando el posterior filtrado de multiclics que se pretende llevar a cabo. Asimismo, debido a las leyes existentes sobre la protección de datos, la ip del usuario no puede ser enviada en texto plano o cifrada en un algoritmo simétrico que permita ser recuperada. Se recomienda el uso de funciones de hash (MD5, SHA1, etc).
EJEMPLO	<requester> <identifier>data:,2e9463d760b56822bc69388fe5edea825c4f644a585b</identifier> ... </requester>			
requester:spatial	Opcional	Es recomendable incluir el elemento spatial indicando el código del país de origen de la petición en formato ISO 3166-1-alpha-2.	El atributo spatial del requester no se encuentra definido o no está conforme al vocabulario.	El atributo spatial permite identificar el país de origen de la petición para obtener estadísticas de acceso por países.
EJEMPLO	<requester> <identifier>data:,2e9463d760b56822bc69388fe5edea825c4f644a585b</identifier> <metadata-by-val> <format>http://dini.de/namespace/oas-requesterinfo</format> <metadata> <requesterinfo> ... <spatial>ES</spatial> ... </metadata-by-val> </requester>			

requester:hashed-c	Opcional	Es recomendable incluir la máscara de subred, formado por la IP a la que se le ha cambiado el último dígito por un .0	No se encuentra definido el campo hashed-c.	La máscara de subred desde donde se realizó la petición. Al encontrarse la IP cifrada, la máscara de subred puede ser usada para obtener el país origen.
EJEMPLO	<requester> <identifier>data:,2e9463d760b56822bc69388fe5edea825c4f644a585b</identifier> <metadata-by-val> <format>http://dini.de/namespace/oas-requesterinfo</format> <metadata> <requesterinfo> ... <hashed-c>192.168.64.0</hashed-c> ... </metadata-by-val> </requester>			
dini:hashed-session	Obligatorio si aplica	ctx:context-object/ctx:requester/ctx:metadata/dini:requesterinfo/dini:hashed-session es obligatorio siempre que sea aplicable y debe contener la sesión del usuario que ha realizado la petición cifrada mediante un algoritmo de hash.	El campo hashed-session no se encuentra definido.	**¿Por qué debe ser obligatorio y no opcional como indica KE?** La ausencia del campo requester:hashed-session no permite identificar al usuario que realizó la petición, imposibilitando el posterior filtrado de multiclics que se pretende llevar a cabo. Es obligatorio si es aplicable porque si se trata de la primera petición de un usuario, ésta no incluirá dicha información. Asimismo, debido a la protección de datos, la sesión del usuario no puede ser enviada en texto plano o cifrada en un algoritmo simétrico que permita ser recuperada. Se recomienda el uso de funciones de hash (MD5, SHA1, etc).
EJEMPLO	<requester> <identifier>data:,2e9463d760b56822bc69388fe5edea825c4f644a585b</identifier> <metadata-by-val> <format>http://dini.de/namespace/oas-requesterinfo</format> <metadata> <requesterinfo> ... <hashed-session>data:,b64841a7d43483f3c742444fd07</hashed-session> ... </metadata-by-val> </requester>			

