Facultad de INFORMÁTICA

UNIVERSIDAD NACIONAL DE LA PLATA

Reportes estadísticos para repositorios digitales desarrollados en DSpace

Autor Adorno, Facundo Gabriel

Directora De Giusti, Marisa Raquel

Asesor Profesional Lira, Ariel Jorge

Objetivo

Implementar una **herramienta** que asista en el **análisis** del uso de un repositorio digital basado en DSpace.

Los elementos del repositorio a analizar típicamente son ítems, comunidades, colecciones y bitstreams.

Las facilidades brindadas por la herramienta propuesta serían:

- Realización de **búsquedas** en los registros de uso.
- **Generación de gráficos** basados en diversos **reportes predefinidos** a partir de los resultados de búsqueda**,** y
- **Exportación** de registros de búsqueda en diversos formatos de texto (JSON, CSV, etc.) para posibilitar su posterior consumo en aplicaciones externas**.**

Marco teórico

- Repositorios
- DSpace
- Estadísticas

Repositorio Institucional

Es un archivo digital provisto de un conjunto de servicios web centralizados encargados de organizar, gestionar, preservar y ofrecer acceso libre a los recursos derivados de la producción científica y académica de una institución.

DSPACE

Es un software de código abierto pensado para la gestión de repositorios digitales que proporciona distintas herramientas y funcionalidades que permiten satisfacer las diferentes necesidades que requieren las instituciones.

SEDICI
sedici.unlp.edu.ar

Su misión es albergar, preservar, difundir y dar visibilidad a nivel mundial a toda la producción científica e intelectual de las distintas unidades académicas que componen la UNLP.

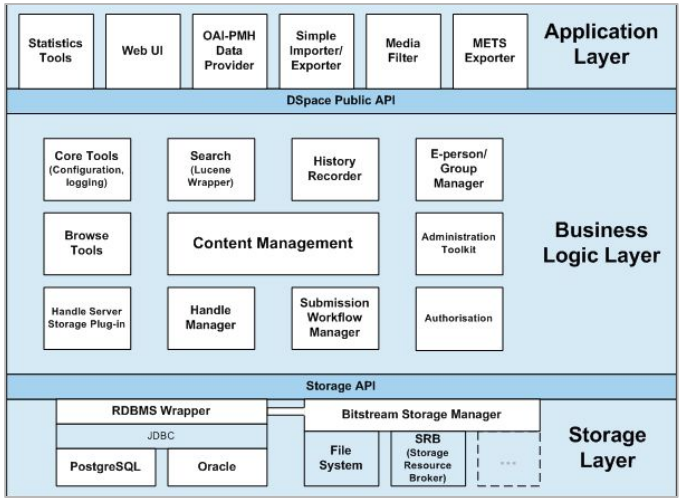
CICDIGITAL

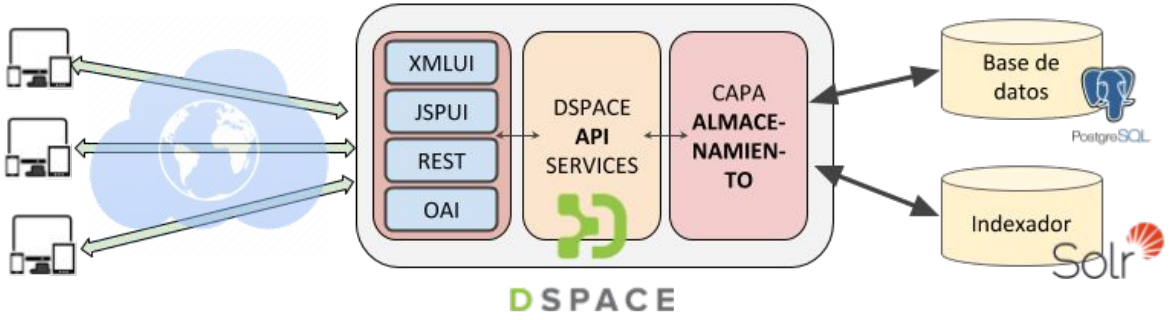
Su misión es reunir, registrar, divulgar, preservar y dar acceso público a toda la producción científico-tecnológica y académica de la CICBA.

Arquitectura en DSpace

La arquitectura en DSpace se divide en 3 grupos:

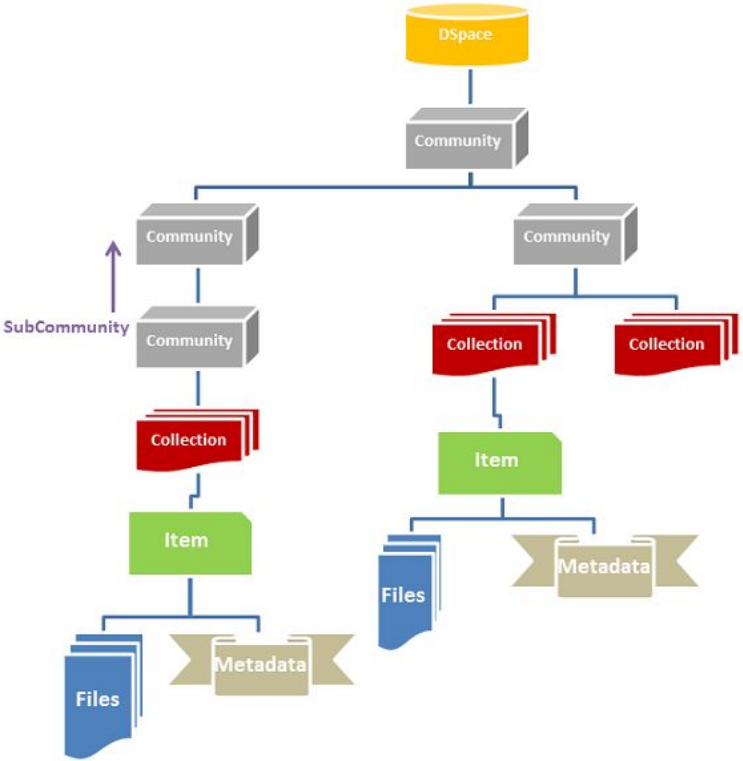
- **Módulos de Aplicación** (XMLUI, JSPUI, OAI-PMH, REST-API, etc.).
- **Módulos de Lógica de negocio** (DSpace API basada en servicios).
- **Módulos de Almacenamiento** (Base de datos, Solr, Assetstore, etc.).





