Prototipo para la exploración y análisis de los datos de uso estadísticos DSpace en el repositorio institucional CIC-Digital

Facundo G. Adorno[1,2], Marisa R. De Giusti[1,2], Ariel J. Lira[1,2]

1. Proyecto de Enlace de Bibliotecas y Servicio de Difusión de la Creación Intelectual (SEDICI - PREBI) de la Universidad Nacional de La Plata
2. Centro de Servicios en Gestión de Información (CESGI) de la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CICPBA)

Resumen

En el presente trabajo se detalla la experimentación en la implementación de una herramienta para facilitar el análisis y comprensión de los datos estadísticos almacenados en un repositorio DSpace, de tal forma de que permita la exploración del uso de la producción científica en el repositorio institucional de la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires llamado «CIC-Digital». A pesar de que el software DSpace dispone del módulo «DSpace Statistics» encargado de la generación de algunos reportes estadísticos, se determinó que el mismo presenta limitaciones que no permitían explotar en mayor profundidad los datos estadísticos indexados en Solr; debido a esto, se decidió implementar un módulo *ad hoc* que solvente las limitaciones del módulo y agregue nuevas funcionalidades. Para realizar este trabajo se analizó la arquitectura del software para repositorios DSpace, de la herramienta complementaria Solr para la indexación de datos, y de los módulos «Statistics» y «Discovery» en DSpace que agregan distintas funcionalidades basados en estos datos, de tal manera de poder abordar en el análisis de las tecnologías que finalmente serían utilizadas para la implementación del prototipo. Posterior a este análisis, se comenzó con la implementación de la herramienta sobre Dspace en su versión 6 utilizando las tecnologías compatibles con esta versión (XMLUI, Apache Cocoon, XSLT, entre otras), y se desarrolló un módulo llamado «Statistics-Discovery» que permite la exploración mediante búsquedas sobre los datos estadísticos indexados en Solr, agregando funcionalidades de exportación de resultados en diversos formatos de texto y generación de reportes y gráficas.

Repositorios digitales

Los repositorios (Jacobs, 2016; Johnson, 2002) son archivos digitales provistos de un conjunto de servicios web centralizados, creados para organizar, gestionar, preservar y ofrecer acceso libre a la producción científica, académica o de cualquier otra naturaleza cultural, en soporte digital, generada por los miembros de una organización o comunidad. Los distintos objetos digitales en un repositorio son representados como recursos, y estos están conformados por un conjunto de metadatos que los describen.

Existen distintos tipos de repositorios («Definición y tipos de repositorios», s. f.) según los servicios que ofrecen y los contenidos que alojan, y en particular para este trabajo son de interés los llamados repositorios institucionales. Un «repositorio institucional» (De Giusti, 2017; «Repositorios digitales - Red Infod», s. f.) es un tipo de repositorio digital donde se depositan recursos derivados de la producción científica y académica de una institución o cualesquiera otros que la institución considera importantes. Estos repositorios ofrecen un punto de acceso único a la información de la institución y de sus autores, trabajando bajo estándares de catalogación, preservación e interoperabilidad de recursos que maximizan su recuperación, compartición y accesibilidad en el tiempo.

Estadísticas de uso en los repositorios digitales

Los repositorios institucionales nacen del Acceso Abierto u *Open Access*, movimiento que aboga por el acceso libre de sus contenidos, sin restricciones o barreras, ya sean éstas económicas o de derechos de explotación, es decir, que el acceso no sólo debe ser gratuito, sino que debe ser libre y permitir la reutilización del material intelectual producido en las instituciones. A medida que este movimiento se arraigó en el mundo científico-académico y fueron surgiendo cada vez más repositorios institucionales, se planteó la necesidad de encontrar criterios que permitan evaluar el estado de los repositorios, la usabilidad de su producción intelectual, o el uso de los servicios que ofrecen, entre otras cosas. Los repositorios institucionales son clave para promover la visibilidad de los resultados de la investigación financiada por una institución. Dado que las organizaciones están interesadas en demostrar el valor y el impacto de los repositorios institucionales, la disponibilidad de las estadísticas vinculadas a las necesidades antes planteadas conforman un servicio de un alto valor.

Como afirma Bernal y Pemau-Alonso (2010) los tipos de estadísticas existentes para medir el impacto de los producción intelectual pueden agruparse en dos perspectivas diferentes: estadísticas desde el punto de vista de los autores y desde el punto de vista de los internautas. Este primer punto de vista, es decir, desde los autores, es el que mayor consolidación presenta y se basa principalmente en las citas de las publicaciones, existiendo sistemas de medición como el «JIF» o *Journal Impact Factor* (sistema basado en la premisa que un alto índice de citaciones de una revista equivale a un alto impacto/calidad de un artículo), y el «h-index» o *Hirsch-index* (que centran su evaluación en la producción del autor y no en el índice

de impacto de las revistas). El otro punto de vista, es decir, el de los internautas, es una evaluación alternativa y se basa principalmente en el uso (por ejemplo, visitas o descargas de publicaciones) que se realiza desde la web al contenido intelectual de un investigador o una institución.

Las estadísticas de uso transmiten una de las informaciones más buscadas por los administradores de un repositorio y los miembros de una institución, ya que indican directamente la actividad y el uso que un usuario del repositorio hace del sistema en sí y de la producción académica de la institución. Entre las estadísticas de uso que más comúnmente interesa medir son:

- cantidad de veces que una página de un servicio fue visitado,
- cantidad de veces que una página de un ítem/recurso fue visitado,
- cantidad de veces que una ítem/recurso fue descargado (con o sin éxito),
- términos de búsqueda mayormente usados en el repositorio,

Existen otras cuestiones potenciales a medir sobre el comportamiento del usuario en el repositorio, y las ya mencionadas forman parte de una área de análisis mayor llamado «Análisis Web» o «Web Analytics». La *web analytics* se define como («Web analytics», 2018) las actividades de medición, recopilación, análisis y generación de informes de datos generados en torno al uso de un sitio web con el fin de comprender y optimizar el/los servicio/s provisto/s por el sitio.

