Benemérita Universidad
Autónoma de San Luis
Potosí (UASLP), México

Conferencia internacional
BIREDIAL-ISTEC
17-18-19 OCTUBRE 2016
congresos.unlp.edu.ar/**biredial-istec**

Un modelo de trabajo para agilizar la generación de documentos de texto para su preservación

Paula Salamone Lacunza
Gonzalo L. Villarreal
Marisa R. De Giusti
Ariel J. Lira

PREBI prebi.unlp.edu.ar
SEDICI sedici.unlp.edu.ar
UNIVERSIDAD NACIONAL DE LA PLATA

Actividades de Preservación

Algunas actividades de preservación típicas:

- Análisis de los formatos de los archivos.
- Selección del mejor formato de transformación o migración.
- Transformación o migración de archivos.
- Verificación de las transformaciones realizadas.
- Validación según las reglas del estándar utilizado.

Requieren una importante carga adicional para los

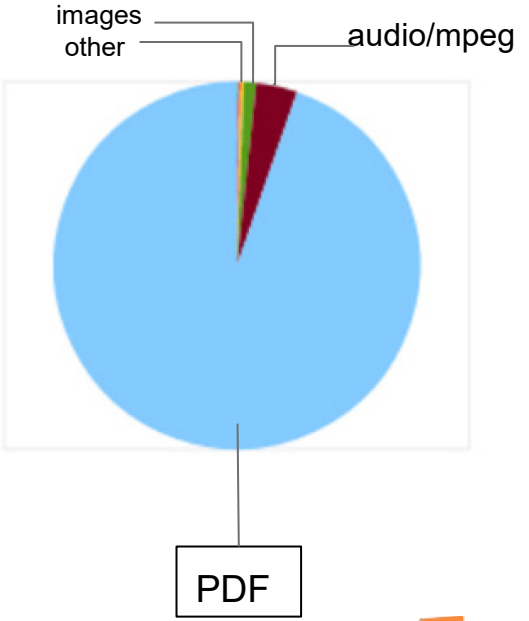
Contexto

Este trabajo se realizó en el marco del repositorio **SEDICI**, perteneciente a la Universidad Nacional de La Plata.

SEDICI cuenta con más de 50 mil recursos.

El **95%** de ellos son documentos de texto en formato PDF.

> ➤ En el caso particular de los documentos de texto, el estándar **PDF/A** (ISO 19005 1-2-3) se impone como el formato más apropiado para su preservación.



sedici.unlp.edu.ar

PDF/A

Basado en el estándar PDF 1.4, al que le incorpora requerimientos adicionales, como ser:

1. Especificaciones sobre los metadatos y la estructura del archivo.
2. La paleta de colores (incluyendo escala de grises y blanco/negro) no deben ser representados en un espacio de color de dispositivo (DeviceRGB, DeviceCMYK, DeviceGray).
3. Las fuentes usadas en texto visibles deben estar embebidas (incluídas dentro del archivo).