Modelo de contenidos en DSpace

1. El repositorio se organiza en una o más **comunidades** de nivel base
   a. Cada comunidad se organiza jerárquicamente en subcomunidades.
   b. Son asimilables a *espacios de trabajo*
2. Las **colecciones** son los “estantes” dentro de las comunidades, que agrupan contenido relacionado.
3. Los **ítems** son las obras que van en los estantes y que se pretende que el público encuentre.
4. Los **metadatos** describen al recurso
5. Los **bitstreams** son la representación digital del recurso.



Estadísticas - ¿Que se mide?

Las estadísticas son una herramienta clave a la hora de medir un repositorio en aspectos como:

- crecimiento de sus contenidos,
- comportamiento de sus usuarios, y
- uso de sus servicios y contenidos

La interpretación de estos datos ayuda a la toma de decisiones para los directores de un repositorio y las autoridades de la institución.

La medición del uso del sitio por parte de los usuarios forma parte de un área de análisis mayor llamada «Web Analytics».

Web Analytics

Es el **estudio del comportamiento** de los visitantes de un sitio web.

Realiza la medición, recopilación, análisis y generación de informes de datos generados en torno al **uso** de un sitio web.

Dispone de diversas técnicas o **herramientas de recolección**: *log analyzers, page tagging, geolocalización de visitantes, click analytics*, etc.

Busca comprender y optimizar los servicios provistos por un sitio web a través de distintos **indicadores**, por ejemplo:

- Hits
- Page Views
- Page View Duration
- Click
- Click Path
- Downloads

Estadísticas de uso en DSpace

DSpace almacena algunos eventos en la interacción entre el usuario y el repositorio a través de las interfaces de aplicación **JSPUI** y **XMLUI**.

Por defecto, se registran eventos relacionados a

- **búsquedas** (en Discovery),
- **vistas** (de Comunidades, Colecciones, e Items),
- **descargas** (de Bitstreams) y
- **workflow** (pasos ejecutados durante el envío de nuevos ítems)

Módulo Statistics

Módulo encargado de indexar los eventos de uso del repositorio.
Dispone de 2 funciones básicas:

- **Registro de eventos en un índice Solr**
- **Generación de reportes:**
  - **Accesos y Descargas** (por Ítem, por Comunidad, por Colección, y totales en el sitio)
  - **Búsquedas** (realizadas en Comunidades, Colecciones y totales en el sitio)
  - **Workflow** (para los pasos activados en los flujos de trabajo)

Módulo Statistics

Términos de Búsqueda mas usados

Total

	Término de Búsqueda	Búsquedas	% del total
1		1074	8.30%
2	has_content_in_original_bundle_keyword:true	1023	7.91%
3	subject_keyword:keyword1	801	6.19%
4	subject_keyword:keyword2	719	5.56%
5	subject_keyword:keyword3	638	4.93%
6	dateIssued_keyword:[1900 TO 1999]	498	3.85%
7	author_keyword:Cat, Lily	441	3.41%
8	subject_keyword:cat	354	2.74%
9	author_keyword:Doe, Jane L	322	2.49%
10	dateIssued_keyword:[1650 TO 1699]	318	2.46%

Total

Búsquedas	% del total	Páginas Vistas / Búsquedas
12940	100.00%	0.12

Limitaciones del módulo Statistics de DSpace

- Los reportes retornan sólo 10 resultados.
- No se puede seleccionar un rango de fecha arbitrario o mayor a un año de antigüedad.
- No permite inspeccionar otros aspecto de los datos de uso indexados más que los que los reportes indican.
- No permite exportar los datos de uso involucrados en un reporte para su posterior evaluación en sistemas estadísticos externos.
- No se ofrecen visualizaciones (gráficas) o*ut-of-the-box* de los reportes generados.
    - Sólo tablas
- Presenta *hardcoding* de algunos datos que podrían estar en configuraciones externas, entre ellos:
    - Rango de tiempo del reporte
    - Cantidad de Filas en tablas por reporte
    - Los filtros que determinan el dataset por reporte

Prototipo desarrollado

- Análisis
- Especificación de requerimientos
- Diseño
- Implementación

Propuesta

El prototipo a implementar debía cubrir una serie de expectativas

EXPLORACIÓN/BÚSQUEDA DE REGISTROS

TIEMPOS DE BÚSQUEDA RAZONABLES

MÚLTIPLES CONTEXTOS DE BÚSQUEDA

EXPORTACIÓN DE REGISTROS

GENERACIÓN DE GRÁFICAS

MAYORMENTE CONFIGURABLE

Tecnologías

Las tecnologías utilizadas fueron

- Apache Cocoon + XSLT + Javascript para la vista (XMLUI)
- JSolr (librería Java) para comunicación con Solr
- Apache Solr para la indexación/recuperación de datos estadísticos

Finalmente se decidió implementar el prototipo sobre DSpace en su versión 6.