La disposición de las estadísticas sobre el uso de la producción intelectual puede conllevar diversos beneficios (Bernal & Pemau-Alonso, 2010), entre ellos:

- Reflejar la visibilidad, la difusión internacional y las tendencias de uso de estos archivos abiertos, que son indicadores de su eventual consolidación.
- Pueden ser un medio persuasivo y elocuente para explicar el porqué de los repositorios abiertos ante la institución de la que dependen y su agencia financiadora –mostrando la relación coste-beneficio del repositorio– y ante la comunidad científica cuya investigación difunden y preservan –demostrando su efectividad en potenciar la accesibilidad de los resultados de investigación de un modo gratuito e inmediato en internet–.
- Son de utilidad para los investigadores que están interesados en saber cuánta atención está recibiendo su investigación y cómo los usuarios están accediendo a este material, comparando el grado de «popularidad» de sus trabajos con el de otros compañeros de profesión. De esta forma sirve como estímulo para los autores depositantes y para captar a otros potenciales.

Iniciativas internacionales en relación a las estadísticas de uso

Más allá del alcance de los repositorios digitales, es decir, en un aspecto más amplio y de mayor alcance, existen iniciativas internacionales que buscan normalizar y unificar la interoperabilidad de los datos de uso generado por los internautas sobre una misma publicación o un contenido alojado en diversos repositorios institucionales. Con el fin de que diferentes repositorios en el mundo informen del uso de sus publicaciones a otros sistemas, diversos estándares y protocolos de interoperabilidad fueron creados y evolucionando en el tiempo, y a su vez han sido creados distintos servicios que hacen uso de estos estándares para

cosechar estos datos desde diversos repositorios y ofrecer estadísticas globales más complejas.

Esta necesidad de normalizar y unificar conceptos se remonta hacia fines del siglo XX, momento en que surgió una explosión en la cantidad de material electrónico disponible en distintas bibliotecas del mundo (Fleming-May & Grogg, 2010), y cuando las inconsistencias en cuanto a la definición de términos relativos al uso por parte de los usuarios hacía imposible la comparación de las estadísticas de uso entre distintos sistemas de bibliotecas. Los elementos medibles en cuestión, tales como sesiones, búsquedas y descargas, eran inconsistentes entre estos sistemas y se distribuían en una variada cantidad de formas y formatos; a ésto último se sumaba la cantidad de tiempo que los bibliotecarios invertían en recolectar, filtrar y archivar las estadísticas de uso para su difusión.

En 2002, en respuesta a la complicada situación que las estadísticas de uso habían creado, surgió una organización internacional sin fines de lucro dedicada a facilitar datos de uso "coherentes, creíbles y comparables": *Counting Online Usage of Networked Electronic Resources*, o el proyecto «COUNTER». En la actualidad, esta organización está respaldada por una comunidad global de miembros de bibliotecas, editores y vendors[1], que contribuyen al desarrollo del «Código de Práctica», o *Code of Practice,* a través de distintos grupos de trabajo. El *Código de Práctica* («COUNTER - Code of Practice», 2012) es un documento que describe y estandariza los reportes de estadísticas de uso para diferentes tipos de recursos en línea, proporcionando orientación sobre los elementos de datos a medir, las definiciones de estos elementos de datos, y el contenido y el formato de los reportes. Permite a los editores y proveedores informar el uso de sus recursos electrónicos de una manera consistente, habilitando a las bibliotecas comparar datos recibidos de diferentes editores y proveedores. El primer lanzamiento del Código de Práctica para reportes estadísticos sobre el uso en revistas y bases de datos fue lanzado en enero del 2003 y, desde entonces, el código fue evolucionando hasta llegar a su versión 4 en el año 2012, ampliando entre otras cosas la cantidad de recursos a ser medidos (revistas, bases de datos, libros, trabajos de referencia y bases de datos multimedia). Actualmente se encuentra en desarrollo la versión 5 del código, próxima a ser lanzada en 2019.

COUNTER además presenta como documento anexo el «Código de Práctica para Artículos», el cual proporciona especificaciones para el registro e informe de estadísticas de uso a nivel de artículo individual, las cuales se basan y son consistentes con el Código de prácticas COUNTER; está destinado a proveedores que cumple con COUNTER, repositorios y otras organizaciones. Este código está basado en el ahora finalizado «PIRUS2», proyecto desarrollado por JISC en el Reino Unido, cuyo objetivo (Fleming-May & Grogg, 2010) era “desarrollar un conjunto de estándares, protocolos y procesos que permitan a los editores, repositorios y otras organizaciones generar y compartir estadísticas de uso fiables y confiables para los artículos individuales y otros temas que alojan”.

[1] Un editor u otro proveedor de información en línea que entrega contenido licenciado al cliente y con quien el cliente tiene una relación contractual.

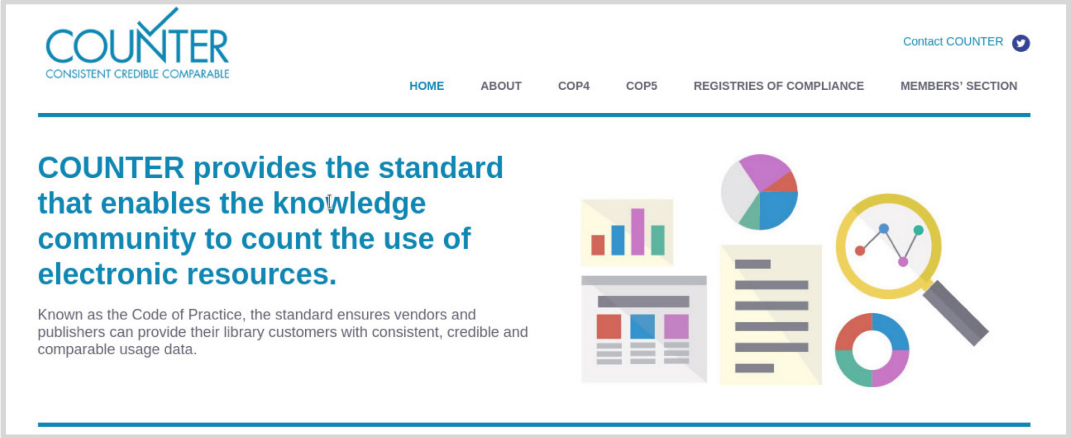

Figura 1. Página web del proyecto COUNTER

A medida que la cantidad de publicadores y proveedores de Europa que exponían datos de uso mediante COUNTER, la recolección y carga de los mismos en los sistemas clientes requería de mucho esfuerzo, por lo que en el año 2005 comenzó la iniciativa conocida como *Standardized Usage Statistics Harvesting Initiative,* o «**SUSHI**», para definir una forma automatizada de intercambio de reportes COUNTER mediante *web services*. Entre los objetivos iniciales de SUSHI (Fleming-May & Grogg, 2010) podemos mencionar:

