FECYT FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA

INFORME DEL TALLER DE ESTADÍSTICAS

Madrid, 1 de Junio de 2016

OBJETIVOS Y FINALIDAD:

Dentro de las actividades que se llevan a cabo en el Departamento de Gestión de la Información Científica y con el objetivo de apoyar en la creación de Repositorios científicos institucionales, se desarrolla el proyecto RECOLECTA, cuyo fin es fomentar entre la comunidad científica la creación de una infraestructura nacional de repositorios científicos digitales interoperables según los estándares de la comunidad internacional, además de promover, apoyar y facilitar la adopción del acceso abierto y dotar de una mayor visibilidad los resultados de la investigación que se realizan en España.

FECYT viene trabajando desde hace años en el objetivo de construir un sistema de recolección de estadísticas de los repositorios de instituciones españolas, que permitan ofrecer de manera agregada información valiosa sobre descargas y visitas de los documentos depositados en los repositorios institucionales. Por ello se ha trabajado en los dos últimos años en desarrollar un sistema que permita por un lado a las instituciones disponer de un sistema de medición, seguimiento y control de los documentos depositados y a FECYT poder ofrecer una información agregada de los repositorios nacionales.

Con vistas a que las instituciones puedan conocer de primera mano el funcionamiento del sistema de estadísticas y fruto de la colaboración que para este proyecto FECYT mantiene con CRUE/Rebiun en el marco del convenio firmado, se ha visto la necesidad de plantear la celebración de un taller, en el que se presente a los responsables funcionales y técnicos de los repositorios, las funcionalidades de la herramienta y se construya de forma conjunta un procedimiento de actuación que les permita adaptarse a la vez que visualizar un resultado a nivel nacional para poder crear, validar y visualizar los metadatos de estadísticas

En este sentido, la Universidad Carlos III ha puesto a disposición sus instalaciones para la celebración del taller el pasado 11 de mayo. El curso ha sido presentado e impartido por FECYT con la colaboración de la empresa de soporte tecnológico contratada por FECYT para este servicio. Los objetivos concretos del mismo han sido:

- Mostrar de manera práctica los procesos de validación y recolección de estadísticas
- Resolver dudas técnicas
- Mostrar las ventajas del módulo para todos los repositorios
- Elaborar material de didáctico de apoyo como guías y manuales de instalación

ASISTENTES:

El taller contó con la asistencia de representantes de 17 instituciones, tanto de forma presencial como online. La lista de universidades participantes es la siguiente:

Instituciones que asistieron de forma presencial:

1. Universidad Carlos III de Madrid
2. Consorcio Madroño
3. Universidad de Salamanca
4. Universidad de Valencia
5. Universidad de Girona
6. Universidad de Córdoba
7. Universidad Complutense de Madrid
8. Universidad Politécnica de Cataluña

Instituciones que asistieron por videoconferencia:

1. Universidad de Navarra
2. UOC
3. Universidad Las Palmas
4. Universidad Autónoma de Barcelona
5. Universidad Jaume I
6. Universidad de Murcia
7. Universidad Politécnica de Cartagena
8. Universidad de Lérida
9. Universidad de Santiago

CONTENIDO IMPARTIDO:

1. **Objetivo del módulo de estadísticas:**

El objetivo principal del módulo de estadísticas es crear y establecer una infraestructura nacional que sea nexo de unión con la infraestructura europea de recolección y tratamiento de estadísticas de uso, permitiendo de esta manera la puesta en común de los datos recogidos de los diferentes repositorios existentes.

Un sistema de explotación de estadísticas pretende servir como medida de evaluación y debe enmarcarse dentro de unas directrices claras que no lo aíslen de otros sistemas similares. Por ello FECYT realizó un análisis de los sistemas estadísticos de recolección de datos de repositorios existentes en Europa, resultado del cual fue la adopción y adaptación a las **directrices europeas de las *Knowledge Exchange Usage Statistics (KE)*** Usage Statistic ya que en la actualidad son las directrices europeas más importantes y más extendidas en materia de recolección de estadísticas de uso. Estas directrices establecen un conjunto normalizado de meta-datos asociados al uso de la plataforma de explotación de estadísticas.