requester:classification conforme a vocabulario	Opcional	Es recomendable incluir la clasificación del usuario.	No se encuentra definido el campo classification o no se encuentra conforme al vocabulario.	Se permiten 3 valores: * **internal**: accesos de herramientas internas (testeadores de servicio, etc.). * **administrative**: accesos por administradores o personal técnico por motivos de mantenimiento (testing, etc.). * **institutional**: accesos al servicio desde la institución donde reside el repositorio en labores de administración.
EJEMPLO	<requester> <identifier>data:,2e9463d760b56822bc69388fe5edea825c4f644a585b</identifier> <metadata-by-val> <format>http://dini.de/namespace/oas-requesterinfo</format> <metadata> <requesterinfo> ... <classification>institutional</classification> ... </metadata-by-val> </requester>			
dini:user-agent	Obligatorio	El atributo ctx:context-object/ctx:requester/ctx:metadata/dini:requesterinfo/dini:classification/dini:user-agent es obligatorio y único, y debe contener la cadena de identificación del agente HTTP utilizado para la petición.	El campo user-agent no se encuentra definido.	**¿Por qué debe ser obligatorio y no opcional como indica KE?** El campo user-agent identifica al agente web que realizó las peticiones al repositorio. La no inclusión de este metadato impide identificar si la petición ha sido realizada por un usuario o por un robot web. El recolector de estadísticas realiza, además del filtro de multiclics, el filtro de peticiones procedentes de robots webs.
EJEMPLO	<requester> <identifier>data:,2e9463d760b56822bc69388fe5edea825c4f644a585b</identifier> <metadata-by-val> <format>http://dini.de/namespace/oas-requesterinfo</format> <metadata> <requesterinfo> ... <user-agent>Mozilla/5.0 (Linux) Firefox/15.0.1</user-agent> ... </metadata-by-val> </requester>			

dcterms:format conforme a vocabulario	Obligatorio	El atributo ctx:service-type/ctx:metadata-by-val/ctx:metadata/dcterms:format es obligatorio, único y debe contener el valor conforme al vocabulario.	El campo dcterms:format no se encuentra definido o el valor no está conforme al vocabulario.	El campo dcterms:format identifica de forma taxativa si el acceso se ha realizado a los metadatos o al objeto digital a texto completo. Distinguir el tipo de acceso utilizado es importante a la hora de realizar las estadísticas.
EJEMPLO	<service-type> <metadata-by-val> <format>http://dublincore.org/documents/2008/01/14/dcmi-terms/</format> <metadata> <dcterms:format>info:eu-repo/semantics/descriptiveMetadata</dcterms:format> </metadata> </metadata-by-val> </service-type>			
resolver:identifier	Obligatorio	El atributo ctx:resolver/ctx:identifier es obligatorio y único.	El campo identifier del resolver no se encuentra definido.	El campo identifier identifica el repositorio responsable donde se encuentra el objeto digital. Este campo es necesario para definir las estadísticas de acceso por repositorio.
EJEMPLO	<resolver> <identifier>http://dlib.org/dspace</identifier> </resolver>			
referrer:identifier	Opcional	Es recomendable incluir el campo referrer identificando el servicio exterior que haya proporcionado la reseña al elemento. Debe estar indicado conforme al formato info:sid	No se encuentra definido el campo referrer.	El atributo referrer permite identificar el lugar a través del cual el usuario ha accedido al recurso (portal de acceso, buscadores, etc.). Permite obtener estadísticas de acceso de usuario a través de buscadores a los que se encuentra conectado el repositorio.
EJEMPLO	<referrer> <identifier>info:sid/dlib.org:dlib</identifier> </referrer>			

Cualquier repositorio que cumpla lo anterior podrá ser recolectado por el almacén central de FECYT.

Veamos ahora un registro completo:

```
<context-objects xsi:schemaLocation="info:ofi/fmt:xml:xsd:ctx
http://www.openurl.info/registry/docs/xsd/info:ofi/fmt:xml:xsd:ctx">
  <context-object timestamp="2013-11-15T07:26:41+00:00" identifier="b06c0444f37249a0a8f748d3b823ef20">
    <administration>
      <oa-statistics>
        <status_code>200</status_code>
        <size>810</size>
        <format>text/html</format>
        <service>http://repositorio.es/dspace</service>
      </oa-statistics>
    </administration>
    <referent>
      <identifier>http://dlib.org/dspace/handle/123456789/6</identifier>
    </referent>
    <referring-entity>
      <identifier>http://dlib.org/dspace/handle/123456789/3</identifier>
    </referring-entity>
    <requester>
      <identifier>data:,2e9463d760b56822bc69388fe5edea825c4f644a585b</identifier>
      <metadata-by-val>
        <format>http://dini.de/namespace/oas-requesterinfo</format>
        <metadata>
          <requesterinfo>
            <hashed-c>192.168.64.0</hashed-c>
            <hostname>fecyt.es</hostname>
            <hashed-session>data:,b64841a7d43483f3c742444fd07</hashed-session>
            <user-agent>Mozilla/5.0 (Linux) Firefox/15.0.1</user-agent>
            <spatial>ES</spatial>
            <classification>institutional</classification>
          </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/descriptiveMetadata</dcterms:format>
        </metadata>
      </metadata-by-val>
    </service-type>
    <resolver>
      <identifier>http://dlib.org/dspace</identifier>
    </resolver>
    <referrer>
      <identifier>info:sid/dlib.org:dlib</identifier>
    </referrer>
  </context-object>
</context-objects>
```

