LibLink
Library Linkages

XVIII ASSEMBLEIA GERAL DO ISTEC
ISTEC GENERAL ASSEMBLY
ASAMBLEA GENERAL DE ISTEC
Porto Alegre - Rio Grande do Sul - Brazil, May 16th to 20th, 2011

VI SIBD
VI Simpósio Internacional de Bibliotecas Digitais

Arquitectura ETL para la recolección de metadatos

Prof. Ing. **Marisa R. De Giusti**
Servicio de Difusión de la Creación Intelectual
Proyecto de Enlace de Bibliotecas
Universidad Nacional de La Plata
Directora de la Iniciativa LibLink

istec

Servicio de Difusión de la Creación Intelectual (SeDiCI)

SeDiCI es el repositorio institucional de la Universidad Nacional de La Plata (UNLP), creado con dos objetivos prioritarios:

•Para atender al rol fundamental de una institución pública de **socializar** el conocimiento.

•Dar mayor visibilidad a la producción académica a través del **acceso libre que posibilita un mayor impacto.**

Arquitectura ETL para la recolección de metadatos

La recolección y agregación de recursos es una de las actividades más comunes en el área de los repositorios digitales.

En general se busca incrementar el volúmen de documentos ofrecidos a los usuarios, para que éstos cuenten con una base documental de consulta más amplia para sus investigaciones y desarrollos.

Este trabajo presenta una aproximación al problema de la recolección de recursos desde distintas fuentes datos

Recolección de recursos

Problemas generales

Es importante notar que la recolección tiene sentido siempre que la información recolectada sea relevante para el repositorio que la realiza, lo cual podría definirse según:

- Temática adecuada
- Tipología documental aceptada
- Calidad mínima

Problemas frecuentes de los agregadores de recursos

- Problemas vinculados a la obtención de los recursos.
- Problemas vinculados a la mejora de los recursos.
- Problemas vinculados al almacenamiento de los recursos.

Recolección de recursos

Problemas generales

Cómo tratar con los múltiples protocolos y técnicas de comunicación y transferencia:

- OAI-PMH: protocolo simple para el intercambio de metadatos
- Web-Crawling: técnica que recorre páginas web, extrayendo contenido
- Web-services: comúnmente sobre SOAP o XML-RPC, ofrecen servicios especiales.

Y las diferentes formas de representación de los mismos:

- XML, HTML, tuplas de una base de datos, documentos no estructurados…

Recolección de recursos

Problemas generales

Cada protocolo o método de recolección es distinto incluso a distintos niveles:

Comunicación y transferencia

Al usar la red cambian los parámetros de conexión, los protocolos de la capa de aplicación utilizados, formato de mensajes, tiempos de espera, etc.

Formato de datos

Dependiendo de cual sea el sistema final con el que se esté realizando la comunicación, los datos se transmiten de distinta forma: archivo binarios o de texto, comprimidos, en porciones, etc.

Interpretación de la información

Una vez que se cuenta con la información hay que interpretarla: tuplas obtenidas de una base de datos, XML bajo algún schema, archivos de texto delimitado por comas, codificación de caracteres diferente, etc.

Recolección de recursos

Problemas generales

Una vez obtenida e interpretada la información se observan grandes problemas derivados de la heterogeneidad. Por ejemplo:

Nombres de Autores: "Gomez, Juan Carlos", "Gomez Juan C.", "Gomez, J. C.", "Juan Carlos Gomez", "Juan C. Gomez"
Formato de fechas: "2011-05-20", "20-05-2011", "20-may-2011", 20/05/2011", "05-20-2011"
Sistemas de Clasificación: LCC, DDC, sistemas de clasificación temáticos, uso de códigos y uso de términos, etc
Tesauros: UNESCO, Eurovoc, DECs, etc.
Idiomas: ISO-639-X, nombre del idioma, nombre en otros idiomas, etc.
Campos ausentes: errores de mapeo (en el origen o el destino), falta de información durante la catalogación en el origen, etc.
Información errónea o concatenada: idioma con "english;eng"; fecha de publicación con "PUB:2011-05-25", autor con "Gomez, Juan C.; Lopez, Mario A."