Estas directrices proponen tres escenarios distintos para la recolección de estadísticas: *Local, Centralizado o Descentralizado*. De los tres se recomienda la utilizacion de los sistemas *Centralizado o Descentralizado*. La principal ventaja del uso de cualquiera de estos modelos, es que el tratamiento de los datos no se hace de forma individual en cada repositorio, dicho

tratamiento se realiza de forma global en el sistema recolector aplicando las mismas reglas a todos los datos de todos los repositorios.

2. **Requisitos previos que tienen que tener los repositorios:**

   - El *repositorio* tiene que contener publicaciones.

   - Tiene que existir una herramienta que sea capaz de capturar cada evento de uso que se produzca en el repositorio, un *parseador* de eventos de uso.

   - Y una herramienta que permita publicar los eventos de uso capturados para su recolección *proveedor de estadísticas.*

3. **Elementos del módulo de estadísticas:**

   - Proveedor de estadísticas.
   - Herramienta de validación y recolección
   - Portal de explotación de estadísticas.

4. **¿Cómo se recolecta toda esta información?:**

FECYT dispone de un sistema recolector centralizado "*almacén central de estadísticas"* que recolecta, analiza, calcula y enriquece, previa validación, la información de las estadísticas de cada uno de los *proveedores de estadísticas* participantes.

La información recolectada debe pasar por un primer proceso para filtrar multiclics y accesos realizados por robots webs, antes de pasar a su explotación.

5. **Instalación de proveedor de estadísticas:**

Partiendo de la elección, se recomienda, como proveedor de datos el *OAS Data Provider,* los puntos principales para su instalación son los siguientes:

   - El *OAS Data Provider* 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.

   - *Acceso a los log de Apache:* el log de Apache debe ser configurado para que contenga la información necesaria y rotado para facilitar su procesamiento.

   - *Planificación*: se recomienda incluir el script de ejecución que iniciará el tratamiento de los logs, en un cron diario.

El OAS Data Provider está formado por dos herramientas diferentes:

   - LogFile parser: Módulo que se encarga de extraer los eventos de uso de entre los logs registrados por Apache.

   - Data Provider: Módulo que publica por OAI-PMH en CTXO las estadísticas de uso recogidas por el módulo anterior.

El software OAS Data Provider puede funcionar en cualquier sistema o distribución de Linux que esté capacitada para dar soporte a la tecnología LAMP (Linux, Apache, MySQL y PHP, Perl o Python) y se necesita:

1. Servidor Apache
2. PHP versión 5.2.x o superior
3. Extensión DOM de PHP
4. Base de datos MySQL

6. Formato de la información:

El sistema de recolección de estadísticas está basado en la recolección de metadatos en formato OpenURL Context Object (CTXO) a través del protocolo OAI-PMH (en su versión 2.0). Estos metadatos aseguran unos umbrales de calidad mínima de las estadísticas y representan cada uno de los eventos de uso.

La recolección se hace mediante una URL de recolección que cada repositorio debe facilitar, que puede ser distinta a la utilizada para la recolección de metadatos en Dublin Core. Es decisión de los responsables 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. 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.

7. Directrices Recolecta para estadísticas basadas en *Knowledge Exchange Usage Statistics (KE)*:

En relación a los metadatos, estos se tienen que establecer siguiendo las directrices Recolecta para estadísticas basadas en Knowledge Exchange Usage Statistics (KE).

Cualquier repositorio que cumpla con estas directrices podrá ser recolectado por el recolector central de FECYT independientemente del software proveedor de estadísticas utilizado.

Regla	Uso	Descripción
Timestamp del context-object	Obligatorio	El atributo timestamp es obligatorio y debe seguir el formato ISO 6801. 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.
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. 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.

referent:identifier	Obligatorio	El atributo referent:identifier es obligatorio, único y debe coincidir con el URI único del documento en el repositorio. La ausencia del campo referent:identifier no permite conocer el objeto digital al que se encuentra asociada la petición.
Regla	Uso	Descripción
referringEntity:identifier	Obligatorio si aplica	El atributo referringEntity permite identificar el lugar a través del cual el usuario ha accedido al recurso (portal de acceso, buscadores, etc.).
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"
requester:identifier	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). 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.
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 permite identificar el país de origen de la petición para obtener estadísticas de acceso por países.
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
dini:hashed-session	Obligatorio si aplica	Es obligatorio si es aplicable porque si se trata de la primera petición de un usuario, ésta no incluirá dicha información. Debe contener la sesión del usuario que ha realizado la petición cifrada mediante un