5. Plataforma de explotación de estadísticas

5.1. Sistema de recolección de estadísticas

El sistema de recolección de estadísticas del proyecto Recolecta de FECYT lo conforman tres elementos principales:

- ***Validador de estadísticas,*** sistema validador de directrices KE y directrices de FECYT. Está basado en el software de validación DRIVER, al que se le incorporó toda la lógica necesaria para que además de la actual validación sobre Dublin Core, también se realice la validación CTXO.
- ***Recolector de estadísticas*** o almacén central de estadísticas (Central Clearing House en adelante CCH). Sin quitarle importancia al fin principal de la plataforma, que no es otro que la publicación para su explotación de las estadísticas recolectadas; el almacén central de estadísticas (CCH) es sin duda una parte muy importante dentro del sistema de recolección de estadísticas. Como ya se ha comentado, se trata del elemento del sistema desde el que se realizará la recolección y el tratamiento de los datos relacionados con los eventos de uso: filtros de multiclics, filtro de robots, unificación de los accesos que se realizan a documentos que pueden estar alojados en varios repositorios.

  Su implementación está basada en el software alemán OAS Service Provider (desarrollado bajo el mandato de las KE Usage Statistics como puede suponerse).
- ***Portal de estadísticas,*** portal basado en DRUPAL que publica las estadísticas de uso recolectadas desde los diferentes repositorios que participan del proyecto Recolecta. Muestra información referente a: visitas, ratio visitas/documentos, visitas únicas, visitas de usuarios únicos, descargas, ratio descargas/documentos, descargas únicas y descargas de usuarios únicos.

5.1.1. Etapas de la recolección

Cuando un repositorio pretende que sus estadísticas de uso sean recolectadas por FECYT, el repositorio debe cumplir con el trámite previo de pasar la validación de directrices y recomendaciones. Para ello FECYT cuenta con una herramienta que realiza la validación de este aspecto.

Podemos identificar las siguientes etapas en este proceso:

1. **Registro.** Como paso inicial el administrador de un repositorio o la persona responsable deberá proceder al registro de la de URL de recolección de estadísticas de uso en el sistema.
2. **Validación.** Seguidamente se procederá a la validación de la interfaz de estadísticas de uso.
3. **Obtención de resultados e interpretación.** Una vez terminado el proceso de validación, se deberán interpretar los resultados obtenidos.
4. **Corrección de las deficiencias.** Etapa de correcciones sobre la interfaz de estadísticas de uso. Si no se ha cumplido con las directrices establecidas habrá que realizar ajustes en el proveedor de datos de estadísticas.
5. **Repetición de la validación.** Etapa de repetición de la validación para el caso de no haber sido calificado como APTO anteriormente.
6. **Solicitud de recolección.** Una vez se obtiene la calificación de APTO, se podrá solicitar la recolección de las estadísticas de uso.

