UNIVERSIDAD NACIONAL DE LA PLATA

SEDICI
REPOSITORIO INSTITUCIONAL DE LA UNLP

PREBI
prebi.unlp.edu.ar

Capacitación IUPA

Clase 4

Pablo de Albuquerque, Santiago Tettamanti, Ariel Lira
{pablo, santit, alira}@sedici.unlp.edu.ar

11 de Mayo de 2022

Índice

- Repaso clase anterior
- Interoperabilidad
- Estadísticas en DSpace
- Administración avanzada
- Instalación
- Actualización
- Requerimientos y recomendaciones para el entorno de producción

Interoperabilidad

Servicios, protocolos y directrices

Servicios de Interoperabilidad

Los repositorios digitales deben pensarse como sistemas interoperables desde el principio.

Interoperabilidad **desde** el repositorio para

- integrarse con otros sistemas de la institución
- ampliar el alcance y difusión de los contenidos
- incorporarse a sistemas o redes regionales e internacionales

Interoperabilidad **hacia** el repositorio para

- facilitar y/o agilizar la ingesta de contenidos
- modificar remotamente los contenidos

Servicios de Interoperabilidad

Tener en cuenta…

- protocolos de comunicación y transferencia
  - OAI (entrada y salida)
  - SWORD (entrada y salida)
  - RSS (salida)
  - OpenSearch (salida)
- codificación de caracteres
- formatos de datos

Que es OAI?

Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)

- es un mecanismo de interoperabilidad entre repositorios.
- Define dos roles
  - Data Providers → repositorios que exponen metadatos estructurados vía OAI-PMH
  - Service Providers → Sistemas que van a cosechar lo expuesto por un data provider.

Cómo funciona?

conceptos de **Filter**, **Transformer**, **Format** y **Context**.

- Context → Define **el conjunto de ítems a exponer**.
- Filter →Define qué condiciones debe cumplir un ítem para ser expuesto en un contexto
- Transformer → Realiza cambios en los metadatos antes de exponerlos en OAI
- Format → mapea los metadatos al perfil expuesto
    - De **dcterms.creator.author** a **dc.creator**

Estadísticas en DSpace

Estadísticas en Dspace

DSpace utiliza la aplicación SOLR subyacente para el registro de las estadísticas.

- SOLR permite la búsqueda y la adición de grandes cantidades de datos (de uso).
- Dedica un core exclusiva para registrar estos eventos, llamado statistics

La activación de las estadísticas no requiere instalación o personalización adicional.

Estadísticas en Dspace - Que se registra?

- Las visitas a las páginas
- Las descargas de los archivos
- Los campos que se registran ante cada evento se configuran en el archivo **dspace/solr/statistics/conf/schema.xml**
    - Se pueden configurar diferentes campos si se trata de una visita a una página, de una descarga de un archivo o del módulo de búsqueda de SOLR

```
<field name="type" type="integer" indexed="true" stored="true" required="true" />
<field name="id" type="integer" indexed="true" stored="true" required="true" />
<field name="ip" type="string" indexed="true" stored="true" required="false" />
<field name="time" type="date" indexed="true" stored="true" required="true" />
<field name="epersonid" type="integer" indexed="true" stored="true" required="false" />
<field name="continent" type="string" indexed="true" stored="true" required="false"/>
```

Estadísticas en Dspace - En donde se visualizan?

- En el home **/statistics-home**

Estadísticas

Número total de visitas

	Accesos
Biophysical effect of climate change on summer crops	13360
Síndrome Ascítico edematoso	8503
Rousseau y el liberalismo	4605
Un análisis experimental de tipo de aplicaciones para dispositivos móviles	3915
Abastecimiento de agua a Bahía Blanca. Ubicación y diseño de la batería de pozos de explotación de agua subterránea	2879

Estadísticas en Dspace - En donde se visualizan?