Recolección de recursos

Problemas generales

Estos problemas de heterogeneidad deben ser disminuídos o resueltos para lograr un aprovechamiento eficiente de la información.

Esto se logra a través de procesos de análisis y transformación, en lo posible automáticos.

Recolección de recursos

Problemas generales

Todos los recursos recolectados y transformados deben ser almacenados en algún lugar para su posterior uso.

El criterio según el cual se determina el destino para esta información depende del uso que se desea dar a dichos datos:

- búsqueda y recuperación → motor de indexación
- backup → servidor de backup
- compartición → datos para un sistema de interoperabilidad
- estadísticas → base de datos relacional

Asimismo, las transformaciones a aplicarse dependen del destino y del uso planeado.

Recolección de recursos

Problemas generales

De forma análoga a la recolección, para el almacenamiento existen muchas alternativas.

- Motor de Indexación
- Bases de datos
- Web-services
- Archivos

Cada una determina protocolos y mecanismos de comunicación y transferencia, formato de datos, y reglas para la interpretación de los datos.

Esto requiere ser capaz de manejar cada tipo de almacenamiento

Abstracción del problema

Desde un nivel de abstracción elevado se observan tres grandes etapas:

•**Extracción**: recolección de recursos desde las distintas fuentes de datos.

•**Transformación**: disminución o anulación de los problemas derivados de la heterogeneidad de la información.

•**Carga**: almacenamiento final de la información.

ETL → ***Extract, Transform & Load***

Arquitectura ETL

ETL es un patrón arquitectural de software del área de Integración de Datos.Usado principalmente en aplicaciones de Data Mining y Bussiness Intelligence.

Sus tres principales componentes:

•**Extract**: obtención de los datos desde bases de datos, archivos, etc.

•**Transform**: unificación y normalización de la información, con el fin de cruzar datos y obtener nueva información no visible inicialmente.

•**Load**: depósito de la información obtenida en un Data Warehouse, para su posterior consulta.

ETL es valorado en los escalafones gerenciales, ya que provee información valiosa para la toma de decisiones.

Aproximación a una solución

Arquitectura ETL en el ámbito de los repositorios digitales

Las tres principales actividades detectadas desde un nivel de abstracción elevado se condicen adecuadamente con las actividades modeladas en el patrón ETL.

Se propuso la creación de una herramienta que implemente este patrón, para finalmente permitir la recolección de datos desde múltiples orígenes, su transformación y normalización, y su posterior almacenamiento en múltiples almacenes de datos.

Aproximación a una solución

El diseño de la aplicación debería cumplir con las siguientes premisas:

- Permitir la recolección desde múltiples tipos de fuentes de datos.
- Permitir el almacenamiento en múltiples tipos de almacenes de datos.
- Permitir la selección y configuración de los filtros de análisis y transformación disponibles en la aplicación.

En los tres casos, la aplicación debería ser extensible a través del uso de componentes conectables.

Aproximación a una solución

El diseño de la aplicación debería cumplir con las siguientes premisas:

- Llevar los recursos a una representación abstracta a fin de normalizar la lógica de los componentes de análisis y transformación (todos los componentes reciben la misma entrada, sin importar el origen de los datos).

- Proveer una interfaz de administración desde la cual se permita manejar todos los aspectos de la herramienta.

- Garantizar la tolerancia a fallos (interrupciones en el servicio eléctrico, problemas de conexión, etc) y la reanudación automática de las actividades interrumpidas.

- Proveer información estadísticas sobre la actividad de la herramienta, las fuentes y almacenes de datos utilizados, etc.

Modelo de Datos

Visión general

Tres elementos más importantes del modelo son:

- Repositorios
- Definiciones de Cosecha
- Colecciones

El modelo de datos completo se desarrolla a partir de estos tres componentes.

