XXVIII Asamblea General del ISTEC 2018
Congreso Internacional
"La influencia de la tecnología en las comunidades del conocimiento"

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

Adorno, Facundo G. (Presentador)
De Giusti, Marisa R.
Lira, Ariel J.

Repositorios digitales

Este trabajo fue desarrollado sobre el software para repositorios institucionales llamado **DSpace.**

DSpace es un software de código abierto desarrollado en Java.

En particular, se creó un prototipo para la exploración y análisis de los **datos de uso** alojados en el repositorio institucional **CIC-Digital**.

CIC-Digital es un repositorio creado sobre la plataforma DSpace.

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».

Estadísticas - 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.

Matomo

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

CIC-Digital

CIC-Digital es el repositorio institucional de la *Comisión de Investigaciones Científicas de la Provincia de Buenos Aires* (CICBA).

Almacena y preserva toda la producción científica-tecnológica de la CICBA:

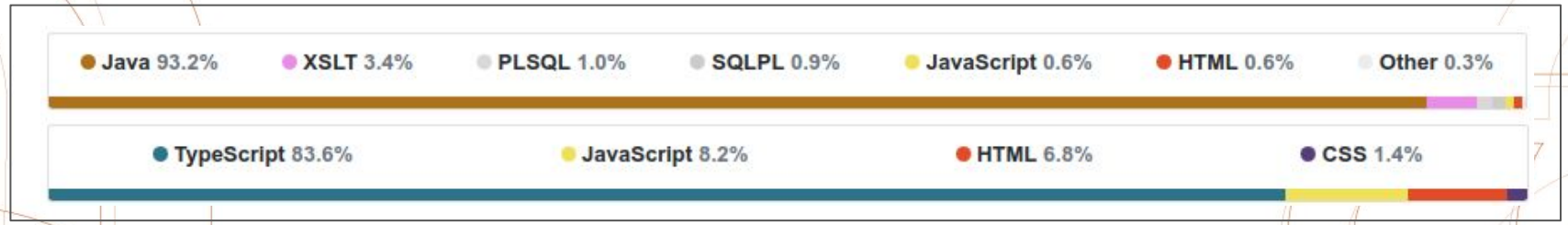
- Informes de investigadores, personal de apoyo y becarios, Tesis de grado y posgrado, Artículos, Publicaciones en congresos, etc.

Compuesto por más de

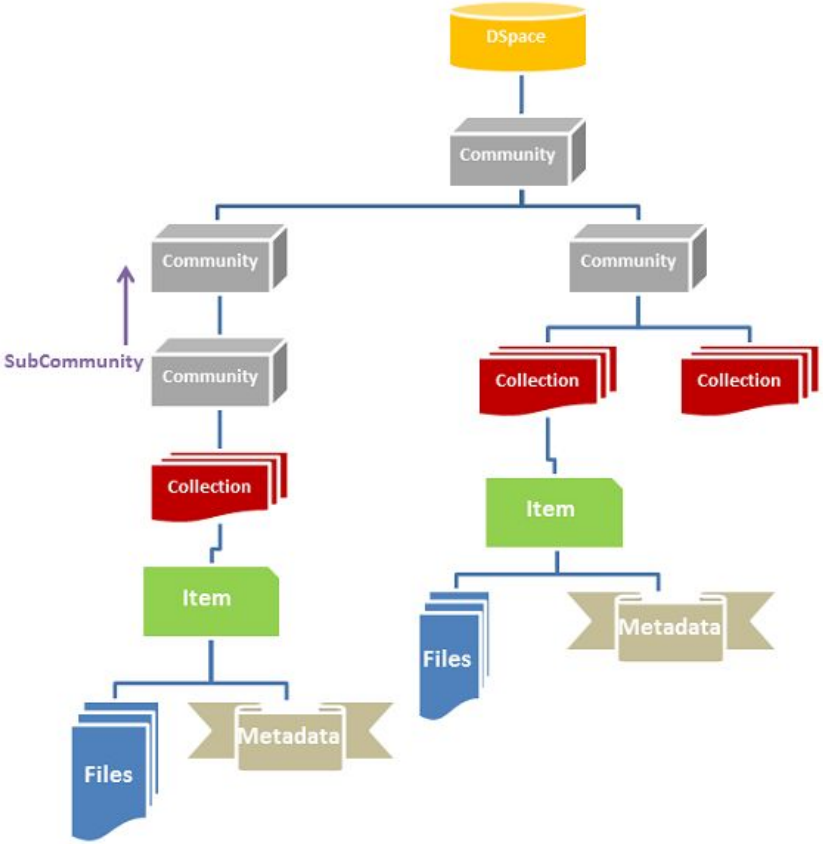
- 7000 items
- 400 colecciones
- 200 comunidades

