

Universidad Nacional de La Plata

Facultad de Informática



Herramientas para la Interoperabilidad y Normalización de datos en RI

Tesina de Licenciatura en Sistemas

Autor: Almazán María Belén

Director: Ing. De Giusti, Marisa

La Plata, Octubre de 2012

AGRADECIMIENTOS

Agradezco a todas aquellas personas que de una manera u otra hicieron posible la realización de esta Tesina, especialmente:

A la Universidad Nacional de La Plata, pública y gratuita, y en particular a la Facultad de Informática, por permitir desarrollarme profesional y personalmente durante estos años.

A la Ing. Marisa De Giusti, directora de esta Tesina, por proporcionarme las condiciones de trabajo necesarias para su adecuado desarrollo y su constante confianza y predisposición para conmigo.

A Néstor Oviedo por su incesante, desinteresada e invaluable ayuda.

Al Lic. Ariel J. Lira, por guiarme en la realización de esta Tesina, por estar siempre dispuesto a responder dudas y proporcionar ideas.

Al Lic. Gonzalo L. Villarreal por prestarse en cualquier momento a colaborar con lo que fuese necesario.

A mi novio que lo amo y es quien está presente en todos los aspectos de mi vida, acompañándome y alentándome en todo momento.

A mis hermanos que los quiero mucho.

A mi familia por acompañarme y celebrar mis logros siempre con alegría.

A mis amigas por entenderme y apoyarme más allá de no poder estar siempre presente.

Y especialmente a mis padres por darme el apoyo necesario en todo momento y estar siempre presentes.

ÍNDICE

Capítulo 1 - Introducción

1.1 Motivación.....	13
1.2 Objetivo.....	15
1.3 Contexto.....	16
1.3.1 Concepto Y Tipología de los Documentos.....	16
1.3.2 Repositorios Institucionales.....	19
1.3.3 Acceso Abierto.....	19
1.4 Estructura de la tesina.....	22
1.5 Bibliografía.....	23

Capítulo 2 – Representación de la Información

2.1 Introducción.....	29
2.2 Lenguaje Natural.....	29
2.3 Lenguaje Documental.....	30
2.3.1 Objetivos y Funciones del Lenguaje Documental.....	31
2.3.2 Tipos de Lenguajes Documentales.....	32
2.3.2.1 Según el Control.....	32
2.3.2.1.1 Lenguajes Libres.....	32
2.3.2.1.2 Lenguajes Controlados.....	33
2.3.2.2 Según la Coordinación.....	33
2.3.2.2.1 Lenguajes Precoordinados.....	33
2.3.2.2.2 Lenguajes Postcoordinados.....	34
2.3.2.2.3 Comparación: Lenguajes Precoordinados vs. Lenguajes Postcoordinados.....	34
2.3.2.3 Según la Estructura.....	35
2.3.2.3.1 Estructura Jerárquica.....	36
2.3.2.3.1.1 Clasificaciones enciclopédicas.....	36
2.3.2.3.1.2 Clasificaciones especializadas.....	37
2.3.2.3.1.3 Clasificaciones de facetas.....	37
2.3.2.3.2 Estructura Combinatoria.....	37
2.3.2.3.3 Estructura Sintáctica.....	37
2.3.3 Lenguajes documentales: definición y características.....	38
2.3.3.1 Listas de palabras clave.....	38
2.3.3.2 Listas de Descriptores Libres.....	38
2.3.3.3 Sistemas de Clasificación.....	38
2.3.3.4 Listas de Encabezamiento de Materias.....	39
2.3.3.5 Tesoros y Descriptores.....	40
2.3.3.6 Ontologías.....	40
2.3.3.7 Diferencias entre los lenguajes documentales	41
2.4 Bibliografía.....	42

Capítulo 3 - Metadatos

3.1 Introducción.....	47
3.2 De la Información a la Metainformación.....	47

3.2.1 Definición y Aplicaciones.....	47
3.2.2 Clasificación.....	49
3.3 Modelos de Metadatos.....	50
3.3.1 Estructura de los Metadatos.....	51
3.3.2 Evolución de los Metadatos.....	51
3.3.3 Protocolos para la Recuperación de Información.....	52
3.3.3.1 Machine Readable Cataloguing (MARC)	52
3.3.3.2 Z39.50.....	53
3.3.4 Lenguajes de Metadatos.....	54
3.3.4.1 Standard Generalized Markup Language (SGML).....	55
3.3.4.2 HyperText Markup Language (HTML)	56
3.3.4.3 XML.....	56
3.3.4.3.1 Definición.....	56
3.3.4.3.2 Objetivos.....	57
3.3.4.3.3 Funciones.....	57
3.3.4.3.4 Ventajas.....	57
3.3.4.4 RDF.....	57
3.3.4.4.1 Definición.....	57
3.3.4.4.2 Aplicaciones.....	59
3.3.4.5 Comparación entre XML y RDF.....	59
3.3.5 Conjunto de Metadatos “Dublin Core”	60
3.4 Bibliografía.....	61

Capítulo 4 - Interoperabilidad

4.1 Introducción.....	67
4.2 Definición.....	67
4.3 Tipos de Interoperabilidad.....	68
4.3.1 Interoperabilidad Técnica.....	69
4.3.2 Interoperabilidad Semántica.....	70
4.3.3 Interoperabilidad Organizativa/Pragmática.....	71
4.3.4 Relación entre los Tipos de Interoperabilidad.....	71
4.4 Estándares de Interoperabilidad.....	72
4.4.1 Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH).....	72
4.4.2 SWORD.....	74
4.4.3 OpenSearch.....	75
4.5 Herramientas de Interoperabilidad	75
4.5.1 Plataformas de Software para Crear Repositorios.....	76
4.5.2 Aplicaciones Generadoras de Proveedores de Datos.....	77
4.5.2.1 Virginia Tech OAI (VTOAI)	78
4.5.2.2 Open Archives In a Box (OAIB)	78
4.5.2.3 OAIcat.....	79
4.5.2.4 OAIbiblio.....	79
4.5.2.5 XMLFile.....	80
4.5.2.6 Rapid Visual OAI Tool (RVOT)	80
4.5.2.7 VOAI.....	80
4.5.2.8 PHP OAI Data Provider.....	81
4.5.2.9 Comparación de Herramientas.....	81
4.5.3 Recolectores y Proveedores de Servicios.....	82
4.5.3.1 Arc.....	83
4.5.3.2 DP9.....	83

4.5.3.3 PKP Open Archives Harvester.....	83
4.5.3.4 Perl Harvester.....	84
4.5.3.5 Net::OAI::Harvester.....	84
4.5.3.6 OAIHarvester2.....	84
4.5.3.7 Iniciativa DRIVER.....	85
4.5.4 Aplicaciones que Incorporan OpenSearch.....	86
4.5.5 Implementaciones de SWORD.....	86
4.6 Bibliografía.....	87

Capítulo 5 – Servicio de Difusión de la Creación Intelectual (SeDiCI)

5.1 Introducción.....	91
5.2 Contexto.....	91
5.3 Harvesting.....	93
5.3.1 Modelo de Datos.....	95
5.3.2 Transformación.....	95
5.4 Bibliografía.....	97

Capítulo 6 – Normalización y Calidad de Datos

6.1 Introducción.....	101
6.2 Normalización.....	101
6.3 Recursos Útiles para Normalizar Nombres de Autor.....	102
6.3.1 IralIS.....	102
6.3.2 SCOPUS.....	103
6.4 Herramientas para la Normalización de Datos.....	104
6.4.1 Google Refine.....	104
6.4.2 Prism Warehouse Manager.....	105
6.4.3 Passport.....	105
6.4.4 Data Reengineering Tool.....	105
6.4.5 Enterprise/Integrator.....	105
6.5 Actividades Para Mejorar la Calidad de los Datos.....	106
6.6 Control de Calidad de la Información.....	107
6.6.1 Detección y Corrección de Inconsistencias.....	107
6.6.2 Detección y Corrección de Datos Incompletos.....	107
6.6.3 Detección y Corrección de Anomalías.....	107
6.7 Enriquecimiento de la Información.....	108
6.8 Técnicas Para Mejorar la Calidad de los Datos en RI.....	108
6.8.1 Metadato Autor.....	108
6.8.2 Metadato Título.....	109
6.8.3 Metadato Lenguaje.....	109
6.8.4 Metadato Fecha.....	110
6.8.5 Otros Metadatos.....	112
6.9 Bibliografía.....	112

Capítulo 7 - Desarrollo

7.1 Introducción.....	117
7.2 Problema Específico.....	117
7.3 Mejora Propuesta.....	117
7.4 Metodología.....	118
7.5 Estandarización de Formato del Texto.....	122

7.5.1 Filtro: UpperCaseField.....	122
7.5.2 Filtro: LowerCaseField.....	122
7.5.3 Filtro: TitleCaseField.....	122
7.6 Detección de Lenguaje del Texto.....	123
7.6.1 Filtro: LanguageDetector.....	123
7.7 Normalización de Nombre de Autor.....	124
7.7.1 Filtro: SurnameDetector.....	124
7.7.2 Filtro: AuthorNormalizer.....	124
7.7.3 Filtro: ScopusAuthorNormalizer.....	125
7.8 Depuración y Normalización de la Fecha.....	126
7.8.1 Filtro: DateCleaner.....	126
7.8.2 Filtro: PeriodDetector.....	126
7.8.3 Filtro: DateNormalizer.....	126
7.9 Pruebas Realizadas.....	127
7.10 Aporte.....	132
7.11 Bibliografía	133

Capítulo 8 – Conclusiones y Posibles Trabajos Futuros

8.1 Introducción.....	137
8.2 Conclusiones.....	137
8.3 Trabajos Futuros.....	138

Capítulo 1

Introducción

Capítulo 1 - INTRODUCCIÓN

El presente trabajo de grado se ubica en el área de las bases de datos, más precisamente en el de las bases de datos documentales, haciendo énfasis en la normalización e interoperabilidad de los datos que residen en las mismas de modo de facilitar los procesos de recuperación de información y catalogación de los documentos.

El objetivo de esta tesina, por tanto, consiste en investigar e implementar métodos que mejoren la calidad de los metadatos provenientes de bases de datos documentales de repositorios institucionales y temáticos, a través de procesos de depuración, asociación, inferencia y normalización, con el fin que se puedan optimizar las técnicas de recuperación de la información e interoperabilidad. Asimismo, se pretende abrir el paso al desarrollo de nuevas técnicas de recuperación basadas en criterios semánticos.

1.1 Motivación

En los últimos años, las bibliotecas digitales han tenido una gran evolución histórica, teórica y práctica tanto en el ámbito tecnológico como en el social, y en la actualidad es uno de los principales temas de estudio, investigación y desarrollo en todo el mundo bibliotecario.

En (1.1) Borgman intenta explicar el significado y la interpretación que se le da a la expresión *biblioteca digital* a través del análisis de diversas definiciones del término. El autor identifica dos sentidos diferentes en su utilización. La definición tecnológica¹ que sostiene que las bibliotecas digitales son un conjunto de recursos electrónicos y capacidades técnicas asociadas para crear, buscar y utilizar la información; se complementa con el punto de vista social² que afirma que las bibliotecas digitales son construidas, recolectadas y organizadas por (y para) una comunidad de usuarios, y sus capacidades funcionales soportan las necesidades y usos de la información de dicha comunidad.

Se puede decir entonces que una biblioteca digital consiste en una serie de recursos digitales, en la cual se facilita el empleo y la recuperación de la información respecto a las bibliotecas y archivos convencionales. Además, cabe destacar el enfoque social del término, el cual apunta a satisfacer las necesidades de una comunidad de usuarios.

Aunque existen numerosas áreas de investigación, se pueden encontrar tres grandes temas dentro de la comunidad de bibliotecas digitales: repositorios digitales, museos digitales y manejo de activos digitales (1.2). En el marco del presente trabajo se profundizará sobre el primero de ellos, es decir, los repositorios digitales, más precisamente sobre los que poseen carácter académico.

El Proyecto de Apoyo a Repositorios (1.3) define a un repositorio digital de la siguiente manera: "Un repositorio digital es un mecanismo para administrar y almacenar contenido

¹ "digital libraries are a set of electronic resources and associated technical capabilities for creating, searching and using information"

² "digital libraries are constructed, collected and organized, by (and for) a community of users, and their functional capabilities support the information needs and uses of that community"

digital. Los repositorios pueden ser temáticos o institucionales. El depósito de contenidos en un repositorio institucional permite a las instituciones gestionarlo y preservarlo y por lo tanto obtener un mayor valor de producción. Un repositorio puede apoyar la investigación, el aprendizaje y los procesos administrativos. Los repositorios utilizan estándares abiertos para garantizar que sus contenidos son accesibles y pueden ser buscados y recuperados para su uso posterior. El uso de estos estándares internacionales permiten mecanismos para importar, exportar, identificar, almacenar y recuperar el contenido digital dentro del repositorio”.

Dentro de un repositorio digital, los objetos son descritos mediante ciertas etiquetas conocidas como metadatos (id, autores, fechas, temas, etc.), que han sido utilizadas por las bibliotecas convencionales y que facilitan su recuperación. De manera concreta, los metadatos no son más que datos estructurados que dan cuenta de otros datos o información; en otras palabras, se trata de datos sobre datos.

Por otra parte, una tendencia vista a nivel mundial es la adopción por parte de los repositorios digitales de las políticas de acceso abierto (*Figura 1.1*).

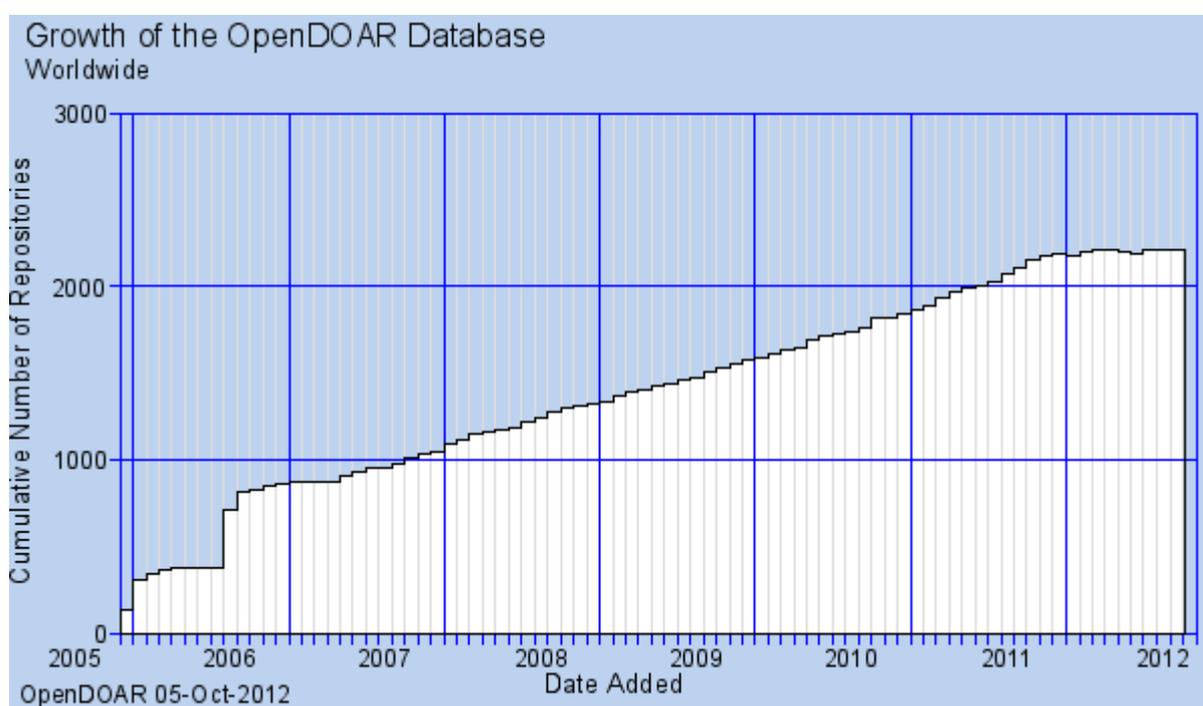


Figura 1.1: Crecimiento de los repositorios institucionales en el mundo (Fuente: OpenDOAR³)

Además, dado que la mayoría de estos repositorios buscan interactuar entre sí para intercambiar información, suelen implementar protocolos estándar de interoperabilidad, permitiendo así la comunicación exitosa y el intercambio de información entre aplicaciones sin que las diferencias en tecnología o plataforma de base sean un obstáculo. OAI-PMH es uno de los protocolos de interoperabilidad más difundidos a nivel mundial y se tratará en detalle más adelante en este trabajo.

Existe un tipo de repositorio digital que constituye una fuente de datos para efectuar estudios vinculados a la producción científica de una institución, y corresponde a los llamados

³ OpenDOAR <<http://www.opendoar.org>>. Consulta: [2012-10-05]

repositorios institucionales. Lynch en (1.6) define el término como una serie de recursos y servicios creados y administrados con el fin de que la comunidad académica pueda manejar y difundir materiales digitales. Su objetivo es capturar y administrar la producción intelectual de una o varias comunidades universitarias y maximizar la visibilidad e impacto en línea. De acuerdo a Heery (1.7) los repositorios deben ofrecer un mecanismo para depositar material por parte del creador, el dueño u otra persona (por ejemplo, un bibliotecario), contar con una arquitectura que maneje tanto el contenido como los metadatos, servicios básicos tales como búsqueda y recuperación, administración, controles de acceso y permisos y ser sustentable a largo plazo es decir, ser administrado y apoyado por una organización confiable.

En resumen, se puede decir que los repositorios institucionales son colecciones digitales que contienen la producción intelectual de cierta comunidad académica y que cuentan con mecanismos de depósito y recuperación de información. Además de ser abiertos e interactivos, se los considera acumulativos y perpetuos. De forma excepcional, pueden contener documentos con acceso restringido o embargo temporal.

Dada la relevancia de los datos, los procesos de calidad de la información adquieren un papel muy importante. Es necesario notar que el sólo registro de los metadatos o descriptores asociados a cada publicación no es suficiente para automatizar la determinación y generación de indicadores de producción científica, sino que para poder explotar la información es necesario que los datos sean almacenados y reconocidos unívocamente, determinando que los procesos técnicos vinculados a la catalogación del material deban ser analizados exhaustivamente, fijando modalidades de trabajo y normas de carga.

Debido a la gran cantidad de información proveniente de diferentes orígenes (catalogación externa, importaciones, cosechas automáticas, entre otras), se dificulta la definición de procesos de normalización de datos y metadatos, y con ello se ven obstaculizadas las tareas automáticas y acciones de recuperación de la información por parte de los usuarios.

1.2 Objetivo

A partir de este trabajo se pretende mejorar la calidad de los datos provenientes de bases de datos documentales de repositorios institucionales y temáticos, a través de procesos de depuración, asociación, inferencia y normalización, a fin de optimizar las técnicas de recuperación de la información e interoperabilidad y obtener como resultado una base de datos más consistente, confiable y concreta. Asimismo, se procura abrir el paso al desarrollo de nuevas técnicas de recuperación basadas en criterios semánticos.

El estudio de caso e implementación serán los contenidos de la base Solr de recursos recolectados bajo OAI pertenecientes al Servicio de Difusión de la Creación Intelectual (SeDiCI), el repositorio institucional central de la UNLP. El marco general (teórico-práctico) será el procesamiento de texto.

1.3 Contexto

1.3.1 Concepto Y Tipología de los Documentos

Para empezar, se analizará el significado de “documento”, debido a que se trata de la unidad mínima que integra una base documental, la cual será el campo de acción del presente trabajo. En términos generales, se habla de documento para referirse a cualquier unidad significativa de información que haya sido registrada en un soporte que permita su almacenamiento y su posterior recuperación. Dicho soporte debe ser accesible y permitir la reutilización del contenido que guarda.

De esta definición, podemos deducir entonces, que si la información se encontrase registrada en un soporte electrónico, también constituiría un documento. En este caso estaríamos hablando de un *documento digital*. Es una realidad ineludible que este tipo de documentos tiene una menor estabilidad que los documentos anteriores. La permanencia del soporte, cualidad tradicional del documento "fuente permanente de información permanente" según Desantes (1.8), es cuestionada en el documento digital, y sobre todo en el que se transmite vía web.

Según Martín Vega (1.9) todo documento incluye tres tipos de componentes:

- 1 Componentes físicos o materiales: Todo documento se asienta en una determinada clase de soporte material. Ofrece, por consiguiente, cualidades de peso, tamaño, sustancia material, etc.
- 2 Componentes formales: Los documentos adquieren una estructura. La materia básica que los sustenta se dispone de una cierta manera (grabando sobre piedra, dando forma gráfica con la tinta, digitalizando datos en un ordenador, etc.), con objeto de mostrar un contenido, de adquirir un sentido, de tener un significado, de transmitir un conocimiento.
- 3 Componentes conceptuales: Todos los documentos proporcionan un significado.

En términos generales, los componentes físicos o materiales serían el soporte, los formales el medio de fijación del mensaje al soporte y, por último, los componentes conceptuales, el contenido.

Podemos decir entonces, que el documento es a la vez medio y mensaje de información y conocimiento. De esta manera, de acuerdo a lo expuesto por López Yepes (1.10) el documento se caracteriza por una triple dimensión: el soporte físico o material, el mensaje informativo y la posibilidad de transmisión o difusión de este conocimiento. Esta triple dimensión que caracteriza al documento ha servido, a su vez, para establecer una tipología de los documentos en la que la mayoría de los autores coinciden. Se puede citar por ejemplo, el criterio de clasificación del autor López Yepes como sigue:

- 1 Por la forma de representación del mensaje en el soporte físico:
 - a Gráfico: libro, revista, etc.
 - b Iconográfico: fotografía, pintura, etc.
 - c Fónico: disco, cinta magnetofónica, etc.

- d Audiovisual: película, vídeo, etc.
 - e Plástico: objetos.
 - f Electrónico: disquete, disco óptico digital, etc.
- 2 Por el nivel de difusión:
 - a Publicado: cualquier documento multiplicado en número suficiente de ejemplares que permiten su difusión pública.
 - b Inédito: manuscrito o documento de archivo no publicado.
 - c Reservado: documento manuscrito o impreso pero no difundido.
 - 3 Por el grado de originalidad en su creación:
 - a Fuentes: los documentos más cercanos a la información o a los acontecimientos que reflejan o que constituyen la materia prima. Ej.: documentos de época, crónicas, estadísticas, legislación, objetos de museo, etc.
 - b Bibliografía: los documentos elaborados desde las fuentes. Ej.: monografía, artículo de revista, etc.
 - 4 Por el grado de modificación de la naturaleza del mensaje, resultado del análisis documental:
 - a Primario: son aquellos que contienen la información original de los autores, es decir, el texto completo. Ej.: libro, artículo de revista, etc. Dentro de esta tipología de documentos es preciso destacar lo que en el mundo bibliotecario se denomina "literatura gris", término que describe el material publicado de forma informal y que no se puede conseguir por medios comerciales (Ej.: informes internos). Este tipo de material a menudo es muy difícil de localizar (1.11)
 - b Secundario: contienen los datos y la información referente a los documentos originales, pero no se accede directamente al documento. Sirven para saber qué documentos primarios hay sobre una materia, escritos por un determinado autor, etc. Ej.: bibliografías, catálogos de bibliotecas, bases de datos documentales, resúmenes, etc.
 - 5 Por el grado de transformación del mensaje documentario soportado en el documento:
 - a Mensaje documentado.
 - b Mensaje marginal.
 - c Mensaje referencial.
 - d Mensaje documental.
 - 6 Por su situación en el sistema de las ciencias:
 - a Jurídico, matemático, médico, etc.
 - 7 Por el grado de comprobación de la verdad del mensaje:
 - a Científico: monografía científica, tesis doctoral, etc.
 - b No científico: artículo de prensa, ensayo, etc.

Este autor evita establecer una sistematización en función del soporte, seguramente por la poca estabilidad que estas clasificaciones tienen. Esto puede explicarse a través de la frase de Desantes (1.12) "la teoría de los soportes en la documentación va a constituir siempre

una síntesis incompleta, susceptible permanentemente de ampliación en todas las direcciones”.

En contraparte, el autor Martínez Comeche (1.13) sí distingue entre el criterio del soporte y del medio empleado para fijar el mensaje a ese soporte. Establece diversas clasificaciones, la mayoría coincidentes con las de López Yepes y las cuales son resumidas por Martín Gavilán (1.14), e inciden en:

- 1 La naturaleza del soporte: Desde la perspectiva del soporte empleado, los documentos utilizan hoy día mayoritariamente el papel (libros, artículos, folletos, etc.), los materiales magnéticos (cintas magnetofónicas y de vídeo, disquetes, etc.) y los soportes ópticos (videodiscos, CD-ROM, discos compactos, etc.), aunque a lo largo de la historia han variado rigurosamente (madera, piel, pergamino, piedra, metal, etc.).
- 2 El código empleado en el mensaje: Esta clasificación distingue entre documentos textuales (cuando los signos corresponden a la lengua escrita), gráficos (mapas, planos), iconográficos (cuadros, diapositivas, fotografías), sonoros o fónicos (cintas, discos), audiovisuales (películas, vídeos), plásticos o tridimensionales (cualquier objeto conservado en un museo, por ejemplo), informáticos (legibles por ordenador), o documentos multimedia (cuando el documento combina varios de los códigos anteriormente expuestos).
- 3 El rigor científico del mensaje: Los documentos suelen dividirse en científicos, técnicos y de divulgación, según va disminuyendo el nivel de profundidad y precisión del mensaje emitido.
- 4 El área del conocimiento que abarca el mensaje: Aquí se sitúan los documentos económicos, sociales, históricos, jurídicos o lingüísticos, entre una enorme variedad de mensajes posibles.
- 5 El tratamiento y consiguiente modificación del mensaje original: La estructuración del documento considera el tratamiento a que es sometido el contenido el mensaje, modificando consiguientemente su presentación. Establece una diferencia entre un documento primario, un documento secundario y un documento terciario. No existe acuerdo entre los estudiosos sobre el documento terciario.
- 6 La capacidad de difusión: el documento puede ser portador de un mensaje publicado o público, inédito, reservado o personal.

Se puede concluir entonces, que si bien el concepto de documento puede ser abordado desde distintos puntos de vista, su definición debe ser lo más amplia posible, ya que tiene que integrar una gran variedad de soportes, formatos y distintas morfologías. La tipología de los documentos también se va extendiendo a medida que surgen nuevas formas y tecnologías de lectura y escritura, nuevas formas de acceso y recuperación de documentos, nuevas formas de estructurar la información y nuevos modos de interacción por parte del usuario.

1.3.2 Repositorios Institucionales

Como se dijo anteriormente, un Repositorio Institucional (RI) no es nada menos que un repositorio digital de carácter académico, es decir, un archivo electrónico de la producción científica de una institución, almacenada en un formato digital, en el que se permite la búsqueda y la recuperación para su posterior uso nacional o internacional.

Bustos y González (1.15) lo definen como “un auténtico sistema de gestión de contenidos ya que, además de los documentos propiamente dichos, el repositorio ofrece a la comunidad académica un conjunto de servicios para la gestión de esa producción. El RI es una vía de comunicación científica, pero no puede ser entendido como un canal de publicación, sino que debe comprenderse como un complemento al proceso de publicación científica formalizado con revisión por pares”.

Las colecciones intelectuales que integran un repositorio varían según cada institución, y pueden incluir tanto la producción científica (artículos, tesis, comunicaciones, software, etc.), los objetos para la enseñanza (guías de estudio, simuladores, apuntes, bibliografía, guías, etc.), así como también los documentos administrativos y/o aquellos documentos que ella misma genera (reglamentos, normas, informes técnicos, etc.), presentándose de diversas formas como ser textos, registros audiovisuales, etc. Generalmente, se requiere que las publicaciones sean abiertas.

En el ámbito docente, el RI facilita el cambio de paradigma en la enseñanza y el aprendizaje, aportando un entorno pedagógico rico en información.

En resumen, las principales características de un repositorio institucional son:

- Su naturaleza institucional
- Que alberga los documentos de manera acumulativa y perpetua.
- Que gestiona una colección organizada: un repositorio no es un mero depósito de documentos, sino que los mismos son descritos utilizando un número suficiente de metadatos mínimamente normalizados, organizados mediante la aplicación de alguna clasificación de contenidos.
- Su carácter científico: posee documentos creados por la institución o alguno de sus miembros como producto de las funciones de investigación que le son propias.
- Su carácter abierto: con el propósito de aumentar la visibilidad e impacto de la investigación que se realiza en una institución.
- Su carácter interoperable con otros sistemas: el repositorio no es un fin en sí mismo sino que su verdadero potencial se descubre cuando sus contenidos se integran en un nivel superior de agregación desde donde se puedan prestar servicios especializados a comunidades concretas.

1.3.3 Acceso Abierto

En los últimos años, el movimiento de Acceso Abierto (Open Access - OA) ha tomado mucha fuerza entre las instituciones académicas y científicas. El objetivo base de este movimiento es mejorar la comunicación científica y eliminar todo tipo de barreras que impidan el acceso a la información, maximizando el acceso a la misma por medio de la creación de repositorios de acceso abierto, accesibles a través de internet.

Existe un compromiso social avalado por declaraciones de ámbito internacional que sostienen y perfilan la definición de Open Access. Las tres más importantes son la Declaración de Budapest (Budapest Open Access Initiative, BOAI) de 2002, seguida de la Declaración de Bethesda (2003) y la Declaración de Berlín (Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities) también del año 2003.

La Budapest Open Access Initiative (1.16) define de la siguiente manera el acceso abierto: “disponibilidad gratuita en la Internet pública, para que cualquier usuario la pueda leer, descargar, copiar, distribuir y/o imprimir, con la posibilidad de buscar o enlazar todos los textos de estos artículos, recorrerlos para indexación exhaustiva, usarlos como datos para software, o utilizarlos para cualquier otro propósito legal, sin barreras financieras, legales o técnicas, distintas de la fundamental de ganar acceso a la propia Internet”. En la Declaración de Bethesda (1.17), además se menciona el archivo inmediato de los trabajos para facilitar este acceso abierto. La Declaración de Berlín (1.18), fue suscrita por diferentes representantes políticos y científicos y en ella, explícitamente se manifiestan las grandes posibilidades que brinda internet en la difusión del conocimiento, avala el paradigma de acceso abierto y recoge los términos de las dos declaraciones anteriores.

El libre acceso proporciona ventajas que se pueden concretar en su aspecto económico, científico y de servicios de valor añadido para el autor:

- Económicas:
 - La investigación es igualmente accesible a todos los científicos.
 - Su creación tiene bajo costo y resultados rápidamente visibles
- Visibilidad máxima – impacto máximo:
 - Alta posibilidad de ser visto, leído y citado
 - Crean recursos de calidad (metadatos) que aseguran la recuperación eficaz y eficiente de la información.
- Rapidez:
 - Sistema en línea para entrega, arbitraje y publicación.
 - Más rápido y directo acceso a los resultados de la investigación.
- Servicios de valor añadido:
 - Consulta y navegación.
 - Acceso al texto completo.
 - Servicio de Alerta en línea.
 - Estadísticas de consultas y descargas.
 - Elaboración del Currículum Vitae.

Se puede ver en la *Figura 1.2* que existe un tipo de vía de publicación denominada ruta roja, la cual comprende las revistas a las que sólo se accede por un medio pago. Funcionan publicando trabajos tras la revisión y aceptación de un pre-print o manuscrito enviado por el/los autores (1.27). Si la revista retiene los derechos de explotación de manera exclusiva (el autor los cede por completo):

- No permiten el autoarchivo, es decir, la acción por parte del autor de depositar un artículo o cualquier otra obra en estos repositorios
- Se debe pedir permiso para cualquier acción con el trabajo.

En la Declaración de Budapest (1.16) se establecen dos rutas para alcanzar el acceso abierto: la *ruta dorada* o la de publicación en revistas de acceso abierto y la *ruta verde* que alude al archivo o depósito de recursos digitales en repositorios institucionales o temáticos.

Además de que la revista retenga los derechos del autor, este último tiene otras posibilidades:

- Puede no ceder de manera exclusiva sus derechos, lo que le habilita la denominada vía verde.
- Publicar en una revista que ofrece sus contenidos (o algunos de ellos) gratis para los lectores desde alguna base, lo que le posibilita otro tipo de vía conocida como la *ruta dorada*.



Figura 1.2. Fuente: fundación Max Planck⁴

En la denominada “ruta o vía verde”, es el autor el responsable de poner a disposición los artículos en un repositorio de forma voluntaria. Para esto, debe conocer el contrato de cesión de derechos que estableció con la editorial, lo que le permitirá saber si, por ejemplo, retiene el permiso para autoarchivar como asimismo el resto de los derechos que puede mantener. A continuación se mencionan algunas de las ventajas de utilizar esta vía:

- El material es acumulativo, se ofrece un punto de acceso uniforme a la información de la institución y del autor.
- El material es catalogado y descrito en profundidad con metadatos, permitiendo mejores formas de acceso a la información.
- Se preserva la integridad de las obras
- Trabaja bajo estándares de intercambio que permiten la exposición automática de los registros, y por tanto, maximiza la difusión de las obras hacia sitios como

⁴ Max Planck <<http://www.mpg.de/de>>. Consulta: [2012-10-05]

OAISTER/OCLC, RECOLECTA, Scientific Commons, BASE, NDLTD (para tesis), entre otros.

- El material se distribuye junto al detalle de sus derechos de uso, notificando a los lectores los usos permitidos.

Por otra parte, se le denomina “ruta dorada” al hecho de que una editorial haga que una publicación sea libremente accesible de forma OA (acceso abierto), o bien porque la revista se expone completamente bajo OA, o bien a partir de un pago por parte del autor o de la institución a la cual está vinculado.

El circuito de la ruta dorada es complejo, por lo que en los últimos años han crecido considerablemente las publicaciones en repositorios. Se recomienda a los autores utilizar la ruta verde, debido que al depositar una copia de la versión presentada al editor en un repositorio institucional/temático tan pronto como sale a la luz se garantiza que la obra gane rápidamente visibilidad y se mantenga permanentemente accesible.