Módulo Discovery

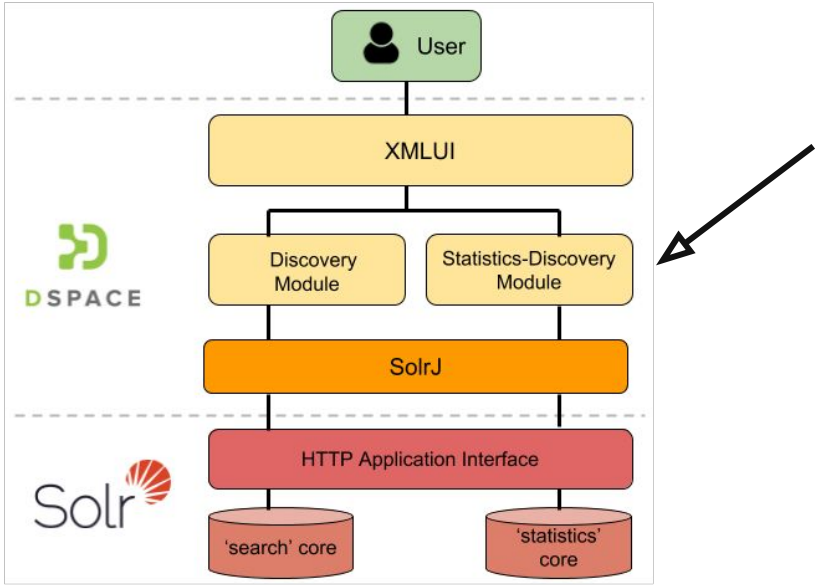
El prototipo se basó en el funcionamiento del módulo Discovery de DSpace, el cual implementa:

- **Servicio** para la búsqueda en el índice «search» en Solr
- Búsqueda de registros a partir de **filtros** (p.e. por fecha, por autor, etc)
- Refinamiento mediante **facets**
- **Paginación** de resultados
- **Ordenamiento** de resultados de búsqueda
- **Contextos de búsqueda** a nivel Comunidad y Colección

