Reportes estadísticos para repositorios digitales a partir de múltiples fuentes basado en el *stack* ELK

PABLO CÉSAR DE ALBUQUERQUE

PREBI-SEDICI, Universidad Nacional de La Plata

CESGI, Comisión de Investigaciones Científicas de la Provincia de Buenos Aires

pablo@sedici.unlp.edu.ar

GONZALO LUJÁN VILLARREAL

PREBI-SEDICI, Universidad Nacional de La Plata

CESGI, Comisión de Investigaciones Científicas de la Provincia de Buenos Aires

gonzalo@prebi.unlp.edu.ar

MARISA RAQUEL DE GIUSTI

PREBI-SEDICI, Universidad Nacional de La Plata

CESGI, Comisión de Investigaciones Científicas de la Provincia de Buenos Aires

marisa.degiusti@sedici.unlp.edu.ar

RESUMEN

Los Repositorios Institucionales (RI) tienen la necesidad de generar métricas de impacto que permitan comprender cómo son utilizados, lo que puede ayudar a generar información que permita asistir a la toma de decisiones; por ejemplo, fundamentar decisiones políticas o simplemente ofrecer servicios a los investigadores que depositan sus recursos y quieren saber cuál es la interacción del público con el repositorio. Actualmente, existen servicios que brindan la posibilidad de generar tableros de control a partir de los accesos registrados al repositorio a través del *tracking* de eventos y si bien estos servicios son muy utilizados tienen ciertas limitaciones al momento de cruzar información con otras fuentes de datos que pueden aportar información valiosa para analizar. En este trabajo se

utilizará el *stack* ELK para desarrollar un prototipo a partir de la base de datos del RI CIC-DIGITAL y se la asociará a los eventos registrados en el *log* de acceso al servidor donde se despliegan sus servicios, con el fin de generar tableros de control que ayuden a visualizar los recursos más accedidos en un intervalo de tiempo y poder interpretar cuáles son los factores que inciden en estos eventos.

PALABRAS CLAVE

Estadísticas; repositorio institucional; ELK. Statistics; Institutional Repository; ELK.

Introducción

Cada vez se necesitan más pruebas cuantitativas que ayuden a demostrar el valor de los servicios que hay en la web y la realidad que atraviesan los Repositorios Institucionales (RI) no difiere en este aspecto. Entre los tantos servicios que brinda un RI se encuentra el de la generación de métricas y estadísticas. Estas métricas pueden utilizarse para comprender mejor cómo se utilizan los repositorios, lo que puede ayudar a fundamentar las decisiones políticas sobre futuras inversiones y las decisiones de política técnica sobre las mejoras de la infraestructura técnica (KELLY *et al*., 2012). También pueden ser de utilidad a la hora de tomar decisiones operativas a través de información que permita comprender los procesos involucrados en la difusión de los recursos almacenados en un repositorio, como por ejemplo:

- qué tipo de recursos se produce
- en qué áreas se investiga
- quiénes realizan estas investigaciones
- desde dónde
- en qué momento se generan los distintos recursos

- qué mecanismos se utilizan para producir o difundir los recursos
- cómo se utilizan tanto de manera interna como externa

Para medir muchos de estos compartimentos en los RI, se utilizan herramientas dedicadas al análisis web o Web Analytics (WA), entre los que mencionaremos a Matomo y Google Analytics debido a su amplia adopción.

Google Analytics

Google Analytics (GA) es un servicio gratuito utilizado por la mayoría de las bibliotecas académicas, basado en el etiquetado de páginas HTML para registrar la actividad de los visitantes en los servidores de Google. Su gran adopción se debe a que es muy fácil su integración, ya que sólo requiere registrarse en el sitio oficial, obtener un código de seguimiento (denominado Google Analytics Tracking Code) e insertarlo en cada una de las páginas que se desean analizar. Tiene la ventaja de que es muy personalizable no sólo desde el conjunto de paneles y alertas que ofrece sino que también permite definir dimensiones personalizadas. Sin embargo, algunos autores argumentan que GA es inapropiado para el uso educativo, ya que fue construido para el comercio electrónico y no para un entorno educativo (Dragoş, 2011) y sumado a la pérdida de control de los datos debido a que son almacenados en servidores de terceros es que aparecen alternativas *open source* como Matomo.