Modelo de Datos

Repositorios

- Representan a los repositorios digitales que serán utilizados como fuente de datos.
- A cada repositorio se asigna al menos un Conector, el cual provee la información necesaria para realizar la conexión y recolección de documentos desde dicho repositorio.
- Dentro de la aplicación existen varios tipos de Conectores, uno para cada protocolo o forma de conexión y recolección.
- Estos conectores son uno de los puntos de extensión de la aplicación, ya que son componentes relativamente independientes.
- En el caso de SeDiCI (como fuente de datos para recolección), se podría configurar un Conector OAI para Harvesting OAI, y un Conector Web-Services para recolectar documentos no expuestos por OAI.

Modelo de Datos

Definiciones de Cosecha

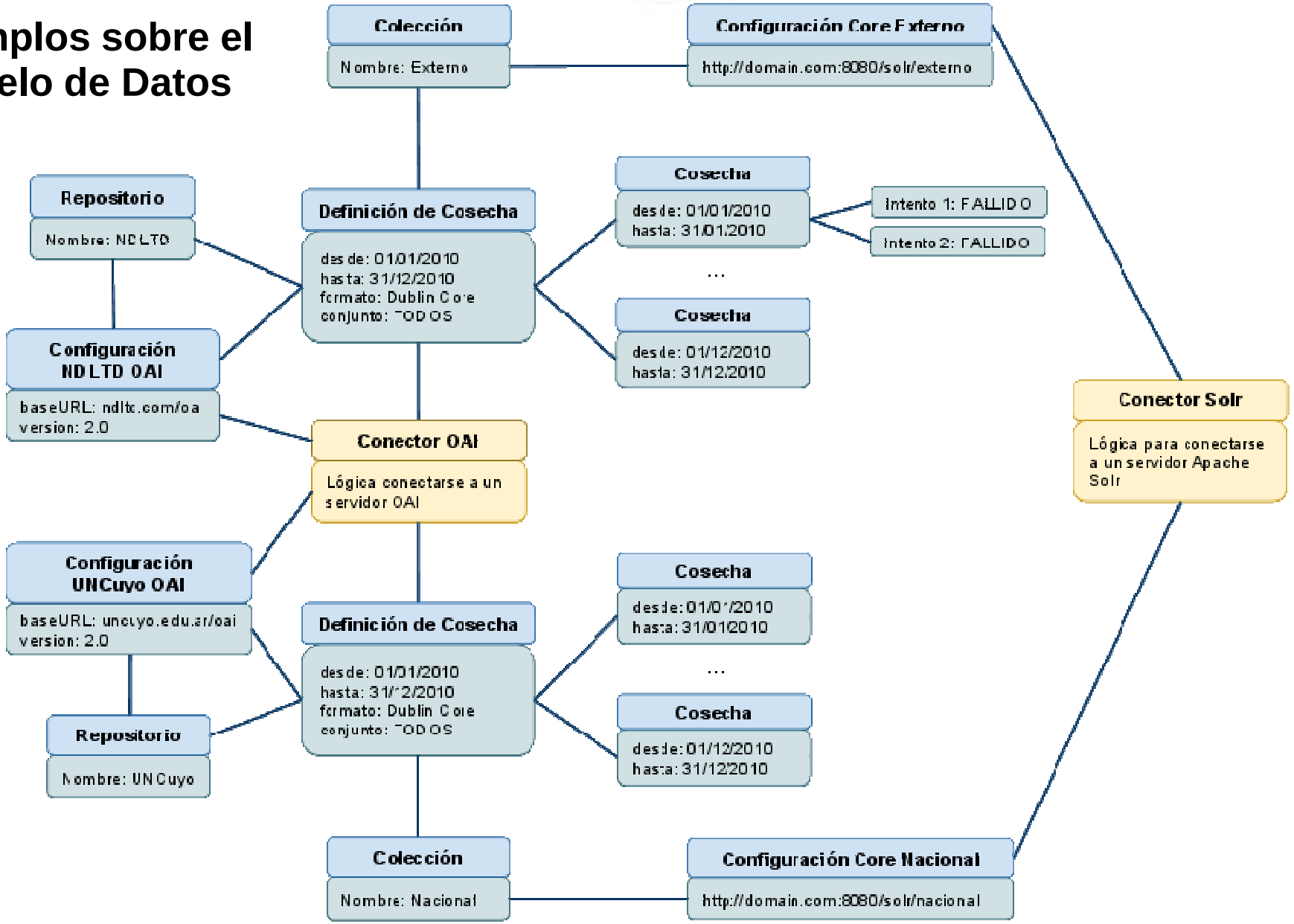
- Creadas a partir de un repositorio y un conector en particular.
- Definen los parámetros correspondientes a una recolección en particular.
- Los parámetros de cada Definición de Cosecha dependen explícitamente del conector al que pertenecen.
- Se permite dividir las tareas de cosecha en partes independientes. La ejecución de la cosecha se considera finalizada sólo cuando todas sus partes fueron completadas (ej. rangos de fechas en OAI).
- Se registra cada intento de cosecha y su resultado: satisfactorio o fallido, para permitir la recuperación ante fallas.
- A partir de estos elementos se genera información estadística.