- Resolver los problemas de cosecha y administración de datos de uso que provenientes de un creciente número de proveedores.
- Promover consistencia en el formato de los datos de uso (XML)
- Automatizar el proceso de cosecha.

SUSHI es un estándar ANSI/NISO (SUSHI Protocol, 2014) que define un protocolo cliente/servidor, modelo donde un cliente solicita un reporte a través de un servicio web, el servidor retorna en la respuesta un reporte COUNTER en formato XML, y luego el cliente procesa la respuesta. El protocolo está diseñado para proporcionar un método automatizado de recuperación de informes de estadísticas de uso COUNTER, utilizando un contenedor XML procesable por máquina y a través de servicios web que se comunican mediante «SOAP[2]». En SUSHI funcionan principalmente 3 formatos XML que posibilitan la interoperabilidad: *SUSHI XML Schema*, *SUSHI WSDL* y los reportes *COUNTER-XML*. SUSHI XML Schema y SUSHI WSDL representan el *contrato de datos* y el *contrato de servicio* entre el cliente y el servidor que operan en un ambiente business-to-business, y el informe COUNTER XML es la carga real o el *payload* de la transacción.

En el año 2015 se lanzó un borrador de una versión reducida o más liviana de SUSHI llamada «SUSHI-Lite». Esta versión de SUSHI se basa en las tecnologías REST para el intercambio de datos (en lugar de SOAP) y JSON para la representación de los informes COUNTER (en lugar de XML).

[2] Protocolo estándar que define cómo dos objetos en diferentes procesos (que podrían correr en diferentes sistemas operativos, con diferentes tecnologias y lenguajes de programación) pueden comunicarse por medio de intercambio de datos XML.

Módulo de estadísticas en DSpace

El presente trabajo fue desarrollado sobre la plataforma de software «DSpace» (en su versión 6) utilizada por el repositorio institucional de la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires, llamado «CIC-Digital», creado a fin de 2014 con el objeto de reunir, registrar, divulgar, preservar y dar acceso público a toda la producción científico-tecnológica y académica de la institución. *DSpace* es un software pensado para la gestión de repositorios digitales que proporciona distintas herramientas y funcionalidades. Entre sus características podemos mencionar: (a) es multiplataforma de código abierto desarrollado en JAVA; (b) es altamente adaptable a las necesidades de cada institución, permitiendo cambios en la vista del repositorio (XMLUI o JSPUI), los formatos de metadatos, los campos de búsquedas, los mecanismos de autenticación, los idiomas soportados, entre otros; (c) permite la administración y preservación de varios tipos de contenidos digitales; (d) dispone de una organización de contenidos jerárquica mediante comunidades y colecciones; (e) soporta gran variedad de estándares: OAI-PMH, OAI-ORE, SWORD, WebDAV, OpenSearch, OpenURL, RSS, ATOM; (f) permite la gestión de permisos, grupos y usuarios; (e) posee una gran comunidad de usuarios y desarrolladores ubicados en todo el mundo.

DSpace («Architecture - DuraSpace», s. f.) posee una arquitectura multicapa modular, con posibilidad de utilizar una amplia variedad en gestores de base de datos relacionales (actualmente con soporte para PostgreSQL y Oracle Database) y motores de indexación (Apache Solr[3] o ElasticSearch) como sistemas secundarios de almacenamiento.

En relación a las estadísticas de uso, DSpace dispone de un módulo específico llamado «DSpace Statistics» que se encarga de generar reportes a partir de los accesos o visitas de páginas del repositorio, las descargas de bitstreams, las búsquedas en el repositorio y los eventos de workflow en el repositorio («SOLR Statistics - DuraSpace», s. f.). DSpace Statistics es una arquitectura cliente/servidor basada en Solr que recopila eventos de uso en las aplicaciones de interfaz de usuario JSPUI y XMLUI de DSpace. El módulo de estadísticas en DSpace registra todos estos eventos de uso en un core Solr específico llamado *statistics,* donde se indexan distintos tipos de eventos y, aunque estén todos en el mismo índice, los campos almacenados para las descargas, consultas de búsqueda y eventos de workflow son diferentes. Por cada documento indexado en este core, un campo en

[3] Apache Solr es una plataforma de búsqueda empresarial de código abierto escrita en Java, del proyecto Apache Lucene. Sus características principales incluyen: búsqueda a texto completo, *hit highlighting*, búsquedas facetada, implementación de una Restful API, flexible y extensible, basado en Apache Lucene (una simple pero poderosa librería de búsqueda e indexación de texto escrita en Java). Además presenta características NoSQL, capacidad de manejo de documentos enriquecidos (por ejemplo, Word, PDF), y la devolución de resultados de búsqueda en variados formatos como XML, CSV, JSON, entre otros.

particular llamado *statistics_type* determina qué tipo de evento de uso está tratando. Los tres valores posibles para este campo son: *view*, *search*, y *workflow*[4].