Hoy en día, el Acceso Abierto cada vez más está siendo utilizado en repositorios para recolectar, archivar y difundir/diseminar la producción científica de una institución, tales como artículos de investigación, actas de conferencias, disertaciones, recopilados de datos, documentos de trabajo e informes.

1.4 Estructura de la Tesina

Aquí se presentará cómo está organizada la tesina capítulo por capítulo.

Capítulo 1: Se presentan los componentes generales de la tesis, introduciendo el contexto sobre el cual se desarrolló. Se plantea el tema central del trabajo con sus objetivos y alcance, el método general y la organización de sus partes.

Capítulo 2: Se describen los aspectos relacionados con la representación de la información en RI. Se analiza al lenguaje como componente fundamental, considerando características, funciones y tipos de lenguajes documentales como elementos normalizadores en los procesos de tratamiento y recuperación de información.

Capítulo 3: Se introducen diferentes definiciones del término metadato, sus aplicaciones y clasificación según el ámbito de aplicación. Partiendo del concepto general de marcado de documentos, se describen los modelos de metadatos más representativos y utilizados, distinguiendo modelos de propósito general, y modelos de propósito específico.

Capítulo 4: Se define el concepto de interoperabilidad y los distintos tipos que existen. Se introduce a OAI como iniciativa para desarrollar y promover estándares de interoperabilidad y su implementación a través del protocolo OAI-PMH. Se describen brevemente sus diferentes implementaciones, tanto las que sirven para proveer servicios, como las que permiten cumplir el rol de proveedores de datos. Se especifican además algunas herramientas de software existentes que ayudan a mejorar la interoperabilidad de los datos en RI.

Capítulo 5: Se presenta el ambiente SeDiCI y el conjunto de datos sobre los cuales se van a trabajar.

Capítulo 6: Se reflexiona sobre cómo debe afrontarse el problema de la normalización de la forma que adopta la información en la Web. Asimismo, se analizan diferentes herramientas para la normalización de bases de datos documentales. Se introducen actividades y técnicas posibles para mejorar la calidad de los datos.

Capítulo 7: Se define el problema principal en SeDiCI, y la metodología de trabajo, presentando las mejoras introducidas y los resultados de su aplicación.

Capítulo 8: Se realiza una conclusión en base al objetivo planteado y se detallan posibles trabajos futuros.

1.5 Bibliografía

(1.1) Borgman, C. L. (1999). What are digital libraries? Competing visions. *Information Processing & Management*, 35 (3), 227-243.

(1.2) ¿Qué es una Biblioteca Digital? Blog Biblioteca Nacional Arturo Prat, Chile, Julio 2007. Disponible en:
<http://bibliopress.wordpress.com/2007/07/27/%C2%BFque-es-una-biblioteca-digital/>
[Consulta: 2012-08-29]

(1.3) Repositories Support Project: What is a Repository? Disponible en:
<http://www.rsp.ac.uk/start/before-you-start/what-is-a-repository/> [Consulta: 2012-08-29]

(1.4) Van de Sompel, H. y C. Lagoze (2000). The Santa Fe Convention of the Open Archives Initiative. *D-Lib Magazine* 6(2).

(1.5) Crow, R. (2002). The Case for Institutional Repositories: A SPARC Position Paper. Washington DC, Scholarly Publishing and Academic Resources Coalition: 37.

(1.6) Lynch, C. (2003). Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. *ARL Bimonthly Report* 226.

(1.7) Heery, R. and S. Anderson (2005). Digital Repositories Review. UKOLN and AHDS: 33.

(1.8) Desantes Guanter, J. M. ^ª. Teoría y régimen jurídico de la documentación. Madrid: EUDEMA, 1987.

(1.9) Martín Vega, A. Fuentes de información general. Gijón: TREA. 1995.

(1.10) López Yepes, J. Teoría de la documentación. Pamplona: EUNSA, 1978.

(1.11) Universidad Politécnica de Cataluña, Tipología documental. Servicio de Bibliotecas y Documentación, Módulo 2

(1.12) Desantes Guanter, J. M. El valor jurídico de los novísimos soportes documentales. *Revista General de Información y Documentación*, 1992, vol. 2, n.o 1, p. 17-31.

- (1.13) Martínez Comeche, J. A. Teoría de la información documental y de las instituciones documentales. Madrid: Síntesis, 1995.
- (1.14) Martín Gavilán, C. Temas de Biblioteconomía: El documento y sus clases. Análisis documental. Indización y resumen., 2009. Disponible en: <http://hdl.handle.net/10760/14605> [Consulta: 2012-08-29]
- (1.15) Bustos Gonzalez, Atilio y Fernandez Porcel, Antonio. Guidelines for the Creation of Institutional Repositories at Universities and Higher Education Organisations. Valparaiso: ALFA Network Babel Library, 2007. Disponible en: http://works.bepress.com/ir_research/18 [Consulta: 2012-08-29]
- (1.16) Budapest Open Access Initiative (2002). Disponible en: <http://www.soros.org/openaccess/read.shtml> [Consulta: 2012-08-29]
- (1.17) Bethesda Statements on Open Access Publishing (2003). <http://www.earlham.edu/~peters/fos/bethesda.htm> [Consulta: 2012-08-29]
- (1.18) Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (2003). <http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.htm>
- (1.19) Bailey Charles W., Jr. (2005). Open Access Bibliography: Liberating Scholarly Literature with E-Prints and Open Access Journals. Washington, DC: Association of Research Libraries. Disponible en: <http://www.digital-scholarship.com/oab/oab.htm>
- (1.20) Wellcome Trust Position Statement in Support of Open and Unrestricted Access to Published Research (2003). Disponible en: <http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Open-access/Policy> [Consulta: 2012-08-29]
- (1.21) The Valparaíso Declaration for Improved Scientific Communication in the Electronic Medium (2004). Disponible en: <https://mx2.arl.org/Lists/SPARC-OAForum/Message/519.html> [Consulta: 2012-08-29]
- (1.22) IFLA Statement on Open Access to Scholarly Literature and Research Documentation (2004). Disponible en: <http://www.ifla.org/V/cdoc/open-access04.html> [Consulta: 2012-08-29]
- (1.23) Washington D.C. Principles for Free Access to Science (2004). Disponible en: <http://www.dcpinciples.org/statement.pdf>. [Consulta: 2012-08-29]
- (1.24) Max Planck Society : Open Access at the Max Planck Society. Disponible en: <http://oa.mpg.de/lang/en-uk/informationen-fur-autoren/open-access-publizieren/> [Consulta: 2012-08-29]
- (1.25) ISI Web of Knowledge. Disponible en: <http://isiwebofknowledge.com> [Consulta: 2012-08-29]
- (1.26) Harnad, Stevan: Maximizing university research impact through self-archiving. 2003. Disponible en: <http://users.ecs.soton.ac.uk/harnad/Temp/che.htm> [Consulta: 2012-08-29]
- (1.27) De Giusti, Marisa Raquel. "Vías de publicación y derechos de autor en la academia". Servicio de Difusión de la Creación Intelectual (SeDiCI). Semana Internacional del Acceso

Abierto 2011 (Argentina): 25 de octubre de 2011. Disponible en: <http://sedici.unlp.edu.ar/handle/10915/5565> [Consulta: 2012-08-29]

(1.28) De Giusti, Marisa Raquel. "El desafío de la difusión abierta de las obras y los derechos de autor en las instituciones académicas". Servicio de Difusión de la Creación Intelectual (SeDiCI). Semana Internacional del Acceso Abierto 2011 (Colombia): 18 de octubre de 2011. Disponible en: <http://sedici.unlp.edu.ar/handle/10915/5564> [Consulta: 2012-08-29]

(1.29) Guédon, Jean-Claude. "Open Access: a symptom and a promise?". Jacobs, N., (Ed) Open Access: Key Strategic, Technical and Economic Aspects. Oxford : Chandos, 2006. cap. 3

(1.30) Lesk, M. Practical digital libraries: Books, bytes and bucks. San Francisco, CA: Morgan Kaufmann. 1997.

(1.31) Les Carr, Alma Swan y Stevan Harnad. "Creación y Mantenimiento del conocimiento compartido: Contribución de la University of Southampton". El profesional de la información, v. 20, n. 1, enero-febrero 2011. Disponible en: <http://sedici.unlp.edu.ar/blog/wp-content/uploads/2011/07/La-contribucion-al-OA-de-la-SH.pdf> [Consulta: 2012-08-29]

Capítulo 2

Representación de la Información

Capítulo 2 - REPRESENTACIÓN DE LA INFORMACIÓN

2.1 Introducción

Los documentos que componen cualquier unidad significativa de información almacenada en soporte, deben ser analizados de manera tal de que sea posible su recuperación. Este proceso, denominado "Análisis documental", tiene por objeto dotar a cada documento de una serie de puntos de acceso, para permitir su posterior recuperación. Estos puntos de acceso pueden consistir en el nombre del autor de la obra, de su título o del tema que trate (en el siguiente capítulo se hará hincapié sobre este tema). Del análisis de los aspectos externos de un documento se podrían tomar, por ejemplo, los datos relativos a su título y autor, lo que constituiría la descripción bibliográfica. El tema, los conceptos que implícita o explícitamente están recogidos en una obra se extraen mediante el "Análisis de Contenido", cuyas operaciones, como la indización, dan al lector un conocimiento más o menos profundo de la información que contiene. La operación de indizar consiste, en efecto en el análisis e identificación de los conceptos del documento y la selección de aquellas nociones que representen con mayor fidelidad la información que contiene. Para normalizar la denominación de dichos conceptos se procede a su traducción a un lenguaje documental, lo cual facilita la recuperación, independientemente del analista que trate el documento.

2.2 Lenguaje Natural

Como se dijo anteriormente, los repositorios poseen documentos, lo que conlleva a la necesidad de describirlos y recuperarlos de alguna forma. Para ello, se recurrió al lenguaje natural.

El lenguaje natural es el lenguaje escrito y hablado por las personas, donde no existen esfuerzos para limitar o definir vocabularios, sintaxis, semánticas e interrelaciones entre los términos. La representación de información y la estructuración de una consulta para la recuperación se realizan sin basarse en un vocabulario controlado, es decir que se compone de términos no predefinidos que se van generando a partir de la realización de procesos de indización.

Este tipo de lenguajes es el que utiliza el discurso científico, técnico o literario. En comparación con el crecimiento de ideas, el desarrollo del lenguaje natural es lento, como consecuencia, es necesario utilizar la misma palabra para expresar dos o más ideas. Ello da lugar a la homonimia, y derivada de dichos accidentes surge la ambigüedad, que crea dificultades en la recuperación: se trata del ruido (en el caso de la homonimia) y silencio documentales (en la sinonimia). Sólo el logro de la deseable entropía puede evitar estos inconvenientes y facilitar una recuperación eficaz de la información.

En un principio se pensaba que el lenguaje natural iba a solventar muchos de los problemas existentes, pero con el aumento exponencial que la producción documental ha venido teniendo en los últimos años, se hizo necesario disponer de un lenguaje que no fuese ambiguo, generando nuevas necesidades, nuevos sistemas y técnicas de tratamiento documental. Los lenguajes documentales, que se describirán a continuación, surgen de la necesidad de eliminar dicha ambigüedad, sirviendo de puente entre la terminología de los autores y la de los usuarios. (2.2)

2.3 Lenguaje Documental

Tal como se describió previamente, los signos del lenguaje natural son las palabras, que representan nuestro conocimiento de la realidad. El lenguaje documental, como es evidente, no es natural, pero se sirve de sus palabras y, en ocasiones, las reemplaza por símbolos cargados de significado preciso e unívoco.

El lenguaje documental, o lenguaje de indización es definido por J. Rowley (2.3) como “una lista de términos o notaciones que pueden ser utilizados como punto de acceso en un índice”, y también como “un conjunto de términos (el vocabulario) y las técnicas para utilizar las relaciones entre ellos en un sistema para dar descripciones de índice”.

Según Blanca Gil (2.4), se puede considerar al lenguaje documental como “todo sistema artificial de signos normalizados, que facilitan la representación formalizada del contenido de los documentos para permitir la recuperación, manual o automática, de información solicitada por los usuarios”, agregando además que “ha de ser unívoco, (...) no puede permitirse la ambigüedad del lenguaje natural”. Asimismo, C. Guinchat y M. Menou (2.5) definen los lenguajes documentales como “lenguajes convencionales utilizados por una unidad de información para describir el contenido de los documentos, para almacenarlos y para recuperar la información que contienen.”.

Por su parte, Roberto Coll-Vinent (2.6) se refiere al lenguaje documental como “conjunto de términos convencionales que representan el contenido de un documento”. Yves Courier (2.7) los define como “lenguajes artificiales que permiten generar la representación formalizada de los documentos y de las consultas que interesan a un grupo de usuarios a fin de recuperar los documentos que responden a las consultas”. Amat Noguera (2.8) utiliza la misma expresión para referirse a “un conjunto de términos o procedimientos sintácticos convencionales utilizados para representar el contenido de un documento con el fin de repetir su recuperación. Se le denomina también lenguaje de indización”.

En síntesis, se puede afirmar que el lenguaje documental es un sistema de signos normalizados (lingüísticos, numéricos) que permiten la representación del contenido de los documentos (convirtiendo el lenguaje natural de uso diario en una serie de términos normalizados) y la organización de estas representaciones con el objetivo de facilitar la recuperación de la información.

Las primeras manifestaciones del lenguaje documental datan de finales del siglo XIX, cuando aparecieron las clasificaciones bibliográficas. Estas clasificaciones, inspiradas en la lógica y en los sistemas filosóficos del conocimiento, se fundan en el principio de precoordinación y son de carácter enciclopédico.

El concepto moderno se consolidó en el siglo XX cuando Cutter introdujo el lenguaje de encabezamientos de materia, basado en los principios de especificidad y de entrada directa, que señala el comienzo del desarrollo de lenguajes documentales especializados.

Estos lenguajes especializados, pues, nacieron como respuesta a una creciente especialización de los conocimientos que no podían ser representados mediante las materias contenidas en los lenguajes multidisciplinarios existentes hasta entonces. Como consecuencia, se crearon numerosas clasificaciones especializadas, así como múltiples tesauros sectoriales, con los que se podían organizar las colecciones a escala institucional, pero, al propio tiempo, provocaron “una confusión babélica a escala mundial” (2.10). Sin embargo, pareciera que actualmente existe una nueva tendencia hacia el enciclopedismo temático.

En las siguientes subsecciones se analizarán los componentes del lenguaje documental, así como sus objetivos y funciones en el proceso de tratamiento de la información.

2.3.1 Objetivos y Funciones del Lenguaje Documental

El lenguaje documental interviene en dos fases del proceso documental, en el momento de la descripción y en el de la recuperación de la información. El objetivo de dichas operaciones es el de facilitar la recuperación de la documentación reduciendo el esfuerzo y gasto de tiempo del usuario.

Este tipo de lenguaje, “sirve fundamentalmente para normalizar la indización”, como lo indica Rodríguez Luna (2.11), el cual es un proceso doble donde se indizan los documentos en la fase de entrada en el sistema y se indizan las consultas de los usuarios en la etapa de salida o recuperación. Adicionalmente, su capacidad de representar sin ambigüedad los mensajes contenidos en los documentos, permite cumplir otro objetivo fundamental, aparte del de la normalización, el de inducción, dado que lo provee al usuario de un instrumento de consulta que le guía a utilizar determinados términos para el concepto requerido, proporcionándole además otros que pueden también interesarle para su búsqueda.

De las dos fases mencionadas en el primer párrafo, la autora Blanca Gil (2.4) indica que se desprenden dos funciones: la descripción del contenido de los documentos y la recuperación de la información. En cuanto a la primera se puede decir que tras la lectura e identificación de los conceptos contenidos en los documentos, éstos se representan mediante un lenguaje documental que proporciona un vocabulario unívoco que permite traducir los conceptos en términos normalizados. El objetivo principal, sin embargo, se cumple en la segunda fase del proceso, el lenguaje documental suministra los conceptos de cada palabra una vez efectuado el análisis, es decir, proporciona instrumentos para efectuar búsquedas a distintos niveles de generalidad o especificidad. En este sentido se puede considerar al lenguaje documental como un metalenguaje o lenguaje intermediario ya que sirve de puente entre la información contenida en los documentos y la información solicitada por los usuarios. Se trata de reducir los términos que aporta el lenguaje documental, como precisa Blanca Gil (2.4): “El lenguaje documental reduce considerablemente el volumen de términos del lenguaje natural no tomando en consideración más que los sustantivos o los sintagmas nominales”.

Por su parte, el autor Van Slype (2.12), observa que los lenguajes de indización pueden intervenir en seis momentos diferentes del proceso de búsqueda:

- 1 Selección de los sistemas documentales que se van a interrogar.

- 2 Enunciado de los conceptos de la pregunta, en lenguaje natural.
- 3 Traducción a un lenguaje de indización.
- 4 Formulación de la ecuación.
- 5 Extensión asistida por el ordenador.
- 6 Apreciación final de la pertinencia.

Además de todas las funciones mencionadas, el lenguaje documental es de gran utilidad para la ordenación o archivo de documentos, resolviendo también problemas de multilingüismo.

2.3.2 Tipos de Lenguajes Documentales

La variada tipología del lenguaje documental lo convierte en un elemento de apoyo a disciplinas como la Biblioteconomía, Documentación, Bibliografía y Archivística, para cuyas necesidades de descripción ofrece posibilidades concretas. En relación con el análisis formal, el lenguaje documental completa el proceso técnico de catalogación dotando al soporte de la descripción de puntos de acceso temáticos.

Existen diversos criterios para establecer la tipología de los lenguajes documentales. Los más generalizados son aquellos que establecen una clasificación según el control ejercido sobre el vocabulario, la coordinación de los términos (el momento en que se combinan los elementos) y la estructura.

2.3.2.1 Según el Control

Dependiendo del control ejercido sobre el vocabulario, los lenguajes pueden organizarse en dos categorías: libres y controlados (*Tabla 2.1*).

Control	Libres	Listas de descriptores libres
	Controlados	Clasificaciones, tesauros, ...

Tabla 2.1: Tipología de los lenguajes documentales en función del control ejercido sobre su vocabulario.

Existe abundante literatura acerca de las ventajas y desventajas que implica el uso del lenguaje libre y del controlado. Del análisis comparativo de uno y otro se puede concluir que el lenguaje controlado neutraliza las deficiencias del lenguaje libre y viceversa, por ello muchas bases de datos combinan la utilización de ambos en las distintas fases del tratamiento documental.

2.3.2.1.1 Lenguajes Libres

Se componen de un vocabulario no predefinido que se va generando a partir de la realización de procesos de indización. Son lenguajes cuya entrada (temas) están tomados del lenguaje natural. Poseen una excesiva ambigüedad semántica.

De este tipo son las listas de descriptores libres y las listas de palabras clave. Los lenguajes libres no son propiamente lenguajes documentales puesto que para que reciban este nombre el vocabulario ha de estar controlado.

2.3.2.1.2 Lenguajes Controlados

Son los lenguajes documentales propiamente dichos: tesauros, listas de encabezamiento de materias y clasificaciones. Presentan un vocabulario previamente elaborado, y admiten un limitado número de modificaciones en el momento de su utilización. Son aquellos que controlan las ambigüedades propias del lenguaje, tanto para la representación como para la recuperación de la información. También el control se realiza a nivel de relaciones semánticas.

En conclusión, el vocabulario controlado proporciona al usuario un punto de búsqueda, en vez de dos o más y reduce la posibilidad de que la búsqueda sea incompleta. Sin embargo, puede perderse información debido a la falta de especificidad y errores en el análisis.

2.3.2.2 Según la Coordinación

La sistematización de los lenguajes documentales según el criterio de coordinación (*Tabla 2.2*) se realiza en función del momento en que se combinan los elementos que los componen. Si los términos se combinan en el momento de la descripción, el lenguaje será precoordinado, y si lo hace en el momento de la recuperación, se tratará de un lenguaje postcoordinado.

Coordinación	Precoordinados	Clasificaciones
		Listas de encabezamientos de materia
	Postcoordinados	Listas de descriptores libres
		Listas de palabras clave
		Tesauros

Tabla 2.2: Tipología de los lenguajes documentales según el criterio de coordinación

2.3.2.2.1 Lenguajes Precoordinados

Combinan los términos en el momento de la descripción, es decir, antes del almacenamiento. Es el caso de los sistemas tradicionales de clasificación y los jerárquicos así como también las listas de encabezamiento de materias.

Son lenguajes precoordinados aquéllos que se han elaborado previamente a su aplicación, e incluso, a la normalización terminológica de las áreas de conocimiento para los que se van a destinar. Son sistemas denominados de tipo sintético, donde las distintas nociones o conceptos que se unen para expresar una materia o un tema ocupan un lugar determinado, es decir se introducen en el momento de la indización en un orden previamente establecido y la recuperación habrá de hacerse secuencialmente, siguiendo ese orden. Suelen

ser muy precisos, pero también muy rígidos. Además, el objetivo en la descripción de contenidos apunta primordialmente a la obtención de materias y no tanto hacia el resultado de conceptos.

El prototipo de los lenguajes precoordinados son las clasificaciones, nacidas de la necesidad de organizar los conocimientos, como se dijo en el párrafo anterior, según un orden establecido.

2.3.2.2.2 Lenguajes Postcoordinados

Combinan los términos libremente en el momento de la recuperación. Son lenguajes postcoordinados los tesauros (lenguaje utilizado para la indización por descriptores), y si incluyéramos los lenguajes naturales podemos mencionar también las listas de descriptores libres y las listas de palabras clave. En estos sistemas los conceptos que se extraen en la indización para expresar el tema o los temas del documento tienen todos la misma categoría y no se expresarán en ningún orden determinado.

Los lenguajes postcoordinados carecen de sintaxis, y, salvo ciertas excepciones, su vocabulario consiste en términos simples o unitérminos cuya combinación se realiza en la fase de recuperación utilizando distintos tipos de operadores, como por ejemplo operadores booleanos o de comparación. Asimismo, son considerados como lenguajes de tipo analítico, ya que desde una perspectiva semántica, y al contrario de los precoordinados, los conceptos ocupan casillas mucho más pequeñas previstas para que cada parte conceptual del documento pueda ser representada. El objetivo en la descripción de contenidos está principalmente dirigido a la obtención de conceptos, y no tanto hacia el resultado de materias.

2.3.2.2.3 Comparación: Lenguajes Precoordinados vs. Lenguajes Postcoordinados

Atendiendo a los procedimientos seguidos para relacionar los conceptos al recuperar la información, podemos concluir que en los lenguajes precoordinados la relación entre los términos es gramatical, mientras que en un lenguaje postcoordinado la relación es lógica.

Los lenguajes precoordinados contribuyen mejor que los postcoordinados a conseguir precisión en la búsqueda, es decir, tienen mayor capacidad para rechazar los documentos irrelevantes en el momento de la recuperación porque las relaciones que se crean mediante operadores booleanos son genéricas y no impiden las falsas combinaciones. Como contrapartida, los postcoordinados tienen la ventaja de dar respuesta a necesidades de indización reales. A continuación se puede ver un ejemplo comparativo de ambos tipos de lenguaje:

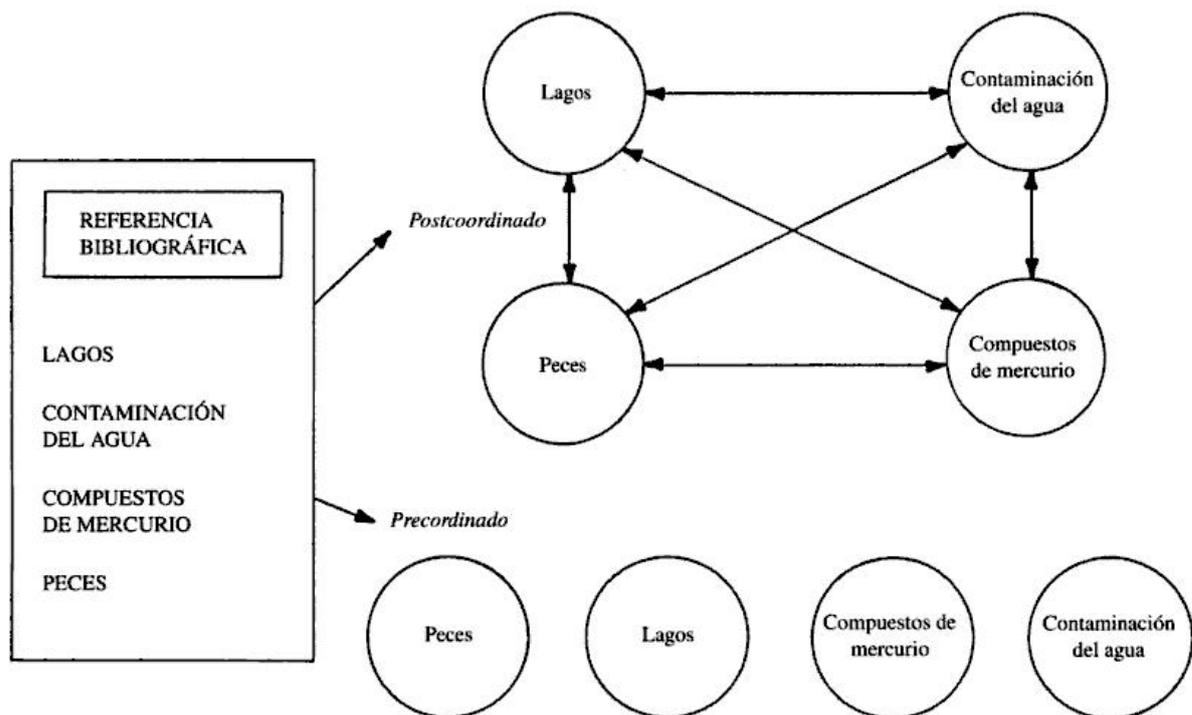


Figura 2.1: comparación lenguajes precoordinados vs. lenguajes postcoordinados

Se puede ver en la *Figura 2.1* la distinción entre un lenguaje precoordinado y otro postcoordinado. Un documento ha sido indizado con cuatro términos (asignado a cuatro clases). En un sistema postcoordinado se conserva la multidimensionalidad de la relación entre las cuatro clases: no es preciso un orden de clases ya que todas tienen el mismo peso y permite recuperar el documento independientemente de cuál sea la combinación de los cuatro términos que se plantee en la búsqueda. Sin embargo, un índice de materias impreso o en forma de fichero convencional pierde la multidimensionalidad. Es posible la confección de una entrada en la que estén presentes todos los términos de indización, pero tendrán que estar ordenados en una secuencia lineal, y sólo podrá accederse al documento a través del primer término de la cadena. Por ejemplo, en la *Figura 2.1* la entrada del índice PECES, LAGOS, COMPUESTOS DE MERCURIO, CONTAMINACIÓN DEL AGUA permite la recuperación sólo si el usuario busca en el índice el término PECES, ya que los otros términos son subdivisiones o modificadores de él. Este tipo de lenguaje es el que llamamos precoordinado: las clases se coordinan (combinan) en una cierta secuencia; el usuario no puede combinar libremente las clases y por tanto, no puede recuperar documentos a partir de aspectos que no estén explícitos en el índice.

2.3.2.3 Según la Estructura

En cuanto a la estructura, se reconoce la existencia de tres tipos: jerárquica, combinatoria y sintáctica.

Estructura	Jerárquica	clasificaciones jerárquicas
	Combinatoria	léxicos documentales
		tesauros
	Sintáctica	de gramática simple
		de gramática elaborada

Tabla 2.3: Tipología de los lenguajes documentales en función de su estructura

2.3.2.3.1 Estructura Jerárquica

También denominada Arbórea o Sistemática, la característica de este tipo de lenguajes es que organizan sistemáticamente el conocimiento, es decir, agrupan las materias en áreas o categorías, donde todos los conceptos dependen de uno superior de significado más genérico. Cada concepto de estas estructuras jerárquicas se halla representado por un símbolo numérico, alfabético o alfanumérico que indica la situación de cada materia, y entre los elementos se establecen relaciones de subordinación, cosubordinación o ambas clases de relaciones, facilitando su localización en la estructura arborescente. En la relación de subordinación, se clasifican algunos conceptos como inferiores en orden con respecto a otros. Por otro lado, la cosubordinación corresponde a una relación que se establece entre los miembros pertenecientes a un mismo nivel de jerarquía

Este tipo de lenguaje se puede emplear para localizar un documento, pero no para indizar con profundidad. Los lenguajes de estructura jerárquica pueden dividirse como se describe a continuación.

2.3.2.3.1.1 Clasificaciones enciclopédicas

Permiten la organización de documentos que tratan sobre cualquier materia: son de ámbito universal y multidisciplinario. Presentan dos inconvenientes: el objetivo de su universalidad limita la descripción de un documento especializado y su rigidez dificulta una puesta al día ágil y rápida. Ejemplos de este tipo de clasificaciones son:

- 1 Clasificación de Dewey (1876):** divide el conjunto de los conocimientos en 9 clases principales, designadas en números arábigos del 1 al 9, reservando el 0 para las generalidades. Cada clase se subdivide sucesivamente en 10 subclases y así sucesivamente, con números que se dividen en grupos de tres por medio de puntos para hacer más fácil la lectura.
- 2 Clasificación expansiva de Cutter (1891):** se compone de 7 tablas o esquemas, cada una de los cuales incluye la totalidad de los conocimientos, pero con una complejidad progresiva.
- 3 Clasificación de la Libray of Congress (LCC 1897):** cuenta con 21 clases principales, tomadas del sistema de Cutter, que designa con otras tantas letras mayúsculas, dejando las restantes para futuras ampliaciones.
- 4 Clasificación bibliográfica de Bliss (1935):** está formada por 4 esquemas generales que son filosofía, ciencias zoológicas, físico y social que se dividen en un total de 26 clases principales. Es muy similar a las facetadas.
- 5 Clasificación Decimal Universal (CDU 1905):** es una ampliación de la clasificación decimal de Melvil Dewey (se basa en su quinta edición). Es numérica hasta cierto punto, precoordinada, universal, multidimensional y arborescente. Está agrupada en 10 clases, reducidas a 9 por la fusión de lingüística, filología y literatura, dejando la clase 4 vacía. Cada una de ellas se subdivide. La tarea de actualizar la CDU está encomendada a la FID (Federación Internacional de Documentación), si bien como señala Inocencia Soria (2.13): "En 1992 se constituyó el consorcio CDU, que asumió las

responsabilidades que antes tuviera la FID sobre su edición, actualización, versiones, etc. Este consorcio, cuyos socios fundadores son Bélgica, España, Holanda, Japón, Reino Unido y la propia FID, se comprometió a organizar y mantener la CDU y sus esfuerzos ya han dado algunos importantes frutos: se ha creado un fichero informático con más de 60.000 entradas que está sirviendo de base para facilitar su manejo y actualización".

2.3.2.3.1.2 Clasificaciones especializadas

Son instrumentos de indización que abarcan disciplinas o campos especializados (medicina, derecho, economía, etc.) que no quedarían profunda y ampliamente representados en una estructura jerárquica de ámbito universal y multidisciplinario. Por ejemplo, la mayoría de las bases de datos disponen de este tipo de clasificaciones.

2.3.2.3.1.3 Clasificaciones de facetas

Son de origen enciclopédico, pero su organización permite construir áreas concretas de los conocimientos, ya que una faceta es cada uno de los aspectos o puntos de vista que pueden incluirse en un área concreta. Se basan en la clasificación colonada de S.R. Ranganathan (1933). Se llama colonada porque utiliza el colon como único signo de síntesis. No es, por tanto, una división lineal y jerárquica como el resto sino la aplicación que tienen algunas materias para descomponerlas. Son muchas sus características, pero en el campo de la biblioteconomía se reducen a cinco: personalidad, materia, energía, espacio y tiempo. Sin embargo, tiene subdivisiones de lengua, geografía, cronología, etc.

2.3.2.3.2 Estructura Combinatoria

Estos lenguajes de estructura combinatoria (o asociativa) se presentan como una lista de términos representativos y permiten la libre combinación entre los mismos, de acuerdo a las necesidades de indización. Los términos se combinan libremente entre sí. Son lenguajes de estructura combinatoria los tesauros y los léxicos documentales, si bien es cierto que en la organización multidimensional de los tesauros participa también la estructura jerárquica.

2.3.2.3.3 Estructura Sintáctica

A este tipo de estructuras, surgido en una época donde no había herramientas para el almacenamiento masivo, pertenecen los lenguajes que recurren a una sintaxis mediante la cual se pueden representar y poner en relación los contenidos de los documentos.

Por el momento, estos lenguajes están en fase de experimentación, y los que han aparecido no han tenido demasiado éxito.

2.3.3 Lenguajes documentales: definición y características

A continuación, se presenta una breve descripción de las características y el funcionamiento de los distintos lenguajes documentales que suelen utilizarse para normalizar la representación y recuperación de los términos que componen un documento.

2.3.3.1 Listas de palabras clave

Son listas no estructuradas (excepto por su ordenación alfabética) de conceptos que han sido seleccionados por métodos automáticos, para describir el contenido de los documentos analizados. El método de determinación de las listas es singular en cada documento de manera tal que no permite la descripción de otros materiales.

La mayoría de las veces son monolingües, aunque pueden contener términos de dos o más lenguas pero sin equivalencias entre si.

Se trata a menudo de términos muy específicos y actuales que permiten seguir de cerca la evolución terminológica de un dominio en particular. A esta ventaja se añaden otras como el ahorro que supone no tener que elaborar y mantener un lenguaje documental, etc. La principal desventaja es la ausencia de control terminológico a causa de la ambigüedad del lenguaje natural.

2.3.3.2 Listas de Descriptores Libres

Este tipo de lenguajes utilizan el llamado sistema unitérmino, basado en el principio de postcoordinación, que permite la elección del término de indización en respuesta a necesidades de información reales.

Los conceptos son expresados por palabras o expresiones extraídas a partir del análisis de documentos, sin verificar si existen en una lista establecida a priori. Las listas de descriptores libres no limitan la incorporación de nuevos conceptos.

