Protección contra escritura de archivos en formato PDF dentro del repositorio "DSpace" en Linux

Washington A. Velásquez Vargas

Gerencia de Tecnologías y Sistemas de Información, Escuela Superior Politécnica del Litoral, Km. 30.5 Vía Perimetral, Guayaqui, Ecuador
wavelasq@espol.edu.ec

Resumen. En este trabajo se presenta una breve solución sobre la protección de documentos con formato PDF en el repositorio DSpace actual de la ESPOL, permitiendo de esta manera parar con la copia ilegal parcial o total de sus contenidos y dando consigo una correcta distribución de los documentos científicos del repositorio. Debido a que el servidor se encuentra bajo sistema operativo Linux se muestra una descripción de las partes más fundamentales del código utilizado en java.

Palabras Clave: Dspace, Repositorios, PDF, Protección de archivos PDF, Encriptación.

1 Introducción

La necesidad de contar con fuentes de información confiables que ofrezcan documentos verídicos a la comunidad científica, permite que las instituciones educativas de renombre a nivel mundial cuenten con un repositorio de sus documentos, para posteriores consultas por parte de investigadores. Pero lastimosamente, algunos hacen mal uso de estas herramientas realizando copias parciales o totales de los documentos sin citar el archivo original.

La principal herramienta usada para la implementación de los repositorios es "Dspace" debido a las grandes ventajas que ofrece, pero siendo una herramienta de código libre no hay preocupación de que archivos se suban al sistema ni quien los puede ver o descargar teniendo el riesgo de una reproducción ilícita. Con todo esto en mente se presenta la oportunidad de proteger los archivos en formato PDF contra escritura logrando tener algo de seguridad al momento de la reproducción de los documentos.

2 DSpace

DSpace es el software de elección por parte de instituciones académicas, sin fines de lucro y organizaciones comerciales que construyen repositorios digitales abiertos. Es gratuito, fácil de instalar y completamente personalizable para adaptarse a las necesidades de cualquier organización. [1]

DSpace permite el acceso fácil y abierto a todo tipo de contenido digital, incluyendo texto, imágenes, imágenes en movimiento, MPEG y conjuntos de datos. Y con una

comunidad cada vez mayor de desarrolladores comprometida con la continua expansión y mejora del software.

A continuación en la (Fig. 1) se describe el funcionamiento del Dspace:

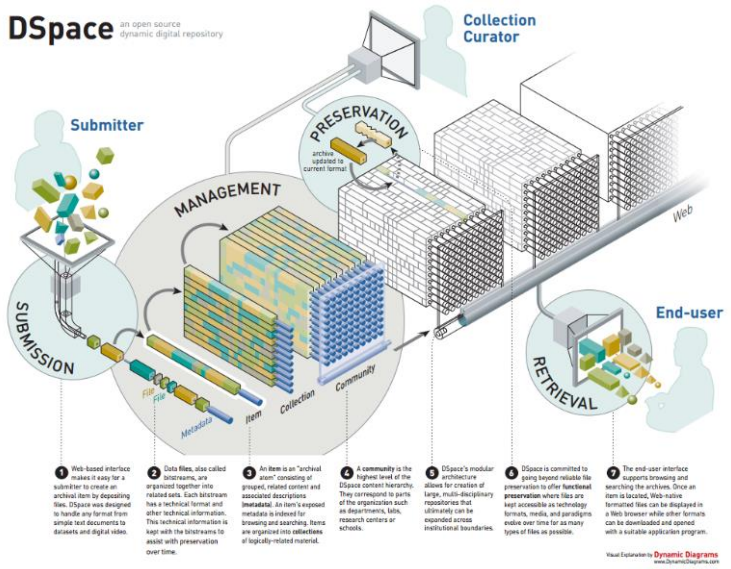

Fig. 1. Funcionamiento del repositorio Dspace [1]

2.1 Estructura del assetstore

Mientras que el modelo de datos de DSpace (metadatos, workflows, estructura del repositorio, usuarios) está soportado por la base de datos Oracle o Postgresql, los contenidos de los ítems se almacenan en el sistema de ficheros denominado assetstore.

La configuración tradicional del assetstore se realiza en el fichero dspace.cfg, mediante el parámetro [2]:

- Para un solo sistema
  ```
  assetstore.dir = [dspace]/assetstore
  ```
- Para más de un sistemas
  ```
  assetstore.dir = [dspace]/assetstore_0
  assetstore.dir.1 = /mnt/other_filesystem/assetstore_1
  ```

La localización física de un objeto se guarda en la base de datos por lo que es de especial importancia NO mover los bitstreams entre assetstores (además, el backup del assetstore tiene que formar parte de cualquier estrategia de backup).

Por defecto, los bitstreams nuevos se guardan en el assetstore 0 (es decir el especificado por la propiedad assetstore.dir). Para usar nuevos assetstores hay que añadir una línea al dspace.cfg que referencie dónde deben ir los nuevos bitstreams:

```
assetstore.incoming = 1
```

Todos estos ficheros se almacenan en el assetstore con una estructura tal como se muestra en la (Fig. 2):

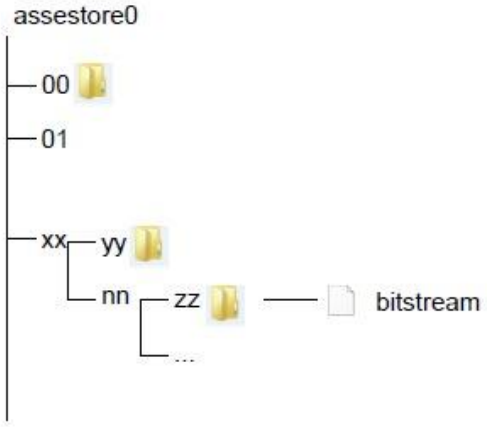

Fig. 2. Estructura del Assetstore [2]

Siguiendo con este mismo ejemplo, la referencia de un ítem a sus ficheros se encuentra en la tabla bitstream, en el campo "internal_id". Tal como se muestra en la (Fig. 3). Ahora, si me encuentro este identificador, 110832826281924074367996140570931140204, este fichero en nomenclatura DSpace, se encuentra buscando los seis primeros dígitos del identificador, que indican en que subdirectorio de tercer nivel está el item (11 >> 08 >> 32) y el nombre real del fichero será 826281924074367996140570931140204, como se puede observar ha desaparecido toda referencia al nombre original del archivo.