Matomo

Matomo es una herramienta *open source* de análisis web que ofrece un servicio que permite evaluar todo el recorrido de los usuarios que visitan un sitio web. Muchas de sus funcionalidades están basadas en el servicio de analítica de Google, a punto tal que permite importar los datos sin perder el

histórico ya procesado. Para utilizarlo se requiere disponer de un servidor web, pero también cuenta con una versión *cloud* de pago.

El uso de las mencionadas herramientas para generar estadísticas es una buena alternativa para comenzar a generar métricas que permitan tener una visión más clara de lo que sucede en torno a un RI. Sin embargo, delegar esta responsabilidad no solo quita control sobre los datos que se manejan, sino fiabilidad en cómo se obtienen esas métricas. Por ejemplo en OBRIEN *et al*., (2016), los autores sostienen que hasta el 58 % de toda la actividad de RI generada por humanos no es reportada por Google Analytics. Otra de las limitaciones que presentan estas soluciones es la dificultad de integrar información proveniente de otras fuentes de datos como pueden ser la misma base de datos del repositorio, otros sistemas que formen parte la institución (como sistemas CRIS o portales de congresos o revistas), así como información generada por el uso del mismo sistema: accesos de usuarios, *logs* del servidor o incluso reportes de seguridad vinculados a cada sistema.

Propuesta

A partir de la problemática planteada, se realizó un prototipo a partir del stack de servicios open source llamado ELK, desarrollado por la empresa Elastic. Este stack de servicios se conforma de 3 herramientas que permiten integrar la información de diversas fuentes de datos en a partir de la implementación de pipelines, para luego realizar tableros de control a partir de los datos procesados.

Fuentes de datos

En este trabajo se utilizarán dos fuentes de datos para realizar los tableros de control. La primera fuente consiste en la base de datos del repositorio CIC-DIGITAL, mientras que la otra se trata del *log* de accesos al servidor sobre la cual se despliegan sus servicios.

El repositorio tomado para realizar estas pruebas está desarrollado sobre DSpace, un proyecto *open source*, que modela los recursos de la institución en

ítems y los organiza en una o más comunidades de nivel base que se organizan jerárquicamente en subcomunidades, como se puede observar en la FIGURA 1.