DSpace - Características

- Software libre y open-source
- Proyecto escrito mayoritariamente en lenguaje **JAVA**.
- Desarrollado y mantenido por una amplia **comunidad** de usuarios en todo el mundo
- Altamente **configurable**
- Puede ser fácilmente **extendido**
- Dispone de una gran cantidad de funcionalidades
- Posee un **modelo** de datos simple, con metadatos no jerárquicos e independencia de los formatos de archivos.



DSpace - Modelo de Contenidos



1. El repositorio se organiza en una o más **comunidades** de nivel base que se organizan jerárquicamente en subcomunidades.
    - Son como 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.

DSpace - Modelo de Contenidos

Módulo Statistics

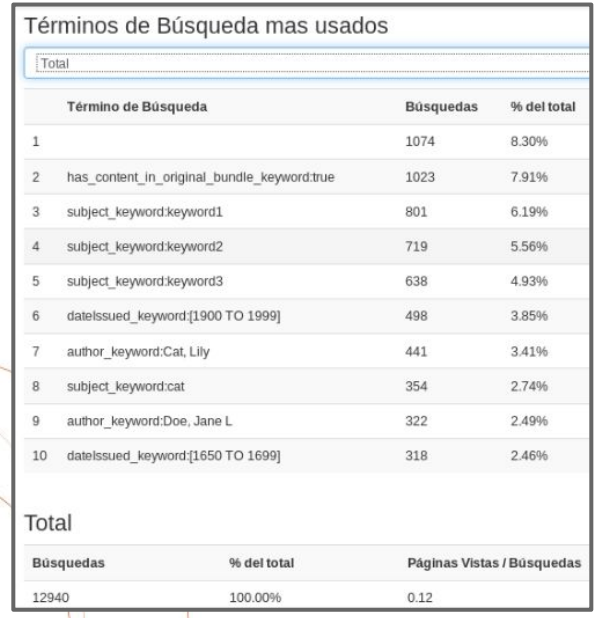
DSpace almacena algunos eventos en la interacción entre el usuario y el repositorio a través de las interfaces de usuario (UI).

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)