5.1.2. Validación de estadísticas

La etapa de validación de estadísticas se divide a su vez en tres etapas. En cada una de éstas se debe obtener la valoración como APTO para que el repositorio sea considerado APTO para su recolección:

- Validación de estadísticas de acuerdo con las directrices RECOLECTA (directrices KE + directrices Recolecta). Durante esta etapa se validan cada uno de los registros que representan a las estadísticas de uso que publica el proveedor de estadísticas del repositorio. Serán válidos si:
  - Existe el atributo timestamp (Obligatorio)
  - Existe el atributo identifier del context-object (Recomendado)
  - Existe el campo referent:identifier (Obligatorio)
  - Existe el campo referringEntity:identifier (Recomendado)
  - El atributo ctx:requester/ctx:metadata-by-val/ctx:format debe contener el valor de las extensiones DINI (Obligatorio)
  - Existe el atributo requester:identifier (Obligatorio)
  - El atributo requester:identifier debe ser un hash (Obligatorio)
  - Existe el atributo requester:spatial (Recomendado)
  - Existe el campo requester:hashed-c (Recomendado)
  - Existe el campo dini:hashed-session y es único (Obligatorio)
  - El campo dini:hashed-session debe ser un hash (Obligatorio)
  - El campo requester:classification se encuentra conforme al vocabulario (Recomendado)
  - Existe el campo dini:user-agent y es único (Obligatorio)
  - Existe el campo dcterms:format, es único y se encuentra conforme al vocabulario (Obligatorio)
  - El campo dcterms:format se encuentra conforme al vocabulario (Obligatorio)
  - Existe el campo resolver:identifier y es único (Obligatorio)
  - Existe el campo referrer:identifier (Recomendado)
- Validación del protocolo OAI-PMH. Que no es más que un conjunto de operaciones encaminas a determinar y valorar la correcta implementación del protocolo OAI-PMH considerando las diferentes cláusulas/sentencias que admite para la parametrización de las peticiones que puede resolver.
- Validación específica de DRIVER sobre el protocolo OAI-PMH. Que consiste en la comprobación de la obligación de cumplir las directrices que establece DRIVER sobre el protocolo OAI-PMH.
  - Es obligatorio utilizar el formato de metadatos CTXO
  - Implementación de una estrategia de eliminación: transitoria o nula
  - Dirección de correo electrónico del administrador del repositorio válida

Si durante la etapa de validación el repositorio obtiene la calificación de APTO, se procederá a su alta en el recolector de estadísticas.

5.1.3. Ejemplos reales de validación

5.1.3.1. Repositorio e-Archivo UC3M

5.1.3.1.1. Resultado de la validación

El repositorio obtiene la valoración de ***NO APTO*** puesto que obtiene los siguientes resultados:

- Validación de metadatos CTXO: ***NO APTO****

  Han sido validados: 368316 registros

Regla	Resultado (cumplimiento)	Observaciones
Existe el atributo timestamp	100%	
Existe el atributo identifier del context-object (Recomendado)	0%	El atributo identifier de la etiqueta contextobject no se encuentra definido. El atributo identifier del campo contextobject permite identificar repeticiones erróneas por fallo del software de captura de estadísticas. Este atributo ayuda en la detección de duplicados, aunque si no aparece se puede detectar mediante la fecha de acceso y la ip/sesión del usuario.
Existe el campo referent:identifier (Obligatorio)	100%	
Existe el campo referringEntity:identifier (Recomendado)	41%	El atributo referringEntity permite identificar el lugar a través del cual el usuario ha accedido al recurso (portal de acceso, buscadores, etc.). Permite obtener estadísticas de acceso de usuario a través de buscadores a los que se encuentra conectado el repositorio. Indica la URL de la entidad a través de la cual se accedió al objeto digital o a sus metadatos.
El atributo ctx:requester/ctx:metadata-by-val/ctx:format debe contener el valor de las extensiones DINI (Obligatorio)	100%	
Existe el atributo requester:identifier (Obligatorio)	100%	
El atributo requester:identifier debe ser un hash (Obligatorio)	100%	
Existe el atributo requester:spatial (Recomendado)	0%	El atributo spatial del requester no se encuentra definido o no está conforme al vocabulario. El atributo spatial permite identificar el país de origen de la petición para obtener estadísticas de acceso por países.
Existe el campo requester:hashed-c (Recomendado)	100%	
Existe el campo dini:hashed-session y es único (Obligatorio)	100%	
El campo dini:hashed-session debe ser un hash (Obligatorio)	100%	
El campo requester:classification se encuentra conforme al vocabulario (Recomendado)	0%	No se encuentra definido el campo classification o no se encuentra conforme al vocabulario.
Existe el campo dini:user-agent y es único (Obligatorio)	99% Sólo falta en 1099 registros	El campo user-agent no se encuentra definido. El campo user-agent identifica al agente web que realizó las peticiones al repositorio. La no inclusión de este metadato impide identificar si la petición ha sido realizada por un usuario o por un robot web, no pudiendo ser filtrada correctamente.
Existe el campo dcterms:format, es único y se encuentra conforme al vocabulario (Obligatorio)	100%	
El campo dcterms:format se encuentra conforme al vocabulario (Obligatorio)	100%	
Existe el campo resolver:identifier y es único (Obligatorio)	100%	
Existe el campo referrer:identifier (Recomendado)	0%	Es recomendable incluir el campo referrer identificando el servicio exterior que haya proporcionado la reseña al elemento. Debe estar indicado conforme al formato info:sid

