

Cómo crear paquetes de información para repositorios digitales

Soloaga Ignacio, Fernández Esteban, De Giusti Marisa

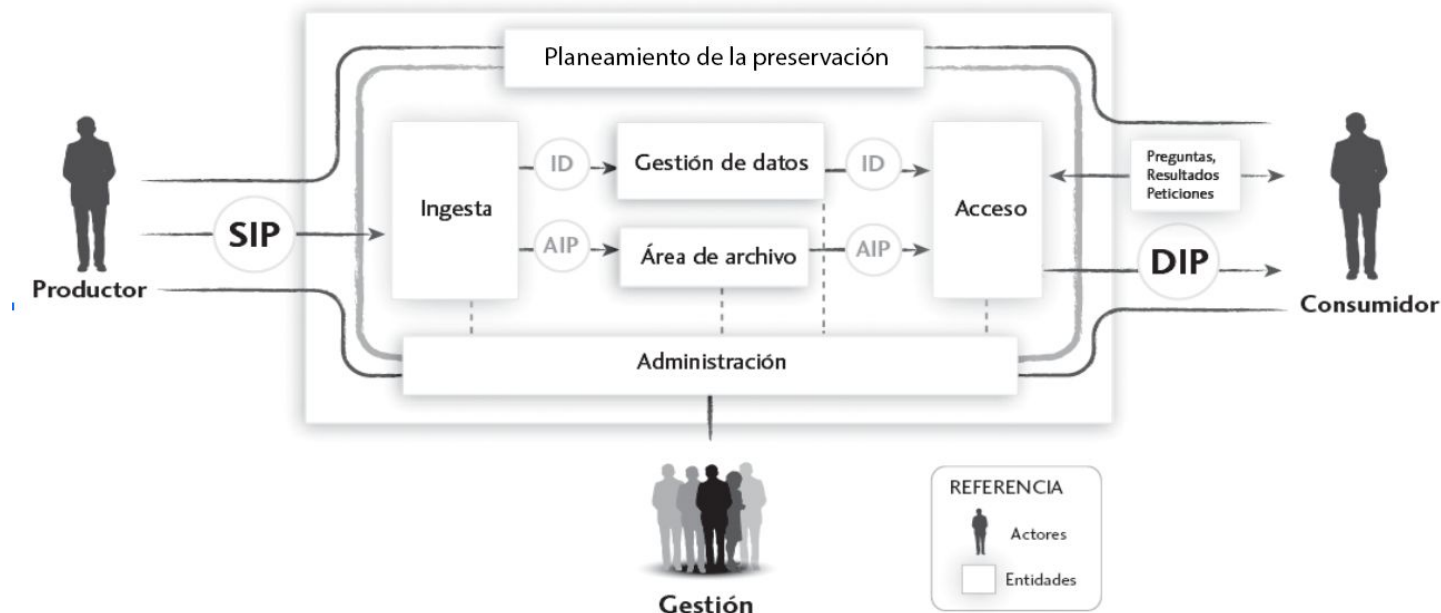
Preservación digital

La preservación digital es el conjunto de estrategias, procesos y técnicas que dan respuesta a los problemas que plantea la conservación de los materiales digitales y de los medios (hardware y software) que se emplean para su almacenamiento y consulta, y que están derivados fundamentalmente de la obsolescencia provocada por la rápida renovación tecnológica y por la inestabilidad de los soportes. Estas técnicas son muy variadas y responden a diferentes situaciones y líneas estratégicas (copias de seguridad, copia de datos en soportes durables, migración, replicación, emulación, etc.), aunque en general están destinadas a mantener los objetos digitales y sus características de acceso a largo plazo.

[Directrices UNESCO](#)

Estándares, Normas, Recomendaciones

El mundo de la PD está plagado de normativas, pero hay un gran modelo abstracto que debe conducir las acciones de un repositorio digital. Modelo Funcional.



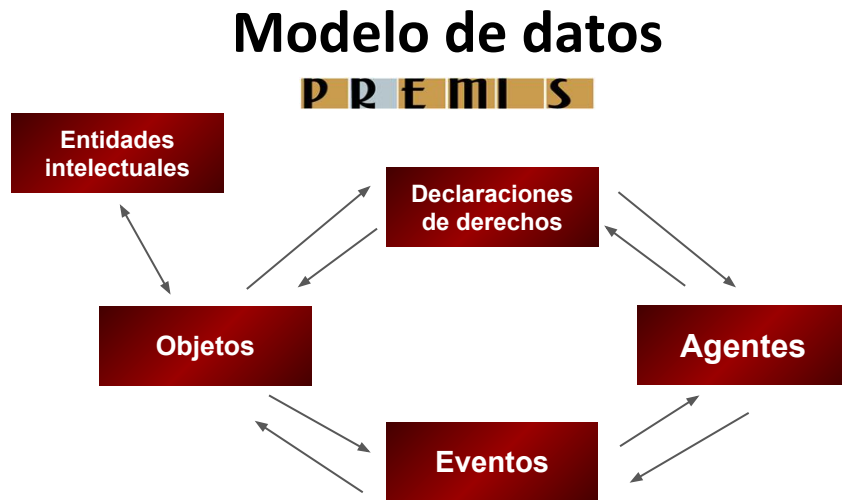
Paquete de información

La ISO 14721 habla de una unidad de información que es el paquete de información.