El encargado de realiza de almacenar estos eventos es el módulo **Statistics.**

Módulo Statistics



Términos de Búsqueda mas usados

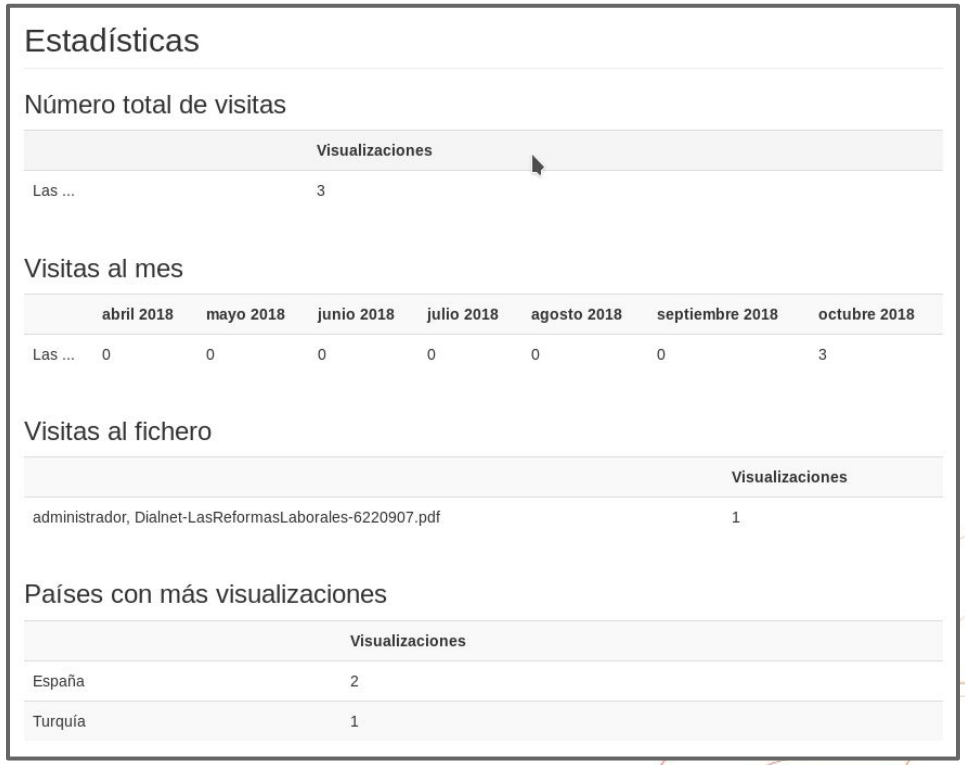
Total

	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

REPORTE DE BÚSQUEDA GLOBAL



Estadísticas

Número total de visitas

	Visualizaciones
Las ...	3

Visitas al mes

	abril 2018	mayo 2018	junio 2018	julio 2018	agosto 2018	septiembre 2018	octubre 2018
Las ...	0	0	0	0	0	0	3

Visitas al fichero

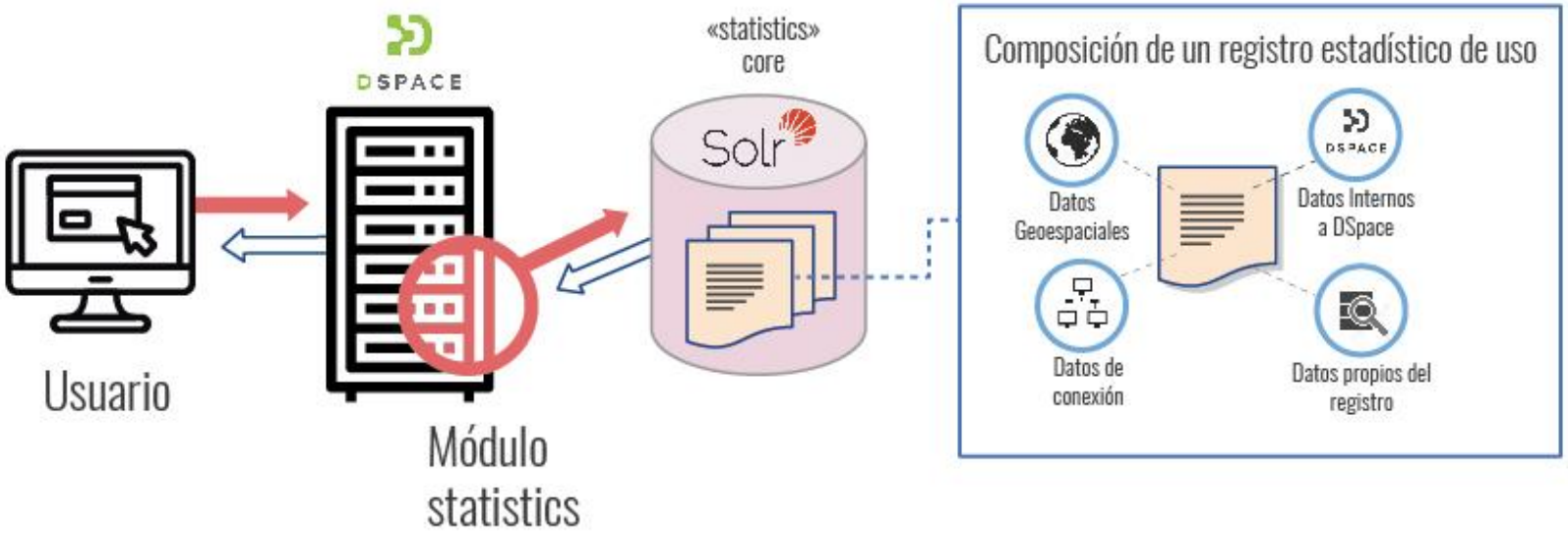
	Visualizaciones
administrador, Dialnet-LasReformasLaborales-6220907.pdf	1

