Herramienta de Validación aplicada a tareas de control de calidad en repositorios digitales

Facultad de Informática
UNIVERSIDAD NACIONAL DE LA PLATA

Autor Franco Agustín Terruzzi
Director Dra. Marisa Raquel de Giusti

Introducción

Motivación

Objetivos

Aporte

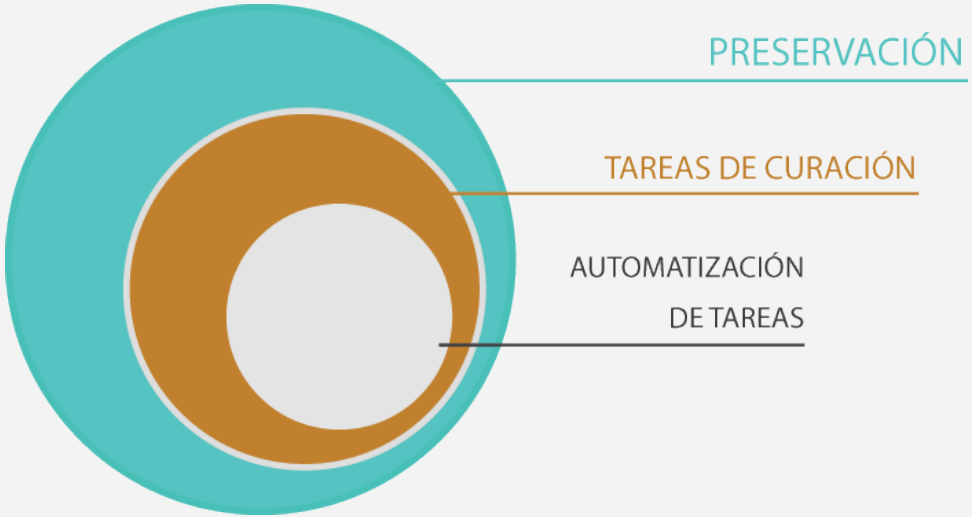
- Especificación de la herramienta
- Implementación para el SEDICI

Conclusiones

Trabajos Futuros

Motivación

Repositorios



Motivación - Contexto

REPOSITORIOS

“Conjunto de servicios centralizados, creados para organizar, preservar y ofrecer acceso a la producción científica, académica, administrativa o de cualquier otra naturaleza, en soporte digital, generada por los miembros de una institución.”

Motivación - Contexto

Repositorios (caso SEDICI)

- Se creó en el año 2003
- Su principal objetivo es socializar el conocimiento
- Más de **40.000** obras que provienen de toda la UNLP
  - Tipologías muy variadas
    - Tesis
    - Audio
    - Cuadros
- Varios campos de estudio involucrados

Motivación - Contexto

PRESERVACIÓN

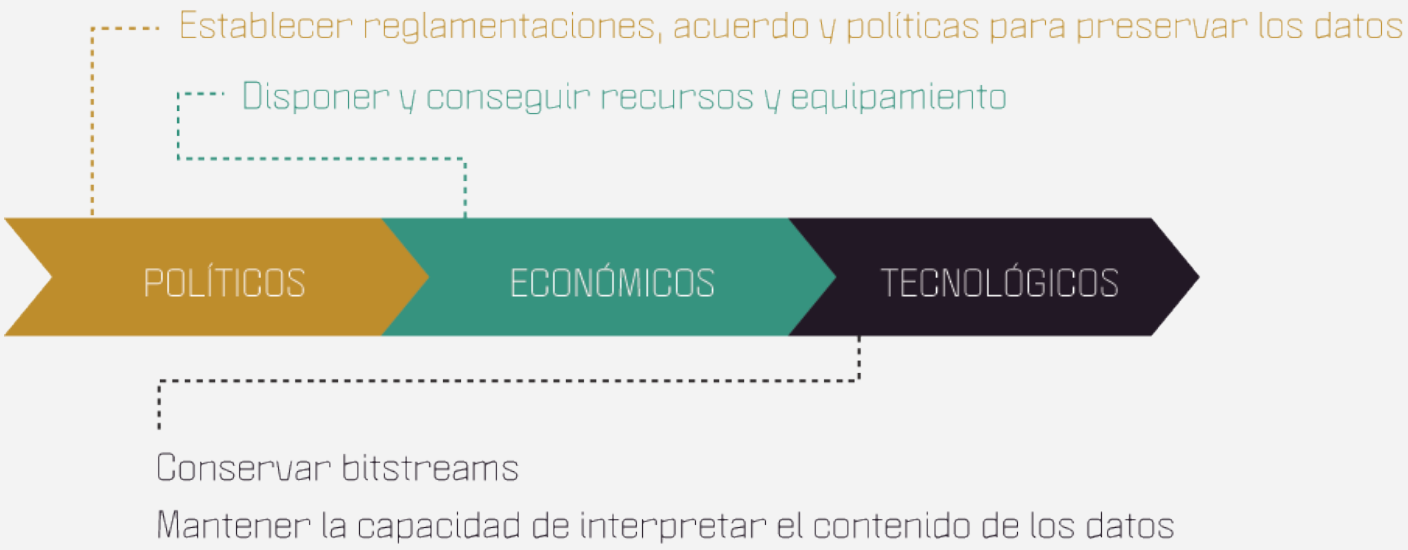
“El conjunto de prácticas de naturaleza política y estratégica, y las acciones concretas destinadas a asegurar el acceso a los objetos digitales a largo plazo.”