Diccionario de datos PREMIS

Un Repositorio Digital debemos tener en cuenta PREMIS. Este diccionario de datos es una traslación del modelo OAIS a unidades semánticas implementables, bajo la forma de un esquema de metadatos específicos para preservación.



Tipo de información a asociar a un OD para asegurar su preservación

Poblamiento de un repositorio

Autoarchivo: estos paquetes se componen de uno o varios archivos y metadatos que deben ser completados en la administración del RD.

Operaciones de la Administración

Ingesta masiva.



La correcta preparación y armado de un SIP es una tarea necesaria para lograr formatos de archivos y metadatos adecuados para la ingesta en un sistema de preservación digital y deberían adecuarse a constar al menos de lo que indica la norma y a tener una sintaxis determinada.

No pasa en la mayoría de los repositorios.

Si se da de baja un repositorio hay que transferir esos contenidos masivamente (AIP)..

Herramientas analizadas

Bagit. Library of Congress.

Pre-ingest Tool de OPF

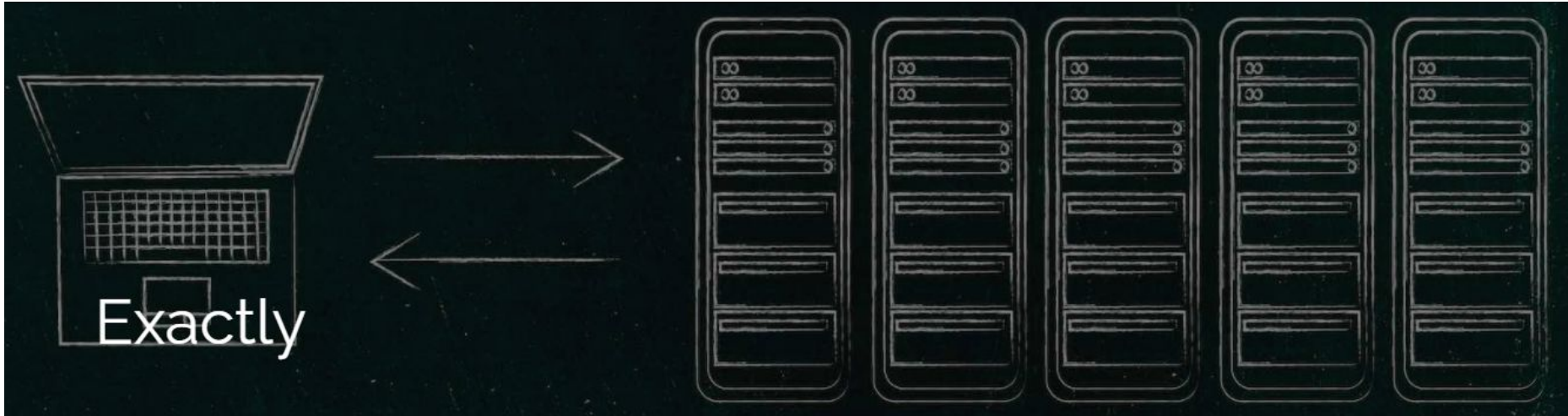
Movelt de Simon Fraser University

Exactly

Modalidad de ingesta para SEDICI a través de paquetes SAF



<https://www.weareavp.com/products/exactly/>



Exactly



Exactly utiliza el formato de empaquetado de archivos **BagIt**, un estándar del Internet Engineering Task-Force desarrollado por Library of Congress y la Biblioteca Digital de California con el apoyo de la Universidad George Washington y la Universidad de Maryland.

Exactly



Descarga:

<https://www.weareavp.com/es/productos/exactly/>

Guia de usuario

<https://www.weareavp.com/wp-content/uploads/2017/08/Quickstart-Guide-Delivery.pdf>

Para la instalación descomprimir el archivo y ejecutar exactly-1.0.exe que se encuentra en la carpeta dis/bundles.

Una vez instalado buscamos el programa en Inicio y lo ejecutamos.

Exactly cuenta con dos partes principales una para **enviar** y otra para **recibir**.