		algoritmo de hash. Su no inclusión imposibilitaría el filtrado de multiclics.
Regla	Uso	Descripción
requester:classification	Opcional	Sirve para clasificar los eventos según el vocabulario: internal, administrative o institutional. En caso de tratarse de eventos de uso generados por cualquiera de esas actuaciones.
dini:user-agent	Obligatorio	Es obligatorio y debe contener la cadena de identificación del cliente HTTP utilizado para la petición.
dcterms:format	Obligatorio	Es obligatorio y debe contener el valor conforme al vocabulario. Identifica de forma taxativa si el acceso se ha realizado a los metadatos o al objeto digital a texto completo.
resolver:identifier	Obligatorio	Identifica al repositorio responsable donde se encuentra el objeto digital.
referrer:identifier	Opcional	Es recomendable incluir el campo referrer identificando el servicio exterior que haya proporcionado la reseña al elemento.

8. Partes y etapas del sistema recolector de estadísticas:

a. *Validador de estadísticas*. El sistema validador de directrices Recolecta para estadísticas 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 realizase la validación CTXO.

Las etapas que tiene que seguir un repositorio para la validación de las estadísticas son:

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. http://validador.recolecta.fecyt.es/

2. *Validación*: seguidamente se procederá a la validación de la interfaz de estadísticas de uso. Esta etapa de validación de estadísticas se divide a su vez en tres sub-etapas. En cada una de éstas se debe obtener la valoración como APTO para que el repositorio sea considerado para su recolección:

    - Validación de estadísticas de acuerdo con las directrices Recolecta. Durante esta etapa se validan uno a uno los registros que representan a las estadísticas de uso que publica el proveedor de estadísticas de cada repositorio.

    - Validación del protocolo OAI-PMH. Validación de un conjunto de operaciones encaminas a determinar y valorar la correcta implementación del protocolo OAI-PMH.

    - 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.

3. *Obtención de los resultados*: una vez terminado el proceso de validación, se deberán interpretar los resultados obtenidos.

4. *Corrección de las deficiencias y repetición de la validación*: si no se ha cumplido con las directrices establecidas y volver a validar.

5. *Solicitud de la recolección*: si es APTO se podrá solicitar la recolección.

b. *Recolector de estadísticas.* Elemento del sistema desde el que se realiza 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. La etapa de la recolección de las estadísticas es la siguiente etapa que sigue a la validación. Para ello FECYT cuenta con el Recolector de estadísticas o almacén central de estadísticas (Central Clearing House en adelante CCH) que es la herramienta que se encarga de:

- *Recolectar*, ya sea de manera puntual o periódica las estadísticas suministradas por los diferentes proveedores de datos. El CCH registra la fecha del último evento recolectado y la utiliza, por defecto, en la siguiente recolección. A la finalización de la recolección, toda la información habrá sido consolidada en la base de datos del sistema recolector. A partir de ese momento el proceso corre por cuenta exclusivamente del CCH.

- *Analizar*, durante esta fase toda la información recolectada en el proceso anterior será verificada para detectar y omitir aquellos registros que no cumplan con los requisitos establecidos (directrices). Los registros incumplidores serán ignorados.

- *Calcular*, hay que hacer mención especial a un hecho que ocurre también durante la fase de cálculo. Durante el cálculo de estadísticas se fusiona la información de cada recurso procedente de los diversos proveedores de estadísticas (entiéndase una publicación que esté depositada en más de un repositorio).

- *Enriquecer*, durante esta fase se agrega la información proveniente de los dos sistemas recolectores de los que dispone FECYT: el recolector de estadísticas y DNET (recolector de metadatos Dublin Core). A cada uno de los identificadores de recursos para los que se han recogido metadatos, se les añade información procedente del otro sistema como: Título de la publicación, Autor y Nombre del repositorio.

  Por lo tanto, **antes de lanzar el enriquecimiento hay que garantizar que el repositorio en cuestión ha sido recolectado por el sistema recolector de metadatos**.

c. *Portal de estadísticas*. Portal web 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.