2.3.3.3 Sistemas de Clasificación

“Clasificar, en términos generales, es el acto de organizar el universo del conocimiento en algún orden sistemático. Ha sido considerada la actividad más fundamental de la mente humana. El acto de clasificar consiste en el dicotómico proceso de distinguir cosas u objetos que poseen cierta característica de aquellos que no la tienen, y agrupar en una clase cosas u objetos que tienen la propiedad o característica en común” (2.14)

Este concepto aplicado al saber que reside en los documentos ha permitido identificar contenedores ideales representados por los temas en los que el saber mismo puede ser dividido. El objetivo de la clasificación en el análisis documental consiste en identificar a qué clase o subclase se atribuye idealmente un documento y concretar este aspecto teórico en la elaboración de un catálogo sistemático de temas, por ejemplo. Para poder realizar esto, las

bibliotecas digitales deben adoptar un sistema de clasificación que proporcione una subdivisión precisa del saber, organizándolo en clases y subclases.

Los sistemas de clasificación son numerosos, y van desde el más rudimentario hasta aquellos que se basan en un fundamento científico y son adoptados institucionalmente tanto por las grandes bibliotecas nacionales como por pequeñas bibliotecas municipales. En cualquier caso, como lenguaje documental que representa de forma formalizada el contenido de los documentos, los sistemas de clasificación permiten la recuperación (manual o automatizada) de información requerida por los usuarios. Por otra parte, su principal inconveniente es su falta de operatividad por la rigidez de su estructura.

Los sistemas de clasificación, además de los requisitos de cualquier lenguaje documental, han de cumplir también las siguientes condiciones:

- No utilizan términos de lenguaje natural, sino signos o códigos normalizados basados en cifras, letras y otros símbolos gramaticales, que pretende ser la descripción sintética del contenido de los documentos.
- Tienen que ser explícitos aunque concisos, es decir, que con el menor número de signos posibles expresen bien el contenido del documento.
- Siempre son lenguajes precoordinados.
- Deben ser completos: han de abarcar todos temas posibles en que se divide un área de conocimiento.
- Deben ser sistemáticos: proceder de lo general a lo particular, formando una estructura donde los conceptos se relacionan y ordenan en función de características específicas.

2.3.3.4 Listas de Encabezamiento de Materias

Lenguaje precoordinado, de estructura asociativa o combinatoria que, según Carrión Gutiez (2.15), son “signos que representan la materia o asunto que trata un documento”. De la misma manera, Miyashiro Anashiro (2.16), indica que los encabezamientos consisten en “una o varias palabras que representan conceptos (...). Intenta, por tanto, condensar el tema sobre el que trata el documento. Está constituido por términos del lenguaje natural, lo que provoca problemas sintácticos y semánticos (sinonimia, por ejemplo), que se resuelven mediante el establecimiento de una serie de relaciones que dará coherencia a las listas, facilitando el control terminológico”.

Cutter (2.17) fue el primero en intentar constituir reglas para los encabezamientos de materia, estableciendo dos reglas básicas: las de especificidad y de entrada directa. La primera se refiere a la importancia de utilizar los conceptos bajo su nombre más concreto y no englobarlos en uno más general. El principio de entrada directa hace referencia a la conveniencia de usar los encabezamientos compuestos por más de una palabra en la forma en que se presentan en el lenguaje natural, evitando la inversión en los términos que lo componen.

2.3.3.5 Tesoros y Descriptores

Los tesauros se pueden definir según su función y según su estructura:

- Por su función, se pueden definir como instrumentos de control terminológico. Controlar el vocabulario significa identificar dentro de un campo semántico todos los conceptos que son representados por más de un término. La identificación de términos equivalentes hace posible minimizar la pérdida de información en las búsquedas realizadas en un sistema documental automatizado.
- Por su estructura, los tesauros permiten conocer todos los términos relacionados con un concepto determinado, lo que ayuda a añadir más términos adecuados para enriquecer tanto los análisis de contenido de los documentos como las estrategias de búsqueda para recuperar información.

El tesoro es un prototipo de lenguaje de indización y recuperación controlado, que se basa en la postcoordinación de sus descriptores. Además representa de manera unívoca los conceptos de los documentos evitando así los problemas de homonimia, sinonimia, polisemia (relaciones de significado), así como el establecimiento de las relaciones jerárquicas y de relación entre los descriptores que lo componen.

Los descriptores son términos del lenguaje natural que por medio de un proceso de selección llega a formar parte del vocabulario de un tesoro. La condición que debe superar un término para adquirir la condición de descriptor es que sea el más representativo del concepto que se quiere representar dentro del tesoro. Con la elección de este término como único representante válido de entre otros muchos con significados idénticos o casi idénticos se consigue la univocidad del lenguaje documental, en este caso del tesoro. Un descriptor, por lo tanto, es un término que representa de manera unívoca un concepto dentro de un lenguaje documental.

Las palabras clave o los unitérminos se diferencian de los descriptores en que, a pesar de compartir la procedencia del lenguaje natural y de tener la función de representar conceptos, las palabras claves carecen del control estricto de la sinonimia, de las relaciones asociativas y, mucho menos de las jerarquías. Los descriptores se caracterizan por presentar relaciones semánticas y mayor control que las palabras clave.

Los tesauros son los elementos de vocabulario controlado más utilizados para la representación y la recuperación de información, debido a su especificidad, flexibilidad y capacidad de manejar conceptos complejos.

2.3.3.6 Ontologías

Como se puede apreciar, existen varios paradigmas y formas concretas de representar del conocimiento, pero las ontologías parecen ser la mejor forma en el ámbito de la web semántica, la cual se trata de una nueva concepción de la web en la que el significado de las cosas y la capacidad de hacerlo inteligible a las máquinas juegan un papel esencial.

Su utilización es clave desde el punto de vista de la reutilización del conocimiento en contextos diferentes al original, ya que por su estructura y capacidad de formalización permiten representar colecciones de sentencias en un lenguaje como puede ser RDF, el cual se describirá en el próximo capítulo, que definen las relaciones entre conceptos y reglas lógicas específicas para razonar acerca de ellos.

Una ontología, tal y como se entiende el término en filosofía, es un registro sistemático de las cosas que existen. Esta idea fue tomada del campo de la inteligencia artificial con algunas modificaciones, de forma que para un sistema basado en el conocimiento, lo que existe es lo que puede ser representado (2.18), y debe ser especificado mediante un formalismo declarativo, que pueda ser procesado no sólo por personas sino también por computadoras. La ontología es conocimiento compartido, fruto del consenso dentro de un grupo.

Existen varias definiciones formales del término ontología, por ejemplo, según Noy y McGuinness (2.18) “es una descripción formal y explícita de los dominios del discurso”. Por otra parte, Gruber (2.19), ofrece otra interpretación más abstracta: “especificación explícita de una conceptualización” siendo ésta “una visión abstracta y simplificada del mundo que queremos representar con algún propósito”.

Las ontologías se han incorporado con fuerza al ámbito de la web (2.20) y son utilizadas con cierta frecuencia en la esfera comercial. En general se puede decir que se usan por personas, bases de datos o aplicaciones que necesitan compartir información acerca de un dominio concreto. Entre las distintas aplicaciones de las ontologías existentes, es de nuestro interés principal la aplicación relacionada con la búsqueda de información. En este caso, las ontologías se utilizan para la anotación de recursos en Internet (documentos, páginas web, imágenes, etc.), y para guiar la búsqueda en un dominio concreto. Esto proporciona una mayor flexibilidad de búsqueda, incrementando la precisión y la recuperación de los documentos buscados.

Otra de las aplicaciones interesantes de ontologías se puede encontrar en las bibliotecas digitales, las cuales proporcionan acceso a grandes cantidades de información en forma de documentos digitales, que pueden tener formatos muy variados y estar distribuidos en sistemas informáticos dispares. Las técnicas basadas en las ontologías permiten tratar esta heterogeneidad mediante la descripción de objetos y repositorios, y posibilitar así un acceso sencillo, consistente y coherente a los recursos digitales.

2.3.3.7 Diferencias entre los lenguajes documentales

Son apreciables las divergencias tanto a nivel semántico como estructural. En el primer caso, ya sean los sistemas de clasificación o las listas de encabezamientos de materias, se limitan a representar mediante términos la información contenida en los documentos. Los tesauros, en cambio, han tratado de representar mediante conceptos dicha información, con la ayuda de un sistema de relaciones que, aunque complejo, no alcanza la capacidad descriptiva que se desarrolla en las ontologías.

En este sentido, se puede decir que los tesauros intentan llegar al nivel conceptual a través de la utilización de relaciones que están fuertemente ancladas a nivel léxico. Es decir, se interpreta que el nivel conceptual representado por los términos que componen un tesoro está directamente asociado a dichos términos. En el caso de las ontologías el significado de la información se explicita a través de atributos, de sus características, y no de su representación léxica.

2.4 Bibliografía

(2.1) Acosta, V. y Moreno, A. M^a. (1999). Dificultades del lenguaje en ambientes educativos. Barcelona, España. Editorial Masson, S.A.

(2.2) UNIVERSIDAD ABIERTA DE CATALUÑA. Lenguajes documentales. <http://docupo.pbwiki.com/Los+lenguajes+documentales> [Consulta: 2012-08-29]. Revista sobre la sociedad del conocimiento, <http://www.uoc.edu/uocpapers/4/esp/index.html> [Consulta: 2012-08-29]

(2.3) ROWLEY, J. Organizing knowledge: an introduction to information retrieval. Aldershot: Gower, 1992.

(2.4) Gil Urdiciain, Blanca. Manual de lenguajes documentales. Madrid : Noesis, 1996. Cap. I.

(2.5) Guinchat, C., Menou, M. Introducción general a las ciencias y técnicas de la información y de la documentación. 20 ed. corr. y aum. Marie France Blanquet. Madrid : CINDOC-Unesco, 1992.

(2.6) Coll-Vinent, R. Teoría y Práctica de la Documentación. 2º ed. Barcelona, A.E.T., 1978, p. 71

(2.7) Courrier, Y. Analyse et Langage Documentaires. Documentaliste, vol 13, nº 5-6, 1976, p. 180

(2.8) Amat Noguera, N. Técnicas Documentales y Fuentes de Información. Barcelona, Biblograf, 1978, p. 155

(2.9) LÓPEZ-HUERTAS P., MARÍA JOSÉ. Lenguajes documentales: terminología para un concepto. En: Boletín de la ANABAD, ISSN 0210-4164, Tomo 41, Nº 2, 1991. <http://dialnet.unirioja.es/servlet/oaiart?codigo=224133> [Consulta: 2012-08-29]

(2.10) RODRIGUEZ DELGADO, R. La integración de los lenguajes documentarios, fin de Babel. Revista Española de Documentación Científica, y. 4, n. 3 (1980).

(2.11) Rodríguez Luna, Cristina: Lenguajes documentales. Universidad de León, 2003.

(2.12) Slype, G. Van: Los lenguajes de indización : concepción, construcción y utilización en los sistemas documentales. Madrid : Fundación Germán Sánchez Ruipérez, 1991, p . 161.

(2.13) Soria González, Inocencia: La organización de la información, los lenguajes documentales y la normalización. Consejo Superior de Investigaciones Científicas.

- (2.14) Chan, L. M. : Cataloging and classification: an introduction. New York: McGraw-Hill, 1981, p.209.
- (2.15) Carrión Gutiez, Manuel: Automatización de Bibliotecas. Madrid: Pirámide, Fundación Germán Sánchez Ruipérez, 1993.
- (2.16) Miyashiro Anashiro, Martha: Técnica del sub-epígrafe y registro sub-epigráfico (Tesis). Lima, 1978. p.1.
- (2.17) Cutter, Charles A. Rules for a Dictionary Catalog. 4th ed. Washington, D.C.: Government Printing Office, 1904.
- (2.18) Noy, Natalya F.; McGuinness, Deborah L. Ontology development 101: A guide to creating your first ontology. Informe técnico. Universidad de Stanford, 2000.
- (2.19) Gruber, Thomas R. "Towards principles for the design of ontologies used for knowledge sharing". En: Guarino, N.; Poli, R. (eds.). Formal ontology in conceptual analysis and knowledge representation. Deventer: Kluwer Academic Publishers, 1993.
- (2.20) Heflin, Jeff. OWL Web ontology language use cases and requirements. Recomendación W3C, febrero 2004. Disponible en: <http://www.w3.org/TR/webont-req/> [Consulta: 2012-08-29]

Capítulo 3

Metadatos

Capítulo 3 - METADATOS

3.1 Introducción

Desde hace ya varios años la cantidad de información disponible en la red ha crecido exponencialmente. Debido a la gran diversidad y volumen de las fuentes y recursos en Internet se hizo necesario recurrir a un mecanismo para etiquetar, catalogar, describir y clasificar los recursos presentes en la red con el fin de facilitar la posterior búsqueda y recuperación de la información. Este mecanismo los constituyen los llamados metadatos: estructuras de base para describir distintos objetos de información distribuidos en la web, de forma tal que la búsqueda basada en esos metadatos disminuyese el problema de la recuperación de información.

Aunque el uso de la palabra “metadato” se masificó en un contexto que se refiere a la era de la información digital, la generación de metadatos data de siglos atrás. Los bibliotecarios han creado metadatos que han tomado la forma de catálogos de libros, catálogos de tarjetas y en la actualidad catálogos en línea. Hoy en día, la generalización del concepto ha cubierto cualquier tipo de información descriptiva (estandarizada) sobre recursos, incluyendo los que no son digitales.

En este marco, surge lo que algunos denominan la Segunda Generación de la Web, propiciada por el desarrollo del XML (eXtensible Markup Language). Sobre la base de XML se han definido distintos lenguajes de marca para los diferentes tipos de documentos, como el lenguaje de marcado semántico RDF (Resource Description Framework).

3.2 De la Información a la Metainformación

Esta sección busca aproximarse a definiciones más pertinentes del término metadato en el ámbito que corresponde a la presente investigación. Se tratan también diferentes aplicaciones de los metadatos y se establecen múltiples clasificaciones según el ámbito de aplicación, la utilidad o función que prestan, la cantidad de contenido informativo que engloban, y la riqueza semántica y complejidad estructural de los modelos de metadatos.

3.2.1 Definición y Aplicaciones

Un metadato no es más que un dato estructurado sobre la información, o sea, información sobre información, o de forma más simple, datos sobre datos. Caplan (3.1) adopta esta última definición y considera a las fichas de los catálogos tradicionales como metadatos, ya que engloban datos bibliográficos (como por ejemplo, autor, título, editorial, etc.) que se refieren a otros documentos. Por otro lado, Xu (3.2), sostiene que es un conjunto de elementos que pueden ser usados para describir y representar objetos de información. Cabe

destacar que en esta definición, el autor habla de objetos de información, independientemente del soporte. Younger (3.3) por su parte, entiende que los metadatos describen recursos, indican dónde están ubicados y qué se requiere para utilizarlos exitosamente. Dempsey y Heery (3.4) los definen como datos que describen los atributos de un recurso. Para Wendler (3.5), la definición apropiada es: "información necesaria para identificar, localizar, manejar y acceder a un recurso electrónico".

Otra definición interesante es la que brinda Tennant en (3.6), exponiendo que los metadatos son "información estructurada sobre información", y destaca que la palabra clave es "estructurada", ya que una descripción en texto libre no es suficiente, es necesario contar con ciertos elementos identificados formalmente y con una codificación para la especificación de una sintaxis dada. En contraparte, Milstead y Feldman (3.7), consideran al registro bibliográfico en sí como metadatos y los cuales no necesariamente deben estar codificados.

Gorman (3.8) diferencia lo que son estándares de estructura de aquellos que prescriben el contenido del registro bibliográfico. Para este autor los metadatos han sido diseñados para responder a las necesidades de: contar con una opción para la catalogación de los recursos electrónicos y, encontrar una alternativa intermedia entre los altos costos de la catalogación tradicional y la simple recuperación por palabra clave. Por esto, sostiene que no es necesario innovar en el tema ya que las técnicas de descripción bibliográfica de la catalogación tradicional cubren satisfactoriamente la catalogación de los recursos en línea.

También cabe mencionar la definición que otorga Taylor en (3.9), la cual introduce dos conceptos, el de contenido y el de codificación, presentes ambos en la acepción del uso común del término de metadatos. Cuando sólo hay contenido, se lo denomina un registro bibliográfico y, cuando hay sólo codificación, es identificado como una "estructura". Para esta autora, un metadato es una descripción codificada de un paquete de información.

En el marco del World Wide Web Consortium (W3 Consortium), se los define como "información sobre objetos de información en la web, que puede ser leída por computadora" (3.10). Es decir, que en el contexto de la web, son datos que se pueden guardar, intercambiar y procesar por medio de una computadora y que están estructurados de tal forma que permiten ayudar a la identificación, descripción, clasificación y localización del contenido de un documento o recurso web y que, por tanto, sirven para su recuperación, así como también para proporcionar cierta información semántica.

Sintetizando, de todas estas definiciones podemos deducir que las funciones primarias de los metadatos son facilitar la identificación, ubicación, recuperación, manipulación y uso de los recursos de información accesibles en línea.

Existen distintos modelos de metadatos, cada uno de ellos con distintos esquemas de descripción. En los distintos modelos, cada objeto se describe por medio de una serie de atributos y el valor de estos atributos es el que puede servir para recuperar la información. Dependiendo de la clase de metadatos puede existir: información sobre elementos de datos o atributos, información sobre la estructura de los datos, información sobre un aspecto concreto, etc. De forma general, podemos encontrar metadatos referidos a: el contenido (concepto), aspectos formales (tipo, tamaño, fecha, lengua, etc.), información del copyright, información de la autenticación del documento o recurso, e información sobre el contexto (calidad, condiciones o características de acceso, uso, etc.).

Los metadatos pueden ser almacenados dentro de una base de datos con una referencia al documento completo o ser incluidos en un encabezado dentro del propio texto.

En el contexto de la Web, los metadatos se forman y almacenan para que puedan ser leídos por los motores de búsqueda. Las grandes ventajas del uso de metadatos radican en que se usa el mismo contenido del documento como un recurso de datos y que los metadatos valen también para recursos que no tienen únicamente la morfología de texto, sino para cualquier tipo de morfologías tales como vídeo, audio o imágenes.

Las aplicaciones del uso de metadatos son muy amplias y van desde la recuperación de información, pasando por la descripción y catalogación de documentos, su uso por parte de robots y agentes de software, comercio electrónico, firmas digitales, derechos de propiedad intelectual; valoración, evaluación y clasificación de contenidos; trabajos bibliométricos e informétricos de todo tipo, etc.

La aplicación de los metadatos en el diseño de páginas web aporta a la descripción de la forma de las páginas, información sobre su contenido. Incluso se pueden definir estructuras de datos e interrelaciones entre los mismos.

El uso de lenguajes para la definición de metadatos estandarizados, tales como XML ó RDF permiten el intercambio de información entre diferentes máquinas, con diferentes sistemas operativos, favoreciendo así la recuperación. Nacen con este propósito diferentes estándares como Dublin Core Metadata Initiative que pretenden definir una serie de vocabularios de metadatos para describir recursos. De esta forma se puede crear un lenguaje estandarizado que defina recursos de forma internacional, lo que facilita el acceso y la recuperación de información.

3.2.2 Clasificación

Se han establecido múltiples y diversas clasificaciones de tipos de metadatos atendiendo a distintos aspectos como su forma, funcionalidad, nivel de estructuración de los datos, persona o entidad que los origina, etc. Las clasificaciones por parte de los diferentes autores e instituciones son muy variadas, y se establecen atendiendo a los distintos aspectos a los que se dé prioridad a la hora de establecer dichas clasificaciones.

De forma general, de acuerdo a la naturaleza de los datos que describen, los metadatos pueden clasificarse, según Lazinger (3.11), en tres amplias categorías con límites no siempre bien definidos y a veces con superposiciones entre sí:

- **Metadatos descriptivos:** son aquellos que sirven para la descripción e identificación de los recursos de información, permiten la búsqueda y recuperación de la información, como también la distinción de un recurso de otro, comprendiendo el asunto o contenido del mismo. Se realizan mediante los estándares como Dublin Core; MARC; Meta tags, HTML, etc.
- **Metadatos estructurales:** son los que más influyen en la recuperación de la información electrónica, facilitan la navegación y presentación de los recursos electrónicos. Así, ofrecen la información sobre la estructura interna de los recursos, estableciendo las relaciones entre ellos, de manera que pueden incluso unir los archivos de imagen y textos que están relacionados. Los estándares más difundidos para ellos son SGML y XML/RDF; EAD (Encoded Archival Description-Descripción Codificada de Archivos).

- Metadatos administrativos: son de carácter más técnico porque incluyen datos sobre la creación y control de calidad, datos sobre la gestión de derechos, requisitos del control de acceso y utilización, información sobre la preservación y permiten la gestión a largo y corto plazo. Ejemplo de los metadatos que se incluyen aquí: tipo y modelo de escáner utilizado, resolución, limitaciones de reproducción, etc.

Además de esta clasificación, también se pueden analizar sus características teniendo en cuenta otros criterios, como por ejemplo, refiriéndose a la función que cumplen, se pueden distinguir los siguientes grupos:

- Metadatos de acceso: permiten la navegación, consulta y recuperación de la información.
- Metadatos semánticos: permiten asignar un significado a la información.
- Metadatos de calidad: permiten un análisis cualitativo de la información.
- Metadatos de transferencia: permiten transferir la información entre aplicaciones.
- Metadatos de almacenamiento: permiten el almacenamiento de la información.

Desde el punto de vista del contenido, se encuentran los siguientes tipos de metadatos:

- Metadatos independientes del contenido: recogen la información que no depende del contenido del documento (localización, fecha de creación y actualización, seguimiento y control de versiones, etc.).
- Metadatos dependientes del contenido: recogen la información que depende del contenido, ya sea de forma directa o indirecta. Este tipo de metadatos permite la interoperabilidad semántica, ya sea que se trate de dominios generales o específicos. Otras clasificaciones similares a la anterior son:
 - Metadatos basados en el recurso: sirven para la identificación y catalogación del recurso digital.
 - Metadatos basados en la materia: representan el contenido y sus relaciones.

Esta proliferación de clasificaciones responde a que los metadatos están en fase de construcción técnica y por eso no existe un consenso generalizado en su conceptualización o sobre los tipos de metadatos existentes (3.12)

3.3 Modelos de Metadatos

Se plantea una aproximación más profunda sobre los principales esquemas de metainformación, partiendo del concepto general de marcado de documentos, de los principales lenguajes (SGML, HTML, XML) y de sus implicaciones para la estructuración y el acceso a la información Web. Se describen los modelos de metadatos más representativos y utilizados, distinguiendo los modelos de propósito general (Dublin Core), y los modelos de propósito específico. Se presenta a RDF como metamodelo de metadatos, describiendo su

3.3.1 Estructura de los Metadatos

Típicamente, los elementos que conforman un metadato están definidos por algún estándar o perfil, donde los usuarios que deseen compartir metadatos están de acuerdo con el significado preciso de cada elemento.

Conforme al nivel de información que brinden sobre un conjunto de datos documentado, los metadatos pueden ser mínimos o detallados:

- **Mínimo:** se restringe sólo a los componentes mas importantes e involucra las siguientes secciones:
 - **Identificación:** Información básica sobre el conjunto de datos (titulo, autoría, propósito, resumen, temática, localización etc.).
 - **Calidad:** Evaluación general de la calidad de un conjunto de datos.
 - **Distribución:** Datos del distribuidor y medios para obtener el conjunto de datos.
- **Detallado:** además de las secciones arriba mencionadas, está compuesta por otras secciones, como ser:
 - **Entidades y atributos:** Información sobre los objetos involucrados y sus atributos.
 - **Referencia del metadato:** Actualidad de la información del metadato y de sus responsables.
 - **Citación:** Datos de soporte sobre las referencias citadas dentro del conjunto de datos.
 - **Contacto:** Información de soporte sobre personas y organizaciones asociadas al conjunto de datos.

3.3.2 Evolución de los Metadatos

En el área de la organización de la información, el uso de metadatos, como parte de la identificación y clasificación es muy anterior a la era de la informática. Los primeros catálogos de libros impresos correspondían a listas ordenadas alfabéticamente sin criterios de clasificación sofisticados. Un avance importante en cuanto a esquemas de clasificación se desarrolla alrededor del año 1900, cuando los catálogos de libros son reemplazados completamente por tarjetas, de las cuales una de sus propiedades es que pueden ser actualizadas. En la década del sesenta, con el surgimiento de la tecnología, se facilitaron los proyectos de organización documental automatizada. La automatización se aplicó con mayor frecuencia a la organización de la información bibliográfica y esto dio como resultado las primeras bases de datos bibliográficas. Los métodos de producción en masa, hicieron necesario disponer de múltiples copias de los catálogos existentes. Es allí cuando surgen masivas colecciones distribuidas de libros y los catálogos de tarjetas no logran satisfacer los

nuevos requerimientos. Fue necesario entonces, desarrollar estándares de codificación, llamados hoy en día metadatos.

Los primeros metadatos digitales y sus bases se desarrollan a finales del siglo XX, cuando emergen, como se dijo anteriormente, múltiples estándares de codificación, así como también variados lenguajes y protocolos, los cuales se utilizan en la generación y uso de catálogos.

La automatización aplicada a las bibliotecas, la generación de programas y las posibilidades de cooperación bibliotecaria basada en la automatización provocaron el interés de organizaciones como la UNESCO, IFLA, FID e ISO, por generar la sistematización de la información bibliográfica orientada al uso de normas en el marco internacional y en los avances de tecnologías de información y telecomunicaciones. El resultado de dicho interés se relaciona con los formatos MARC, USMARC, UNIMARC y CCF, las normas ISO aplicadas a la documentación, las Reglas de Catalogación Angloamericanas (2da. edición) y las ISBD (International Standard Bibliographic Description).

La función de los formatos bibliográficos, como los señalados anteriormente, han sido un soporte metodológico en la representación estructural en ambiente automatizado de registros bibliográficos, para su intercambio entre unidades de información, y en la orientación del diseño de las bases de datos bibliográficas (3.13).

En las últimas décadas, el desarrollo tecnológico trajo aparejada la posibilidad de presentar y transmitir textos electrónicos a través de redes de telecomunicación, o de incorporar al texto electrónico imagen, sonido y movimiento, lo que se le conoce hoy en día como hipertexto. A partir de este avance, aparecen nuevas formas para la representación de la información electrónica. Surgieron entonces los “formatos digitales”, como HTML (HyperText Markup Language), SGML (Standard Generalized Markup Language) y XML (EXTENSIBLE MARKUP LANGUAGE), lenguajes que proponen diferentes sintaxis en la que puede ser representada la información de carácter electrónico (3.14).

A continuación, se describen dos de los primeros protocolos para la recuperación de información: el protocolo Machine Readable Cataloguing (MARC) y el Z39.50. Luego, se exponen diferentes lenguajes involucrados en la evolución del marcado de metadatos en la era digital.

3.3.3 Protocolos para la Recuperación de Información

3.3.3.1 Machine Readable Cataloguing (MARC)

El formato MARC (3.15) es un conjunto de normas que permite almacenar información en registros de cualquier tipo, para posteriormente, poder tratarla, localizarla, intercambiarla o ponerla a disposición del usuario. Fue desarrollado para ayudar a las bibliotecas en el uso,

desarrollo y mantenimiento de sus bases de datos, y precisamente dicho desarrollo ha hecho realidad la catalogación compartida y la automatización de bibliotecas.

La Biblioteca del Congreso de los E.E.U.U. inició a finales de los años '50 la investigación para desarrollar un formato legible por "máquina" para los registros bibliográficos. Otras bibliotecas comenzaron a cooperar en este proyecto que incluía el desarrollo y uso del MARC I. La Biblioteca del Congreso (Library of Congress) distribuía registros MARC I a los miembros del proyecto y las bibliotecas trataban estos registros en sus ordenadores locales.

El proyecto piloto MARC I aportó información para establecer una norma para registros legibles por máquina. El formato MARC I fue refinado y extendido a partir de 1967 en el formato conocido hoy como MARC II, concebido para intercambio de datos, capaz de almacenar información bibliográfica sobre toda clase de materiales.

Este nuevo formato, es un formato de comunicaciones, cuyo uso principal es permitir que distintas bibliotecas y organizaciones, independientemente de sus sistemas, puedan transmitirse registros entre ellas para ser usados en un sistema automatizado. La estructura del formato se aceptó por la Organización Internacional de Normalización convirtiéndose en norma ISO 2709. Esta norma, junto a la norma ANSI Z39.2, norma nacional americana, definen la estructura del formato MARC.

Desde sus inicios, MARC ha sido sometido a continuos perfeccionamientos para adaptarse a las necesidades de las bibliotecas y sus usuarios. La Biblioteca del Congreso coordina la investigación y proyectos sobre MARC, sirve como última autoridad sobre estos formatos y es la editora de la documentación de USMARC (hoy en día MARC 21). Otras organizaciones con gran influencia en el desarrollo del MARC son el Comité de la American Library Association, "MARBI", y el MARC Advisory Comité.

3.3.3.2 Z39.50

"Z39.50" es el nombre de un estándar definido por ANSI/NISO, basado en la estructura cliente/servidor, que permite comunicar sistemas que funcionan en distinto hardware y usan distinto software. Fue diseñado para solucionar los problemas asociados a la búsqueda en múltiples bases de datos con diferentes lenguajes y procedimientos (3.16). En este sentido, el protocolo permite tanto la realización de búsquedas simultáneas a múltiples bases de datos, utilizando una única interfaz de usuario, así como también recuperar la información, ordenarla, y exportar los registros bibliográficos (3.17).

Con el Z39.50 el proceso de consulta de la información es más sencillo y ágil, por eso también son más fluidas otras funciones y servicios habituales en las bibliotecas y centros de documentación, como los trabajos de referencia e información bibliográfica, puesto que una misma interfaz puede ser utilizada con diferentes motores de búsqueda y bases de datos.

También se facilitan la catalogación cooperativa, ya que el protocolo permite la descarga a menudo gratuita de registros MARC de distintas fuentes, y el préstamo interbibliotecario de un documento, solicitado a partir de los datos de ejemplares suministrados por un servidor Z (3.18).

La primera versión del estándar se liberó en 1988. Dos años después se formaron dos grupos de trabajo que garantizaron el desarrollo y evolución de la norma: un grupo de

implementadores ZIG (Z39.50 Implementors Group) y una agencia para el soporte del estándar (Z39.50 Maintenance Agency), fruto de su trabajo se aprueba la versión 2 en 1992 que, además de numerosas mejoras, evita las incompatibilidades con el protocolo de ISO "Search and Retrieve" SR (ISO 10162 y 10163).

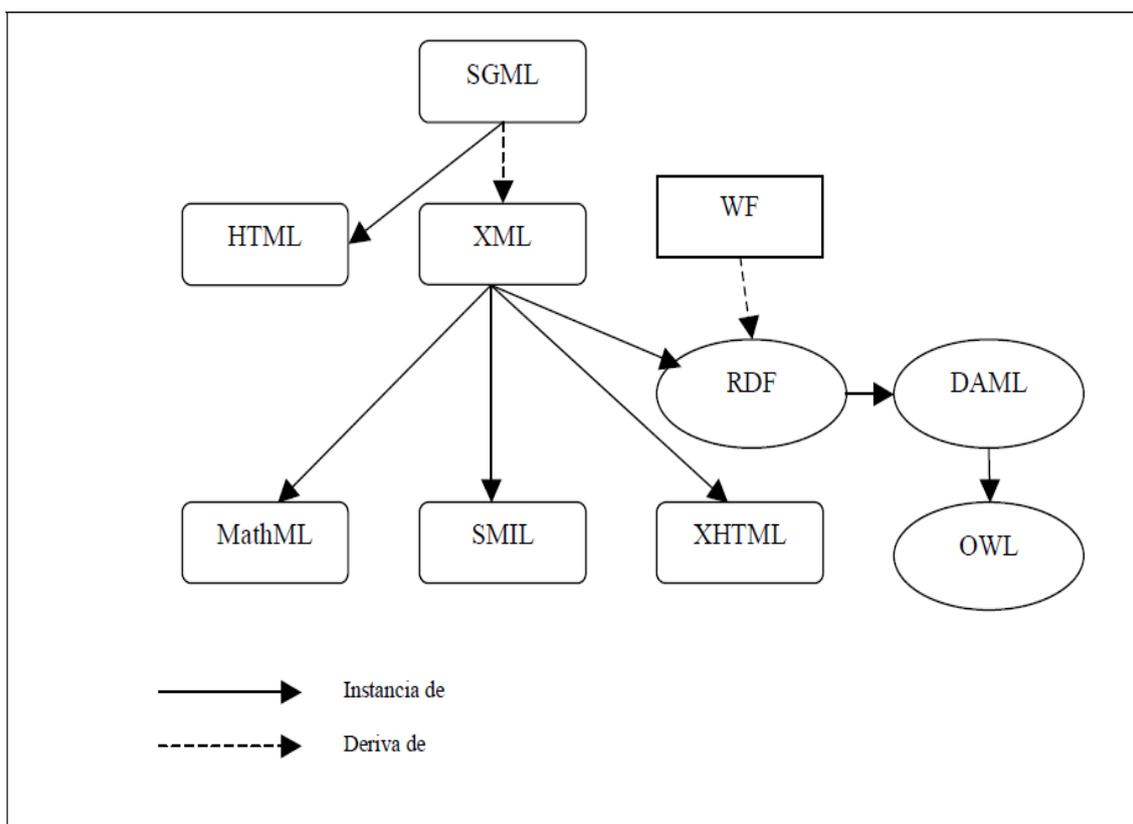
En 1995 se aprobó la versión 3, que fue aceptada como estándar de la ISO (ISO 23950) en marzo de 1997. Las nuevas facilidades que se han ido añadiendo tienen un carácter modular, y pueden irse implementando progresivamente de forma independiente. Los sistemas de gestión documental, evolucionan gradualmente hacia la versión 3, manteniendo además la compatibilidad con versiones anteriores del estándar. (3.19)

En la actualidad, Z39.50 es un estándar maduro, con una amplia presencia en la comunidad bibliotecaria, y se lo puede considerar como la norma más importante para el mundo de las bibliotecas y la documentación desde la aparición del formato MARC.