Enviar

Al presionar el botón Browse del campo **Source** podremos seleccionar la o las carpetas desde donde se importarán los archivos.

Archivo Zip o FTP

☐ Zip files? ☐ FTP delivery ☐ SFTP delivery

Si la entidad receptora ha especificado FTP como el modo preferido de entrega, y ha proporcionado la

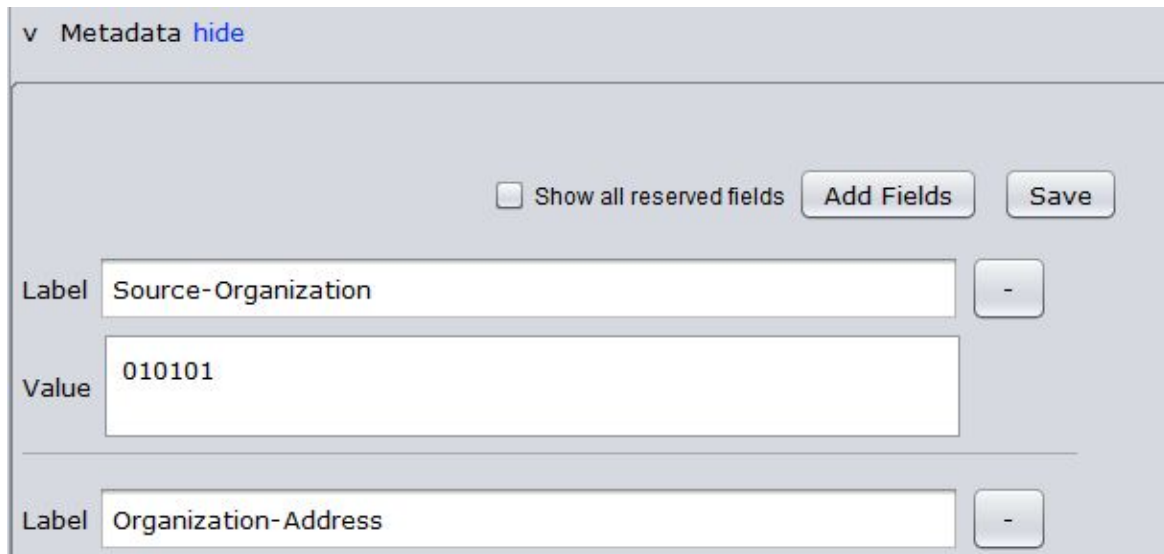
configuración adecuada en Admin, asegúrese de que la entrega FTP esté marcada.

The screenshot shows the 'Exactly 0.1.5' application window with the 'Transfer' tab selected. The interface includes a 'Title' field, a 'Source' field with a 'Browse' button and a '+' icon, and a 'Destination' field with a 'Browse' button. Below these fields are three checkboxes: 'Zip files?', 'FTP delivery', and 'SFTP delivery'. The 'Destination' field is populated with 'C:\Users\Contex\Desktop\SEDICI\pdf\prueb.'. On the right side, there is a large 'Transfer' button and a 'clear log' button. At the bottom, there is a 'Current Template: None' label. The 'Metadata' section is collapsed, showing a 'hide' link. The 'Show all reserved fields' checkbox is also present, along with 'Add Fields' and 'Save' buttons. The 'Label' field is set to 'Source-Organization' and the 'Value' field is set to '010101'. The 'Organization-Address' label is also visible.

La entidad receptora especificará si debe comprimir sus archivos o no.

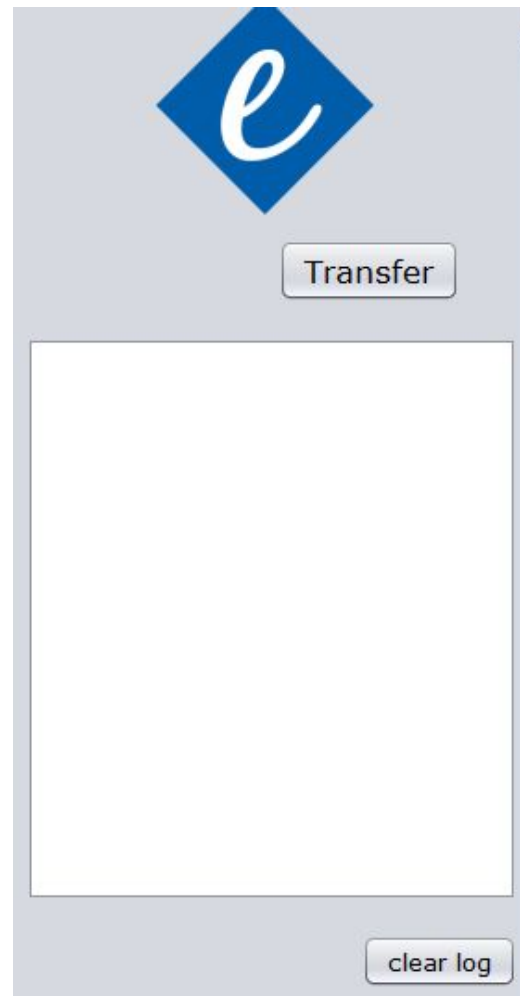
Destination nos permite seleccionar la carpeta de destino

Metadata aquí podemos crear los campos para los metadatos que van a describir el paquete, podemos modificar los que trae por defecto o podemos agregar más si los necesitamos.