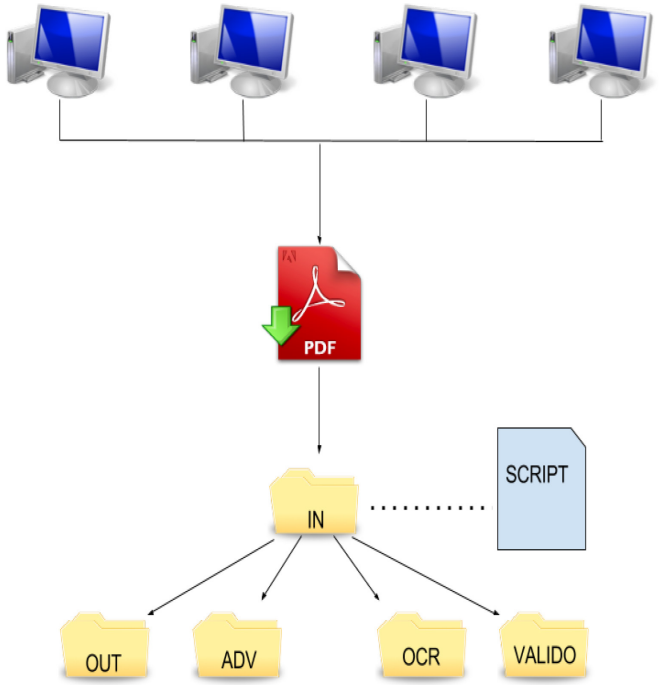
Un modelo de trabajo centralizado

- La propuesta de este trabajo se basa en un modelo de trabajo centralizado.
- Se concentra el procesamiento de los documentos en un equipo dedicado.
- El procesamiento es desatendido.
- La implementación de este modelo se basa en un sistema de red estilo cliente-servidor:
  - los clientes (administradores del repositorio) acceden a un directorio compartido por una máquina de la red y depositan los documentos que deben ser procesados.
  - Esta máquina de la red (proceso servidor) toma los documentos
    - analiza su formato para determinar las tareas que deben realizarse

Flujo de trabajo de la Herramienta

Secciones (directorios):

- **IN**: directorio de entrada de documentos.
- **OUT**: correcta transformación.
- **Válido**: documentos PDF/A válidos.
- **OCR:** documentos que no se pueden convertir/reparar.
- **ADV**: documentos que



Módulos de la Herramienta

- Detección de archivos.
- Pre-Verificación:
  - Archivo PDF
  - Archivo no corrupto
- Conversión:
  - Elección del estándar PDF/A.
  - Llamada al programa conversor.
- Post-Verificación:

report-File-In.txt

```
- Opening file In.pdf.
- Analyzing In.pdf.
- Copied output intent from input file.
- Performing post analysis for In-PDFA.pdf.
- Post analysis for In-PDFA.pdf has been successful.
- File In.pdf converted successfully.

Codigo: 5
```

Reporte de conversión de un archivo satisfactorio (código de aceptación 5).

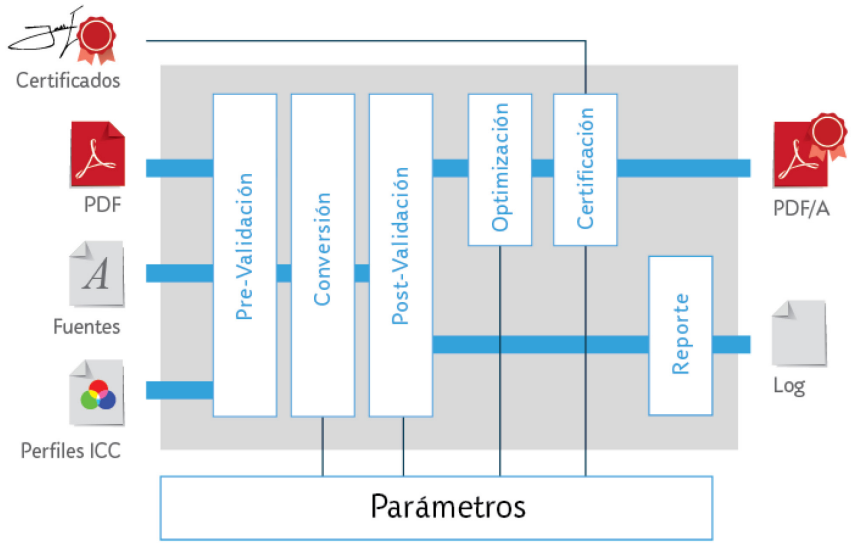
Módulo de Conversión

1) Programas de escritorio: Acrobat DC, pdfaPilot, pdfToolbox.

2) Programas CLI: pdfaPilot, 3Height

Módulos de Verificación

- Programas de escritorio: Acrobat DC.
- Programas CLI: Exiftool, 3Height



Conclusiones

Flexibilidad

Ingreso de nuevas estaciones de trabajo

El script puede ser fácilmente cambiado de computador.

Independencia

Entre cada estación de trabajo

Respecto a las herramientas de procesamiento de documentos

Paula Salamone Lacunza, paula@sedici.unlp.edu.ar
Gonzalo L. Villarreal, gonzalo@prebi.unlp.edu.ar
Marisa R. De Giusti, marisa.degiusti@sedici.unlp.edu.ar
Ariel J. Lira, alira@sedici.unlp.edu.ar

sedici.unlp.edu.ar