3.3.4 Lenguajes de Metadatos

Para que los metadatos se materialicen es necesaria la existencia de lenguajes que permitan especificar las sintaxis en la que se definen las estructuras, además de proveer medios para las especificaciones semánticas necesarias (que especifiquen lo que las expresiones sintácticas significan en términos de un modelo). Estos modelos y sintaxis son los que permiten representar las expresiones, hechos, reglas y consultas sobre las descripciones.

Cada uno de estos distintos lenguajes son derivaciones o instancias de los lenguajes (o esquemas) que los preceden. Como puede observarse en el siguiente diagrama extraído de (3.20):



3.3.4.1 Standard Generalized Markup Language (SGML)

Es un lenguaje estándar generalizado para marcado de documentos, que unifica la aplicación de los conceptos de anotación estructural. Sus raíces se remontan a 1969 cuando en los laboratorios de IBM se desarrolló el Generalized Markup Language (GML), lenguaje que fue evolucionando hasta 1974 donde pasó a llamarse SGML. La International Organization for Standardization (ISO) aprobó y publicó dicho lenguaje en 1984 con el nombre de estándar ISO 8879.

SGML no es un formato de almacenamiento ni un procesador de texto. Por el contrario, se trata de un metalenguaje con el que se pueden definir lenguajes de anotación que permitan almacenar y procesar texto.

Este estándar internacional consta de un conjunto de reglas para describir la estructura de un documento de tal forma que puedan ser intercambiados a través de diferentes plataformas. SGML es extremadamente flexible y es la base de los lenguajes de marcado más utilizados hoy en día.

En SGML un documento está definido en función de la estructura de las entidades que lo conforman. Estas entidades se organizan en una estructura lógica de manera jerarquizada determinando la estructura de los elementos del documento. Las entidades pueden ser compartidas por distintos documentos. El marcado se lleva a cabo mediante delimitadores y etiquetas las cuales pueden estar anidadas y se representan mediante el conjunto de caracteres básicos de acuerdo al estándar ISO 8879.

En el contexto histórico de los metadatos, la introducción de SGML jugó un papel fundamental, pues estableció un nuevo paradigma, en que los datos dejan de ser sólo datos. Los documentos SGML contienen separadamente (en el sentido lógico) los contenidos, la estructura y el formato.

3.3.4.2 HyperText Markup Language (HTML)

El hipertexto es un método de organización de la información en el cual los diferentes elementos se enlazan a través de otros elementos del propio texto. En pocas palabras, hipertexto significa texto almacenado en forma electrónica con vínculos de referencias cruzadas entre páginas, y tiene como característica que en lugar de leer un texto siguiendo una estructura rígida y lineal (como un libro), es posible avanzar de un punto a otro fácilmente, y desplazarse (navegar) por el texto.

El HTML es un “lenguaje de marcas”, que permite construir documentos hipertexto, es decir, se añaden marcas a los documentos que definen la presentación gráfica de los mismos y

los enlaces entre sus páginas. Los recursos de la presentación incluyen resaltados, separación de párrafos, negrita, subrayado, etc.

Este lenguaje se basa en la teoría de que todos los documentos tienen ciertos elementos en común como son los títulos, los párrafos, las listas y las ilustraciones.

Los documentos HTML no son más que documentos de texto con una serie de etiquetas, las cuales le son de utilidad al navegador para interpretar la forma en que tiene que ser representado el texto, las imágenes o los sonidos en la pantalla.

3.3.4.3 XML

3.3.4.3.1 Definición

A medida que el número de materiales disponible en soporte digital aumentaba, también se hacían mayores las dificultades para acceder a los mismos. Para solucionar este problema, se comenzó a trabajar a favor de la normalización de formatos, con el propósito de diseñar un lenguaje de marcas optimizado uniendo la simplicidad de HTML con la capacidad expresiva de SGML.

Tal normalización llevó al surgimiento de XML, siglas en inglés de eXtensible Markup Language (lenguaje de marcas extensible), un metalenguaje extensible de etiquetas desarrollado por el World Wide Web Consortium (W3C). Es una simplificación y adaptación del SGML y permite definir la gramática de lenguajes específicos (de la misma manera que HTML es a su vez un lenguaje definido por SGML). Por lo tanto XML no es realmente un lenguaje en particular, sino una manera de definir lenguajes para diferentes necesidades.

3.3.4.3.2 Objetivos

Dentro de los objetivos que persigue XML podemos nombrar la necesidad de distinguir el contenido y la estructura de los documentos de su presentación en papel o en pantalla; de hacer explícita su estructura y sus contenidos informativos; y crear documentos portables, que puedan intercambiarse y procesarse con facilidad en sistemas informáticos heterogéneos.

Para lograr estos objetivos XML propone un formato de documentos en texto plano (evitando las complejidades de los documentos binarios) e intercalar marcas con el objetivo de distinguir las distintas partes o elementos estructurales que conforman cada tipo de documento.

3.3.4.3.3 Funciones

Una vez definidos los objetivos de XML podemos decir que la funcionalidad de XML consiste en representar y distribuir tanto documentos como información textual; intercambiar datos e información estructurada a través de Internet y la World Wide Web; integrar datos

procedentes de fuentes heterogéneas; y eliminar la barrera entre información estructurada e información textual.

3.3.4.3.4 Ventajas

- Es extensible, lo que quiere decir que una vez diseñado un lenguaje y puesto en producción, es posible extenderlo con la adición de nuevas etiquetas de manera de que los usuarios de la vieja versión todavía puedan entender el nuevo formato.
- El analizador es un componente estándar, por lo que no es necesario crear un analizador específico para cada lenguaje. De esta manera se evitan bugs y se acelera el desarrollo de la aplicación.
- Si un tercero decide usar un documento creado en XML, es sencillo entender su estructura y procesarlo. Esto mejora la compatibilidad entre aplicaciones.

3.3.4.4 RDF

3.3.4.4.1 Definición

RDF (del inglés Resource Description Framework), es un framework para metadatos digitales, desarrollado por el World Wide Web Consortium (W3C), y basado en XML, siendo un estándar flexible para la estructuración de la información en Internet.

RDF es una especificación del W3C para la definición mediante metadatos, generalmente en XML, de los recursos que se pueden encontrar en un sitio a fin de proporcionar un marco estándar para la interoperabilidad en la descripción de dichos contenidos. Como señala Miller (3.22), XML impone la necesidad de una restricción estructural para proporcionar métodos inequívocos de expresión semántica. RDF no es más que la infraestructura que permite esa restricción gracias a la codificación, reutilización e intercambio de metadatos estructurados.

Un documento o recurso se describe a través de un conjunto de propiedades. La esencia concreta de RDF es brindar un modelo formal para la representación de dichas propiedades y los valores de las mismas (Fig.). El modelo RDF se constituye sobre principios bien establecidos en otras comunidades de metadatos, como por ejemplo el Warwick Framework (3.23): las propiedades de RDF se pueden entender como atributos de los recursos y en este sentido corresponden a los pares tradicionales de atributo-valor (3.24). Además estas propiedades también representan las relaciones entre los distintos recursos de información.



Figura 3.2

Por lo tanto, según (3.25) el modelo de datos que propone RDF consiste en tres tipos de objetos (*Figura 3.2*):

- Recursos: cualquier objeto web identificable unívocamente por un URI, es decir, un identificador uniforme de recursos como un URL. Un recurso puede ser un documento HTML; una parte de una página web como por ejemplo un elemento HTML o XML dentro de un documento fuente, una colección de páginas, un sitio web completo; y en síntesis, cualquier recurso entendido como objeto de información.
- Propiedades: son aspectos específicos, características, atributos o relaciones utilizadas para describir recursos. Cada tipo de propiedad tiene sus valores específicos, define los valores permitidos, los tipos de recursos que puede describir y las relaciones que existen entre las distintas propiedades.
- Descripciones: como se puede ver en la *Figura 3.2*, consisten en la combinación de un recurso, una propiedad y el valor de dicha propiedad. Estas partes son conocidas como sujeto, predicado y el objeto respectivamente.

La sintaxis básica, como se dijo anteriormente, es la de XML 1.0. Además se pueden distinguir dos tipos de construcciones sintácticas para codificar RDF: por un lado, la serializada, que expresa de una forma muy regular todas las capacidades de un modelo de datos RDF, y por otro, la sintaxis abreviada, que incluye construcciones adicionales.

Sin embargo, el modelo de datos y la sintaxis, no facilitan los mecanismos para definir las propiedades ni las relaciones entre predicados y otros recursos o sujetos; por ello se ha establecido también una especificación para definir los esquemas (3.26). Un esquema RDF es un conjunto de información relativa a las clases de recursos que sirve para explicitar las relaciones jerárquicas que se establecen entre ellos, o bien para matizar el carácter obligatorio u opcional de las propiedades y otras restricciones como el número de ocurrencias, etc.

3.3.4.4.2 Aplicaciones

Por todo lo mencionado en la sección anterior, las aplicaciones o proyectos actuales del RDF no sólo responden a contextos bibliotecarios o del estricto mundo de la investigación, sino que en muchos casos, parten del ámbito del software que gira en torno a Internet.

Dentro de los proyectos del contexto bibliotecario, merecen mención especial:

- AGORA⁵ del Göttingen Digitization Zenter (GDZ) que ha elegido RDF/XML como formato de metadatos por defecto para desarrollar una biblioteca digital. El GDZ ha seleccionado a RDF como modelo de la metainformación de sus recursos, para soportar la interoperabilidad entre diferentes plataformas y distintos formatos de metadatos, previendo la distribución de su colección digitalizada.
- MANTIS⁶, un proyecto del OnLine Computer Library Center (OCLC) —estrechamente relacionado con CORC (Cooperative Online Resource Catalogue)— para construir sistemas de catalogación basados en la web que emplean distintas interfaces y

⁵AGORA <<http://hosted.ukoln.ac.uk/agora>>

⁶MANTIS <<http://purl.oclc.org/mantis>>

definiciones de metainformación. Utiliza RDF como modelo estándar para codificar e intercambiar formatos de metadatos diferentes.

Se pueden encontrar también, importantes proyectos de aplicación de este formato en iniciativas como por ejemplo la propia de Mozilla-Netscape, que ha presentado el software Aurora, la próxima generación del software cliente desarrollado para integrar información de Internet en el PC, y que tendrá las ventajas de una infraestructura estándar RDF lo que permitirá un simple mecanismo de organización, descripción y navegación por la web; asimismo IBM, está trabajando con RDF en su Java Central Station⁷, un buscador global de recursos Java, cuyo robot de búsqueda en la web usa RDF para describir las colecciones de datos que recopila.

Los objetivos del Resource Description Framework son amplios, y las oportunidades potenciales que ofrece lo son aun mas. Esto conlleva a augurar su éxito e implantación, ya que en este caso hay un entusiasmo implícito por parte de los máximos exponentes en el mundo del software cliente para la web —Netscape y Microsoft— en desarrollar y adoptar esta infraestructura de descripción de recursos.

3.3.4.5 Comparación entre XML y RDF

Mientras que XML es un lenguaje para modelar datos, RDF es un lenguaje para especificar metadatos. XML falla en la escalabilidad de los datos puesto que el orden de los elementos es antinatural y su mantenimiento es muy difícil y costoso, por el contrario, RDF permite la interoperabilidad entre aplicaciones que intercambian información comprensible por el navegador, para proporcionar una infraestructura que soporte actividades de metadatos.

3.3.5 Conjunto de Metadatos “Dublin Core”

Ante el advenimiento de la información digital, el estándar de Metadatos Dublin Core (DCMI) se ha convertido en un simple pero eficaz conjunto de elementos que sirven para describir una amplia gama de recursos de Internet. Actualmente es la iniciativa de catalogación más extendida en el mundo electrónico, al tiempo que es considerada un estándar internacional (ISO-15836-2003).

Existe una serie de características propias de este estándar, las cuales pueden resumirse en :

⁷IBM: Java Central Station <<http://www.ibm.com/developerworks/java/>>

- Simplicidad: Puede ser utilizado tanto por bibliotecarios como por cualquier autor que desee describir sus documentos y aumentar su visibilidad.
- Interoperabilidad Semántica: Contiene un conjunto de descriptores que permiten la unificación con otros estándares de datos.
- Reconocimiento Internacional.
- Extensibilidad: Permite la elaboración de descripciones de modelos tales como el MARC completo. Cuenta también con suficiente flexibilidad y extensibilidad para limitar la estructura, además de una semántica más elaborada y un amplio estándar de descripción.
- Flexibilidad: Nada es obligatorio, todos los elementos son opcionales y repetibles, así el usuario elige la profundidad de una descripción.

La norma Dublin Core (DC) promueve dos niveles de codificación: simple y cualificado. El Dublin Core simple comprende quince elementos; el Dublin Core cualificado implica el mismo número de elementos más un subgrupo de éstos denominados cualificadores, que refinan la semántica de los primeros a fin de recuperar y localizar de mejor modo los recursos en Internet.

Cada elemento del conjunto es opcional y repetible, y pueden clasificarse en tres tipos: los que tienen que ver con el contenido del recurso, los referentes a la propiedad intelectual y los relacionados con la creación e identidad del material, tal como aparece en la tabla a continuación.

Elementos Dublin Core Simple	
<p><i>Contenido</i></p> <ul style="list-style-type: none"> • Título • Tema • Descripción • Fuente • Lengua • Relación • Cobertura 	<p><i>Propiedad intelectual</i></p> <ul style="list-style-type: none"> • Creador • Editor o editorial • Colaborador • Derechos <hr/> <p><i>Creación e identidad</i></p> <ul style="list-style-type: none"> • Fecha • Tipo • Formato • Identificador

Tabla 3.1: Elementos Dublin Core Simple

Comúnmente los repositorios institucionales utilizan el esquema de metadatos DC para describir el contenido de sus objetos, estándar que se ha generalizado en la medida que se ha vuelto indispensable para cumplir el protocolo OAI-PMH, el cual se describirá más adelante, dado que promueve la interoperabilidad entre repositorios estructurados, debido a que es un formato aceptado globalmente.

Lo interesante de la Iniciativa de Metadatos Dublin Core es que permite establecer formas normalizadas para matizar cada uno de sus elementos a partir del uso y promoción de esquemas de codificación y vocabularios. Sin embargo, DC sigue presentando cierta ambigüedad al momento de codificar información en elementos como Título, Creador, Colaborador y Editor, que curiosamente no presentan ningún esquema que ayude a la codificación y asignación de los metadatos.

3.4 Bibliografía

- (3.1) Caplan, P. (1995). You call it corn, we call it syntax-independent metadata for document-like objects. *The public access computer systems review*, 6 (4), 19-23
- (3.2) Xu, Amanda. (1997). Metadata conversion and the library OPAC. IFLA. Disponible en: <http://archive.ifla.org/documents/libraries/cataloging/metadata/xu.pdf> [Consulta: 2012-08-29]
- (3.3) Younger, J. (1997). Resources description in the digital age. *Library trends* 45, 462-81
- (3.4) Dempsey, L. & Heery, R. (1998). Metadata: a current view of practice and issues. *Journal of Documentation*, 54 (2), 145-172
- (3.5) Wendler, R. (2000). Diversificación de actividades: habilidades y funciones catalográficas en la era digital. En: F.F. Martínez Arellano & L. Escalona Ríos (Compos.). *Internet, metadatos y acceso a la información en bibliotecas y redes en la era electrónica* (pp. 36-48). México: Universidad Nacional Autónoma.
- (3.6) Tennant, R. (1998). 21^o Century cataloguing. *Library Journal*, 123 (7), 30-31
- (3.7) Milstead, J. y Feldman, S. (1999). Metadata: cataloging by any other name. *Online*, 23 (1), 24-31
- (3.8) Gorman, M. (2000). ¿Metadatos o catalogación? Un cuestionario erróneo. En: F.F. Martínez Arellano & L. Escalona Ríos (Compos.). *Internet, Metadatos y acceso a la información en bibliotecas y redes en la era electrónica* (pp., 1-20). México: Universidad Nacional Autónoma.
- (3.9) Taylor, A. (2004). *The Organization of Information*. (2^o ed.) Englewood, CO: Libraries Unlimited.
- (3.10) Swick, R. (1997). Metadata: A W3C Activity. W3 Consortium. Disponible en: <http://www.w3.org/Metadata/Activity.html> [Consulta: 2012-08-29]
- (3.11) Lazinger, S. (2001). *Digital preservation and metadata: history, theory, practice*. Englewood, Colorado: Libraries Unlimited.
- (3.12) Daudinot Fournier I. (2006). Organización y recuperación de información en Internet: teoría de los metadatos. *ACIMED*; 14(5). Disponible en: http://bvs.sld.cu/revistas/aci/vol14_5_06/aci02506.htm
- (3.13) Garduño Vera, Roberto. "Organización de la información documental y su utilidad social." *La Información en el inicio de la era electrónica : Organización del conocimiento y sistemas de información*. México : UNAM, Centro Universitario de Investigaciones Bibliotecológicas, 1998 V.1, p.48
- (3.14) Martínez Ortega, Patricia y Juárez Santamaría, Beatriz. *Los Metadatos y la información digital*. Dirección General de Bibliotecas, Universidad Nacional Autónoma de México, Área de la Investigación Científica.

- (3.15) Library of Congress Network Development and MARC Standards Office
<http://www.loc.gov/marc/> [Consulta: 2012-08-29]
- (3.16) CORMENZA, Fernando. Normas y estructuras para automatizar la información, resumen sobre el protocolo Z39.5. p 2.
- (3.17) Carrión Gútiérrez, Alejandro. De las virtudes del catálogo virtual. Dossier 2. Boletín de la SEDIC. p 2-3. Disponible en: <http://www.sedic.es/z3950.pdf> [Consulta: 2012-08-29]
- (3.18) Arango, Marta Elena. El Z39.50 En el Ambiente de Transferencia y Recuperación de Información. Bogotá: Universidad pontificia Javeria
- (3.19) Benitez Sanchez, H., & Robayo Romero, F.S. Protocolo Z39.50 una Herramienta Importante en la Recuperación de Información, 2007. pp.1-17. Disponible en: <http://hdl.handle.net/10760/9556> [Consulta: 2012-08-29]
- (3.20) Baeza-Yates & Ribeiro-Neto: Modern Information Retrieval. p. 174.
- (3.21) Gradmann, Stefan. "Cataloguing vs. Metadata : old wine in new bottles ?", Paper of the 64th IFLA General Conference, 1998.
- (3.22) Miller, Eric. An introduction to the Resource Description Framework. D-Lib Magazine, 1998. Disponible en: <http://www.dlib.org/dlib/may98/miller/05miller.html> [Consulta: 2012-08-29]
- (3.23) Lagonze, Carl, Clifford A. Lynch, Ron Daniel. The Warwick Framework: A container architecture for aggregating sets of metadata. Networked Computer Science Technical Reference Library, 1996.
- (3.24) Tim Berners-Lee. Metadata architecture: document, metadata and links. Design Issues. World-Wide Web Consortium, 1997, rev. 1998. Disponible en: <http://www.w3.org/DesignIssues/Metadata.html> [Consulta: 2012-08-29]
- (3.25) World Wide Web Consortium. Resource Description Framework (RDF): Model and Syntax Specification. W3C Recommendation, Ora Lassila y Ralph R. Swich, eds, 1999. Disponible en: <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222> [Consulta: 2012-08-29]
- (3.26) World Wide Web Consortium. Resource Description Framework (RDF): Schema Specification: W3C Proposed Recommendation, Dan Brickley y R. V. Guha, eds, 1999. Disponible en: <http://www.w3.org/TR/1999/PR-rdf-schema-19990303> [Consulta: 2012-08-29]
- (3.27) Méndez Rodríguez, Eva. RDF: Un modelo de Metadatos flexible para las Bibliotecas Digitales del próximo milenio. Dpto. de Biblioteconomía y Documentación, Universidad Carlos III de Madrid. Disponible en: <http://rayuela.uc3m.es/~mendez/publicaciones/7jc99/rdf.htm> [Consulta: 2012-08-29]

Capítulo 4

Interoperabilidad

Capítulo 4 - INTEROPERABILIDAD

4.1 Introducción

El proceso de investigación constituye un esfuerzo internacional y distribuido, que implica una variedad de partes interesadas, tales como científicos, autores y beneficiarios de las subvenciones, centros de investigación, editoriales y agencias de financiación de la investigación, cada una con sus propios intereses.

Cada repositorio es un valor limitado para la investigación, por lo que el poder real del Acceso Abierto recae en la posibilidad de conectarse y establecer un lazo entre repositorios, surgiendo la necesidad de que exista interoperabilidad entre los mismos. Para crear una capa transparente de contenido a través de repositorios conectados alrededor del mundo, el Acceso abierto se basa en interoperabilidad, la capacidad para comunicar sistemas entre ellos e intercambiar información en un formato utilizable.

Lograr un acceso uniforme a los recursos en repositorios heterogéneos requiere tratar con el problema de la interoperabilidad de los metadatos.

En la actualidad existen muchas herramientas, técnicas y estándares de interoperabilidad, con variados potenciales para resolver estas heterogeneidades estructurales y semánticas que pueden existir entre los metadatos almacenados en diferentes repositorios.

Este capítulo tiene por objetivo establecer una visión general sobre el concepto de Interoperabilidad, a fin de mostrar cómo impacta en la prestación de servicios de un repositorio institucional. A continuación se introducen algunas herramientas y estándares de discutiendo sus principales características y el aporte que brindan en pos de resolver el problema mencionado.

4.2 Definición

Actualmente el término Interoperabilidad es comúnmente usado en la jerga de los profesionales de información para indicar que se están compartiendo recursos, productos y servicios de información especializada, presentes en sus sistemas de información, con otras personas y sistemas de información ubicados en cualquier parte del planeta. De esta forma se está evidenciando la construcción de una gran red global de conocimiento académico y científico que permitirá acercar el conocimiento humano a todas las personas y reducir la brecha digital ocasionada por la dificultad de acceder oportunamente a información actualizada y de calidad.

Si bien existen múltiples definiciones de la interoperabilidad, una de las más citadas, y que define la interoperabilidad de la información a un alto nivel conceptual, es la que propuso el IEEE (4.1) en 1990: “la habilidad de dos o más sistemas, redes de comunicación, aplicaciones o componentes para intercambiar información entre ellos y para usar la información que ha sido intercambiada”.

Por su parte, en la Interoperability Technical Framework (ITF), el gobierno australiano (4.2) define la interoperabilidad enmarcándola en el ámbito de las tecnologías de la información, como “la capacidad de transferir y utilizar información de una manera uniforme y eficiente a través de múltiples organizaciones y sistemas de tecnologías de la información. Permite asegurar el nivel de beneficios que recaudan las empresas, gobierno y la economía en general a través del comercio electrónico”.

De las definiciones recogidas se puede decir entonces que la interoperabilidad hace referencia a la relación entre dos o más recursos o sistemas (dispositivos hardware y de comunicación o componentes de software) de tecnologías de la información y la comunicación (TIC's) que precisan trabajar conjuntamente de forma fácil o automática. Asimismo, las definiciones coinciden en señalar dos capacidades con las que deben contar los sistemas implicados: comunicarse entre ellos (para poder transferir información) y entender la estructura de la información que se transfiere entre las entidades (para poder utilizarla).

De esta manera se puede observar que la interoperabilidad se manifiesta por:

- La capacidad de los sistemas para trabajar entre sí en tiempo real o programado.
- La capacidad de los programas para trabajar en diferentes sistemas.
- La capacidad de los datos para ser intercambiados entre diferentes sistemas (portabilidad).

Así en el ámbito de una institución, estudiantes e investigadores que buscan información no necesitan saber donde fue publicado un elemento específico o donde está almacenado un artículo. En cambio, los usuarios confían en los motores de búsqueda para recuperar los artículos, y son capaces de descubrir información que de otra manera habrían localizado.

Por otra parte, basarse en motores de búsqueda proporciona una unión de los repositorios entre sí, el Acceso Abierto no necesita depender de un único repositorio para recoger los resultados de la investigación en el mundo. La estructura descentralizada y el valor agregado de los servicios y herramientas que se han diseñado en la parte superior de los repositorios son posibles debido a la interoperabilidad. La calidad de estos servicios depende de la información proporcionada por los repositorios y la estandarización de los datos.

La interoperabilidad es el adhesivo técnico que hace esta integración posible, y permite que se logren alcanzar las metas del Acceso Abierto.

4.3 Tipos de Interoperabilidad

La interoperabilidad puede ser vista desde distintas perspectivas, que pretenden determinar los tipos, vertientes, enfoques o dimensiones de la interoperabilidad.

UKOLN (4.3), organización británica de investigación especialmente enfocada en informar prácticas e influenciar políticas en el área de las bibliotecas digitales, sistemas de información, gestión bibliográfica y tecnologías web, distingue un conjunto de dimensiones de la interoperabilidad (tecnológica, semántica, política y humana, intercomunitaria, legal, e incluso internacional) que abarcan distintas cuestiones necesarias para el intercambio de

información entre sistemas. Por su parte, el proyecto europeo LIFE⁸, financiado por la Comisión Europea área de Educación y Cultura, publicó en 2006 un informe detallado sobre la interoperabilidad de la educación en Europa, el cual adopta un marco semiótico para ayudar a entender los distintos aspectos de la interoperabilidad, que contempla varias capas de la interoperabilidad: física, empírica, sintáctica, semántica, pragmática y social.

En general, los distintos enfoques coinciden en señalar como dimensiones fundamentales la interoperabilidad técnica, la semántica, y la organizativa, asociando los aspectos sintácticos bien a la interoperabilidad técnica, bien agrupándolos con la semántica como aspectos de interoperabilidad conceptual. A continuación se describen los principales aspectos de estas tres dimensiones.

4.3.1 Interoperabilidad Técnica

Se corresponde con los modelos lógicos comunes y la capacidad de los sistemas de información para comunicarse e interactuar en ambientes heterogéneos. En general, la interoperabilidad técnica cubre los aspectos técnicos de la conexión y comunicación entre equipos, dispositivos informáticos y aplicaciones. Esto incluye aspectos claves como interfaces abiertos, servicios de interconexión, integración de datos y mediadores, presentación e intercambio de datos, localización y recuperación de recursos, accesibilidad y servicios de seguridad. En cada una de estas áreas se pueden identificar distintos estándares o especificaciones de uso extendido, como los siguientes ejemplos.

- Interconexión: precisa de políticas y tecnologías para conectar sistemas mediante protocolos de comunicación, como pueden ser HTTP, FTP, SMTP, SOAP, CORBA y otros protocolos de amplia aplicación en Internet.
- Intercambio de datos: se basa en tecnologías y estándares para la descripción de la estructura y codificación de los datos para permitir el intercambio de la información entre sistemas que pueden tener distintas formas de representación interna de los datos. Algunos de los estándares más utilizados en la actualidad son el lenguaje de marcado XML (eXtensible Mark-Up Language), sistemas universales de codificación de caracteres como Unicode, mecanismos para la transformación y presentación de documentos como el lenguaje XSL (eXtensible Stylesheet Language), el uso de extensiones consistentes como S/MIME (Secure/Multipurpose Internet Mail Extensions), entre otros.
- Presentación de la información: mediante el uso de formatos de ficheros estandarizados como TXT, PDF, JPEG, PNG, HTML, XML, etc., que puedan ser entendidos y utilizados convenientemente por sistemas heterogéneos para representar con exactitud información proveniente de múltiples fuentes.
- Metadatos para la descripción de procesos y datos: siguiendo mecanismos de descripción como el Modelo Entidad Relación empleado para estructurar la información en las bases de datos relacionales, o esquemas XML que también permiten dotar de una determinada estructura y contenidos a los distintos tipos de documentos
- Localización y recuperación de información: hacen referencia a los mecanismos de búsqueda y localización de recursos (servicio de directorio como DNS, o protocolos

⁸ Life Project: <<http://life.eun.org/>> [Consulta: 2012-08-14]

para la consulta de redes como LDAP), así como a los estándares de metadatos y vocabularios controlados que permiten la descripción consistente de recursos, como es el caso del marco de descripción RDF (Resource Description Framework).

Otros aspectos técnicos destacados son los dirigidos a la identificación única de recursos (como los URI, Universal Resource Identifier), o de elementos y atributos empleados en esquemas XML y otros esquemas de metadatos (como los espacios de nombres XML). Y por último, en relación con la seguridad, es importante tener en cuenta otras tecnologías y estándares para el encriptado de datos, firmas digitales y otros protocolos de transmisión segura de información.

4.3.2 Interoperabilidad Semántica

Se puede entender como la capacidad de los sistemas de información (Bibliotecas Digitales y Repositorios Institucionales) para intercambiar información basándose en un significado común de los términos y expresiones contenidos en los metadatos y documentos, con el fin de asegurar la consistencia, representación y recuperación de los contenidos.

Lograr la interoperabilidad semántica es considerado uno de los mayores retos en la integración de sistemas de información. Básicamente, esto se debe al hecho de que el significado puede cambiar para cada contexto y a lo largo del tiempo, y de que distintos requisitos en dominios diferentes dan como resultado distintos modelos de información. Mientras que la interoperabilidad técnica está próxima de ser alcanzada mediante estándares abiertos, el logro de la interoperabilidad semántica (así como de la organizativa) es más problemático, puesto que afecta a múltiples niveles, funciones y procesos de los sistemas de información.

La interoperabilidad semántica debe ser tenida en cuenta en dos contextos de comunicación: hombre-máquina, y máquina-máquina. En el segundo contexto, presenta distintas dimensiones como las que distingue el proyecto Qualipso⁹, en el que los sistemas de información pueden presentar: incompatibilidades estructurales y de representación en el modelado de conceptos, incompatibilidades lingüísticas como el uso del mismo término para designar un concepto distinto o el empleo de términos distintos para un mismo concepto, así como incompatibilidades conceptuales donde sistemas de información distintos emplean conceptos con un cierto grado de solapamiento en su significado.

Una vía para lograr la interoperabilidad semántica es contar con un modelo conceptual común (esquema de metadatos), en el que se describe la información que se intercambia, en términos de conceptos, sus propiedades y las relaciones entre estos conceptos. Asimismo, las propiedades de un concepto pueden tener distintos valores que requieren un entendimiento común, y para asegurar la correcta interpretación de los datos intercambiados entre los sistemas es necesario el uso de vocabularios controlados. Pero, para lograr el entendimiento de la información intercambiada de una forma más dinámica, el sólo uso de vocabularios controlados no es suficiente, siendo necesarios modelos de conocimiento más ricos

⁹ Qualipso: <<http://www.qualipso.org>>. [Consulta: 2012-08-29]

semánticamente, como taxonomías y ontologías, que definen los conceptos de un determinado dominio así como sus relaciones.

En este sentido, y aunque los esquemas de metadatos ya han sido referidos en relación con la interoperabilidad técnica, estos tienen una importancia fundamental en el logro de la interoperabilidad semántica, junto con los vocabularios, taxonomías y ontologías.

4.3.3 Interoperabilidad Organizativa/Pragmática

El proyecto LIFE considera que la interoperabilidad pragmática, que englobaría aspectos organizativos y políticos, se refiere a que las organizaciones y grupos deben compartir un mínimo de objetivos comunes, como podría ser un objetivo pedagógico común, y asumir las responsabilidades y consecuencias compartidas, para ser capaces de colaborar y construir servicios interoperables e intercambiar información.

En relación con la interoperabilidad pragmática, la interoperabilidad organizativa se refiere a la definición de objetivos y procesos de negocio y a la elaboración de modelos comunes para lograr la colaboración de las organizaciones que deseen intercambiar información, pero que puedan tener una organización y estructura interna diferente. Además, la interoperabilidad organizativa tiene por objeto atender las necesidades de la comunidad de usuarios proporcionando servicios disponibles, localizables, accesibles y orientados los mismos.

4.3.4 Relación entre los Tipos de Interoperabilidad

La interoperabilidad técnica puede ser considerada un prerrequisito para alcanzar la interoperabilidad semántica y organizativa. Los sistemas y componentes deben conectarse primero en lo físico, en el nivel de protocolos, antes de que se pueda establecer una interoperabilidad a nivel de datos y semántica.

Por su parte, la interoperabilidad semántica precisa de acuerdos a nivel técnico y organizativo. En primer lugar, cualquier medida tomada tiene que tener en cuenta aspectos de la interoperabilidad técnica, y ser construida en base a estándares, guías y soluciones dadas. En segundo lugar, como el significado de los datos depende de la finalidad o contexto en el que es empleado, las medidas en el contexto de la interoperabilidad semántica están estrechamente vinculadas, e incluso requieren e implican medidas en el contexto de la interoperabilidad organizativa.

De la misma manera, la interoperabilidad organizativa precisa de las dos anteriores. Por un lado, es necesario que se proporcione un mínimo nivel de seguridad en el nivel técnico para poder iniciar la interoperabilidad a nivel organizativo, definiendo políticas, estrategias y procedimientos. Y por otro, el entendimiento común, y el procesado e intercambio adecuado de los conjuntos de datos se basa en acuerdos sobre conceptos o su relación mutua. Estos acuerdos y el proceso de alcanzarlos sólo pueden ser afrontados desde un nivel organizativo.

4.4 Estándares de Interoperabilidad

Uno de los medios para superar la incompatibilidad de infraestructura y contenidos es la flexibilidad en las soluciones tecnológicas, a través de la implementación de metodologías comunes basadas en estándares.

Un estándar se puede definir como “un acuerdo documentado sobre especificaciones técnicas u otros criterios precisos para ser utilizados consistentemente como reglas, guías, O definiciones de características, para asegurar que materiales, productos, procesos y servicios se ajusten a sus propósitos”¹⁰.

En la práctica, un estándar implica el reconocimiento de un problema común; la reunión de consejeros y expertos; la discusión, revisión y acuerdos respecto a una tecnología; la publicación de las especificaciones y el desarrollo e implementación de las especificaciones en software, que al desarrollarse de forma cíclica aseguren la interoperabilidad.

En la actualidad, varios organismos se han preocupado por definir estándares, normas, metodologías, herramientas y especificaciones que faciliten el desarrollo tecnológico de los sistemas y su integración con otros, la gestión de los recursos, etc.; que repercutan en el almacenamiento, intercambio y búsqueda de los contenidos.