- En las páginas de cada comunidad/colección

  /[handle-coleccion o comunidad]/statistics


CICBA

Estadísticas

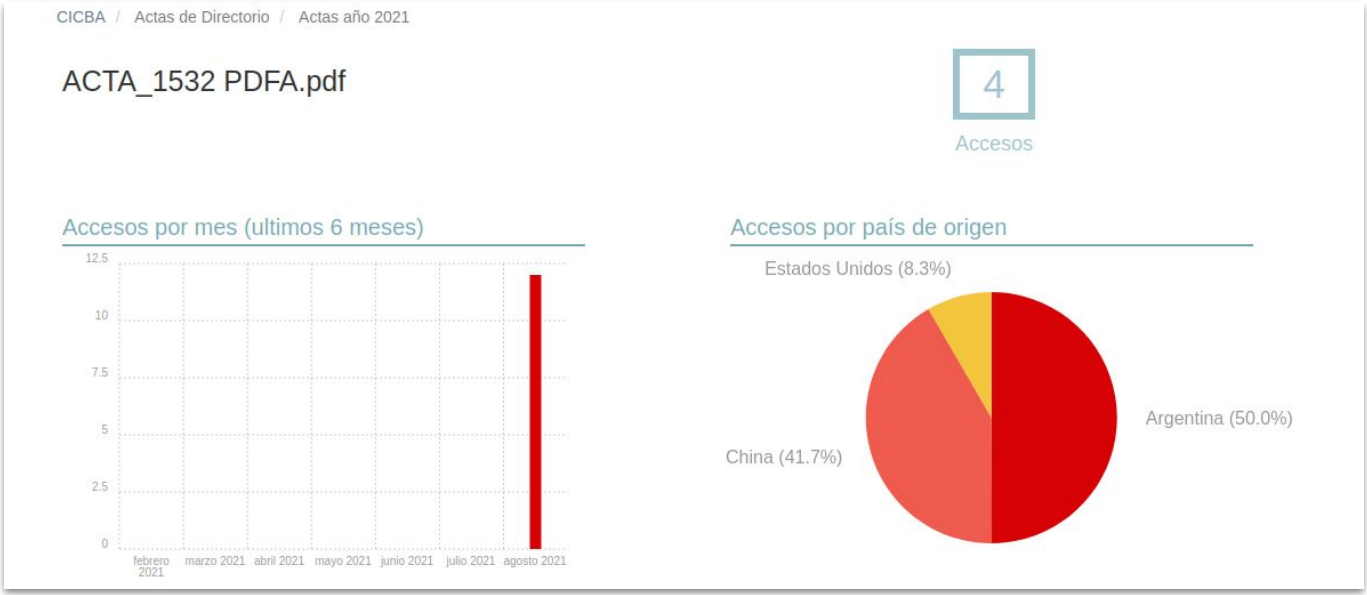
Número total de visitas

	Accesos
CICBA	92637

	Accesos
Argentina	35258
Francia	4919

Estadísticas en Dspace - En donde se visualizan?

- En las páginas de cada ítem **/[handle-item]/statistics**



Estadísticas en DSpace - Configuración

La configuración del core de statistics a interfaz que muestra las estadísticas mencionadas anteriormente se encuentra en usage-statistics.cfg

Ejemplo:

```
# Controla si las páginas de estadísticas deben mostrarse sólo a los usuarios autorizados.

Si se activa, sólo los administradores del DSpace Object podrán ver las estadísticas.
Si se desactiva, cualquiera con permisos de LECTURA del DSpace Object podrá ver las estadísticas.

usage-statistics.authorization.admin.usage=false # Usuarios con LECTURA puede ver estadísticas de uso
usage-statistics.authorization.admin.search=true # Solo admin puede ver estadísticas de búsqueda
usage-statistics.authorization.admin.workflow=true #Solo admin puede ver estadísticas del workflow
```

Estadísticas en DSpace - Google Analytics

Es posible configurar DSpace para que Google Analytics registre las estadísticas de uso. Se deben seguir los siguientes pasos:

- Registrar una cuenta cuenta de Google analytics desde https://analytics.google.com/analytics
- Una vez completados los datos obtenemos el ID de seguimiento (UA-XXXXXXXX-X)
- Habilitar el uso de Google analytics desde la propiedad **xmlui.google.analytics.key**, el valor de la propiedad es el ID de seguimiento obtenido anteriormente
- Ingresar en https://console.developers.google.com/project utilizando la misma cuenta utilizada en Google analytics y configurar allí un nuevo proyecto.