Modelo de Datos

Colecciones

- Representan los almacenes de datos que serán el destino de la información recolectada y transformada.
- Cada Colección tiene asociado un conector, que contiene los parámetros y la lógica específica para un tipo de almacén en particular.
- La aplicación provee varios tipos de Conectores, uno para cada protocolo o forma de almacenamiento.
- Estos conectores son otro de los puntos de extensión de la aplicación, ya que son componentes relativamente independientes.
- Un ejemplo en SeDiCI es el uso del motor de indexación Apache Solr, para el cual existe un Conector Solr que contiene la lógica de conexión y transferencia, y cada Colección que lo usa especifica los parámetros del servidor Solr que se debe utilizar.

Ejemplos sobre el Modelo de Datos



ETL

Etapa de Extracción

- Selección de Definiciones de Cosecha a ejecutar, según información de programación.
- Generación de cosecha a ejecutar (particionamiento), o selección de una cosecha existente, cuyo último intento fue fallido.
- Ejecución de la recolección (a partir de información del conector y la Definición de Cosecha).
- Registro de resultado de la cosecha (recuperación ante fallas y estadísticas).

ETL

Etapa de Transformación

- Transformación de los recursos a una representación común.

- Ejecución de cadena de filtros sobre cada recurso. Entre otros, existen filtros de:
- Vocabularios controlados
- Eliminación
- Duplicación
- Tokenization
- Valor por Defecto
- La ejecución de cada filtro modifica el recurso en algún aspecto.