Estadísticas

Número total de visitas

	Visualizaciones
Collection ...	10

Visitas al mes

	febrero 2018	marzo 2018	abril 2018	mayo 2018	junio 2018	julio 2018	agosto 2018
Collection ...	0	0	0	0	0	0	10

Países con más visualizaciones

	Visualizaciones
India	3
Emiratos Árabes Unidos	1
Argentina	1

Figura 2. Vista de estadísticas de visitas en el módulo DSpace-Statistics

El módulo de estadísticas implementa dos funcionalidades básicas: indexación y búsqueda. Un servicio llamado *SolLoggerServiceImpl* se encarga de registrar los distintos eventos de uso que suceden en las aplicaciones de interfaz de usuario (JSPUI o XMLUI) que existen en DSpace. Cada una de estas aplicaciones va registrando acciones de eventos de uso que, posteriormente, serán procesadas por un listener[5] implementado por el módulo, llamado *SolrLoggerUsageEventListener.* Acorde al tipo de evento de uso del que se trate, el listener invoca alguna de las siguientes implementaciones del *SolLoggerServiceImpl:*

- `postSearch()`: método para indexar y registrar un evento del tipo *"search"*..
- `postView()`: método para indexar y registrar un evento del tipo *"view"*. Utilizado tanto para descargas de Bitstreams como vistas de Ítems, Comunidades y Colecciones.
- `postWorkflow()`: método para indexar y registrar un evento del tipo *"workflow"*.

Además de brindar funciones de indexación, este servicio también ofrece algunas funcionalidades de búsqueda sobre el core de Solr, así como de optimización. Una de las funciones más importantes que ofrece es la de detectar y marcar aquellos registros del core que son bots, determinados a partir de la *ip* o *userAgent* de cada

[4] Para gestionar la carga de ítems en un repositorio, DSpace ofrece un mecanismo de revisión que determina cuándo un envío de una publicación está apto o no para su depósito en el repositorio. Este mecanismo recibe el nombre de «Workflow» y básicamente define un conjunto de pasos a seguir. Cuando estos pasos son ejecutados sin interrupción hasta el final, entonces el envío es archivado en el repositorio.

[5] Un *listener,* o también llamado manejador de evento, es un objeto que recibe la notificación de que un evento fue lanzado en algún lugar de la aplicación, y actúa en respuesta a ese evento.

registro en Solr. Por último, este servicio utiliza bases de datos de geolocalización gratuitas, ofrecidas por el servicio *Maxmind GeoLite databases*, a partir de las cuales se determinan distintos datos geográficos (latitud, longitud, país, ciudad, etc.) derivados desde la ip de un registro.

A partir de los datos de uso indexados y las búsquedas realizadas en el índice «*statistics»*, el módulo genera un conjunto de reportes (ver Figura 2) por cada tipo de evento registrado, es decir, reportes para estadísticas de accesos y descargas, para estadísticas de consultas de búsquedas, y para estadísticas de eventos de workflow. Por defecto, estos reportes son tablas, algunas de rankings con 10 filas como máximo (por ejemplo, los países y las ciudades con más accesos), otras permiten ver los eventos registrados en un determinado periodo de tiempo fijo (1 año, 6 meses o un mes atrás).

Limitaciones del módulo

A pesar de que DSpace cuenta con un mecanismo para la generación de reportes basadas en el uso del repositorio y sus contenidos, el módulo de estadísticas presenta algunas limitaciones que impide a los usuarios y administradores del repositorio explotar en mayor profundidad estos datos almacenados en Solr. Entre estas limitaciones podemos mencionar las siguientes:

- **Poca cantidad de filas por tabla:** la mayoría de los reportes retorna solo 10 de resultados, por ejemplo, el top 10 de términos más buscados, el top 10 de ítems más accedidos, y no permite seleccionar una mayor cantidad de resultados de manera arbitraria.
- **Rangos de fecha fijos:** no se puede seleccionar un rango de fecha arbitrario o mayor a un año de antigüedad. Los reportes sólo pueden generarse en un escueto rango de fechas predeterminado, a saber: el mes anterior, rango de 6 meses antes, rango de un año antes. Además no se puede especificar una mayor granularidad de tiempo, por ejemplo, para reportar los eventos de uso diarios.
- **Ausencia de gráficas:** no existen visualizaciones de reportes *out-of-the-box*, es decir, que deben implementarse utilizando alguna librería de graficación y se requiere de conocimiento informático para ésto.
- **Aspectos a explorar limitados**: no ofrece ninguna funcionalidad para inspeccionar algún otro de los posibles aspectos registrados en el core de statistics, como por ejemplo: IPs, Referrers, etc.
- **Datasets para generación de reportes inaccesibles:** no ofrece exportación de reportes ni permite exportar los registros involucrados en la generación de los reportes para un análisis en mayor profundidad fuera del sistema.
    - Podría servir de mucho disponer de esta capacidad y utilizarlos como datasets para análisis futuros en materia de detección de bots, tendencias en el uso del contenido del repositorio, detección de picos de actividad a lo largo del tiempo, entre otros.

- **Hardcoding**[6]: Algunas limitaciones se deben a la codificación explícita de ciertos datos en las clases que implementan la vista y el modelo del módulo de DSpace Statistics, por lo que cambiar algunas de los mismos requeriría de una completa compilación del código de la aplicación y la posterior actualización de la instalación para que surtan efecto.

Las limitaciones anteriormente mencionadas llevaron a que, en distintas situaciones en que se requería un mayor nivel de detalle en determinado informe institucional, se tuviese que realizar una consulta directa a Solr, recopilar la información retornada por este sistema, y finalmente generar manualmente tablas y gráficas mediante algún programa externo que la procese. Las consultas realizadas a Solr son poco intuitivas, aunque muy potentes, y es necesario leer mucha documentación al respecto para poder realizarlas correctamente, cuestión que representaba un problema para las personas que querían obtener la información necesaria para generar estos reportes personalizados.

Implementación de un prototipo de búsqueda basado en Discovery

Teniendo en cuenta las limitaciones del módulo de estadísticas de DSpace, surgió la necesidad de crear otra herramienta que permitiese mayor libertad al momento de inspeccionar los datos de uso del repositorio y generar reportes. La herramienta debía cumplir con una serie de expectativas, enumeradas en la Tabla 1.