4.4.1 Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH)

OAI (Open Archives Initiative) es una iniciativa para desarrollar y promover estándares de interoperabilidad que faciliten la difusión de contenidos así como el intercambio de formatos bibliográficos entre distintos repositorios digitales. No está específicamente orientada a los contenidos educativos sino a cualquier contenido digital. La idea básica que fomenta OAI es crear una forma simple y sencilla de intercambiar información (concretamente metadatos) entre repositorios heterogéneos que alberguen cualquier objeto que contenga metadatos asociados.

Surgió a finales de los años noventa a partir de los servidores de documentos en acceso abierto que habían aparecido en distintas disciplinas científicas: arXiv en Física, RePEc en Economía, CogPrints en Psicología, NCSTRL en Informática y NDLTD para tesis. Su objetivo inicial fue estudiar la interoperabilidad de los distintos servidores con objeto de facilitar el intercambio de datos entre los mismos. El nacimiento de la iniciativa se sitúa en la Convención de Santa Fe celebrada en la ciudad norteamericana del mismo nombre en Octubre de 1999.

La iniciativa se concretó en el desarrollo del protocolo OAI-PMH (Open Archives Initiative - Protocol for Metadata Harvesting) para permitir el intercambio de estos metadatos entre repositorios, y cuya primera versión apareció en Enero de 2001. Aunque inicialmente se creó para ser aplicado a depósitos de documentos en acceso abierto, muy pronto se vio que podía implementarse sobre cualquier material almacenado en soporte electrónico que requiriese la comunicación de metadatos. Según (4.4), la importancia de OAI-PMH se puede resumir en una frase: “OAI-PMH está llamado a ser a las bibliotecas digitales lo que HTTP es

¹⁰ International Standard Organization (ISO)

hoy al web". Es importante matizar que OAI-PMH trata exclusivamente de la comunicación de metadatos, no de los textos completos de los documentos que son referenciados.

Se pueden destacar tres características fundamentales de este protocolo:

- **Simplicidad:** se concibió bajo la premisa de la sencillez. Conscientes de los problemas de implementación que habían tenido otras iniciativas anteriores como Z39.50 o Dients¹¹, los creadores buscaron una fórmula simple que estuviera al alcance de cualquier potencial implementador.
- **Normalización:** basado en estándares ampliamente utilizados en Internet como son el protocolo HTTP (HiperText Transport Protocol) para la transmisión de datos y órdenes y XML (eXtended Markup Language) para la codificación de los metadatos.
- **Recolección:** frente a otros sistemas de agregación de contenidos como la búsqueda distribuida (Z39.50) o los sistemas de sindicación de contenidos¹² vía RSS, OAI-PMH ha optado por la recolección de metadatos. En este caso, existe una entidad que pone a disposición de los interesados información bibliográfica sobre los documentos que almacena. Estos, normalmente agregadores de contenidos, recogen periódica y sistemáticamente todos o parte de los metadatos expuestos para, localmente, implementar servicios de valor añadido.

OAI-PMH sigue el principio de que existen múltiples proveedores de datos ("Data Providers") que comparten información con múltiples proveedores de servicios ("Service Providers") a través de un protocolo común. Los primeros son los depósitos de documentos que proporcionan los metadatos de los documentos que almacenan. Por otra parte, el rol de "Service Provider" implica recolectar información académica desde distintos repositorios, con el objetivo de incorporarle algún valor añadido y almacenarla en motores de bases de datos o indexadores que admitan una buena performance y un costo de mantenimiento/optimización adecuado. Entre los valores añadidos que se pueden ofrecer se encuentran: sistema de búsqueda e identificación, filtrado, alertas temáticas, medición del uso e impacto de los documentos, etc. El rol de Service Provider es la cara visible al usuario final, y de acuerdo a estudios internacionales hay una relación 1:5 entre la cantidad de Service y Data Providers, tal número muestra que representa una innovación en cuanto a los servicios que debe prestar una biblioteca digital y especialmente, una biblioteca digital temática.

Para minimizar los problemas derivados de conversiones entre múltiples formatos, OAI-PMH requiere que todos los proveedores de datos expongan sus recursos utilizando mínimamente el esquema Dublin Core sin calificar, descrito en el capítulo anterior. Además de este formato, cada servidor es libre de ofrecer los registros en otro/s formatos adicionales (como por ejemplo el formato MARCXML). Por lo tanto, el protocolo OAI-PMH puede

¹¹Dienst: Protocolo que permite la comunicación entre servicios que forman una biblioteca digital. Fue diseñado y probado con el sistema NCSTRL (Networked Computer Science Technical Report Library). Más información en: <http://www.cs.cornell.edu/cdlrg/dienst/protocols/DienstProtocol.htm> [Consulta: 2012-08-29]

¹²Sindicación de contenidos: Es la forma que tienen algunos portales de internet para distribuir sus contenidos de forma automática a aquellos usuarios que acceden frecuentemente a sus páginas. Este contenido se distribuye a través de ciertos canales temáticos que se pueden leer a través de programas llamados "lectores de noticias" (se conocen también como agregadores, lectores de canales, newsreaders o feed reader).

combinarse con otros protocolos y normas de bibliotecas digitales para facilitar un amplio rango de funcionalidades.

De esta forma, este protocolo se convierte en una opción viable y sencilla para que los proveedores de datos puedan poner sus metadatos a disposición de diferentes servicios de información, utilizando para ello estándares abiertos como el HTTP (Hypertext Transport Protocol) y XML (eXtensible Markup Language).

4.4.2 SWORD¹³

Antes de su aparición en el año 2007, no existía ninguna interfaz estándar para etiquetado, empaquetamiento o herramientas de autoría para cargar objetos en un repositorio, ni tampoco para la transferencia de objetos digitales entre repositorios. No había manera de realizar un depósito desde el exterior de un repositorio, ni de depositar en más de uno a la vez.

La ausencia de un estándar de depósito llevó a que JISC (Joint Information Systems Committee) propusiera una solución denominada SWORD (Simple Web-service Offering Repository Deposit). La misma se trata de un protocolo liviano diseñado para facilitar el depósito interoperable de recursos principalmente en repositorios, pero potencialmente en cualquier sistema en el cual se pretenda recibir contenido de fuentes remotas.

El acrónimo de las siglas que lo conforman es el siguiente:

- Simple: liviano, ágil y apropiado para sus fines
- Web-service (Servicio Web): independiente del software propietario, soporta interfaces estándar
- Offering (de Oferta): el cliente brinda contenido al servidor.
- Repository (Repositorio): o cualquier otro sistema en el cual se quiera depositar o recibir contenido
- Deposit (Depositar): poner, enviar, registrar o añadir, es un paso en el flujo del consumo.

Fue creado por las siguientes razones:

- Facilitar la interoperabilidad entre las aplicaciones.
- Simplificar el proceso de identificación, hallar la opción apropiada de contribución y colocación de metadatos mínimos.
- Intentar dotar a las herramientas comunes usadas por el usuario para la creación de materiales digitales, de las capacidades de contribución con los RI.

Casos de uso:

- Depósito desde una herramienta de escritorio o en línea
- Múltiples depósitos simultáneos
- Depósito de Máquina. Por ejemplo, depósito automático desde una máquina de laboratorio
- Migración / transferencia. Por ejemplo, a un servicio de preservación

¹³ SWORD <<http://www.swordapp.org/>> [Consulta: 2012-08-15]

- Depósito mediado. Por ejemplo, depositar a través de un representante designado, en repositorios adicionales

4.4.3 OpenSearch¹⁴

Es un conjunto de tecnologías que permiten publicar resultados de búsqueda en un determinado formato (estándar), asegurando que los motores de búsqueda y los resultados de búsqueda tengan interoperabilidad entre plataformas.

OpenSearch consiste en:

- Archivos descriptivos OpenSearch (en inglés OpenSearch Description Documents): Archivos en formato XML que identifican y describen un motor de búsqueda.
- OpenSearch Query Syntax: Contiene la información que los clientes requieren para realizar búsquedas.
- OpenSearch Response (Open Search 1.1): Es la respuesta que se genera a partir de una búsqueda. Además de la lista de resultados contiene información de paginación.
- Agregadores: Sitios que pueden mostrar resultados de las búsquedas realizadas mediante OpenSearch.

Este protocolo fue creado por A9¹⁵ (propiedad de Amazon) y publicada en su versión 1.0 en marzo de 2005, actualmente es soportado por muchos navegadores (principalmente Internet Explorer 7 y Firefox 2) como medio para la incorporación de nuevos motores de búsqueda. Es una tecnología fácil de implementar y hay varias soluciones, entre ellas utilizar el software de Open Search Server¹⁶ desarrollado bajo licencia de software abierto GPL v.3.

La ventaja de añadir OpenSearch a un sitio web es que esta tecnología lo hará más visible y útil en la red, en la medida que puede ser utilizado como motor de búsqueda por parte de las personas interesadas en su contenido. En el entorno de los repositorios institucionales, las publicaciones pueden ganar visibilidad y prestigio gracias a esta tecnología, minimizando además los tiempos de búsqueda.

Tiene poco sentido añadir OpenSearch a un sitio pequeño o de escaso contenido, pues será difícil encontrar material aprovechable en él, pero incluso en tales circunstancias, el administrador de un sitio puede añadirle OpenSearch para utilizarlo como recurso con el que desarrollar técnicas y estrategias de búsqueda de información en la red, con la ventaja añadida de que los resultados serán siempre perfectamente controlables y verificables.

4.5 Herramientas de Interoperabilidad

Desde los orígenes de OAI-PMH, se han ido creando una serie de herramientas software que soportan distintos aspectos, roles y funciones de la arquitectura del protocolo,

¹⁴ OpenSearch <<http://www.opensearch.org>> [Consulta: 2012-08-15]

¹⁵ A9 <<http://a9.com>> [Consulta: 2012-08-15]

¹⁶ OpenSearchServer <<http://www.open-search-server.com/>> [Consulta: 2012-08-15]

todas ellas con el objetivo final de facilitar la distribución de información científica y académica, así como de cualquier tipo de contenido electrónico, a través de la Web.

Para conocer estas herramientas resultan de gran utilidad los listados que ofrece tanto el sitio web de la propia Open Archives Initiative¹⁷ como el proyecto europeo Open Archives Forum (OAF)¹⁸. Los programas recogidos en estas fuentes se caracterizan por ser software libre y haber sido desarrollados por implementadores de la comunidad OAI, principalmente en el seno de proyectos de investigación en universidades pero también en centros de investigación, bibliotecas y consorcios bibliotecarios, proyectos de bibliotecas digitales e incluso por programadores individuales.

Dentro de estas aplicaciones se incluyen plataformas para la creación de repositorios desde su origen y que proporcionan una serie de funcionalidades básicas, como DSpace o EPrints (sección 4.5.1), como aplicaciones generadoras de proveedores de datos a partir de colecciones ya existentes (sección 4.5.2), y proveedores de servicios o recolectores metadatos (sección 4.5.3).

Para finalizar, se introducen aplicaciones que incorporan la funcionalidad del estándar OpenSearch (sección 4.5.4) e implementaciones del protocolo SWORD (sección 4.5.5).

4.5.1 Plataformas de Software para Crear Repositorios

Con el objetivo de lograr la interoperabilidad en un repositorio institucional, si bien se puede realizar una implementación propia de OAI-PMH (como es el caso de Celsius-DL, el cual se describirá más adelante), han aparecido una serie de programas que permiten a cualquier institución (universidad o centro de investigación) crear su propio repositorio de acceso abierto al mismo tiempo que hacerlo compatible con OAI-PMH.

Estos sistemas, una vez instalados y configurados en un servidor web determinado, deben ofrecer una serie de funcionalidades básicas para el mantenimiento y gestión del repositorio. Una de sus funcionalidades principales de gestión es el soporte al flujo documental (remisión de documentos; evaluación, aceptación o rechazo de documentos); la edición, revisión y evaluación de metadatos sobre los recursos; y la transformación de formatos de fichero, entre otras. En general, ofrecen interfaces de usuario, de autor y de administrador, y permiten la creación de diversos grupos de usuario y niveles de acceso. En cuanto al contenido, facilitan la creación de colecciones de materiales y generalmente aceptan múltiples formatos de archivo. Las funcionalidades relativas al uso del archivo se limitan a diversas opciones de búsqueda y navegación entre los registros del repositorio (palabras clave, búsqueda por campos de metadatos), así como funciones para su visualización y descarga.

La mayoría de los programas y sobre todo los más conocidos utilizan software libre, aunque también han aparecido empresas comerciales que también han desarrollado sus propios programas (como por ejemplo ProQuest, Innoate, etc.). Los programas más extendidos son:

¹⁷ Listado de herramientas compatibles con el protocolo OAI-PMH, recogidas por la Open Archives Initiative: <<http://www.openarchives.org/pmh/tools/tools.php>>. [Consulta: 2012-08-24]

¹⁸ Inventario de productos software para implementar archivos abiertos, de la Open Archives Forum: <<http://www.oaforum.org/resources/tvtools.php>>. [Consulta: 2012-08-24]



Figura 4.1: Principales programas para la creación de repositorios.

Proveedores de servicios:

- Eprints¹⁹, desarrollado por la Universidad de Southampton, es un software desarrollado en el seno del Open Citation Project dirigido por Stevan Harnad en la Universidad de Southampton (UK). Permite la ejecución centralizada (basada en disciplinas) de archivos de publicaciones académicas, así como también distribuida (basada en instituciones). Es compatible con OAI, es decir, los metadatos pueden ser cosechados desde los repositorios ejecutando el software con el protocolo de cosecha de metadatos OAI.
- DSpace²⁰: Es una plataforma nacida de una alianza entre HP y el MIT que permite gestionar archivos en cualquier formato, desde texto plano hasta video digital. Los archivos son indexados de modo tal que su búsqueda y recuperación vía web es extremadamente simple. A su vez, son preservados a largo plazo.
- CDSware²¹: CERN Document Server Software, permite ejecutar un servidor propio de preprint electrónicos. Implementa el protocolo OAI-PMH y utiliza MARC 21 como estándar bibliográfico subyacente.
- Fedora²²: Desarrollado por Cornell University, es un arquitectura de acceso abierto de repositorios digitales que permite el empaquetado de contenidos y servicios distribuidos asociados. Fedora soporta peticiones OAI-PMH sobre el contenido del repositorio.

Proveedores de datos:

- DSpace
- E-LIS²³: Es el proyecto más reciente dado que aun no se ha hecho público. Es un esfuerzo internacional para crear un archivo multinacional y multilingüe de documentos científicos en las áreas de Biblioteconomía y Documentación. Ha sido financiado parcialmente por el Ministerio de Educación.

4.5.2 Aplicaciones Generadoras de Proveedores de Datos

En esta sección se discute la situación actual en lo que a la construcción automatizada de proveedores de datos respecta.

¹⁹ Eprints University of Southampton <<http://www.eprints.org>>. [Consulta: 2012-08-29]

²⁰ Dspace <<http://www.dspace.org>>. [Consulta: 2012-08-29]

²¹ CDSware <<http://cdsware.cern.ch>>

²² Fedora <<http://fedoraproject.org>>. [Consulta: 2012-08-29]

²³ E-LIS (Eprints in Library and Information Science) <<http://eprints.rclis.org>>. [Consulta: 2012-08-29]

Se han desarrollado múltiples proyectos para proveer herramientas a administradores de bases de datos que permitan la construcción de servidores OAI de forma parcial o total.

A continuación, se describen algunas de las herramientas disponibles capaces de generar fácilmente proveedores de datos a partir de colecciones de documentos electrónicos ya existentes, organizadas en razón de la infraestructura previa de la colección. Los lenguajes de programación empleados para desarrollar estas aplicaciones son en su mayoría PHP, Java o Perl; así como tecnologías web ASP o JSP.

Entre los desarrolladores de las aplicaciones seleccionadas merece especial mención el Laboratorio de Investigación en Bibliotecas Digitales (DLRL) del Instituto Politécnico de Virginia (Virginia Tech), el proyecto OAI-PMH de la Biblioteca de la Universidad de Illinois en Urbana Champaign, así como el Laboratorio de Tecnologías Interactivas y Cooperativas (ICT) y el Centro de investigación en Tecnologías de Información y Automatización (CENTIA), de la Universidad de Las Américas Puebla, México. Estos grupos de investigación han creado varias herramientas generadoras de proveedores de datos, junto a otras herramientas que soportan otros aspectos del protocolo OAI-PMH. También se recogen herramientas desarrolladas por universidades, como las estadounidenses Old Dominion y Chapel Hill (en colaboración con el proyecto *ibiblio*), o la Universidad de Oldenburg (Alemania); el departamento de Investigación de Online Computer Library Center (OCLC Research) e incluso de empresas como ZZ/OSS Information Networking.

Para finalizar, en la sección 4.5.2.9 se resaltan y comparan las características más importantes de los proyectos citados.

4.5.2.1 Virginia Tech OAI²⁴ (VTOAI)

Esta herramienta, desarrollada en el lenguaje PERL, implementa solamente la estructura de un documento OAI-PMH vacío a partir de metadatos almacenados en bases de datos relacionales, dejando la generación del contenido a los usuarios. Es decir, cada usuario debe programar la transformación de los registros al formato de metadatos a exponer. Luego, VTOAI toma el contenido generado por el usuario y completa la respuesta.

Es un producto del laboratorio de investigación en bibliotecas digitales de Virginia Tech, el DLRL (Digital Library Research Laboratory), y no se actualiza desde junio de 2002.

Se requiere la intervención del usuario, el cual debe estar familiarizado con dicho lenguaje para editar ciertos módulos y especificar la forma de acceder a los metadatos almacenados en una base de datos local. Gracias a esto, es posible compartir cualquier colección local sin importar su estructura interna.

4.5.2.2 Open Archives In a Box (OAIB)

Fue creada en 2001 por DLT (Digital Libraries Technologies), grupo de tecnologías para bibliotecas digitales del NCSA (National Center for Supercomputing Applications)²⁵, en Estados Unidos.

²⁴ Virginia Tech OAI (VTOAI) <<http://www.dlib.vt.edu/projects/OAI/software/vtoai/vtoai.html>> Hussein Suleman, Junio 2002. [Consulta: 2012-08-15]

Esta herramienta permite la generación automática de servidores OAI para colecciones de metadatos almacenadas en una base de datos relacional. Le brinda al usuario la facilidad de configurar e instalar su servidor a través del uso de ciertas de herramientas que se distribuyen con la herramienta principal. Puede usarse en cualquier plataforma, puesto que está programada en Java.

Las desventajas de esta herramienta radican en que se debe de poder acceder a todos los metadatos de un registro (sin importar cuál sea el formato de metadatos a diseminar) a través de una sola consulta SQL. De lo contrario, se recomienda contar con una sola tabla que contenga todos los metadatos necesarios. Otra desventaja considerable se observa al momento de querer agregar un nuevo formato de metadatos, ya que el usuario necesita apegarse a una sintaxis establecida y un tanto limitada, provocando en determinadas ocasiones que sólo puedan definirse formatos de metadatos cuya estructura en XML sea lo más simple posible.

4.5.2.3 OAI Cat²⁶

OAI Cat es un producto de OCLC, quien lo distribuye con una licencia propia (OCLC Research Public License), y desde 2006 también mediante licencia Apache, versión 2.

Es una herramienta muy similar a VTOAI, en el sentido de que también genera el esqueleto o estructura principal de un servidor OAI. OAI Cat requiere de trabajo adicional pues es necesaria la intervención del usuario para especificar parámetros en un archivo de configuración, proceso que podría tornarse complejo si solamente conoce la estructura de la colección que desea compartir.

Para poder definir otros formatos de metadatos se requiere la intervención del usuario para programar la forma de generar esos nuevos formatos.

4.5.2.4 OAIbiblio

Es una aplicación implementada en PHP que permite generar de forma automática servidores OAI para colecciones almacenadas en una base de datos relacional y que específicamente usa MySQL como su manejador (4.5). OAIbiblio es un desarrollo de ibiblio: The Public's Library and Digital Archive, una colección de colecciones de Internet elaborada en colaboración por el Center for the Public Domain y la Universidad de Carolina del Norte en Chapel Hill.

Es requerida la instalación de PHP y de DOM (Document Object Model) para administrar los objetos XML. La principal desventaja de esta herramienta radica en el hecho de tener que configurar ciertos módulos que la instalación inicial de PHP no tiene configurados.

4.5.2.5 XMLFile

²⁵ El grupo DLT (Digital Libraries Technologies) <<http://www.dlt.its.psu.edu/>> [Consulta: 2012-08-15] forma parte del NCSA (National Center for Supercomputing Applications), un centro nacional de investigación supercomputación dependiente de la National Science Foundation (Estados Unidos) y con sede en la Universidad de Illinois en Urbana-Champaign.

²⁶ OAI Cat <<http://www.oclc.org/research/software/oai/cat.htm>> [Consulta: 2012-08-15]

Esta otra herramienta también está implementada en PERL y usa algunos módulos de VTOAI. Es una herramienta de utilidad para generar servidores OAI para colecciones de documentos XML almacenados en una estructura de directorios.

No soporta colecciones almacenadas en bases de datos y solamente utiliza la estructura de directorios para la definición de conjuntos. Además de eso, se requiere la intervención del usuario para modificar ciertas subrutinas definidas en PERL, ya sea para la configuración o para la definición de nuevos formatos de metadatos.

4.5.2.6 Rapid Visual OAI Tool²⁷ (RVOT)

Está desarrollada en Java y permite generar fácil y gráficamente servidores OAI para colecciones almacenadas en un sistema de archivos, lo que implica que no soporta bases de datos relacionales.

Los desarrolladores de esta herramienta recomiendan su uso para colecciones pequeñas, ya que la herramienta tiene que hacer una transformación del formato “nativo” de cada archivo almacenado al formato Dublin Core.

Aunque esta aplicación representa una excelente oportunidad para poder compartir cualquier archivo de una colección que se encuentra almacenado en el disco duro, RVOT tiene un problema de escalabilidad.

Los formatos “nativos” que soporta son: RFC1807, MARC y COSATI. Cada uno de estos tiene asociado un “traductor” para generar su correspondiente representación en el formato Dublin Core. Si se requiere manejar otros formatos de metadatos, es necesario que el usuario implemente su propio traductor, de tal manera que si se desean compartir colecciones que consistan de una gran cantidad de registros (miles por ejemplo), el proceso de especificación de registros se vuelve sumamente tedioso.

4.5.2.7 VOAI

Esta herramienta, creada en 2005 por el Laboratorio de Tecnologías Interactivas y Cooperativas (ICT) y el Centro de investigación en Tecnologías de Información y Automatización (CENTIA), de la Universidad de Las Américas Puebla, México, permite la generación automática de servidores OAI para colecciones almacenadas en cualquier base de datos relacional (4.6). No limita a tener que almacenar todos los metadatos en una sola tabla, ya que da libertad al usuario de indicar las consultas en SQL que se utilizarán para recuperar los metadatos.

Aleja al usuario de cualquier detalle de implementación y el código fuente que se genera puede ser utilizado en cualquier plataforma, ya que utiliza Java como lenguaje de programación.

²⁷ Rapid Visual OAI Tool (RVOT): <<http://rvot.sourceforge.net/>> [Consulta: 2012-08-15]

4.5.2.8 PHP OAI Data Provider²⁸

Es un desarrollo de la Universidad de Oldenburg, Alemania. Cumple completamente con el protocolo OAI-PMH 2.0, incluyendo la compresión de respuestas en XML lo que reduce significativamente el volumen de datos a transferir.

Es una aplicación de configuración sencilla y se puede conectar con varias bases de datos existentes empleando la capa de abstracción de bases de datos en PEAR²⁹.

Puede instalarse en plataformas Linux, Unix y Windows, requiere tecnologías Apache y PHP y soporta bases de datos relacionales como Oracle o MySQL.

Este proveedor de datos es utilizado por el repositorio de partituras musicales de la Universidad de Indiana (Indiana University Sheet Music)³⁰ así como 'IFEApub'³¹, repositorio de publicaciones del Instituto Francés de Estudios Andinos.

4.5.2.9 Comparación de Herramientas

Luego de introducir varias aplicaciones que permiten la creación de servidores OAI, se presenta un resumen a través de la siguiente Tabla, la cual provee una referencia rápida para comparar fácilmente las principales características de las aplicaciones antes mencionadas.

Herramienta	Soporta BD Relacionales	Soporta BD Relacionales Multitablas	Sin Modificación de la BD	No Requiere Implementación del Usuario
OAIB	✓	×	×	✓
VTOAI	✓	✓	✓	×
OAIBiblio	✓	✓	✓	×
RBOT	×	×	✓	✓
OAIcat	✓	✓	✓	×

Tabla 4.1: Comparación de Herramientas de Interoperabilidad

Como se puede ver, cada una de las aplicaciones mencionadas en este capítulo tiene sus propias características que, según el contexto, se traducen en ventajas y desventajas. De esta manera, la tabla puede interpretarse como sigue:

²⁸ PHP OAI Data Provider <<http://physnet.uni-oldenburg.de/oai/>>. [Consulta: 2012-08-27]

²⁹ PEAR Database Abstraction Layer <<http://pear.php.net/package/DB>>. [Consulta: 2012-08-27]

³⁰ Repositorio de partituras musicales de la Universidad de Indiana (Indiana University Sheet Music):

<<http://www.dlib.indiana.edu/collections/inharmony>>. [Consulta: 2012-08-27]

³¹ Repositorio de publicaciones del Instituto Francés de Estudios Andinos (IFEA), en: <<http://www.ifeanet.org/>>. [Consulta: 2012-08-27]

- OAIB: soporta bases de datos relacionales y no requiere de ningún tipo de implementación por parte del usuario. Sin embargo su principal desventaja es que no soporta bases de datos multitaslas, y por tal motivo es necesario hacerles ajustes a las bases de datos.
- VTOAI, OAIBiblio, OAICAT: soportan bases de datos relacionales multitaslas, y además tienen la ventaja de no requerir que se modifiquen las bases de datos, aunque sí requieren de una implementación por parte del usuario, lo cual constituye una desventaja.
- RVOT: su principal desventaja es la falta de soporte para bases de datos relacionales. Por otra parte, tiene la ventaja de que no requiere de implementación alguna por parte del usuario.

4.5.3 Recolectores y Proveedores de Servicios

Como se dijo anteriormente, un proveedor de servicios recolecta los metadatos expuestos por los proveedores de datos, es decir lleva a cabo cosechas OAI-PMH, y emplea los metadatos recolectados como base para la creación de servicios de valor añadido.

Un recolector de recursos digitales generalmente forma parte de una herramienta que provee servicios de valor agregado, recolectando metadatos desde diferentes repositorios (proveedores de datos) para que esta herramienta disponga su organización y uso en función de distintos intereses. Se actualiza de forma periódica, ampliando su base de datos de forma continua e incremental.

Entre los más conocidos recolectores se encuentran OAister³² y SDL³³ (se especializa en el área de la documentación). De igual forma se destaca Driver³⁴, el recolector de la red europea de repositorios de investigación.

A continuación se describe una serie de aplicaciones capaces de generar fácilmente proveedores de servicios a partir de sus recolectores correspondientes. La mayor parte de las herramientas generadoras de proveedores de servicios y los paquetes software recolectores han sido desarrolladas por los mismos creadores de aplicaciones para proveedores de datos, como el Laboratorio de Investigación en Bibliotecas Digitales (DLRL) del Instituto Politécnico de Virginia (Virginia Tech), el proyecto OAI-PMH de la Biblioteca de la Universidad de Illinois en Urbana Champaign, el grupo de Bibliotecas Digitales de la Universidad de Old Dominion o el departamento de Investigación de la OCLC (OCLC Research). Por otra parte, se introducen los aspectos principales de la iniciativa DRIVER.

En general, las herramientas recogidas en esta sección fueron desarrolladas en los comienzos de la iniciativa OAI, y en muchas ocasiones han dejado de mantenerse e incluso de distribuirse³⁵. Él ultimo en aparecer, y el más recientemente actualizado es el recolector del Public Knowledge Project, el cuál está siendo adoptado por un número cada vez mayor de proveedores de servicios.

³² OAister <<http://www.oaister.org/index.html>>. [Consulta: 2012-08-27]

³³ SDL (Search Digital Libraries In Library and Information Science) <<http://drtc.isibang.ac.in/sdl/>>. [Consulta: 2012-08-27]

³⁴ Driver <<http://www.driver-support.eu/en/index.html>>. [Consulta: 2012-08-27]

³⁵ Se han descartado herramientas como My.OAI que recientemente han dejado de mantener su sitio web y de distribuir el software recolector, que en este caso había sido desarrollado en Perl por la empresa FS Consulting, Inc., también desaparecida.

4.5.3.1 Arc

Es un recolector de metadatos desarrollado por y para Arc³⁶, un servicio de búsqueda federado, basado en el protocolo OAI-PMH, que permite buscar información en varios repositorios OAI desde una única interfaz. Incluye un recolector de metadatos que funciona con repositorios que cumplan cualquiera de las versiones del protocolo OAI-PMH, así como un motor de búsqueda con una sencilla interfaz para consultar una base de datos relacional en la que se han almacenado previamente los metadatos recolectados de los distintos repositorios.

Fue desarrollado por el Grupo de Bibliotecas Digitales de la Universidad de Old Dominion³⁷, y se basa en tecnologías Java requiriendo de un kit de desarrollo (JDK 1.4), un servidor web (Tomcat 4.0x), y un servidor de bases de datos relacionales (como Oracle o MySQL). Puede configurarse para una comunidad específica, y para ello es posible aplicar ciertas extensiones y adaptaciones a cada comunidad.

4.5.3.2 DP9

Al igual que Arc, fue desarrollado por el Grupo de Bibliotecas Digitales de la Universidad de Old Dominion, en colaboración con el Laboratorio Nacional de Los Álamos³⁸ en 2001.

Se trata de un servicio de código abierto que permite a motores de búsqueda indexar archivos OAI. Estos buscadores no son capaces de indizar los contenidos de un repositorio y no gestionan los ficheros XML con facilidad. Para resolver estas dificultades, DP9 asigna una URL persistente a los registros del repositorio, y los convierte en una consulta contra el repositorio apropiado cuando se requiere esa URL. No almacena los registros OAI, sino que tan sólo reenvía las consultas a los proveedores correspondientes, de manera que la calidad del servicio depende de la disponibilidad de estos proveedores. Se basa en la lista de proveedores del sitio web de la Open Archives Initiative.

Asimismo, soporta un servicio de conversión de nombres: dado un identificador OAI, este responde con una página HTML, un fichero XML o reenvía la consulta al proveedor de datos apropiado.

4.5.3.3 PKP Open Archives Harvester³⁹

Es un recolector y proveedor de servicios desarrollado en PHP por el proyecto Public Knowledge Project, una iniciativa conjunta de las universidades canadienses Simon Fraser y British Columbia con el apoyo de SPARC, con el objetivo general de fomentar la expansión y mejorar del acceso a la investigación. Con este fin, el proyecto ha desarrollado otras dos herramientas, Open Journal Systems (OJS) y Open Conference Systems (OCS), que permiten crear proveedores de datos OAI para colecciones de revistas y conferencias respectivamente.

A partir de su versión 2, PKP Harvester permite recolectar metadatos en una variedad de esquemas. Además, permite una recolección selectiva mediante el uso de conjuntos y

³⁶ ARC, disponible para su descarga en Sourceforge: <<http://oaiarc.sourceforge.net/>>. [Consulta: 2012-08-27]

³⁷ Old Dominion University - Digital Library Research Group <<http://dlib.cs.odu.edu/>>. [Consulta: 2012-08-27]

³⁸ Los Alamos National Laboratory <<http://lib-www.lanl.gov/>>. [Consulta: 2012-08-27]

³⁹ Public Knowledge Project (PKP) <<http://pkp.sfu.ca/?q=harvester>>. [Consulta: 2012-08-27]

rangos de fechas. De forma posterior a la recolección, se puede llevar a cabo un filtrado o normalización de los metadatos recopilados.

Presenta un interfaz flexible que permite una búsqueda simple y avanzada en todos los archivos recolectados, ya que realiza un mapeado o correspondencia de los campos de metadatos de los distintos esquemas que éstos emplean. Si además los repositorios comparten el mismo esquema se puede buscar por campos comunes.

Éste recolector resulta sencillo de instalar y configurar, empleando una base de datos MySQL o PostgreSQL, un servidor web Apache o Microsoft IIS, y pudiéndose instalar en cualquier sistema operativo.

4.5.3.4 Perl Harvester⁴⁰

Este recolector se desarrolló en lenguaje Perl por la DLRL (Digital Library Research Laboratory) de Virginia Tech, y su última versión es la 2.0, de agosto de 2002. Es un recolector orientado a objetos y no requiere ningún tipo de proceso de instalación, tan sólo precisa de los módulos estándar de Perl y de un servidor capaz de funcionar con módulos CGI. Una misma instancia puede ser utilizada fácilmente para recolectar metadatos de distintos sitios con propósitos diferentes. Todas las extensiones, configuraciones y contenedores se especifican empleando esquemas XML.

4.5.3.5 Net::OAI::Harvester⁴¹

Es una extensión en Perl, desarrollada por los programadores Ed Summers y Martin Emmerich (4.7), que permite consultar repositorios OAI-PMH de forma sencilla y recolectar metadatos de éstos.

En cuanto sus características, se destaca el empleo de un analizador o parser (XML::SAX) para dividir los resultados de una consulta OAI-PMH con posible respuesta larga. Al dividir el documento XML de la respuesta se crean una serie de objetos Perl, que se almacenan en el disco duro de forma seriada y facilitan que se utilice menos memoria de procesamiento. Además, los filtros XML::SAX permiten que cada implementador desarrolle sus propios paquetes para la indización de metadatos, de manera que se acepte que los mismos posean cualquier formato.

4.5.3.6 OAIHarvester2

Se trata de una aplicación Java que proporciona un recolector de metadatos de repositorios que cumplen OAI-PMH versión 1.1 y 2.0. Es un desarrollo de software libre de la OCLC⁴², que se distribuye de forma gratuita mediante licencia Apache versión 2.