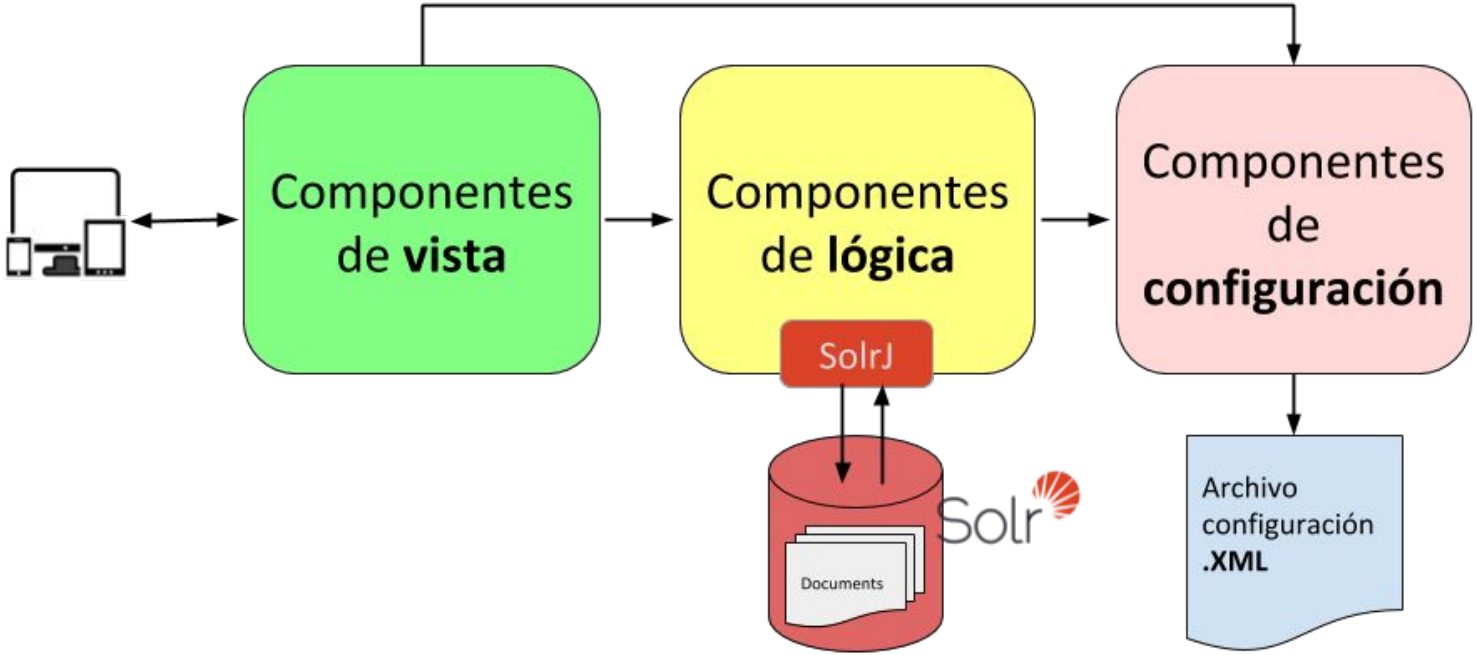
Módulo Discovery

Diseño - Arquitectura

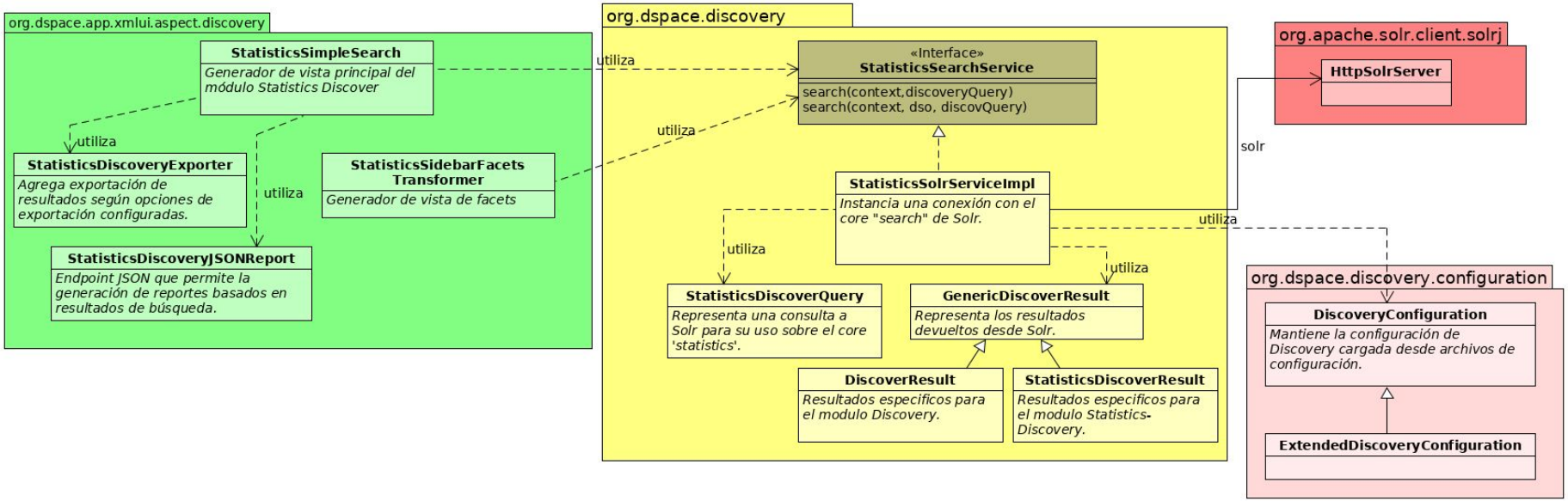
El prototipo funciona de una manera similar a como lo hace Discovery.