Países con más visualizaciones

	Visualizaciones
España	2
Turquía	1

REPORTE VISUALIZACIÓN DE ITEM

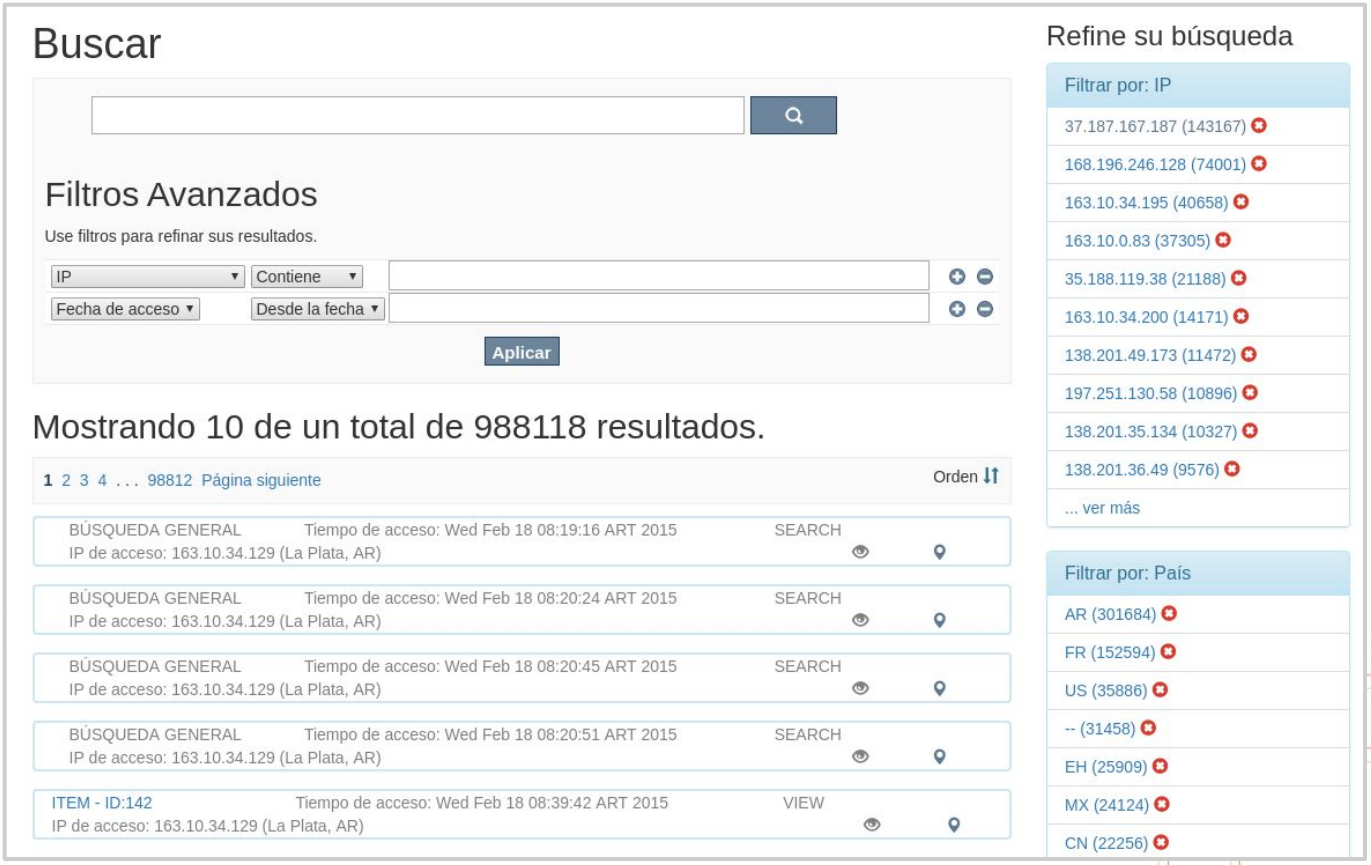
Módulo Statistics - Funcionamiento



Módulo Statistics - Limitaciones

- 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 - Módulo Statistics-Discovery



Statistics-Discovery

Se decidió crear un nuevo módulo experimental que permita facilitar la **exploración** y el **análisis** de los datos de uso en DSpace.

Entre sus funcionalidades implementadas están

- Búsqueda de registros de uso
- Aplicación de contextos de búsqueda
- Exportación de registros en diversos formatos textuales
- Generación de reportes y gráficas basadas en los registros

Este módulo está basado en el módulo de búsqueda de DSpace, llamado **Discovery**.

Statistics-Discovery - Tecnologías utilizadas

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.