OAIHarvester2 puede correr como una aplicación independiente o integrarse en aplicaciones existentes. Mientras que la antigua aplicación tenía un interfaz en Java que permitía el procesamiento de los registros recolectados sobre la marcha, la nueva versión

⁴⁰ Perl O-O Harvester <<http://www.dlib.vt.edu/projects/OAI/software/harvester/harvester.html>>. [Consulta: 2012-08-27]

⁴¹ Net::OAI::Harvester <<http://search.cpan.org/~thb/OAI-Harvester-1.15/lib/Net/OAI/Harvester.pm>>. [Consulta: 2012-08-27]

⁴² OCLC: <<http://www.oclc.org/>>. [Consulta: 2012-08-27]

simplemente agrupa las respuestas OAI en un archivo XML para su posterior procesamiento (p. ej. con XSLT), haciéndola mucho mas ligera.

4.5.3.7 Iniciativa DRIVER

La iniciativa DRIVER (Digital Repository Infrastructure Vision for European Research) tiene como principal objetivo recolectar los contenidos digitales de los repositorios abiertos de investigación de ámbito europeo, posibilitando la creación de servicios globales de búsqueda y localización de contenidos.



Figura 4.2: Driver

Los aspectos a destacar de esta iniciativa son:

- Está basado en estándares existentes para recolectar recursos almacenados en repositorios abiertos: protocolo OAI-PMH.
- Intenta aportar visibilidad y accesibilidad a fondos de alcance europeo.
- Las directrices DRIVER indican los requerimientos que deben cumplir los repositorios que desean ser recolectados, pero además sirven de guía a los administradores de nuevos repositorios para definir su política de administración de datos y a los administradores de repositorios existentes en pasos a seguir para obtener un servicio de mayor calidad.

Según el grado de conformidad con las directrices, el estado del repositorio se considerará como validado (cumple los puntos obligatorios) o con futuro (cumple además puntos recomendados).

Las directrices DRIVER (4.8) se elaboran a partir de la experiencia práctica y de otras directrices existentes a nivel internacional: HAL (Francia), DARE (Países Bajos), Certificado DINI (Alemania) o SHERPA (Reino Unido).

Los componentes principales de DRIVER son:

- Recursos textuales: contempla los requerimientos que deben cumplir los recursos. Se destaca la obligación de utilizar conjuntos (“sets”) que definan las colecciones accesibles a texto completo.
- Metadatos: marca y define los elementos Dublin Core obligatorios y recomendados.
- Implementación de OAI-PMH: define características obligatorias y recomendadas para solucionar problemas en distintas implementaciones de repositorios de acceso abierto.

DRIVER ofrece un estándar de compatibilidad e interoperabilidad que facilita que el contenido de un repositorio sea más fácilmente recuperable y visible desde la propia institución y desde la comunidad científica y el conjunto de la sociedad.

4.5.4 Aplicaciones que Incorporan OpenSearch

Como se dijo anteriormente, OpenSearch es una tecnología fácil de instanciar, y existen actualmente varias aplicaciones que lo incluyen dentro de sus funcionalidades. Entre ellas se pueden mencionar las siguientes:

- Mediawiki Extension: OpenSearch⁴³: permite a MediaWiki enviar resultados de una búsqueda a través del formato OpenSearch, así como anunciar la existencia de OpenSearch a través de un archivo de descripción.
- Drupal OpenSearch Results⁴⁴: añadiendo la capacidad de OpenSearch, las búsquedas se exportan como palabras clave y pueden ser descubiertas automáticamente a través de archivos OpenSearch XML de descripción.
- Pyopensearch⁴⁵: ejemplo de aplicación Python que implementa las características básicas de OpenSearch 1.1 junto con html/javascript para navegadores web, de manera tal que los formularios HTML puedan hacer uso de las funciones de ayuda de OpenSearch (sugerencias/autocompletado/buscar tal cual se ingresa).
- WordPress OpenSearch⁴⁶: permite añadir fácilmente la funcionalidad RSS de OpenSearch en un sitio de WordPress 2.0.
- PLOS (Plone OpenSearch)⁴⁷: agrega la posibilidad de generar resultados OpenSearch compatibles con una búsqueda a un sitio Plone⁴⁸.

4.5.5 Implementaciones de SWORD

⁴³MediaWiki Extension:OpenSearch <<http://www.mediawiki.org/wiki/Extension:OpenSearch>>. [Consulta: 2012-08-27]

⁴⁴Drupal OpenSearch Results <<http://drupal.org/project/opensearch>>. [Consulta: 2012-08-27]

⁴⁵Pyopensearch <<http://code.google.com/p/pyopensearch/>>. [Consulta: 2012-08-27]

⁴⁶WordPress OpenSearch, v1.1 <<http://www.williamsburger.com/wb/archives/opensearch-v-1-1>>. [Consulta: 2012-08-27]

⁴⁷Plone Open Search (PLOS) <<http://plone.org/products/plos>>. [Consulta: 2012-08-27]

⁴⁸Plone CMS: Open Source Content Manager <<http://plone.org/>>. [Consulta: 2012-08-27]

Dentro de los límites del proyecto SWORD han surgido numerosas herramientas, que van desde clientes de depósito de escritorio en línea y fuera de línea, hasta implementaciones de clientes Java.

- Servidor Java Genérico: es un Servlet Java creado para permitir que cualquier sistema Web escrito en Java pueda agregar una interfaz SWORD fácilmente. El Servlet SWORD maneja las interacciones con el usuario, pero delega el manejo de la autenticación, las solicitudes de servicio de documentos y depósitos de paquetes en el repositorio. El repositorio envía mensajes de vuelta al Servlet SWORD para remitirlo al usuario en un formato apropiado. Como se trata de acciones estándar que se realizan en un repositorio, integrarlo con este Servlet resulta muy sencillo.
- DSpace: La implementación DSpace SWORD ha sido escrita como un 'add-on' para DSpace haciendo uso del nuevo sistema modular de construcción introducido en la versión 1.5 de DSpace. Esto permite que el módulo SWORD sea instalado con facilidad junto a una instalación existente de DSpace. Una vez instalado, se utiliza un Servidor Java Genérico de SWORD, el cual interactúa con DSpace. Debido a la estructura jerárquica de comunidades y colecciones utilizada en DSpace, las respuestas a las solicitudes de servicio de documentos son fáciles de cotejar. Cualquier colección en la que un usuario pueda depositar está incluida en el documento de servicio devuelto. En el caso de colecciones con una licencia predefinida asociada, SWORD la respeta, asociándola de forma automática a los documentos depositados.
- Fedora: Al igual que con DSpace, Fedora hizo uso de la implementación genérica de SWORD en Java. La aplicación es muy flexible en cuanto a los tipos de objeto que acepta para su depósito.
- Eprints 3: eprints introdujo una API jerárquica en la versión 3, lo que permite que el sistema sea fácil de personalizar. Debido a esto, la instalación de SWORD es simplemente una cuestión de copiar algunos archivos Perl en el directorio EPrints y puede ser fácilmente activado o desactivado mediante la edición de un archivo de configuración de Apache.
- arXiv 1.3-compliant endpoint⁴⁹: La API de depósito arXiv SWORD permite la presentación programática de material para su incorporación en la base de datos de arXiv alojada en arXiv.org. Con el fin de mantener simple la implementación del lado del cliente, arXiv no requiere ni apoya formatos complejos de empaquetado.
- OfficeSWORD⁵⁰: es un plugin que permite cargar documentos de Office en un repositorio directamente desde las aplicaciones Office utilizando el protocolo SWORD.

4.6 Bibliografía

(4.1) Institute of Electrical and Electronics Engineers. IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries. New York: IEEE, 1990.

(4.2) Australian Government Information Management Office. Interoperability Technical Framework for the Australian Government, Version 2, Junio 2003. Disponible en: <http://www.agimo.gov.au/publications/2005/04/agtifv2>. [Consulta: 2012-08-14]

(4.3) UKOLN. Interoperability Focus: About. Última actualización: 03-07-2006. Disponible en: <http://www.ukoln.ac.uk/interop-focus/about>. [Consulta: 2012-08-14]

⁴⁹ arXiv <http://arxiv.org/help/submit_sword>. [Consulta: 2012-08-27]

⁵⁰ OfficeSword <<http://officesword.codeplex.com/>>. [Consulta: 2012-08-27]

(4.4) Barrueco Cruz, José Manuel; Subirats Coll, Imma. Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH): descripción, funciones y aplicación de un protocolo. El Profesional de la Información, Vol. 12, No 2, pp. 99-106, 2003. Disponible en: <http://hdl.handle.net/10760/4093>. [Consulta: 2012-08-14]

(4.5) Burton, A. 2004. OAIbiblio: PHP OAI-PMH data provider

(4.6) Villegas, A. 2005. Voai: un generador automático de servidores de metadatos bajo el protocolo OAI-PMH. B. Eng. Thesis. Dept. of Comp. Sysys. Eng., UDLA Puebla. (Español).

(4.7) SUMMERS, Ed (2004). "Building OAI-PMH Harvesters with Net::OAI::Harvester". Ariadne, n. 38. Disponible en: <http://www.ariadne.ac.uk/issue38/summers>. [Consulta: 2012-08-27]

(4.8) Digital Infrastructure Vision for European Research, 2008. Directrices Driver 2.0: Directrices para proveedores de contenido – Exposición de recursos textuales con el protocolo OAI-PMH. Disponible en: http://www.driver-support.eu/documents/DRIVER_2_0_Guidelines_Spanish.pdf. [Consulta: 2012-08-14]

(4.9) Barrueco Cruz, José Manuel; Caballos Villar, Almudena; Campos Rodríguez, Ángeles; Casaldàliga, Núria; Combarro Felpeto, Pilar; Cívico Martín, Rafaela; Domènech Luisa; García Gil, Ma Angeles; Losada, Marina; Morillo Moreno, José Carlos. "Guía para la evaluación de repositorios institucionales de investigación". FECYT, RECOLECTA y CRUE, 2010. Disponible en: <http://www.recolecta.net/buscador/documentos/GuiaEvaluacionRecolectav1.0-1.pdf>. [Consulta: 2012-08-14]

(4.10) Julio Alonso Arévalo, Imma Subirats Coll y M.ª Luisa Martínez Conde Gijón. "Informe APEI sobre acceso abierto". Asociación Profesional de Especialistas en Información, 2008.

Capítulo 5

Servicio de Difusión de la Creación Intelectual (SeDiCI)

Capítulo 5 - SERVICIO DE DIFUSIÓN DE LA CREACIÓN INTELECTUAL (SeDiCI)

5.1 Introducción

Se presenta el ambiente sobre el cual se trabajará, incluyendo una selección de un conjunto de repositorios digitales que aportan registros de contenido al repositorio digital de SeDiCI (Servicio de Difusión de la Creación Intelectual). Se presenta una breve descripción de la herramienta de cosecha que se utiliza en el repositorio, orientada al manejo y la producción de datos y metadatos.

5.2 Contexto

El Servicio de Difusión de la Creación Intelectual (SeDiCI) es el repositorio institucional de la Universidad Nacional de La Plata (UNLP). Por lo tanto, tal como se estudió en capítulos anteriores, el hecho de que sea un repositorio institucional significa que alberga, preserva y difunde mundialmente, a través de su sitio web, las creaciones y producciones intelectuales, científicas y artísticas de los diversos actores de la universidad (alumnos, profesores e investigadores). Además de contener, como lo hacen las bibliotecas digitales, referencias de catalogación de cada uno de los recursos, también ofrece, en una buena parte de su acervo y en número creciente en los últimos años, el texto completo de los mismos. El acceso a estos textos es completamente abierto y gratuito, debido a que adhiere a las políticas del *Open Access*.

Entre los servicios que presta se pueden destacar las búsquedas de diferentes tipos (inclusive en repositorios externos⁵¹), asesoramiento sobre derechos de autor, y principalmente el servicio de auto-archivo, a través del cual el autor (ya sea alumno, docente o investigador de la universidad) sube sus trabajos y se asegura la publicación y difusión de sus trabajos en forma rápida y sencilla.

SeDiCI utiliza un formato de metadatos propio, internacionalizable, y los transforma al formato Dublin Core para que sean accedidos vía OAI.

⁵¹ Actualmente SeDiCI cuenta con acceso a 30 repositorios digitales de diversas partes del mundo y de muy variadas temáticas: <<http://sedici.unlp.edu.ar/search/searchoai.php>>. [Consulta: 2012-08-29]



Qlr

SeDiCI es el Repositorio Institucional de la Universidad Nacional de La Plata creado para albergar, preservar y dar visibilidad a las producciones de las Unidades Académicas de la Universidad.

Navegue por nuestras colecciones

<p>Tesis</p>  <p>Tesis de grado, post-gradado y otros documentos</p>	<p>Revistas</p>  <p>Publicaciones en revistas científicas</p>	<p>Eventos</p>  <p>Ponencias realizadas en congresos y conferencias</p>	<p>Libros</p>  <p>Libros digitalizados y e-books</p>	<p>Red UNCI</p>  <p>Artículos y ponencias de la Red UNCI</p>
---	--	--	--	---

Noticias

- 14/07 Convocatoria BIREDIAL-SIBD-CIPECC 2012
 - 10/07 Presentación SNRD
 - 10/07 Artículo de Peter John
 - 02/07 Entrevista a Marisa De Giusti
 - 27/06 Conferencia Internacional Acceso Abierto, Comunicación Científica y Preservación Digital
 - 15/06 Webinar auspiciado por NECOBELAC y RSP
 - 05/06 Resumen del Foro "Las Universidades Latinoamericanas frente a los rankings"
 - 28/05 SeDiCI en "Científicos Industria Argentina"
 - 16/05 WSIS Forum 2012: taller temático
 - 11/05 Subsidios de apoyo a la edición de revistas científicas de la UNLP
- [Ver todas las noticias](#)

Últimos documentos agregados

- Assessment scheme-based service selection for SOC-based applications
- Aspectos esenciales de condicionales para los contratos sensibles al contexto
- Análisis comparativo de estimación de esfuerzo en el desarrollo de software
- Ambiente para la ayuda a la mejora de procesos en las PyMEs
- Algunas consideraciones para la transformación de Semántica de Negocios SBVR en el Lenguaje de Ontologías Web OWL2

Navegar por

Tipo de documento

- Artículo (7999)
- Capítulo de libro (2)
- Comunicación (747)
- Contribución a revista (545)
- Documento de trabajo (264)
- Documento institucional (2)
- Imagen en movimiento (52)
- Imagen fija (9)
- Informe técnico (33)
- Libro (56)
- Objeto de conferencia (2674)
- Preprint (7)
- Revisión (1041)
- Tesis de doctorado (2469)
- Tesis de grado (873)
- Tesis de maestría (338)
- Trabajo de especialización (128)

Fecha de publicación

- 2000 - 2012 (11504)
- 1900 - 1999 (5618)
- 1800 - 1899 (6)
- 1794 - 1799 (7)

Materia

- Humanidades (4829)
- Ciencias Naturales (2505)
- Farmacia (2000)
- Letras (1750)
- Ciencias Económicas (1221)
- Ciencias Informáticas (1172)
- Filosofía (845)
- Ciencias Jurídicas (829)
- Antropología (661)
- Zoología (652)
- ... Ver más

Autor

- Lopez, Hugo Luis (59)
- Gasparini, Leonardo (52)
- Ponte Gomez, Justina (50)
- Porto, Alberto (50)
- De Giusti, Marisa Raquel (46)
- Baran, Enrique Jose (42)
- Dipierri, Jose Edgardo (40)
- Mandrile, Eloy L. (40)
- Consani, Norberto E. (39)
- Oyhenart, Evelia Edith (39)
- ... Ver más

Palabra Clave

- Literatura (1341)
- Historia (1047)
- Economía (817)
- Educación (730)
- Filosofía (687)
- Informática (595)
- Farmacología (527)
- Herpetología (445)
- Reptiles (425)
- Antropología biológica (362)
- ... Ver más



Figura 5.1: Portal SeDiCI

Además, contempla una gran diversidad de tipos documentales, cada uno de los cuales posee una convención de carga (entrada de metadatos) propia. La tipología documental abarca desde Tesis (de grado y de postgrado), artículos de publicaciones periódicas, documentos multimediales (sonidos e imágenes), libros electrónicos, proyectos de investigación, etcétera.

A partir de mayo de 2012 SeDiCI ha cambiado el software para la gestión de su acervo y ha comenzado a utilizar la plataforma DSpace, la cual fue definida en el capítulo 4. Actualmente, DSpace es una de las plataformas más utilizadas por los repositorios institucionales más importantes en el mundo, por lo que esta migración representa una gran oportunidad para que SeDiCI evolucione y se siga posicionando como uno de los principales repositorios de Latinoamérica.

Más allá del cambio de interfaz, se le adicionaron a SeDiCI nuevas funcionalidades. El cambio más importante radica en la gestión y recuperación de la información. DSpace cuenta con un ordenamiento jerárquico de los archivos que los dispone en “comunidades” y dentro de éstas en “colecciones”. Así, dentro de las comunidades pueden encontrarse las unidades académicas que componen la UNLP o los eventos (congresos, simposios, encuentros, etc.) y dentro de ellos habrá colecciones de tesis, revistas, artículos o lo que corresponda dado el carácter de la comunidad que los alberga. Por otra parte, a partir de DSpace el proceso de autoarchivo de los documentos se aligera notablemente: los autores podrán cargar los documentos que deseen y luego los operadores completarán el proceso de carga de los mismos.

5.3 Harvesting

Con el propósito de poner a disposición los recursos digitales de la producción científica de la institución, así como también aquellos que potencialmente podrían querer utilizar quienes realizan las consultas, el equipo de SeDiCI plantea diferentes estrategias para maximizar la cantidad de documentos ofrecidos, intentando a la vez, minimizar el esfuerzo de procesamiento y conexión que implica esta tarea. En esta sección la estrategia que nos ocupa es la referida al proceso de “harvesting” (o cosecha).

Para estas tareas de cosecha, SeDiCI ha desarrollado una herramienta general de recolección de recursos que actualmente se encuentra instalada y funcionando como parte de la administración del portal de búsqueda de recursos académicos obtenidos desde múltiples instituciones pertenecientes al consorcio ISTE⁵². Dado que este servicio se encuentra en funcionamiento y con casi 150.000 recursos recolectados, representa un elemento de estudio concreto para este trabajo.

Esta herramienta busca seguir el patrón de arquitectura de software ETL (5.1), por lo general relacionado con el almacenamiento de datos (data warehousing). Esta arquitectura se aplica principalmente para unificar la información que se utiliza en la lógica de negocios para la toma de decisiones. De esta forma, el proceso de cosecha se inicia con la extracción de datos de diferentes fuentes, luego aplica un conjunto de transformaciones definidas según reglas y

⁵² ISTE: <<http://www.istec.org/es>>. [Consulta: 2012-08-29]

validaciones, para luego finalizar con la carga de los datos en un Data Warehouse (5.2) o Data Mart (5.3) .

Su diseño está basado en las siguientes premisas:

- Permitir el uso de diferentes fuentes y almacenes de datos, encapsulando su lógica particular en componentes conectables.
- Permitir la extensión de la herramienta con nuevos componentes de fuentes y almacenes de datos.
- Permitir la selección y configuración de los filtros de análisis y transformación suministrados por la herramienta, encapsulando la lógica particular en componentes conectables.
- Permitir la extensión de la herramienta mediante la adición de nuevos filtros de análisis y transformación filtros.
- Utilizar una representación abstracta para los recursos, con el fin de unificar los procesos de transformación.
- Proporcionar una interfaz de usuario sencilla e intuitiva para la herramienta de gestión.
- Proveer una interfaz para gestionar la cosecha y el almacenamiento.
- Lograr la tolerancia a errores y reanudar procesos interrumpidos luego de la ocurrencia problemas externos.
- Proporcionar información estadística sobre el estado del proceso y la cosecha de información.

Esta herramienta de cosecha fue desarrollada intentando mantener los componentes tan desacoplados como sea posible. La *Figura 5.2* muestra un diagrama de la arquitectura de la herramienta.

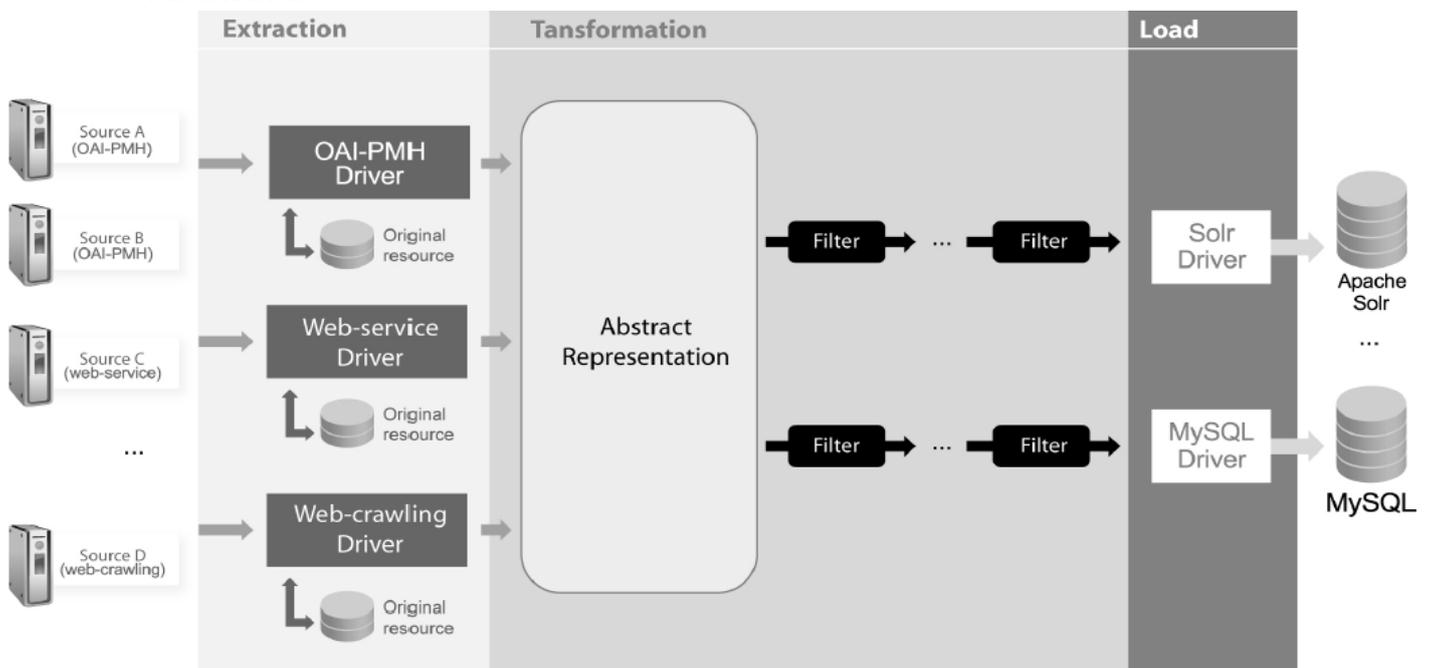


Figura 5.2: Diagrama de Arquitectura - Harvester

5.3.1 Modelo de Datos

El modelo de datos se basa principalmente en tres elementos: Repositorios, Definiciones de cosecha (Harvest Definitions) y colecciones.

Un Repositorio es una entidad abstracta que no determina la forma de obtener recursos, sólo registra información de carácter general tal como el nombre de la institución de origen, correo electrónico de contacto, sitio Web, etc. Con el fin de cosechar los recursos de un repositorio específico, primero deben ser asociados los drivers de conexión -componentes con la lógica requerida para establecer conexiones-, determinando los parámetros pertinentes.

Un elemento "Definición de cosecha" está compuesto por todas las especificaciones requeridas para llevar a cabo una instancia de la cosecha (o una cosecha en particular). Una definición de cosecha se crea a partir de un conector asociado a un repositorio, especificando el protocolo o el método de cosecha utilizada. Esto permite la creación de múltiples cosechas en un único repositorio, usando diferentes enfoques de comunicación.

Las "Colecciones" son el tercer elemento importante. Representan los diferentes destinos para la información generada después de aplicar los procesos de transformación y análisis a los recursos cosechados. Como los repositorios, las colecciones son un elemento abstracto dentro del sistema, y esto significa que cada colección tiene un conector asociado que determina el método de almacenamiento y sus parámetros correspondientes. El objetivo principal es permitir el uso de diferentes opciones de almacenamiento, no sólo basándose en el tipo de almacenamiento, sino también el tipo de información a ser almacenada.

Sumado a estos tres componentes importantes, el modelo de datos se completa con los elementos asociados a los conectores y a las definiciones de cosechas, información complementaria sobre los depósitos y elementos adicionales para controlar y realizar un seguimiento de los métodos de cosecha.

5.3.2 Transformación

De las tres fases de la arquitectura ETL del harvester mencionadas anteriormente, la que nos atañe aquí es la que corresponde a la etapa de transformación de los datos.

Esta fase inicialmente transforma los recursos cosechados en una simple representación abstracta que permite procesar de la misma manera todos los recursos. Esta transformación se realiza mediante conectores, ya que los mismos tienen información acerca de la representación original y de las reglas que deben aplicarse para llevarlo a un nivel abstracto. Cada recurso, en su representación abstracta, pasa a través de una cadena de filtros para analizar particularidades y modificar datos, si fuese necesario. El sistema comprende un conjunto predeterminado de filtros independientes, que son componentes simples y reutilizables que actúan de acuerdo a los parámetros especificados en el archivo de configuración del filtro.

Dado que las colecciones representan los posibles destinos de la información recolectada, y son generadas a partir de necesidades de información específicas (colecciones de registros con determinadas características), es en dichas colecciones donde se especifican el conjunto de filtros que deben ser aplicados antes de insertar un recurso en la misma, donde el orden de selección determina el orden de aplicación.

La ejecución de los filtros puede dar lugar a la modificación, adición o borrado de metadatos específicos, dependiendo de la lógica implementada y la configuración.

Actualmente la herramienta dispone de los siguientes filtros:

- CopyField: copia el contenido de un campo a otro. Si el campo objetivo es inexistente, el filtro lo crea.
- DefaultValue: determina si hay un campo inexistente o nulo. Si es así, crea uno nuevo con un valor predeterminado.
- FieldRemover: toma una lista de campos y los elimina.
- Tokenizer: toma los valores de un campo y los divide a través de una serie de caracteres específicos, generando valores adicionales.
- Stack: agrega filtros; define una lista de filtros (con configuración y ordenamiento), para garantizar el orden de la aplicación.
- ISOLanguage: se aplica a un campo que especifica el recurso del idioma, buscando el valor del campo en una lista de idiomas y reemplazando el valor original con el código de idioma ISO-639 encontrado.
- YearExtractor: se aplica a un campo que contiene una fecha, extrayendo el año y guardándolo en un campo nuevo.
- Vocabulary: toma los valores de un campo y los compara contra los de un diccionario, unificando las variaciones de palabras y sinónimos en una sola palabra.

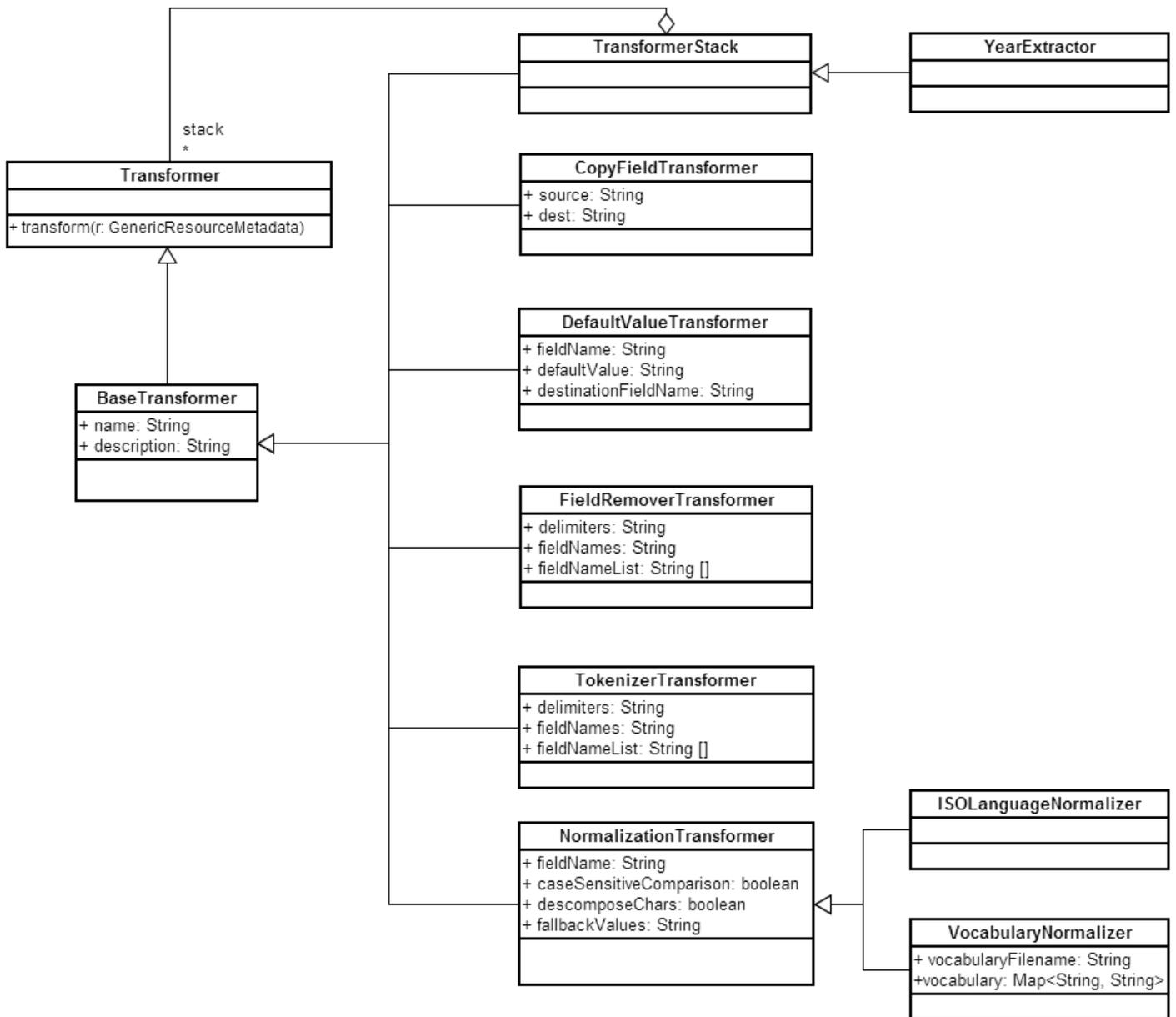


Figura 5.3: Diagrama de Clases - Transformer

5.4 Bibliografía

- (5.1) P. Minton, D. Steffen, "The Conformed ETL Architecture", DM Review, 2004. Disponible en: <http://amberleaf.net/content/pdfs/ConformedETL.pdf>
- (5.2) Data Warehouse, http://en.wikipedia.org/wiki/Data_warehouse. [Consulta: 2012-08-29]
- (5.3) Data Mart, http://en.wikipedia.org/wiki/Data_mart. [Consulta: 2012-08-29]

(5.4) De Giusti, Marisa Raquel; Lira, Ariel Jorge; Oviedo, Néstor, “Extract, transform and load architecture for metadata collection”, VI Simposio Internacional de Bibliotecas Digitales, 17 de mayo de 2011, Servicio de Difusión de la Creación Intelectual (SeDiCI). Disponible en: http://sedici.unlp.edu.ar/bitstream/handle/10915/5529/Documento_completo.pdf?sequence=1. [Consulta: 2012-08-29]

(5.5) SeDiCI – Servicio de Difusión de la Creación Intelectual. “Preguntas Frecuentes”. Disponible en: <http://sedici.unlp.edu.ar/pages/FAQ>. [Consulta: 2012-08-29]

(5.6) De Giusti, Marisa Raquel; Marmonti, Emiliano Horacio; Vila, Maria Marta; Lira, Ariel Jorge; Sobrado, Ariel, “Experiencia en el harvesting de documentos OAI en el proyecto SeDiCI”, III Simposio Internacional de Bibliotecas Digitales (San Pablo, Brasil), 2005, Servicio de Difusión de la Creación Intelectual (SeDiCI). Disponible en: <http://sedici.unlp.edu.ar/handle/10915/5533>. [Consulta: 2012-08-29]

(5.7) De Giusti, Marisa Raquel; Sobrado, Ariel; Lira, Ariel Jorge; Vila, María Marta; Villarreal, Gonzalo Luján, “SeDiCI | Servicio de Difusión de la Creación Intelectual – UNLP”, Servicio de Difusión de la Creación Intelectual (SeDiCI), Revista Interamericana de Bibliotecología; vol. 31, no. 2, 2008. Disponible en: <http://sedici.unlp.edu.ar/handle/10915/5524>. [Consulta: 2012-08-29]

Capítulo 6

Normalización y Calidad de Datos

Capítulo 6 - NORMALIZACIÓN Y CALIDAD DE DATOS

6.1 Introducción

Las bases de datos correspondientes a los repositorios institucionales constituyen una importante fuente de información para efectuar estudios vinculados a la producción científica de una institución y analizar por ejemplo su estado actual.

El sólo registro de metadatos o descriptores asociados a cada publicación no es suficiente para automatizar la determinación y generación de indicadores de producción científica, sino que para poder explotar la información es necesario que los datos sean almacenados y reconocidos unívocamente.

Con el fin de optimizar las técnicas de recuperación de la información e interoperabilidad, se torna indispensable mejorar la calidad de los metadatos de las bases de datos documentales, a través de ciertos procesos que pueden ser de normalización, depuración, asociación o inferencia, entre otros.

6.2 Normalización

El proceso de normalización consiste, básicamente, en la aplicación de un conjunto de atributos o dimensiones para definir adecuadamente los datos. K. Laudon (6.1) reconoce cada una de esas dimensiones de la siguiente manera:

- Exactitud: Referido al grado de correctitud de los registros de información de acuerdo a su correspondencia con el mundo real.
- Integridad: esta dimensión se refiere a que estén presentes todos los datos necesarios que debe contener un sistema de información.
- Consistencia: Se refiere a la definición de estándares y protocolos para guardar los datos. Todos los datos se representan en un formato compatible, que además es el más adecuado para la tarea que se está desarrollando. Debe definirse una forma (o estructura) común de almacenamiento de los datos.
- Temporalidad: Los datos deben estar disponibles siempre que se los requiera. Si los datos no están disponibles o son erróneos cuando se debe tomar una decisión, se podrán generar grandes perjuicios.