* Se indica en rojo el incumplimiento causante de la calificación como NO APTO.

- Validación OAI-PMH: ***APTO***
- Validación específica DRIVER sobre OAI-PMH: ***APTO***

5.1.3.2. Repositorio URJC

5.1.3.2.1. Resultado de la validación

El repositorio obtiene la valoración de ***APTO*** puesto que obtiene los siguientes resultados:

- Validación de metadatos CTXO: ***APTO***

  Han sido validados: 224 registros

Regla	Resultado (cumplimiento)	Observaciones
Existe el atributo timestamp	100%	
Existe el atributo identifier del context-object (Recomendado)	0%	El atributo identifier de la etiqueta contextobject no se encuentra definido. El atributo identifier del campo contextobject permite identificar repeticiones erróneas por fallo del software de captura de estadísticas. Este atributo ayuda en la detección de duplicados, aunque si no aparece se puede detectar mediante la fecha de acceso y la ip/sesión del usuario.
Existe el campo referent:identifier (Obligatorio)	100%	
Existe el campo referringEntity:identifier (Recomendado)	93%	El atributo referringEntity permite identificar el lugar a través del cual el usuario ha accedido al recurso (portal de acceso, buscadores, etc.). Permite obtener estadísticas de acceso de usuario a través de buscadores a los que se encuentra conectado el repositorio. Indica la URL de la entidad a través de la cual se accedió al objeto digital o a sus metadatos.
El atributo ctx:requester/ctx:metadata-by-val/ctx:format debe contener el valor de las extensiones DINI (Obligatorio)	100%	
Existe el atributo requester:identifier (Obligatorio)	100%	
El atributo requester:identifier debe ser un hash (Obligatorio)	100%	
Existe el atributo requester:spatial (Recomendado)	0%	El atributo spatial del requester no se encuentra definido o no está conforme al vocabulario. El atributo spatial permite identificar el país de origen de la petición para obtener estadísticas de acceso por países.
Existe el campo requester:hashed-c (Recomendado)	100%	
Existe el campo dini:hashed-session y es único (Obligatorio)	100%	
El campo dini:hashed-session debe ser un hash (Obligatorio)	100%	
El campo requester:classification se encuentra conforme al vocabulario (Recomendado)	0%	No se encuentra definido el campo classification o no se encuentra conforme al vocabulario.
Existe el campo dini:user-agent y es único (Obligatorio)	100%	
Existe el campo dcterms:format, es único y se encuentra conforme al vocabulario (Obligatorio)	100%	
El campo dcterms:format se encuentra conforme al vocabulario (Obligatorio)	100%	
Existe el campo resolver:identifier y es único (Obligatorio)	100%	
Existe el campo referrer:identifier (Recomendado)	0%	Es recomendable incluir el campo referrer identificando el servicio exterior que haya proporcionado la reseña al elemento. Debe estar indicado conforme al formato info:sid

- Validación OAI-PMH: ***APTO***
- Validación específica DRIVER sobre OAI-PMH: ***APTO***