Más info en Wiki sedici/DSpace -> Estadísticas#google-analytics

Estadísticas en DSpace - Google Analytics

Lo registrado en analytics es más confiable que lo obtenido desde solr

- Filtrado de bots y estadísticas en tiempo real

Administración avanzada del repositorio

Administración avanzada - DSpace CLI

- El comando ***dspace*** (`DSPACE_INSTALL/bin/dspace`) permite:
  - administrar índice solr de statistics:
    - **`stats-util`** → actualiza el índice con los últimos cambios de metadatos
    - **`stats-util --mark-spiders`** → marca los accesos de IPs considerados como spiders con el flag *isBot*
    - **`stats-util --delete-spiders-by-flag`** → elimina los accesos marcados con el flag *isBot*
  - administrar índice solr de oai:
    - **`oai import`** → actualiza el índice con los últimos cambios de metadatos
    - **`oai import -c`** → reindexa todos los ítems del repositorio
    - **`oai clean-cache`** → limpia la caché del OAI
  - administrar índice solr de search:
    - **`index-discovery`** → actualiza el índice con los últimos cambios de metadatos
    - **`index-discovery -b`** → regenera el índice

Administración avanzada - DSpace CLI

- crear usuarios administradores:
  - `create-administrator`
- crear usuarios
  - `user --add -l es -g Nombre -s Apellido -m email@dominio.com -p <password>`
- Ver el valor actual de una propiedad de configuración
  - `dsprop -p db.schema`
- Actualizar metadatos, agregar ítems y/o mover ítems de colección →Ver wiki para más info
  - `metadata-import -f <archivo.csv> [-w]` →-w agregar los ítems al workflow
- Exportar metadatos
  - `metadata-export -f <archivo_destino> [-i (id o handle a exportar)]`
- y muchos más: aquí

IMPORTANTE: **SIEMPRE** ejecutar el comando dspace bajo el usuario designado para DSpace y tomcat.

Administración avanzada - Importación/Exportación

Exportar o importar comunidades, colecciones o ítems en DSpace.

- Solo metadatos → Archivo csv
- Metadatos + objeto digital → formato SAF
- Desde UI → como administrador del DSpace Object
- Desde la consola → bin/dspace

Administración avanzada - Importación/Exportación - Solo Metadatos

- Metadata-import
  - `[dspace]/bin/dspace metadata-import -f name_of_file.csv`
  - `-f` →nombre del archivo
  - `-w` → Para enviar los ítems a workflow
  - `-e` → email del usuario que realiza la importación

Más info en la wiki de DSpace

- Metadata-export
  - `[dspace]/bin/dspace metadata-export -f name_of_file.csv -i 1023/24`
  - `-i` → handle o id del DSpace Object, si no se especifica se exporta **todo el repo**
- La primera fila del CSV debe definir los valores de los metadatos
  - La primera columna debe "id" → ID interno de la base de datos del ítem
- Es posible realizar ediciones en masa fácilmente:
  1. Exportar en csv la colección/items deseados
  2. Corregir o agregar el metadato deseado
  3. Importar el mismo csv

Administración avanzada - Importación/Exportación - SAF

- Formato SAF
  - Una carpeta por ítem
  - Uno o más archivos xml con metadatos
  - Archivos que describen el contenido
  - Los Objetos digitales del ítem

```
archive_directory/
    item_000/
        dublin_core.xml
        metadata_[prefix].xml
        contents
        collections

        file_1.doc
        file_2.pdf
```

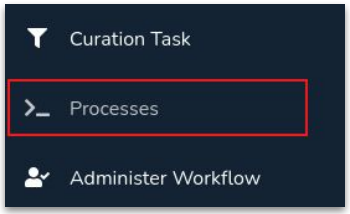
- Import
  - `[dspace]/bin/dspace import --add --eperson=joe@user.com --collection=CollectionID --source=items_dir --mapfile=mapfile`
- Export
  - `[dspace]/bin/dspace export --type=COLLECTION --id=collectionID_or_handle --dest=/path/to/destination --number=seq_num`