Tabla 1 - *Expectativas a cubrir por el prototipo*

Exploración/ Búsqueda de registros	Permitir realizar búsquedas avanzadas sobre el índice *statistics,* abstrayendo al usuario de la complejidad de consulta a Solr subyacente, de tal manera de poder analizar los datos en crudo que se indexan.
Tiempos de respuesta razonables	Se espera una interacción fluida entre la aplicación y el usuario, de tal manera que ante cada petición no tarde mucho en resolver. Inicialmente, se considera razonable que ante mayor cantidad de datos a consultar (posiblemente millones y en aumento constante), mayor sean los tiempos de respuestas.
Configurable	Se espera que algunas funcionalidades de la herramienta sean configurables a través de archivos de configuración, de tal manera de ofrecer mayor flexibilidad y que sea pueda adaptar a las necesidades de cada Institución que quiera utilizarla.
Múltiples contextos de búsqueda	Permitir definir diversos contextos sobre el que realizar búsqueda de registros, de tal manera que pueda obtenerse el uso realizado sobre cualquier objeto o conjunto de objetos en DSpace (ítems, colecciones y comunidades). Esto tendría que permitir seleccionar un conjunto de objetos a partir de sus características.

[6] El *hardcode* es una mala práctica en el desarrollo de *software* que consiste en incrustar datos directamente en el código fuente del programa, en lugar de obtener esos datos de una fuente externa como un fichero de configuración o parámetros de la línea de comandos, o un archivo de recursos.

Generación de gráficas	Disponer de una sección de graficación para visualizar diferentes características o variables de los registros resultantes de una búsqueda. Estos gráficos no deberían representar un cálculo estadístico complejo, simplemente deberían facilitar un paneo de diversos aspectos del conjunto de registros explorados.
Exportación de registros	Permitir exportación de los registros resultantes de una búsqueda para su posterior uso en sistemas externos especialmente dedicados al análisis estadístico (como el software *MatLab* o el ambiente estadístico *R*).

Analizando las diferentes alternativas de implementación, se determinó reutilizar un módulo ya existente en DSpace llamado *«Discovery».* Discovery es el servicio que permite («Discovery - DuraSpace», s. f.) la búsqueda de contenidos en los repositorios DSpace, el cual se conecta con el índice «search» en Solr para encontrar determinados objetos que coincidan con las términos de búsqueda aplicados. Este módulo habilita la exploración mediante la búsqueda de coincidencias en los metadatos que describen a los ítem, comunidades y colecciones en DSpace, a través de la aplicación de filtros (p.e. por fecha, por autor, etc.), refinamiento mediante *facets*, paginación y ordenamiento de resultados. Además permite determinar contextos específicos de búsqueda, en una comunidad o colección específica.

Considerando las ventajas de implementación de la herramienta basada en el módulo de Discovery, se creó un nuevo módulo llamado «Statistics-Discovery». Su arquitectura (visualizada de forma simplificada en la Figura 3) es similar a la definida por el módulo Discovery aunque, como se verá más adelante, tuvo que ser readecuada para su funcionamiento con el índice «statistics» en vez del índice «search» en Solr. Al igual que en Discovery, tanto la construcción de la vista como el funcionamiento de algunos fragmentos de la lógica en este nuevo módulo se configuran desde archivos de configuración XML, configuraciones instanciadas mediante *inversion of control*[7] (IoC) por «Spring», un framework *open-source* escrito en Java para el desarrollo de aplicaciones y contenedor de inversión de control. La comunicación con el servidor «Apache Solr» desde DSpace se realiza mediante «SolrJ», una API que facilita la comunicación entre las aplicaciones Java y Solr a través de HTTP.

[7] Es un principio de diseño de software en el que el flujo de ejecución de un programa se invierte respecto a los métodos de programación tradicionales. Basado en el patrón de *dependency injection*, donde un objeto provee/inicializa las dependencias a/de otro objeto.

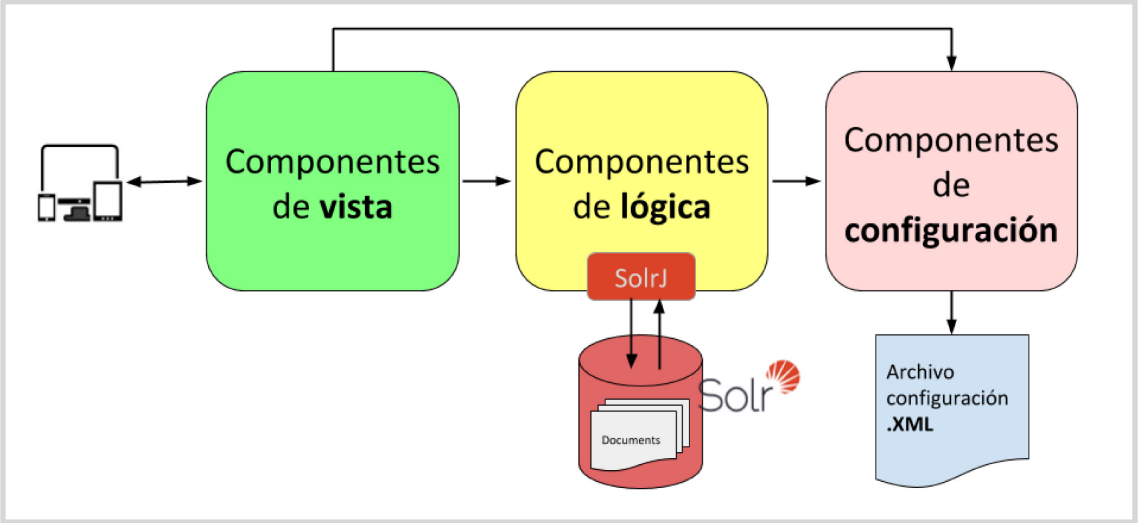

Figura 3. Capas que componen el módulo Statistics-Discovery

Extensiones al modelo base

Para el cumplimiento de las expectativas esperadas sobre el prototipo, algunas extensiones fueron realizadas al modelo original derivado de Discovery. Entre ellas podemos mencionar las siguientes:

- Opciones ampliadas para la selección de un contexto de búsqueda.
- Configuración de nuevos filtros de búsqueda y opciones de refinamiento mediante facets basados en los metadatos que componen los registros de uso en el core «statistics».
- Creación de un modelo extensible para la exportación de resultados en diversos formatos de texto.
- Creación de un *endpoint* de consulta JSON para la creación de reportes y generación de gráficas basadas en estos reportes.