Búsqueda de registros - Contextos y Filtros

Se permitió definir como **contextos de búsquedas**

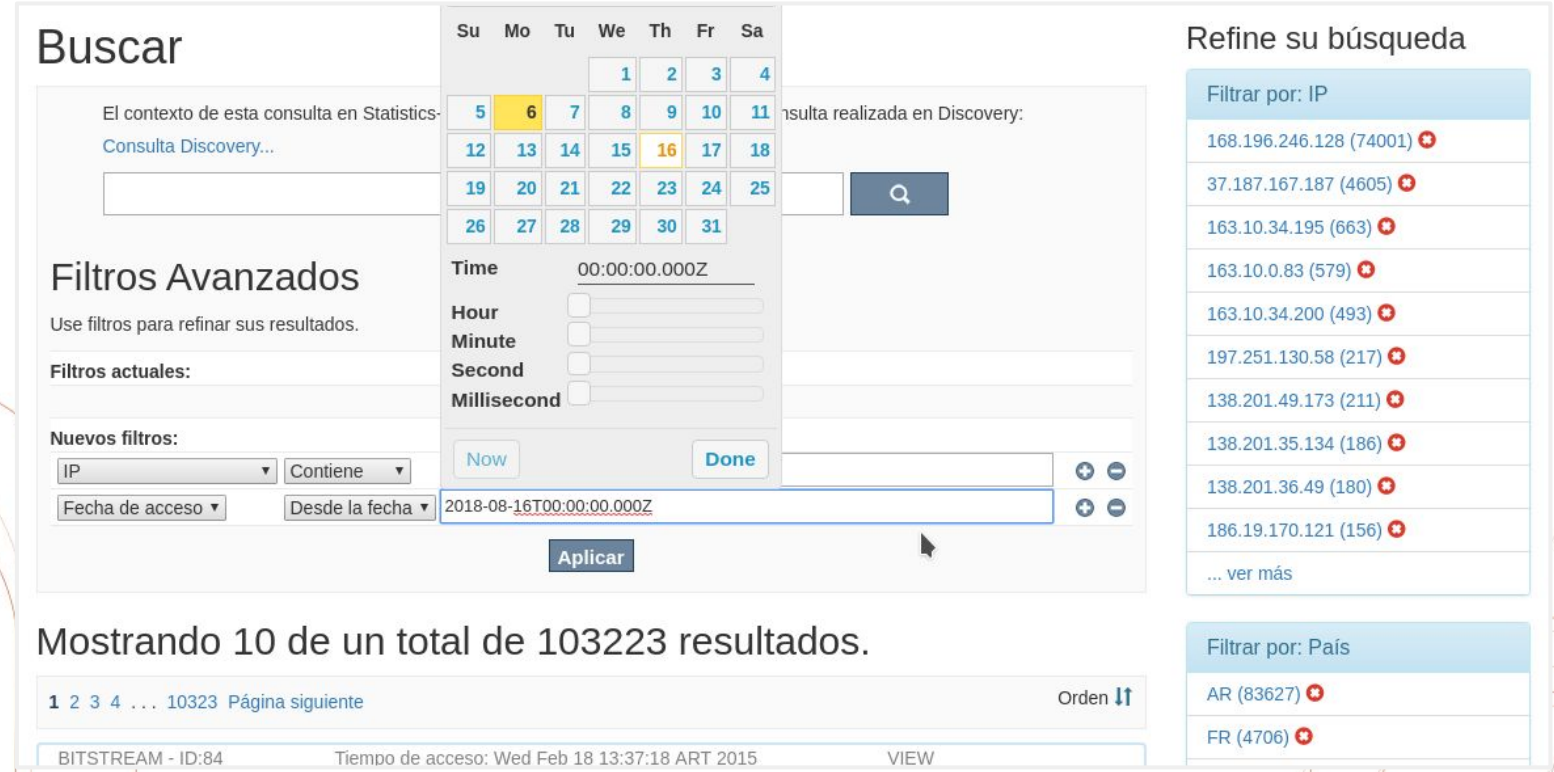
- una comunidad
- una colección
- un ítem
- un conjunto de objetos DSpace resultantes de una consulta Discovery.
    - *Por ejemplo: los ítems cuyo autor sea "Juan Perez"*

Para las búsquedas se pueden definir distintos de **filtros y facets**, entre ellos:

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

Para la **búsquedas por campos de fechas** se agregaron nuevos operadores para la definición de rangos de fecha.