Fig. 3. Ejemplo de la Estructura del Assetstore [2]

Conociendo la forma en que guarda los archivos el DSpace, podemos pasar a explicar la solución de sobrescribir los archivos .PDF contra escritura en la carpeta assetstore.

3 Encriptación de archivos PDF

Para realizar la encriptación de los archivos se debe tener acceso al servidor, en nuestro caso es un servidor en Linux. Por lo que se implementará una aplicación en java que realice la encriptación de los documentos.

Para la implementación se tuvo que utilizar las siguientes librerías:

- *Dspace-api-1.8.2.jar*.- API del DSpace, que utilizamos como guía para obtener el directorio donde se almacenan los archivos. [4]
- *Jargón-2.2.1-jar y Log4j-1.2.17.jar*.- Librerías que utiliza el API del DSpace para su funcionamiento. [5]
- *Postgresql-9.3.1102.jdbc41.jar*.- Librería para la conexión con la base de datos postgresql. [7]
- *Itext-1.4.6.jar*.- Librería que permite el manejo de archivos .PDF, con la cual realizamos la encriptación de los archivos. [6]

3.1 Búsquedas de Archivos

Para obtener archivos en formatos .PDF, realizamos una búsqueda en la tabla "bitstream" en el campo "name", donde el nombre del documento termine con extensión .PDF. Luego de esto debemos conocer la localización de donde se encuentran los documentos en el servidor. En este punto, es donde entra en función el api del DSpace que mediante el método "getFile" nos devuelve el archivo.

Hay que tener en cuenta, la búsqueda de este método, debido a que realiza búsquedas en el sistema local o en el Almacenamiento SRB. [3] Todo va a depender de la ruta que se tenga en la tabla bitstream.

Asumiendo que tenemos archivos en el directorio local, cuando se invoca el método "getFile" obtendremos objetos que tendrán como formato el siguiente:

```
/var/dspace/assetstore/52/75/20/527520035822039551128059493698010
32815
```

3.2 Encriptación

La encriptación de archivos se la lleva a cabo mediante la librería itext, en donde lo único que debemos especificar es las rutas del archivo original y el nuevo archivo que se generará encriptado, tal como se muestra en el siguiente código:

```
try {
  PdfReader reader = new PdfReader(Source);
  if(!reader.isEncrypted()){
     PdfEncryptor.encrypt(reader,  new  FileOutputStream(Target),
null, null, PdfWriter.AllowPrinting, false);
  }
} catch (Exception e) {
      System.out.println(e.getMessage());}
```

3.3 Cheksum MD5

Algo a tener muy en cuenta en esta solución es que al momento de realizar la encriptación estamos modificando el pdf original, por lo tanto; el hash se ha modificado. Para no tener ningún problema al momento de visualizar los archivos encriptados debemos obtener un nuevo "HASH" y actualizar de la tabla "bitstream" los siguientes campos:

- size_bytes: El nuevo tamaño del archivo encriptado.
- checksum: El nuevo checksum md5, para lograr esto utilizamos el siguiente código, especificando el path del archivo encriptado:

```
BufferedInputStream is = new BufferedInputStream(new
FileInputStream(path));
DigestInputStream dis = null;
dis = new DigestInputStream(is,
MessageDigest.getInstance("MD5"));
```

4 Conclusiones

Mediante la aplicación implementada en java y teniendo el acceso al servidor donde se encuentran los archivos del assetstore para ejecutar la aplicación con permisos de root, se logra tener archivos en formato .PDF con protección contra escritura. Logrando así tener los documentos cifrados, evitando el plagio parcial o completo de texto de los archivos sin citar al original.

Referencias

1. Dspace, «DSpace,» [En línea]. Available: http://dspace.org/introducing. [Último acceso: 15 11 2014].
2. DSpace, «Grepcode,» [En línea]. Available: http://grepcode.com/snapshot/repo1.maven.org/maven2/org.dspace/dspace-api/1.8.2. [Último acceso: 17 11 2014].
3. Dspace. [En línea]. Available: http://dspace.org/sites/dspace.org/files/archive/1_6_2Documentation/ch09.html. [Último acceso: 17 11 2014].
4. elorenzo, «Arvo - Hablando del DSpace,» [En línea]. Available: http://www.arvo.es/dspace/la-estructura-del-assetstore/. [Último acceso: 15 11 2014].
5. Itext, «Java2s,» [En línea]. Available: http://www.java2s.com/Code/Jar/i/Downloaditext146jar.htm. [Último acceso: 17 11 2014]
6. jargon, «Java2s,» [En línea]. Available: http://www.java2s.com/Code/Jar/j/Downloadjargon221jar.htm. [Último acceso: 17 11 2014].
7. Postgresql, «Postgresql,» [En línea]. Available: http://jdbc.postgresql.org/download.html. [Último acceso: 17 11 2014]