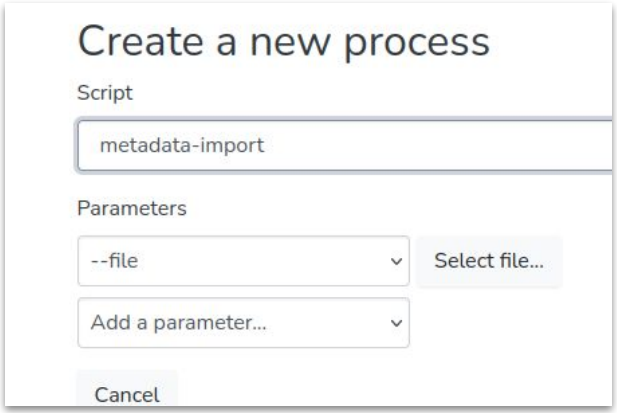
Más info en la wiki de DSpace

Administración avanzada - Importación/Exportación - UI

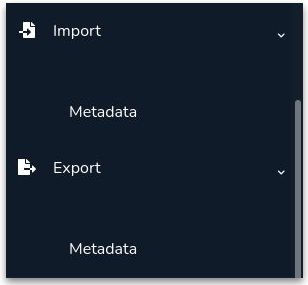
Por ahora solo metatados en CSV → más adelante SAF (7.2 o 7.3)

- Desde el menú de procesos





- Desde la opción import/export



Instalación en servidor

Servidor - Instalación de dspace

Dependencias (apt-get install --no-install-recommends)

DSpace: git maven ant ant-contrib openjdk postgresql-client ghostscript imagemagick

DB: postgresql postgresql-contrib

Descarga de Fuentes

```
git clone https://github.com/DSpace/DSpace/ [dspace-source]
```

Más info en la wiki

Creación de Base de datos

```
createdb -U dspace_user -O dspace_user -E UNICODE dspace_repo
psql dspace_repo -c "CREATE EXTENSION pgcrypto;"
```

Compilacion e Instalacion DSpace

```
mvn clean && mvn package
cd [dspace-source]/dspace/target/dspace-installer && ant fresh_install
```

Servidor - Web container

- Web Container:
  - se puede usar Jetty o Tomcat.
  - se recomienda tomcat 9.0.x o superior
- Usar mismo usuario para tomcat y DSpace
- JAVA_OPTS para todos
  - -Xmx2048m → heap máxima
  - -Xms1024m → heap mínima
  - -XX:MaxPermSize=1024m → classloading

Más info en la wiki

Servidor - Web container - Configuración

Configuración:

- **/etc/tomcat9/server.xml** (configuraciones del servicio)
  - Escuchar solicitudes HTTP en el puerto 8080 (default)

```
<Connector port="8080" protocol="HTTP/1.1" connectionTimeout="20000"
    URIEncoding="UTF-8" redirectPort="8443"/>
```

  - Recibir solicitudes AJP al puerto 8009

```
<Connector port="8009" protocol="AJP/1.3" URIEncoding="UTF-8"
redirectPort="8443"/>
```

- **/etc/default/tomcat9**
  - TOMCAT_USER, TOMCAT_GROUP
  - JAVA_OPTS

Servidor - Proxy reverso

- Proxy reverso
  - **MUY** recomendado
  - se puede usar nginx, varnish o apache2+mod_proxy
  - Si se elige apache se recomienda usar AJP (mod_proxy_ajp)

Más info en Wiki

Servidor - Base de datos

- Base de datos
  - Se puede usar cualquier DB compatible con Hibernate
  - Se recomienda PostgreSQL 9.4+
  - Para PG es necesaria la extensión pgcrypto
- Ubicación
  - misma máquina que DSpace o en otra con baja latencia
  - se recomienda usar una vm separada para simplificar actualizaciones y escalabilidad

Más info en Wiki

Servidor - Solr

- **Usos**
  - Búsqueda (core *search*)
  - Data Provider OAI-PMH (core oai)
  - Estadísticas (core *statistics*)

- **Ubicación:**
  - vm separada: permite regular de manera independiente la cantidad de Memoria asignada y designar hardware según conveniencia
  - en misma vm que DSpace: mismo tomcat, actualización más simple

Servidor - Handle

1. Conseguir un prefijo handle
2. Actualizar configuración en local.cfg con el prefijo asignado
3. Generar un archivo sitebndl.zip en el servidor de producción:
   a. [dspace]/bin/dspace make-handle-config [dspace]/handle-server
4. Subir sitebndl.zip a http://handle.net/prefix_request.html y completar el registro
5. Iniciar el servidor handle y verificar su funcionamiento
   a. ejecutar [dspace]/bin/start-handle-server
   b. revisar [dspace]/log/handle-server.log
   c. probar resolución de identificadores http://dominio-dspace:8000
6. Actualizar identificadores de registros preexistentes
   a. [dspace]/bin/dspace update-handle-prefix
7. Configurar inicio automático

Más info en la wiki de Dspace

Administración avanzada - Cronjobs

Un Cron Job ejecuta tareas, Jobs, a intervalos regulares.