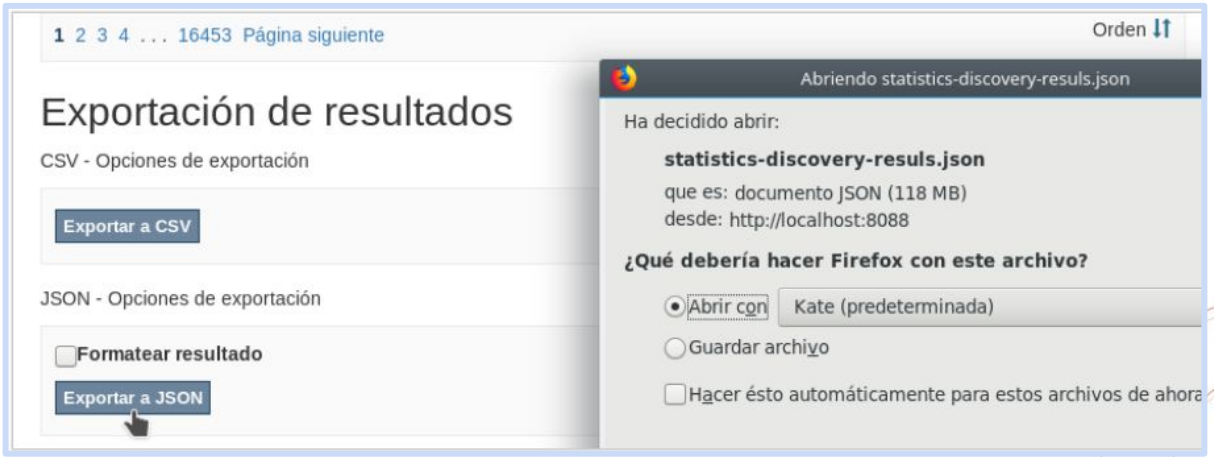
Búsqueda de registros - Contextos y Filtros



Exportación de resultados

Se implementó un modelo extensible para la exportación de registros mediante distintas estrategias y transformación de resultados.

Cada **estrategia** debe implementar el método *export().* Por defecto, se implementaron 2 estrategias de exportación: **CSV** y **JSON.**



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)

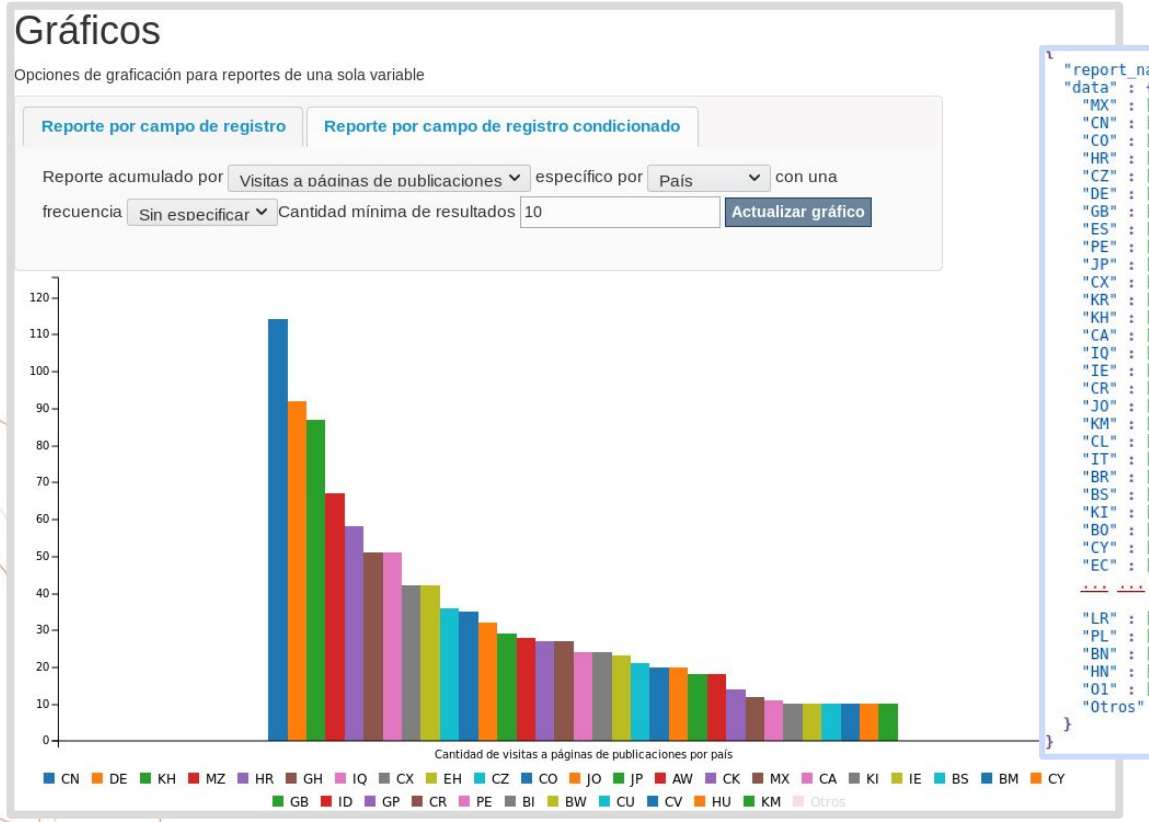
Se agregó capacidad para determinar un **lapso de tiempo** por reporte: *mensual* o *anual*.