Diseño - Separación en capas



Modelo

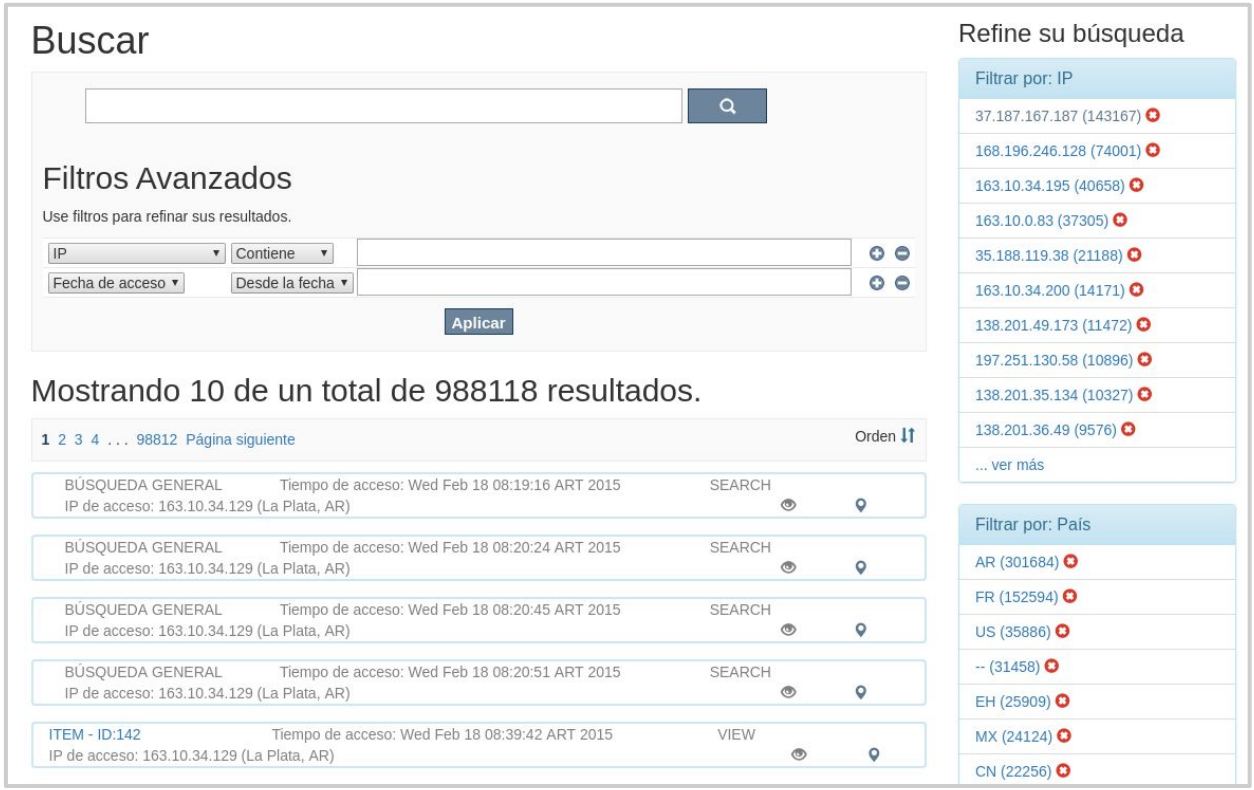

org.dspace.app.xmlui.aspect.discovery
StatisticsSimpleSearch
Generador de vista principal del módulo Statistics Discover
utiliza
StatisticsDiscoveryExporter
Agrega exportación de resultados según opciones de exportación configuradas.
utiliza
StatisticsSidebarFacets Transformer
Generador de vista de facets
StatisticsDiscoveryJSONReport
Endpoint JSON que permite la generación de reportes basados en resultados de búsqueda.
org.dspace.discovery
utiliza
«Interface»
StatisticsSearchService
search(context,discoveryQuery)
search(context, dso, discovQuery)
utiliza
StatisticsSolrServiceImpl
Instancia una conexión con el core "search" de Solr.
utiliza
utiliza
StatisticsDiscoverQuery
Representa una consulta a Solr para su uso sobre el core 'statistics'.
GenericDiscoverResult
Representa los resultados devueltos desde Solr.
DiscoverResult
Resultados especificos para el modulo Discovery.
StatisticsDiscoverResult
Resultados especificos para el modulo Statistics-Discovery.
org.apache.solr.client.solrj
HttpSolrServer
solr
utiliza
org.dspace.discovery.configuration
DiscoveryConfiguration
Mantiene la configuración de Discovery cargada desde archivos de configuración.
ExtendedDiscoveryConfiguration

Extensiones al modelo base

Se crearon las siguientes **extensiones al modelo original** de Discovery para cumplir con las expectativas de la herramienta:

- Diversos contextos de búsqueda
- Filtros y facets
- Exportación de resultados en diversos formatos de texto
- Generación de diversos reportes y gráficas
- Utilidades de vistas

1.Herramienta de búsqueda



1. Herramienta de búsqueda - Contextos

- Se permitió definir como contextos de búsquedas una comunidad, una colección o un ítem.
- Además se agregó la capacidad de definir un conjunto de objetos DSpace como contexto, resultantes a partir de una consulta Discovery.
  a. *Por ejemplo: los ítems cuyo autor sea “Juan Perez”*

1. Herramienta de búsqueda - Filtros y Facets

Las búsquedas de registros en Solr se hacen a partir de filtros y facets.

Se configuraron filtros por:

- IP
- Ciudad
- Código de país
- Agente de usuario
- Tipo de estadística
- Referer
- Tipo de objeto DSpace (combinado)
- Código de Continente

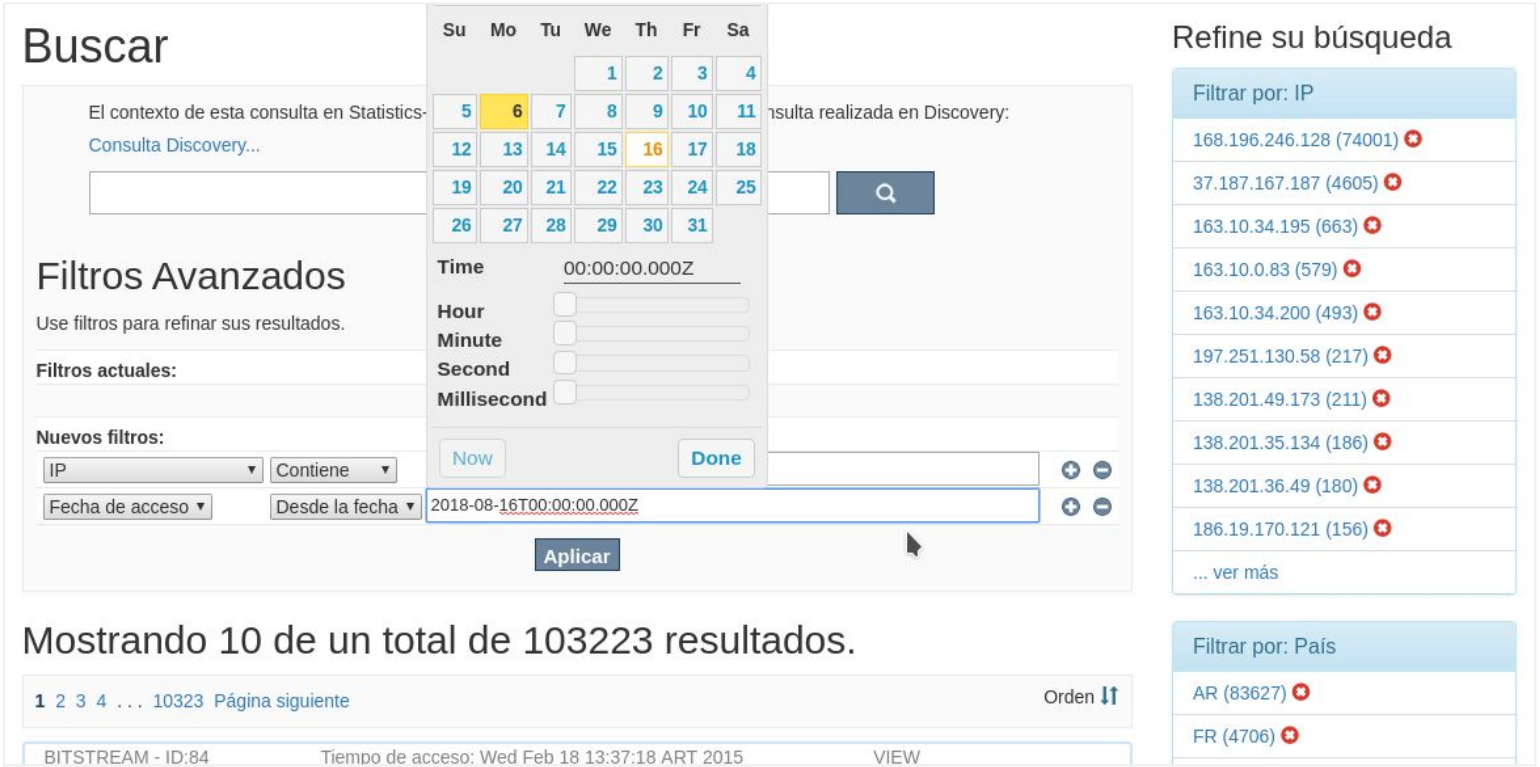
Se crearon **filtros por fecha** sobre el campo «fecha de acceso»