Se depositan en **/etc/cron.d y en crontab**

Se recomienda usar 2 archivos de cronjob:

- backups:
  - tareas de generación de copias de respaldo de todos los componentes de dspace (db, índices, datos, configuraciones, etc)
  - tareas de duplicación de backups offsite
- dspace
  - mantenimiento de índices solr
  - media-filter
  - etc

Administración avanzada - Cronjobs - Ejemplos

Alguna de las tareas ejecutadas a través de cronjobs son las siguientes:

- Generar AIPs una vez por semana
- Actualizar el árbol de páginas o sitemap para crawlers
- Actualizar y optimizar el índice de búsqueda ('search').
- Envío automático de emails a los usuarios suscriptos a una colección o comunidad.
- Ejecución de media-filters sobre el assetstore
- Actualiza el índice oai

Administración avanzada - Cronjobs - Ejemplos

Los dos primeros números indican la hora, los siguientes los días.

```
# Regenerate DSpace Sitemaps every 8 hours (12AM, 8AM, 4PM).
# SiteMaps ensure that your content is more findable in Google, Google Scholar, and other major search engines.
0 0,8,16 * * *  dspace  $DSPACE/bin/dspace generate-sitemaps > /dev/null

#----------------
# DAILY TASKS
#----------------

# Clean and Update the Discovery indexes at midnight every day
# (This ensures that any deleted documents are cleaned from the Discovery search/browse index)
0 0 * * *       dspace  $DSPACE/bin/dspace index-discovery > /dev/null

# Re-Optimize the Discovery indexes at 12:30 every day
# (This ensures that the Discovery Solr Index is re-optimized for better performance)
30 0 * * *      dspace  $DSPACE/bin/dspace index-discovery -o > /dev/null
```

https://crontab.guru/ para jugar un poco con el formato de las cron

Administración avanzada - Backups - Riesgos

¿Por qué hacer backups?

Riesgos!

Es importante contar con un **plan de riesgos** con pasos a seguir ante cada situación

- Error humano → delete o update from sin el where!!!
- Ataques → Malware, ransomware, captura del servidor
- Entorno → Inundación, incendio, robo
- Necesidad de acceso a estados previos de la base de datos → Base de datos corrupta

Administración avanzada - Backups

- Dump de postgres
  - `pg_dump -U $DBUSER -h $DBHOST -f /tmp/$FILENAME.tar -F t $DBNAME`
  - `pg_restore -Ft -O -f /tmp/$FILENAME.sql /tmp/$FILENAME.tar`
  - `tar --directory /tmp -cvzf $DESTFILENAME $FILENAME.sql`
- Cronjobs para backups diarios y rotación de backups
  - `0 21 * * * root sh $DSPACE_BACKUP_DIR/script_pgdump`
  - `0 5 * * 1 dspace find $DSPACE_BACKUP_DIR -maxdepth 1 -name "*-20??-??-??*.tgz" -mtime +12 -not -name "*-20??-??-01*.tgz" -exec rm {} \;`
- Medidas adicionales
  - Enviar los backups a otro server con **rsync** u otra herramienta similar
  - Copiar datos a otro edificio con menor o al menos diferente riesgo ante catástrofe.
  - Copiar backups en modo pull, para evitar encriptación de backups viejos por parte de ataques ransomware

Servidor - Logs Revisión de Errores

Logs principales

- DSpace log: **[dspace]/log/dspace.log**
- Solr log: **[dspace]/log/solr.log**
- Tomcat log directory: **/var/log/tomcat8/***

IMPORTANTE:

- **activar y configurar logrotate para estos servicios.**
- **NO descartar dspace logs y access.logs**

Actualización

Actualización

Dspace se puede actualizar de dos maneras.