   El portal de estadísticas es la herramienta utilizada para explotar las estadísticas recolectadas. Desde este portal se muestra la información relativa a las visitas (accesos a los metadatos), las descargas de los documentos adjuntos y se identifican aquellos eventos que hayan sido llevados a cabo por usuarios únicos

PROPUESTAS DE MEJORAS TÉCNICAS PARA DESARROLLAR POR FECYT:

Tras la exposición del proceso, se recogen a continuación las diferentes aportaciones de mejoras por parte de los asistentes:

- Se puede eliminar la llamada php - f harvest-identifiers.php -- -c config-dspace.php en el update-script.sh

- Los identificadores del referent, no apuntan a la url del item. Esto lo cambiaron el Consorcio Madroño y la Universidad Carlos III en las modificaciones que hicieron locales.

- A pesar de eliminar el paso del harvest, es necesario instalar el sqlite y el php5-sqlite. En ubuntu: apt-get install sqlite php5-sqlite. Esto debería de ir en la documentación (o averiguar porque se usa sqlite en la línea 28 de./lib/identifiers/lib-dspace.php y, si realmente es necesario).

- Sería conviene eliminar la obligatoriedad de la cookie de sesión de los logs si su único fin es detectar dobles clicks. Aunque una institución grande salga por un proxy que comparta la misma IP, los casos de que se acceda al mismo item, desde la misma institución en un intervalo de 10s, serán residuales y tiene dos problemas grandes:
  - hay plugins de DSpace y e-prints, que esperan el log combined.
  - e-prints no genera cookies de sesión y, quizá otro software tampoco.

  Esto no tiene una solución fácil, pero sería conveniente buscar una forma de mejorar el rendimiento de la aplicación para los repositorios grandes con mucho uso (¿nosql?, ¿no meter el xml entero en la BD?, ¿optimizar / eliminar algún paso? ¿...?). En cualquier caso, si se hace un fork grande, estoy de acuerdo en lo que has dicho que conviene antes hablar con los desarrolladores originales y ver porqué han tomado las decisiones de diseño que han elegido.

- Sería valioso que publicarais una "lista oficial" de robots. En esto está trabajando también COAR y, probablemente se usara no sólo para las estadísticas de repositorios.

- Las direcciones de tipo "administrative" para requesterinfo->classification, están hardcode en ./lib/oasparser-webserver-dspace.php. Sería conveniente que estuvieran en un fichero de configuración.

- Se podría sacar a un fichero de configuración (o por lo menos variables) las direcciones de los logs y del logfile-parser en el update-script.sh.

ESTADO ACTUAL DE LOS REPOSITORIOS:

Tras el taller hay repositorios que están trabajando de forma activa y se comunican en la plataforma https://recolectastats.slack.com . En esta plataforma están compartiendo los desarrollos y las modificaciones que han realizado sobre el desarrollo del cliente servidor.

La Universidad Politécnica de Cataluña y el Consorcio Madroño han realizado desarrollo de la adaptación del OAS que ha desarrollado FECYT.

Instituciones que han conseguido crear el servicio web de consulta de estadísticas en el formato OpenURL Context Object (CTXO):
- Universidad de Valladolid
- Universidad Carlos III de Madrid
- Universidad Politécnica de Cataluña
- Universidad de Girona
- Universidad Complutense de Madrid
- Consorcio Madroño

Instituciones que han completado el proceso, a falta de colgar datos reales:
- La Universidad Carlos III de Madrid
- La Universidad de Valladolid

PASOS A SEGUIR:

- Actualización de la adaptación del programa, basado en el software OAS Data Provider, que ha adaptado FECYT y ha puesto a disposición de los repositorios para que estos puedan crear las estadísticas para que puedan ser recolectadas y visualizadas, así para y el manual necesario para su instalación.
- Actualización de la web de Recolecta en lo referente a la información sobre el proceso de estadísticas
- Seguimiento y asistencia a las instituciones para que implementen el software y puedan validar, recolectar y visualizar sus metadatos. En primer lugar hacer un seguimiento de las seis instituciones que han logrado constituir el ctxo, en un segundo paso dirigirse al resto de las instituciones que asistieron al taller
- Revisión y actualización del validador, recolector y portal de explotación de estadísticas de acuerdo con lo recogido en este informe en el apartado de las propuestas de mejora resultado de las necesidades de los repositorios asistentes.