- Se definen operadores especiales *«desde la fecha»* y *«hasta la fecha»*

Se crearon **facets** para el refinado

- IP
- Código de país
- Tipo de registro
- Tipo de objeto Dspace
- Fecha de acceso (combinado)

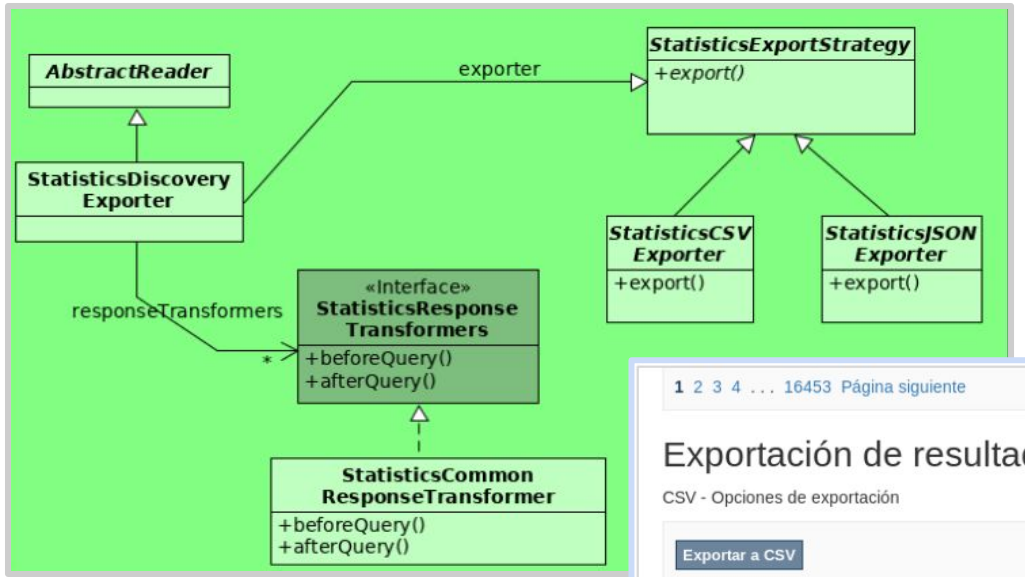
1. Herramienta de búsqueda - Filtros y Facets

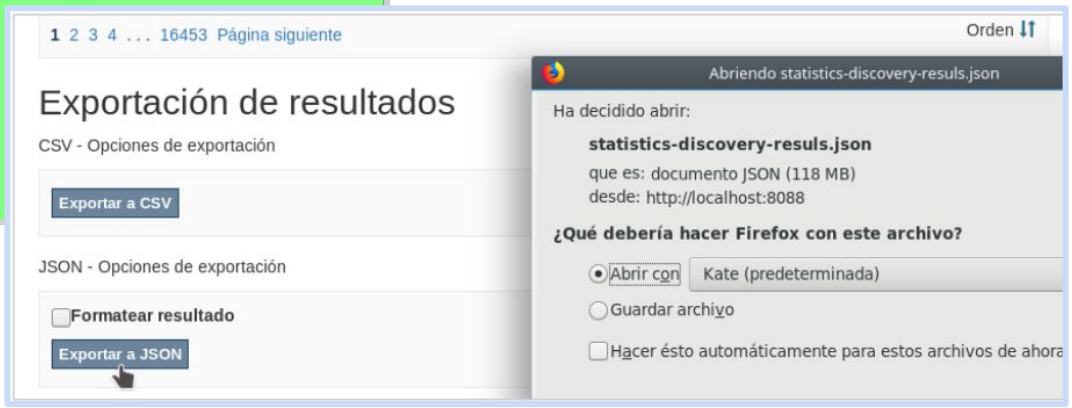


2. Exportación de resultados

Se implementó un modelo extensible para la exportación de registr

2. Exportación de resultados





3. Generación de reportes

Se creó un **endpoint de consulta JSON** para la generación de reportes predefinidos.
Los reportes hasta ahora implementados son:

- *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)