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

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" ],
    "OI" : [ "10" ],
    "Otros" : [ "1244" ]
  }
}
```

Generación de reportes

Gráficos

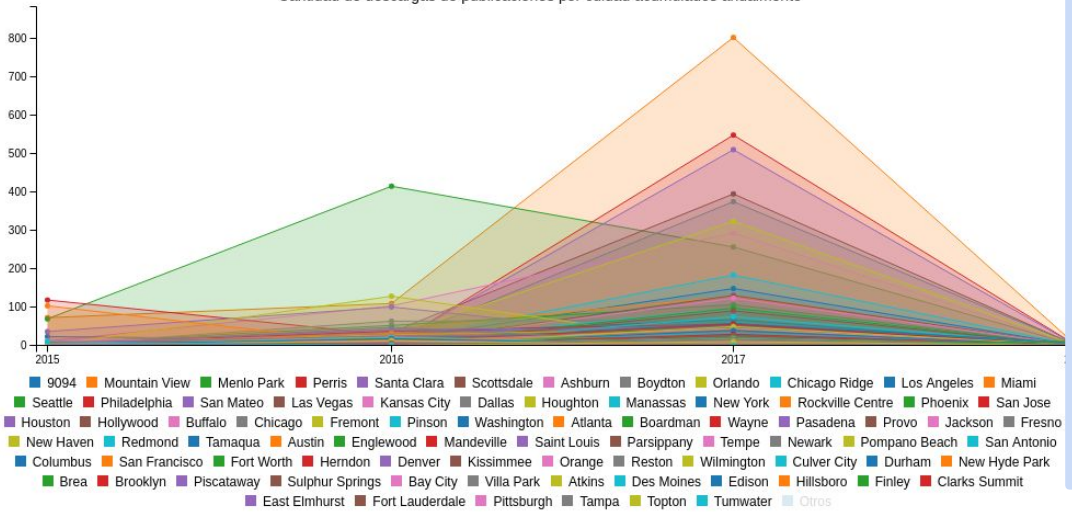
Opciones de graficación para reportes de una sola variable

Reporte por campo de registro | Reporte por campo de registro condicionado

Reporte acumulado por Descargas de publicaciones ▾ específico por Ciudad ▾ con una frecuencia

Anual ▾ Cantidad mínima de resultados 10 Actualizar gráfico

Cantidad de descargas de publicaciones por cuidad acumulados anualmente




```
{
  "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" ]
  }
}
```

Código Fuente del prototipo

El código fuente del prototipo se encuentra libre en Github para su descarga, inspección y contribución:

https://github.com/FacundoAdorno/DSpace

¡Muchas gracias!

Por consultas:

facundo@sedici.unlp.edu.ar

Nuestros sitios

http://sedici.unlp.edu.ar

http://digital.cic.gba.gob.ar/

http://cesgi.cic.gba.gob.ar/

http://prebi.unlp.edu.ar

http://www.istec.org/liblink/

http://revistas.unlp.edu.ar/cientificas/

http://revistas.unlp.edu.ar

http://congresos.unlp.edu.ar

http://ibros.unlp.edu.ar