En el módulo de Statistics-Discovery, un contexto de búsqueda se define como el conjunto de objetos en DSpace sobre el que se van a obtener registros de uso a partir de una búsqueda. Las opciones originalmente permitidas era la definición de contexto por colección y comunidad (objetos que componen el modelo de datos en DSpace(«Data Model - DuraSpace», s. f.)), a las que se agregaron las opciones de selección de contexto por ítem y por agrupamiento arbitrario de objetos DSpace. Esta última opción se traduce como la definición de un conjunto de objetos que cumplen con ciertas características (por ejemplo, ítems cuyo autor sea *«Juan Pérez»* y su fecha de publicación sea la década pasada), y la forma en que fue implementada en el prototipo fue mediante el uso de búsquedas realizadas en Discovery (ver Figura 4). A partir de una búsqueda Discovery, el prototipo determina los objetos utilizados como contexto.

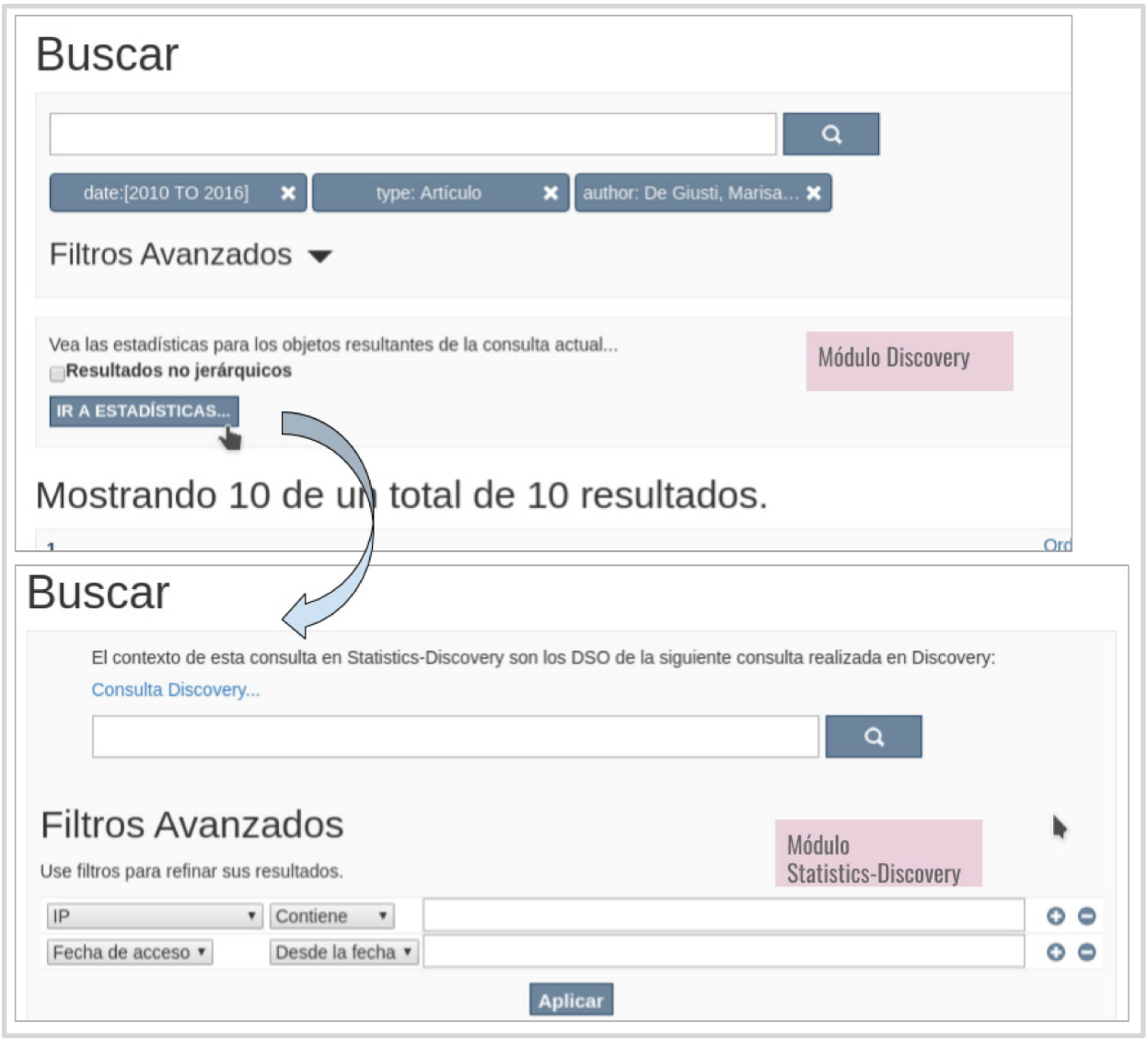

Figura 4. Contexto de búsqueda derivado de una consulta en Discovery

La búsquedas en el nuevo módulo se realizan a través de la aplicación de filtros y facets. Los filtros buscan coincidencias en diversos campos de los registros de uso, mientras que los facets permiten el refinamiento de resultados a partir de las características de los mismos (por ejemplo, aquellos cuyo valor en el campo «IP» sea «W.X.Y.Z»). Entre las opciones de filtrado y facets configuradas figuran las siguientes: *IP, Código de país, Código de Continente, Ciudad, Agente de usuario, Referer, Tipo de registro estadístico, Tipo de objeto DSpace, Fecha de acceso.*

La exportación de resultados se realiza mediante la definición de un modelo extensible de formatos de exportación; por defecto se implementaron exportaciones en formatos CSV y JSON. Cada formato se implementa como una *estrategia* en el modelo (ver Figura 5), donde cada estrategia debe implementar el método *hook* («export») para definir cómo debe hacerse la exportación a partir de los resultados de búsqueda. Además, el mecanismo de exportación permite manipular los valores de salida de los campos exportados mediante el uso de *transformaciones*, de tal forma de permitir el control de los valores exportados. Un ejemplo de transformación

puede ser el cambio del valor de código de país, en donde en vez de exportar el valor en el formato ISO 3166-1 alpha-2[8] se exporte utilizando el nombre del país completo (es decir, “Argentina” en vez de “AR”).