5.1.3.3. Conclusiones

Aunque no se cuenta con una muestra lo suficiente amplia de repositorios como para poder extraer unas conclusiones finales, podemos aseverar que el nivel de cumplimiento de las directrices de obligado cumplimiento es bueno en general:

- Sólo en el caso del repositorio de estadísticas de e-Archivo de la UC3M se incumple en una pequeñísima parte de los registros la directriz obligatoria que indica que “debe existir el campo dini:user-agent y que además debe ser único”. Este valor es muy importante puesto que el campo user-agent identifica al agente web que realizó las peticiones al repositorio. La no inclusión de este metadato impide identificar si la petición ha sido realizada por un usuario o por un robot web, no pudiendo ser filtrado correctamente el evento de uso registrado. Es decir, esta directriz pretende sobre todo evitar que las estadísticas registradas puedan ser alteradas, sea intencionalmente o no, por sistemas automáticos.

En el caso de las reglas que atienden a las recomendaciones sí se observa en los resultados que necesitan de más ajustes para conseguir que se consiga una puntuación de 100 en la validación:

- Podemos comprobar que ningún repositorio cumple con la directriz recomendada que indica que “debe existir el atributo identifier del context-object”. El atributo identifier del campo contextobject permite identificar repeticiones erróneas por fallo del software de captura de estadísticas. Este atributo ayuda en la detección de duplicados, aunque si no aparece se puede detectar mediante la fecha de acceso y la ip/sesión del usuario.
- También podemos ver que en todos los casos existen registros que no incluyen información sobre el referringEntity, que permite identificar, mediante la URL, a la entidad a través de la cual se accedió al objeto digital o a sus metadatos. Téngase en cuenta que el propio repositorio que incluye al recurso puede aparecer en este metadato si fuese el caso.
- Con el metadato spatial del requester ocurre que para ningún registro se encuentra definido.
    - El atributo spatial permite identificar el país de origen de la petición para obtener estadísticas de acceso por países.
- No existe ningún registro con el metadato classification del requester. El atributo classification permite clasificar el tipo de acceso y consigna tres valores si se produce alguno de los siguientes supuestos:
    - internal: accesos de herramientas internas (testeadores de servicio, etc.).
    - administrative: accesos por administradores o personal técnico por motivos de mantenimiento (testing, etc.).
    - institutional: accesos al servicio desde la institución donde reside el repositorio en labores de administración.
- Por último podemos observar también que ningún registro incluye información respecto al metadato referrer que permite identificar al servicio exterior que ha proporcionado la reseña al recurso. Los valores de este metadatos deben estar indicados conforme al formato info:sid (explicación). Su omisión debe entenderse como que el evento no procede de ningún servicio exterior.

5.1.4. Recolección de estadísticas

Previamente a la ejecución del proceso de recolección de estadísticas hay que establecer el modo de recolección: puntual o periódico; así como el rango de fechas que se quiere recolectar si es que no se desea realizar la recolección completa de las estadísticas del repositorio. Hay que hacer un inciso para aclarar que el sistema de recolección de estadísticas registra la fecha de la última fecha recolectada para, por defecto, continuar desde dicha fecha en una futura recolección.

En caso de establecer una recolección periódica, además del rango del fechas, estará disponibles dos opciones que permiten definir:

- Intervalo entre cada recolección
- Tiempo máximo que durará la recolección

Los intervalos de tiempo se expresan en formato PHPs DateInterval. Por ejemplo:

- para 5 minutos PT5M
- para 2 días P2D
- para 6 horas y 5 minutos es PT6H5M
- para 1 día y 2 horas es P1DT2H

Cuando termina la ejecución de la recolección, la información recolectada habrá sido consolidada en la base de datos del sistema recolector por lo que a partir de ese momento el sistema recolector no dependerá de la disponibilidad o no de los data providers de los repositorios para continuar.

Será en una segunda fase, el proceso de análisis, cuando toda la información será verificada con objeto de detectar y omitir aquellos que no cumplan con los requisitos establecidos. Es decir, si fuese el caso de una recolección de un repositorio que no obtuvo la valoración de APTO para los metadatos de estadísticas, en esta etapa esos registros incumplidores serán ignorados.