The screenshot shows a web interface for configuring metadata. At the top, there is a tab labeled 'v Metadata' with a 'hide' link. Below the tab, there is a checkbox labeled 'Show all reserved fields' which is currently unchecked. To the right of the checkbox are two buttons: 'Add Fields' and 'Save'. The main area contains two rows of input fields. The first row has a 'Label' field with the text 'Source-Organization' and a small square button with a minus sign to its right. Below the label field is a 'Value' field containing the text '010101'. The second row has a 'Label' field with the text 'Organization-Address' and a similar minus sign button to its right.

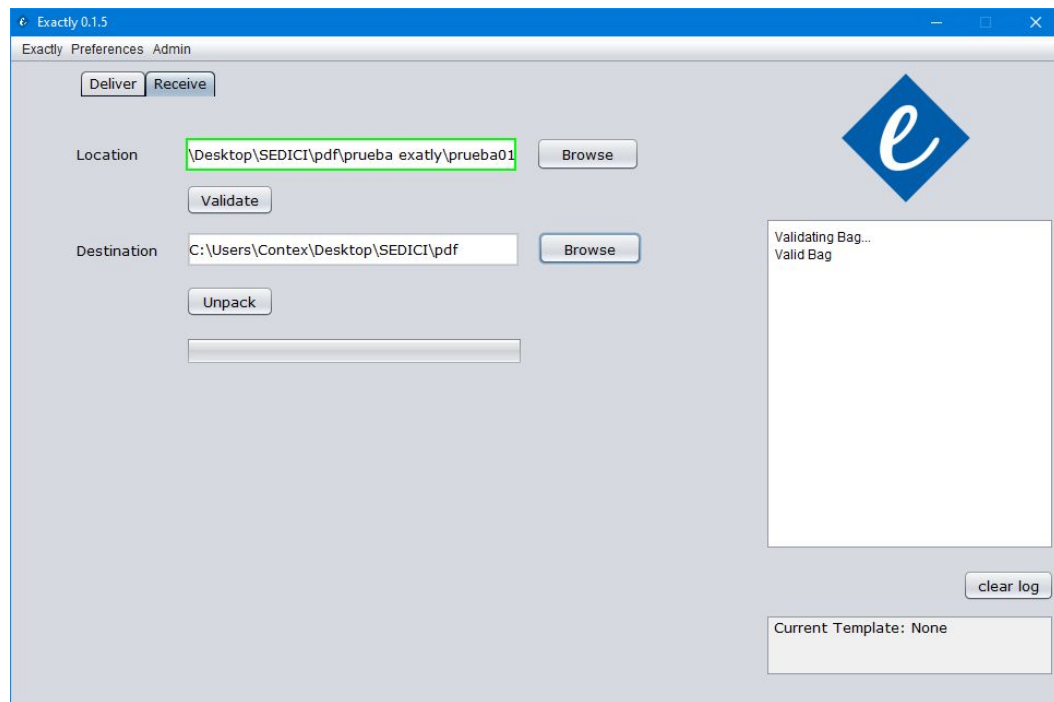
Por último presionamos **Transfer** para iniciar la transferencia, nos mostrará el estado en la ventana de log