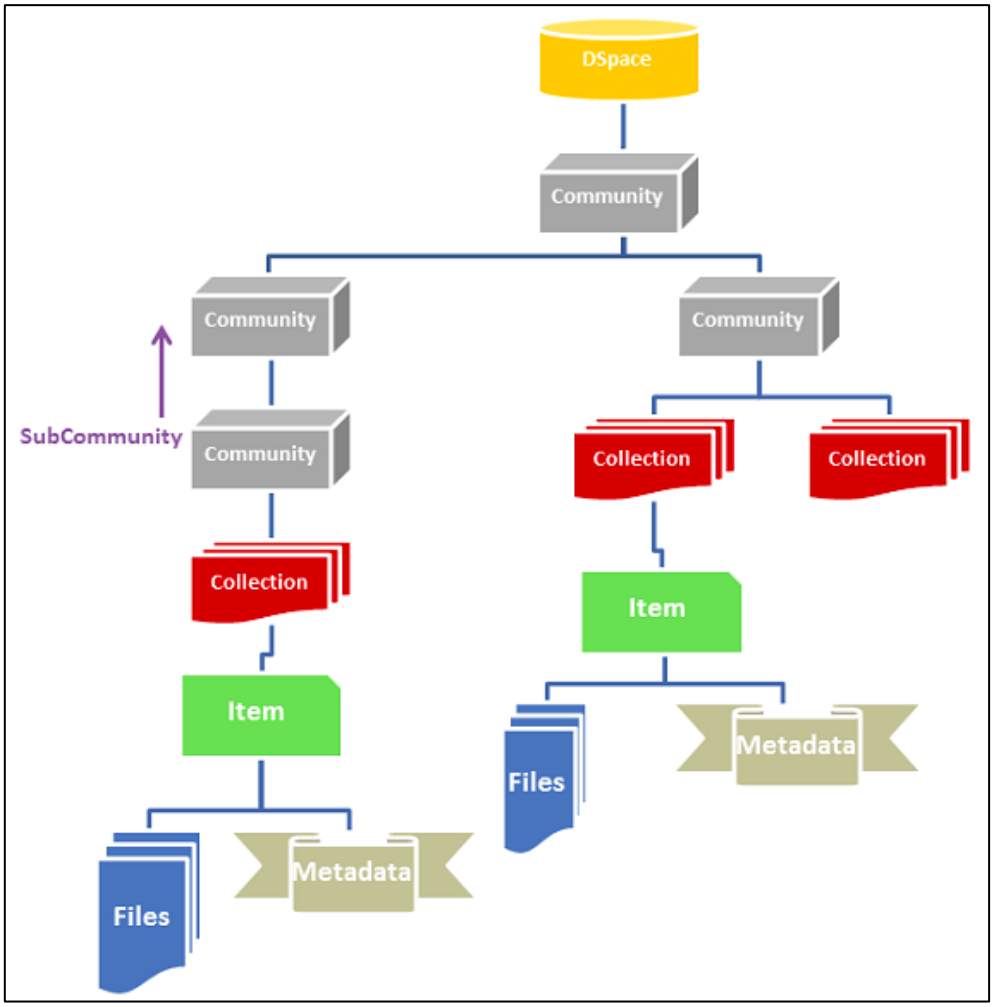

FIGURA 1

Fuente: Functional Overview DSpace 6.x Documentation - LYRASIS Wiki (2018)

En esta fuente de datos cada recurso, comunidad y colección del repositorio se asocia a una tabla de metadatos, que describen no solo los recursos que representan sino que también almacenan información de eventos, como la fecha en la que fue modificado un ítem por última vez, la fecha en la que ingresó al sistema y la fecha en la que el recurso se publicó. Otra aspecto a destacar es que tanto los ítems, como las comunidades y colecciones tienen asociado un identificador persistente, implementado en Handle, que será utilizado para poder interoperar con otras fuentes de datos. En la FIGURA 2 se

puede observar un diagrama del modelo de datos usado en DSpace, donde se puede apreciar la presencia del handle tanto en los ítems como en las comunidades y colecciones.