El proceso de análisis también es configurable para ser ejecutado de manera puntual o periódica, al igual que podrá ser ejecutado para un sólo data provider o para todos los existentes.

5.1.5. Calculo de estadísticas y enriquecimiento

Durante el proceso de cálculo de estadísticas se generan las estadísticas definitivas para cada recurso sobre el que se han recolectado datos. Por lo tanto, durante este proceso se agregará la información procedente de los diferentes data providers (supóngase un recurso que tenga estadísticas de uso recogidas de diferentes repositorios), se detectarán los multiclics y se filtrarán los accesos que hayan sido producidos por robots webs. Para la detección de multiclics se determina el siguiente umbral que fue establecido por COUNTER (iniciativa predecesora y alternativa a KE).

Propuesta	Tiempo
COUNTER	Sólo se consideran eventos diferentes si al menos transcurren entre ellos: · 10 segundos para recursos HTML · 30 segundos para recursos PDF *En el supuesto que tuvieran la misma IP y/o la misma sesión.*

Para ello necesitamos que disponer información sobre la IP y la sesión del usuario que generó el evento. De ahí la importancia de incluir dentro de los `<contextobject>` que representan a los eventos de uso, los metadatos que permitirán su identificación: `<requester:identifier>` y `<dini:hashed-session>`

El proceso de cálculo de estadísticas puede ejecutarse parcialmente, para ello se puede establecer un rango de fechas para el que calcular las estadísticas. También se pueden calcular las estadísticas para un único recurso si se conoce su identificador.

Tras la finalización de este cálculo, la información ya podrá verse reflejada parcialmente en el portal de estadísticas.

Llegados a este punto, se debe ejecutar el proceso de enriquecimiento con metadatos. Este proceso representa un punto clave para el sistema de recolección de estadísticas puesto que en este proceso se fusiona la información procedente del sistema de recolección de estadísticas, junto con la información procedente del sistema recolector de producción (metadatos). Por lo tanto, previa a la ejecución de este proceso, debemos asegurar que se ha llevado a cabo la recolección (metadatos) del repositorio en cuestión.

Existe un último proceso disponible en el sistema recolector que generará las estadísticas agregadas por usuarios únicos. Este último proceso sólo está disponible para personal administrador de la herramienta y se puede ejecutar para calcular estadísticas mensuales y anuales. A la finalización de este proceso el portal de estadísticas dispondrá completamente de toda la información.

5.2. Portal de explotación de estadísticas

El portal de estadísticas es la herramienta final utilizada para poder explotar los eventos de uso recolectados desde los diferentes data providers. Desde este portal se muestra la información relativa a las visitas o accesos a los metadatos de los recursos, las descargas de los mismos y se identifican aquellos eventos que sean llevados a cabo por usuarios únicos.

5.2.1. Visitas

En el apartado de visitas el portal muestra los siguientes apartados:

- Top 10 de visitas en repositorios ordenados por el ratio de visitas/nº de documentos en orden descendente.
- Top 10 de visitas a objetos digitales (artículos) ordenados por el número de visitas en orden descendente.
- Gráfico de tarta de todos los repositorios repartidos según el ratio de visitas/nº de documentos.
- Gráfico de evolución temporal del top 10 por visitas a los repositorios.
- Distribución geográfica de los accesos en gráfica de geolocalización.
- Top 10 de distribución de los buscadores/portales desde donde se accede a los recursos.

Además, el portal incluye la posibilidad de mostrar la información por repositorios o por objetos digitales:

- Para repositorios dispondremos de:
  - Listado de repositorios. Con la posibilidad de marcarlos todos.
  - Tipos de recursos: todos, artículo, libros, etc.
  - Idiomas: todos, español, inglés, etc.
  - Fechas: Totales, agregadas por año, agregadas por mes, y fecha desde y hasta.

  Los vocabularios anteriores los obtiene del sistema recolector de metadatos (DNET).
- Para objetos digitales dispondremos de:
  - URI del Artículo. Para indicar la URL persistente del objeto digital a consultar.
  - Fechas: Totales, agregadas por año, agregadas por mes, y fecha desde y hasta.

