Experimentación en el desarrollo de una herramienta para la exploración de datos estadísticos en DSpace

Adorno, Facundo Gabriel | Lira, Ariel Jorge | De Giusti, Marisa Raquel

Resumen extendido

INTRODUCCIÓN

Los repositorios digitales institucionales que nacen del Acceso Abierto son herramientas para el depósito, acceso y preservación de la producción académica de una Institución generalmente financiada con fondos públicos. A medida que la cantidad de material en el repositorio aumenta en volumen y antigüedad, también crece el conjunto de comunidades/colecciones, las interrelaciones entre sus objetos, y el acceso y uso por parte del internautas. La medición de la usabilidad de estos recursos disponibles en la web por parte del público resulta de importancia para que estas Instituciones demuestren el valor y el impacto de estos repositorios así como de la producción que alberga, ya que dicha medición puede servir como un factor más al momento de determinar la continuidad en el financiamiento de la producción en acceso abierto de la Institución. Considerando lo anterior, para este trabajo se pretende compartir los resultados de experimentación realizados sobre el repositorio institucional de la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires llamada CIC-DIGITAL en el desarrollo de un prototipo de herramienta que facilite el análisis de los «datos estadísticos» alojados en repositorios DSpace, es decir, datos derivados a partir del uso de los objetos alojados en DSpace y alojados en sistemas de almacenamiento Solr, como una manera de habilitar el acceso a la exploración y búsqueda sobre la usabilidad realizada en un repositorio DSpace.

MATERIALES Y METODOLOGÍA

DSpace, en su última versión estable (la versión 6.X al momento de escritura de este trabajo), ofrece un módulo de estadísticas 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. DSpace Statistics presenta una arquitectura cliente/servidor basada en una herramienta de indexación llamada «Solr», que recopila eventos de uso en las interfaces de usuario «JSPUI» y «XMLUI» de DSpace y los almacena en un índice o core Solr específico llamado «statistics». Luego de analizar la funcionalidad de este módulo, se determinó que el mismo presenta limitaciones que no permiten explotar en mayor profundidad los datos estadísticos indexados en Solr: no puede modificarse la cantidad de resultados por reporte, no puede especificarse libremente un rango de fechas como contexto temporal en la definición del reporte, se basan en un conjunto reducido de campos indexados y no permite especificar otros campos de interés, presenta características definidas mediante hardcoding, y no ofrece visualizaciones de los reportes out-of-the-box ya que requieren de una implementación específica.

Estas limitaciones del software en el módulo de estadísticas llevó al desarrollo de una herramienta ad hoc que solvente las mismas y permita una mayor flexibilidad al momento de generar reportes a partir de los datos estadísticos indexados en Solr. Esta herramienta fue desarrollada sobre el DSpace en su versión 6 y, en particular, sobre el software de la plataforma que subyace al repositorio institucional CIC-Digital. La funcionalidad de esta herramienta está inspirada principalmente en la metodología de búsqueda definida por el módulo de búsqueda «Discovery» en DSpace, el cual mediante un conjunto de componentes visuales (caja de búsquedas, scopes, filtros, facets, opciones de ordenamiento, paginados, etc.) permite la exploración del core «search» en Solr, un índice de datos generado a partir de los objetos troncales de DSpace

(Comunidades, Colecciones, Ítems, y Bitstreams) y de los metadatos que los componen. La mayoría de estos componentes fueron reutilizados en la implementación de la herramienta propia, los cuales son configurados mediante beans «Spring» (framework para la construcción de aplicaciones de «Inversión de Control» o IoC) en archivos de configuración XML, lo cual brinda mayor flexibilidad y extensibilidad a la herramienta al momento de su personalización.

RESULTADOS Y DISCUSIÓN

Previo a la implementación del prototipo fue necesario un análisis para determinar la configuración de tecnologías a utilizar sobre la éste se desarrollaría, analizando ventajas y desventajas de la plataforma DSpace en su versión 7 y en su versión 6. Luego de decidir la base tecnológica, y con miras de la reutilización de algunas de las herramientas provistas por el software, se decidió que la herramienta fuera diseñada a partir del modelo definido por el módulo Discovery: módulo que a lo largo de los continuos lanzamientos de DSpace fue mejorada y optimizada acorde a las necesidades de búsqueda. Sin embargo, el modelo derivado de Discovery tuvo que ser redefinido y adecuado en diferentes secciones para que funcione específicamente sobre el core de datos estadísticos y atienda las necesidades concretas para la exploración de los mismos. Además, se agregaron funcionalidades de exportación y generación de gráficas sencillas.

CONCLUSIONES

Concretamente, la herramienta se desarrolló sobre la versión 6 del software DSpace mediante la creación de un módulo específico basado en el módulo Discovery de la plataforma, y la utilización de las tecnologías que funcionan sobre esta versión: Solr, Apache Cocoon, Spring Framework, XSLT, Javascript, entre otras. Si bien la alternativa de la versión 7 del software era la mejor a largo plazo por sus ventajas a nivel tecnológico, ésta todavía se encuentra en una etapa reciente de desarrollo y sin una fecha definitiva de lanzamiento. Por último, la herramienta fue probada en el repositorio institucional CIC-Digital, que se encuentra en la versión 6 de la línea de desarrollo de DSpace.

Entre las funcionalidad desarrollada por la herramienta se encuentra: la exploración y búsqueda mediante aplicación de términos de búsquedas, filtros, y facets, se agregaron funcionalidades de exportación de resultados en formatos JSON y CSV mediante un modelo de clases extensible (es decir, con posibilidad de implementaciones propias a otros formatos), de tal manera de posibilitar la realización de análisis más específicos por parte del usuario en herramientas estadísticas más potentes (por ejemplo, R o Matlab), y se agregó la generación de gráficas basados en los resultados de búsquedas mediante el uso de la librería Javascript «c3.js» y un endpoint de generación de reportes JSON.

Finalmente, se describen algunos de los problemas encontrados durante la generación de la herramienta, entre ellos: problemas generados por la definición del esquema de campos en el core «statistics» en Solr, problemas derivados a partir de la extensión de las consultas realizadas a Solr mediante su interfaz HTTP.