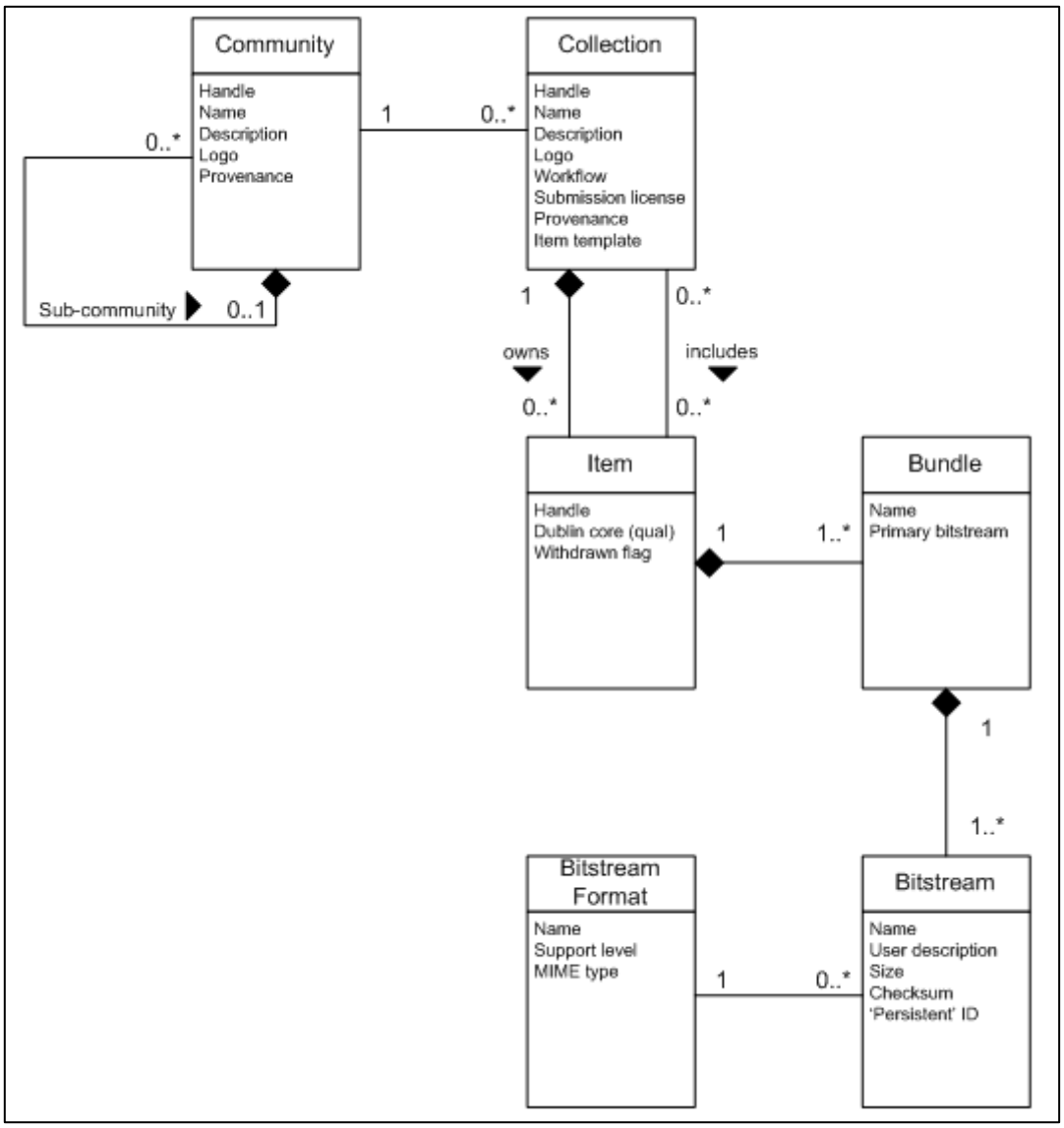

FIGURA 2

Fuente: Functional Overview - DSpace 6.x Documentation - LYRASIS Wiki (2018)

La base de datos utilizada en este trabajo está implementada sobre PostgresSQL, y en ella se modelan, entre otras cosas, los recursos, las colecciones, las comunidades y los metadatos que describen a los recursos.

Los servicios que ofrece este repositorio están desplegados en un servidor Apache, que registra sus accesos un archivo de *log* que será utilizado como segunda fuente de datos. En este archivo de *log*, se registra la IP que realiza la consulta al servidor, el User Agent del cliente, el código HTTP de la respuesta del servidor, el método HTTP con el que se realizó la consulta, el día, la fecha y la hora en que se atendió la petición, el tamaño de la respuesta y el recurso solicitado. En la Figura 3 se deja un ejemplo de cómo se registra un acceso al servidor en el archivo de *log*.

```
186.0.176.252 - - [17/Jun/2021:00:00:11 -0300] "GET
/bitstream/handle/10915/72076/Documento_completo.pdf?isAllowed=y&sequenc
e=1 HTTP/1.1" 200 7470507 "http://sedici.unlp.edu.ar/handle/10915/72076"
"Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML,
like Gecko) Chrome/91.0.4472.101 Safari/537.36"
```

FIGURA 3. EJEMPLO DE REGISTRO EN *LOG*

Stack ELK

Para realizar la integración de estas fuentes de datos se utilizará el *stack* ELK, desarrollado por Elastic, que se compone de tres partes. En primera instancia se encuentra Logstash, que será el encargado de recuperar la información de las diversas fuentes de datos y de procesarlas a través de un conjunto de *pipelines* definidos con el fin de normalizar los datos que finalmente se integrarán más adelante. Luego se encuentra Elasticsearch, un motor de búsqueda en el que se almacenarán los datos procesados por Logstash. Y finalmente se encuentra Kibana, una herramienta de visualización de datos que consume los datos indexados en Elasticsearch para generar diversos tableros de control.

Para poder levantar estos servicios se realizó un *fork* de un repositorio GitHub (LAPENNA, 2014/2021), que consiste de un proyecto *docker-compose* donde se definen cada uno de los servicios involucrados en el *stack* ELK, al cual se le realizaron algunas modificaciones, como definir nuevos servicios para las fuentes de datos. El código del proyecto se encuentra en **GitHub**.

Para poder integrar los datos en Elasticsearch, en primera instancia se crearon dos índices para almacenar, por un lado los datos provenientes de PostgreSQL y por otro los logs de accesos. Estos índices llevan como nombre *items* y *logs*. La creación y la gestión de los datos de cada uno de estos índices es responsabilidad de instancias de Logstash independientes entre sí, que tomarán los datos correspondientes, los procesarán y los almacenarán en un índice en particular.