UNESCO (United Nations Educational, Scientific and Cultural Organization)

Motivación - Contexto

Preservación

Actividades



Motivación - Contexto

Preservación

Aspectos Importantes

CONTROL DE CALIDAD

Realizar validaciones sobre los distintos elementos de un repositorio y detectar los que no las cumplen

CURADURÍA

Realizar actividades vinculadas al ciclo de vida de los objetos digitales

- Ej: Reparar los posibles elementos que se han encontrado como inválidos

Motivación - Planteo de la necesidad

Permitir que los administradores puedan determinar ***qué*** *elementos evaluar* y ***qué*** *validaciones realizar*. Además de **cómo** arreglar aquello que no ha pasado estas validaciones.

A partir de:

- Controles semiautomáticos de calidad en los distintos elementos del repositorio.
- Tareas de curaduría sobre los elementos en los que se localizan problemas

Objetivos

Realizar un aporte para los mecanismos de control de calidad dinámicos en un repositorio institucional

- Construir y especificar una herramienta de validación dinámica para facilitar tareas recurrentes en la preservación y la curaduría.
- Dar una posible solución al planteo del problema de este trabajo, a través de esta herramienta.
- Utilizar la herramienta desarrollada dentro de un contexto real, sobre un caso de uso concreto

Aporte

Formulación

¿QUÉ?

Construcción y especificación de una herramienta que permita determinar ciertas tareas en un repositorio institucional

¿PARA QUÉ?

Para intentar cumplir con los objetivos planteados anteriormente

Aporte

Formulación

¿CÓMO?

1. Estudio de un Contexto para la solución del problema
2. Especificación de la herramienta
3. Alternativas de Implementación
4. Implementación

Aporte - Marco Teórico

¿POR QUÉ?

Se intenta asegurar la expresividad de la herramienta y su cercanía con los términos del dominio en cuestión

Metodología

Dos aportes

Especificación de una herramienta:

- Planteo de una solución general
- Utilización de patrones de diseño y herramientas análisis del ámbito de DSL.
- Caracterización con propiedades de lenguajes generales
- Ortogonalidad

Implementación para SEDICI-DSPACE:

- Desarrollo de una solución específica
- Utilización de técnicas y herramientas de implementación provenientes al ámbito de DSL
- Utilización de conceptos de dominio DSpace.

Especificación de la Herramienta

Requisitos funcionales

- Validación de condiciones

Dado un objeto **O** y una regla **R** que define una o más condiciones C_1 C_2..C_n, se aplicará **R** sobre **O**, evaluando si **O** cumple C_i para todo **i = 1..n** o para al menos uno, según sea una conjunción o disyunción respectivamente.

- Selección de Objetos

Seleccionar un subconjunto (**S**) de objetos que cumplan la(s) condición(es) **P**, donde **P** podría ser una validación.