Figura 5. Mecanismo de exportación de resultados

Por último, se habilitó la generación de reportes a partir de un conjunto de reportes predefinidos. Para posibilitar ésto, se creó un endpoint de consulta, que mediante ciertos parámetros recibidos retorna un archivo en formato JSON acorde al tipo de reporte solicitado. La *población de datos* utilizada para generar los reportes está restringida al conjunto de resultados retornados a partir de la búsqueda en el módulo.

Los reportes implementados para el prototipo fueron:

- Cantidad de registros (por IP, País, Ciudad, Continente, Tipo de registro, Tipo de Objeto DSpace).
- Visitas a publicaciones/Colecciones/Comunidades (por IP, País, Continente, Ciudad).
- Búsquedas en todo el repositorio/Colecciones/Comunidades (ídem arriba).
- Eventos de workflow (por IP, País, Continente, Ciudad).

Además, se agregó la capacidad de determinar un *lapso de tiempo* por reporte: mensual o anual.

Para la graficación (ver Figura 6), se utilizó la librería javascript llamada «c3js», que a partir de los datos devueltos por el endpoint de consulta genera distintos tipos de gráficas. c3js ofrece una API muy sencilla de utilizar y con la que se puede generar gráficas mediante pocas líneas de código. Está construida sobre otra librería más compleja llamada «d3js», que mediante el uso de tecnologías bien sustentadas como SVG, HTML5, y CSS permite producir infogramas dinámicos e interactivos en navegadores web.

[8] https://en.wikipedia.org/wiki/ISO_3166-1_alpha-2

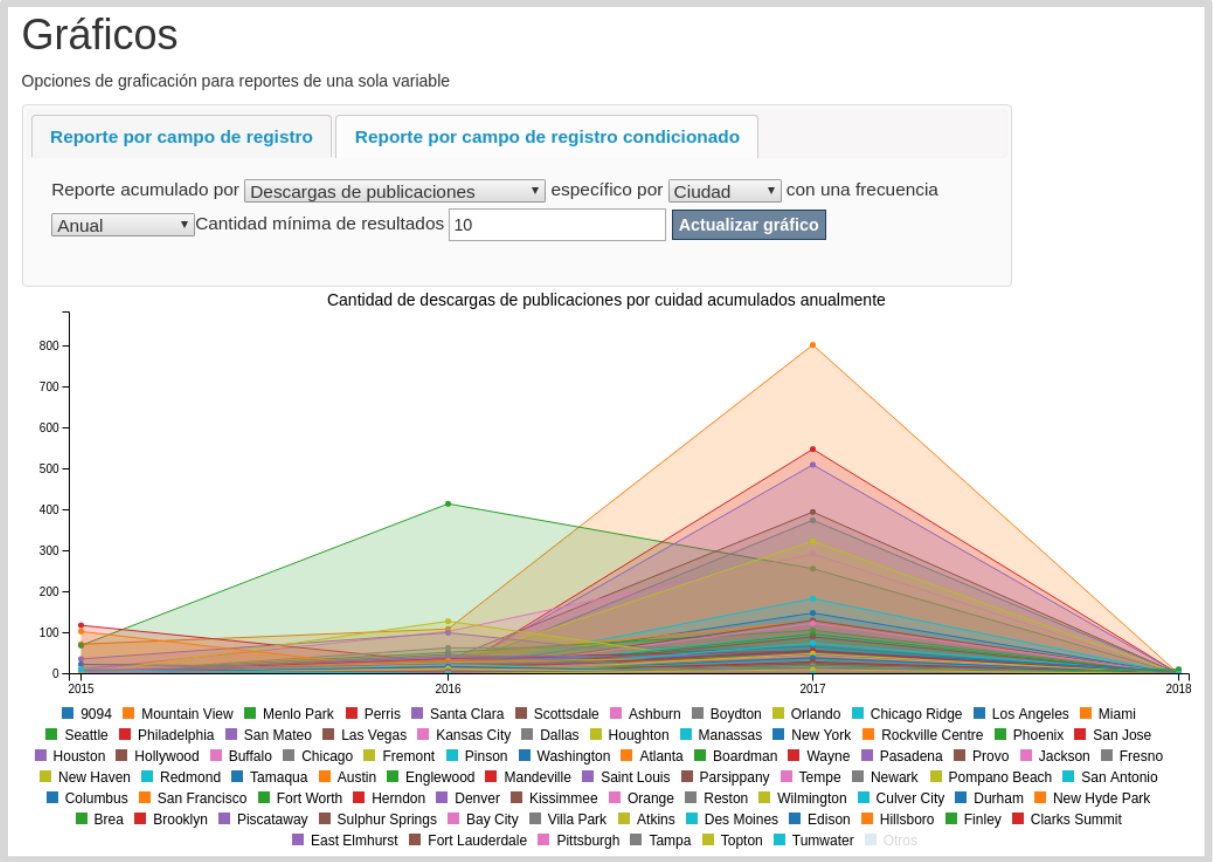

Figura 6. Generación de reportes y gráficas

Conclusiones

Luego de analizar las limitaciones del módulo de generación de reportes estadísticos en DSpace, y vista necesidad de explotar de mejor manera el cúmulo de datos de uso que constantemente son almacenados en el repositorio, se determinaron una serie de expectativas para la creación de un prototipo de herramienta para la exploración de los datos de uso en repositorios DSpace. Finalmente, se desarrolló un prototipo sobre la versión 6 de DSpace, cuyo modelo arquitectónico fue derivado a partir del módulo de búsqueda de contenidos llamado «Discovery». Mediante una serie de adaptaciones para su utilización sobre los datos de uso y el agregado de un conjunto de extensiones, se posibilitó: la búsqueda sobre registros de uso, le exportación de los resultados de búsqueda en formato de texto CSV y JSON, y la generación de reportes y gráficas, entre otras cosas. La herramienta fue integrada y probada sobre los datos de uso en el repositorio institucional CIC-Digital.