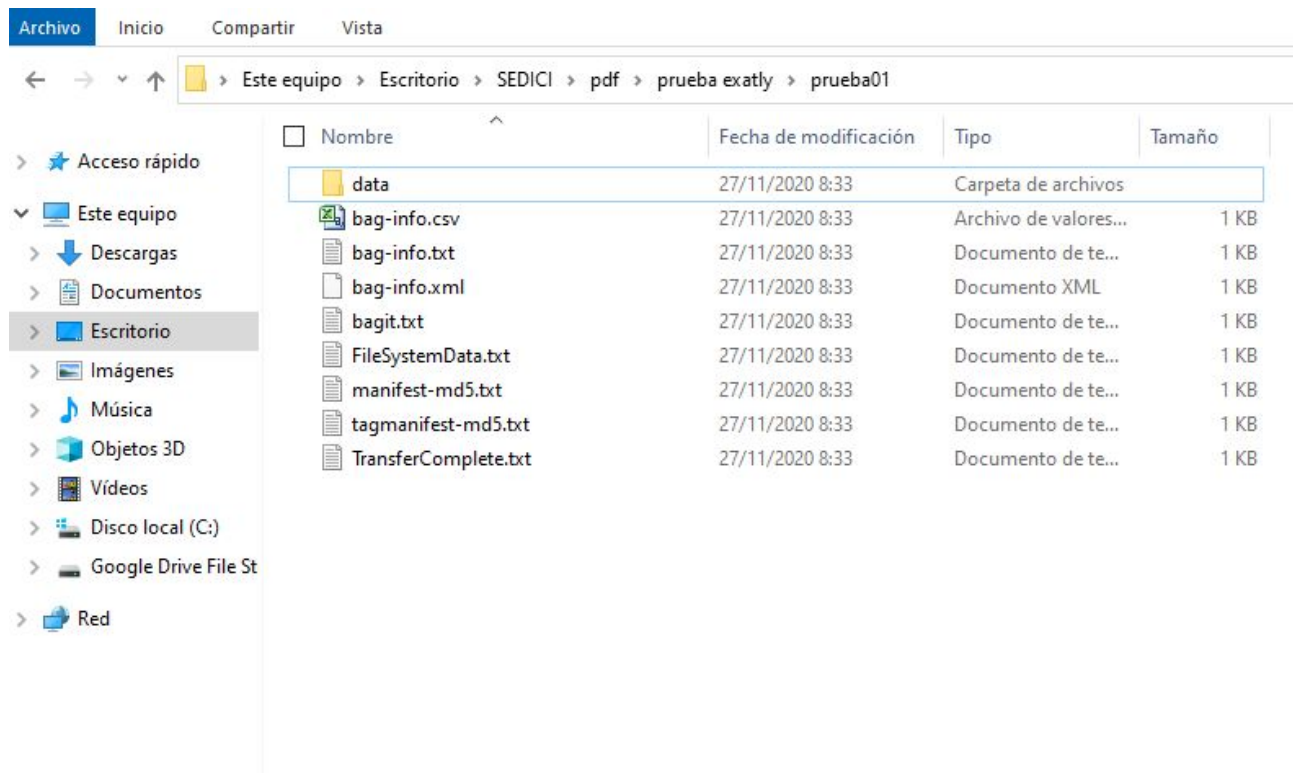
Recibir

Seleccionamos la carpeta donde se encuentra el conjunto de archivos que componen el bag, presionamos “**Validate**” y nos verificará si cumple con la estructura que debe tener un bag.

Luego en **Destination** colocamos la carpeta donde se guardará. Por último presionamos **Unpack**



Carpeta de salida



Data: contiene los archivos empaquetados

Bag-info.csv, Bag-info.txt, Bag-info.xml: Contienen los metadatos y los valores de los mismos

Bagit.txt: contiene la versión de bagit y la codificación utilizada (ej. UTF-8)

FileSystemData.txt: información sobre el volumen donde se guardó, directorio, y archivos

manifest-md5.txt: la suma de verificación de cada archivo (checksum)

tagmanifest-md5.txt: el checksum de los archivos bagit.txt, manifest-md5.txt, bag-info.txt

Transfercomplete.txt: información sobre la transferencia (hora, nombre, directorio donde se guardó)

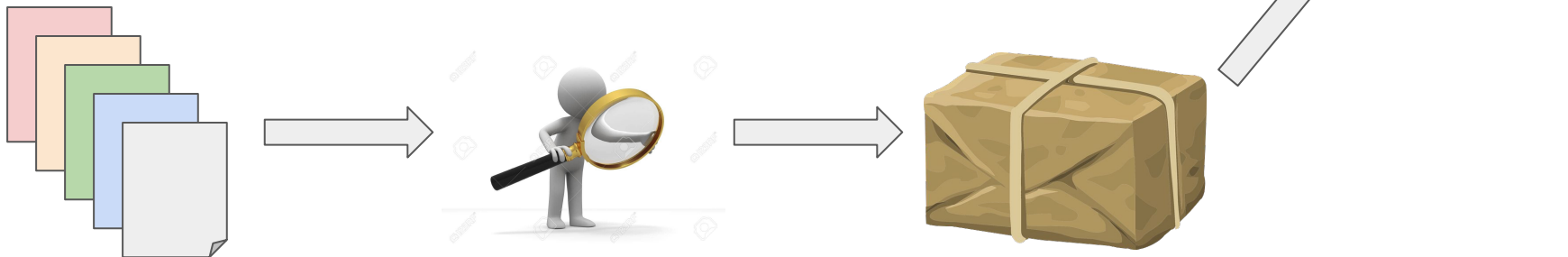
Importación masiva en SEDICI

Los repositorios a menudo precisan recuperar contenidos que pertenecen a autores de la institución y no han sido autoarchivados. Cuando se reconoce un espacio que cuenta con mucho contenido propio se realiza una operación de cosecha e ingesta masiva por procesos informáticos.