El índice de *items* contiene los siguientes campos:

- *@timestamp*: fecha en la que se indexó el documento
- *availabledate*: fecha en la que el ítem se publica en el repositorio
- *collection_handle*: handle de la colección a la que pertenece el ítem
- *community_handle*: handle de la comunidad a la que pertenece la colección
- *handle*: handle del ítem
- *hasfulltext*: si el ítem tiene texto completo
- *item_title*: título del ítem
- *last_modified*: fecha de la última modificación que sufrió el ítem
- *owning_collection*: colección a la que pertenece el ítem
- *subtype*: subtipo de recurso del ítem
- *type*: tipo de recurso del ítem

Los campos en el Índice *logs* son:

- *@timestamp*: fecha en la que se indexó el documento
- *agent*: User Agent del cliente
- *bytes*: tamaño de la respuesta expresado en bytes
- *clientip*: IP que realiza la consulta al servidor
- *handle*: handle del recurso accedido

- *referrer*: campo Referrer de la petición HTTP
- *request*: recurso solicitado al servidor
- *response*: código HTTP de la respuesta del servidor
- *timestamp*: día, fecha y hora en que se atendió la petición

Para recuperar los datos, las instancias de Logstash debieron usar distintos *plugins*. En el caso del Logstash dedicado a recuperar los datos de PostgreSQL, se utilizó el JDCB input plugin, mientras que para recuperar los datos del *log* de Apache, se utilizó Filebeats, un cliente ligero que es utilizado para enviar archivos a Logstash o Elasticsearch. Algo importante a aclarar es que en el Logstash que procesa los eventos de acceso se define un filtro, que a través del filtro grok, se encarga de obtener el handle a partir del *request*. Esto es importante ya que es a partir de este dato que se asociará un acceso al servidor con un ítem.

Una vez realizada esta primera carga inicial, una nueva instancia de Logstash se encarga de recuperar los datos del índice de *logs*, y en el proceso de filtrado se enriquecen estos registros a partir de los datos en el índice *items*. Para poder recuperar los datos se usó el Elasticsearch input plugin, mientras que la etapa de filtrado se realiza a través del filtro de Elasticsearch, que se encarga de asociar a partir del handle un evento de acceso con un item. Una vez realizada la asociación, esta instancia de Logstash depositará los eventos de acceso enriquecidos en un nuevo índice llamado *accesed-items* como se puede observar en la FIGURA 4.