Las ventajas de disponer de esta herramienta en un repositorio son varias. Entre ellas se podría mencionar la fácil exploración de los datos de uso indexados en el repositorio (que pueden a llegar a crecer a una cantidad de millones de datos), y todo ésto sin requerir de conocimientos técnicos sobre Solr, la plataforma de indexación utilizada para almacenar estos datos. Además, la posibilidad de exportación en diversos formatos de los datos explorados desde la herramienta, permite la realización de análisis más avanzados y complejos en herramientas externas de mayor potencia y dedicadas al análisis estadístico de datos (como por ejemplo, Matlab o R). Otra de las ventajas visibles es la posibilidad de vinculación

entre los datos administrados por los módulos Discovery y Statistics-Discovery, de tal forma de poder determinar los registros de uso de los ítems resultantes de una búsqueda en Discovery; esto posibilita la rápida obtención de los datos de uso de los objetos en DSpace a partir de sus características (por ejemplo, para los ítems del tipo «Tesis de grado», los ítems del autor «Juan Pérez», o los ítems cuya año de publicación fue el 2018).

Queda como trabajo futuro la mejora de la herramienta en cuanto diversas cuestiones, entre ellas mejora de la performance, integración al sistema de permisos para el acceso a la utilización de la herramienta, agregado opciones de compartición de reportes, capacidad de depuración de los datos de uso almacenados en Solr a partir de inconsistencias detectadas mediante la herramienta.

Bibliografía

ANSI/NISO Z39.93-2014 The Standardized Usage Statistics Harvesting Initiative (SUSHI) Protocol. (2014) (American National Standards Institute). Recuperado de https://www.niso.org/publications/z3993-2014-sushi

Architecture - DSpace 6.x Documentation - DuraSpace Wiki. (s. f.). [Wiki]. Recuperado 12 de junio de 2018, de https://wiki.duraspace.org/display/DSDOC6x/Architecture

Bernal, I., & Pemau-Alonso, J. (2010). Estadísticas para repositorios: sistema métrico de datos en *Digital.CSIC*. *El Profesional de la Informacion*, *19*(5), 534-544. https://doi.org/10.3145/epi.2010.sep.15

counterart_cop_October2015.pdf. (s. f.). Recuperado de https://www.projectcounter.org/wp-content/uploads/2016/11/counterart_cop_October2015.pdf

Data Model - DSpace 6.x Documentation - DuraSpace Wiki. (s. f.). Recuperado 23 de agosto de 2018, de https://wiki.duraspace.org/display/DSDOC6x/Functional+Overview#FunctionalOverview-DataModel

De Giusti, M. R. (2017). Curso de posgrado: Bibliotecas y repositorios digitales. Tecnología y aplicaciones. Presentado en Curso de posgrado de repositorios digitales (Facultad de Informática, 2017). Recuperado de http://hdl.handle.net/10915/62871

Definición y tipos de repositorios. (s. f.). [PoliScience]. Recuperado 30 de julio de 2018, de http://poliscience.blogs.upv.es/open-access/repositorios/definicion-y-tipos/

Discovery - DSpace 6.x Documentation - DuraSpace Wiki. (s. f.). [Wiki]. Recuperado 12 de junio de 2018, de https://wiki.duraspace.org/display/DSDOC6x/Discovery

Fleming-May, R. A., & Grogg, J. E. (2010). Chapter 2: Standards, Tools, and Other Products. *Library Technology Reports*, *46*(6), 11-16.

Jacobs, N. (2016, mayo). What is a repository? | Jisc scholarly communications. Recuperado 22 de junio de 2018, de https://scholarlycommunications.jiscinvolve.org/wp/2016/05/31/what-is-a-repository/

Repositorios digitales - Red Infod. (s. f.). Recuperado 12 de junio de 2018, de https://red.infd.edu.ar/articulos/repositorios-digitales/

The COUNTER Code of Practice for e-Resources: Release 4. (2012, abril). Recuperado de https://www.projectcounter.org/wp-content/uploads/2016/01/COPR4.pdf

Web analytics. (2018). En *Wikipedia*. Recuperado de https://en.wikipedia.org/w/index.php?title=Web_analytics&oldid=841158888

Autores: resúmenes biográficos

Facundo G. Adorno es Licenciado en Sistemas por la Universidad Nacional de La Plata (UNLP), Argentina. Trabaja en el PREBI-SEDICI Proyecto de Enlace de Bibliotecas y Servicio de Difusión de la Creación Intelectual (SEDICI) de la UNLP desde 2013, y en el repositorio institucional CIC-Digital desde 2014. Es miembro del Comité Asesor del el Centro de Servicios en Gestión de Información (CESGI) desde 2016 y parte del personal del Observatorio Medioambiental La Plata. Es integrante del staff del Centro de Servicios en Gestión de Información (CESGI) y parte del personal del Observatorio Medioambiental La Plata (OMLP).

Marisa R. De Giusti es Doctora en Informática por la por la Universidad Nacional de La Plata, Argentina. También es Ingeniera en Telecomunicaciones y Profesora de Letras. Ha dirigido el Servicio de Difusión de la Creación Intelectual (SEDICI) de la UNLP desde 2003, y el repositorio CIC-Digital desde 2014, y el Centro de Servicios en Gestión de Información (CESGI) desde 2016. También coordina el Observatorio Medioambiental La Plata. Es profesora en la Facultad de Informática de la UNLP y sus intereses de investigación incluyen el campo de repositorios digitales, preservación digital y acceso abierto.

Ariel J. Lira es Licenciado en Sistemas por la Universidad Nacional de La Plata, Argentina. Trabaja en el PREBI-SEDICI Proyecto de Enlace de Bibliotecas y Servicio de Difusión de la Creación Intelectual (SEDICI) de la UNLP, y en el repositorio CIC-Digital desde 2014. Es miembro del Comité Asesor del el Centro de Servicios en Gestión de Información (CESGI) y parte del personal del Observatorio Medioambiental La Plata (OMLP).