En ambos casos los resultados se mostrarán con la posibilidad de seleccionar 5, 10 o 20 elementos. Además, el portal dará la posibilidad de hacer “drilldown” para desglosar la información por años y meses a partir de los resultados obtenidos.

5.2.2. Descargas

En el apartado de descargas el portal muestra los siguientes apartados:

- Top 10 de descargas en repositorios ordenados por el ratio de descargas/nº de documentos en orden descendente.
- Top 10 de descargas a objetos digitales (artículos) ordenados por el número de descargas en orden descendente.
- Gráfico de tarta de todos los repositorios repartidos según el ratio de descargas/nº de documentos.
- Gráfico de evolución temporal del top 10 de descargas en los repositorios.
- Distribución geográfica de las descargas en gráfica de geolocalización.
- Top 10 de distribución de los buscadores/portales desde donde se accede a las descargas.

Además, el portal incluye la posibilidad de mostrar la información por repositorios o por objetos digitales:

- Para repositorios dispondremos de:
  - Listado de repositorios. Con la posibilidad de marcarlos todos.
  - Tipos de recursos: todos, artículo, libros, etc.
  - Idiomas: todos, español, inglés, etc.
  - Fechas: Totales, agregadas por año, agregadas por mes, y fecha desde y hasta.

  Los vocabularios anteriores los obtiene del sistema recolector de metadatos (DNET).
- Para objetos digitales dispondremos de:
  - URI del Artículo. Para indicar la URL persistente del objeto digital a consultar.
  - Fechas: Totales, agregadas por año, agregadas por mes, y fecha desde y hasta.

En ambos casos los resultados se mostrarán con la posibilidad de seleccionar 5, 10 o 20 elementos. Además, el portal dará la posibilidad de hacer “drilldown” para desglosar la información por años y meses a partir de los resultados obtenidos.

5.2.3. Usuarios únicos

Concepto de usuario único: Cuando un usuario accede a un recurso para descargar su información adjunta, cualquier data provider registrará, como norma general, dos eventos diferentes: la visita a los metadatos del recurso y la descarga del documento adjunto. Sin embargo a la hora de computarlo como usuario único esto se debe contabilizar como un sólo evento de uso.

Para mostrar la información filtrada por usuarios únicos se disponen de las mismas herramientas visuales que para los apartados de visitas y descargas:

- Top 10 de usuarios únicos en repositorios ordenados por el ratio de usuarios únicos/nº de documentos en orden descendente.
- Top 10 de usuarios únicos a objetos digitales (artículos) ordenados por el número de usuarios únicos en orden descendente.
- Gráfico de tarta de todos los repositorios repartidos según el ratio de usuarios únicos/nº de documentos.
- Gráfico de evolución temporal del top 10 de usuarios únicos en los repositorios.

- Distribución geográfica de los usuarios únicos en gráfica de geolocalización.
- Top 10 de distribución de los buscadores/portales desde donde acceden los usuarios únicos.

Además, el portal incluye la posibilidad de mostrar la información por repositorios o por objetos digitales:

- Para repositorios dispondremos de:
  - Listado de repositorios. Con la posibilidad de marcarlos todos.
  - Tipos de recursos: todos, artículo, libros, etc.
  - Idiomas: todos, español, inglés, etc.
  - Fechas: Totales, agregadas por año, agregadas por mes, y fecha desde y hasta.

  Los vocabularios anteriores los obtiene del sistema recolector de metadatos (DNET).
- Para objetos digitales dispondremos de:
  - URI del Artículo. Para indicar la URL persistente del objeto digital a consultar.
  - Fechas: Totales, agregadas por año, agregadas por mes, y fecha desde y hasta.

En ambos casos los resultados se mostrarán con la posibilidad de seleccionar 5, 10 o 20 elementos. Además, el portal dará la posibilidad de hacer “drilldown” para desglosar la información por años y meses a partir de los resultados obtenidos.