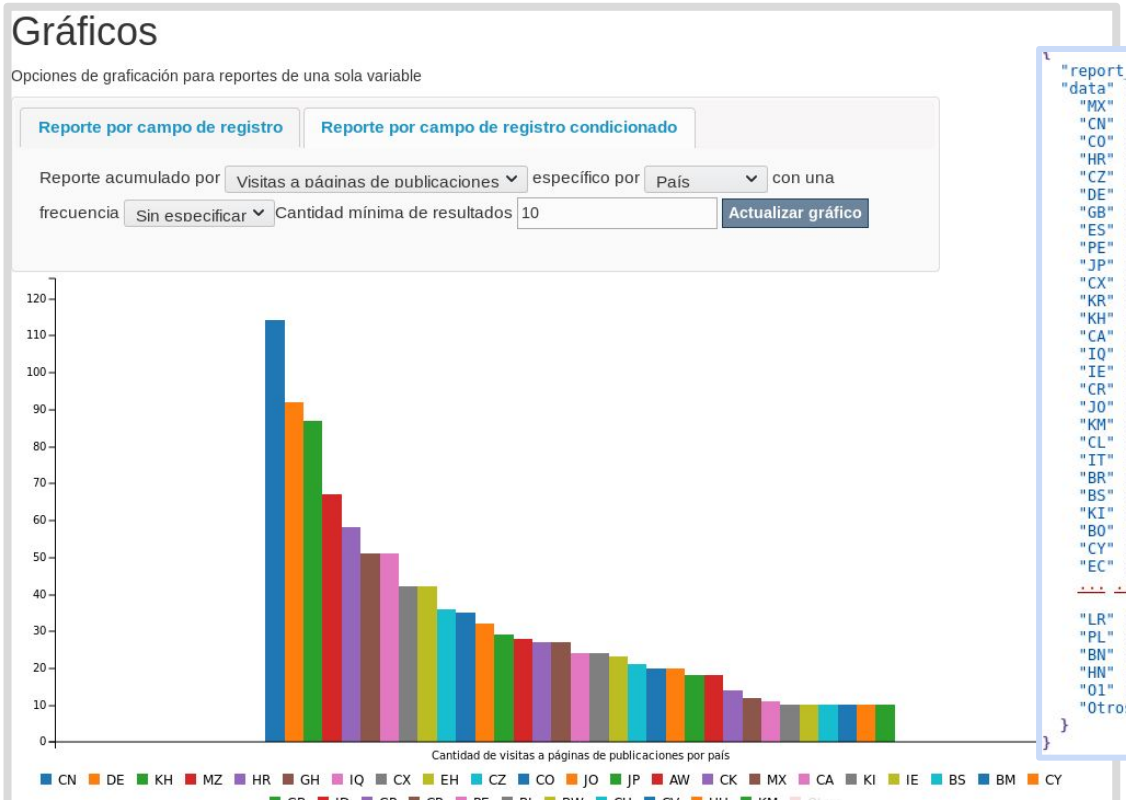
Se agregó capacidad para determinar un **lapso de tiempo** por reporte: *mensual* o *anual*.

La población de datos para la generación de reportes se restringe a los resultados de búsqueda.

Se utiliza la librería javascript **c3.js** para la generación de las gráficas.

C3.js

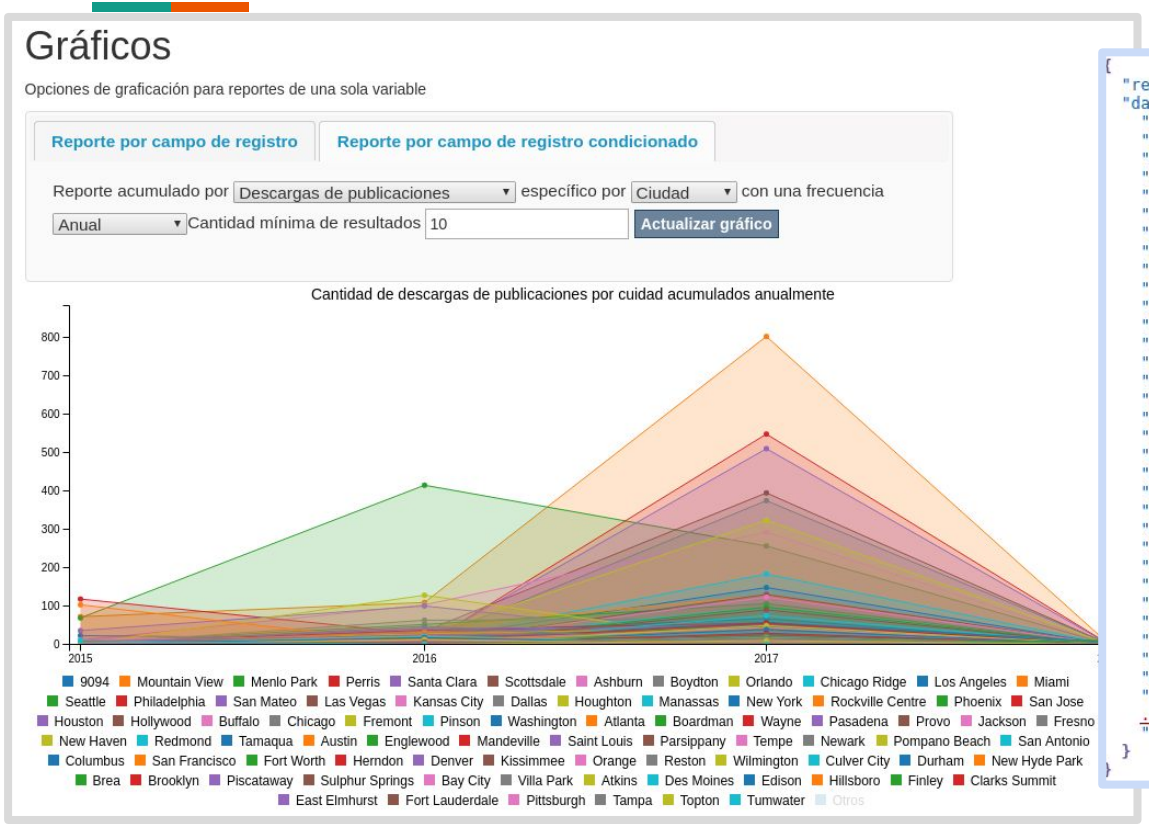
3. Generación de reportes




```
{
  "report_name" : "Cantidad de descargas de publicaciones por país",
  "data" : {
    "MX" : [ "349" ],
    "CN" : [ "342" ],
    "CO" : [ "316" ],
    "HR" : [ "298" ],
    "CZ" : [ "276" ],
    "DE" : [ "272" ],
    "GB" : [ "257" ],
    "ES" : [ "253" ],
    "PE" : [ "233" ],
    "JP" : [ "211" ],
    "CX" : [ "195" ],
    "KR" : [ "191" ],
    "KH" : [ "175" ],
    "CA" : [ "165" ],
    "IQ" : [ "163" ],
    "IE" : [ "152" ],
    "CR" : [ "151" ],
    "JO" : [ "151" ],
    "KM" : [ "146" ],
    "CL" : [ "142" ],
    "IT" : [ "142" ],
    "BR" : [ "125" ],
    "BS" : [ "123" ],
    "KI" : [ "113" ],
    "BO" : [ "97" ],
    "CY" : [ "89" ],
    "EC" : [ "89" ],
    ... ... ... ...
    "LR" : [ "11" ],
    "PL" : [ "11" ],
    "BN" : [ "10" ],
    "HN" : [ "10" ],
    "O1" : [ "10" ],
    "Otros" : [ "1244" ]
  }
}
```