1. Manualmente
2. A través de un script
3. Servicios de automatización (ej Jenkins)

Más info en la wiki

Etapas en la Actualización de DSpace

1. Preparar el entorno
2. Recuperar los cambios del repositorio de código fuente
3. Compilar el proyecto
4. Instalar el código

Actualización de DSpace - Preparar el entorno

1. Cambiar a un usuario no root (Ejemplo: DSpace)
2. **Backups** BBDD y archivos de configuración importantes. (local.cfg, etc).
3. Buscar archivos que no pertenezcan al usuario o grupo dspace
   a. `find -not -user dspace -or -not -group dspace`

Para no buscar en el assetstore: `find -type d \( ! -name assetstore \) -not -user dspace -or -not -group dspace`

Si hay archivos no pertenecientes a **dspace:dspace** → `sudo chown dspace:dspace <file>`

Actualización de DSpace - Traer los cambios del repositorio

4. **git status** - chequear si existen cambios locales
    a. **commit** o **stash**
5. **git pull -** traernos nuevos cambios
6. **Si se hizo git stash en el punto 3:**
    a. **git stash pop -** recuperar los cambios locales

Actualización de DSpace - Compilar el proyecto

7. **mvn clean package -** en el directorio root de DSpace (*/[DSPACE-DIR]/)*
    a. El compilado se genera en **[DSPACE-DIR]/dspace/target/dspace-installer**
8. **service tomcat stop -** parar la instancia de tomcat.

Actualización de DSpace - Instalar el código

8. Posicionarse en el directorio donde se encuentra el compilado
    a. **cd [DSPACE-DIR]/dspace/target/dspace-installe**r
9. **ant clean_backups**
    a. Para limpiar los backups realizados por ant en instalaciones previas
10. **ant update** (**ant install** para la primera vez)
    a. Para instalar o actualizar el código de DSpace
    b. El código queda en el directorio definido por la variable de config **dspace.dir**
11. **service tomcat start** - inicializa nuevamente el servidor

Requerimientos y recomendaciones para el entorno de producción

Servidor - Requerimientos de hardware

- Desarrollo - Solo backend
  - hardware: 4Gb RAM, 5Gb Almacenamiento
  - configuración: 1Gb es para tomcat, postgres default
- Mínimo producción
  - hardware: 4Gb RAM, 20Gb Almacenamiento
  - configuración: tomcat con 2Gb de -Xmx, postgres default
- Recomendado producción
  - máquina 1:
    - hardware: 3Gb RAM, 20GB almacenamiento (o superior)
    - configuración: postgres configurado para usar todos los recursos
  - máquina 2:
    - hardware: 6GB RAM, 100Gb Almacenamiento (o superior)
    - configuración: tomcat con 4Gb de -Xmx

Ver Recomendaciones de DSpace

Servidor - Recomendaciones

- En entornos de producción se recomienda:
  - raid con replicación: 1,10,5, 6, etc.
  - discos ssd o mecánico, en lo posible SAS hot plug.
  - RAM ECC
  - virtualización
- Sistema operativo
  - recomendado: Linux*
  - super testeado en debian, ubuntu y red hat
- Otras
  - en caso de muchos accesos y/o hardware lento, separar el servidor solr en una máquina independiente.

Servidor - Seguridad

Mínimamente se recomienda:

- Firewall
  - Web container accesible sólo desde DMZ
  - Apache2 accesible desde internet
  - Handle.net server accesible desde internet
- sshd:
  - PermitRootLogin no
  - accesible sólo desde DMZ o en su defecto aplicar control de acceso con fail2ban o similar
- Anti-DOS
  - detección y bloqueo de bots con comportamiento agresivo. Ej: accesos concurrentes.

Servidor - Monitoreo

- Watchdogs
  - control de disponibilidad de recursos
  - ej: almacenamiento libre, RAM libre, % CPU, etc
- Monitores externos
  - control de uptime y de tiempo de respuesta
  - ej nagios, pingdom, etc
- Implementar mecanismo de detección de bots abusivos
  - bots que no respetan robots.txt
    i. Disallow en robots.txt
    ii. connection drop
  - bloqueo automático de ips con demasiados accesos por minuto

Servidor - Errores frecuentes

- Ataques (DOS, bots, etc)
- Cantidad de conexiones a BD
- Espacio en disco por bitstreams y logs
- Superposición de cronjobs en una misma ventana de tiempo
- OutOfMemory en tomcat o cli:
    - aparece en catalina.out → requiere reiniciar tomcat

Actividades

Actividades

- (opcional) Realizar import, export, reindex de discovery y de oai
- Diseñar entorno de producción:
  - Hardware usado: físico o virtual, bare metal o cloud.
  - cantidad y características de cada vm/ct
  - Reglas de firewall (puertos, protocolo, ip)
  - backups: esquema de rotación y niveles de almacenamiento
  - Herramientas de monitoreo