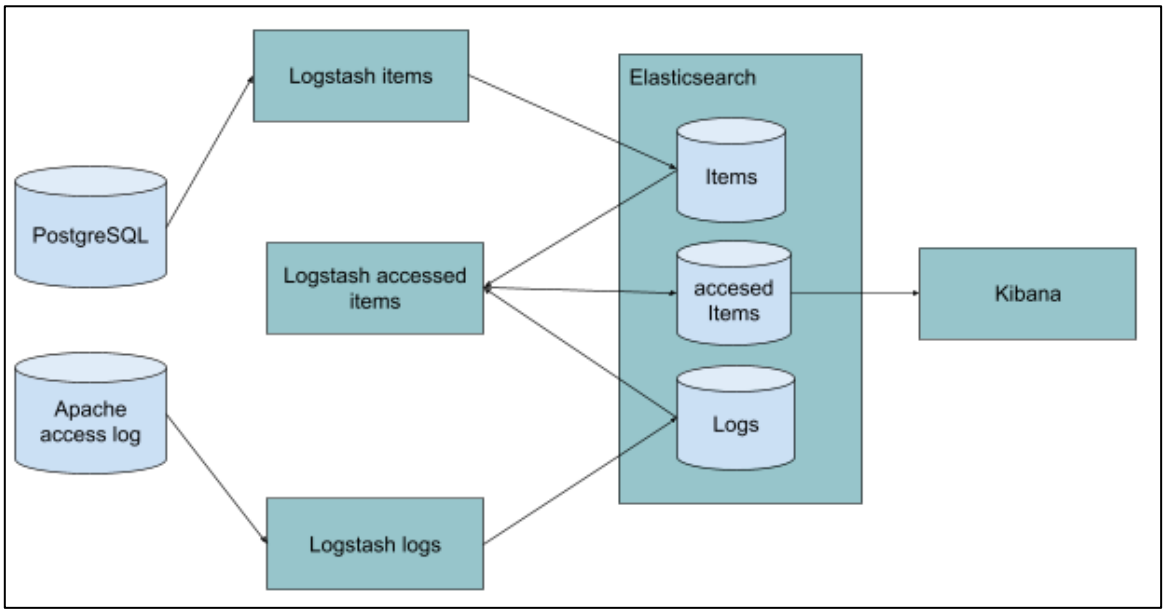

FIGURA 4

Una vez que se consolida el índice de *accessed-items* ya es posible realizar tableros en Kibana como el que se muestra a continuación.

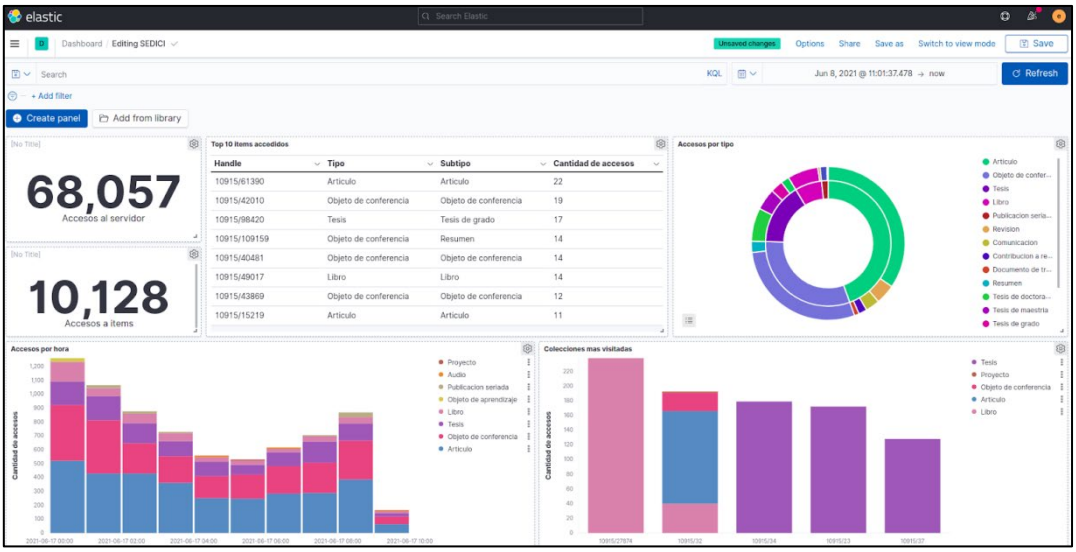

FIGURA 5. VISTA GENERAL DE TABLERO DE CONTROL

En la FIGURA 6 se puede observar dos métricas que muestran la cantidad de accesos que se han registrado en el archivo de *logs* y cuántos de esos accesos se corresponden con accesos a ítems, dejando de lado otras *requests*, como búsquedas en el repositorio o descargas de recursos.

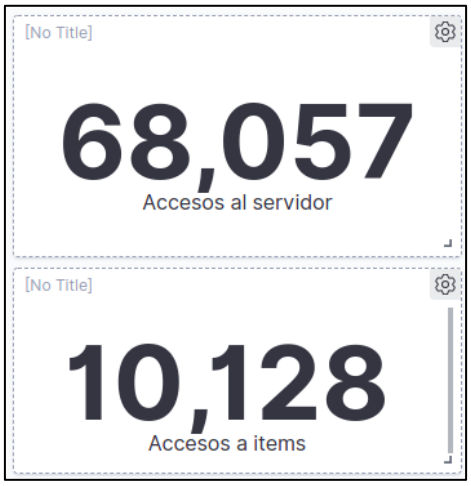

FIGURA 6. CANTIDAD DE ACCESOS AL SERVIDOR Y ACCESOS A ITEMS

Otra de las visualizaciones que se han generados es la de los 10 ítems más accedidos, junto con el tipo y subtipo de documento como se ven en la FIGURA 7.