Respuesta generada desde el ENDPOINT JSON

3. Generación de reportes




```
{
  "report_name" : "Cantidad de descargas de publicaciones por cuidad acu
  "data" : {
    "dateLabel" : [ "2015-03-03", "2016-03-03", "2017-03-03", "2018-03-0
    "Mountain View" : [ "72", "109", "801", "0" ],
    "Menlo Park" : [ "69", "414", "256", "0" ],
    "Perris" : [ "0", "0", "547", "0" ],
    "Santa Clara" : [ "1", "3", "509", "0" ],
    "Scottsdale" : [ "0", "34", "394", "0" ],
    "Ashburn" : [ "11", "103", "292", "0" ],
    "Boydton" : [ "0", "9", "374", "0" ],
    "Orlando" : [ "0", "25", "323", "0" ],
    "Chicago Ridge" : [ "0", "10", "183", "0" ],
    "Los Angeles" : [ "23", "14", "148", "0" ],
    "Miami" : [ "1", "43", "128", "0" ],
    "Seattle" : [ "4", "52", "106", "0" ],
    "Philadelphia" : [ "118", "27", "12", "0" ],
    "San Mateo" : [ "36", "100", "6", "0" ],
    "Las Vegas" : [ "4", "5", "130", "0" ],
    "Kansas City" : [ "4", "5", "123", "0" ],
    "Dallas" : [ "1", "63", "66", "0" ],
    "Houghton" : [ "0", "128", "0", "0" ],
    "Manassas" : [ "6", "8", "97", "0" ],
    "New York" : [ "7", "30", "67", "0" ],
    "Rockville Centre" : [ "103", "0", "0", "0
    "Phoenix" : [ "1", "1", "97", "0" ],
    "San Jose" : [ "3", "37", "56", "0" ],
    "Houston" : [ "5", "41", "46", "0" ],
    "Hollywood" : [ "0", "1", "90", "0" ],
    "Buffalo" : [ "11", "21", "43", "0" ],
    "Chicago" : [ "6", "20", "49", "0" ],
    "Fremont" : [ "4", "28", "43", "0" ],
    "Pinson" : [ "0", "0", "75", "0" ],
    "Washington" : [ "1", "11", "57", "6" ],
    "Atlanta" : [ "3", "32", "35", "0" ],
    ... ... ... ...
    "Tumwater" : [ "10", "0", "0", "0" ]
  }
}
```


Respuesta generada desde el ENDPOINT JSON

4. Otras utilidades

Botones de rápida vinculación con las estadísticas.

Botón para visualización completa de un registro.

Botón de vinculación con los datos geo-espaciales y el servicio de mapas OpenStreetMap.

Conclusiones

Resultados

Se desarrolló un prototipo de herramienta para el análisis del uso en repositorios DSpace.

Se analizaron las alternativas de implementación entre DSpace 6 y DSpace 7.

Finalmente, el prototipo se implementó en DSpace 6, basándose en el módulo Discovery.

Se pudo implementar con éxito la mayoría de las funcionalidades propuestas, permitiendo:

- Exploración fluida de los datos de uso mediante sencillas búsquedas a través de la interfaz.
- Aplicación de filtros y facets sobre registros de uso
- Exportación de registros para uso en entornos estadísticas externos
- Fácil vinculación entre objetos DSpace y sus respectivos datos de uso
- Generación de reportes y gráficas

Problemas encontrados

Entre algunos de los problemas encontrados, la mayoría derivan de la configuración de **Solr** en DSpace:

- Distintos campos que representan el mismo dato. Caso de *type* y *dsoType.*
- Errores en la definición de campos para el índice «statistics» en Solr.
- Problemas con la longitud de las peticiones a Solr.

Trabajos futuros

- Definir permisos para el acceso al servicio mediante una capa de seguridad.
- Permitir la compartición de los reportes generados.
  - Por correo, embebimiento HTML, botones sociales, exportación
- Ampliar la cantidad de reportes a generar así como los tipos de gráficas.
- Agregar datos estadísticos (frecuencias, media, mediana, etc) calculados a partir de los registros.
- Generar estadísticas de crecimiento del repositorio.
- Integrar una herramienta para la depuración de registros estadísticos.
- Aportar el código de la herramienta a la comunidad de DSpace.
- Migrar la herramienta a DSpace 7.

Facultad de INFORMÁTICA

UNIVERSIDAD NACIONAL DE LA PLATA

¡Muchas Gracias!

¿Preguntas?

Facundo Gabriel Adorno