- Transformación

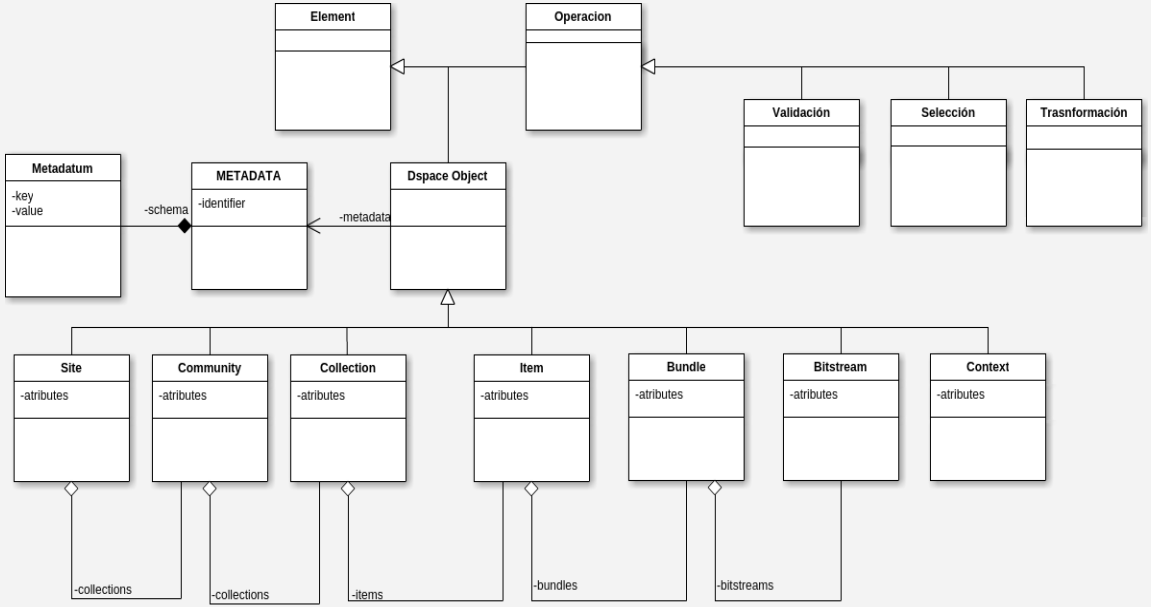
Dado un objeto **O** seleccionado, aplicar una función de transformación **T(O)**, que lleva al objeto **O** a un objeto **O'** a partir del cambio de al menos uno de los metadatos M_1, M_2,..,M_n que pertenecen a **O**.

Especificación de la Herramienta

Representación de la herramienta

Modelo estructural

- Ortogonalidad
- Amplitud
- Reutilización
- Adaptabilidad



Implementación para el SEDICI

Solución específica

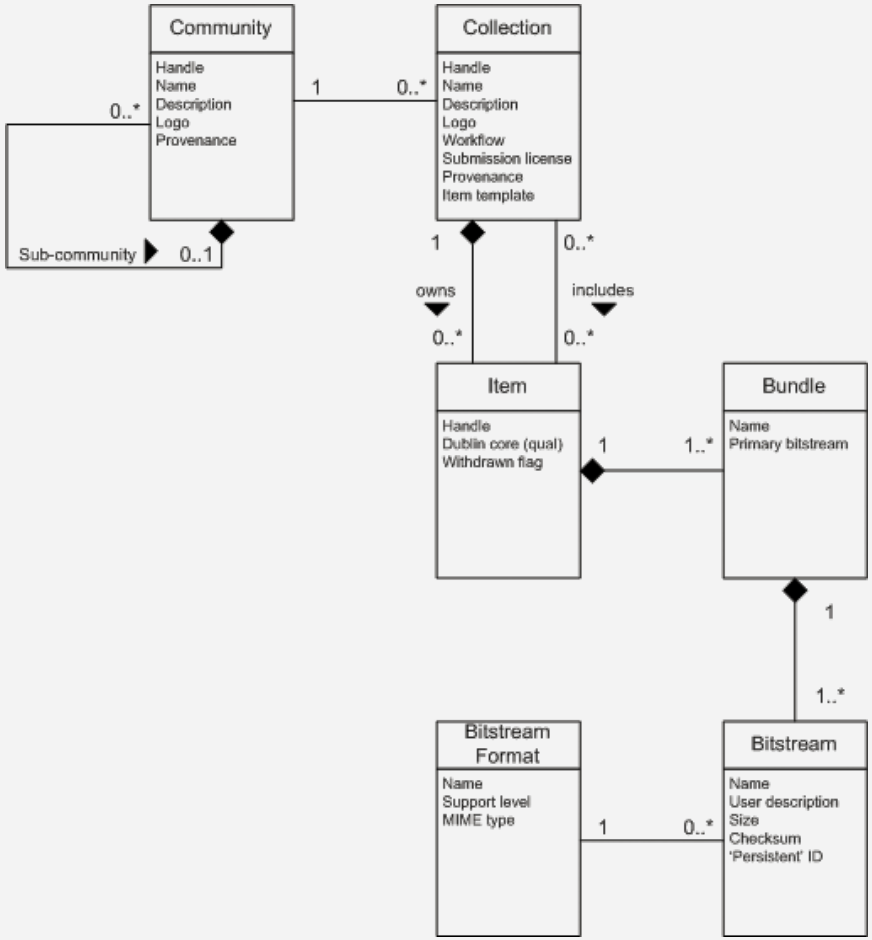
- Una implementación posible para la especificación
- Adaptada para ejecutarse en el contexto de **DSpace**
- Personalizada para los términos utilizados en **SEDICI**
- Implementada con patrones y herramientas para implementar **DSL**