Importación masiva a un repositorio

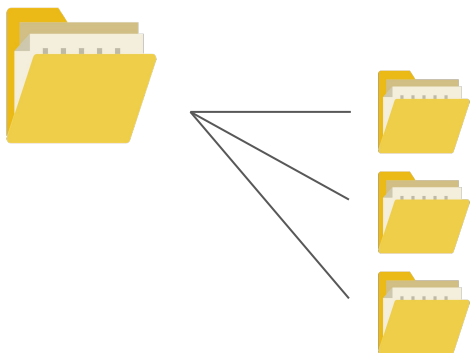
- Se obtienen registros desde múltiples fuentes
- Se realizan mapeos entre esquemas de metadatos
- Se limpian y normalizan los datos
- Se obtienen los objetos digitales asociados
- Se genera un paquete SAF
- Se importa el paquete SAF al repositorio



¿Qué es un paquete SAF?



- Simple Archive Format, definido por DSpace
- Permite empaquetar muchos objetos digitales (junto con sus metadatos) para ser importados o exportados en DSpace.
- Se estructura en base a directorios y subdirectorios



```
archive_directory/  
  item_000/  
    dublin_core.xml  
    metadata_[prefix].xml  
    contents  
    collections  
  
    file_1.doc  
    file_2.pdf  
  item_001/  
    dublin_core.xml  
    contents  
    file_1.png  
    ...
```

¿Cómo se arma un paquete SAF?

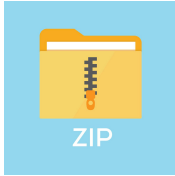
Existe una herramienta de código abierto llamada 'dspace-csv-archive' que permite generar un paquete SAF a partir de un archivo CSV.

dc.title	dc.type	sedici.subtype	mods.originInfo.place	files
Nombre de artículo 1	Article	Article	Institución 1	C:/ruta/al/pdf1
Nombre de artículo 2	Article	Article	Institución 2	C:/ruta/al/pdf2
Nombre de artículo 3	Article	Article	Institución 3	C:/ruta/al/pdf3

A partir de un comando en consola se genera el paquete SAF asociando cada fila del archivo CSV con el archivo correspondiente, cuya ruta se especifica en la columna 'files' del CSV.

Guía de instalación *dspace-csv-archive*

- Descargar la herramienta *dspace-csv-archive* desde el siguiente enlace:
<https://github.com/weilandp/dspace-csv-archive/archive/master.zip>
- Descomprimir el archivo zip en una carpeta



- En caso de no tener Python instalado en la computadora, descargar la última versión desde: <https://www.python.org/downloads/>.



Guía de uso *dspace-csv-archive*

- Abrir la consola del sistema operativo.
- Situarse en el directorio donde se encuentra la carpeta descomprimida durante la instalación.
 - `cd directorio/destino/de/instalación`. En *Windows*, reemplazar `cd` por **CD**.

```
→ ~ cd Sedici/Presentacion
→ Presentacion ls
dspace-csv-archive      ejemplo-presentacion.csv pdfA-solo
→ Presentacion █
```

En Linux, el comando **ls** permite listar los archivos del directorio actual. El equivalente en *Windows* es el comando **DIR**.

Guía de uso *dspace-csv-archive* (cont.)

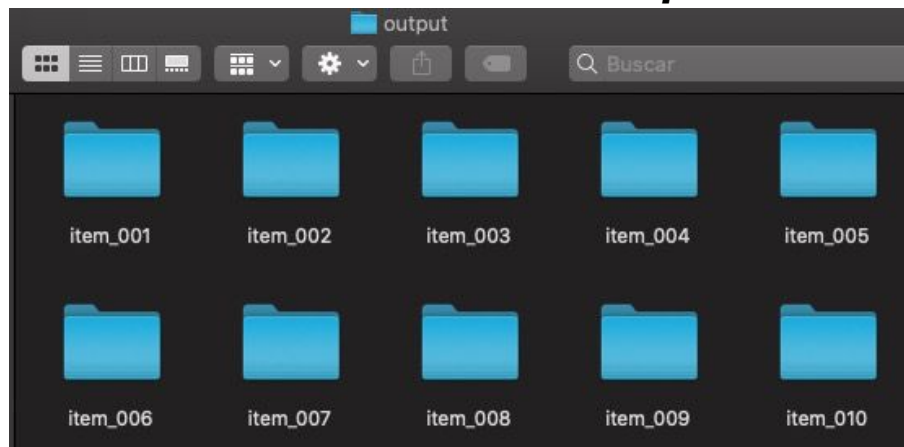
- Ejecutar el comando **python** o **python3** seguido de la ruta al archivo ***dspace-csv-archive.py*** (se encuentra dentro de la carpeta descomprimida en la instalación) y el archivo csv con los registros a importar.

```
[→ Presentacion python3 dspace-csv-archive/dspace-csv-archive.py ejemplo-presentacion.csv
[→ Presentacion ls
dspace-csv-archive      ejemplo-presentacion.csv output          output.zip      pdfA-solo
[→ Presentacion █
```