Top 10 items accedidos

Handle	Tipo	Subtipo	Cantidad de accesos
10915/61390	Articulo	Articulo	22
10915/42010	Objeto de conferencia	Objeto de conferencia	19
10915/98420	Tesis	Tesis de grado	17
10915/109159	Objeto de conferencia	Resumen	14
10915/40481	Objeto de conferencia	Objeto de conferencia	14
10915/49017	Libro	Libro	14
10915/43869	Objeto de conferencia	Objeto de conferencia	12
10915/15219	Articulo	Articulo	11

FIGURA 7. TOP TEN ITEMS ACCEDIDOS

También se realizaron dos visualizaciones para ver como es la distribución de tipos de documento. En la FIGURA 8 se puede observar un gráfico donde en el primer anillo se puede observar la cantidad de accesos por tipos de documentos, mientras que en el anillo externo se pueden visualizar los accesos por subtipos.

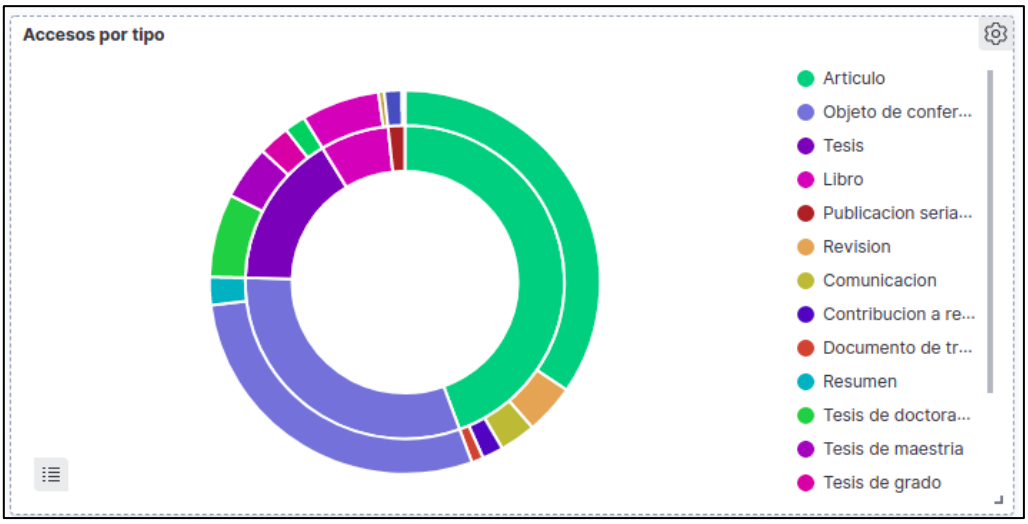

FIGURA 8. ACCESOS POR TIPO DE RECURSO

En la FIGURA 9 se pueden observar la cantidad de accesos por hora y la distribución de tipos de documentos.

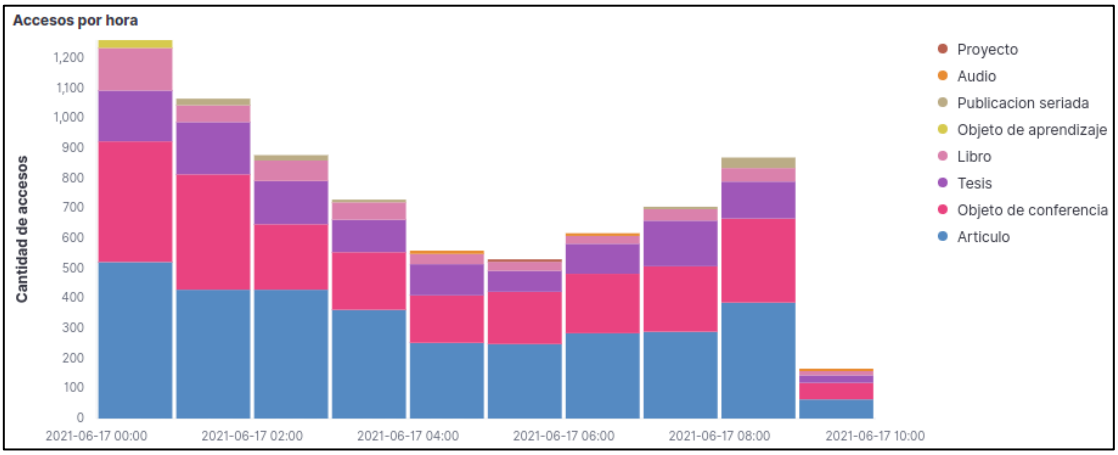

FIGURA 9. ACCESOS POR HORA

En la FIGURA 10 se muestra las colecciones más visitadas, a partir del handle de la colección.