Implementación para el SEDICI

Conceptos de DSpace

Definición del Software

“Sistema de administración ampliamente utilizado alrededor del mundo para la gestión de repositorios digitales.”



Implementación para el SEDICI

Herramientas utilizadas

JSR-341 - Brinda

- Una sintaxis concreta simple, restringida a la evaluación de expresiones
- Operadores relacionales, lógicos, aritméticos, condicionales y de vacío (empty)
- Posibilidad de implementar funciones propias, para ampliar el comportamiento de la librería.
- Una leve comprobación semántica, realizada a partir de valores por defecto, y chequeos y conversiones de tipos nativos de Java.
- Mecanismos para validar expresiones mediante una herramienta autónoma

Implementación para el SEDICI

Módulo de expresiones

Expresiones posibles

- Expresiones de validación
  - *bundle.name == 'ORIGINAL'*
  - *!empty(item.handle)*
  - *item.getMetadata.stream().anyMatch(m->m.name == "dc.creator")*
- Expresiones de selección
  - *collection.items.stream().filter(i->empty(i.getMetadata("dc.date.issued"))*
  - `item.getMetadata().stream().filter(m->m.language == 'ES')`
  - `item.bundles.stream().forEach(b->b.bitstreams.stream().filter(bits->bits.getChecksumAlgorithm()!='MD5'))`

Implementación para el SEDICI

Ejecución

Tareas de curación

Se definieron dos tareas de curación que, en la actualidad, son ejecutadas por los administradores del Repositorio Institucional de la UNLP.

- *nonEmptyCollectionTask*
  - Se detectaron colecciones vacías en distintas comunidades del repositorio
- *itemHasBitstream*
  - Se detectaron ítems que no tenían ningún archivo asignado, que se continúan corrigiendo en la actualidad

Conclusiones

Distintos niveles de abstracción

- Especificación de la Herramienta
  - La sintaxis abstracta lograda y el metamodelo obtenido permiten cierto grado de ortogonalidad
- Módulo de Expresiones
  - El módulo permite ejecutar validaciones y selecciones sobre cualquier elemento del repositorio, particularmente mediante tareas de curación que evalúan los resultados de las expresiones.
- Caso de Uso: Tareas de Curación
  - Son ejecutadas desde una consola de administración y pueden personalizarse con mayor facilidad gracias a la incorporación de las expresiones para el módulo

Trabajos Futuros

Estudio de otras alternativas de implementación

ANTES

```
collection.items().stream().filter(i->!empty(i.getMetadata('dc.type')))
```

DESPUÉS

```
SELECTION (ITEMS from Collections) WITH VALIDATION :Item.hasMetadatada(dc.type)
```

Trabajos Futuros

Mejoras al módulo de expresiones

- Objetos Genéricos
  - `#handle(10915/30522).getMetadata(dc.type)`
  - `#DOI(10.1000/182).getType == ‘item’`
- Transformaciones
  - `item.setMetadatada(dc.type = ‘Article’)`
  - `bundle.createBitstream(stream)`

Trabajos Futuros

Casos de uso nuevos

- Mejorar las tareas de curación
- Módulo de exportación basado en expresiones
- Ejecución por lotes
- Utilización en el data provider

¡Muchas Gracias!

¿PREGUNTAS?

Franco Agustín Terruzzi

agustinterruzzi@sedici.unlp.edu.ar