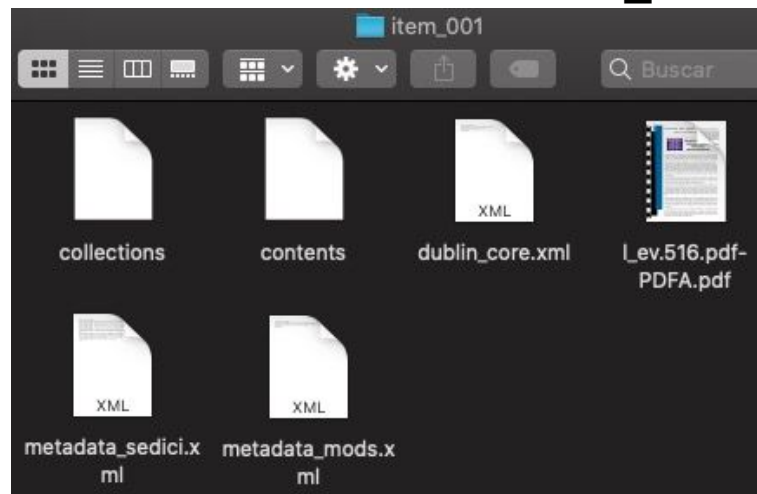
Se genera el archivo *output.zip* (paquete SAF comprimido) y el directorio *output* (paquete SAF sin comprimir). El archivo que será importado al repositorio es ***output.zip***.

Estructura de los directorios

Contenido del directorio *output*



Contenido del directorio *item_001*



- Los metadatos de cada ítem se encuentran en los archivos *.xml* (uno para cada esquema utilizado)
- El archivo *collections* contiene los *handles* de las colecciones destino del ítem en el repositorio
- El archivo *contents* enumera los objetos digitales asociados, en este caso: I_ev.516.pdf-PDFA.pdf.