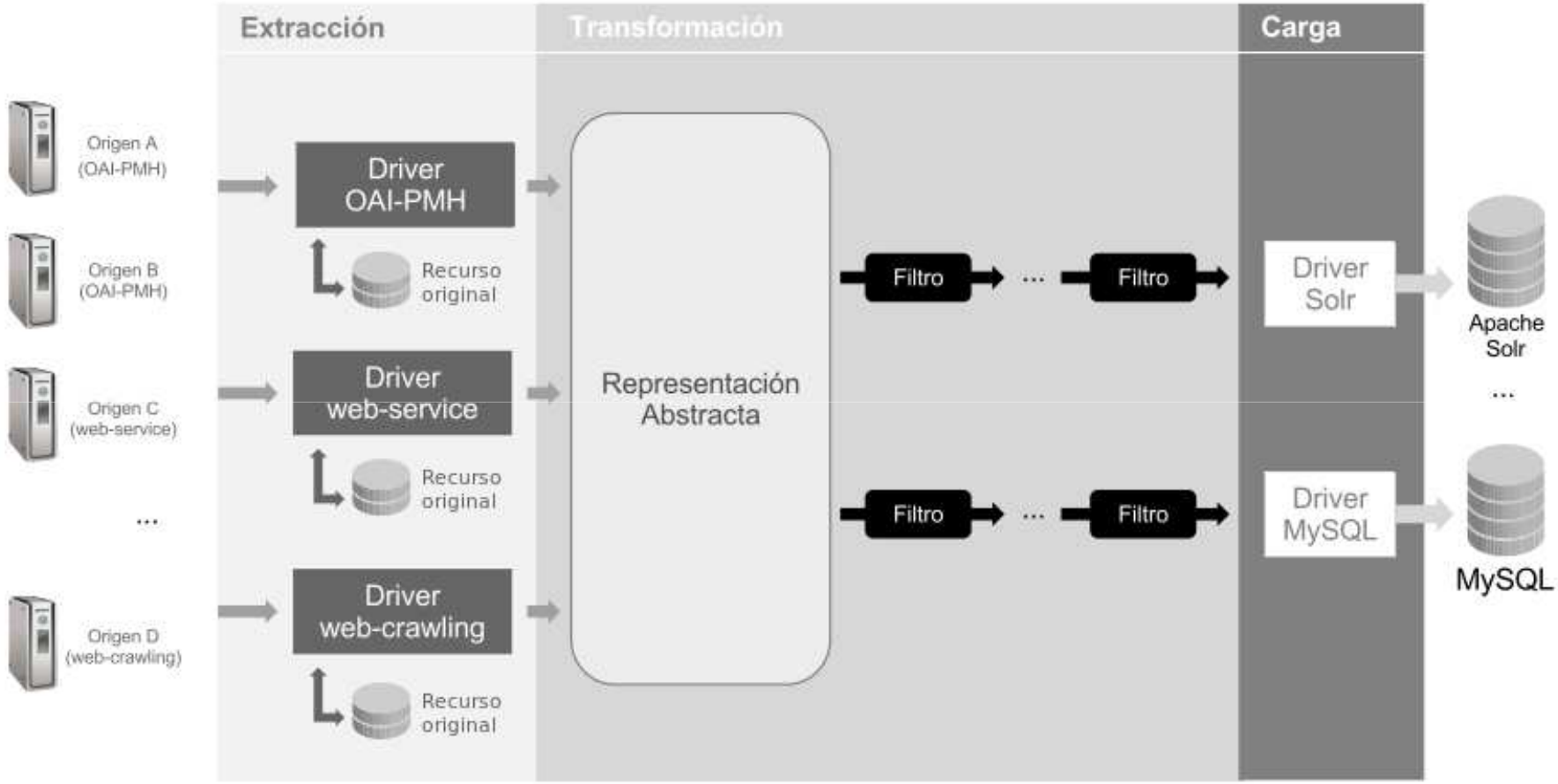
ETL

Etapa de Carga

- Los recursos son enviados al conector de almacenamiento, en donde:
- Se transforman a la representación adecuada según el almacenamiento
- Se conecta al almacén y se envían los recursos
- Se registra el resultado de la inserción (COMPLETADA o FALLIDA) para recuperación ante fallas y generación de estadísticas.

ETL

Diagrama de Arquitectura



Administración

Se provee una interfaz web de administración que permite:

- Administración de Repositorios, Colecciones, Definiciones de Cosechas, etc.
- Selección de Filtros a aplicar (por Colección).
- Generación de estadísticas.
- Control (Iniciar/Finalizar) sobre los procesos de recolección.

Puntos de extensión y trabajos futuros

- Nuevos Filtros y Tranformaciones
- Descarga automática del Fulltext
- Normalización de autores
- Detección de duplicados
- Extracciones semánticas

MUCHAS GRACIAS!!!

Ing. Marisa R. De Giusti
marisa.degiusti@sedici.unlp.edu.ar

Nestor F. Oviedo
nestor@sedici.unlp.edu.ar

Lic. Ariel J. Lira
alira@sedici.unlp.edu.ar

SeDiCI – Servicio de Difusión de la Creación Intelectual
http://sedici.unlp.edu.ar

PrEBi – Proyecto de Enlace de Bibliotecas
http://prebi.unlp.edu.ar

Universidad Nacional de Plata