A partir de esto se persigue:

- Minimizar redundancias
- Simplificar el mantenimiento de los datos
- Permitir la recuperación sencilla de los datos en respuesta a solicitudes de los usuarios
- Evitar datos no identificables

6.3 Recursos Útiles para Normalizar Nombres de Autor

El nombre de un autor de un trabajo suele conocerse también como la “firma” del autor, y se caracteriza por la forma en la que se escribe, es decir, puede aparecer completo, con las iniciales de los nombres, primero el apellido y luego los nombres, etc.

Estas variaciones en las firmas de un mismo autor, en distintos trabajos, genera el inconveniente de que sus trabajos aparecen referenciados de diferentes formas en los buscadores, en los depósitos OAI, en las bases de datos internacionales, e incluso en otros trabajos, por lo que recopilar su bibliografía personal es en ocasiones muy difícil.

Uno de los pilares fundamentales para la recuperación de la información es justamente el nombre de los autores y es por ello, que su normalización es un aspecto muy relevante.

A continuación se presentan algunas organizaciones que proveen al autor de cierto medio para normalizar su firma.

6.3.1 IraLIS

IraLis⁵³ (International Registry for Authors Links to Identify Scientists) es un proyecto surgido en noviembre de 2006 a partir de las relaciones entre E-LIS (Eprints in library and information science), EXIT⁵⁴ (Directorio de expertos en el tratamiento de la información) y la revista EPI (El profesional de la información).

Consiste en un sistema de estandarización de firmas de autores científicos, cuyo objetivo es ayudar a reducir en lo posible la grave distorsión que se presenta en la recopilación bibliográfica de los autores, de tres formas (6.2):

- Creando un registro de nombres de autores que ayude a localizar sus diferentes variantes. El registro incorpora tanto las que puede haber usado un autor, como las que haya interpretado el productor, agregador, buscador, etc., de las diversas fuentes de información.
- Concientizando a los autores de la importancia de la firma para ser citados correctamente y para poder recuperar toda su bibliografía. Y redactando además criterios de firma normalizada para que sean indexados correctamente y se distingan de otros autores con nombres iguales.
- Creando un formato de firma propio, que permite ser interpretado adecuadamente y sin confusiones también por las fuentes de información de cultura anglosajona.

⁵³ IraLIS <<http://www.iralis.org>>. Consulta [2012-09-12]

⁵⁴ EXIT <<http://www.directorioexit.info/>>. Consulta [2012-09-12]

Así, IraLIS no es únicamente un registro de la forma estandarizada del nombre, sino que se basa en la interoperabilidad de los sistemas y en la recuperación del nombre del autor desde diferentes bases de datos abiertas. Por ejemplo, IraLIS es capaz de contestar en XML a solicitudes hechas bajo el protocolo OpenURL⁵⁵, y el campo iralis del directorio EXIT muestra de forma dinámica los datos que están registrados en IraLIS. Esta funcionalidad permitirá igualmente que desde repositorios como E-LIS pueda validarse la introducción de autores mediante consultas directas a IraLIS.

IraLIS recomienda:

- Firmar siempre igual
- Firmar con el nombre de pila completo y no con la inicial
- Adoptar el formato internacional Nombre Apellido uniendo los nombres de pila o los dos apellidos con un guion - en este sentido se asemeja mucho a las recomendaciones de la FECYT (6.3) -
- Conservar los acentos. Esto evitará que se multipliquen las variantes de firmas y permitirá que se recuperen los nombres en buscadores sensibles a los acentos.

6.3.2 SCOPUS

SCOPUS es la mayor base de datos de resúmenes hasta ahora vista en el mundo, con 13.450 publicaciones (85% de las cuales están indizadas con vocabulario controlado) procedentes de más de 4.000 editoriales internacionales. Con un acceso a más de 25 millones de resúmenes (desde 1966) y 5 años retrospectivos de referencias (llegando a alcanzar 10 años en 2005). Representa aproximadamente un 80% de las publicaciones internacionales revisadas por especialistas, permitiendo asegurar un contenido actualizado gracias a sus actualizaciones semanales.

Scopus permite unificar de una sola vez todas las variantes de nombres de un autor bajo una única firma, utilizando un identificador (Author-ID) que es asignado a cada autor que tiene artículos publicados en su plataforma. Esto es, si un autor determinado detecta que su nombre ha sido indexado de varias formas distintas, scopus le permite seleccionar cada una de esas entradas y solicitar a través de un formulario la unificación de las mismas bajo un solo nombre de autor, de forma tal que su producción científica no aparezca separada. Asimismo ofrece perfiles de autor que cubren afiliaciones, número de publicaciones y sus datos bibliográficos, referencias y detalles del número de citas que ha recibido cada documento publicado.

Todos estos datos pueden ser consultados y descargados de la base de datos de scopus para algún uso específico como puede ser normalizar los datos de un repositorio local a partir de los datos obtenidos.

⁵⁵ OpenURL: tipo de dirección de Internet (URL) que contiene metadatos.

6.4 Herramientas para la Normalización de Datos

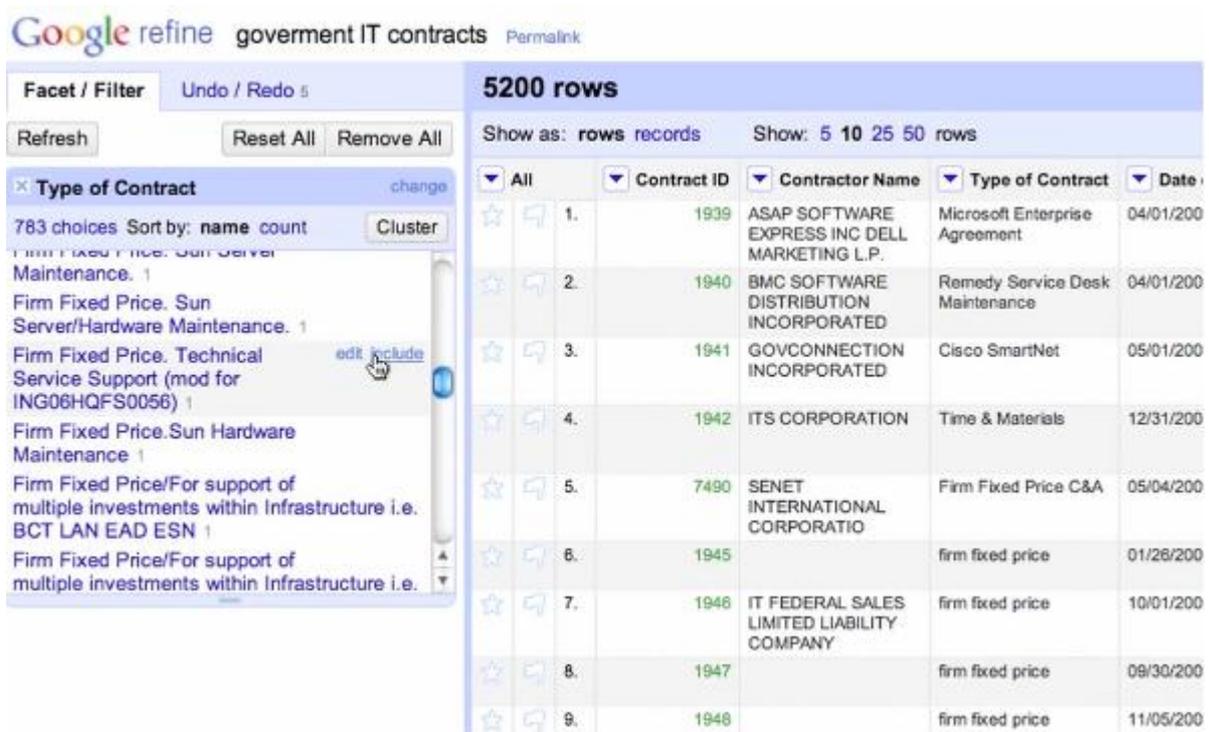
Existen diversas herramientas tanto comerciales como gratuitas, automatizadas e inteligentes, que sirven para el análisis y mejoramiento de la calidad de la información contenida en un repositorio. A continuación se introducen algunas de ellas junto con sus principales características.

6.4.1 Google Refine

Google Refine 2.0 es una herramienta gratuita de Google que permite organizar y transformar datos, ya sea de la web o de las propias bases de datos.

Es muy útil para los casos en que la información con la que se trabaja está desorganizada y resulta difícil encontrar la manera de definirla nuevamente para que tenga sentido. Su principal función es la de organizar datos que no estén bien estructurados y además tiene capacidad de trabajar hasta con cientos de miles de líneas de información.

Una de las opciones que ofrece esta herramienta es agrupar conceptos similares que tienen nombre distinto ya sea de forma manual o por medio de filtros.



The screenshot shows the Google Refine interface for a dataset of 'government IT contracts'. The interface includes a 'Facet / Filter' panel on the left with a 'Type of Contract' filter applied, showing 783 choices. The main table displays 5200 rows of data with columns for Contract ID, Contractor Name, Type of Contract, and Date. The table is sorted by Contract ID.

	Contract ID	Contractor Name	Type of Contract	Date
1.	1939	ASAP SOFTWARE EXPRESS INC DELL MARKETING L.P.	Microsoft Enterprise Agreement	04/01/200
2.	1940	BMC SOFTWARE DISTRIBUTION INCORPORATED	Remedy Service Desk Maintenance	04/01/200
3.	1941	GOVCONNECTION INCORPORATED	Cisco SmartNet	05/01/200
4.	1942	ITS CORPORATION	Time & Materials	12/31/200
5.	7490	SENET INTERNATIONAL CORPORATIO	Firm Fixed Price C&A	05/04/200
6.	1945		firm fixed price	01/28/200
7.	1946	IT FEDERAL SALES LIMITED LIABILITY COMPANY	firm fixed price	10/01/200
8.	1947		firm fixed price	09/30/200
9.	1948		firm fixed price	11/05/200

Figura 6.1: Google Refine

La aplicación también posibilita encontrar incoherencias en los datos, algo muy útil en el caso de manejar cifras (los datos pueden estar escritos en medidas muy distintas). Por otro lado, también trabaja con datos normalizados, pero que se desean organizar en un formato distinto, como listas o tablas. Una función que puede resultar muy útil se llama

“reconciliación”, y permite asociar palabras del conjunto de datos con palabras clave dentro de otras bases de datos. De esta forma, se puede añadir más información a los mismos.

6.4.2 Prism Warehouse Manager

Desarrollado por Prism Solutions, es una herramienta comercial que se enfoca en reducir el tiempo que se emplea en construir y mantener un Data Warehouse. Para eso genera un código para extraer datos operacionales y externos desde ciertas bases de datos, integrando los datos desde variados orígenes, transformándolos y cargándolos en un repositorio específico.

6.4.3 Passport

De Carleton’s Passport, es un software comercial de extracción y transformación de base de datos utilizada en la etapa de migración y conversión de los datos. Ésto automatiza el proceso de extracción, transformando y limpiando los datos desde el origen de los sistemas de producción y otros ambientes de aplicaciones. También es usado en la preparación de datos desde otras aplicaciones, tal como proyectos Data Warehouse y administración de Data Marts. Tanto esta herramienta como la anterior utilizan la técnica de limpieza de datos moderada, apropiada para mejorar campos que son conocidos como nombres o direcciones, o con información que no es utilizada como llaves o índices.

6.4.4 Data Reengineering Tool

De Vality Technology, es una herramienta paga que adopta un método de limpieza bottom-up, esto es, analizando los datos carácter por carácter, surgiendo patrones y reglas del negocio automáticamente. Además provee un diseño de datos y condiciones para ayudar a estandarizar y consolidar los datos. Este método tiende a dejar pocas excepciones para que sean manejadas manualmente y se caracteriza porque el proceso consume poco tiempo. Esta herramienta se focaliza exclusivamente en la limpieza de datos, comenzando con archivos planos. No extrae datos desde las bases de datos operacionales, pero carga datos en el Data Warehouse, replica y sincroniza datos, o administra metadatos.

6.4.5 Enterprise/Integrator

Desarrollada por Apertus Technologies y accesible únicamente a través de un medio pago, soporta un método de limpieza top-down, y se encarga de proveer reglas para la limpieza de datos. No sólo se encarga de esto, sino que también de la extracción, transformación, carga, replicación, sincronización, y administración de los metadatos.

6.5 Actividades Para Mejorar la Calidad de los Datos

Las actividades relativas a la calidad de datos se refieren a cualquier proceso (o transformación) que se aplica a los datos con el objetivo de mejorar su calidad. Entre ellas se pueden nombrar las siguientes:

- Obtención de nueva información: es el proceso de refrescar la información almacenada en la base con datos de mayor calidad (por ejemplo ingresar datos más precisos, de mayor actualidad).
- Estandarización: es el proceso de “normalizar” los datos almacenados, de manera que queden almacenados respetando cierto formato (por ejemplo: la fecha debe tener el formato dd/mm/yy)
- Identificación de Objetos: es el proceso por el cual se identifican registros (dentro de una misma tabla, o entre tablas) que hacen referencia al mismo objeto de la realidad.
- Integración de datos: hace referencia a la actividad de unificar datos provenientes de distintas fuentes, resolviendo los problemas que esto trae aparejados (redundancias, problemas de consistencia, duplicación).
- Confiabilidad de las fuentes: implica “calificar” a las distintas fuentes de información de acuerdo a la calidad de los datos que proveen.
- Composición de calidad: hace referencia a la definición de un álgebra para calcular la composición (o agregación) de las medidas de las dimensiones de calidad de datos. Por ejemplo, calcular la completitud de una unión de relaciones, a partir de la completitud de cada relación.
- Detección de errores: dada una o más tablas, y ciertas reglas que los registros de dichas tablas deben cumplir, este es el proceso de detectar qué registros no cumplen con dichas reglas.
- Corrección de errores: luego de la detección, esta actividad se encarga de corregir los registros con errores, de manera que se respeten todas las reglas correspondientes.
- Optimización de costos: implica obtener la mejor relación costo-beneficio al aplicar procesos de mejora de la calidad de los datos.
- Depuración de datos: intenta resolver la problemática de la detección y corrección de errores e inconsistencias que ocurren en los datos, con el fin de mejorar su calidad. Estas actividades son de mayor importancia en las bases de datos en las cuales la información se ingresó de alguna manera que deja lugar a la aparición de errores. Por ejemplo, cuando la información se ingresa desde el teclado, cuando se obtiene de fuentes no muy confiables o cuando se integran diferentes fuentes de información. En este último caso se vuelve necesario también consolidar los datos cuyo significado es el mismo (pero varían en su representación), así como descartar aquellos datos que se encuentren duplicados.

Un ejemplo de ello son Data Warehouses (almacenes de datos) y sistemas de información basados en web.

Existen variadas herramientas que dan soporte a la limpieza de datos. Sin embargo, es importante tener en cuenta que esta tarea implica, además de la utilización de herramientas, un arduo trabajo manual o de programación de bajo nivel para su resolución.

6.6 Control de Calidad de la Información

En el presente trabajo llamaremos control de calidad de la información a la detección de errores y su posterior corrección. Utilizar el término error puede resultar demasiado amplio, teniendo en cuenta el concepto multifacético con el que se define la calidad de datos. Por lo tanto, aquí se pone foco en: detectar y corregir inconsistencias, datos incompletos y anomalías.

6.6.1 Detección y Corrección de Inconsistencias

Se intentan detectar registros que no cumplan con determinadas reglas, y luego modificar los datos, por ejemplo a partir de la obtención de nueva información, para que sí cumplan con las mismas. Esta tarea incluye asegurar que la información se encuentra consistente (sin contradicciones) y libre de redundancias.

Existen varias formas de corregir los errores detectados, se puede o bien refrescar la base de datos con nuevos datos, o bien utilizar las reglas definidas de manera tal que cuando no se cumple alguna, se asigna un valor que haga que la misma sea verdadera.

6.6.2 Detección y Corrección de Datos Incompletos

Si se consideran las tablas de las bases de datos relacionales, el primer caso de incompletitud a tener en cuenta son los valores nulos. En este caso si bien es muy simple detectar los datos incompletos, puede que corregir sea difícil (en el caso de no tener forma de obtener la información faltante).

Aquí se distinguen dos tipos de fuentes de incompletitud: datos truncados, que corresponden a aquellos datos que son eliminados por no ser significantes para la realidad en cuestión, por ejemplo, y datos censurados, que corresponden a aquellos datos que se sabe que no fueron obtenidos, ya sea porque no se pudo o porque se omitió.

6.6.3 Detección y Corrección de Anomalías

Este es el caso de datos cuyo valor difiere en gran medida con respecto a los demás datos. La situación puede ser alguna de las siguientes:

- El valor fue mal medido, o mal ingresado en la base.
- El valor corresponde a una “muestra” distinta a la de todos los demás.
- El valor es correcto y simplemente corresponde a algún suceso inusual de la realidad.

Estos datos se pueden identificar a partir de dos medidas distintas: midiendo la distancia de los valores registrados a los valores que se espera que haya (desviación interna), o

midiendo la variación de los datos en el tiempo con respecto a otros datos (desviación relativa). Existen varias técnicas para ello. Una de ellas, calcula el valor promedio y la desviación estándar de cierto conjunto de datos, para identificar aquellos valores que se desvíen “demasiado” del valor promedio. Se podría definir por ejemplo un valor límite a partir del cual el dato es sospechoso de estar incorrectamente registrado. Otras técnicas utilizan también el factor tiempo para identificar datos anómalos, partiendo de la base que datos medidos o registrados en cierto lapso de tiempo pueden estar altamente relacionados, y también teniendo en cuenta posibles ciclos donde se registren valores inusuales.

Lidiar con estas anomalías implica un doble esfuerzo: primero se deben identificar, y luego decidir si corresponden a datos correctos de sucesos de la realidad poco comunes, o si corresponden a datos incorrectos y deben ser corregidos.

6.7 Enriquecimiento de la Información

En este caso, podemos hablar tanto de completar los datos almacenados en un repositorio con información más precisa, como también de añadir información inexistente, de buena calidad, proveniente de distintas fuentes de datos, y así incrementar la potencialidad de los datos residentes en el repositorio local.

6.8 Técnicas Para Mejorar la Calidad de los Datos en RI

Aquí se presentan posibles aplicaciones de técnicas de normalización y depuración sobre el contenido de algunos de los metadatos más comunes en un repositorio institucional.

6.8.1 Metadato Autor

Un nombre de autor indica la persona u organización responsable de la creación del contenido intelectual de cierto documento y puede encontrarse rotulado con la etiqueta “author”.

La aparición de numerosos proveedores de datos y de servicios usando el protocolo OAI-PMH convierte a la normalización de nombres en un aspecto clave para la recuperación de la información. No es extraño encontrar que el nombre de un mismo autor aparece citado de modos muy diversos, lo que puede llevar a la confusión y a considerarlo, a los efectos de la catalogación, como dos -o más- personas distintas.

Junto a las recomendaciones a los autores sobre la importancia de firmar las publicaciones de una forma normalizada y estable a lo largo del tiempo, surgen también hoy indicaciones dirigidas a las revistas y a las bases de datos (ver por ej. 6.4), principalmente a las internacionales, las cuales tienen que enfrentar el problema de las diferencias que existen entre los diferentes países al estructurar los nombres personales. Así, la estructura de nombre

personal predominante en las bases de datos internacionales es la formada por una o dos iniciales de nombres, seguida de un solo apellido (por ej. M. B. Almazán), pero con frecuencia son mal recogidos los nombres hispanos si los autores incluyen dos apellidos (por ejemplo, J. García Sánchez, puede ser recogido como J.G.Sánchez). En SeDiCI, a la hora de ingresar el nombre de un autor a la base de datos, se siguen las Reglas de Catalogación Angloamericanas (o AACR2), si bien se hacen algunas excepciones, como no utilizar guiones para unir los apellidos dobles.

Es indudable el interés de propuestas a priori, orientadas a dar recomendaciones a autores, revistas y bases de datos con la intención de lograr una mayor normalización de los nombres de autores en las publicaciones y bases de datos. Sin embargo, también resulta interesante plantear soluciones a posteriori, una vez introducidos los datos en la base de datos. A continuación se detallan posibles soluciones al problema planteado:

- Estandarizar el formato en que está escrito el nombre de autor, esto puede ser, o bien poniendo todo el texto en mayúsculas (ej: ALMAZÁN, MARÍA BELÉN) o bien capitalizado (ej: Almazán, María Belén)
- Detectar y diferenciar apellidos de nombres para luego poder normalizarlos por ejemplo de la forma Apellidos, Iniciales. Se puede utilizar una base de datos de autores “normalizada” (nada asegura que haya una base 100% normalizada, pero sí en su mayoría) para reemplazar el nombre del autor por el valor que aparece según el criterio de la base elegida .

6.8.2 Metadato Título

Representado por la etiqueta “title”, corresponde al nombre dado a un recurso, en este caso a un documento, habitualmente por el autor.

Sería provechoso estandarizar los títulos de los documentos, para que todos aparezcan escritos de la misma forma. Por ejemplo, se podría optar por poner todo el texto en mayúsculas, o de forma capitalizada. También es conveniente quitar los espacios en blanco antes del texto, como así también los que sobren entre las palabras que lo componen.

6.8.3 Metadato Lenguaje

Se alberga en la etiqueta “language”, e indica el/los idioma/s del contenido intelectual del documento.

Cuando un documento no tiene asociado el metadato lenguaje, puede ser útil detectar el idioma del texto en los casos en que sea posible, de modo tal de corregir posteriormente la falta de completitud del mismo.

Una solución interesante consiste en utilizar la API “language” de google⁵⁶ para detectar el idioma de un texto. Consiste en una API cliente Java simple, no oficial, y actualmente paga, correspondiendo a la versión 2 de la misma. Se requiere una clave (key) para utilizar el servicio, y el número de solicitudes diarias es limitada. Si se implementara un

⁵⁶ google-api-translate-java: <<http://code.google.com/p/google-api-translate-java/>>. [Consulta: 2012-08-29]

transformer con esta forma de detección, deberían ser configurables parámetros como la clave de la API, de manera tal de poder instanciar el harvester y establecer en su instalación la clave propia del usuario obtenida de google, y de esta forma aplicar la detección de idioma.

Otra librería muy prometedora y de código abierto (licencia Apache 2.0) es la desarrollada por Ciboza Labs, denominada "Language Detection Library". La misma posee un 99% de precisión en la detección de lenguajes, sobre 50 posibilidades diferentes.

Para la detección podría utilizarse el campo título, o si tuviese, la descripción del documento, la cual proporciona mas elementos para abordar el lenguaje en el que está escrito el texto completo.

6.8.4 Metadato Fecha

Indica la fecha en la cual el recurso se puso a disposición del usuario en su forma actual, y está representada por la etiqueta "date".

Podría intentar normalizarse también el formato del metadato fecha, estandarizándolo por ejemplo a la notación "yyyy-MM-dd" (estándar ISO 8601), donde yyyy es el año en el habitual calendario gregoriano, MM es el mes del año entre 01 (enero) y 12 (diciembre), y dd es el día del mes entre 01 y 31. Otras anotaciones utilizadas comúnmente son por ejemplo, 2/4/95, 4/2/95, 95/2/4, 4.2.1995 04-FEB-1995, 4-February-1995, y muchos más. Sobre todo los dos primeros ejemplos son peligrosos, ya que como ambos se utilizan con bastante frecuencia en los EE.UU. y en Gran Bretaña y ambos no pueden distinguirse, no está claro 2/4/95 significa 1995-04-02 o 1995-02-04. La notación 2/4/5 tiene por lo menos seis interpretaciones razonables (suponiendo que sólo el siglo XX y XXI son candidatos razonables).

Una manera de abordar este problema de falta de normalización del metadato Fecha, puede ser definir un detector de fechas, que sirva para un conjunto de casos: diferentes combinaciones de año, mes y día, en representación numérica o textual, en inglés o español, con hora y/o timezone, contemplando diferentes caracteres de separación (/,-), abreviaturas (como ser ene, feb, o para de días Lun, Mon), períodos (2002-2003), o fechas abiertas (como ser 201?, que indica "un año de 2010-2019").

Deben contemplarse la mayoría de las diferentes notaciones de fechas posibles, esto es:

- Little-endian gregoriano, comenzando con el día: Esta secuencia es común a la gran mayoría de los países del mundo. Este formato de fecha se origina en el oeste por la costumbre de utilizarla en documentos religiosos y legales, que en un tiempo eran la mayoría de los documentos creados. El formato se ha acertado en el tiempo, pero el orden de los elementos se ha mantenido constante. A continuación se presentan algunos ejemplos de fechas escritas con este formato:
 - "8 November 2003" o "8. November 2003" (Este último suele ser común en regiones de habla hispana)
 - 8/11/2003, 08.11.2003 o 8-11-2003
 - 08-Nov-2003
 - 08Nov03

- [The] 8th [of] November 2003
 - Sunday, 8 November 2003
 - 8/xi/03, 8.xi.03, 8-xi.03, o 8.XI.2003 (utilizando números romanos para el mes) – Esto generalmente se limita a la escritura y se asocia con una serie de escuelas y universidades. También ha sido utilizado por el Vaticano y por el correo de Canadá como una alternativa bilingüe del mes.
 - 8 November AD 2003
- Big-endian gregoriano, comenzando con el año: En este formato, en consistencia con el sistema decimal Indio, el dato más significativo está escrito antes de los menos significativos, es decir, el año antecede al mes y este último antecede al día. Es un estándar en los países asiáticos, Hungría, Suecia, las fuerzas armadas de Estados Unidos y parte de Canadá. Ejemplos de fechas utilizando este otro formato son:
 - 2003-11-09: utilizado por el estándar ISO 8601 que también suele utilizar adicionalmente ceros a la izquierda, por ejemplo, 0813-03-01, para que sea fácil de leer y ordenar por una computadora. Se lo puede encontrar también junto con la hora UTC en el formato de Internet de fecha/hora. Este formato también se puede encontrar en ciertos países de Asia, principalmente del este asiático, así como también en algunos países europeos.
 - 2003 November 9
 - 2003Nov9
 - 2003Nov09
 - 2003-Nov-9
 - 2003-Nov-09
 - 2003-Nov-9, Sunday
 - 2003. november 9. – El formato oficial en Hungría. También puede utilizarse: 2003. nov. 9., 2003. 11. 9., 2003. XI. 9.
 - 2003.11.9 utilizando puntos y sin ceros a la izquierda. Es común en China.
 - 9 November 2003, 18h 14m 12s
 - 2003/11/9/18:14:12
 - 2003-11-09T18:14:12 (ISO 8601)

Otro formato ya no utilizado es el middle-endian que comienza con el mes. Esta secuencia era de uso común en el Reino Unido y en periódicos británicos como *The Times* a principios del siglo XX, por lo que no sería necesario contemplar esta variante.

6.8.5 Otros Metadatos

A veces, un documento trae aparejado el título de la revista donde fue publicado o el nombre del evento en donde se presentó (metadatos dc:source o dc:isPartOf de Dublin Core). Sería útil normalizar este lugar de publicación, buscando por ejemplo la revista, la editorial o el evento en una base de datos, de la misma manera en la que se planteó para normalizar los

nombres de autor. Por ejemplo, podría utilizarse la base de sherpa-romeo⁵⁷ que incluye información relevante sobre más de 16 mil revistas.

Otra mejora sería intentar distinguir entre Autores e Instituciones (“affiliation”), y luego colocarlos en metadatos independientes o bien en un mismo metadato separado por algún delimitador (ej: Estivariz, F.E., Centro de Estudios Endocrinos, Universidad Nacional de La Plata, 1900 La Plata, Argentina⁵⁸). Además, para enriquecer la información, si no existiese este metadato afiliación, podría agregarse de alguna fuente en donde aparezca el autor y su afiliación, como puede ser SCOPUS.

6.9 Bibliografía

(6.1) Laudon, K. C. “Data quality and due process in large interorganizational record system”. 4-11, 1986.

(6.2) Baiget, Tomàs; Rodríguez-Gairín, Josep-Manuel; Peset, Fernanda; Subirats, Imma; Ferrer, Antonia. “Normalización de la información: la aportación de IralIS”. En: El profesional de la información, 2007, noviembre-diciembre, v. 16, n. 6, pp. 636-643. Disponible en: <http://www.elprofesionaldeinformacion.com/contenidos/2007/noviembre/10.pdf>. [Consulta: 2012-09-12]

(6.3) FECYT: Normalización de autores e instituciones en las publicaciones científicas. Disponible en: http://www.accesowok.fecyt.es/wp-content/uploads/2009/06/normalizacion_nombre_autor.pdf. [Consulta: 2012-08-29]

(6.4) Ruíz-Pérez, R.; Delgado López-Cózar, D. y Jiménez Contreras, E. “Spanish personal name variations in national and international biomedical databases: implications for information retrieval and bibliometric studies”, Journal of Medical Library Association, 90 (4), 411-30, 2002. Disponible en: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC128958/>. [Consulta: 2012-08-28]

(6.5) Baiget, Tomás; Rodríguez-Gairín, Josep-Manuel; Peset, Fernanda; Subirats, Imma; Ferrer, Antonia. “La normalización de la información: la aportación de IralIS”. International Registry for Authors in Library and Information Science. Disponible en: <http://www.elprofesionaldeinformacion.com/contenidos/2007/noviembre/10.pdf>

(6.6) Hoffer, George & Valacich Benjamin. “Modern Systems Analysis and Design”. Cummings Publishing, 1996

(6.7) Ballou D.P. & Pazer H.L. “Modeling data and process quality in multi-input, multi-output information systems”. 1985.

(6.8) Batini, C. and Scannapieca, M. (2006) Data Quality: Concepts, Methodologies and Techniques, Springer - ISBN-13 978-3-540-33172-8.

⁵⁷ Sherpa-Romeo <<http://www.sherpa.ac.uk/>>. [Consulta: 2012-08-29]

⁵⁸ Extraído de la base de SCOPUS <<http://www.scopus.com/>>, metadato affiliation. [Consulta: 2012-08-29]

(6.9) Caballero, I., Calero, C., Al-Hakim, L. and Serrano, M.Á. (2009) 1st International Workshop on Data Quality for Software Engineering (DQ4SE 2009). Disponible en: <http://alarcos.esi.uclm.es/DQ4SE/> . [Consulta: 2012-08-29]

(6.10) Erhard Rahm, H.H.D. (2000) Data Cleaning: Problems and Current Approaches, Universidad de Leipzig, Alemania: IEEE Data Engineering Bulletin, Vol. 23(4): 3-13, 2000.

(6.11) Marotta, A. (2009) Material Curso Calidad de Datos, Instituto de Computación, Facultad de Ingeniería de la UdelaR.

Capítulo 7

Desarrollo

Capítulo 7 - DESARROLLO

7.1 Introducción

En este capítulo se presenta el problema existente y se exponen las mejoras introducidas a partir de métodos descritos en capítulos anteriores. Asimismo se describen los resultados de las pruebas realizadas.

7.2 Problema Específico

Como se expuso en el capítulo 5, SeDiCI realiza actualmente la catalogación de recursos mediante un conjunto de metadatos de uso interno, algunos de los cuales son obligatorios y otros recomendados u opcionales, con el objeto de minimizar toda posibilidad de error y de maximizar la inmediata recuperación del recurso buscado.

Sin embargo, debido a que se incluyen recursos bibliográficos pertenecientes a una gran diversidad de fuentes (catalogación externa, importaciones, cosechas automáticas, entre otras), la heterogeneidad de los datos suma nuevas dificultades al momento del intercambio de metadatos. Al utilizar el protocolo OAI-PMH con formato Dublin Core para el proceso de recolección de recursos (o “harvest”), SeDiCI necesita realizar mapeos y transformaciones sobre la información importada, lo cual puede llevar a la pérdida de información o incluso a la generación de documentos incompletos (7.1). Por ejemplo, el metadato que identifica el tipo de documento (metadato “dc:type” de Dublin Core) es completado por cada repositorio según sus criterios particulares, con lo cual puede suceder que muchos valores distintos encontrados en este metadato en realidad representen el mismo valor: valores como *Article*, *Artículo* y *ART* significan que el tipo de recurso es un artículo científico. Algo similar sucede con el metadato de idioma.

Por lo tanto, de acuerdo a lo expuesto en el presente y en capítulos previos, se hace evidente la necesidad de normalizar la información contenida en una base de datos documental, en este caso la utilizada por SeDiCI.

7.3 Mejora Propuesta

El principal desafío es diseñar e implementar una serie de filtros que generen de manera eficaz y eficiente datos normalizados. Los mismos se ejecutarán en la etapa de Transformación del harvester descrita previamente apuntando a dos áreas:

- Control de calidad de la información: detección de errores y posteriores correcciones sintácticas.
- Enriquecimiento de la información

A partir de esto se espera:

- Fomentar la reusabilidad de datos sin tener que recurrir a la fuente inicial. Aunque este trabajo parezca un coste adicional, el valor que los datos tomarán será mucho mayor ya que la información que se posea sobre los documentos será mucho mas rica y confiable.
- Proporcionar información que ayude a la transferencia de los datos: facilitar el acceso a los datos, su adquisición y una mejor utilización de los mismos logrando una mejor interoperabilidad de la información cuando esta procede de fuentes diversas. Esto le será de gran utilidad al usuario u organización que los recibe en el procesamiento, interpretación, y almacenamiento de los datos en repositorios internos.

7.4 Metodología

Los cambios se introducirán de manera gradual o evolutiva en la mayoría de los casos, esto significa que serán desarrollos dentro de todo pequeños y reutilizables que se enfocarán en resolver un problema determinado en cada caso, pudiendo ser la salida de uno la entrada de otro, o bien utilizarse individualmente.

Se define una clase Java para cada uno de los filtros y sus respectivos parámetros de acuerdo a los cuales actuará, encontrándose estos especificados en el archivo de configuración transformerContext.xml, el cual además indica sobre que campo será aplicado y los parámetros de entrada correspondientes.