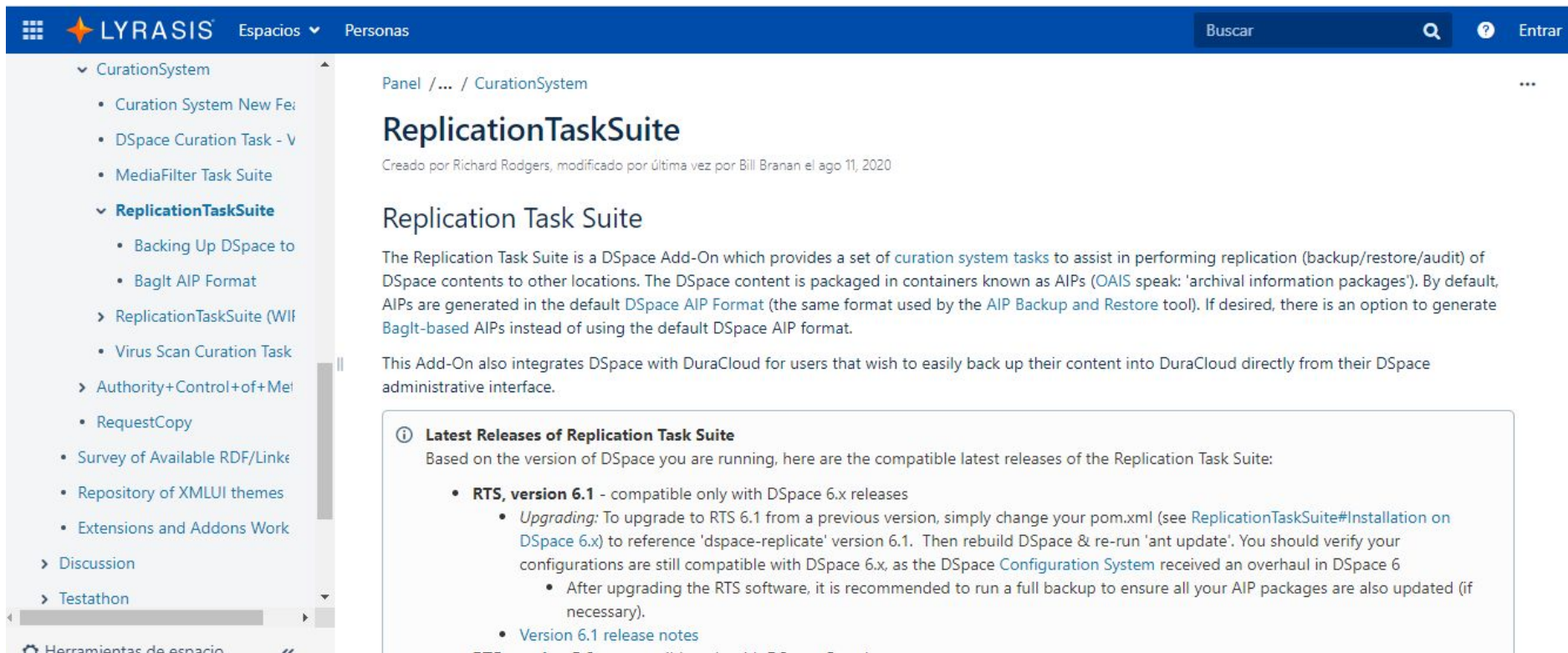
Desventajas paquete SAF

- Desde el punto de vista de la preservación digital, un paquete SAF no es suficiente.
 - No se generan checksums de los objetos digitales.
 - No se cumple con el estándar PREMIS.
 - DSpace utiliza el metadato dc.provenance y no es de gran utilidad.
- Resulta necesaria la posibilidad de incorporar una gran cantidad de paquetes SIP al repositorio.



¿Se puede acoplar una herramienta como Exactly al software de repositorio?

Backup por fuera del repositorio:AIP



The screenshot shows the LYRASIS website interface. The top navigation bar is blue with the LYRASIS logo, a search bar, and a login button. The left sidebar contains a tree view of the site structure, with 'ReplicationTaskSuite' selected. The main content area displays the 'ReplicationTaskSuite' page, which includes a breadcrumb trail, a title, a creation/modification date, a description of the suite, and a section for the latest releases.

LYRASIS Espacios Personas Buscar Entrar

Panel / ... / CurationSystem

ReplicationTaskSuite

Creado por Richard Rodgers, modificado por última vez por Bill Branan el ago 11, 2020

Replication Task Suite

The Replication Task Suite is a DSpace Add-On which provides a set of [curation system tasks](#) to assist in performing replication (backup/restore/audit) of DSpace contents to other locations. The DSpace content is packaged in containers known as AIPs (OAIS speak: 'archival information packages'). By default, AIPs are generated in the default [DSpace AIP Format](#) (the same format used by the [AIP Backup and Restore](#) tool). If desired, there is an option to generate Baglt-based AIPs instead of using the default DSpace AIP format.

This Add-On also integrates DSpace with DuraCloud for users that wish to easily back up their content into DuraCloud directly from their DSpace administrative interface.

Latest Releases of Replication Task Suite

Based on the version of DSpace you are running, here are the compatible latest releases of the Replication Task Suite:

- RTS, version 6.1** - compatible only with DSpace 6.x releases
 - Upgrading:** To upgrade to RTS 6.1 from a previous version, simply change your pom.xml (see [ReplicationTaskSuite#Installation on DSpace 6.x](#)) to reference 'dspace-replicate' version 6.1. Then rebuild DSpace & re-run 'ant update'. You should verify your configurations are still compatible with DSpace 6.x, as the DSpace [Configuration System](#) received an overhaul in DSpace 6
 - After upgrading the RTS software, it is recommended to run a full backup to ensure all your AIP packages are also updated (if necessary).
 - [Version 6.1 release notes](#)

<https://wiki.lyrasis.org/display/DSPACE/ReplicationTaskSuite>

Trabajos futuros

- Buena parte de los repositorios del mundo están en DSPACE
- DSPACE permite la ingesta de un paquete SAF
- Entre los repositorios Dspace se pueden intercambiar paquetes, incluso AIPs. Esta es una buena opción para backups: <https://wiki.lyrasis.org/display/DSPACE/ReplicationTaskSuite>
- DSPACE está vinculado con Bagit
 - <https://github.com/nye-duo/BagItLibrary>

¡Muchas gracias!

Consultas

marisa.degiusti@sedici.unlp.edu.ar

Presentación disponible en:

<http://sedici.unlp.edu.ar/handle/10915/25293>

<https://digital.cic.gba.gob.ar/handle/11746/4083>



Esta obra está bajo una [Licencia Creative Commons Atribución-NoComercial-CompartirIgual 4.0 Internacional](#)