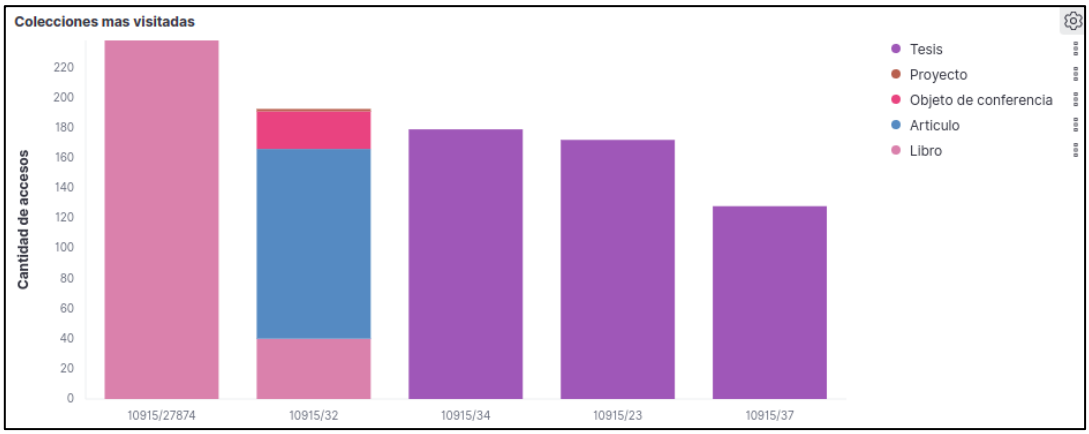

FIGURA 10. COLECCIONES MÁS VISITADAS

Conclusión

En este trabajo se han repasado algunas de las funciones y necesidades que atiende un repositorio, con énfasis en la generación de métricas y estadísticas destinadas a la asistencia en la toma de decisiones. Como se ha mencionado, el uso de herramientas como GA y Matomo, si bien provee muchas funcionalidades importantes para tener una visión más amplia de lo que sucede en un RI, la dificultad que se presenta al momento de integrar otras fuentes y a la falta de control sobre los datos (en el caso de GA) dan lugar a la implementación de soluciones como la propuesta, basada en el *stack* ELK.

En este caso, la información que se puede obtener a partir de estas visualizaciones puede ayudar a entender cómo la tipología de un recurso puede afectar en el interés de los usuarios de un repositorio, de forma tal que es posible realizar acciones como la de la promoción de los ítems más accesibles o bien ofrecer estas métricas como un nuevo servicio a los responsables del repositorio o autoridades de una institución.

Si bien los datos utilizados se corresponden a un lapso de tiempo muy acotado, sumado a que no muchos de los accesos registrados se corresponden con *bots* (como los utilizados por Google que indexan el sitio para ofrecer distintos servicios), resulta interesante cómo este *stack* permite definir diversos mecanismos de ingesta para cada una de las fuentes de datos

a utilizar, pudiendo generar métricas complementarias a las obtenidas por servicios de terceros, disponiendo siempre de los datos procesados.

Bibliografía

DONOHUE, T. (2018). Functional Overview—DSpace 6.x Documentation—LYRASIS Wiki. https://wiki.lyrasis.org/display/DSDOC6x/Functional+Overview

DRAGOŞ, S.-M. (2011). Why Google Analytics cannot be used for educational web content. 7th International Conference on Next Generation Web Services Practices, 113-118. https://doi.org/10.1109/NWeSP.2011.6088162

KELLY, B., SHEPPARD, N., DELASALLE, J., DEWEY, M., STEPHENS, O., JOHNSON, G., & TAYLOR, S. (2012, julio 9). Open Metrics for Open Repositories.

LAPENNA, A. (2021). Deviantony/docker-elk [Shell]. https://github.com/deviantony/docker-elk (Original work published 2014)

OBRIEN, P., ARLITSCH, K., STERMAN, L., MIXTER, J., WHEELER, J., & BORDA, S. (2016). Undercounting File Downloads from Institutional Repositories. *Journal of Library Administration, 56*(7), 854-874. https://doi.org/10.1080/01930826.2016.1216224