Por ejemplo:

```
<entry key="author_uc">
  <bean
    class="ar.edu.unlp.sedici.harvester.transformers.trans.UpperCaseFieldTransformer">
    <property name="fieldName" value="author"/>
    <property name="name" value="author_uc"/>
    <property name="makeBackup" value="true"/>
  </bean>
</entry>
```

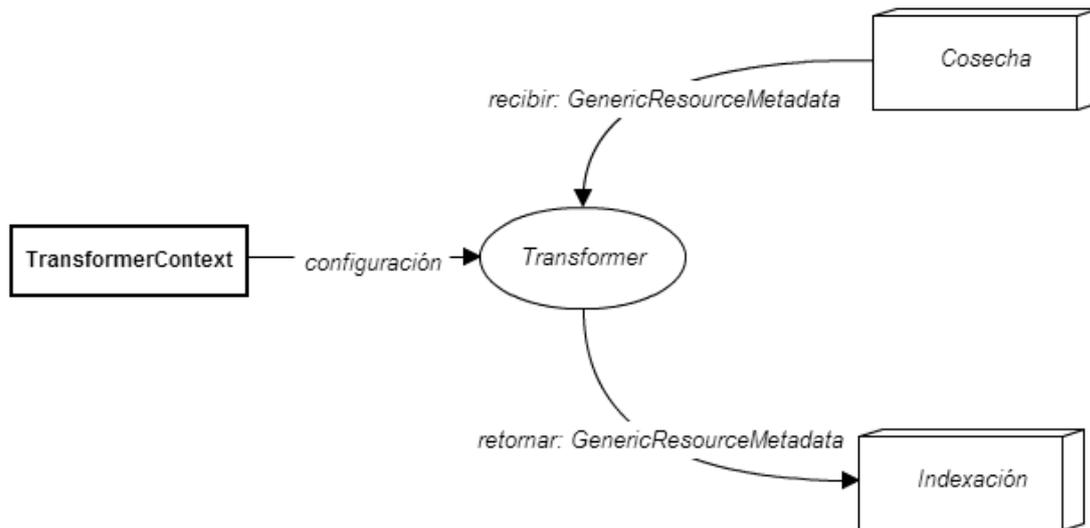


Figura 7.1

Se define un bean de Spring⁵⁹ para cada Transformer. Cada bean recibe como parámetro una configuración determinada, la cual fue definida en el archivo de configuración TransformerContext.xml. Por lo tanto obtiene una clave de entrada (“key”), que generalmente se corresponde con el nombre del filtro, es decir la propiedad “name” que aparecerá en el harvester para su selección. Por su parte, “class” se refiere a la clase del transformer que se va a ejecutar y “properties” corresponde a los parámetros de entrada, donde por ejemplo puede encontrarse: “fieldName” con el nombre del campo sobre el cual va a aplicarse el filtro, “name” indicando el nombre del filtro, la propiedad “makeBackup” que indica que debe hacerse un backup si existiese algún valor previo alojado en dicho campo, etc. Estos parámetros varían según el Transformer a ejecutar, pudiendo existir n parámetros diferentes si el filtro lo requiere.

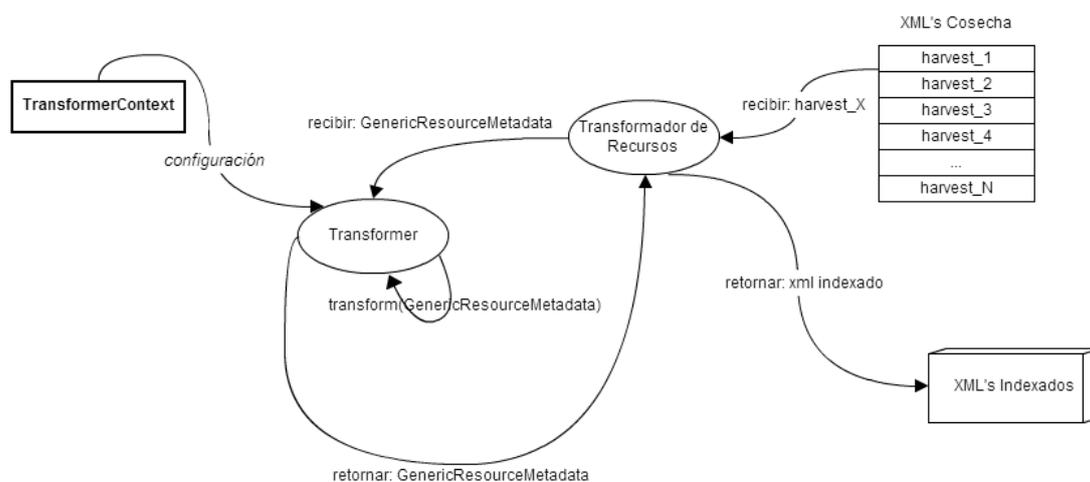


Figura 7.2

⁵⁹ Spring Framework <<http://www.springsource.org/>>. Consulta: [2012-10-11]

Inicialmente se transforman los recursos cosechados en una representación abstracta simple denominada *GenericResourceMetadata* que permite procesar de la misma manera todos los recursos independientemente de su representación original, en este caso XML.

Cada recurso, en su representación abstracta, pasa a través de uno o más filtros de transformación para analizar particularidades y realizar modificaciones sobre los datos, si fuese necesario. En la *Figura 7.2* puede apreciarse el proceso de transformación de los datos cuando estos pasan por un determinado filtro.

Las clases Java que implementan los filtros, heredan de *BaseTransformer* (ver *Figura 7.3*) que implementa la interface *Transformer* cuyo método principal, y el que tiene relevancia en el presente trabajo es el siguiente:

- `public void transform(GenericResourceMetadata r)`: Es el método de transformación en sí, el cual es invocado en la indexación y se redefine en cada filtro, cada uno de los cuales trabaja los datos de alguna forma en particular y los devuelve con ciertos cambios aplicados. Recibe como parámetro un “*GenericResourceMetadata*” que, como se dijo anteriormente, corresponde a la representación abstracta de la cosecha realizada. De ese recurso, se obtienen los campos que se deseen transformar (los cuales fueron definidos en el archivo de configuración), iterando sobre los mismos.

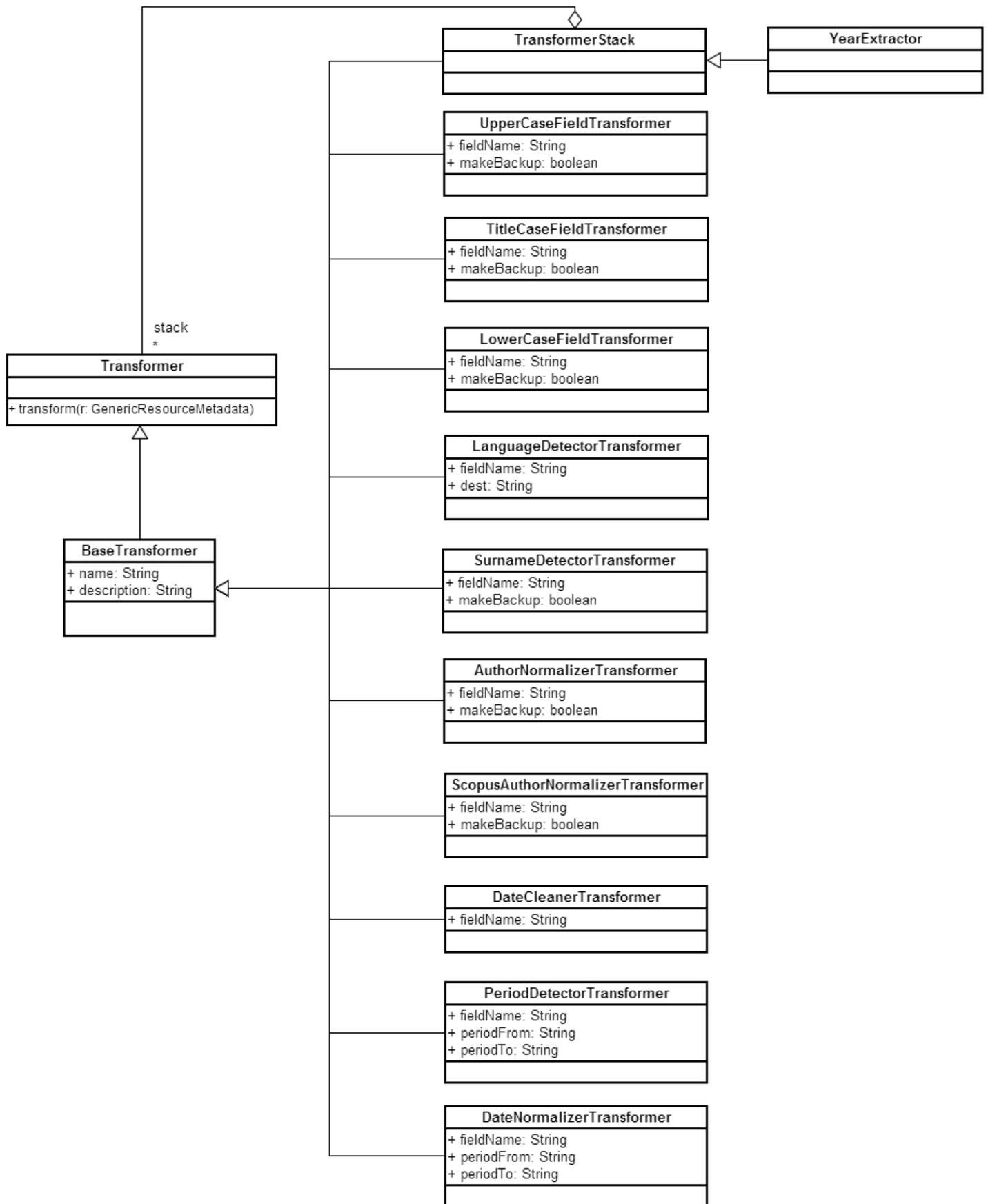


Figura 7.3: Diagrama de Clases - Transformers Implementados

A continuación se detallan cada uno de los filtros que fueron incorporados a la herramienta existente, los cuales se pueden apreciar en el diagrama de clases de la *Figura 7.3*.

7.5 Estandarización de Formato del Texto

Se implementaron diferentes filtros que permiten transformar el formato del texto con alguna de las siguientes propiedades: "title" (la primera letra con mayúscula), "uppercase" (todas las letras con mayúsculas), "lowercase" (todas las letras con minúsculas).

Además, quitan los espacios en blanco que se encuentren al principio y/o al final del texto, y reducen a uno la cantidad de espacios entre palabra y palabra en caso de que haya dos o más.

Por defecto el texto no posee un formato estándar, y estos filtros son aplicables a cualquier campo que contenga texto.

7.5.1 Filtro: UpperCaseField

Filtro que estandariza el formato de determinado campo recibido como parámetro, pasando todo el texto a mayúsculas. Se puede utilizar por ejemplo para el título y/o para el nombre del autor.

La clase que implementa este filtro se llama UpperCaseFieldTransformer, el cual utiliza para la transformación el método toUpperCase() de java.lang.String, el cual convierte todos los caracteres de un String recibido como parámetro a mayúsculas, usando las reglas del lenguaje por defecto.

7.5.2 Filtro: LowerCaseField

Filtro que estandariza el formato de determinado campo recibido como parámetro, pasando todo el texto a minúsculas. Se puede utilizar para los mismos casos que el filtro anterior.

La clase que implementa este filtro se llama LowerCaseFieldTransformer, el cual utiliza para la transformación el método toLowerCase() de java.lang.String, el cual convierte todos los caracteres de un String recibido como parámetro a minúsculas, usando las reglas del lenguaje por defecto.

7.5.3 Filtro: TitleCaseField

Filtro que estandariza el formato de determinado campo recibido como parámetro, llevándolo a la forma de "Título", lo que se conoce como "capitalizado" (esto es, mayúsculas

en la primer letra de cada palabra y minúsculas para el resto). También puede utilizarse para el título, para el nombre del autor, etc.

La clase que implementa este filtro se llama `TitleCaseFieldTransformer`, y lo que hace es convertir todas las palabras contenidas en un `String` delimitadas por espacios, en palabras con formato tipo título utilizando el método `capitalizeFully` de la clase `org.apache.commons.lang.WordUtils`, esto es, cada palabra conformada por la primer letra en mayúsculas y el resto de las letras que la componen en minúsculas. Un espacio en blanco es definido por el método `Character.isWhitespace(char)`. Un `String` de entrada nulo retornará nulo.

7.6 Detección de Lenguaje del Texto

En este caso, se desarrolló un filtro capaz de identificar el idioma de un texto cuando el mismo no contiene el metadato "language" asociado, o bien cuando no está especificado o tiene un valor que no encaja en ningún idioma (como por ejemplo: "otro").

7.6.1 Filtro: LanguageDetector

La clase que implementa este filtro se denomina `LanguageDetectorTransformer`, la cual utiliza la librería "Language Detection" de Cybozu Labs, para detectar el idioma de los documentos a partir del título de los mismos y solucionar la falta de completitud del dato en ciertos casos. La misma proporciona un método que consiste en calcular las probabilidades de los diferentes idiomas sobre las características de ortografía de un texto específico. Esto se realiza utilizando el algoritmo clasificador de bayes, una técnica de clasificación descriptiva y predictiva basada en la teoría de la probabilidad del análisis de Tomas Bayes la cual busca correlaciones entre atributos.

Primero clasifica los documentos en categorías de lenguajes, y luego actualiza las probabilidades de acuerdo a las características del texto dado.

$$p(C_k|X)^{(m+1)} \propto p(C_k|X)^{(m)} \cdot p(X_i|C_k)$$

Donde: C_k = categoría, X = documento, X_i = característica de documento.

Finaliza el proceso de detección si la probabilidad máxima (normalizada) es mayor a 0.99999. Este algoritmo tiene una precisión del 90%, y tiene probabilidades muy bajas de detectar lenguajes como el japonés, el chino tradicional, el ruso y el persa, debido al sesgo y el ruido que generan las características que poseen los mismos.

Por lo expuesto anteriormente, la librería de Cybozu Labs implementó una mejora que añade un filtro de ruido y la normalización de caracteres. El primero, elimina caracteres de idiomas independientes (cifras numéricas, símbolos, URL's y direcciones de correo electrónico), caracteres latinos en textos no latinos y latinos (acrónimos, nombres de personas, etc.). Para la normalización, clasifica frecuencias similares de texto y normaliza cada

grupo en una representación determinada. De esta manera, a través de esta librería, pueden ser detectados 49 lenguajes diferentes con una precisión de 99.8%.

7.7 Normalización de Nombre de Autor

Lograr la normalización de los nombres de autores requiere un análisis un tanto complejo, por lo que se decidió realizar una descomposición en filtros simples de manera tal que cada uno tenga un objetivo distinto y claramente definido, reduciendo la cantidad de operaciones. Los mismos se desarrollaron de manera evolutiva, esto es, la transformación lograda como salida en uno de ellos sirve como entrada para el siguiente filtro, donde este último es capaz de utilizar todas las transformaciones de los filtros que lo preceden.

7.7.1 Filtro: SurnameDetector

La clase que implementa este filtro se llama SurnameDetectorTransformer, y lo que hace es detectar los apellidos de un autor, según su ocurrencia en la base de autores de Scopus o según su sintaxis:

- Si el nombre está separado por comas, es decir *Palabra1, Palabra2 Palabra3*, entonces *Palabra1* es el apellido y el nombre se lleva a la forma *Apellidos, Nombres*.
- De lo contrario, si viene separado por espacios, lo que hace es buscar combinaciones de palabras en la base de autores de Scopus para ver si las mismas componen un apellido. Procede de la siguiente manera: si la entrada es P1 P2 P3 P4...PN, busca primero si la combinación de todas las P conforman un apellido, o sea P1 P2 P3 P4...PN, sino, continua buscando P2 P3 P4...PN, luego P3 P4...PN hasta PN. El criterio para realizar las combinaciones de este modo, es porque es más probable que al estar el nombre separado por espacios, el apellido esté al final del mismo. Al finalizar, si no se encontró ninguna combinación, continúa buscando pero ahora combinando P1 P2 P3 P4... PN-1, luego P1 P2 P3 P4... PN-2 y así hasta llegar a P1. Si no encuentra ninguna de las combinaciones posibles, la salida es igual a la entrada, ya que no se cuenta con ninguna herramienta para deducir cuál es el apellido.

7.7.2 Filtro: AuthorNormalizer

Este filtro es implementado en la clase AuthorNormalizerTransformer, la cual emplea el filtro "SurnameDetector" en primera instancia para detectar los apellidos de un autor, llevando el nombre a la forma *Apellidos, Nombres*. Luego normaliza los nombres dejando sólo sus iniciales (de acuerdo las directrices que utiliza Scopus) para darle salida con el formato *Apellidos, Iniciales*.

Si el nombre recibido como entrada ya se encuentra bien formado, esto es, sigue el patrón *Apellidos, Inicial1. Inicial2. InicialN.*, se le da salida a ese valor directamente.

7.7.3 Filtro: ScopusAuthorNormalizer

La clase ScopusAuthorNormalizerTransformer implementa este filtro, el cual se aplica sobre un campo que contiene un nombre, y lo que hace es reemplazar ese nombre por el valor que encuentra en Scopus en los casos que se consideren apropiados, según cierto nivel de confianza. En la siguiente tabla se representan las salidas según un valor de entrada obtenido luego de aplicar “SurnameDetector” y “AuthorNormalizer” y finalmente normalizándolo según el valor que aparece en la base de Scopus. Además se indica el nivel de confianza para cada salida obtenida.

Los niveles de confianza pueden ser:

- Alto: si todas las iniciales de los nombres de la salida están contenidas en la entrada, respetando el orden que tienen en la misma.
- Medio: si todas las iniciales de los nombres de la salida están contenidas en la entrada, sin respetar el orden que tienen en la misma.
- Bajo: si alguna de las iniciales de los nombres de la salida no está contenida en la entrada.

Aquí se tomaron sólo los casos de confianza Alta y Media, descartando las combinaciones posibles que resulten en un nivel de confianza Bajo.

Entrada	Aparición en Scopus (Apellido, A. B.*)	Salida (Apellidos, A. C.*)	Nivel Confianza (Bajo, Medio, Alto)
Apellidos, A. B.	-	Apellidos, A. B.	Alto
Apellidos, A.	Apellidos, A. B.	Apellidos, A.	Alto
Apellidos, A.	Apellidos, B. A.	Apellidos, A.	Alto
Apellidos, A. B.	Apellidos, A.	Apellidos, A.	Alto
Apellidos, A. B.	Apellidos, B. A.	Apellidos, B. A.	Medio
Apellidos, A. B.	Apellidos, A. B.	Apellidos, A. B.	Alto
Apellidos, A. B.	Apellidos, C.	Apellidos, A. B.	Alto

Tabla 7.1

7.8 Depuración y Normalización de la Fecha

7.8.1 Filtro: DateCleaner

Se implementa en la clase `DateCleanerTransformer` y se aplica a un campo que contiene una fecha, y su objetivo es eliminar caracteres, como ser '[]', blancos repetidos, y convirtiendo signos como '?', '/' o '.' a '-', y todo tipo de construcciones inválidas o aclaraciones, por ejemplo '(4° fecha)'.

7.8.2 Filtro: PeriodDetector

Es implementado por la clase `PeriodDetectorTransformer` y se aplica a un campo que contiene una fecha, y su objetivo es detectar períodos en el dato entrante, empezando por limpiar el dato con el filtro "DateCleaner". Luego, revisa el dato de entrada intentando detectar períodos de acuerdo a determinados formatos (ej.: 'dd' al 'dd' de MMMM de yyyy, yyyy-yyyy, from yyyy to yyyy, etc.). En el caso de encontrar un período, separa el campo en dos, un período inicial y uno final (`dateFrom`, `dateTo`).

7.8.3 Filtro: DateNormalizer

Este filtro se implementa en la clase `DateNormalizerTransformer`, y el mismo se aplica a un campo que contiene una fecha. Primero emplea el filtro "DateCleaner" para limpiar valores indeseados, luego "PeriodDetector", para detectar períodos en el caso que hubiese, y por último normaliza tanto el campo `date`, como `dateFrom` y `dateTo` para los documentos que posean un período como fecha, al formato `yyyy-MM-dd`.

Se decidió normalizar la fecha según el estándar ISO 8601 (YYYY-MM-DD), el cual especifica representaciones numéricas para fecha y hora. Esta notación estándar ayuda a evitar la confusión causada por la gran variedad de notaciones existentes y aumenta la portabilidad en interfaces de usuario. Se contemplaron todas las variantes de fechas expuestas en 6.8.4, comprendiendo los formatos big y little endian, y variantes según los distintos países.

A continuación se detallan algunas ventajas de la notación estándar para la fecha de la norma ISO 8601 en comparación con otras variantes de uso común:

- De fácil lectura y escritura por software (no hay necesidad de usar tablas de referencia para entradas como 'Jan', 'FEB', etc.)
- De fácil comparación y ordenamiento utilizando comparaciones de cadenas triviales.
- Independiente del idioma.
- No se confunde con otras notaciones de fecha populares.
- Coherente con el sistema de notación común de 24 horas, donde las horas también se escriben anteponiéndose a los minutos y segundos, lo cual permite que dos cadenas se comparen y ordenen fácilmente.
- La notación es corta y tiene una longitud constante, lo que hace que tanto la entrada de datos por teclado, como el alta en una tabla sea más sencilla.

7.9 Pruebas Realizadas

- **UpperCaseField:**

Entrada	Salida
Muschietti, María Emilia	MUSCHIETTI, MARÍA EMILIA
Juan Solomin	JUAN SOLOMIN
Structural testing with use cases	STRUCTURAL TESTING WITH USE CASES
Cellular memetic algorithms	CELLULAR MEMETIC ALGORITHMS

Tabla 7.2

- **TitleCaseField:**

Entrada	Salida
Muschietti, maría emilia	Muschietti, María Emilia
Juan solomin	Juan Solomin
Structural testing with use cases	Structural Testing With Use Cases
Cellular memetic algorithms	Cellular Memetic Algorithms

Tabla 7.3

- **LowerCaseField:**

Entrada	Salida
Títol obtingut de la portada digitalitzada	títol obtingut de la portada digitalitzada
Hipovitaminosis D y obesidad mórbida efectos de la cirugía bariátrica	hipovitaminosis d y obesidad mórbida efectos de la cirugía bariátrica
Structural testing with use cases	structural testing with use cases
Cellular memetic algorithms	cellular memetic algorithms

Tabla 7.4

- **LanguageDetector:**

Entrada	Lenguaje
Títol obtingut de la portada digitalitzada	es
Hipovitaminosis D y obesidad mórbida efectos de la cirugía bariátrica	es
High-speed polymerase chain reaction in CMOS-compatible chips	en
Hibridació genòmica comparada en oòcits aplicabilitat al diagnòstic genètic preimplantacional	es

Herpevirus humà 8 infecció i patogènia en relació amb el virus d'Epstein Barr	es
Guatemala: segona oportunitat	it

Tabla 7.5

- **SurnameDetector:**

Entrada	Salida
Muschietti, María Emilia	Muschietti, María Emilia
Ronderos, Ramón	Ronderos, Ramón
Juan Ernesto Solomin	Solomin, Juan Ernesto
Juan Solomin	Solomin, Juan
Martínez Bravo Orlando	Martínez Bravo, Orlando
Orlando Martínez Bravo	Martínez Bravo, Orlando
Orlando Martínez	Martínez, Orlando
Ariel Orlando Martínez	Martínez, Ariel Orlando
Belén Almazán	Belén Almazán

Tabla 7.6

- **AuthorNormalizer:**

Entrada	Salida
Ronderos, Ramón	Ronderos, R.
Muschietti, M. A.	Muschietti, M.A.
Muschietti, María Emilia	Muschietti, M.E.
Juan Ernesto Solomin	Solomin, J.E.
Juan Solomin	Solomin, J.
Martínez Bravo Orlando	Martínez Bravo, O.
Orlando Martínez Bravo	Martínez Bravo, O.
Orlando Martínez	Martínez, O.
Ariel Orlando Martínez	Martínez, A.O.
Belén Almazán	Belén Almazán

Tabla 7.7

- **ScopusAuthorNormalizer:**

Entrada	Salida
Martin, María Teresa	Martin, M.T.
Martin, María Teresa Juana	Martin, M.T.
Muschietti, M.A.	Muschietti, M.A.
Muschietti, María	Muschietti, M.
Juan Ernesto Solomin	Solomin, J.E.
Solomin, Ernesto Juan	Solomin, J.E.
Ernesto Solomin	Solomin, E.
Martínez Bravo Orlando	Martínez Bravo, O.
Orlando Agustín Martínez Bravo	Martínez Bravo, O.
González, Carlos	González, C.
González, María Cecilia	González, C.
González, María	González, M.
González, María Marta	González, M.M.
Belén Almazán	Belén Almazán

Tabla 7.8

- **DateCleaner:**

Entrada	Salida
19--	19--
2009 (4º época)	2009
2009 (4º época) Febrero	2009 Febrero
200?	200-
2002/2003	2002-2003
2007-03-00	2007-03
00-03-2007	03-2007
03-00-2007	03-2007

[1980?]	1980
2007/12/04	2007-12-04
8.xi.03	8-xi-03
02001.July.04 AD 12:08 PM	02001-July-04 AD 12:08 PM
2003. nov. 9.	2003- nov- 9-

Tabla 7.9

- **PeriodDetector:**

Entrada	date	dateFrom	dateTo
1 al 3 de octubre de 2008		1 de octubre de 2008	3 de octubre de 2008
from 2008 to 2009		2008	2009
desde el 2004 hasta el 2005		2004	2005
desde el 4/05/2004 hasta el 4/05/2005		4-05-2004	4-05-2005
desde el 2004-05-4 hasta el 2005-05-04		2004-05-4	2005-05-4
desde el 4 de Mayo de 2004 hasta el 04 de Mayo de 2005		4 de Mayo de 2004	04 de Mayo de 2005
2004-2005		2004	2005
2007/12/04	2007-12-04		

Tabla 7.10

- **DateNormalizer**

Entrada	date	dateFrom	dateTo
1 al 3 de octubre de 2008		2008-10-01	2008-10-03
desde el 4/05/2004 hasta el 4/05/2005		2004-05-04	2005-05-04
desde el 4 de Mayo de 2004 hasta el 04 de Mayo de 2005		2004-05-04	2005-05-04
2007.12.04	2007-12-		

	04		
1997-07-16T19:20+01:00	1997-07-16		
1997-07-16T19:20:30+01:00	1997-07-16		
1997-07-16T19:20:30.45+01:00	1997-07-16		
2001.07.04 AD at 12:08:56 PDT	2001-07-04		
2001.07.04 AD	2001-07-04		
Wed, Jul 4, '01	2001-07-04		
02001.July.04 AD 12:08 PM	2001-07-04		
010704120856-0700	2001-07-04		
2010-12-07T17:53:04Z	2010-12-07		
01 de febrero de 2010	2010-02-01		
Sun, 07 Nov 2004 09:55:41 +0100	2004-11-07		
Jan 12, 1952	1952-01-12		
Lunes, 01 de febrero de 2010	2010-02-01		
Lun, 01 de febrero de 2010	2010-02-01		
Mar 01 de febrero de 2010	2010-02-01		
04/05/1988	1988-05-04		
12/31/2005	2005-12-31		
2003 November 9	2003-11-		

	09		
2003Nov9	2003-11-09		
2003-Nov-09	2003-11-09		
2003-Nov-9, Sunday	2003-11-09		
2003. november 9	2003-11-09		
2003. nov. 9.	2003-11-09		
2003. XI. 9.	2003-11-09		
9 November 2003, 18h 14m 12s	2003-11-09		
2003/11/9/18:14:12	2003-11-09		
2003-11-09T18:14:12	2003-11-09		
The 8th of November 2003	2003-11-08		
8.XI.2003	2003-11-08		
08Nov2003	2003-11-08		

Tabla 7.11

7.10 Aporte

La complejidad de compartir recursos en redes más amplias, la concentración de recursos digitales de información propios y obtenidos por procesos de cosecha, fuerza a una normalización imprescindible para la posterior recuperación de la información por parte de los usuarios. De este panorama es que las mejoras introducidas implican un aporte fundamental para la normalización de la información proveyendo de este modo la posibilidad de: 1) exponerla vía portales para usuarios y a través de protocolos para interoperabilidad, 2) consultarla por medio de técnicas de búsqueda avanzada basadas en criterios semánticos y con estadísticas bibliométricas, y finalmente 3) vincularla adecuadamente, favoreciendo la integridad y preservación de los datos.

7.11 Bibliografía

(7.1) De Giusti, M., Oviedo, N., Lira, J. (2011) Extract, Transform and Load architecture for metadata collection. En Ibero-American Science and Technology Education Consortium (ISTEC): "XVIII Ibero-American Science and Technology Education Consortium General Assembly 2011". ISBN 978-987-595-146-4 <http://sedici.unlp.edu.ar/handle/10915/15948> .
[Consulta: 2012-08-29]

Capítulo 8

Conclusiones y Posibles Trabajos Futuros

Capítulo 8 - CONCLUSIONES Y POSIBLES TRABAJOS FUTUROS

8.1 Introducción

Se presentan estimaciones del impacto de la mejora introducida en el área de los repositorios digitales. Se indican posibles desarrollos futuros a partir de las bases sentadas.

8.2 Conclusiones

En el presente trabajo se han integrado diversas cuestiones que van desde conceptos del área de la bibliotecología hasta el estudio de normas y diferentes herramientas de interoperabilidad. El objetivo principal en este sentido, fue el de crear un marco teórico sobre el cual está inmerso un desarrollo que conlleva a la normalización de la información de un repositorio institucional, en este caso el perteneciente a SeDiCI (Servicio de Difusión de la Creación Intelectual de la UNLP).

Los repositorios institucionales, además de su cometido de almacenamiento, preservación y de las facilidades que adicionan para la recuperación de la información, constituyen una fuente de datos para estudios y estadísticas vinculadas a la producción científica de una institución, que se utilizan por las Secretarías de Ciencia y Técnica y otros organismos científicos superiores.

A pesar de estas aptitudes, la gran cantidad de información proveniente de diversas fuentes dificulta la normalización de los datos y metadatos y con ello se ven obstaculizadas las tareas de recuperación de la información por parte de los usuarios interesados en obtener documentos de un sistema de este tipo. Como se pudo apreciar, entre los problemas que se pueden detectar en la información contenida en un repositorio institucional, se encuentran: registros duplicados, clasificaciones duplicadas, información incompleta, información errónea, entre otros. En este tipo de repositorios, los datos de autores, títulos y temas son particularmente importantes para analizar la producción científica de la institución, por lo que para poder explotar una fuente de este tipo y automatizar la generación de indicadores de producción científica, por ejemplo por área, surge como necesidad que los datos estén normalizados, lo cual implica un análisis exhaustivo de los procesos técnicos vinculados a la catalogación del material.

Para la toma de decisiones, las instituciones deben basarse en la neutralidad y objetividad de los datos, más que en intuiciones, deseos y/o esperanzas. Las decisiones acertadas, se basan en datos objetivos y fiables.

Los métodos y algoritmos introducidos en el presente trabajo mejoran la calidad de los datos y por lo tanto, la calidad de la información obtenida. De este modo, con una buena calidad de información es posible realizar estudios a futuro y obtener avances a corto plazo, permitiendo así nuevas vías de exploración. Al aplicar las transformaciones desarrolladas, se optimiza el uso y se maximiza el aprovechamiento del material con que

cuenta la biblioteca digital, con la garantía de que el tratamiento que se realizó está apegado a reglas y estándares bibliotecarios, así como a las políticas internas del repositorio. Esto conlleva a facilitar los procesos de recuperación de información por parte de los usuarios, aumentando la cantidad de información relevante de las búsquedas, es decir, simplificando la tarea de acceder a la información y detectar qué datos son útiles y cuales no lo son.

Asimismo, el desarrollo aquí propuesto permite optimizar los procesos de intercambio de información, debido a que la misma ahora estará en su mayoría normalizada. En un mundo cada vez más interconectado, y particularmente en un momento en que los repositorios institucionales tienden a abrir y compartir cada vez más su información en búsqueda de una mayor difusión, contar con datos que garanticen cierto nivel de calidad asegura una mejor exposición de la producción científica de la institución.

8.3 Trabajos Futuros

Dado que los procesos de normalización se aplican sobre todos los recursos recolectados, es primordial que los mismos sean implementados de forma eficiente. De lo contrario, es probable que las etapas de transformación y normalización se conviertan en un cuello de botella para el funcionamiento normal de aplicación. Debido a esto es importante considerar la realización de un estudio que permita medir y cuantificar la performance sobre los algoritmos que realizan la ejecución de los filtros, considerando un volumen de registros elevado, a fin de establecer la escalabilidad de la implementación, y determinar en caso de ser necesario, los cambios en las implementaciones, buscando primordialmente disminuir el impacto de estos componentes durante la ejecución de la aplicación de recolección (harvester). Esta cuantificación debe considerar el tiempo de ejecución de las transformaciones así como también los recursos de hardware que las mismas requieren durante ese tiempo, ya que podrían llegar a degradar el normal funcionamiento del repositorio en caso de un excesivo consumo de recursos.

Por otra parte, podrían implementarse nuevos filtros de transformación, por ejemplo, un filtro que incorpore a cierto documento la/s afiliación/es de su/s autor/es a partir de una base de datos de afiliaciones normalizadas. Asimismo, sería provechoso normalizar el dato que contiene el lugar de publicación de un documento, a partir de información tomada de una base de datos de revistas y publicaciones normalizada, como por ejemplo SHERPA-RoMEO. Otro caso que se puede considerar es realizar una normalización avanzada del tipo de documento (dc:type), el cual según el caso posee información sobre la versión o el tipo del ítem. En este sentido es viable desarrollar por ejemplo un nuevo filtro que utilice varios diccionarios de datos aceptando términos traducidos.

Utilizando los filtros desarrollados hasta el momento, podría pensarse en incorporar una herramienta gráfica al harvester que permita la creación de filtros complejos y compuestos utilizando por ejemplo “drag and drop”, de manera de que los mismos puedan combinarse de diferentes maneras cumpliendo distintas funciones dependiendo de cómo son fusionados.

Por último, el presente trabajo conduce hacia un futuro análisis respecto de las necesidades de normalización de datos requeridas a fin de lograr una integración con tecnologías semánticas.